Données industrielles et analyse de survie, comment ne pas se tromper ?

Industrial Data and Survival Analysis: How to Avoid Mistakes?

FIGUEROA Lionel
RTE Réseau de Transport d'Éléctricité
Puteaux
lionel.figueroa@rte-france.com

GUILLON Thomas
RTE Réseau de Transport d'Éléctricité
Puteaux
thomas.guillon@rte-france.com

OUESLATI Abdullah
SNCF Réseau
Saint-Denis
abdullah.oueslati@reseau.sncf.fr

Résumé – L'analyse de survie est une branche des statistiques utilisée pour caractériser le temps avant l'apparition d'un événement unique à partir des données issues du retour d'expérience. Dans le domaine de la fiabilité, cet événement correspond le plus souvent à la défaillance non réparable d'un matériel. L'information du temps avant cette défaillance joue un rôle crucial dans de nombreux domaines industriels, en particulier pour les gestionnaires d'actifs. Ces derniers cherchent à appliquer l'analyse de survie en se basant sur leurs données industrielles issues du leur système d'information opérationnel d'entreprise, notamment les outils de Gestion de Maintenance Assistée par Ordinateur (GMAO). Dans ce contexte, cet article soulève les difficultés les plus courantes rencontrées lors de la constitution d'un échantillon de données de durées de vie et présente des moyens de détecter et corriger certaines erreurs. Chaque étape est illustrée par des cas d'applications sur données réelles et simulées d'actifs électriques haute tension. Une méthodologie itérative, synthétisée dans un logigramme, est proposée afin de s'assurer de la pertinence du jeu de données constitué.

Mots-clefs – analyse de survie, biais d'observation, estimateur de Kaplan-Meier, gestion d'actifs, données de durée de vie, données industrielles, GMAO

Abstract – Survival analysis is a branch of statistics used to characterize the time until the occurrence of a single event based on data from experience feedback. In the field of reliability, this event most often corresponds to the non-repairable failure of equipment. Information about time until failure plays a crucial role in many industrial sectors, especially for asset managers. They seek to apply survival analysis based on their industrial data from the enterprise information system, especially the Computerized Maintenance Management System (CMMS) software. In this context, this article raises the most common difficulties encountered when building a lifetime dataset, presents methods for detecting specific errors, and proposes a flowchart to help analysts ensure the relevance of the created dataset.

Keywords – survival analysis, observation bias, Kaplan-Meier estimator, asset management, lifetime data, industrial data, CMMS

I. Introduction

Les gestionnaires d'infrastructures sont confrontés à des décisions d'investissement cruciales dans un contexte de transition énergétique et d'actifs vieillissants. La dynamique de renouvellement des actifs industriels est liée à leur durée de vie incertaine, souvent représentée par une distribution de probabilités. L'analyse de survie est la branche des statistiques qui permet de caractériser les distributions de durées de vie en tenant compte des biais d'observation inhérents aux durées. Cependant, l'application de l'analyse de survie aux problématiques industrielles requiert, d'une part, une bonne compréhension des données issues des systèmes d'information, notamment de la Gestion de Maintenance Assistée par Ordinateur (GMAO) et, d'autre part, la maitrise des concepts statistiques et des hypothèses de modélisation spécifiques à l'observation des durées.

Au travers de l'expérience acquise sur les actifs électriques haute tension de SNCF Réseau et de RTE, l'objectif de cet article est de présenter les principales difficultés relatives aux points clés de la modélisation des données de durées de vie à partir de données industrielles et met en évidence l'importance de l'analyse de survie pour détecter certaines erreurs et interpréter les résultats.

Cet article s'adresse aux ingénieurs ou analystes du monde industriel souhaitant mener des analyses de survie ou qui ont déjà une expérience élémentaire. La première partie suivante introduit les fondements théoriques de l'analyse de survie. La deuxième partie de l'article est dédiée aux difficultés principales rencontrées et les précautions à prendre lors de l'application de l'analyse de survie aux données industrielles. Cette partie est organisée selon la chronologie des différentes étapes d'une étude reposant sur l'analyse de survie. Certaines recommandations formulées sont illustrées par des résultats obtenus avec les estimateurs statistiques non-paramétriques et paramétriques. En conclusion, un logigramme est proposé pour guider les analystes dans la démarche itérative d'analyse de données de durées vie des matériels.

II. L'ANALYSE DE SURVIE ET SES APPLICATIONS

A. *Introduction à l'analyse de survie*

L'analyse de survie est définie comme un ensemble de procédures statistiques pour laquelle la variable d'intérêt est le temps jusqu'à l'apparition d'un événement (Kleinbaum & Klein, 1996) (Kalbfleisch & Prentice, 2011). Cet événement correspond à un événement unique, comme le décès d'un patient en épidémiologie, ou la défaillance non réparable d'un matériel en fiabilité. Dans ce dernier cas, la variable d'intérêt concerne le temps avant défaillance et cela correspond à la durée entre l'installation du matériel et l'apparition de la défaillance. Cette durée s'exprime généralement en temps calendaires (e.g., années ou mois), mais peut aussi s'exprimer en distance parcourue (e.g., mètres, kilomètres) ou en nombre de cycles d'utilisation en fonction du contexte (e.g., nombre de manœuvres d'un disjoncteur).

Une question classique en fiabilité concerne la caractérisation du temps avant défaillance : sur un grand nombre de matériels similaires, quelle est la proportion de matériels défaillants après 5 ans d'exploitation ? après 10 ans ? et plus généralement après t années ? Pour répondre à ces questions, le temps avant défaillance est décrit à l'aide d'une variable aléatoire T, et nous cherchons à estimer :

$$F(t) = P(T \leq t) = 1 - P(T > t) = 1 - S(t), \tag{1}$$

Avec

- $F(t)$ la fonction de répartition, représentant la probabilité que le temps avant défaillance du matériel soit inférieur ou égal à t, strictement croissante entre 0 et 1 ;
- $S(t)$ la fonction de survie, aussi appelée fonction de fiabilité, représentant la probabilité que le temps avant défaillance du matériel soit supérieur à t, strictement décroissante entre 1 et 0.

