

N° d'ordre : 3657



THÈSE

PRÉSENTÉE A

L'UNIVERSITÉ BORDEAUX 1

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUE

Par Julie BERTHON

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : MATHÉMATIQUES APPLIQUÉES

NOUVELLE APPROCHE DE LA FIABILITÉ OPÉRATIONNELLE

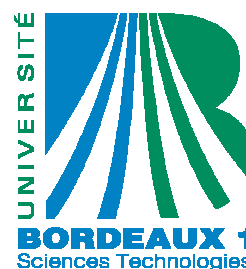
Directeur de recherche : M. Jean-Marc DESHOUILLERS, Professeur à l'Université Bordeaux II

Soutenue le : 7 novembre 2008

Devant la commission d'examen formée de :

M. Jaromir ANTOCH	Professeur, Université Charles de Prague	Examineur
M. Bernard BERCU	Professeur, Université Bordeaux I	Examineur
M. Stéphane CHARRUAU	Expert fiabilité, THALES AVIONICS	Examineur
M. Jean-Marc DESHOUILLERS	Professeur, Université Bordeaux II	Directeur de thèse
M. Yves DUTUIT	Professeur, Université Bordeaux I	Co-directeur de thèse
M. Olivier GAUDOIN	Professeur, Institut Polytechnique de Grenoble	Rapporteur
M. Fabrice GUERIN	Professeur, Université d'Angers	Rapporteur
Mme Hélène MISSON	Expert Sûreté de fonctionnement, THALES AVIONICS	Examineur

N° d'ordre : 3657



THÈSE

PRÉSENTÉE A

L'UNIVERSITÉ BORDEAUX 1

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUE

Par Julie BERTHON

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : MATHÉMATIQUES APPLIQUÉES

NOUVELLE APPROCHE DE LA FIABILITÉ OPÉRATIONNELLE

Directeur de recherche : M. Jean-Marc DESHOUILLERS, Professeur à l'Université Bordeaux II

Soutenue le : 7 novembre 2008

Devant la commission d'examen formée de :

M. Jaromir ANTOCH	Professeur, Université Charles de Prague	Examineur
M. Bernard BERCU	Professeur, Université Bordeaux I	Examineur
M. Stéphane CHARRUAU	Expert fiabilité, THALES AVIONICS	Examineur
M. Jean-Marc DESHOUILLERS	Professeur, Université Bordeaux II	Directeur de thèse
M. Yves DUTUIT	Professeur, Université Bordeaux I	Co-directeur de thèse
M. Olivier GAUDOIN	Professeur, Institut Polytechnique de Grenoble	Rapporteur
M. Fabrice GUERIN	Professeur, Université d'Angers	Rapporteur
Mme Hélène MISSON	Expert Sûreté de fonctionnement, THALES AVIONICS	Examineur

REMERCIEMENTS

Trois années viennent de s'écouler... Une expérience scientifique étonnante, une expérience humaine particulièrement enrichissante.

Je tiens ici à remercier toutes celles et ceux qui ont, de près ou de loin, participé à cette aventure.

J'exprime mes remerciements les plus sincères à Monsieur le Professeur Jean-Marc Deshouillers, qui fut pour moi un directeur de thèse formidable. Je lui suis extrêmement reconnaissante pour sa grande disponibilité et pour l'attention qu'il a bien voulu m'accorder tout au long de ma thèse. Je veillerai à conserver la rigueur scientifique qu'il a tentée de me transmettre et à appliquer les nombreux conseils qu'il m'a prodigués. Je conserverai, c'est certain, de nombreux souvenirs... Certaine soirée sardines-pommes de terres au syndicat des pêcheurs d'Estoril ; certaine séance shopping dans les rues de Hong-Kong ; un exemplaire unique de La Disparition ; j'en passe...

Je remercie celui qui a éveillé mon intérêt pour la fiabilité : Monsieur le Professeur Yves Dutuit. Ce fut d'abord un professeur passionnant, expert dans l'art de présenter tout sujet d'une manière à la fois ludique et rigoureuse ; ce fut ensuite un « co-directeur » exemplaire, qui a consacré beaucoup de temps et d'énergie à mon travail. J'espère être à la hauteur de tout ce qu'il m'a appris.

Merci à vous deux, Jean-Marc et Yves, pour m'avoir si bien accompagnée durant ces trois années. Plus qu'un honneur, ce fut un plaisir de travailler avec vous. J'espère que notre collaboration, et notre amitié, ne s'arrêteront pas là...

Merci à Stéphane Charruau ; la confiance qu'il m'a accordée pendant ces trois années m'a permis de progresser rapidement. Il est l'instigateur de la « méthode convolutive » qui m'a tant résisté... Qu'il soit assuré de toute mon estime et de ma reconnaissance.

Je remercie Messieurs les Professeurs Olivier Gaudoin et Fabrice Guerin de me faire l'honneur d'être les rapporteurs de ma thèse. Je les remercie pour leurs conseils et leurs suggestions, qui ont permis l'amélioration de ce manuscrit. Le regard critique et avisé qu'ils ont porté sur mes travaux ne peut que m'encourager à être encore plus perspicace et engagée dans mes recherches.

Que Monsieur le Professeur Jaromir Antoch soit remercié pour l'intérêt qu'il a porté à mon travail et pour son implication depuis le début de ma thèse ; les discussions que nous avons eues lors de mes séjours à l'Université Charles de Prague m'ont été d'une aide précieuse.

Je remercie Hélène Misson d'avoir accepté d'examiner mon travail. J'espère que notre collaboration future sera brillante !

Je remercie Bernard Bercu d'avoir accepté de faire partie de mon jury.

J'exprime ma plus profonde sympathie aux personnes du service SIH, qui m'ont gentiment soutenue et encouragée ces derniers jours, ainsi qu'à tous les « thalésiens » qui ont participé à

cette aventure, en m'apportant aide technique ou soutien. Je ne les cite pas car ils sont nombreux...

Je remercie Messieurs Christophe Labille et Denis Bonnet, respectivement mon « ancien » et mon « actuel » chef à Thales Avionics, pour la confiance qu'ils ont bien voulu m'accorder et qu'ils continuent à m'accorder aujourd'hui.

Merci à Elodie, Alexandra, Elisa et Marie.

Un très grand merci à mes parents, à Sylvain, Céline, Nathalie, Camille et Marie, qui n'ont cessé de m'encourager pendant ces trois années.

Et un merci tout particulier à Vincent, qui me soutient tous les jours, et encore plus patiemment depuis quelques semaines...

TABLE DES MATIERES

REMERCIEMENTS.....	4
LISTE DES FIGURES	8
LISTE DES TABLEAUX	10
RESUME	12
ABSTRACT.....	14
Chapitre 1.....	16
I. Introduction.....	17
II. Difficulté du problème	18
III. Formulation mathématique.....	20
III.1. Définitions	20
III.2. Deux modèles typiques.....	21
IV. Un exemple simple.....	22
IV.1. Approche combinatoire	22
IV.2. Formule de Poincaré	24
IV.3. Approche basée sur le formalisme des Diagrammes de Décision Binaires	27
<i>IV.3.a. Une brève présentation des DDB.....</i>	<i>27</i>
<i>IV.3.b. Application au calcul de la probabilité d'occurrence d'un « cluster »</i>	<i>29</i>
IV.4. Approche markovienne.....	33
IV.5. Conclusion.....	36
V. Balayage discret ou continu : modèles usuels d'intérêt pratique	36
Chapitre 2.....	40
I. Résolution par simulation numérique.....	41
I.1. Simulation de Monte-Carlo directe	41
<i>I.1.a. Principe</i>	<i>41</i>
<i>I.1.b. Application au cas étudié</i>	<i>42</i>
<i>I.1.c. Modèle de Bernoulli</i>	<i>42</i>
<i>I.1.d. Modèle exponentiel renouvelé.....</i>	<i>45</i>
<i>I.1.e. Résultats de la simulation de Monte Carlo classique</i>	<i>46</i>
I.2. Réseau de Petri animé par simulation de Monte Carlo.....	47
II. Les méthodes markoviennes	48
II.1. Première approche	50
<i>II.1.a. Modèle de Bernoulli</i>	<i>51</i>
<i>II.1.b. Modèle de Bernoulli : application à l'exemple ($T=365, w=10, k=3$)</i>	<i>55</i>
<i>II.1.c. Modèle de Poisson discret.....</i>	<i>57</i>
<i>II.1.d. Modèle de Poisson discret : application à l'exemple ($T=365, w=10, k=3$)</i>	<i>62</i>
II.2. Modèle markovien à double fenêtre de balayage	64
<i>II.2.a. Déclinaison au modèle de Bernoulli</i>	<i>67</i>
<i>II.2.b. Modèle de Bernoulli : application à l'exemple ($T=365, w=10, k=3$)</i>	<i>70</i>
II.3. Modèle markovien complet	72
<i>II.3.a. Modèle de Bernoulli</i>	<i>73</i>
<i>II.3.b. Modèle de Bernoulli : application à l'exemple ($T=365, w=10, k=3$)</i>	<i>77</i>
<i>II.3.c. Modèle de Poisson discret.....</i>	<i>79</i>
<i>II.3.d. Modèle de Poisson discret : application à l'exemple ($T=365, w=10, k=3$)</i>	<i>81</i>

III. Récapitulatif.....	83
ANNEXE : SIMULATION DE MONTE-CARLO SOUS VISUAL BASIC.....	84
Chapitre 3.....	88
I. Fiabilité : la problématique industrielle	89
II. Définitions.....	90
II.1. Fiabilité.....	90
II.2. Espérance et variance	92
III. Lois de probabilité usuelles en fiabilité	93
III.1. La loi exponentielle.....	93
III.2. Loi de Weibull.....	94
IV. Les méthodes statistiques d'exploitation du Retour d'EXpérience.....	96
IV.1. Le problème des données censurées	96
IV.2. Méthodes tenant compte des données censurées	101
V. Les limites des méthodes statistiques	109
Chapitre 4.....	112
I. Problématique générale.....	113
II. Modélisation des retours opérationnels	114
II.1. Contexte et hypothèses.....	114
II.2. Terminologie	114
II.3. Notations	115
II.4. Relation entre W, Q et R	117
II.5. Un cas particulier : les données groupées	118
II.6. Cas où les données sont agrégées	119
III. Vérification de la formule de convolution sur deux exemples élémentaires	121
III.1. Exemple 1 : loi de Weibull, flux de production constant.....	121
III.2. Exemple 2 : loi exponentielle et production linéairement croissante	123
IV. Détermination du taux de défaillance	124
IV.1. Estimation de W(t) à l'aide de N(t).....	125
IV.1.a. Cas particulier.....	125
IV.1.b. Cas général	126
IV.2. Méthodologie de résolution	127
IV.3. Hypothèse et choix des fonctions.....	128
V. Vérification de la méthode sur des jeux de données simulées.....	132
V.1. Simulation des données.....	133
V.2. Application de la méthode	134
V.2.a. Loi de Weibull de paramètres $\beta=0.3$ et $\eta=60$	134
V.2.b. Loi de Weibull de paramètres $\beta=1$ et $\eta=120$	137
V.2.c. Loi de Weibull de paramètres $\beta=2$ et $\eta=36$	138
VI. Application à des données industrielles.....	138
VII. Conclusion.....	142
CONCLUSION ET PERSPECTIVES	144
BIBLIOGRAPHIE	146

LISTE DES FIGURES

Figure 1 : la série d'accidents aériens du mois d'août 2005	17
Figure 2 : balayage de la fenêtre d'observation	21
Figure 3 : un (3,2)-cluster.....	22
Figure 4 : clusters multiples, clusters se chevauchant.....	23
Figure 5 : Intersection du type $W_u \cap W_{u+1}$, cas où $y_{u+1} = 0$ et $y_{u+2} = 1$	25
Figure 6 : Arbre de Shannon de la formule $f(x) = x_1 \vee (x_2 \wedge x_3)$	27
Figure 7 : DDB associé à $f(x)$ après application des règles de simplification	28
Figure 8	29
Figure 9	29
Figure 10	29
Figure 11 : DDB correspondant à la première fenêtre	30
Figure 12 : DDB de la séquence S	32
Figure 13 : balayage de la séquence S.....	33
Figure 14 : Graphe de la chaîne de Markov pour l'exemple proposé	34
Figure 15 : discrétisation de la fenêtre d'observation	42
Figure 16 : algorithme de la simulation sous Excel	44
Figure 17 : génération des dates d'accidents.....	45
Figure 18 : modèle RdP relatif au modèle de Bernoulli.....	47
Figure 19 : modèle RdP relatif au modèle exponentiel renouvelé	47
Figure 20 : balayage de la fenêtre d'observation	48
Figure 21 : dynamique d'une transition	49
Figure 22 : Hypothèse pour la condition initiale.....	50
Figure 23 : graphe de la chaîne de Markov pour le modèle binomial, premier modèle markovien.....	54
Figure 24 : Chaîne de Markov : modèle binomial, premier modèle markovien, $T=365$, $w=10$, $k=3$	55
Figure 25 : Graphe de la chaîne de Markov pour le modèle de Poisson discret, premier modèle markovien.....	61
Figure 26 : Chaîne de Markov pour le modèle de Poisson discret, premier modèle markovien, $T=365$, $w=10$, $k=3$	62
Figure 27 : temps de séjour dans la fenêtre de balayage	65
Figure 28 : des transitions irréalisables admises par la première approche	65
Figure 29 : transition dans le cadre du modèle à double fenêtre.....	66
Figure 30 : les états « i » et « (1,i) » sont équivalents.....	73
Figure 31 : transition dans le cadre du modèle poissonnien	79
Figure 32 : transition	80
Figure 33: courbe en "baignoire"	95
Figure 34: densité de probabilité	95
Figure 35: fiabilité.....	95
Figure 36: taux de défaillance	96
Figure 37 : Comportement général des cartes électroniques. Les données de défaillance et les données censurées sont visibles.	98
Figure 38: Diagramme en bâtons représentant les temps de défaillance et les temps de censure pour la population de l'exemple 2	99
Figure 39 : Comportement général des visualisations, illustrant les mises en service échelonnées ainsi que les défaillances et les censures	99

Figure 40 : Données translatées pour un recalage en zéro.	100
Figure 41 : Droite de Weibull	101
Figure 42 : Graphe de l'estimateur de Kaplan-Meier de la fiabilité $R(t)$ pour les données de l'exemple 2, et loi de Weibull obtenue par moindres carrés.....	106
Figure 43 : Estimateur de Kaplan-Meier vs estimateur de Johnson.....	109
Figure 44 : Base de données des pannes	110
Figure 45 : Flux cumulé de livraison	Figure 46 : Flux cumulé de défaillances ..
Figure 47 : Flux cumulés de livraison et de défaillances	116
Figure 48 : Données groupées	118
Figure 49 : données agrégées	119
Figure 50 : données agrégées pour l'exemple considéré.....	120
Figure 51 : Flux de production cumulé	122
Figure 52 : Flux de production cumulé	123
Figure 53 : flux de production / flux de défaillances cumulés	129
Figure 54 : $P(t)$, $P(t+D)$ et $P(t)\gamma(t-D)$	130
Figure 55 : Flux de production cumulé	133
Figure 56 : Flux cumulé des défaillances.....	134
Figure 57 : Comparaison des fonctions $t \rightarrow N(t)$ et $t \rightarrow -Q(t)\ln\left(1 - \frac{N(t)}{Q(t)}\right)$	135
Figure 58 : Histogramme des défaillances pour les 36 premières classes d'âge.....	136
Figure 59 : Histogramme du nombre de défaillances sur les derniers mois d'observations	136
Figure 60 : Flux cumulé de livraison	139
Figure 61 : Dépotes observées en clientèle.....	139
Figure 62 : Flux cumulé des dépotes	140
Figure 63 : $X(t)$ et régression	141
Figure 64 : taux de défaillance obtenu par la méthode de Kaplan-Meier	142

LISTE DES TABLEAUX

Tableau 1 : table de vérité pour l'exemple(3,2). Le dégradé de couleurs est fonction du nombre de « 1 » dans la séquence.	23
Tableau 2 : Table de vérité de la formule $f(x) = x_1 \vee (x_2 \wedge x_3)$	27
Tableau 3 : caractéristiques des modèles	39
Tableau 4 : Valeur moyenne et écart-type pour un échantillon de 100 jeux de données comportant 1000 simulations chacun.....	46
Tableau 5 : Résultats obtenus par RdP.....	48
Tableau 6 : Double fenêtre de balayage, transitions réalisables depuis un état transitoire (i,j) .	69
Tableau 7 : transitions réalisables	75
Tableau 8 : matrice de transition	76
Tableau 9 : matrice de transition	78
Tableau 10 : matrice de transition	82
Tableau 11: Temps de défaillance de cartes électroniques en heures de fonctionnement	98
Tableau 12 : Contributions à la vraisemblance selon le type de donnée.....	102
Tableau 13	104
Tableau 14 : estimateur de Kaplan-Meier	105
Tableau 15 : Estimateur de Johnson de la fiabilité.....	108
Tableau 16 : Comparaisons des estimateurs de Kaplan-Meier et de Johnson	109
Tableau 17 : flux de livraison d'un matériel A sur l'année 2007.....	119
Tableau 18 : Flux de livraison agrégé au mois.....	120
Tableau 19 : Quelques transformées de Laplace utiles	130
Tableau 20 : Résultats obtenus avec la méthode proposée	135
Tableau 21 : Résultats obtenus avec la méthode de Kaplan-Meier - 1	137
Tableau 22 : Résultats obtenus, $\beta=1$ et $\eta=120$	137
Tableau 23 : Résultats obtenus, $\beta=2$ et $\eta=36$	138
Tableau 24 : Flux de livraison.....	138
Tableau 25 : Flux de déposes	140

RESUME

Le travail de recherche exposé dans ce mémoire de thèse a été accompli dans le cadre d'une convention CIFRE¹ associant l'Université Bordeaux I et l'entreprise Thales Avionics. Il constitue une analyse originale de la fiabilité de matériels complexes, dans une perspective de maîtrise et d'amélioration.

Le sujet traité comporte deux parties disjointes qui trouvent leur justification dans les problématiques rencontrées par l'industriel.

La première thématique porte sur l'analyse des « clusters » d'évènements (accidents en série, succession de défauts sur un îlot de production, ...). Elle fait appel aux statistiques de balayage pour évaluer la probabilité d'occurrence d'un agrégat d'évènements. Plusieurs méthodologies sont envisagées. Une simulation de Monte Carlo directe, puis un Réseau de Petri animé par simulation de Monte Carlo, sont tout d'abord proposés. Des approches markoviennes sont ensuite développées.

La seconde thématique se place dans le cadre classique de l'analyse statistique des données du retour d'expérience en fiabilité. L'originalité du travail présenté est d'explorer le cas particulier où les durées de vie des matériels étudiés ne peuvent être observées ; les seules informations disponibles sont le flux des livraisons (nombre de produits livrés par unité de temps) et le flux des déposes (nombre de défaillances par unité de temps).

Une méthodologie innovante, permettant la construction d'un estimateur du taux de défaillance du matériel considéré, est exposée.

Mots-clés : statistiques de balayage, clusters, Monte-Carlo, Réseau de Petri, chaînes de Markov ; Retour d'Expérience, durées de vie, flux de livraison, flux de déposes, produit de convolution, transformée de Laplace, méthode convolutive.

¹ Convention Industrielle de Formation par la REcherche

ABSTRACT

The thesis went within the scope of an agreement between the University Bordeaux I and the Thales Avionics company. It constitutes an original analysis of the reliability of complex materials equipments, with the prospect of control and improvement.

The thesis consists of two separate parts connected to the problems met by the manufacturer: The first part deals with the analysis of "clusters" of undesirable events (chain of disasters, series of failures,...). It appeals to the scan statistics in order to estimate the probability of occurrence of a cluster of events. A Monte Carlo simulation implemented in a dedicated algorithm, then a Monte Carlo simulation supported by a Petri net model, are proposed. Several markovian approaches are then developed.

The second part deals with the analysis of feedback in a non common context when the only information available is the number of equipments which are delivered during each period and the number of those which are removed during each period. An innovative approach, allowing to obtain the intrinsic failure rate of the materials under study according to the production flow and the removal flow, is explained.

Chapitre 1

Etude préparatoire aux Statistiques de balayage

*« Comment analyser des clusters d'évènements :
coïncidences plausibles ou exceptionnelles ? »*

I. Introduction

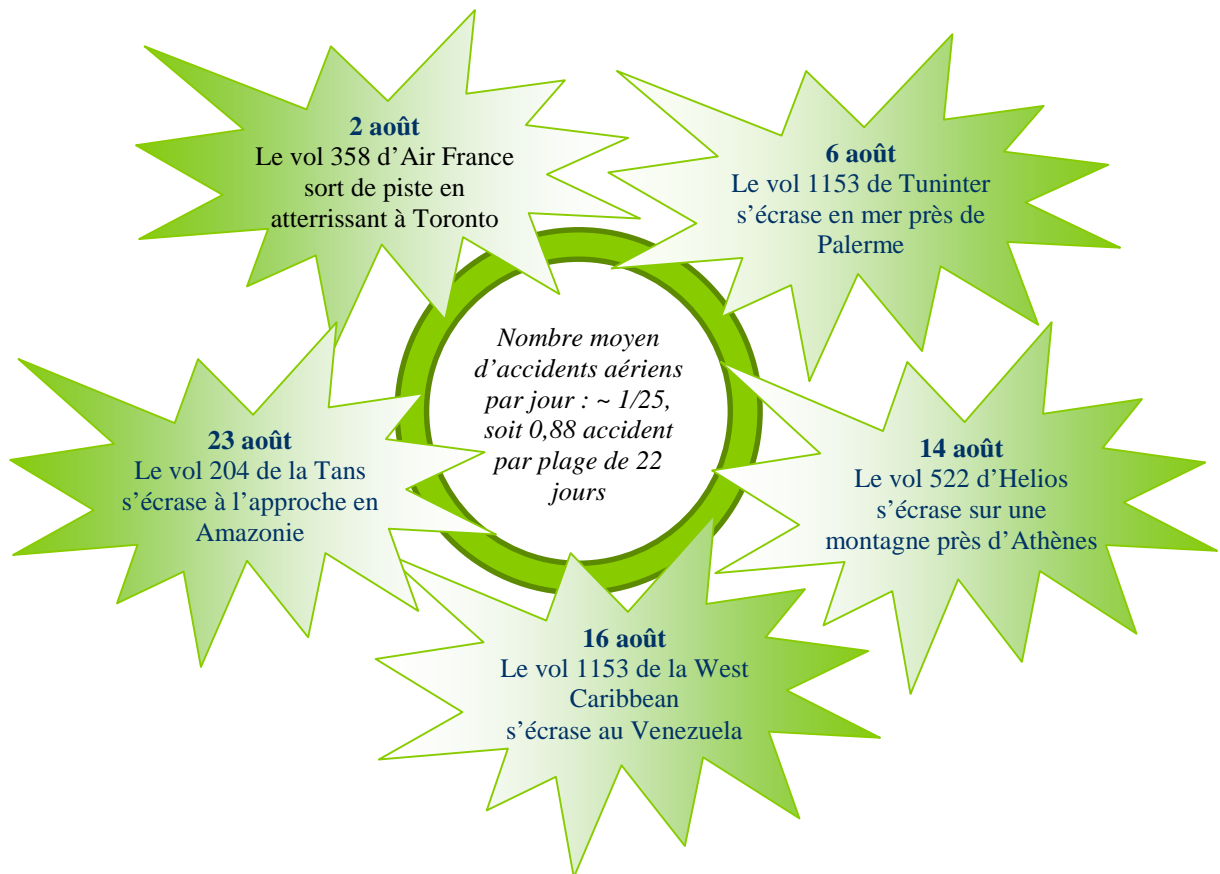
Lorsque nous observons des évènements sur une période donnée, il arrive que nous soyons frappés par des séries de faits particulièrement rapprochés dans le temps. Le réflexe commun est alors de se demander si ces agrégats, que nous appellerons par la suite des **clusters**, sont « normaux ».

D'un point de vue statistique, cela revient à se demander si la probabilité de tels clusters est suffisamment élevée pour que nous admettions que ces « séries » sont des coïncidences ou si, au contraire, cette probabilité est tellement faible que le hasard ne peut en être responsable.

Pour illustrer ces premiers propos, nous reprenons un exemple qui a récemment alerté l'opinion publique.

Le mois d'août 2005 a vu une recrudescence du nombre d'accidents aériens : 5 accidents graves ont eu lieu en 22 jours.

Figure 1 : la série d'accidents aériens du mois d'août 2005



Si ce chiffre a soulevé certaines interrogations, c'est qu'il se situait largement au-dessus du nombre moyen d'accidents observé sur les dix années précédentes. En effet, 147 accidents graves avaient été recensés entre le début de l'année 1995 et le début de l'année 2005 soit une moyenne de 0,04 accident par jour ou encore 0,88 accident en 22 jours.

Cette « série noire » a soulevé nombre de questions relatives à la sécurité aérienne : l'augmentation observée du trafic aérien était-elle correctement maîtrisée ? Le parc d'avions n'entrait-il pas en phase de vieillissement ? Le contrôle et la maintenance étaient-ils adaptés aux nouvelles orientations du trafic aérien ?

L'objectif des **statistiques de balayage** est de repérer et/ou d'analyser les clusters d'évènements afin de distinguer ceux qui résultent du hasard et ceux qui, au contraire, sont la manifestation d'une situation très improbable et donc anormale.

Ajoutons que l'enjeu est fort puisque, dans ce dernier cas, il faudra expliquer les facteurs qui ont conduit à la réalisation du cluster.

Enfin, nombreux sont les domaines où l'on s'intéresse aux clusters d'évènements.

Le premier secteur concerné est celui de la santé publique. L'observation localisée d'un nombre élevé de cas d'une maladie donnée conduit à s'interroger sur la présence de causes communes (habitudes alimentaires, pollution industrielle,...); les épidémiologistes tentent d'anticiper les vagues de virus et définissent des seuils d'alerte au-delà desquels l'état d'épidémie est déclaré, etc...

Mais aujourd'hui, les statistiques de balayage intéressent de plus en plus d'industriels : les ingénieurs en télécommunications, qui cherchent à dimensionner les réseaux de telle sorte qu'ils puissent non seulement supporter un trafic moyen mais aussi faire face à une hausse soudaine de la demande ; les spécialistes des transports, qui cherchent à expliquer certaines séries d'accidents ; les usines de production, qui traquent les défauts en série souvent synonymes d'un relâchement dans les procédés de fabrication, de contrôle ou de maintenance, et précurseurs d'une baisse de qualité, etc...

II. Difficulté du problème

Reprenons l'exemple des accidents aériens et tentons de répondre à la question : *comment peut-on voir un cluster de 5 accidents en 22 jours ? Comme la manifestation aléatoire, au sens probabiliste du terme, d'une série d'évènements ?*

Pour répondre à cette question, et ainsi juger de la « normalité » d'un tel cluster, il faut déjà avoir une idée de la façon dont sont distribués les évènements « en temps normal ».

Pour ce faire, nous utilisons l'historique des accidents aériens sur les dix années précédant notre cluster : nous supposons que les accidents se produisent selon un processus de Poisson dont l'intensité λ n'est rien d'autre que le nombre moyen d'accidents par jour constaté sur la décennie précédant la période qui nous intéresse : soit $\lambda = 0,04 \text{ j}^{-1}$.

Une première approche... incorrecte !

Il est possible de donner une première réponse à la question posée en tentant d'estimer la probabilité que 5 accidents se succèdent en 22 jours.

Soit $N(t)$ la variable aléatoire décrivant le nombre d'accidents survenant durant une période de t jours. Un résultat classique [Foata et Fuchs, 2004] [Saporta, 2006] nous dit que cette variable aléatoire suit une loi de Poisson d'intensité λt .

La probabilité $P(N(t) = k)$ d'observer exactement k accidents sur une période de t jours au plus vaut :

$$P(N(t) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!} .$$

La quantité $P(N(t) \geq k)$ est la probabilité d'observer au moins k accidents sur une période de t jours au plus. On a :

$$P(N(t) \geq k) = 1 - \sum_{i=0}^{k-1} P(N(t) = i) = 1 - \sum_{i=0}^{k-1} e^{-\lambda t} \frac{(\lambda t)^i}{i!} .$$

Dans le cas où $k = 5$ et $t = 22$, le calcul donne une probabilité de 0,0022.

La première conclusion qui s'offre à nous est donc que la probabilité d'observer plus de 5 accidents sur une période donnée de 22 jours est très faible puisqu'inférieure à 0,3%.

Mais la probabilité calculée ci-dessus est la probabilité d'observer une série de 5 accidents sur une plage fixe de 22 jours, par exemple entre le 1^{er} et le 22 décembre. Or, sur une année de 365 jours, les plages de 22 jours s'enchaînent (et se chevauchent) au nombre de 344 et n'importe laquelle de ces plages peut révéler un cluster d'accidents.

Nous pouvons donc supposer que la probabilité d'observer un cluster sur une année complète est en réalité bien supérieure à 0,3%...

Une seule bonne approche !

La probabilité d'observer un cluster sur une période quelconque de 22 jours est difficile à évaluer, la difficulté résidant dans le fait que les contributions des plages ne sont pas nécessairement indépendantes car lesdites plages sont non disjointes.

En effet, si l'on note T la durée de l'observation (T est égal à 365 jours dans notre exemple), le nombre de plages de t jours à considérer (t est égal à 22 jours dans notre exemple) est $T - t + 1$. Sur chacune des plages de t jours, la probabilité d'observer un cluster de k accidents au moins est celle calculée plus haut. Sur la période complète de T jours, la probabilité d'observer au moins un cluster de k accidents sur une plage quelconque de t jours s'écrit comme la probabilité d'une réunion, ou bien, en considérant l'évènement complémentaire (« n'observer aucun cluster sur la période d'observation »), comme le complément à 1 de la probabilité d'une intersection. Soit :

$$P \left(\bigcup_{\substack{t_{\text{fin}} - t_{\text{deb}} \leq t \\ t_{\text{fin}}, t_{\text{deb}} \in \{1, \dots, T\}}} N(t_{\text{fin}} - t_{\text{deb}}) \geq k \right) = 1 - P \left(\bigcap_{\substack{t_{\text{fin}} - t_{\text{deb}} \leq t \\ t_{\text{fin}}, t_{\text{deb}} \in \{1, \dots, T\}}} N(t_{\text{fin}} - t_{\text{deb}}) < k \right),$$

t_{fin} et t_{deb} désignant respectivement le début et la fin de la fenêtre courante.

D'un côté, nous avons une réunion d'évènements non disjointes ; le calcul de la probabilité correspondante fait appel à la formule de Poincaré et le nombre de termes impliqués croît très vite avec T . D'un autre côté, la probabilité de l'intersection n'est pas plus simple à évaluer car les évènements $\{N(t_{\text{fin}} - t_{\text{deb}}) < k\}$ ne sont pas indépendants.

Dans les deux cas, les calculs sont vite impraticables.

La solution réside dans l'utilisation des « statistiques de balayage » qui permettent d'évaluer la probabilité qu'une série d'évènements se produise dans une fenêtre (temporelle, spatiale ou spatio-temporelle) de taille donnée pouvant occuper toutes les positions possibles d'une région.

III. Formulation mathématique

III.1. Définitions

Soit (Ω, τ, P) un espace probabilisé et $(X_i)_{i \in I}$ une famille finie ou dénombrable de variables aléatoires réelles positives sur Ω .

Une réalisation de X_i , c'est-à-dire une valeur $X_i(\omega)$ pour $\omega \in \Omega$, sera notée ξ_i .

Soit $T > 0$, $0 < w \leq T$ et $k \in \mathbb{N}^*$.

Définition. Une famille de réalisations $(\xi_i = X_i(\omega))_{i \in I}$ des X_i présente un (w, k) -cluster (ou (w, k) -agrégat) sur l'intervalle $[0, T[$ s'il existe $u \in [0, T - w]$ tel que l'intervalle $[u, u + w[$ contienne au moins k des valeurs des ξ_i .

Hypothèse de finitude. Pour tout $t > 0$, un nombre fini des X_i sont à valeurs dans $[0, t[$ presque sûrement.

L'hypothèse de finitude permet d'introduire la statistique d'ordre $(X_{(n)})$, où n parcourt l'ensemble $\{1, 2, \dots, |I|\}$ si I est fini et \mathbb{N}^* sinon.

Nous définissons deux variables aléatoires :

1. La fonction de comptage définie pour $u \in [0, T - w]$ par

$$N_w(u) = \text{Card} \{i \in I \mid X_i \in [u, u + w[\} = \text{Card} \{n \mid X_{(n)} \in [u, u + w[\} ;$$

2. La statistique de balayage :

$$S_w = \max_{u \in [0, T - w]} N_w(u).$$

Avec les notations ci-dessus, la probabilité d'apparition d'un (w, k) -cluster pour la famille $(X_i)_{i \in I}$ est $P\{S_w \geq k\}$ et s'écrit :

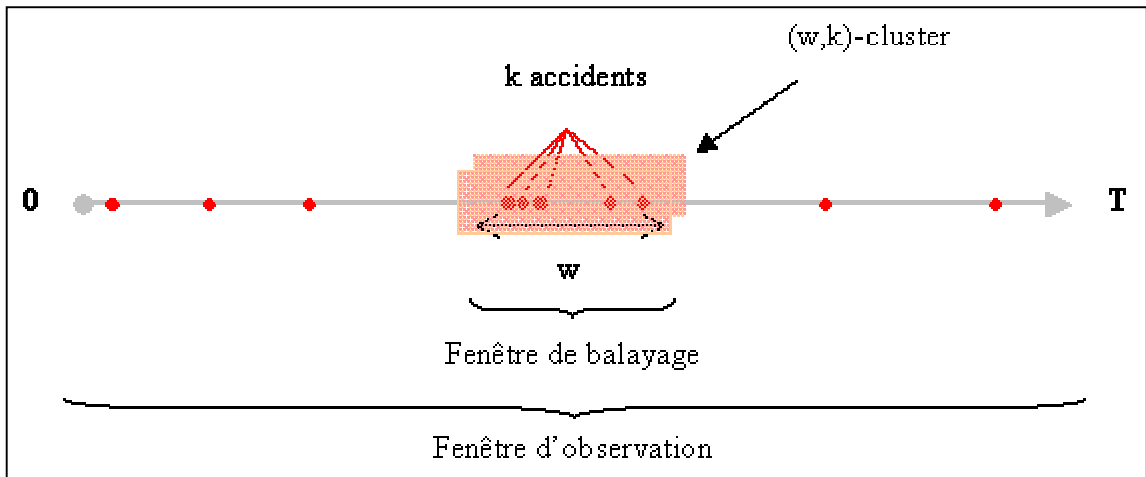
$$P\{S_w \geq k\} = P \left\{ \bigcup_{u \in [0, T - w]} \{N_w(u) \geq k\} \right\},$$

que l'on peut réécrire en tenant compte de la statistique d'ordre :

$$P\{S_w \geq k\} = P \left\{ \bigcup_{\substack{m \\ X_{(m)} \leq T - w}} \{N_w(X_{(m)}) \geq k\} \right\}.$$

Remarque : avec l'hypothèse de finitude, S_w et $N_w(u)$ sont presque sûrement à valeurs dans \mathbb{N} ou, en d'autres termes, presque sûrement finies.

Figure 2 : balayage de la fenêtre d'observation



III.2. Deux modèles typiques

Un modèle continu : modèle « exponentiel avec renouvellement »

Soit $\alpha > 0$.

Soit Z_1, \dots, Z_m, \dots une suite de variables aléatoires indépendantes de loi exponentielle de paramètre α , et $X_m = Z_1 + \dots + Z_m$, $m \in \mathbb{N}^*$.

Proposition. La variable aléatoire $N_w(u)$ suit dans ce cas une loi de Poisson de paramètre αw [Foata et Fuchs, 2003].

Remarque : l'hypothèse de finitude est satisfaite ; nous avons $X_{(m)} = X_m$.

Un modèle discret : modèle de Bernoulli

Soit $T \in \mathbb{N}^*$.

$I = \{0, 1, \dots, T-1\}$ ($= [0, T[\cap \mathbb{N}$).

$w \in \{1, \dots, T\}$.

$p \in [0, 1]$.

$1 \leq k \leq w$.

Soit Y_0, \dots, Y_{T-1} une suite de variables aléatoires indépendantes de loi de Bernoulli de paramètre p .

Posons $X_t = \begin{cases} t & \text{si } Y_t = 1 \\ T+1 & \text{si } Y_t = 0 \end{cases}$.

La fonction de comptage s'écrit alors :

$$N_w(u) = \text{Card}\{t \mid X_t \in [u, u+w[\} = \text{Card}\{t \mid X_t \in [u, u+w[\text{ et } Y_t = 1\} = \sum_{t=u}^{u+w-1} Y_t .$$

Proposition : la variable aléatoire $N_w(u)$ suit dans ce cas une loi binomiale de paramètres w et p [Foata et Fuchs, 2003].

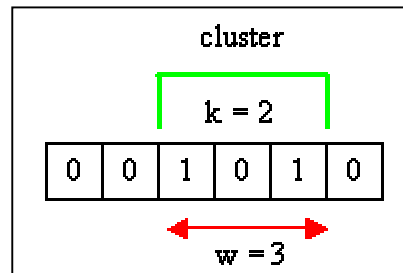
IV. Un exemple simple...

Pour comprendre le problème, regardons l'exemple simple suivant, relevant du modèle de Bernoulli, avec $T = 6$, $w = 3$ et $k = 2$.

Soit la séquence $S = (y_0, y_1, y_2, y_3, y_4, y_5)$ résultant de $T = 6$ épreuves de Bernoulli prenant la valeur 1 avec la probabilité p et la valeur 0 avec la probabilité $q = 1 - p$. Cherchons la probabilité P qu'il y ait au moins une fenêtre de longueur trois contenant au moins deux 1.

La séquence suivante est une des séquences possibles. Elle contient le cluster (1,0,1).

Figure 3 : un (3,2)-cluster



Les clusters que nous pouvons rencontrer dans ce problème sont les triplets (0,1,1), (1,0,1), (1,1,0) ou (1,1,1).

De par sa petite taille, ce problème peut aisément être résolu « à la main », et ce de plusieurs façons. Nous présentons ci-après quatre d'entre elles.

Nous n'avons pas trouvé trace dans la littérature de la quatrième approche, de nature markovienne.

IV.1. Approche combinatoire

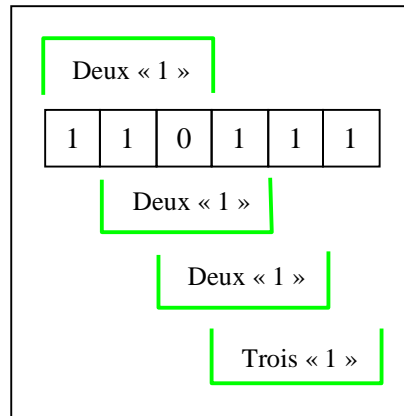
Il y a $2^6 = 64$ configurations possibles pour la séquence S . Un dénombrement exhaustif est donc tout à fait raisonnable. Une table de vérité (voir Tableau 1) nous permet d'identifier les remplissages possibles et d'isoler ceux qui contiennent des clusters.

Nous pouvons même aller plus loin et observer :

- Que certaines séquences contiennent plusieurs clusters ;
- Que le nombre de « 1 » dans un cluster peut être 2 ou 3 ;
- Que plusieurs clusters peuvent se chevaucher.

Par exemple, la séquence suivante contient 3 clusters constitués de deux « 1 » et d'un « 0 » se chevauchant et 1 cluster constitué de trois « 1 » :

Figure 4 : clusters multiples, clusters se chevauchant



La table de vérité est présentée dans le tableau 1.

Tableau 1 : table de vérité pour l'exemple(3,2). Le dégradé de couleurs est fonction du nombre de « 1 » dans la séquence.

ε_1	ε_2	ε_3	ε_4	ε_5	ε_6	cluster
1	1	1	1	1	1	Oui
1	1	1	1	1	0	Oui
1	1	1	1	0	1	Oui
1	1	1	1	0	0	Oui
1	1	1	0	1	1	Oui
1	1	1	0	1	0	Oui
1	1	1	0	0	1	Oui
1	1	1	0	0	0	Oui
1	1	0	1	1	1	Oui
1	1	0	1	1	0	Oui
1	1	0	1	0	1	Oui
1	1	0	1	0	0	Oui
1	1	0	0	1	1	Oui
1	1	0	0	1	0	Oui
1	1	0	0	0	1	Oui
1	1	0	0	0	0	Oui
1	0	1	1	1	1	Oui
1	0	1	1	1	0	Oui
1	0	1	1	0	1	Oui
1	0	1	1	0	0	Oui
1	0	1	0	1	1	Oui
1	0	1	0	1	0	Oui
1	0	1	0	0	1	Oui
1	0	1	0	0	0	Oui
1	0	0	1	1	1	Oui
1	0	0	1	1	0	Oui
1	0	0	1	0	1	Oui
1	0	0	1	0	0	Non
1	0	0	0	1	1	Oui
1	0	0	0	1	0	Non
1	0	0	0	0	1	Non
1	0	0	0	0	0	Non

0	1	1	1	1	1	Oui
0	1	1	1	1	0	Oui
0	1	1	1	0	1	Oui
0	1	1	1	0	0	Oui
0	1	1	0	1	1	Oui
0	1	1	0	1	0	Oui
0	1	1	0	0	1	Oui
0	1	1	0	0	0	Oui
0	1	0	1	1	1	Oui
0	1	0	1	1	0	Oui
0	1	0	1	0	1	Oui
0	1	0	1	0	0	Oui
0	1	0	0	1	1	Oui
0	1	0	0	1	0	Non
0	1	0	0	0	1	Non
0	1	0	0	0	0	Non
0	0	1	1	1	1	Oui
0	0	1	1	1	0	Oui
0	0	1	1	0	1	Oui
0	0	1	1	0	0	Oui
0	0	1	0	1	1	Oui
0	0	1	0	1	0	Oui
0	0	1	0	0	1	Non
0	0	1	0	0	0	Non
0	0	0	1	1	1	Oui
0	0	0	1	1	0	Oui
0	0	0	1	0	1	Oui
0	0	0	1	0	0	Non
0	0	0	0	1	1	Oui
0	0	0	0	1	0	Non
0	0	0	0	0	1	Non
0	0	0	0	0	0	Non

Nombre de séquences contenant au moins un cluster

51

Chaque y_i étant le résultat d'une épreuve de Bernoulli de paramètre p , la probabilité d'une séquence donnée, constituée de j fois « 1 » et $6-j$ fois « 0 », est $P_j = p^j q^{6-j}$, ce que l'on peut encore voir comme la probabilité d'obtenir exactement i succès et j échecs à des places déterminées de la séquence S .