L'analyse de survie s'attache donc à estimer la fonction de survie $S(t)$ à partir des données de durées de vie observées sur la population concernée.

B. *Estimateurs de la fonction de survie*

Dans cet article, nous considérons deux estimateurs statistiques de la fonction de survie $S(t)$, dont l'estimation est notée $\hat{S}(t)$ (Lawless, 2011) :

- **Estimateur de Kaplan-Meier (EKM)** : l'EKM est un estimateur non-paramétrique de $S(t)$ à partir des données de durées de vie et ne nécessite pas d'hypothèses sur les modèles. Il est recommandé de commencer l'analyse par ce type de méthode (Meeker, Escobar, & Pascual, 2022) : *« De telles méthodes permettent d'interpréter les données sans distorsion qui pourrait être causée par l'utilisation d'hypothèses de modèle inadéquates. [...] Une analyse non paramétrique constitue une étape intermédiaire vers un modèle plus structuré, permettant des inférences plus précises ou plus étendues, à condition que les hypothèses supplémentaires de ce modèle soient valides » ;*
- **Estimateur du Maximum de Vraisemblance (EMV)** : l'EMV est un estimateur paramétrique permettant d'inférer les paramètres de différentes distributions de probabilité en maximisant leur vraisemblance par rapport aux observations. Les distributions régulièrement utilisées en fiabilité sont l'Exponentielle, Weibull, et Gompertz. L'estimation de ces paramètres permet de calculer la fonction de survie $S(t)$ sur $t \in [0; +\infty[$ et d'en déduire des informations telles que l'espérance de vie, $E[T]$, l'espérance de vie résiduelle sachant que le matériel fonctionne à l'instant t_0, $E[T - t_0 | T > t_0]$. L'estimation paramétrique de $S(t)$ permet de construire d'autres fonctions utiles, telles que la probabilité de défaillance annuelle (2) ou la probabilité de défaillance annuelle conditionnelle (3) :

$$P(t < T \leq t + 1) = S(t) - S(t + 1) \tag{2}$$

$$P(T \leq t + 1 | T > t) = 1 - \frac{S(t + 1)}{S(t)} \tag{3}$$

Dans la suite de cet article, les résultats fournis par les EKM et EMV sont réalisés à partir de la bibliothèque open source Python « ReLife » (https://rte-france.github.io/relife/). Les données de durées de vie utilisées ont été générées avec ReLife, en cohérence avec les résultats obtenus sur certains matériels HT.

C. *Période d'observation et données de durées de vie*

Les estimations de $S(t)$ sont basées sur des données de durées de vie observées sur un échantillon de n matériels. Le statut de l'observation correspond à la collecte de l'information relative à l'apparition ou non de la défaillance du matériel. Ces informations sont généralement consignées dans des bases de données issues d'outils de GMAO. Pour le matériel d'indice $i \in \{1, \ldots, n\}$, la période d'observation correspond à l'intervalle de temps entre le début de l'observation, notée t_{start} et la fin de son observation t_{end_i}. Généralement, t_{start} correspond à la date à partir de laquelle les observations sont collectées et archivées. La durée d'observation t_{obs_i} est égale à $t_{end_i} - t_{start}$. La durée de vie t_i est définie comme la durée entre l'installation ou la mise en service du matériel t_{init_i} et sa défaillance t_{fail_i} : $t_i = t_{fail_i} - t_{init_i}$. Si ce matériel a été observé depuis sa mise en service jusqu'à

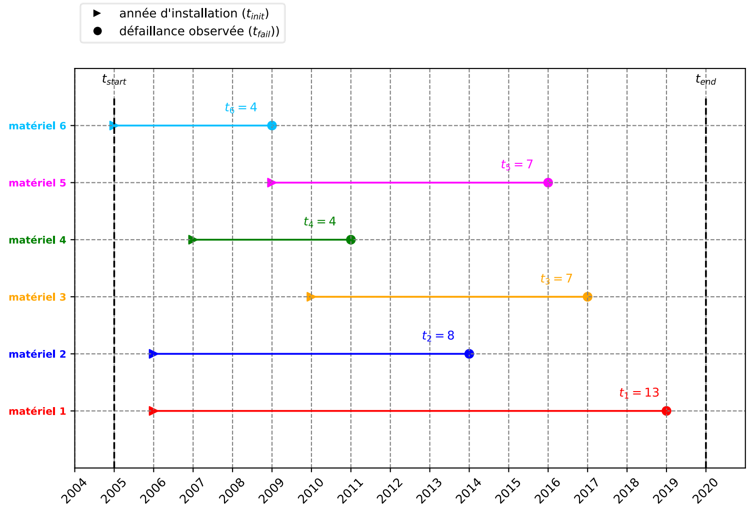

Fig. 1. Schéma d'observation pour six matériels (exemple élémentaire)

TABLE I. JEU DE DONNEES DE DUREES DE VIE POUR SIX MATERIELS (EXEMPLE ELEMENTAIRE)

i	t_i	δ_i	Ensembles
1	13	1	D
2	8	1	D
3	7	1	D
4	4	1	D
5	7	1	D
6	4	1	D

sa défaillance, alors $t_{start} < t_{init_i} < t_{fail_i} \leq t_{end_i}$. Cette information peut être illustrée par un schéma d'observation.