La table de vérité (voir Tableau 1) nous permet de dénombrer :

- 9 clusters de taille 2 ayant chacun la probabilité $P_2 = p^2 q^4$.
- 20 clusters de taille 3 ayant chacun la probabilité $P_3 = p^3 q^3$.
- 15 clusters de taille 4 ayant chacun la probabilité $P_4 = p^4 q^2$.
- 6 clusters de taille 5 ayant chacun la probabilité $P_5 = p^5 q$.
- 1 cluster de taille 6 ayant la probabilité $P_6 = p^6$.

Par conséquent, la probabilité d'observer un cluster dans la séquence S est :

$$P = 9P_2 + 20P_3 + 15P_4 + 6P_5 + P_6,$$

d'où

$$P = 9p^2 q^4 + 20p^3 q^3 + 15p^4 q^2 + 6p^5 q + p^6,$$

et en tenant compte du fait que $q = 1 - p$:

$$P = 9p^2 - 16p^3 + 9p^4 - p^6.$$

Remarque : cette approche est à l'évidence peu adaptée aux cas « réalistes » qui correspondent à des valeurs de T , w et k bien plus élevées que celles du cas élémentaire que nous venons de détailler.

IV.2. Formule de Poincaré

Pour la concision des formules, nous notons provisoirement W_u l'évènement $\{N_w(u) \geq 2\}$, ou encore $y_u + y_{u+1} + y_{u+2} \geq 2$, c'est-à-dire l'évènement « la séquence (y_u, y_{u+1}, y_{u+2}) contient au moins deux 1 ». La probabilité recherchée est alors :

$$P \left\{ \bigcup_{u=0}^3 W_u \right\},$$

soit, en recourant à une formulation explicite :

$$P = \sum_{u=0}^3 P(W_u) - \sum_{0 \leq u < u' \leq 3} P\{W_u \cap W_{u'}\} + \sum_{0 \leq u < u' < u'' \leq 3} P\{W_u \cap W_{u'} \cap W_{u''}\} - P\{W_0 \cap W_1 \cap W_2 \cap W_3\}.$$

Premier terme : évènement simple

Pour $0 \leq u \leq 3$, $P(W_u)$ est la probabilité que la séquence (y_u, y_{u+1}, y_{u+2}) contienne un cluster, c'est-à-dire deux ou trois 1. Soit : $P(W_u) = \binom{3}{2} p^2 q + \binom{3}{3} p^3 = 3p^2 - 2p^3$.

Par conséquent, le premier terme de la formule de Poincaré vaut :

$$\sum_{u=0}^3 P(W_u) = 12p^2 - 8p^3.$$

Deuxième terme : intersection de deux évènements

Nous avons deux types d'intersection à considérer :

- Les intersections du type $W_u \cap W_{u+1}$ pour $0 \leq u \leq 2$. Une intersection de ce type est non vide si les séquences (y_u, y_{u+1}, y_{u+2}) et $(y_{u+1}, y_{u+2}, y_{u+3})$ contiennent chacune un cluster.

Conditionnons par y_{u+1} et y_{u+2} qui sont communs aux deux séquences :

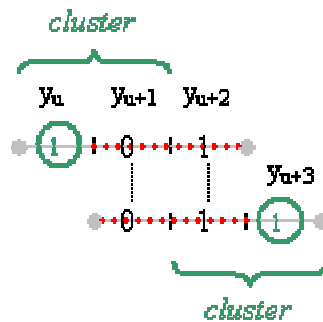
$$P(W_u \cap W_{u+1}) = \sum_{0 \leq m \leq n \leq 1} P(W_u \cap W_{u+1} | (y_{u+1}, y_{u+2}) = (m, n)) \times P((y_{u+1}, y_{u+2}) = (m, n)).$$

- Si $(y_{u+1}, y_{u+2}) = (0, 0)$, aucune des deux fenêtres ne contient de cluster et l'intersection est vide.
- Si $(y_{u+1}, y_{u+2}) = (1, 1)$ alors W_u et W_{u+1} contiennent un cluster quelles que soient les valeurs prises par y_u et y_{u+3} .
- Si $(y_{u+1}, y_{u+2}) = (0, 1)$ ou $(1, 0)$ alors W_u et W_{u+1} contiennent chacune un cluster (cf. figure 5) si et seulement si $y_u = 1$ et $y_{u+3} = 1$ (voir figure 5).

Il vient donc pour la probabilité de l'intersection :

$$\begin{aligned} P(W_u \cap W_{u+1}) &= 0 \times P((y_{u+1}, y_{u+2}) = (0, 0)) + 1 \times P((y_{u+1}, y_{u+2}) = (1, 1)) + 2 \times p^2 \times P((y_{u+1}, y_{u+2}) = (0, 1)) \\ &= 1 \times p^2 + 2 \times p^2 \times pq \\ &= p^2 + 2p^3(1-p) \\ &= p^2 + 2p^3 - 2p^4 \end{aligned}$$

Figure 5 : Intersection du type $W_u \cap W_{u+1}$, cas où $y_{u+1} = 0$ et $y_{u+2} = 1$



- Les intersections du type $W_u \cap W_{u+2}$ pour $u = 0, 1$. Une intersection de ce type est non vide si les séquences (y_u, y_{u+1}, y_{u+2}) et $(y_{u+2}, y_{u+3}, y_{u+4})$ contiennent chacune un cluster.

En conditionnant par y_{u+2} qui est commun aux deux séquences, il vient :

$$P(W_u \cap W_{u+2}) = \sum_{m \in \{0,1\}} P(W_u \cap W_{u+2} | y_{u+2} = m) \times P(y_{u+2} = m).$$

- Si $y_{u+2} = 0$, alors W_u et W_{u+2} contiennent chacune un cluster si et seulement si $y_u = y_{u+1} = y_{u+3} = y_{u+4} = 1$.
- Si $y_{u+2} = 1$ alors W_u et W_{u+2} contiennent chacune un cluster si et seulement si $(y_u, y_{u+1}) \neq (0,0)$ et $(y_{u+3}, y_{u+4}) \neq (0,0)$.

D'où

$$\begin{aligned} P(W_u \cap W_{u+2}) &= P(y_u = 1, y_{u+1} = 1, y_{u+3} = 1, y_{u+4} = 1) \times P(y_{u+2} = 0) \\ &\quad + (1 - P(y_u = 0, y_{u+1} = 0))(1 - P(y_{u+3} = 0, y_{u+4} = 0)) \times P(y_{u+2} = 1); \end{aligned}$$

$$\begin{aligned} P(W_u \cap W_{u+2}) &= p^4 \times q + (1 - q^2)^2 \times p \\ &= p^4 \times (1 - p) + (1 - (1 - p)^2)^2 \times p \\ &= 4p^3 - 3p^4. \end{aligned}$$

- L'intersection du type $W_0 \cap W_3$. Elle est non vide si les séquences (y_0, y_1, y_2) et (y_3, y_4, y_5) , qui sont disjointes, contiennent chacune un cluster.

Il vient :

$$\begin{aligned} P(W_0 \cap W_3) &= \left(\binom{2}{3} p^2 (1 - p) + \binom{3}{3} p^3 \right)^2 \\ &= 9p^4 - 12p^5 + 4p^6. \end{aligned}$$

Par conséquent, le deuxième terme de la formule de Poincaré vaut :

$$\begin{aligned} \sum_{0 \leq u < u' \leq 3} P\{W_u \cap W_{u'}\} &= 3 \times [p^2 + 2p^3 - 2p^4] + 2 \times [4p^3 - 3p^4] + [9p^4 - 12p^5 + 4p^6] \\ &= 3p^2 + 14p^3 - 3p^4 - 12p^5 + 4p^6. \end{aligned}$$

En raisonnant de manière similaire, nous pouvons évaluer le troisième et le quatrième terme de la formule de Poincaré.

Troisième terme : intersection de trois évènements

Il s'agit des intersections du type $W_u \cap W_{u+1} \cap W_{u+2}$ pour $u = 0, 1$, de probabilité $3p^3 - p^4 - p^5$, $W_0 \cap W_2 \cap W_3$ et $W_0 \cap W_1 \cap W_3$ de probabilité $7p^4 - 8p^5 + 2p^6$.

Le troisième terme de la formule de Poincaré vaut :

$$\begin{aligned} \sum_{0 \leq u < u' < u'' \leq 3} P\{W_u \cap W_{u'} \cap W_{u''}\} &= 2 \times (3p^3 - p^4 - p^5) + 2 \times (7p^4 - 8p^5 + 2p^6) \\ &= 6p^3 + 12p^4 - 18p^5 + 4p^6. \end{aligned}$$

Quatrième terme : intersection de quatre évènements

Il s'agit de l'intersection $W_0 \cap W_1 \cap W_2 \cap W_3$ de probabilité $6p^4 - 6p^5 + p^6$.

Nous trouvons donc pour la probabilité recherchée, probabilité que la séquence $S = (y_0, y_1, y_2, y_3, y_4, y_5)$ contienne au moins un (3,2)-cluster :

$$P = 9p^2 - 16p^3 + 9p^4 - p^6.$$

Remarque : le nombre de termes à évaluer pour la formule de Poincaré est 2^{T-w+1} , contre 2^T pour le dénombrement classique. Malgré tout, le nombre de termes deviendra vite ingérable dans la majorité des cas rencontrés (par exemple $T=365$ jours et $w \ll T$).

Nous constatons de plus, sur un exemple pourtant élémentaire, que le calcul des probabilités d'intersections de fenêtres qui se recouvrent devient vite complexe.

Le recours à la formule de Poincaré ne constitue donc pas une option praticable, ce qui était prévisible et déjà maintes fois constaté dans le domaine des statistiques de balayage [Dupin, 2007], comme dans celui du traitement des arbres de défaillances [Dutuit et Rauzy, 2002], [Linnios, 2005].

IV.3. Approche basée sur le formalisme des Diagrammes de Décision Binaires

Dans le domaine de la sûreté de fonctionnement, et tout particulièrement dans celui du traitement des arbres de défaillances, une avancée importante a été réalisée en codant ces arbres, ou plus exactement les formules booléennes sous-jacentes, à l'aide des diagrammes de décision binaires (DDB).

Appliquons, en l'adaptant, cette technique au cas élémentaire traité ci-dessus.

IV.3.a. Une brève présentation des DDB

Les DDB [Bryant, 1986] [Rauzy, 1996] sont des structures de données permettant de coder et de manipuler les fonctions booléennes très efficacement. Un DDB, associé à une formule booléenne, code d'une manière concise la table de vérité de cette formule. En choisissant un ordre total sur les variables et en appliquant de manière récursive la décomposition pivotale, la table de vérité peut être représentée graphiquement par un arbre binaire (dit arbre de Shannon) tel que celui de la figure 6, associé à la formule $f(x) = x_1 \vee (x_2 \wedge x_3)$.

Figure 6 : Arbre de Shannon de la formule

$$f(x) = x_1 \vee (x_2 \wedge x_3)$$

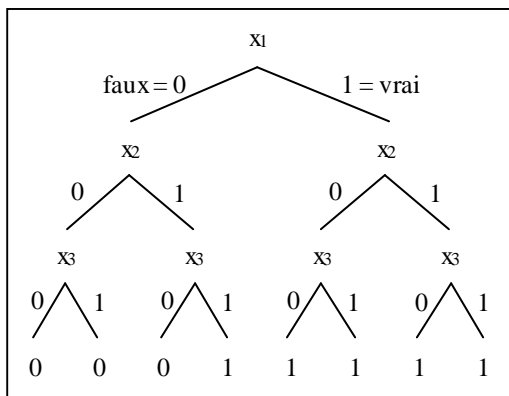


Tableau 2 : Table de vérité de la formule

$$f(x) = x_1 \vee (x_2 \wedge x_3)$$

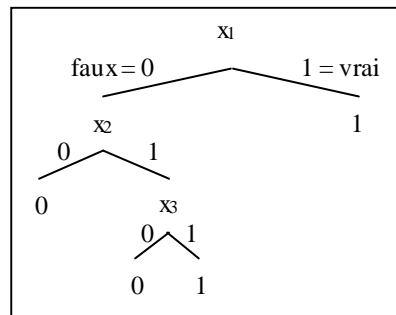
x_1	x_2	x_3	$x_2 \wedge x_3$	$f(x)$
1	1	1	1	1
1	1	0	0	1
1	0	1	0	1
1	0	0	0	1
0	1	1	1	1
0	1	0	0	0
0	0	1	0	0
0	0	0	0	0

Bien évidemment, la taille d'une telle représentation arborescente croît très rapidement avec le nombre de variables, mais il est possible de la réduire en appliquant deux règles :

- la fusion des sous-arbres isomorphiques ;
- la suppression des nœuds dotés de deux « fils » identiques et leur remplacement par l'un d'eux.

L'application de ces règles à l'arbre précédent conduit au DDB suivant, qui est donc également un graphe orienté acyclique :

Figure 7 : DDB associé à f(x) après application des règles de simplification



Dans la pratique, l'obtention des DDB codant une formule booléenne $f \odot g$ ne résulte pas de la simplification de l'arbre de Shannon correspondant mais d'une procédure récursive qui permet de le construire en combinant les DDB qui codent f et g , selon la logique du connecteur \odot impliqué (conjonction, disjonction, négation). Cette procédure s'appuie sur les deux égalités suivantes :

$$\text{ite}(x, F, G) \odot \text{ite}(y, H, L) = \text{ite}(x, F \odot \text{ite}(y, H, L), G \odot \text{ite}(y, H, L)),$$

$$\text{ite}(x, F, G) \odot \text{ite}(x, H, L) = \text{ite}(x, F \odot H, G \odot L),$$

dans lesquelles x et y sont deux variables ordonnées ($x < y$) et F, G, H, L quatre formules booléennes codées par un DDB,

ite signifiant « if-then-else ».

Remarque: l'expression $\text{ite}(x, F, G)$, que l'on peut lire « if x then F else G », représente la même entité que :

$$\begin{array}{c} 1 \\ x \text{ --- } F \\ 0 | \\ G \end{array} \quad \text{ou} \quad x.F + \bar{x}.G$$

Le codage d'une formule booléenne au moyen d'un DDB permet de calculer directement la probabilité qui lui est associée au terme d'un seul parcours de ce DDB, au cours duquel la décomposition de Shannon est appliquée à chacun de ses nœuds $\alpha = \text{ite}(x, F, G)$. L'algorithme qui effectue ce calcul est entièrement décrit par les équations récursives suivantes :

$$\begin{aligned} P(1) &= 1 ; P(0) = 0 ; \\ P(\alpha) &= P(x).P(F) + (1 - P(x)).P(G) ; \end{aligned}$$

IV.3.b. Application au calcul de la probabilité d'occurrence d'un « cluster »

Nous reprenons l'exemple simple traité dans les paragraphes précédents. Cette fois cependant, afin de réduire le phénomène d'explosion combinatoire qui annihile l'intérêt des deux tentatives précédentes, nous proposons de calculer la probabilité de l'évènement complémentaire de celui retenu initialement. Nous calculons donc la probabilité $P'=1-P$ qu'aucune des quatre fenêtres de longueur trois couvrant la période d'observation ne contienne plus d'un évènement.

La démarche s'articule en étapes successives que nous décrivons succinctement, et de manière graphique, ci-après.

- Nous construisons d'abord le DDB relatif à la première fenêtre, constituée des trois premiers tirages de la séquence S.
Le DDB élémentaire représenté sur la figure 8 correspond aux éventualités inhérentes à un tirage J donné. Il rend compte de l'union (disjonction) exclusive des deux évènements J_0 et J_1 qui indiquent respectivement que le résultat du tirage est 0 et 1.

Figure 8

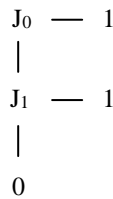


Figure 9

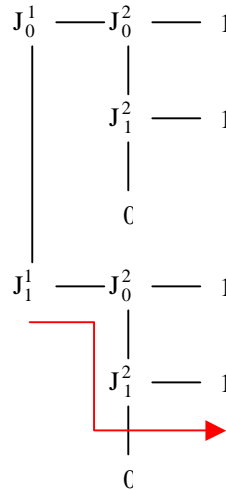
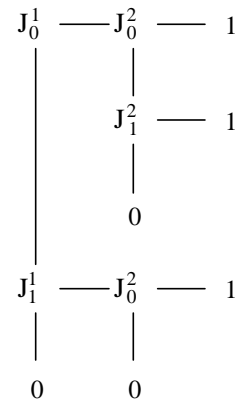


Figure 10

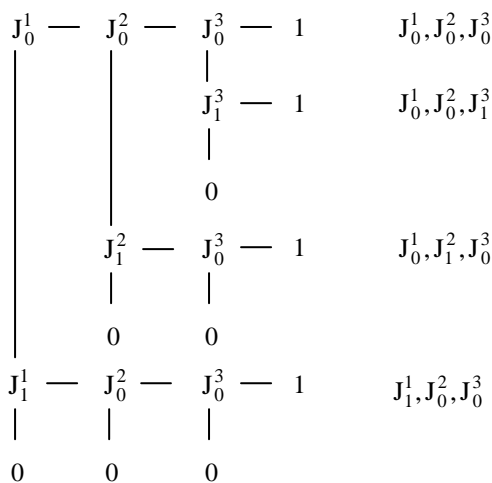


Ce formalisme étant retenu, nous combinons par conjonction les DDB relatifs au premier et au deuxième tirages. Nous obtenons le DDB de la figure 9.

Nous remarquons alors que le sous-arbre signalé par une flèche n'est pas autorisé car le premier tirage et le second tirage sont des 1. Nous remplaçons donc la feuille « 1 » par « 0 » ce qui entraîne la suppression de la séquence. Le DDB résultant est représenté sur la figure 10.

Le DDB représentatif de la première fenêtre s'obtient en faisant la conjonction du DDB représenté figure 10 et du DDB correspondant au troisième tirage. Après simplification, c'est-à-dire après suppression des clusters de deux évènements, il reste le DDB représenté sur la figure 11.

Figure 11 : DDB correspondant à la première fenêtre



Les seules séquences autorisées sont donc J_0^1, J_0^2, J_0^3 , J_0^1, J_0^2, J_1^3 , J_0^1, J_1^2, J_0^3 , J_1^1, J_0^2, J_0^3 . Nous pouvons dès à présent remarquer que ces séquences renvoient aux états E_i , $i = 1$ à 4, de la chaîne de Markov du prochain paragraphe (figure 14).

- Nous déduisons du graphe de la figure 11 le DDB de chacune des fenêtres suivantes, en incrémentant à chaque fois d'une unité le rang de chacun des jours inclus dans le DDB précédent.
- Nous procédons ensuite à la conjonction du premier et du deuxième DDB, puis du DDB résultant avec le suivant, et ainsi de suite jusqu'à obtenir le DDB représentatif de la séquence S (figure 12).
Ce DDB regroupe toutes les réalisations souhaitées de S : celles qui ne contiennent aucune fenêtre de longueur 3 contenant deux événements ou plus. Elles sont au nombre de treize.
- Enfin, nous exploitons le DDB final pour calculer la valeur de la probabilité P' qui nous intéresse. Nous pouvons le faire de deux manières :
 - En recouvrant à la méthode récursive citée précédemment ;
 - En calculant la probabilité de réalisation de chacun des 13 tirages et en sommant ces probabilités.

Vu la taille modeste du DDB considéré, c'est cette seconde méthode que nous mettons en œuvre. Pour cela, nous affectons les probabilités p et q respectivement aux nœuds J_1^i et J_0^i , et pour chaque séquence, nous effectuons le produit des probabilités rencontrées lors de son parcours.

Pour l'unique séquence ne contenant aucun événement, la probabilité calculée est q^6 ; pour chacune des six séquences contenant une seule occurrence, la probabilité calculée est pq^5 ; pour chacune des six séquences contenant deux occurrences, la probabilité calculée est p^2q^4 .

Ces probabilités sont renseignées sur la figure 12.

Nous trouvons $P' = q^6 + 6q^5p + 6q^4p^2$ soit, en tenant compte du fait que $q = 1 - p$,
 $P' = -9p^2 + 16p^3 - 9p^4 + p^6 + 1$.

Il vient :

$$P = 9p^2 - 16p^3 + 9p^4 - p^6,$$

ce qui est conforme aux résultats obtenus avec les précédentes méthodes.

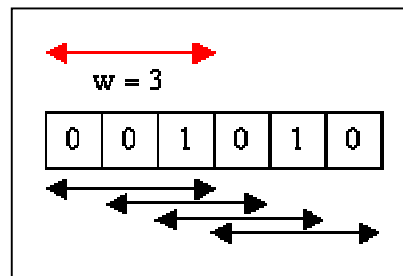
Remarque : l'avantage de l'approche par DDB est double : d'une part, elle réduit le nombre de termes (séquences ou probabilités) à prendre en compte pour le calcul de la probabilité recherchée (13 séquences à considérer contre 64 pour l'approche combinatoire classique et 15 probabilités conjointes pour la formule de Poincaré) ; d'autre part, elle ne nécessite pas le calcul des probabilités conjointes de fenêtres non disjointes. Il convient cependant de noter que la mise en œuvre de DDB sur des cas réalistes n'est envisageable que si l'on dispose d'un moteur de calcul informatique dédié. Son intérêt est donc pour l'instant limité, comme celui des deux approches proposées précédemment.

IV.4. Approche markovienne

Nous restons sur le même exemple. Soit une séquence S donnée résultant de $T=6$ épreuves de Bernoulli. Pour détecter la présence de cluster(s) la démarche la plus intuitive consiste à parcourir cette séquence au moyen d'une « fenêtre » de longueur $w=3$ et à contrôler le contenu de cette fenêtr e à chaque déplacement.

La fenêtr e de balayage est initialement positionnée au début de la séquence. Le balayage de la séquence est interrompu soit lorsqu'un cluster est détecté, soit lorsque la fin de la séquence est atteinte.

Figure 13 : balayage de la séquence S



Cette dynamique de « balayage » peut être représentée par une chaîne de Markov (voir figure 14) ayant les caractéristiques suivantes :

- *Etats*

La chaîne de Markov présente cinq états E_1, E_2, E_3, E_4 et A qui sont associés aux 8 triplets de $\{0,1\}^3$ de la façon suivante :

$(0,0,0) \rightarrow E_1,$

$(0,0,1) \rightarrow E_2,$

$(0,1,0) \rightarrow E_3,$

$(1,0,0) \rightarrow E_4,$

$(0,1,1), (1,0,1), (1,1,0), (1,1,1) \rightarrow E_5 = A$, c'est-à-dire que l'état A correspond à tout triplet comportant un cluster.

- *Condition initiale*

Initialement, la chaîne de Markov peut être en n'importe lequel des états de E selon la composition de la première fenêtr e c'est-à-dire, dans notre exemple, selon la valeur du triplet (y_0, y_1, y_2) .

Le vecteur des probabilités initiales est donc le vecteur :

$$\Pi = [q^3 \quad pq^2 \quad pq^2 \quad pq^2 \quad 1 - q^3 - 3pq^2]^t.$$

- *Matrice de transition*

La dynamique de notre chaîne traduira l'évolution de la composition de la fenêtr e courante tant que l'on n'a pas rencontré de cluster ; dans le cas contraire, elle se contentera de se souvenir que l'on rencontre un cluster, ou que l'on en a rencontré un antérieurement.

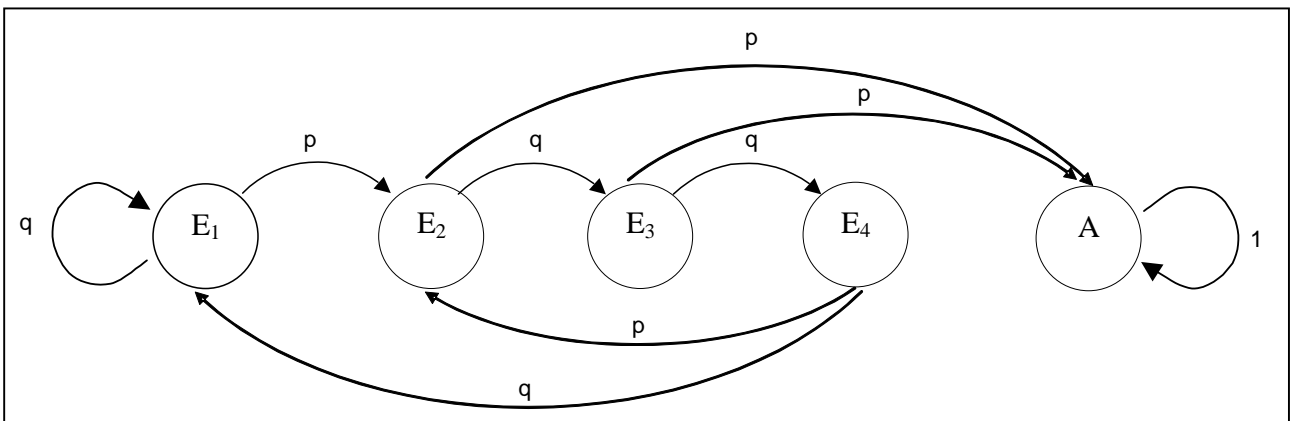
Ce dernier point se traduit par le fait que l'état A est absorbant. Les autres probabilités de transition dépendent directement de la nature du nouvel élément à introduire dans un triplet. A titre d'exemple, la probabilité de passer du triplet (0,1,0) au triplet (1,0,1) est p : c'est la probabilité pour que le nouvel élément entrant soit 1. En revanche, la probabilité de passer du triplet (0,1,0) au triplet (0,0,0) est nulle, car le second élément 1 du premier triplet doit nécessairement se retrouver comme premier élément du second triplet.

Nous obtenons la matrice de transition :

$$M = \begin{bmatrix} & E_1 & E_2 & E_3 & E_4 & A \\ E_1 & q & 0 & 0 & q & 0 \\ E_2 & p & 0 & 0 & p & 0 \\ E_3 & 0 & q & 0 & 0 & 0 \\ E_4 & 0 & 0 & q & 0 & 0 \\ A & 0 & p & p & 0 & 1 \end{bmatrix}.$$

Nous pouvons également voir la dynamique de la chaîne de Markov sur son graphe :

Figure 14 : Graphe de la chaîne de Markov pour l'exemple proposé



Remarque : afin de parcourir la séquence S dans son intégralité, $T - w = 3$ itérations sont nécessaires.

Pour obtenir la quantité recherchée, c'est-à-dire la probabilité P que la séquence S contienne un cluster au moins, il nous faut élever la matrice M à la puissance 3 puis la multiplier par le vecteur distribution initiale. Le dernier terme du vecteur ainsi obtenu nous donnera la quantité recherchée.

Preuve :

Notons $X_n = (x_n, x_{n+1}, x_{n+2})$ pour $0 \leq n \leq 3$: X_n renvoie à la composition de la fenêtre de balayage à la n^{ième} itération.

Notons $\Pi = (p_1, p_2, p_3, p_4, p_5)$ le vecteur distribution initiale.

- o La matrice de transition est la matrice :

$$M = (p_{ij})_{1 \leq i, j \leq 5} = (P(X_{n+1} = E_i | X_n = E_j))_{1 \leq i, j \leq 5}.$$

- o Pour tout $k \geq 1$, M^k est la matrice :

$$M^k = (p_{ij}^k)_{1 \leq i, j \leq 5} = (P(X_{n+k} = E_i | X_n = E_j))_{1 \leq i, j \leq 5},$$

matrice de transition en k pas.

Remarquons maintenant que la quantité qui nous intéresse est la probabilité $P(X_3 = A)$, que nous pouvons réécrire en tenant compte de la condition initiale :

$$P(X_3 = A) = \sum_{j=1}^5 P(X_3 = A, X_0 = E_j).$$

Il vient

$$P(X_3 = A) = \sum_{j=1}^5 P(X_3 = A | X_0 = E_j) P(X_0 = E_j) = \sum_{j=1}^5 p_{51}^3 p_j.$$

La somme que nous avons écrite plus haut est bien le dernier terme du vecteur produit de la matrice M^3 et du vecteur distribution initiale Π .

CQFD

Nous avons :

$$M^3 = \begin{bmatrix} q^3 & q^3 & q^3 & q^3 & 0 \\ pq^2 & pq^2 & pq^2 & pq^2 & 0 \\ pq^2 & 0 & pq^2 & pq^2 & 0 \\ pq^2 & 0 & 0 & pq^2 & 0 \\ p^2 + 2p^2q & p + pq & p + p^2q & p^2 + 2p^2q & 1 \end{bmatrix},$$

$$M^3 \Pi = \begin{bmatrix} q^6 + 3pq^5 \\ pq^5 + 3p^2q^4 \\ pq^5 + 2p^2q^4 \\ pq^5 + p^2q^4 \\ 1 + 2p^2q^2 - 3pq^2 - q^3 + 2p^2q^3 + p^3q^2 + 2p^2q^4 + 3p^3q^3 \end{bmatrix},$$

d'où

$$P = 1 + 2p^2q^2 - 3pq^2 - q^3 + 2p^2q^3 + p^3q^2 + 2p^2q^4 + 3p^3q^3.$$

Soit, en tenant compte du fait que $q = 1 - p$:

$$P = 9p^2 - 16p^3 + 9p^4 - p^6.$$

Nous retrouvons bien le même résultat qu'avec les trois approches précédentes.

IV.5. Conclusion

L'intérêt de cette section était double.

D'une part, il s'agissait de présenter l'approche par chaîne de Markov qui, à notre connaissance, est originale.

D'autre part, cela nous a permis de mettre en évidence la difficulté du problème. En effet, si nous avons pu appliquer différentes méthodes avec succès, c'est avant tout que l'exemple choisi était très simple, la période d'observation petite ($T = 6$) et son parcours rapide (quatre déplacements de la fenêtre de balayage). Malgré tout, même sur cet exemple simple, les calculs n'étaient pas immédiats, et nous imaginons bien que chacune de ces méthodes deviendra inexploitable avec des données plus réalistes.

Un des objectifs du deuxième chapitre est de montrer que la méthode de Markov permet de conduire à des algorithmes qui certes, ne mènent pas à des formules closes, mais qui constituent des approximations algorithmiques simples et efficaces (Berthon et al., 2008a ou Berthon et al., 2008b).

V. Balayage discret ou continu : modèles usuels d'intérêt pratique

Dans le paragraphe III, proposant une formulation mathématique au problème posé, nous avons donné deux exemples de modèles renvoyant chacun à une loi régissant l'arrivée des événements : un modèle continu dit « exponentiel renouvelé », et un modèle discret de Bernoulli. Nous les reprenons ici en détail et rappelons pour chacun d'eux la définition d'un (w,k) -cluster.

Nous proposons par ailleurs plusieurs variantes de ces modèles et précisons dans quels cas de figure l'une ou l'autre doit être utilisée.

Modèle continu « exponentiel renouvelé »

Soient :

- $Z_1, Z_2, \dots, Z_n, \dots$ des variables aléatoires de loi exponentielle de paramètre α .
- T un réel strictement positif : $T > 0$.
- w un réel inférieur ou égal à T : $w \leq T$.
- k un entier plus grand que 1 : $k \geq 1$.

Posons $X_m = Z_1 + Z_2 + \dots + Z_m$.

La suite de réalisations $\xi_1, \xi_2, \dots, \xi_n, \dots$ présente un (w,k) -cluster sur l'intervalle $[0, T[$ s'il existe $u \in [0, T - w]$ tel que l'intervalle $[u, u + w[$ contienne au moins k des valeurs des ξ_i .

D'un point de vue pratique, nous pouvons voir les Z_i comme les temps inter-événements, et X_m comme le temps écoulé avant le $m^{\text{ième}}$ événement. En d'autres termes, nous sommes en train de considérer un processus de Poisson de paramètre α .

Dans ce modèle, le nombre d'événements sur la période d'observation $[0, T[$ est inconnu a priori. C'est la situation que rencontre l'observateur qui, adoptant un point de vue prédictif, souhaite anticiper d'éventuels agrégats d'accidents, estimer le risque (la probabilité) d'occurrence d'un (w,k) -cluster. Nous parlerons de balayage « *prospectif* ».

Modèle de Poisson continu

Soient :

- T un réel strictement positif : $T > 0$.
- Soit N une variable aléatoire de loi de Poisson de paramètre γ et X_1, X_2, \dots, X_N des variables aléatoires indépendantes de loi uniforme sur $[0, T[$.
- w un réel inférieur ou égal à T : $w \leq T$.
- k un entier plus grand que 1 : $k \geq 1$.

La suite de réalisations $\xi_1, \xi_2, \dots, \xi_N$ présente un (w, k) -cluster sur l'intervalle $[0, T[$ s'il existe $u \in [0, T - w]$ tel que l'intervalle $[u, u + w[$ contienne au moins k des valeurs des ξ_i .

Dans ce modèle, le nombre d'évènements sur la période d'observation $[0, T[$ est la réalisation d'une variable aléatoire de Poisson. Les évènements se répartissent ensuite selon une loi uniforme sur $[0, T[$. Ce modèle correspond à la situation que rencontre un observateur qui, adoptant un point de vue « *rétrospectif* », souhaite analyser et expliquer un agrégat d'accidents qui se sont produits dans le passé.

Lorsque $\alpha T = \gamma$, le modèle exponentiel renouvelé et le modèle de Poisson continu sont équivalents dans le sens suivant : S_w et S'_w , statistiques de balayage pour le premier et le second modèle cité respectivement, ont même loi [Billingsley, 1986].

Modèle de Poisson discret

Soient :

- T un entier strictement positif : $T \geq 1$.
- Y_0, Y_1, \dots, Y_{T-1} des variables aléatoires indépendantes de loi de Poisson de paramètre λ , dont les réalisations sont notées y_0, y_1, \dots, y_{T-1} .
- w un entier inférieur ou égal à T : $w \in \{1, \dots, T\}$.
- k un entier plus grand que 1 : $k \geq 1$.

Pour $t \in \{0, 1, \dots, T-1\}$ et $y \in \mathbb{N}^*$, posons $X_{t,y} = \begin{cases} t & \text{si } 1 \leq y \leq Y_t \\ T+1 & \text{si } y > Y_t \end{cases}$, de réalisation $\xi_{t,y}$.

Avec cette notation, la valeur $t=i$ se produit exactement Z_i fois.

La suite de réalisations $(\xi_{t,y})_{t,y}$ présente un (w, k) -cluster sur l'intervalle $[0, T[$ s'il existe $u \in [0, T - w]$ tel que $y_u + \dots + y_{u+w-1} \geq k$.

Ce modèle est la version discrétisée du modèle exponentiel renouvelé.

Modèle de Bernoulli

Soient :

- T un entier strictement positif : $T \geq 1$.
- $p \in [0, 1]$

- Y_0, Y_1, \dots, Y_{T-1} des variables aléatoires indépendantes de loi de Bernoulli de paramètre p . Nous notons y_0, y_1, \dots, y_{T-1} leurs réalisations.
- w un entier inférieur ou égal à T : $w \in \{1, \dots, T\}$.
- k un entier plus grand que 1, plus petit que w : $1 \leq k \leq w$.

Posons $X_t = \begin{cases} t & \text{si } Y_t = 1 \\ T+1 & \text{si } Y_t = 0 \end{cases}$.

La suite de réalisations $\xi_0, \xi_1, \dots, \xi_{T-1}$ présente un (w, k) -cluster s'il existe $u \in [0, T - w]$ tel que $y_u + \dots + y_{u+w-1} \geq k$.

Ce modèle est une version approchée du modèle de Poisson discret.

Remarque : dans les deux modèles discrets, nous pouvons relier la variables aléatoire Y_u à la fonction de comptage introduite au paragraphe III. : nous avons en effet $Y_u = N_1(u)$ pour tout $0 \leq u \leq T - 1$.

L'utilisation de l'un ou l'autre des modèles proposés ci-dessus dépend essentiellement de la nature des données considérées, de la nature du balayage envisagé, et de la précision souhaitée.

Si les données sont discrètes, nous choisirons un des deux derniers modèles, Poisson discret ou Bernoulli, le deuxième cité étant une version approchée du premier.

Si les données sont continues, nous choisirons un des deux premiers modèles, exponentiel renouvelé ou Poisson continu, selon que l'on adopte un point de vue rétrospectif ou prospectif.

Le tableau 3 résume les caractéristiques des différents modèles proposés.

Tableau 3 : caractéristiques des modèles

<i>Modèle</i>	<i>Exponentiel renouvelé</i>	<i>Poisson continu</i>	<i>Poisson discret</i>	<i>Bernoulli</i>
<i>Nombre espéré d'évènements sur $[0, T[$</i>	αT	γ	λT	pT
<i>Continu ou discret</i>	Continu	Continu	Discret	Discret
<i>Caractéristiques</i>	Le nombre d'évènements sur $[0, T[$ est inconnu à priori	Le nombre d'évènements sur $[0, T[$ suit une loi de Poisson	<ul style="list-style-type: none"> - La période d'observation $[0, T[$ est discrétisée - Sur chaque unité de temps, le nombre d'évènements suit une loi de Poisson 	<ul style="list-style-type: none"> - La période d'observation $[0, T[$ est discrétisée - Sur chaque unité de temps, le nombre d'évènements suit une loi de Bernoulli (0 ou 1)

Chapitre 2

Statistiques de balayage :
Les méthodes

Le problème que nous souhaitons résoudre se pose de la manière suivante : nous observons des événements sur une période de longueur T , dite *période d'observation*, et nous souhaitons évaluer la fréquence d'apparition d'un (w,k) -cluster qui est, rappelons-le, un agrégat de k événements sur une *fenêtre* de longueur w .

Chacune des méthodes proposée dans ce chapitre sera présentée dans un cadre général avant d'être déclinée à un même exemple. Nous ne reprendrons pas l'exemple des accidents aériens évoqué dans le premier chapitre, exemple trop complexe (en termes de taille des données) pour le modèle markovien dit « complet » présenté ici. Toutefois, nous précisons que les résultats que nous avons pu voir dans la littérature ainsi que ceux obtenus avec certaines de nos méthodes donnent une probabilité légèrement supérieure à 11%.

L'exemple que nous choisissons est le suivant :

Exemple.

La période d'observation est une année de 365 jours. La probabilité recherchée est la probabilité d'observer un $(10,3)$ -cluster, c'est-à-dire un agrégat de 3 événements en moins de 10 jours. Le choix de la loi régissant l'arrivée des événements, et du modèle sous-jacent, reposera sur l'information suivante : la fréquence du type d'accident considéré est, « en temps normal », de 8 par an. Soit : $T = 365$, $w = 10$, $k = 3$ et une fréquence, en jours, de $\frac{8}{365}$.

I. Résolution par simulation numérique

Dans cette partie, nous montrons comment il est possible de résoudre le problème posé en ayant recours à des méthodes de simulation numérique. Nous proposons tout d'abord une simulation de Monte-Carlo directe, puis un réseau de Petri animé par simulation de Monte-Carlo.

Nous nous placerons successivement dans le cadre du modèle de Bernoulli et du modèle exponentiel renouvelé.

I.1. Simulation de Monte-Carlo directe

La première approche consiste à obtenir une estimation numérique de la probabilité d'apparition d'un (w, k) -cluster sur la période d'observation en réalisant une simulation de Monte-Carlo.

I.1.a. Principe

Nous rappelons brièvement le principe de la simulation de Monte-Carlo : soit g une fonction d'une variable aléatoire X dont on souhaite estimer l'espérance $\theta = E(g(X))$.

La simulation de Monte-Carlo consiste à générer un échantillon (x_1, x_2, \dots, x_N) de réalisations de X et à estimer la quantité recherchée à partir de cet échantillon :

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^N g(x_i).$$

L'échantillon est généré de manière aléatoire suivant une loi de probabilité spécifiée. Autrement dit, les x_i sont les réalisations de variables aléatoires X_i indépendantes et identiquement distribuées selon ladite loi.

Remarque : l'estimateur de Monte-Carlo est simplement construit à partir de la moyenne empirique, que l'on sait être un estimateur sans biais de l'espérance.

I.1.b. Application au cas étudié

Nous souhaitons évaluer la probabilité que, sur une période de longueur T , il existe (au moins) une fenêtre de longueur (au plus) t contenant (au moins) k accidents.

La simulation de Monte-Carlo consiste donc dans un premier temps à générer aléatoirement des dates d'accidents sur une durée T . Cet échantillon « fictif » de dates étant tiré, il est facile de contrôler la présence d'un cluster, c'est-à-dire d'un groupement de k accidents sur une durée inférieure à t .

La probabilité recherchée est alors estimée de la manière suivante :

$$\theta = \frac{\text{Nombre de tirages comportant au moins un cluster}}{N}$$

où N est le nombre de tirages réalisés.

L'algorithme de la simulation est le suivant :

Nb_Cluster = 0 .

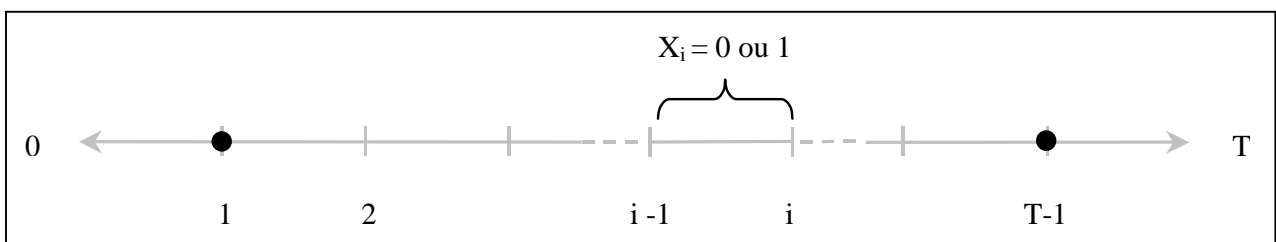
1. Tirer au hasard les dates d'accidents $\varepsilon_1, \dots, \varepsilon_s$ sur la période d'observation $[0, T[$:
 $s = \operatorname{argmax}_i \varepsilon_i \leq T$.
2. Balayer la liste des dates d'accidents ainsi générée. S'il y a un (ou des) cluster(s), incrémenter Nb_Cluster de 1.
3. Itérer N fois les deux pas précédents.
4. $p = \frac{\text{Nb_Cluster}}{N}$

I.1.c. Modèle de Bernoulli

La période d'observation $[0, T[$ étant exprimée selon l'unité de temps choisie (minute, heure, jour, ...), nous considérons les instants entiers $\{0, 1, \dots, T-1\}$ et nous regardons le nombre d'évènements à chacun de ces instants. Ce nombre suit une loi de Bernoulli de paramètre p : il vaut 1 avec la probabilité p ou 0 avec la probabilité $q = 1 - p$.

Parallèlement, le nombre d'évènements se produisant sur une fenêtre de longueur t , disons $\{j, j+1, \dots, j+t-1\}$, suit une loi binomiale de paramètres t et p .

Figure 15 : discrétisation de la fenêtre d'observation



Génération aléatoire des variables de Bernoulli

Soit Y_i la variable aléatoire binaire qui vaut 1 ou 0 selon qu'un évènement se produise ou non à l'instant i .