La Fig. 1 présente un exemple élémentaire de schéma d'observation pour six matériels dont les observations correspondent à la situation précédente : la durée de vie de chaque actif a été entièrement observée. Pour faciliter la représentation graphique, la période d'observation est la même pour les six actifs, avec $t_{start} = 2005$. Dans cet article, D représente l'ensemble des indices pour lesquels les défaillances sont observées.

Une donnée de durée de vie correspond à la représentation de l'observation, sous un formalisme spécifique. Nous notons, pour chaque actif, sa durée de vie observée t_i et le statut de son observation δ_i. Par convention, si la défaillance est observée pour le matériel i, alors $\delta_i = 1$. L'ensemble des données de durées de vie observées pour l'échantillon de n matériels est appelé le jeu de données de durées de vie (voir exemple au Tab. 1 correspondant au schéma d'observation de la Fig. 1).

D. Application de l'analyse de survie aux problématiques industrielles

Cette partie présente quelques exemples de problématiques de gestion d'actifs pour lesquelles l'analyse de survie est utilisée :

- **Aide au retour d'expérience** : les industriels peuvent intégrer l'analyse de survie parmi les méthodes utilisées pour réaliser leurs bilans de retour d'expérience, à partir des données d'observation collectées. Ces analyses permettent au gestionnaire de mieux saisir le comportement des différentes populations de matériels, et de les comparer selon différentes caractéristiques (e.g., fournisseurs, technologies). Cette comparaison peut se focaliser sur des indicateurs spécifiques, tels que l'espérance de vie ou la médiane. Ces informations peuvent aussi être utilisées pour s'assurer de la cohérence entre les performances réelles en exploitation et les spécifications de durées de vie ;

- **Établissement des durées d'amortissement comptables** : la connaissance des distributions de durées de vie des différentes populations de matériels permet à l'industriel de mieux déterminer ses durées d'amortissement comptables et contribue donc à favoriser l'alignement des décisions entre les fonctions financières et non financières, en cohérence avec les recommandations de la norme ISO 55010 (International Organization for Standardization, 2019) ;

- **Calcul de l'âge de remplacement préventif optimal** : si l'industriel s'intéresse au remplacement préventif par âge d'une population de matériels, la question centrale concerne le choix de l'âge de remplacement préventif optimal, c'est-à-dire celui qui minimise une certaine fonction de coût. Celle-ci peut tenir compte de facteurs économiques tels que l'actualisation ou l'inflation mais aussi du coût unitaire du remplacement préventif, des coûts réels et sociétaux liées aux conséquences des défaillances au regard des objectifs de gestion d'actifs (International Organization for Standardization, 2014), et de la probabilité d'apparition de cette défaillance. L'ingénierie économique et la fiabilité permettent de construire ce problème d'optimisation et de le résoudre ;

- **Évaluation socio-économique :** cette évaluation quantitative permet la comparaison socio-économique de différentes options et participe à la justification des dépenses d'investissement auprès des décisionnaires et des régulateurs. Par exemple, dans le cas de la justification d'un plan de gestion d'actifs d'une population de matériel, cela permet de comparer plusieurs options de réduction des risques, telles que le remplacement sur défaillance et le remplacement préventif à l'âge optimal, et conduit à l'estimation des ressources moyennes à prévoir pour le déploiement le chaque option : coûts réels et sociétaux, main d'œuvre, matériels à approvisionner en maintenances préventives et correctives.

III. Application aux donnees industrielles

A. *Nature des défaillances*

L'analyse de survie s'intéresse à la modélisation d'événements uniques. En fiabilité, cela correspond le plus souvent aux défaillances non réparables entraînant le remplacement du matériel ou dont la réparation est qualifiée de *« as good as new »*. D’autres modèles de fiabilité sont utilisés pour représenter des événements récurrents, par exemple les processus non-homogène de Poisson sont adaptés aux défaillances réparables mais dont la réparation est qualifiée de *« as bad as old »* car elle n’améliore pas l’état de l’actif. La norme IEC 61703 (International Electrotechnical Commission, 2016) distingue le caractère non réparable ou réparable du matériel à la suite de l’apparition d’une défaillance :

- **Réparable** : dans des conditions données après une défaillance, le matériel peut être remis dans un état lui permettant de fonctionner tel que requis [IEV 192-01-11] ;

- **Non réparable** : dans des conditions données après une défaillance, le matériel ne peut pas être remis dans un état lui permettant de fonctionner tel que requis [IEV 192-01-12]. Dans ce cas, le choix est généralement fait de remplacer le matériel.

Cependant, comme noté dans la norme IEC 61703, il faut bien distinguer le caractère réparable de la défaillance de l'action corrective réalisée. En effet, dans le cas de l'apparition d'une défaillance réparable, l'industriel peut faire le choix de remplacer le matériel pour diverses raisons (e.g., absence de gestion d’un stock de pièces de réparation, pas de maintien de compétences spécialisées). Dans ce cas, si le problème de décision nécessite d’estimer la probabilité d'occurrence d’événements entraînant le remplacement du matériel, alors cette défaillance de nature réparable doit être intégrée à l’échantillon. Dans la suite, ces événements sont qualifiés d’événements uniques.

Pour commencer l’étude, l’analyste peut vérifier que chaque identifiant de matériel est unique dans son jeu de données. L'inverse pourrait correspondre à des observations de défaillances réparables sur un même matériel. Dans le cas où l’analyste n'est pas certain de la nature de certaines défaillances enregistrées, il peut décider d'engager des ressources pour collecter et corriger les informations nécessaires en allant interroger, par exemple, les équipes de maintenance en charge des défaillances concernées ou en considérant différentes sources de données disponibles (e.g., rapports d'avarie, bases de données comptables). La valeur ajoutée de ce travail est démontrée en comparant les coûts prévisionnels des options reposant sur les deux hypothèses suivantes : la première suppose que toutes les données incertaines sont des défaillances irréparables et sont donc intégrées au jeu de données, tandis que la seconde suppose que ces données sont toutes des défaillances réparables et ne sont donc pas intégrées au jeu de données.

B. *Données de durée de vie et biais d'observation*

1) *Prise en compte des observations censurées à droite*

La section précédente a permis de s’assurer que le problème de décision nécessitait bien de s’attacher à la question du temps avant l’apparition d’un événement unique. Il est alors nécessaire de déterminer les sources de données industrielles nécessaires à la constitution de l’échantillon.

Dans la grande majorité des cas, l'industriel n'observe pas l’événement unique sur chaque matériel de son échantillon. La plupart des matériels sont toujours en fonctionnement à la fin de leur période d'observation : l'observation est censurée à droite et l’ensemble CD correspond aux indices de ces matériels. Ainsi, pour $i \in CD, t_{fail_i} > t_{end_i}$. Ces observations apportent une information importante : la durée de vie est supérieure à l'âge obtenu à la fin de l’observation : $t_i > t_{end_i} - t_{init_i}$, et par convention $\delta_i = 0$.

Les observations censurées à droite correspondent généralement aux matériels toujours en fonctionnement au moment de l’étude. Néanmoins, d'autres matériels sont à considérer dans l’ensemble CD : ce sont les matériels remplacés préventivement dans le cadre d'une stratégie de renouvellement ou d'une dépose de matériel (e.g., pour cause de restructuration de réseau). Dans certains Système d’Information d’entreprise, il est difficile de récolter l’information du remplacement préventif des actifs. La Fig. 2 et le Tab. 2 présentent le schéma d’observation et le jeu de données associé, basés sur un exemple d’observation plus réalistes que celui de la Fig. 1. Les matériels 1 et 5 sont toujours en fonctionnement à la fin de leur observation, $CD = \{1; 5\}$ et pour $i \in CD, \delta_i = 0$.

Si l’analyste considère uniquement les défaillances observées, et n'intègre donc pas l'ensemble des informations apportées par les matériels toujours en fonctionnement : c’est le *biais de mortalité*. Ce dernier conduit à une surestimation systématique des probabilités de défaillance pour un âge donné. Cette erreur entraîne des surinvestissements lors de la mise en place de stratégies de remplacements préventifs, une mauvaise estimation des nombres des défaillances à venir et des ressources associées. Il est donc nécessaire de tenir compte de ces informations lors de l'analyse de survie. Ce biais de mortalité est illustré sur la Fig. 3. Lorsque l’analyste considère uniquement les défaillances dans son échantillon (courbe orange), les résultats obtenus sont pessimistes par rapport à l’analyse tenant compte des biais d’observation (courbe bleue). Dans cet exemple, l’espérance de vie estimée à l’aide de l’EMV passe de 38 ans (analyse correcte) à 16 ans (biais de mortalité).

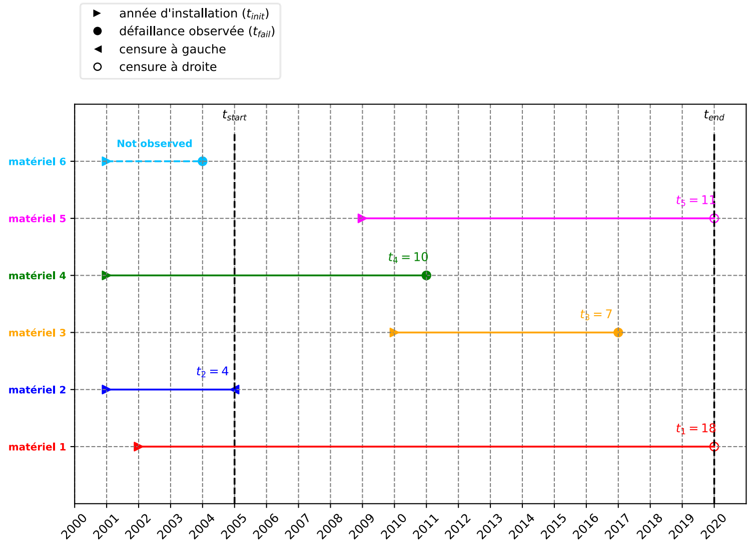

Fig. 2. Schéma d'observation pour six matériels (exemple réaliste)

TABLE II. JEU DE DONNEES DE DUREES DE VIE POUR SIX MATERIELS (EXEMPLE REALISTE)

i	t_i	δ_i	e_i	**Ensembles**
1	18	0	3	$TG \cap CD$
2	4	2	0	CG
3	7	1	0	D
4	10	1	4	$TG \cap D$
5	11	0	0	CD

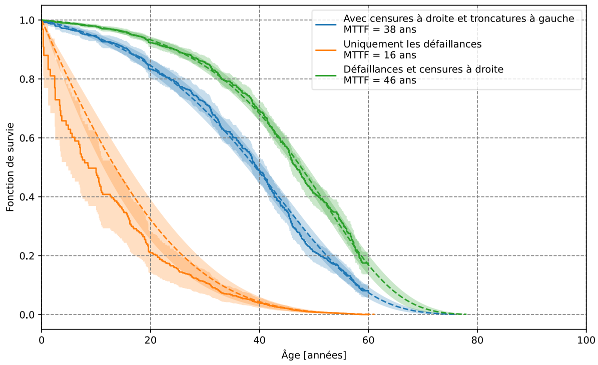

Fig. 3. Résultats des estimateurs selon la prise en compte des biais d'observation