Pour tout $i \in \{0, 1, \dots, T-1\}$, nous générons un nombre aléatoire u_i selon la loi Uniforme sur $[0, 1]$. Si $u_i \leq p$, nous affectons la valeur 1 à la variable aléatoire Y_i , ce qui signifie qu'un évènement se produit à l'instant i ; sinon nous lui affectons la valeur 0, ce qui signifie qu'il n'y a pas d'accident à l'instant i :

$$\text{Pour } i \in \{0, 1, \dots, T-1\}, Y_i = \begin{cases} 0 & \text{si } u_i > p \\ 1 & \text{si } u_i \leq p \end{cases}.$$

On obtient bien ainsi la réalisation d'une variable de Bernoulli de paramètre p .

Récupération des dates d'évènements

Soit $Y = (Y_0, \dots, Y_{T-1})$ le vecteur binaire dont la $i^{\text{ème}}$ composante vaut 1 si un évènement s'est produit à l'instant i et 0 sinon. Nous lui faisons correspondre le vecteur de taille variable V contenant les « dates » d'évènements.

Exemple

Supposons que la période d'observation soit de 365 jours et qu'une simulation ait produit le vecteur binaire $Y = (0, \dots, 0, 1, 1, 0, \dots, 0, 1, 0, \dots, 0, 1, 1, 1, 0, 0, 0)$, la liste des dates d'évènements est donnée par le vecteur $V = (17, 18, 120, 360, 361, 362)$.

Algorithme de la simulation

La simulation a été réalisée sous Visual Basic et les résultats renvoyés dans un classeur Excel. Bien qu'il soit élémentaire, nous présentons ci-dessous l'algorithme :

Pour i allant de 1 à N

 indice_deb=1

 indice_fin=k

 CPT_ACC = 0

! Génération aléatoire des dates d'accidents

 Pour j allant de 1 à T

$u = \text{Rnd}()$

 Si $u \leq p$ Alors

$Y_j=1$

$\text{CPT_ACC} = \text{CPT_ACC} + 1$

$V_{\text{cpt}}=j$

 Sinon $Y_j=0$

! Parcours de la liste des dates d'accidents et identification des clusters

 Tant que $\text{ind_fin} \leq \text{CPT_ACC}$

 Si $V_{\text{indice_fin}} - V_{\text{indice_deb}} \leq t$ Alors

$\text{CPT_YES} = 1$

$\text{indice_fin} = \text{indice_fin} + 1$

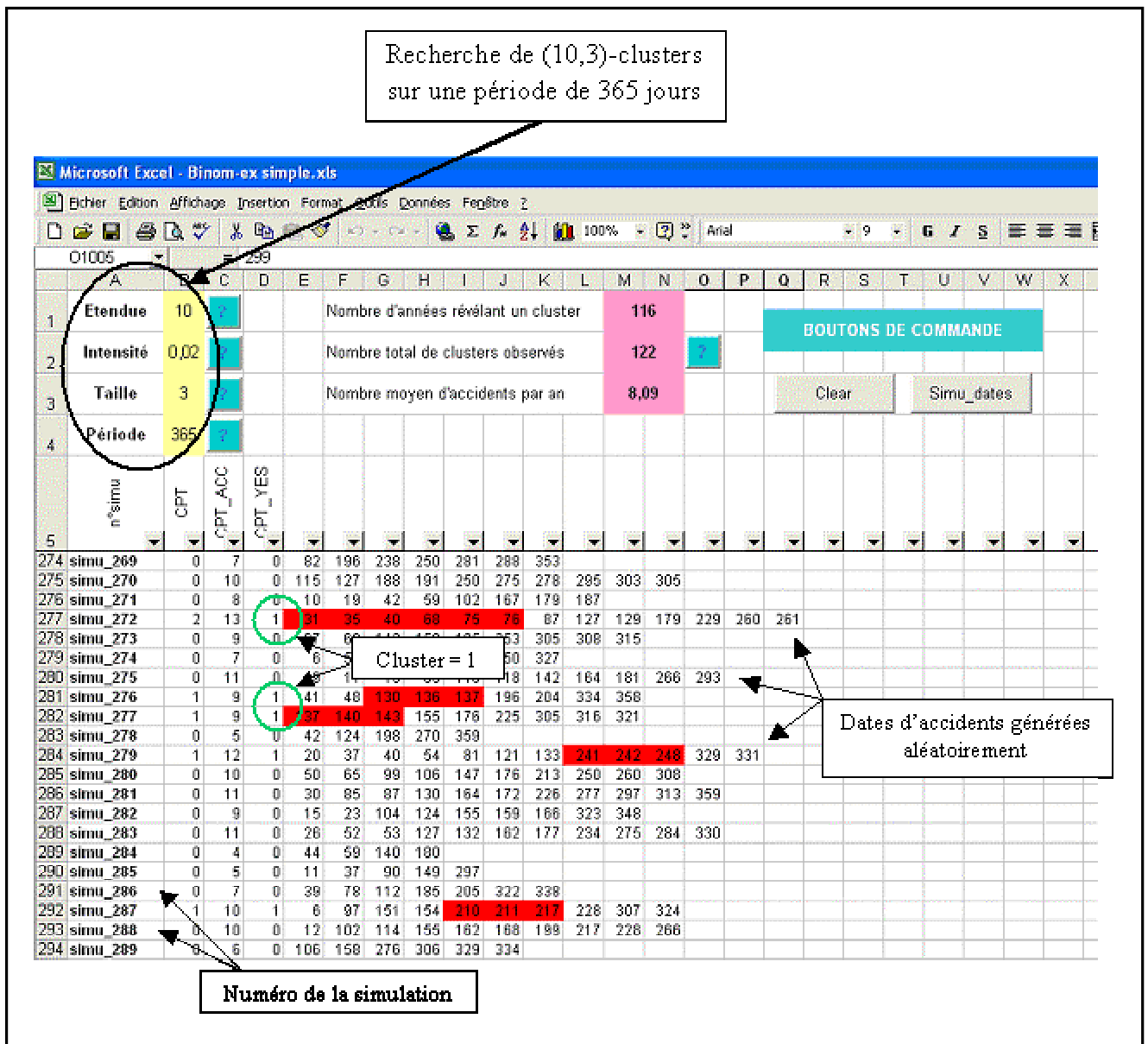
$\text{indice_deb} = \text{indice_deb} + 1$

 Sinon $\text{CPT_YES} = 0$

Le programme détaillé construit sous Visual Basic est donné en annexe. Il est légèrement différent de l'algorithme ci-dessus. Nous y avons rajouté des lignes de programme permettant d'évaluer le nombre de clusters observés sur chaque période d'observation (CPT) puisqu'il arrive que plusieurs agrégats se succèdent.

La capture d'écran ci-dessous montre l'interface graphique développée sous Excel.

Figure 16 : algorithme de la simulation sous Excel



Remarque : Pour une meilleure visualisation des clusters, l'algorithme stocke une à une les dates d'évènements couvrant la période d'observation avant de les balayer pour contrôler la présence de clusters. Bien évidemment, garder en mémoire les k dernières dates serait suffisant et aussi simple à coder.

I.1.d. Modèle exponentiel renouvelé

Les évènements se répartissent sur la période d'observation $[0,T]$ selon un processus de Poisson de paramètre λ . Autrement dit, la distribution des temps entre deux évènements consécutifs est exponentielle de paramètre λ et le nombre d'accidents sur une période de longueur t suit une loi de Poisson de paramètre λt .

Rappelons les caractéristiques de la loi exponentielle :

Densité : $f(t) = \lambda e^{-\lambda t}$.

Fonction de répartition : $F(t) = 1 - e^{-\lambda t}$.

Valeur moyenne : $\frac{1}{\lambda}$.

Génération aléatoire des évènements

Pour $i \geq 1$, Y_i est la variable aléatoire représentant le temps écoulé entre l'évènement i et l'évènement $i+1$ et y_i sa réalisation ; $X_m = Y_1 + \dots + Y_m$ est le temps écoulé avant l'évènement m et x_m sa réalisation.

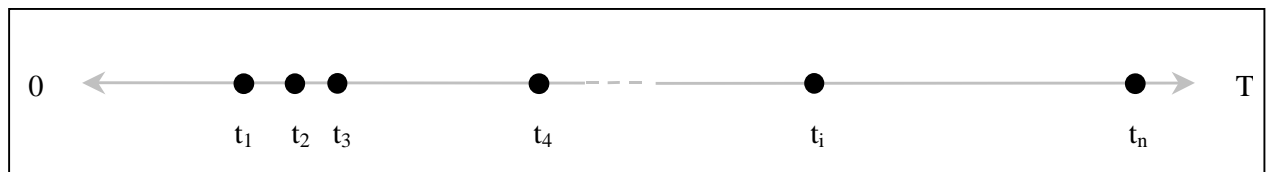
Les dates d'évènements sont générés de la manière suivante :

- la première selon la formule $x_1 = \frac{1}{\lambda}(-\ln(u))$,
- les suivantes selon la formule $x_{i+1} = x_i + \frac{1}{\lambda}(-\ln(u))$,

où u est la réalisation d'une variable aléatoire uniforme sur $[0,1]$, et tant que la valeur générée x_i est inférieure strictement à T .

On obtient bien ainsi un processus de Poisson de paramètre λ .

Figure 17 : génération des dates d'accidents



Algorithme de la simulation

L'algorithme de la simulation réalisée sous Visual Basic est le suivant :

Pour i allant de 1 à N

 indice_deb=1

 indice_fin=k

 CPT_ACC = 0

 j = 1

 ! Génération aléatoire de la première date d'accident

 dat_acc = $\frac{1}{\lambda}(-\ln(\text{Rnd}()))$

 tj=dat_acc

 j=j+1

! Génération aléatoire des dates suivantes

Tant que $\text{dat_acc} \leq T$

$$\text{dat_acc} = \text{dat_acc} + \frac{1}{\lambda} (-\ln(\text{Rnd}()))$$

$t_j = \text{dat_acc}$

$j=j+1$

! Parcours de la liste des dates d'accidents et identification des clusters

Tant que $\text{ind_fin} \leq \text{CPT_ACC}$

Si $t_{\text{indice_fin}} - t_{\text{indice_deb}} \leq t$ Alors

$\text{CPT_YES} = 1$

$\text{indice_fin} = \text{indice_fin} + 1$

$\text{indice_deb} = \text{indice_deb} + 1$

Sinon $\text{CPT_YES} = 0$

Le programme détaillé de la simulation, construit sous Visual Basic est donné en annexe.

Remarque : il est temps de faire une remarque importante. Dans le cas du modèle exponentiel renouvelé, nous avons généré les dates d'accidents en continu et nous avons parcouru la période d'observation en continu. Or, toutes les méthodes markoviennes que nous proposons par la suite sont discrétisées. Ainsi, la probabilité que nous obtiendrons par simulation et selon le modèle précédemment cité sera la plus élevée, et c'est elle qui nous servira de repère. Malgré tout, et afin de pouvoir juger de la pertinence des méthodes markoviennes, nous avons réalisé une simulation discrète avec respectivement un pas de discrétisation d'un jour et d'une heure. Les résultats sont présentés dans le tableau ci-dessous.

I.1.e. Résultats de la simulation de Monte Carlo classique

Le tableau suivant présente les résultats que nous avons obtenu pour l'exemple présenté plus haut - $T = 365$, $(w, k) = (10, 3)$, fréquence (p ou λ) = $\frac{8}{365}$ - dans le cas binomial et dans le cas poissonnien.

Tableau 4 : Valeur moyenne et écart-type pour un échantillon de 100 jeux de données comportant 1000 simulations chacun.

Modèle	Bernoulli		Poisson		
	1 jour	1 heure	1 jour	1 heure	continu
Moyenne	0.1025	0.1310	0.1220	0.1301	0.1329
Ecart-type	0.0097	0.0097	0.0091	0.0090	0.0095
Intervalle de confiance ² à 95%	0.1025 ± 0.0019	0.1310 ± 0.0019	0.1220 ± 0.0018	0.1301 ± 0.0018	0.0095 ± 0.0019

² Intervalle de confiance calculé selon la formule $\bar{X} \pm t_{1-\alpha/2} \frac{s}{\sqrt{n}}$ où $\alpha = 0.05$, $t_{1-\alpha/2} = 1.96$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite, $n = 100$, \bar{X} est la moyenne et s l'écart-type.

I.2. Réseau de Petri animé par simulation de Monte Carlo

La pertinence de l'introduction d'un modèle réseaux de Petri (RdP) animé par simulation de Monte-Carlo se pose immédiatement : quel serait son apport par rapport à la précédente démarche basée sur la même technique ?

La réponse tient en deux points : d'une part, le formalisme des RdP facilite la mise en oeuvre des simulations mais aussi le repérage des clusters (nous l'expliquons brièvement dans les paragraphes ci-dessous) ; d'autre part, il existe sur le marché des logiciels performants dédiés aux réseaux de Petri.

Le modèle de réseau de Petri (RdP) que nous présentons ici appartient à la catégorie des réseaux de Petri stochastiques à prédicats [Dutuit et al., 1997] [Malhotra et Trivedi, 1995]. Ces réseaux de Petri sont semblables aux RdP traditionnels, constitués de places (cercles) et de transitions (traits), mais ont une capacité de calcul étendue.

Par exemple, des attributs peuvent être alloués aux différentes transitions. Ils peuvent être des messages pré-conditionnels (identifiés dans notre modèle par un double point d'interrogation) autorisant la transition, ou des messages post-conditionnels (identifiés dans notre modèle par un double point d'exclamation) qui mettent à jour les valeurs des variables utilisées dans le modèle.

Le RdP présenté sur les graphes 1 et 2 ont été construits pour vérifier si, dans une période d'observation de longueur T donnée contenant divers évènements, il existait des groupements d'au moins k dans une fenêtre de longueur w.

Figure 18 : modèle RdP relatif au modèle de Bernoulli

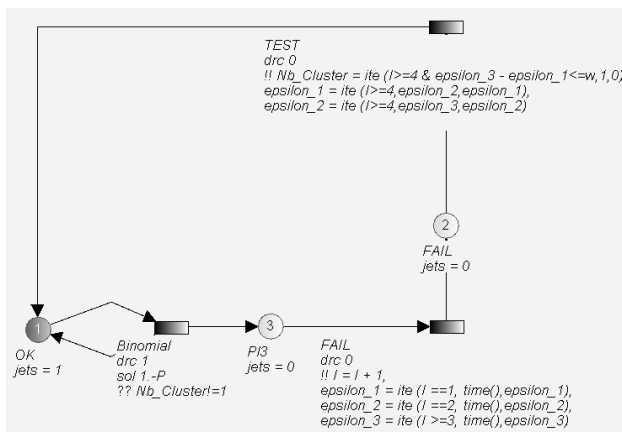
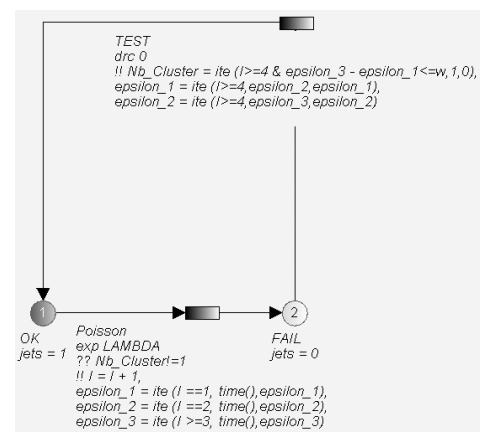


Figure 19 : modèle RdP relatif au modèle exponentiel renouvelé



Le RdP de la figure 19 est constitué de 2 places et 2 transitions. Initialement, seule la place 1 est marquée d'un jeton.

- Les variables ϵ_{i-1} (i allant de 1 à k) indiquent les instants de réalisation de k accidents successifs, ces instants étant « tirés » aléatoirement suivant la loi spécifiée ;
- L'index I permet de calculer, de manière continue, le temps écoulé entre les occurrences i et (i+k-1) ;
- La variable Nb_Cluster est initialisée à zéro, et devient égale à 1 dès que les k accidents sont contenus dans une période de longueur inférieure ou égale à w.

- Afin d'estimer la probabilité d'occurrence d'un (w,k) -cluster, l'évolution du RdP sur une période de longueur T est itérée un grand nombre de fois ($N = 10^6$) et la valeur de la variable $Nb_Cluster$, qui est soit 0 soit 1, vient s'ajouter à celle résultant de l'histoire précédente.

Après N itérations, la valeur attendue est estimée par le quotient $(Nb_Cluster / N)$.

Pour l'exemple considéré :

$$\begin{cases} T = 365 \\ (w, k) = (10, 3) \\ \text{fréquence } (\lambda \text{ ou } p) = \frac{8}{365} \end{cases},$$

nous obtenons les résultats résumés dans le tableau ci-dessous :

Tableau 5 : Résultats obtenus par RdP

Modèle	Bernoulli		Poisson
Discrétisation	1 jour	1 heure	continu
Moyenne	0.1225	0.1252	0.1317

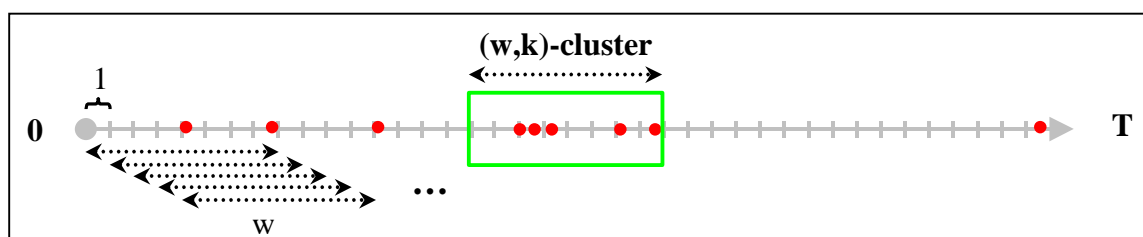
II. Les méthodes markoviennes

Dans les paragraphes qui suivent, nous proposons plusieurs approches markoviennes permettant de modéliser le problème de l'évaluation de la probabilité d'occurrence d'un (w,k) -cluster sur une période de temps donnée. Ces approches sont semblables à celle que nous avons appliquée sur l'exemple simple du premier chapitre.

Pour chacune d'entre elles, le principe est identique : nous balayons la période d'observation par glissement d'une fenêtre de longueur w ; nous arrêtons soit lorsque la fenêtre de balayage contient un agrégat de k évènements, soit lorsque nous avons atteint la fin de la période d'observation.

La période d'observation est discrétisée : à chaque itération, nous la déplaçons d'une unité selon le pas choisi : de jour en jour si l'on raisonne en jours, d'heure en heure si l'on raisonne en heures, etc...

Figure 20 : balayage de la fenêtre d'observation



\mathcal{E} , l'ensemble des états, est constitué :

- d'une part, de tous les éléments codant des fenêtres qui ne comportent pas de cluster, que nous noterons E_0, E_1, \dots, E_{r-1} ;
- d'autre part, d'un élément noté A ou E_r codant n'importe quelle fenêtre comportant un cluster.

Le nombre d'états est $r+1$. La variable r prendra différentes valeurs suivant l'approche considérée.

Pour chacune des trois approches proposées, nous appliquerons successivement le **modèle de Bernoulli** et le **modèle de Poisson discret**.

Rappelons tout d'abord les notations qui seront communes aux trois approches:

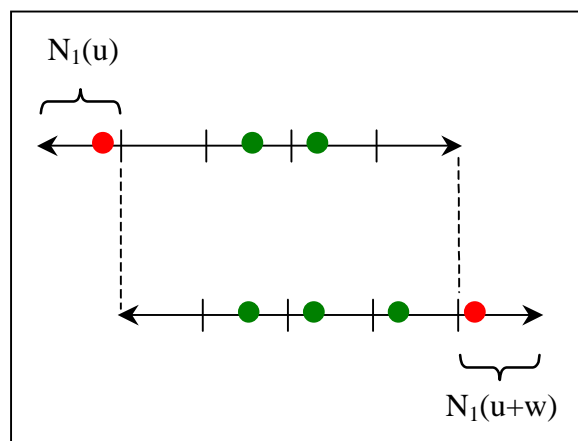
- ✓ La période d'observation est l'intervalle ouvert $[0, T[$.
- ✓ La fenêtre de balayage est de taille w , $0 < w \leq T$.
- ✓ Le nombre d'évènements qui nous intéresse est $k \in \mathbb{N}^*$.
- ✓ Pour $u \in [0, T - t]$, l'objet « fenêtre » de longueur t démarrant en u est noté $C_t(u)$.
- ✓ Pour $u \in [0, T - t]$, le nombre d'évènements dans la fenêtre $C_t(u)$ est :

$$N_t(u) = \text{Card} \{i \in I \mid X_i \in [u, u + t[\} = \text{Card} \{n \mid X_{(n)} \in [u, u + t[\} .$$
- ✓ Le nombre de déplacements nécessaires pour parcourir l'intégralité de la période d'observation est :

$$N = T - w .$$

Regardons maintenant la dynamique d'une transition:

Figure 21 : dynamique d'une transition



Le passage de la fenêtre $C_w(u)$ à la fenêtre $C_w(u+1)$ se fait en "perdant" la variable aléatoire $N_1(u)$ et en "gagnant" la variable aléatoire $N_1(u+w)$:

$$N_w(u+1) = N_w(u) - N_1(u) + N_1(u+w) ,$$

où $N_1(u+w)$ est indépendante de $N_w(u)$ tandis que $N_1(u)$ en est dépendante.