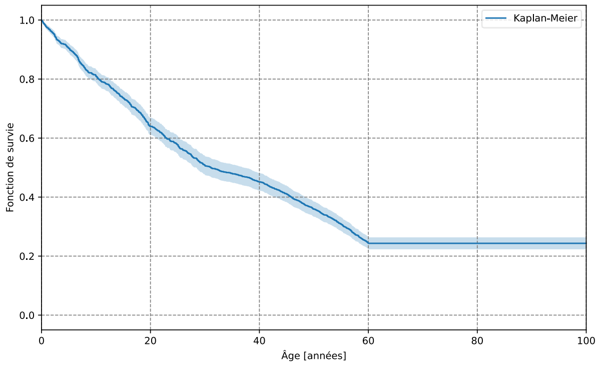

Fig. 4. Exemple d'EKM basé sur un échantillon contenant une durée observée aberrante

2) Prise en compte des observations censurées à gauche

Dans certains cas, l'industriel possède des informations sur des défaillances de matériels précédant le début de la période d'observation : l'observation est censurée à gauche. L'ensemble CG correspond aux indices de ces matériels. Ainsi, pour $i \in CG, t_{fail_i} < t_{start}$. Ces observations apportent une indication supplémentaire : $t_i < t_{start} - t_{init_i}$. Par convention, cette donnée de durée de vie est notée $t_i = t_{start} - t_{init_i}$ et $\delta_i = 2$.

Cette situation peut arriver lorsqu'une nouvelle base de données de collecte des observations est mise en place à t_{start}. Dans ce cas, des informations relatives aux défaillances antérieures peuvent parfois être intégrées à cette nouvelle base de données, mais l'âge exact de défaillance n'est pas conservé. Dans l'exemple présenté sur la Fig. 2, c'est le cas du matériel 2 ($CG = \{2\}$) : sa défaillance a eu lieu entre $t_{init_2} = 2001$ et $t_{start} = 2005$, donc $t_2 < 4$ et $\delta_2 = 2$ (voir Tab. 2).

3) Prise en compte des observations tronquées à gauche

Parfois, le début de la période d'observation des défaillances est postérieur à l'année d'installation de certains matériels. Dans ce cas l'observation est conditionnée par le fait que celui-ci soit toujours en fonctionnement au début de la période d'observation : l'observation est tronquée à gauche. L'ensemble TG correspond aux indices de ces matériels. Ainsi, pour $i \in TG$, $t_{init_i} < t_{start}$ et $t_{fail_i} > t_{start}$. Dans le jeu de données, une information d'entrée est ajoutée, notée e_i, correspondant à l'âge du matériel au début de la période d'observation : $e_i = t_{start} - t_{init_i}$. Si $j \notin TG$, alors $e_j = 0$. Sur l'exemple présenté sur la Fig. 2, $TG = \{1; 4\}$, $e_1 = 3$ et $e_4 = 4$.

Ce biais d'échantillonnage peut être très important pour les industriels dont les matériels ont été massivement installés longtemps avant la mise en place de leurs bases de données. Le *biais du survivant* consiste à ne considérer que les matériels toujours en fonctionnement au début de la période d'observation, en oubliant que leur observabilité est conditionnelle au fait d'avoir survécu suffisamment longtemps. Ce biais et a donc tendance à ne pas considérer les matériels de durées de vie plus courtes, ce qui conduit à sous-estimer la probabilité de défaillance. Ainsi, pour deux matériels installés au même instant, avant t_{start}, le premier peut avoir été observé (e.g., Matériel 4, $t_{fail_4} > t_{start}$) alors que le second, de durée de vie plus courte, n'aura jamais été observé et n'est donc pas présent dans l'échantillon (e.g., Matériel 6).

La prise en compte de la troncature à gauche dans les estimateurs réduit le biais du survivant grâce à l'information apportée par e_i. Ce biais du survivant est illustré sur la Fig. 3. Lorsque l'analyste ne tient pas compte des troncatures à gauche dans son échantillon (courbe verte), les résultats obtenus sont optimistes par rapport l'analyse tenant compte des biais d'observation (courbe bleue). Dans cet exemple, l'espérance de vie estimée à l'aide de l'EMV passe de 38 ans (analyse correcte) à 46 ans (biais du survivant).

C. Informations apportées par l'estimateur non paramétrique

1) Présence de données incohérentes

Lors de l'estimation de la fonction de survie par l'EKM, l'analyste peut se rendre compte de données incohérentes dans le jeu de données utilisé :

- **Erreurs ou avertissements renvoyés par l'algorithme** : les outils d'analyse de survie afficheront des erreurs ou avertissements lorsque les données sont incohérentes : durées de vie nulles ou négatives ($t_i \leq 0$), ou entrée supérieure à la valeur de la durée observée ($e_i \geq t_i$). L'origine du problème peut venir d'une mauvaise application des règles de construction du jeu de données à partir des bases de données industrielles ou de défauts de saisie dans l'application de GMAO, de recensement du patrimoine ou du retour d'expérience ;

- **Présence de valeurs aberrantes** : lorsque $\hat{S}(t)$ est constante sur de longues périodes (plateau), cela signifie qu'il n'y a aucune nouvelle information de durée de vie entre le début et la fin du plateau. Ce résultat doit être questionné, car il peut résulter de la présence de valeurs aberrantes de durées de vie observées. La Fig. 4 illustre cette situation avec un matériel k toujours en fonctionnement à l'âge de 100 ans ($t_k = 100$ ans et $\delta_k = 0$), alors qu'aucune défaillance n'a été observée entre 60 ans et 100 ans. Cela peut provenir d'erreurs dans les données industrielles relatives aux dates d'installation. C'est le cas lorsque le choix est fait de considérer une valeur par défaut (e.g., $t_{init_i} = 1900$), lorsque l'information n'est pas disponible. Il est alors nécessaire de statuer sur la validité de cette valeur avec les experts concernés. Par exemple, la question suivante peut se poser : la technologie étudiée était-elle disponible à cette époque-là ? Si non, il est nécessaire de corriger l'information ou de la retirer de l'échantillon. À noter que la conservation d'une valeur aberrante peut mener à une incohérence graphique entre les estimateurs paramétriques et non paramétriques ;

- **Présence de défaillances précoces** : une autre erreur similaire au point précédent, concerne les défaillances très précoces par exemple inférieures à 1 an qui doivent également être vérifiées par l'analyste et l'expert. Il est possible que ces défaillances correspondent à des défaillances réparables ou d'un défaut fugitif qui ne doivent pas être considérés dans l'échantillon.

2) Mélange de populations

Habituellement, l'estimation de la fonction de survie par l'EKM à partir d'un jeu de données portant sur un échantillon homogène comporte un seul point d'inflexion. L'existence de plusieurs points d'inflexion très prononcés doit questionner l'analyste sur la présence d'un mélange de populations dont les comportements sont distincts. Cette différence de comportement peut être lié :

- **Aux types de matériels** : il est nécessaire de vérifier la cohérence des types de matériels considérés dans l'échantillon. Par exemple, il existe plusieurs types de transformateurs de mesure (TdM) (e.g., transformateurs de courant, inductifs de tension, capacitifs de tension, combinés) qui présentent des caractéristiques différentes, tant au niveau des fonctions que des technologies utilisées. Il serait incohérent de considérer l'ensemble des TdM comme un échantillon homogène. De même, pour l'étude de disjoncteurs Haute Tension ayant la même fonction dans une sous-station d'alimentation, il convient de distinguer les disjoncteurs à coupure dans l'huile des disjoncteurs à coupure dans le gaz. En première approche, l'analyste et l'expert peuvent convenir de regroupements d'échantillons de populations homogènes à considérer. Ce cas est illustré sur la Fig. 5 : la fonction de survie estimée par l'EKM basée sur l'échantillon initial, présente deux points d'inflexion. À partir d'une analyse par les experts et les analystes, cet échantillon a été séparé en deux sous-échantillons A et B, distincts en nombre n et en espérance $E(T)$: $n_A = 1514$, $E_A(T) = 38$ ans et $n_B = 1360$, $E_B(T) = 63$ ans. Les fonctions de survie de ces deux échantillons A et B sont représentées sur la Fig. 6 ;

- **Aux conditions externes** : ces conditions peuvent différer assez largement au sein d'une même population. Elles peuvent être d'ordre environnementales (e.g., polluants, températures ambiantes, humidité) ou d'exploitation (e.g., utilisation du matériel aux limites). Dans ce cas, il est nécessaire de déterminer les facteurs explicatifs du vieillissement (e.g., types de polluants, unité de mesure) et de les prendre en compte sous forme de covariables dans des modèles de régression paramétriques, tels que les modèles *Accelerated Failure Time* (AFT) ou *Proportional Hazard Model*. Chez RTE, c'est le cas des chaînes d'isolateurs : leur vieillissement est accéléré lorsqu'elles sont soumises à des concentrations élevées de polluants (e.g., HCl, H_2SO_4) généralement présentes dans des zones industrielles ou des régions côtières. Ainsi, le résultat obtenu avec l'EKM présente plusieurs points d'inflexion (Fig. 7). La Fig. 8 présente l'estimation de la probabilité de survie en fonction de l'âge de la chaîne d'isolateurs, selon qu'elle soit située dans les conditions environnementales les plus défavorables (courbe rouge) ou les plus favorables (courbe verte).

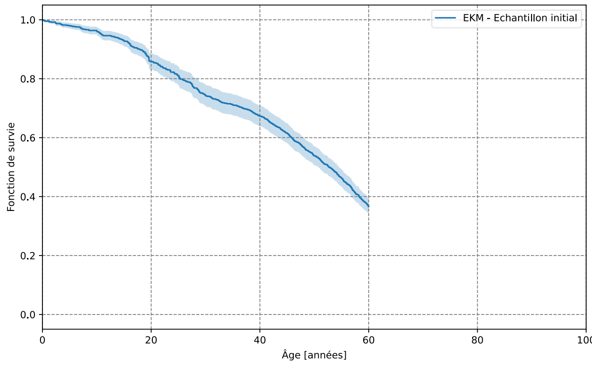

Fig. 5. Résultat d'un EKM avec deux points d'inflexion

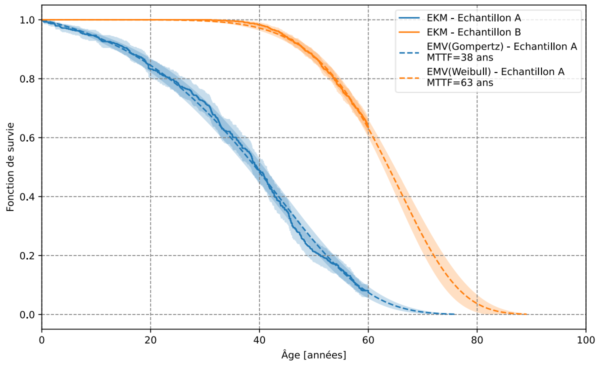

Fig. 6. Résultats obtenus après séparation en deux échantillons distincts

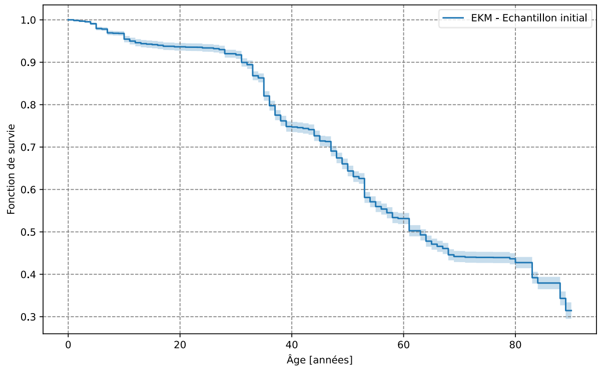

Fig. 7. Résultat d'un EKM avec deux points d'inflexion (chaînes d'isolateurs)