Pour les deux modèles discrets que nous considérons, le nombre d'évènements à l'instant u (modèle de Bernoulli) ou sur la $u^{\text{ième}}$ unité de temps (modèle de Poisson discret) est noté Y_u . Nous pouvons donc réécrire : $N_w(u+1) = N_w(u) - Y_u + Y_{u+w}$.

Remarque : le fait que les variables aléatoires Y_i sont indépendantes et identiquement distribuées assurera l'homogénéité des chaînes de Markov considérées : le taux de transition sera en effet constant dans le temps.

II.1. Première approche

Nous proposons une chaîne de Markov discrète dont les états renseignent sur le nombre d'évènements dans la fenêtre de balayage : la chaîne visite l'état E_i lorsque la fenêtre courante comporte exactement i accidents.

- *Etats*

Nous considérons une chaîne de Markov à $r = k + 1$ états décrits ci-dessous :

- Etat E_0 : la fenêtre courante ne contient aucun évènement;
- Etat E_1 : la fenêtre courante contient un évènement exactement;
- Etat E_2 : la fenêtre courante contient deux évènements exactement;
- ... ;
- Etat $A=E_k$: la fenêtre courante contient au moins k évènements.

L'état A est absorbant. C'est ce dernier état qui nous intéresse puisque lorsque la chaîne de Markov l'atteint, cela signifie que la fenêtre courante contient un (w,k) -cluster.

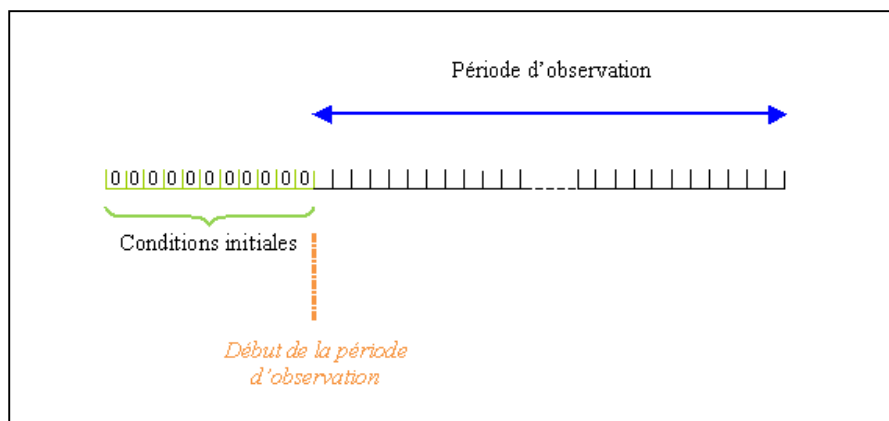
- *Condition initiale*

Il semble logique d'adopter comme condition initiale la composition de la première fenêtre rencontrée : $[0,w[$. Cette fenêtre peut contenir indifféremment $0, 1, 2, \dots, k$ ou plus de k évènements, ce qui revient à dire que la chaîne est initialement en n'importe lequel des états E_i , $0 \leq i \leq k$. Cela conduit au vecteur conditions initiales :

$$\begin{bmatrix} P(N_w(0) = 0) \\ P(N_w(0) = 1) \\ \dots \\ P(N_w(0) = k-1) \\ P(N_w(0) \geq k) \end{bmatrix}.$$

Mais nous pouvons simplifier la condition initiale en introduisant une fenêtre fictive, de longueur w , à « gauche » de la période d'observation.

Figure 22 : Hypothèse pour la condition initiale



Cette fenêtre correspond au w -uplet de variables aléatoires $(Y_{-w}, Y_{-w+1}, \dots, Y_{-1})$ dont chaque composante prend, par convention, la valeur zéro. Au début de l'observation, la fenêtre de balayage ne contient alors aucun évènement et se trouve en l'état E_0 avec la probabilité 1. Le vecteur conditions initiales est alors :

$$\Pi = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}.$$

C'est ce vecteur que nous prendrons par la suite.

Remarque : la fenêtre de balayage étant initialement positionnée à « gauche » de la période d'observation, le nombre de déplacement nécessaires pour parcourir l'intégralité de la période d'observation n'est plus $N = T - w$ mais $N = T$.

- *Matrice de transition*

C'est la matrice M dont l'élément m_{ij} est la probabilité de transition de l'état E_j vers l'état E_i . Elle s'écrit :

$$m_{ij} = P\{N_w(u+1) = i | N_w(u) = j\},$$

et dépend de la loi régissant l'arrivée des évènements.

Les probabilités de transition sont calculées dans les paragraphes qui suivent pour les deux modèles considérés : le **modèle de Bernoulli** et le **modèle de Poisson discret**.

La probabilité recherchée sera donnée par le produit $M^T \Pi$.

II.1.a. Modèle de Bernoulli

Soit $p \in [0,1]$.

La variable aléatoire $Y_u = N_1(u)$ compte le nombre d'évènements à l'instant u . C'est une épreuve de Bernoulli de paramètre p , c'est-à-dire qu'elle prend la valeur 1 avec la probabilité p ou la valeur 0 avec la probabilité $q = 1 - p$. Soit :

- ✓ Probabilité qu'il y ait un évènement à l'instant u : $P\{Y_u = 1\} = p$;
- ✓ Probabilité qu'il n'y ait aucun évènement à l'instant u : $P\{Y_u = 0\} = q$.

Nous supposons que les variables aléatoires Y_u , $0 \leq u \leq w-1$, sont deux à deux indépendantes.

Le nombre d'évènements se produisant sur une période de longueur t suit alors une loi binomiale de paramètres t et p . Soit :

- ✓ Pour tout $u \in [0, T-t]$: $P\{N_t(u) = i\} = \binom{t}{i} p^i (1-p)^{t-i}$.

Nous reprenons ci-dessous, pour le modèle de Bernoulli, les trois points évoqués dans le paragraphe précédent : états, condition initiale et matrice de transition.

- *Etats*

Nous avons k états transitoires E_i , $0 \leq i \leq k-1$, et un état absorbant $A = E_k$.

Remarquons que, puisqu'il y a au plus 1 évènement par unité de temps, le nombre d'évènements k constituant la fenêtre de balayage est nécessairement inférieur ou égal à la longueur de la fenêtre d'observation : $k \leq w$.

- *Condition initiale*

Selon le choix établi au paragraphe précédent, le vecteur conditions initiales est le vecteur de taille $k+1$:

$$\Pi = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}.$$

- *Matrice de transition*

Les transitions réalisables sont :

- Les transitions d'un état E_i , $i=0, \dots, w-1$, vers un état E_{i+1} qui se produisent lorsque la variable aléatoire Y_u vaut 0 et que la variable aléatoire Y_{u+w} vaut 1. La probabilité d'une telle transition est

$$m_{i+1,i} = p \times \left(1 - \frac{i}{w}\right);$$

- Les transitions d'un état E_i , $i=0, \dots, w-1$, vers un état E_i , ou boucles sur l'état E_i . Ces transitions se produisent dans deux cas de figure : soit lorsque la variable aléatoire Y_u et la variable aléatoire Y_{u+w} valent simultanément 1, soit lorsqu'elles valent simultanément 0, de sorte qu'au final le nombre d'évènements dans la fenêtre courante reste inchangé. La probabilité d'une telle transition est

$$m_{i,i} = p \times \left(\frac{i}{w}\right) + q \times \left(1 - \frac{i}{w}\right);$$

Pour $i=k$, nous avons par convention $m_{k,k} = 1$.

- Les transitions d'un état E_i , $i=1, \dots, w$, vers un état E_{i-1} qui se produisent lorsque la variable aléatoire Y_u vaut 1 et que la variable aléatoire Y_{u+w} vaut 0. La probabilité d'une telle transition est

$$m_{i-1,i} = q \times \left(\frac{i}{w}\right).$$

Preuve

Etape 1. Nous commençons par évaluer la probabilité qu'il y ait un accident sur la première unité de temps de la fenêtre de balayage connaissant le nombre total d'accidents sur la fenêtre :

$$P(Y_u = 1 | N_w(u) = n) = \frac{n}{w},$$

et, respectivement, la probabilité qu'il n'y ait aucun accident sur la première unité de temps de la fenêtre de balayage connaissant le nombre total d'accidents sur la fenêtre :

$$P(Y_u = 0 | N_w(u) = n) = 1 - \frac{n}{w}.$$

En effet, nous avons

$$P(Y_u = 1 | N_w(u) = n) = P(Y_u = 1 | Y_u + \dots + Y_{u+w-1} = n),$$

que nous pouvons réécrire :

$$P(Y_u = 1 | N_w(u) = n) = \frac{P(Y_u = 1, Y_{u+1} + \dots + Y_{u+w-1} = n-1)}{P(Y_u + Y_{u+1} + \dots + Y_{u+w-1} = n)}.$$

Les variables aléatoires Y_i étant indépendantes, identiquement distribuées selon une loi de Bernoulli, il vient :

$$P(Y_u = 1 | N_w(u) = n) = \frac{p \times \binom{w-1}{n-1} \times p^{n-1} \times (1-p)^{w-n}}{\binom{w}{n} \times p^n \times (1-p)^{w-n}}.$$

D'où finalement :

$$P(Y_u = 1 | N_w(u) = n) = \frac{n}{w}.$$

Comme

$$P(Y_u = 0 | N_w(u) = n) = 1 - P(Y_u = 1 | N_w(u) = n),$$

nous avons encore

$$P(Y_u = 0 | N_w(u) = n) = 1 - \frac{n}{w}.$$

Etape 2. Pour la transition d'un état E_i vers un état E_{i+1} , nous écrivons :

$$m_{i+1,i} = P(N_w(u+1) = i+1 | N_w(u) = i).$$

Il vient naturellement :

$$\begin{aligned} m_{i+1,i} &= \frac{P(Y_u = 0, Y_{u+w} = 1, N_w(u) = i)}{P(N_w(u) = i)} \\ &= \frac{P(Y_{u+w} = 1)P(Y_u = 0, N_w(u) = i)}{P(N_w(u) = i)} \\ &= P(Y_{u+w} = 1)P(Y_u = 0 | N_w(u) = i) \\ &= p \times \left(1 - \frac{i}{w}\right). \end{aligned}$$

Etape 3. Pour la boucle sur l'état i , nous écrivons :

$$m_{i,i} = P(N_w(u+1) = i | N_w(u) = i).$$

Il vient naturellement :

$$\begin{aligned}
m_{i,i} &= \frac{P(Y_u = 1, Y_{u+w} = 1, N_w(u) = i)}{P(N_w(u) = i)} + \frac{P(Y_u = 0, Y_{u+w} = 0, N_w(u) = i)}{P(N_w(u) = i)} \\
&= \frac{P(Y_{u+w} = 1)P(Y_u = 1, N_w(u) = i)}{P(N_w(u) = i)} + \frac{P(Y_{u+w} = 0)P(Y_u = 0, N_w(u) = i)}{P(N_w(u) = i)} \\
&= P(Y_{u+w} = 1)P(Y_u = 1|N_w(u) = i) + P(Y_{u+w} = 0)P(Y_u = 0|N_w(u) = i) \\
&= p \times \left(\frac{i}{w}\right) + q \times \left(1 - \frac{i}{w}\right).
\end{aligned}$$

Étape 4. Pour la transition d'un état E_i vers un état E_{i-1} , nous écrivons :

$$m_{i-1,i} = P(N_w(u+1) = i | N_w(u) = i-1).$$

Il vient naturellement :

$$\begin{aligned}
m_{i-1,i} &= \frac{P(Y_u = 1, Y_{u+w} = 0, N_w(u) = i)}{P(N_w(u) = i)} \\
&= \frac{P(Y_{u+w} = 0)P(Y_u = 1, N_w(u) = i)}{P(N_w(u) = i)} \\
&= P(Y_{u+w} = 0)P(Y_u = 1|N_w(u) = i) \\
&= q \times \left(\frac{i}{w}\right).
\end{aligned}$$

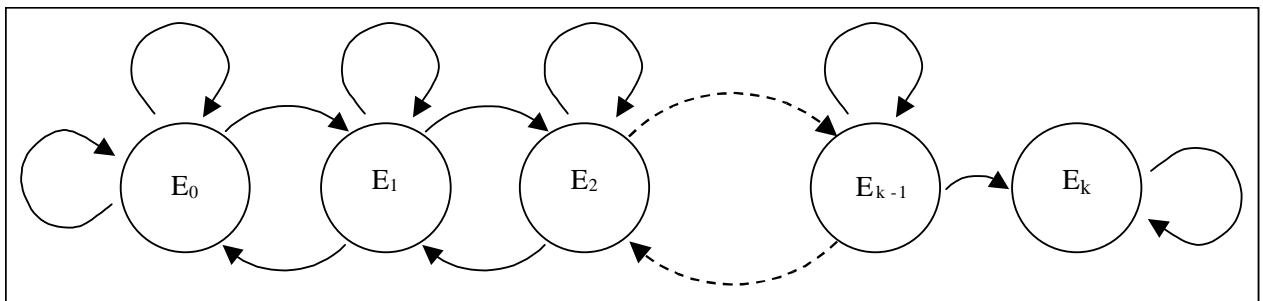
CQFD

Remarque : il suffit en fait de calculer les deux probabilités les plus simples, disons $m_{i-1,i}$ et $m_{i+1,i}$, la troisième $m_{i,i}$ en découlant facilement puisque $m_{i-1,i} + m_{i,i} + m_{i+1,i} = 1$.

La matrice de transition est donc une matrice tridiagonale. Les termes de la diagonale principale sont les $m_{i,i}$ définis plus haut ; ceux de la diagonale supérieure sont les $m_{i-1,i}$ et ceux de la diagonale inférieure sont les $m_{i+1,i}$.

Nous pouvons voir la dynamique de la chaîne de Markov sur son graphe :

Figure 23 : graphe de la chaîne de Markov pour le modèle binomial, premier modèle markovien



La probabilité recherchée sera donnée par le produit $M^T \Pi$.

II.1.b. Modèle de Bernoulli : application à l'exemple (T=365, w=10, k=3)

Nous allons appliquer la méthode à l'exemple cité plus haut, avec successivement un pas de discrétisation d'un jour, puis d'une heure.

La chaîne de Markov correspondant à notre exemple est une chaîne à 4 états :

- les états notés E_0 , E_1 et E_2 sont des états transients et correspondent respectivement aux cas où la fenêtre courante comporte 0, 1 et 2 accidents ;
- l'état noté E_3 est absorbant et correspond au cas où la fenêtre courante comporte au moins 3 accidents.

Le vecteur des conditions initiales est le vecteur de taille 4

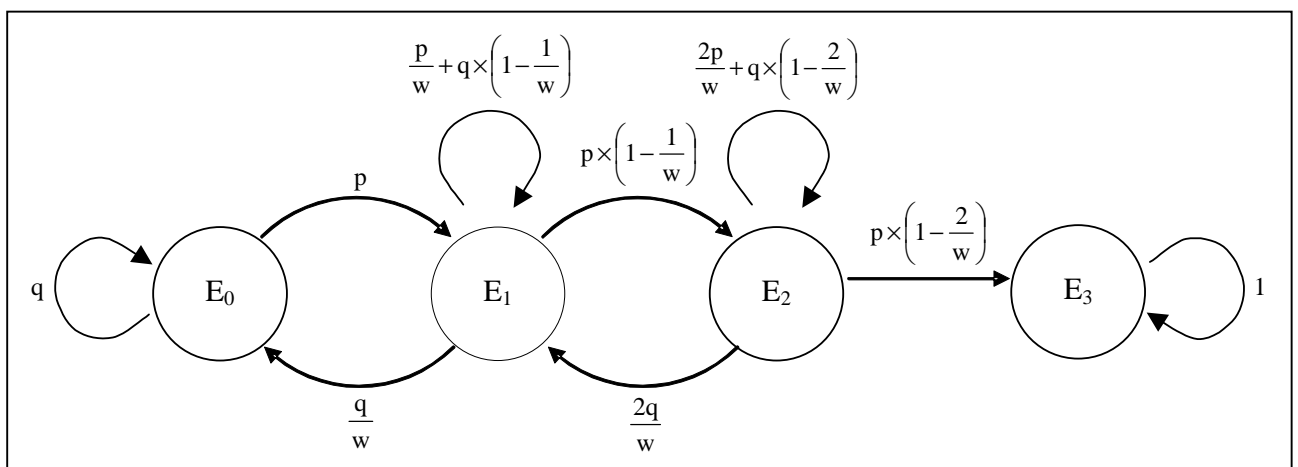
$$\Pi = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

La matrice de transition est la matrice :

$$M = \begin{bmatrix} q & \frac{q}{w} & 0 & 0 \\ p & \frac{p}{w} + q \left(1 - \frac{1}{w}\right) & \frac{2q}{w} & 0 \\ 0 & p \left(1 - \frac{1}{w}\right) & \frac{2p}{w} + q \left(1 - \frac{2}{w}\right) & 0 \\ 0 & 0 & p \left(1 - \frac{2}{w}\right) & 1 \end{bmatrix}.$$

Le graphe de la chaîne de Markov a l'allure suivante :

Figure 24 : Chaîne de Markov : modèle binomial, premier modèle markovien, T=365, w=10, k=3



La probabilité recherchée est donnée par le dernier terme du vecteur $M^T \Pi$.

Application numérique

Avec pas de discrétisation d'un jour

Rappelons que nous avons :

$$\begin{cases} T = 365 \\ (w, k) = (10, 3) \\ \text{fréquence } (\lambda \text{ ou } p) = \frac{8}{365} \end{cases}$$

Nous obtenons :

$$M = \begin{bmatrix} 0,97808 & 9,7808 \times 10^{-2} & 0 & 0 \\ 2,1918 \times 10^{-2} & 0,88247 & 0,19562 & 0 \\ 0 & 1,9726 \times 10^{-2} & 0,78685 & 0 \\ 0 & 0 & 1,7534 \times 10^{-2} & 1 \end{bmatrix},$$

$$\Pi = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Le nombre d'itérations nécessaire pour couvrir la période d'observation est $T=365$.

Résultats

$$M^T \Pi = \begin{bmatrix} 0,72537 \\ 0,16041 \\ 1,4866 \times 10^{-2} \\ 9,9070 \times 10^{-2} \end{bmatrix}.$$

La probabilité recherchée est **0,09907**.

Avec pas de discrétisation d'une heure

Nous cherchons maintenant à affiner le résultat en prenant un pas de discrétisation plus fin. Si nous raisonnons en heures, il vient :

$$\begin{cases} T = 8760 \\ (w, k) = (240, 3) \\ \text{fréquence } (\lambda \text{ ou } p) = \frac{8}{8760} \end{cases}, \text{ toutes les données devant être exprimées en heures.}$$

Notons que la taille de la matrice de transition n'évolue pas puisqu'un état est caractérisé par le nombre d'évènements dans la fenêtre courante (0, 1, 2 ou plus de 3). Par contre le nombre d'itérations nécessaires pour parcourir la période d'observation augmente considérablement : $T=8760$.

Nous avons :

$$M = \begin{bmatrix} 0,99909 & 4,1629 \times 10^{-3} & 0 & 0 \\ 9,1324 \times 10^{-4} & 0,99493 & 8,3257 \times 10^{-3} & 0 \\ 0 & 9,0944 \times 10^{-4} & 0,99077 & 0 \\ 0 & 0 & 9,0563 \times 10^{-4} & 1 \end{bmatrix},$$

$$\Pi = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Résultats

$$M^T \Pi = \begin{bmatrix} 0,72539 \\ 0,15639 \\ 1,5430 \times 10^{-2} \\ 0,1274 \end{bmatrix}.$$

La probabilité recherchée est **0,1274**.

Remarques :

- La probabilité est plus élevée avec un pas de discrétisation d'une heure qu'avec un pas de discrétisation d'un jour. En effet, dans le premier cas nous admettons (0 ou) 1 évènement par heure, tandis que dans le second cas nous n'admettons que (0 ou) 1 évènement par jour.
- Notons que ces résultats sont cohérents avec ceux obtenus par simulation de Monte-Carlo (Tableau 3).
- Enfin, remarquons que le pas de discrétisation d'une heure nous rapproche du résultat obtenu par la simulation de Monte-Carlo dans le cadre du modèle « exponentiel renouvelé ». C'est d'ailleurs conforme à l'intuition : si l'on fait tendre le pas de discrétisation vers zéro dans le modèle de Bernoulli, on doit converger vers le modèle « exponentiel renouvelé » ; si la propriété de convergence de la loi binomiale vers la loi de Poisson est connue, il n'était pas évident qu'elle soit vérifiée dans le particulier où la fenêtre d'observation est balayée.

II.1.c. Modèle de Poisson discret

La période d'observation $[0, T[$ étant discrétisée, nous supposons que le nombre d'évènements susceptibles de se produire sur la $(u+1)^{\text{ième}}$ unité de temps $[u, u+1[$ est décrit par la variable aléatoire $Y_u = N_1(u)$ qui suit une loi de Poisson de paramètre λ :

✓ Probabilité qu'il y ait n évènements sur la $u^{\text{ième}}$ unité de temps : $P\{Y_u = n\} = \frac{\lambda^n e^{-\lambda}}{n!}$.

Les variables aléatoires Y_u , pour $0 \leq u \leq w-1$, sont les accroissements d'un processus de Poisson et sont de ce fait indépendants. La variable aléatoire $N_w(u)$ compte le nombre d'évènements sur la fenêtre de longueur w démarrant en u :

✓ Probabilité de n évènements sur la fenêtre $[u, u+w[$: $P\{N_w(u) = n\} = \frac{(\lambda w)^n e^{-\lambda w}}{n!}$.