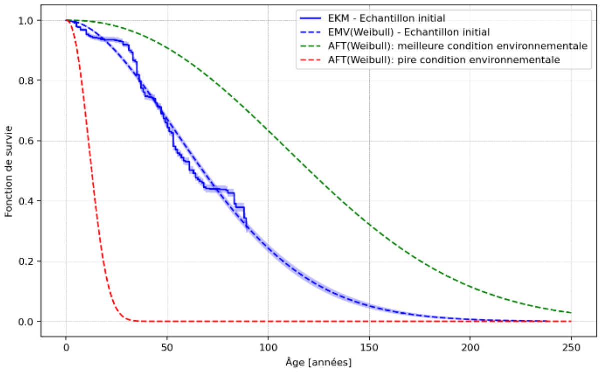

Fig. 8. Résultats obtenus en utilisant un modèle AFT pour tenir en compte des covariables environnementales (chaînes d'isolateurs)

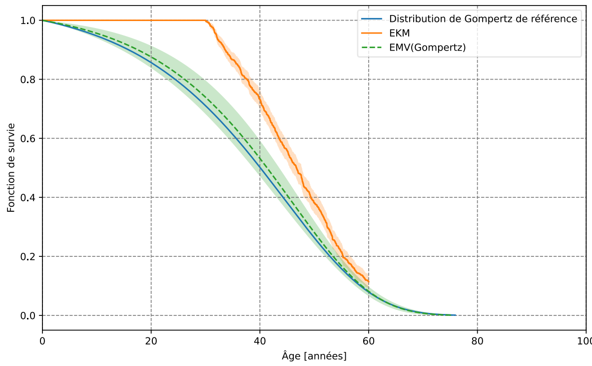

Fig. 9. Illustration du biais de l'EKM pour des données fortement tronquées à gauche

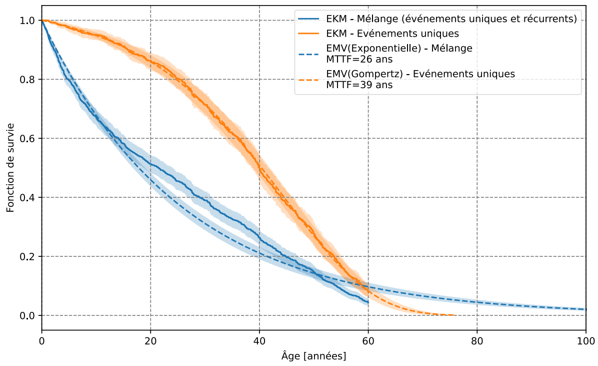

Fig. 10. Impact de la prise en compte d'événements récurrents sur le modèle paramétrique

3) *Données fortement tronquées à gauche*

Pour les gestionnaires d'infrastructures vieillissantes, avec des matériels ayant de longues durées de vie (tels que les appareils haute tension), il arrive que certains ne soient ni produits ni installés depuis plusieurs années. De plus si les systèmes d'informations permettant de collecter les défaillances ont été implémentés plus récemment, les équipements concernés sont fortement tronqués à gauche. Dans ce cas, il faut être particulièrement prudent avec l'interprétation de l'EKM qui peut être fortement biaisé. Par exemple, si l'équipement n'a pas été installé pendant une période de 30 ans précédant le début de la collecte des observations, alors aucune information sur les défaillances entre 0 à 30 ans n'est intégrée au calcul de l'EKM, et sa valeur est constante et égale à 1 sur cette période. À partir de données générées selon une distribution de Gompertz de référence (courbe bleue) et tronquées entre 0 et 30 ans, la Fig. 9 illustre le biais de l'EKM (courbe orange), alors que l'EMV, pour la distribution de Gompertz (courbe verte), s'ajuste de manière très satisfaisante.

D. *Informations apportées par l'estimateur paramétrique*

Il arrive que certaines données relatives aux défaillances correspondent à des événements d'exploitation (e.g., disjonction), sans savoir si le matériel a été réparé ou remplacé. Ainsi, malgré les précautions prises au III.A, des événements uniques et des événements récurrents peuvent être mélangés dans l'échantillon. Dans ce cas, l'âge aura un impact faible sur la calibration d'une distribution de durée de vie. Cette caractéristique se traduit par une bonne adéquation avec la distribution Exponentielle. Pour rappel, cette distribution est caractéristique d'une absence de vieillissement : la fonction du taux de défaillance $h(t)$ est constante en fonction de l'âge t, ou dit autrement, la probabilité de défaillance conditionnelle dans l'intervalle $]t, t + \Delta t]$, $P(T \leq t + \Delta t | T > t)$, est constante $\forall t > 0$. Cette dérive vers une distribution Exponentielle, conduit, par exemple, à rendre l'option du remplacement préventif inefficace pour un nombre infini de cycles, et contribue à privilégier l'option du remplacement sur avarie. En effet, le remplacement d'un matériel âgé par un neuf n'est alors jamais rentable, puisque le nouveau conserve la même probabilité de défaillance conditionnelle que l'ancien. Ainsi, une mauvaise sélection des événements observés peut mener à une mauvaise estimation de la fonction de survie et donc à de mauvaises décisions de gestion d'actifs. Lorsque le résultat de l'EMV est favorable au modèle Exponentielle, l'analyste doit s'interroger sur la présence d'événements récurrents dans l'échantillon.

La Fig. 10 illustre cette situation : lorsque des événements récurrents sont mélangés aux événements uniques dans l'échantillon, le résultat de l'EMV est conforme à une distribution Exponentielle (courbe bleue), dont l'espérance de vie est de 26 ans. Lorsque ces événements récurrents sont retirés de l'échantillon, le résultat de l'EMV est très différent : la distribution la plus adaptée est Gompertz, avec une espérance de vie de 39 ans, et sa la concavité est caractéristique d'un phénomène de vieillissement.