Regardons ce que deviennent, pour ce modèle, les états, la condition initiale et la matrice de transition :

- *Etats*

Cette fois le nombre d'évènements par unité de temps n'est pas limité ; le nombre d'évènements à l'intérieur de la fenêtre de balayage peut donc prendre n'importe quelle valeur entière.

Nous avons ainsi k états transients E_0, E_1, \dots, E_{k-1} et l'état absorbant $A=E_k$.

- *Condition initiale*

Le vecteur conditions initiales est associé à une fenêtre fictive de longueur w située à gauche de la période d'observation et ne contenant aucun évènement, c'est-à-dire que la chaîne se trouve initialement dans l'état E_0 avec la probabilité 1. Nous avons donc :

$$\Pi = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \\ 0 \end{bmatrix}.$$

- *Matrice de transition*

Les transitions réalisables sont :

- La transition d'un état E_n (« n évènements dans la fenêtre courante ») vers l'état E_{n+s} pour $1 \leq s < k-n$.

Cette transition se produit avec la probabilité :

$$m_{n+s,n} = \sum_{\ell=0}^n \binom{n}{\ell} \left(1 - \frac{1}{w}\right)^{n-\ell} \left(\frac{1}{w}\right)^\ell \frac{\lambda^{\ell+s}}{(\ell+s)!} e^{-\lambda}.$$

- La boucle sur un état E_n pour $0 \leq n < k$ qui se produit avec la probabilité :

$$m_{n,n} = \sum_{\ell=0}^n \binom{n}{\ell} \left(1 - \frac{1}{w}\right)^{n-\ell} \left(\frac{1}{w}\right)^\ell \frac{\lambda^\ell}{\ell!} e^{-\lambda}.$$

Pour $n=k$, nous avons $p_{k,k} = 1$.

La transition d'un état E_n vers l'état E_{n-s} pour $1 \leq s \leq n$, qui se produit avec la probabilité :

$$m_{n,n} = \sum_{\ell=s}^n \binom{n}{\ell} \left(1 - \frac{1}{w}\right)^{n-\ell} \left(\frac{1}{w}\right)^\ell \frac{\lambda^{\ell-s}}{(\ell-s)!} e^{-\lambda}.$$

- La transition d'un état E_n vers l'état absorbant E_k avec la probabilité :

$$m_{k,n} = 1 - \sum_{s=1}^n m_{n-s,n} - \sum_{s=1}^{k-n-1} m_{n+s,n} - m_{n,n}.$$

Preuve : calcul des probabilités de transition

Étape 1. Notons tout d'abord :

- ✓ $p(i, \lambda, t)$ la probabilité qu'il y ait i évènements sur la fenêtre de longueur t démarrante en u , lorsque la loi évènementielle est une loi de Poisson de paramètre λt :

$$p(i, \lambda, t) = P(N_t(u) = i) = \frac{(\lambda t)^i e^{-\lambda t}}{i!}.$$

- ✓ $P(i, \lambda, t)$ la probabilité qu'il y ait au plus i évènements sur la fenêtre de longueur t démarrante en u , lorsque la loi évènementielle est une loi de Poisson de paramètre λt . C'est donc la fonction de répartition d'une loi de Poisson de paramètre λt :

$$P(i, \lambda, t) = P(N_t(u) \leq i) = \sum_{j=0}^i \frac{(\lambda t)^j e^{-\lambda t}}{j!}.$$

- ✓ $\pi(\ell, n)$ la probabilité conditionnelle que le premier sous-intervalle de temps de la fenêtre courante $[u, u+w[$ comporte ℓ évènements sachant que ladite fenêtre en comporte exactement n :

$$\pi(\ell, n) = P(Y_u = \ell | N_w(u) = n) = \binom{n}{\ell} \left(1 - \frac{1}{w}\right)^{n-\ell} \left(\frac{1}{w}\right)^\ell.$$

En effet,

$$\begin{aligned} \pi(\ell, n) &= P(Y_u = \ell | Y_u + \dots + Y_{u+w-1} = n) \\ &= \frac{P(Y_u = \ell, Y_{u+1} + \dots + Y_{u+w-1} = n - \ell)}{P(Y_u + \dots + Y_{u+w-1} = n)}, \end{aligned}$$

d'où :

$$\begin{aligned} \pi(\ell, n) &= \frac{\frac{\lambda^\ell e^{-\lambda}}{\ell!} (\lambda(w-1))^{n-\ell} e^{-\lambda(w-1)}}{\frac{(\lambda w)^n e^{-\lambda w}}{n!}} \\ &= \binom{n}{\ell} \frac{(w-1)^{n-\ell}}{w^n} \\ &= \binom{n}{\ell} \left(1 - \frac{1}{w}\right)^{n-\ell} \left(\frac{1}{w}\right)^\ell. \end{aligned}$$

Signalons que ce résultat est classique : si N_t est un processus de poisson, la loi de N_s sachant $N_t = n$ est la loi binomiale de paramètres n et $\frac{s}{t}$. Dans notre cas, $\frac{s}{t} = \frac{1}{w}$, et nous retrouvons bien le résultat ci-dessus.

Étape 2. Transition d'un état E_n vers l'état E_{n+s} , pour $1 \leq s < k - n$

Ce genre de transition se produit lorsque le nombre d'évènements que l'on « gagne » est supérieur de s au nombre d'évènements que l'on « perd » :

$$Y_{u+w} - Y_u = s$$

de sorte que

$$N_w(u+1) - N_w(u) = s.$$

Pour ℓ allant de 0 à n , nous écrivons la probabilité de « perdre » ℓ évènements sur la première unité de temps de la fenêtre $C_w(u)$, conditionnellement au fait qu'elle en comporte n , et d'en « gagner » $\ell + s$ sur la première unité de temps de la fenêtre $C_w(u+1)$:

$$P(Y_u = \ell | N_w(u) = n)P(Y_{u+w} = \ell + s) = \pi(\ell, n)p(\ell + s, \lambda, 1).$$

La probabilité de passer d'une fenêtre $C_w(u)$ comportant n évènements à une fenêtre $C_w(u+1)$ en comportant $n+s$ est alors :

$$m_{n+s, n} = \sum_{\ell=0}^n \pi(\ell, n)p(\ell + s, \lambda, 1).$$

Étape 3. Boucle sur un état E_n .

Il y a boucle sur l'état E_n lorsque le nombre d'évènements que l'on « gagne » est égal au nombre d'évènements que l'on « perd » :

$$Y_{u+w} - Y_u = 0$$

de sorte que

$$N_w(u+1) = N_w(u).$$

Pour ℓ allant de 0 à n , nous écrivons la probabilité de « perdre » ℓ évènements sur la première unité de temps de la fenêtre $C_w(u)$, conditionnellement au fait qu'elle en comporte n , et d'en « gagner » ℓ sur la première unité de temps de la fenêtre $C_w(u+1)$:

$$P(Y_u = \ell | N_w(u) = n)P(Y_{u+w} = \ell) = \pi(\ell, n)p(\ell, \lambda, 1).$$

Pour $n < k$, la probabilité de passer d'une fenêtre $C_w(u)$ comportant n évènements à une fenêtre $C_w(u+1)$ en comportant encore n est alors :

$$m_{n, n} = \sum_{\ell=0}^n (\pi(\ell, n)p(\ell, \lambda, 1)).$$

Étape 4. Transition d'un état E_n vers l'état E_{n-s} pour $1 \leq s \leq n$.

Cette transition se produit lorsque le nombre d'évènements que l'on « gagne » est inférieur de s au nombre d'évènements que l'on « perd » :

$$Y_u - Y_{u+w} = s$$

de sorte que

$$N_w(u+1) - N_w(u) = -s.$$

Pour ℓ allant de s à n , nous écrivons la probabilité de « perdre » ℓ évènements sur la première unité de temps de la fenêtre $C_w(u)$, conditionnellement au fait qu'elle en comporte n , et de n'en « gagner que » $\ell - s$ sur la première unité de temps de la fenêtre $C_w(u+1)$:

$$P(Y_u = \ell | N_w(u) = n)P(Y_{u+w} = \ell - s) = \pi(\ell, n)p(\ell - s, \lambda, 1).$$

La probabilité de passer d'une fenêtre $C_w(u)$ comportant n évènements à une fenêtre $C_w(u+1)$ en comportant $n-s$ est alors :

$$m_{n-s, n} = \sum_{\ell=0}^n (\pi(\ell, n)p(\ell - s, \lambda, 1)) ;$$

Étape 5. Transition d'un état E_n vers l'état absorbant E_k .

Cette transition se produit avec la probabilité :

$$m_{k,n} = P(N_w(u+1) \geq k | N_w(u) = n) = 1 - P(N_w(u+1) < k | N_w(u) = n) = 1 - \sum_{s=1}^n m_{n-s,n} - \sum_{s=1}^{k-n-1} m_{n+s,n} - m_{n,n}.$$

CQFD

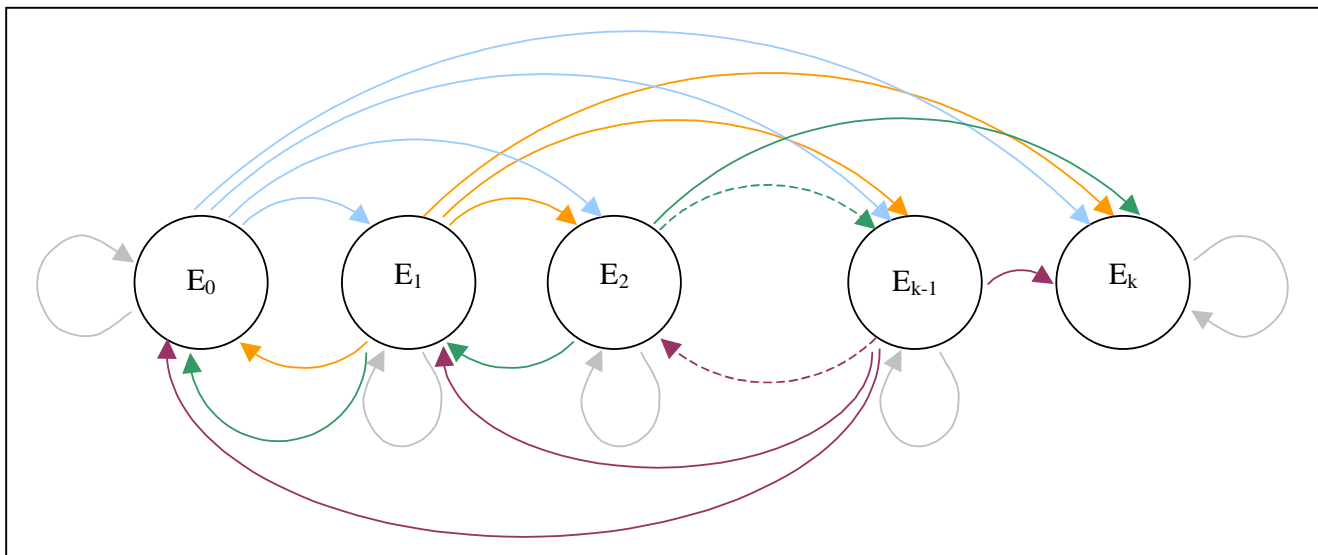
La matrice de transition est donc une matrice carrée de taille $k+1$:

- les k premiers termes diagonaux sont les $m_{n,n}$ et le dernier terme diagonal est $m_{k,k}=1$;
- les termes situés au-dessous de la diagonale sont les $m_{n+s,n}$;
- les termes au-dessus de la diagonale sont les $m_{n-s,n}$, sauf le terme situé en dernière ligne qui est le complément à 1 des autres termes .

Nous avons donc la matrice M suivante :

$$M = \begin{bmatrix} m_{0,0} & m_{0,1} & m_{0,2} & \dots & m_{0,k-1} & 0 \\ m_{1,0} & m_{1,1} & m_{1,2} & \dots & m_{1,k-1} & 0 \\ m_{2,0} & m_{2,1} & m_{2,2} & \dots & m_{2,k-1} & 0 \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\ m_{k-1,0} & m_{k-1,1} & m_{k-1,2} & \dots & m_{k-1,k-1} & 0 \\ 1 - \sum_{\ell=0}^{k-1} m_{\ell,0} & 1 - \sum_{\ell=0}^{k-1} m_{\ell,1} & 1 - \sum_{\ell=0}^{k-1} m_{\ell,2} & \dots & 1 - \sum_{\ell=0}^{k-1} m_{\ell,k-1} & 1 \end{bmatrix}$$

Figure 25 : Graphe de la chaîne de Markov pour le modèle de Poisson discret, premier modèle markovien



II.1.d. Modèle de Poisson discret : application à l'exemple (T=365,w=10,k=3)

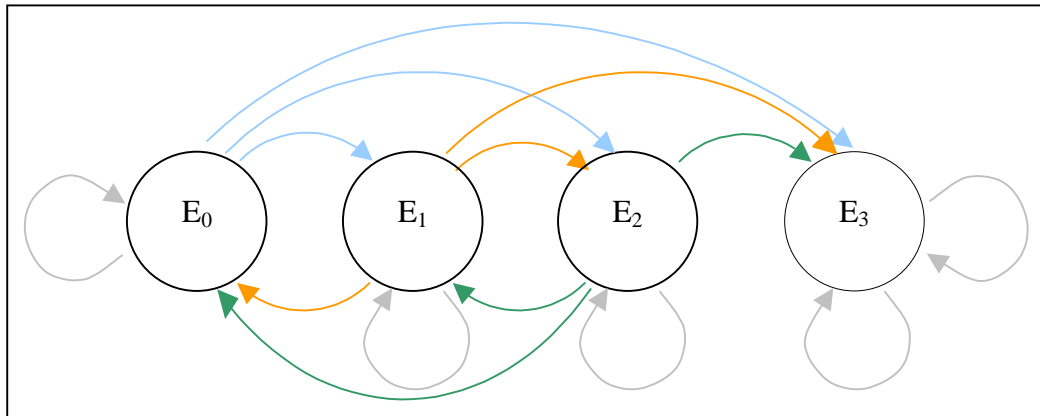
Nous allons appliquer la méthode à l'exemple cité plus haut, avec successivement un pas de discrétisation d'un jour, puis d'une heure.

La matrice M à considérer est une matrice (4×4) qui a l'allure suivante :

$$M = \begin{bmatrix} m_{0,0} & m_{0,1} & m_{0,2} & 0 \\ m_{1,0} & m_{1,1} & m_{1,2} & 0 \\ m_{2,0} & m_{2,1} & m_{2,2} & 0 \\ 1 - m_{0,0} - m_{1,0} - m_{2,0} & 1 - m_{0,1} - m_{1,1} - m_{2,1} & 1 - m_{0,2} - m_{1,2} - m_{2,2} & 1 \end{bmatrix}$$

Sa chaîne de Markov est la chaîne à 4 états dont le graphe est reproduit sur la figure ci-dessous :

Figure 26 : Chaîne de Markov pour le modèle de Poisson discret, premier modèle markovien, T=365, w=10, k=3



Le vecteur conditions initiales est le vecteur

$$\Pi = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

La probabilité recherchée est donnée par le dernier terme du vecteur $M^T \Pi$, avec T=365.

Application numérique

Avec pas de discrétisation d'un jour

Dans l'exemple proposé nous avons :

$$\left\{ \begin{array}{l} T = 365 \\ (w, k) = (10, 3) \\ \text{fréquence } (\lambda \text{ ou } p) = \frac{8}{365} \end{array} \right. .$$

Par conséquent, nous écrivons la matrice M et le vecteur Π :

$$M = \begin{bmatrix} 0,97832 & 9,7832 \times 10^{-2} & 9,7832 \times 10^{-3} & 0 \\ 2,1443 \times 10^{-2} & 0,88263 & 0,17631 & 0 \\ 2,3499 \times 10^{-4} & 1,9322 \times 10^{-2} & 0,7963 & 0 \\ 1,7263 \times 10^{-6} & 2,1322 \times 10^{-4} & 1,7603 \times 10^{-2} & 1 \end{bmatrix},$$

$$\Pi = \begin{bmatrix} 0,80318 \\ 0,17604 \\ 1,9292 \times 10^{-2} \\ 1,4902 \times 10^{-3} \end{bmatrix}.$$

Résultats

$$M^T \Pi = \begin{bmatrix} 0,71277 \\ 0,15388 \\ 1,5445 \times 10^{-2} \\ 0,11765 \end{bmatrix}.$$

La probabilité recherchée est **0,11765**.

Avec pas de discrétisation d'une heure

Le dernier résultat a été obtenu en discrétisant la fenêtre d'observation de 365 jours selon un pas de discrétisation d'un jour. Considérons maintenant un pas de discrétisation d'une heure, ce qui correspond à modifier l'unité des données d'entrée de sorte que :

$$\begin{cases} T = 8760 \\ (w, k) = (240, 3) \\ \text{fréquence } (\lambda \text{ ou } p) = \frac{8}{8760} \end{cases}.$$

Le nombre d'itérations nécessaires pour parcourir la période d'observation est alors $T = 8760$.

Nous avons :

$$M = \begin{bmatrix} 0,99909 & 4,1629 \times 10^{-3} & 1,7345 \times 10^{-5} & 0 \\ 9,1241 \times 10^{-4} & 0,99493 & 8,2911 \times 10^{-3} & 0 \\ 4,1662 \times 10^{-7} & 9,0861 \times 10^{-4} & 0,99079 & 0 \\ 1,2686 \times 10^{-10} & 4,1502 \times 10^{-7} & 9,0524 \times 10^{-4} & 1 \end{bmatrix},$$

$$\Pi = \begin{bmatrix} 0,80318 \\ 0,17604 \\ 1,9292 \times 10^{-2} \\ 1,4902 \times 10^{-3} \end{bmatrix}.$$

Résultats

$$M^T \Pi = \begin{bmatrix} 0,72255 \\ 0,15563 \\ 1,5408 \times 10^{-2} \\ 0,12803 \end{bmatrix}.$$

La probabilité recherchée est **0,12803**.

Remarque : Les résultats sont très proches de ceux obtenus par la simulation de Monte-Carlo discrète (Tableau 3). Là encore, la probabilité est plus élevée avec un pas de discrétisation d'une heure qu'avec un pas de discrétisation d'un jour, et le pas de discrétisation d'une heure nous rapproche du résultat obtenu par la simulation de Monte-Carlo continue.

II.2. Modèle markovien à double fenêtre de balayage

L'approche markovienne présentée au paragraphe précédent ne fournit pas une représentation très précise du phénomène observé. En effet, l'état du système à un instant donné renseigne sur le nombre d'évènements dans la fenêtre courante mais en aucun cas sur la position de ces évènements dans ladite fenêtre. Ainsi, nous observons certaines transitions qui ne cadrent pas avec le phénomène que nous observons, comme illustré dans l'exemple ci-dessous.

Exemple : Supposons qu'après la $(i-1)^{\text{ième}}$ transition, la fenêtre courante ne contienne aucun évènement : la chaîne de Markov est sur l'état E_0 . Après tirage de la $i^{\text{ième}}$ transition, disons qu'un évènement rentre dans la fenêtre : la chaîne de Markov passe alors sur l'état E_1 .

En pratique, cet évènement ne peut quitter la fenêtre courante avant w nouvelles transitions : dans l'exemple ($T=365, w=10, k=3$), un évènement donné quitte la fenêtre courante après exactement $w=10$ transitions, comme illustré sur la figure 27.

Par conséquent la chaîne de Markov peut, soit rester sur l'état E_1 pendant au moins $(w-1)$ transitions dans le cas où aucun nouvel évènement ne rentre dans la fenêtre courante, soit passer sur un état E_i , $i \geq 2$, dans le cas où un nouvel évènement intègre la fenêtre courante.

C'est là que nous constatons le biais de notre modèle : il autorise la chaîne de Markov à retourner sur l'état E_0 , correspondant au cas où la fenêtre ne contient aucun évènement, et ce dès la $(i+1)^{\text{ième}}$ transition, car le passage de l'état E_1 vers l'état E_0 est valide à tout instant (voir figure 28).