IV. CONCLUSION

L'analyse de survie est un ensemble de méthodes statistiques permettant de caractériser les durées de vie des matériels. Cet article a introduit les outils élémentaires nécessaires à l'analyse de survie : définitions probabilistes, estimateurs statistiques, notations des différents temps et durées modélisés, principaux champs d'application de l'analyse de survie. L'article a mis en évidence l'importance de la bonne qualification :

- De la nature de l'événement étudié ;
- Du schéma d'observation ;
- De la population représentée dans l'échantillon.

Une erreur ou un oubli de l'un de ces aspects peut mener à introduire :

- Un *biais de mortalité* conduisant à sous-estimer la fiabilité des actifs analysés ;
- Un *biais du survivant* conduisant à surestimer leur fiabilité ;
- Une adéquation erronée à la distribution Exponentielle conduisant à négliger à tort le vieillissement des actifs.

En outre, de mauvaises interprétations peuvent également résulter de défauts de qualité des données sources issues du Système d'Information opérationnel d'entreprise.

L'article a proposé une méthodologie pour détecter et rectifier les points cités précédemment, synthétisée dans le logigramme de la Fig. 11. Notons que ces recommandations ne sont pas exhaustives : les spécificités des bases de données (e.g., structures, interactions entre bases, méthodes de collecte, qualité des données) peuvent générer d'autres traitements particuliers. De plus, d'autres outils de l'analyse de survie et l'expérience de l'analyste sont nécessaires pour approfondir l'analyse de survie (e.g., calibration de plusieurs distributions, analyse des intervalles de confiance, sélection du modèle).

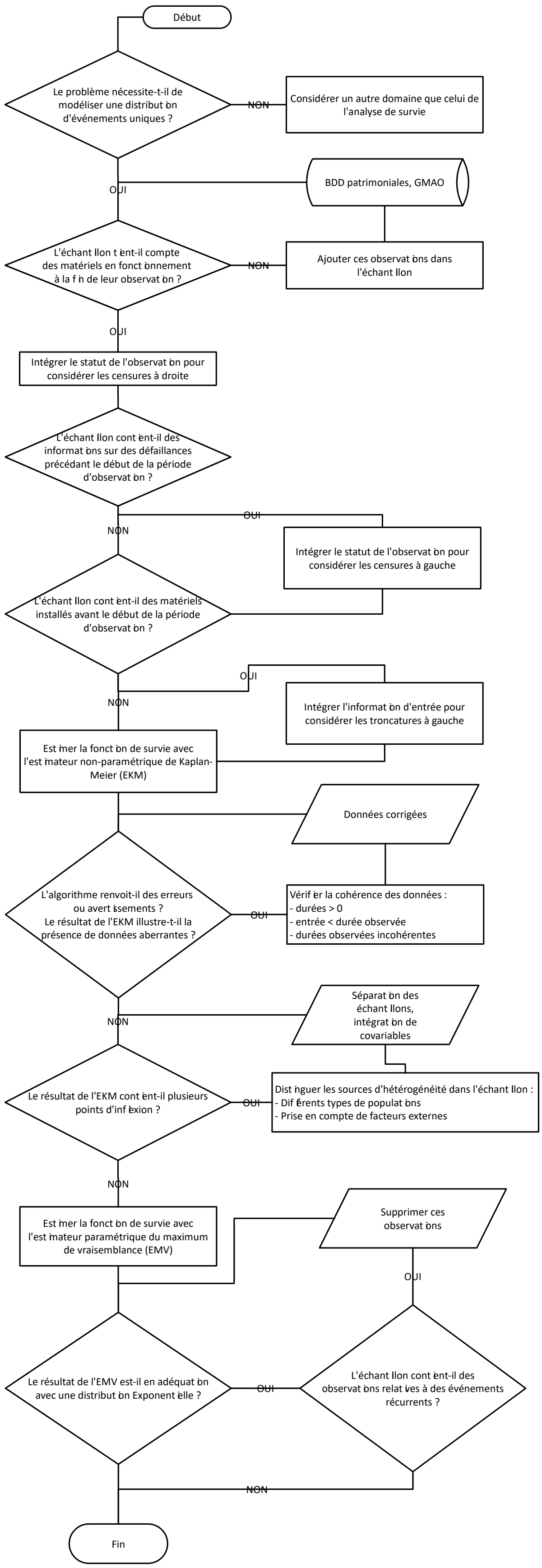

Fig. 11. Logigramme d'aide à la réalisation d'une analyse de survie à partir de données industrelles

V. References

International Electrotechnical Commission. (2016). *Mathematical expressions for reliability, availability, maintainability and maintenance support terms (IEC standard No. 61703:2016).* IEC. Récupéré sur https://webstore.iec.ch/publication/25646

International Organization for Standardization. (2014). *Asset management - Management systems - Requirements (ISO Standard No. 55001:2014).* Récupéré sur https://www.iso.org/standard/55089.html

International Organization for Standardization. (2019). *Asset management - Guidance on the alignment of financial and non-financial functions in asset management (ISO/TS No. 55010:2019).* ISO. Récupéré sur https://committee.iso.org/sites/tc251/home/projects/published/isots-55010.html

Kalbfleisch, J., & Prentice, R. (2011). *The statistical analysis of failure time data.* John Wiley & Sons.

Kleinbaum, D. G., & Klein, M. (1996). *Survival analysis a self-learning text.*

Lawless, J. (2011). *Statistical models and methods for lifetime data.* John Wiley & Sons.

Meeker, W., Escobar, L., & Pascual, F. (2022). *Statistical methods for reliability data.*