Figure 27 : temps de séjour dans la fenêtre de balayage

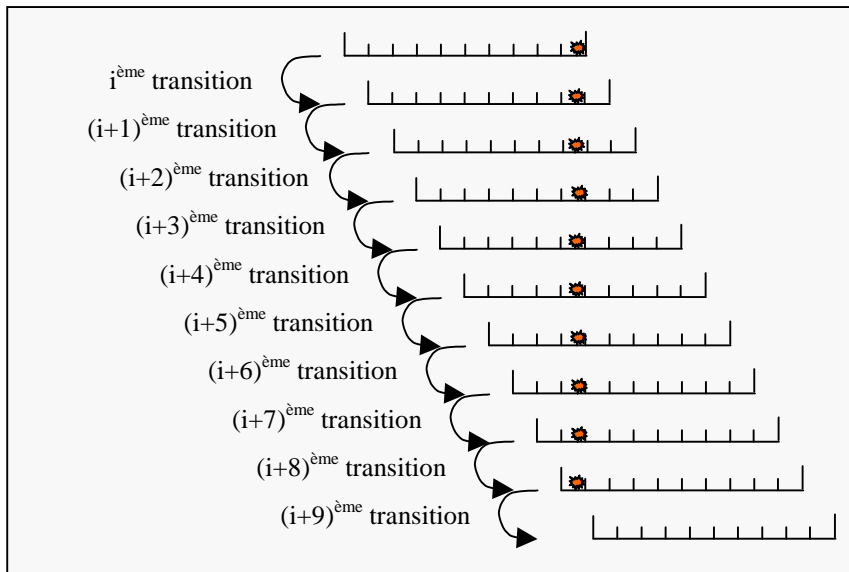
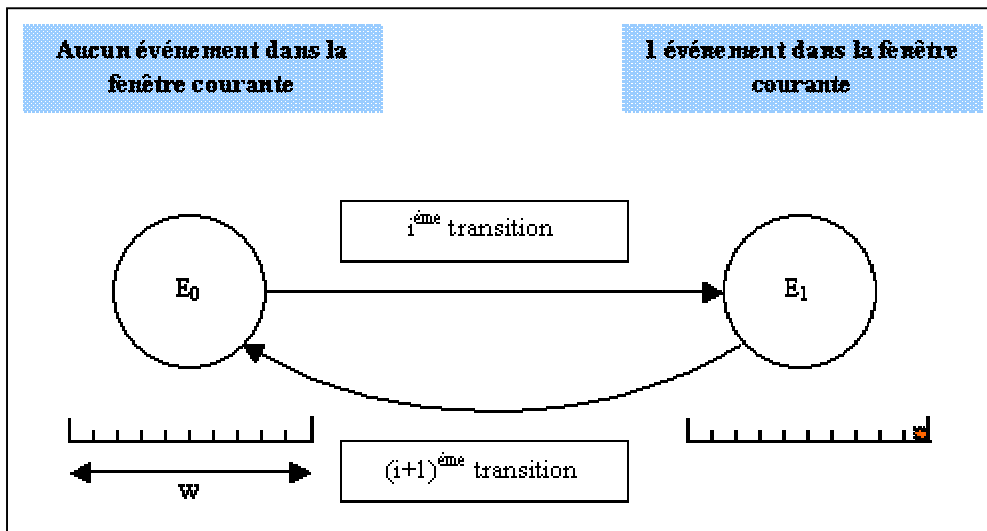


Figure 28 : des transitions irréalisables admises par la première approche



Nous proposons donc une approche plus fine que la précédente, c'est-à-dire contenant « un peu plus » d'information, dans le sens suivant : lorsque la fenêtre courante contient des événements, nous ne connaissons toujours pas leur position exacte mais nous savons dire s'ils se situent dans la première ou dans la seconde moitié de ladite fenêtre.

L'approche proposée est développée ci-dessous pour le modèle de Bernoulli puis appliquée à l'exemple ($T=365, w=10, k=3$). Nous ne la développerons pas pour le modèle de Poisson discret, les calculs étant plus fastidieux.

Dans un premier temps, rappelons les notations introduites précédemment et que nous utiliserons dans cette section :

- ✓ La période d'observation est l'intervalle ouvert $[0, T[$.
- ✓ La fenêtre de balayage est de taille w , $0 < w \leq T$.
- ✓ Le nombre d'évènements qui nous intéresse est $k \in \mathbb{N}^*$.
- ✓ Pour $u \in [0, T - t]$, l'objet « fenêtre » de longueur t démarrant en u est noté $C_t(u)$.
- ✓ Pour $u \in [0, T - t]$, le nombre d'évènements dans la fenêtre $C_t(u)$ est $N_t(u)$.
- ✓ La variable aléatoire comptant le nombre d'évènements à l'instant u est notée Y_u , et n'est rien d'autre que $N_1(u)$.

Regardons maintenant les notations dont nous aurons spécifiquement besoin dans cette section :

- ✓ Nous supposons que w est pair³ et nous noterons $h = \frac{w}{2}$.
- ✓ Nous considérons une « double fenêtre de balayage », associée à un couple de variables aléatoires $(N_h(u), N_h(u+h))$ prenant des valeurs entières et vérifiant $N_w(u) = N_h(u) + N_h(u+h)$.

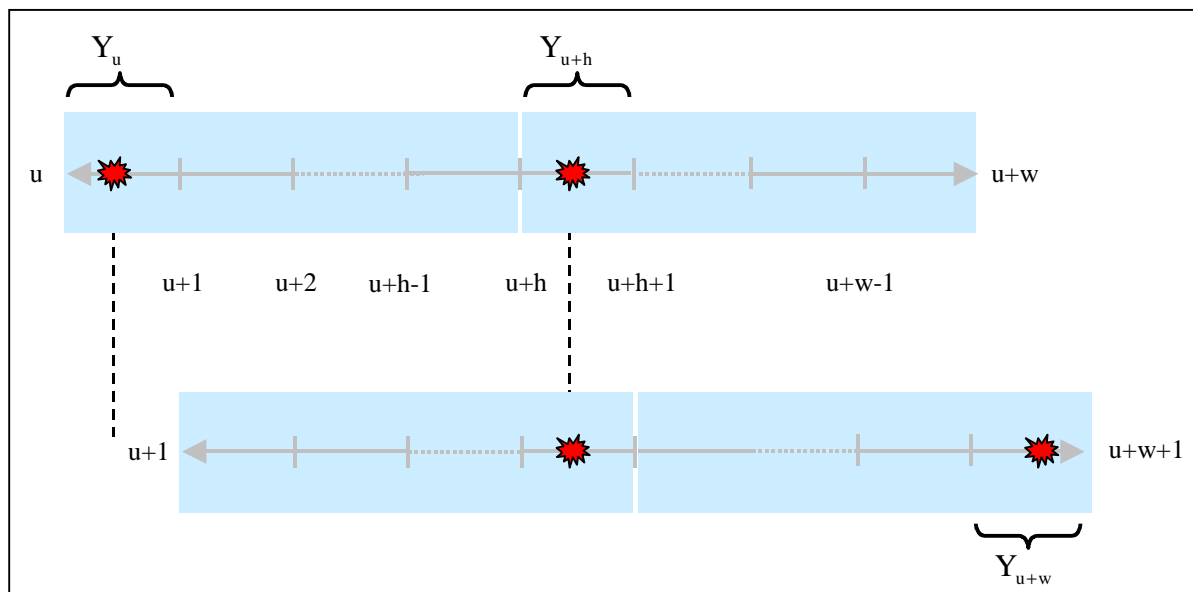
La dynamique d'une transition, que nous avons écrit jusque là sous la forme générale $N_w(u+1) = N_w(u) - Y_u + Y_{u+w}$, peut maintenant s'exprimer de la façon suivante :

$$(N_h(u+1), N_h(u+h+1)) = (N_h(u) - Y_u + Y_{u+h}, N_h(u+h) - Y_{u+h} + Y_{u+w}),$$

ce qui illustre bien le fait que la première demi-fenêtre « perd » Y_u et « gagne » Y_{u+h} tandis que la seconde « perd » Y_{u+h} et « gagne » Y_{u+w} .

Cette dynamique est illustrée sur la figure 29.

Figure 29 : transition dans le cadre du modèle à double fenêtre



³ Si le nombre w de sous-intervalles constituant la fenêtre de balayage est impair, la première moitié de la fenêtre sera de longueur $\left\lfloor \frac{w}{2} \right\rfloor$ et la seconde moitié sera de longueur $w - \left\lfloor \frac{w}{2} \right\rfloor = \left\lceil \frac{w}{2} \right\rceil$.

La chaîne de Markov que nous construisons est alors basée sur les éléments suivant :

- *Etats*

Les états transitoires, notés E_0, E_1, \dots, E_{r-1} , renseignent sur le nombre d'évènements dans chacune des deux demi-fenêtres constituant la fenêtre courante. Un état est donc un couple d'entiers (i,j) : i est le nombre d'évènements dans la première moitié de la fenêtre, j le nombre d'évènements dans la seconde moitié.

L'état absorbant, noté A ou E_r , correspond au cas où la fenêtre courante contient plus de k évènements, quelle que soit leur répartition sur les deux demi-fenêtres.

- *Condition initiale*

Nous conservons la logique précédente : le vecteur des probabilités initiales est associé à une fenêtre fictive de longueur w , située à gauche de la période d'observation, et ne contenant aucun événement. En notant E_0 l'état $(0,0)$, le vecteur des probabilités initiales est donc le vecteur de taille $r+1$:

$$\Pi = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}.$$

- *Matrice de transition*

C'est la matrice M dont l'élément m_{ij} est la probabilité de transition de l'état E_j vers l'état E_i .

La probabilité recherchée sera donnée par le produit $M^T \Pi$.

II.2.a. Déclinaison au modèle de Bernoulli

- *Etats*

Les états transitoires E_0, E_1, \dots, E_{r-1} sont des couples (i,j) tels que :

- ~ $i + j \leq k - 1$, puisque k est le nombre d'évènements au-delà duquel nous avons un cluster.
- ~ $i \leq h$ et $j \leq h$, puisque le nombre d'évènements par unité de temps est 0 ou 1, ce qui limite le nombre d'évènements sur chaque demi-fenêtre à sa taille h .

L'état absorbant est noté $A = E_r$ et correspond au cas où la fenêtre courante comprend plus de k accidents.

Le nombre d'états $r + 1$ s'exprime en fonction des quantités w et k :

$$r + 1 = \sum_{S=0}^k (S+1) + \sum_{S=\kappa+1}^{k-1} (w - S + 1) + 1 \text{ où } \kappa = \min(h, k-1) = \min\left(\frac{w}{2}, k-1\right).$$

Preuve :

Soit $S \leq k - 1$. Regardons le nombre de couples (i,j) vérifiant « $i + j = S$ » (condition 1) et « $i \leq h$ et $j \leq h$ » (condition 2).

- Si $S \leq h$, la condition 2 est automatiquement vérifiée et le nombre de couples vérifiant la condition 1 est égal à $S+1$: ce sont les couples $(0,S), (1,S-1), \dots, (S,0)$.

- Si $S > h$, alors les couples (i,j) tels que $i > h$, au nombre de $S-h$, ou ceux tels que $j > h$, au nombre de $S-h$, ne vérifient pas la condition 2 (le cas $i > h$ et $j > h$ est exclu car $i + j \leq k-1 \leq w = 2h$). Si bien que le nombre de couples vérifiant simultanément les conditions 1 et 2 est $(S+1) - 2(S-h) = 2h - S + 1 = w - S + 1$.

Au final, en sommant sur S , pour S allant de 0 à $k-1$, et en distinguant le cas $S \leq h$ et le cas $S > h$, il vient :

$$r = \sum_{S=0}^{\kappa} (S+1) + \sum_{S=\kappa+1}^{k-1} (w - S + 1) \text{ où } \kappa = \min(h, k-1) = \min\left(\frac{w}{2}, k-1\right).$$

Enfin, en ajoutant 1 pour l'état absorbant, il vient $r+1$ pour le nombre d'états.

CQFD

Les états seront ordonnés selon l'ordre lexicographique, c'est-à-dire :

$$E_0 = (0,0), E_1 = (0,1), \dots, E_{\kappa} = (0, \kappa), E_{\kappa+1} = (1,0), \text{ etc...}$$

- *Condition initiale*

Le vecteur condition initiale est le vecteur de taille $r+1$:

$$\Pi = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}.$$

- *Matrice de transition*

- Le Tableau 6 recense toutes les transitions réalisables à partir d'un état (i,j) qui n'est pas l'état absorbant, vers un état (i,j) qui n'est pas l'état absorbant, en fonction de la valeur (0 ou 1) prise par Y_u , Y_{u+h} et Y_{u+w} . En d'autres termes, il s'agit des transitions entre états transitoires. Les probabilités qui leur sont associées sont renseignées dans le tableau, et la preuve détaillée du calcul est donnée pour deux d'entre elles.

- La probabilité de transition d'un état E_i vers l'état absorbant est le complément à 1 de la somme des probabilités de transitions de l'état E_i vers les autres états :

$$m_{ir} = 1 - \sum_{j=0}^{r-1} m_{ji}.$$

- La probabilité de transition de l'état absorbant vers un état E_i qui n'est pas l'état absorbant est nulle :

$$m_{ir} = 0 \text{ pour tout } i \neq r.$$

Preuve des deux premières probabilités

Transition de l'état (i,j) vers l'état (i,j)

La chaîne réalise une boucle sur l'état (i,j) dans deux cas de figure : d'une part, lorsque $(Y_u = 0, Y_{u+h} = 0, Y_{u+w} = 0)$ si bien que le nombre d'évènements dans chacune des deux demi-fenêtres est inchangé ; d'autre part, lorsque $(Y_u = 1, Y_{u+h} = 1, Y_{u+w} = 1)$ si bien que chacune des deux demi-fenêtres « perd » un évènement et en « gagne » un autre, en conservant ainsi le même nombre.

Cela se traduit par la probabilité suivante :

$$P(N_w(u) = (i, j) | N_w(u) = (i, j)) = P(Y_u = 1 | N_h(u) = i) P(Y_{u+h} = 1 | N_h(u+h) = j) P(Y_{u+w} = 1) \\ + P(Y_u = 0 | N_h(u) = i) P(Y_{u+h} = 0 | N_h(u+h) = j) P(Y_{u+w} = 0)$$

Or, pour tout $u \in [0, T - w[$, nous avons :

- ✓ $P(Y_{u+w} = 1) = p$ et $P(Y_{u+w} = 0) = q = 1 - p$.
- ✓ $P(Y_u = 1 | N_h(u) = i) = \frac{i}{h} = \frac{2i}{w}$ et $P(Y_u = 0 | N_h(u) = i) = 1 - \frac{i}{h} = 1 - \frac{2i}{w}$.
- ✓ $P(Y_{u+h} = 1 | N_h(u+h) = j) = \frac{j}{h} = \frac{2j}{w}$ et $P(Y_{u+h} = 0 | N_h(u+h) = j) = 1 - \frac{j}{h} = 1 - \frac{2j}{w}$.

Nous obtenons bien le résultat escompté.

Transition de l'état (i,j) vers l'état (i,j-1)

La chaîne réalise une transition de l'état (i,j) vers l'état (i,j-1) dans un seul cas de figure : lorsque $(Y_u = 1, Y_{u+h} = 1, Y_{u+w} = 0)$, si bien que la première demi-fenêtre « perd » un évènement et en « gagne » un autre, en conservant ainsi le même nombre, tandis que la deuxième demi fenêtre « perd » un évènement et n'en gagne aucun, passant ainsi de j à j-1 évènements.

Cela se traduit par la probabilité suivante :

$$P(N_w(u) = (i, j-1) | N_w(u) = (i, j)) = P(Y_u = 1 | N_h(u) = i) P(Y_{u+h} = 1 | N_{u+h}(h) = j) P(Y_{u+w} = 0).$$

En remplaçant par les probabilités adéquates, mentionnées ci-dessus, nous obtenons le résultat escompté.

CQFD

Tableau 6 : Double fenêtre de balayage, transitions réalisables depuis un état transitoire (i,j)

Y_u	Y_{u+h}	Y_{u+w}	Etat initial	Etat final	Probabilité de transition
1	1	1	(i, j)	(i, j)	$q \times \left(1 - \frac{2i}{w}\right) \times \left(1 - \frac{2j}{w}\right) + p \times \left(\frac{2i}{w}\right) \times \left(\frac{2j}{w}\right)$
0	0	0			
1	1	0		(i, j-1) si $j \geq 1$	$q \times \left(\frac{2i}{w}\right) \times \left(\frac{2j}{w}\right)$
1	0	1		(i-1, j+1) si $i \geq 1, j \leq k-2$	$p \times \left(\frac{2i}{w}\right) \times \left(1 - \frac{2j}{w}\right)$
1	0	0		(i-1, j) si $i \geq 1$	$q \times \left(\frac{2i}{w}\right) \times \left(1 - \frac{2j}{w}\right)$
0	1	1		(i+1, j) si $i \leq k-2$	$p \times \left(1 - \frac{2i}{w}\right) \times \left(\frac{2j}{w}\right)$
0	1	0		(i+1, j-1) si $i \leq k-2, j \geq 1$	$q \times \left(1 - \frac{2i}{w}\right) \times \left(\frac{2j}{w}\right)$
0	0	1		(i, j+1) si $j \leq k-2$	$p \times \left(1 - \frac{2i}{w}\right) \times \left(1 - \frac{2j}{w}\right)$

II.2.b. *Modèle de Bernoulli : application à l'exemple ($T=365, w=10, k=3$)*

Nous appliquons le modèle proposé à l'exemple que nous rappelons : nous cherchons la probabilité d'observer, sur une année donnée de 365 jours, un agrégat de 3 accidents en moins de 10 jours, sachant que la fréquence du type d'accident considéré est de 8 par année.

Les états sont au nombre de 7 :

- l'état $E_0=(0,0)$: la fenêtre courante ne comporte aucun accident ;
- l'état $E_1=(0,1)$: la fenêtre courante comporte un accident et cet accident est situé dans la seconde moitié de la fenêtre ;
- l'état $E_2=(0,2)$: la fenêtre courante comporte deux accidents et les deux accidents sont situés dans la seconde moitié de la fenêtre ;
- l'état $E_3=(1,0)$: la fenêtre courante comporte un accident et cet accident est situé dans la première moitié de la fenêtre ;
- l'état $E_4=(1,1)$: la fenêtre courante comporte deux accidents, un sur chacune des deux demi-fenêtres ;
- l'état $E_5=(2,0)$: la fenêtre courante comporte deux accidents et les deux accidents sont situés dans la première moitié de la fenêtre ;
- l'état $A=E_6$ qui est absorbant, et regroupe tous les cas de figure correspondant au fait que la fenêtre courante contient trois accidents au moins.

Le vecteur des probabilités initiales est le vecteur de taille 7 :

$$\Pi = [1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]^t.$$

La matrice de transition est la matrice 7×7 remplie avec les probabilités de transitions telles que nous les avons explicitées au paragraphe précédent :

$$M = \begin{bmatrix} q & 0 & q \cdot \left(\frac{2}{w}\right) & 0 & 0 & 0 & 0 \\ p \cdot \left(1 - \frac{2}{w}\right) & q \cdot \left(\frac{2}{w}\right) & p \cdot \left(\frac{2}{w}\right) & 0 & q \cdot \left(\frac{2}{w}\right) \cdot \left(1 - \frac{2}{w}\right) & 0 & 0 \\ 0 & q \cdot \left(\frac{2}{w}\right) & q \cdot \left(1 - \frac{2}{w}\right) & 0 & q \cdot \left(\frac{2}{w}\right) \cdot \left(\frac{2}{w}\right) & q \cdot \left(\frac{4}{w}\right) & 0 \\ 0 & p \cdot \left(1 - \frac{2}{w}\right) & 0 & q \cdot \left(1 - \frac{4}{w}\right) & p \cdot \left(\frac{2}{w}\right) \cdot \left(1 - \frac{2}{w}\right) & 0 & 0 \\ 0 & p \cdot \left(\frac{2}{w}\right) & p \cdot \left(1 - \frac{2}{w}\right) & q \cdot \left(\frac{4}{w}\right) & p \cdot \left(\frac{2}{w}\right) \cdot \left(\frac{2}{w}\right) + q \cdot \left(1 - \frac{2}{w}\right) \cdot \left(1 - \frac{2}{w}\right) & p \cdot \left(\frac{4}{w}\right) & 0 \\ 0 & 0 & 0 & 0 & q \cdot \left(\frac{2}{w}\right) \cdot \left(1 - \frac{2}{w}\right) & q \cdot \left(1 - \frac{4}{w}\right) & 0 \\ 0 & 0 & 0 & p & p \cdot \left(1 - \frac{2}{w}\right) & p \cdot \left(1 - \frac{4}{w}\right) & 1 \end{bmatrix}$$

Application numérique

Avec pas de discrétisation d'un jour

Nous avons :

$$\begin{cases} T = 365 \\ (w, k) = (10, 3) \\ \text{fréquence } (\lambda \text{ ou } p) = \frac{8}{365} \end{cases}$$

Il vient :

$$M = \begin{bmatrix} 0.97808 & 0 & 0.19562 & 0 & 0 & 0 & 0 \\ 2.1918 \times 10^{-2} & 0.78247 & 4.3836 \times 10^{-3} & 0 & 0.15649 & 0 & 0 \\ 0 & 0.19562 & 0.78247 & 0 & 3.9123 \times 10^{-2} & 0.39123 & 0 \\ 0 & 1.7534 \times 10^{-2} & 0 & 0.58685 & 3.5068 \times 10^{-3} & 0 & 0 \\ 0 & 4.3836 \times 10^{-3} & 1.7534 \times 10^{-2} & 0.39123 & 0.62685 & 8.7671 \times 10^{-3} & 0 \\ 0 & 0 & 0 & 0 & 0.15649 & 0.58685 & 0 \\ 0 & 0 & 0 & 2.1918 \times 10^{-2} & 1.7534 \times 10^{-2} & 1.3151 \times 10^{-2} & 1 \end{bmatrix}$$

Résultats

Les probabilités de chaque état après $T = 365$ itérations sont données par le vecteur :

$$M^T \Pi = \begin{bmatrix} 0.72278 \\ 0.080625 \\ 0.079892 \\ 0.0034959 \\ 0.0084485 \\ 0.0032024 \\ 0.10144 \end{bmatrix}.$$

La probabilité que l'on ait un cluster de 3 accidents en 10 jours sur une année de 365 jours est donc de **0,10144**.

En comparaison à une approche binomiale classique, ce modèle a légèrement amélioré le résultat : nous sommes passés d'une probabilité de 9,91% à une probabilité de 10,14%, nous rapprochant ainsi du résultat obtenu par simulation.

Avec pas de discrétisation d'une heure

Le dernier résultat a été obtenu en discrétisant la fenêtre d'observation de 365 jours selon un pas de discrétisation d'un jour.

Considérons maintenant un pas de discrétisation d'une heure, ce qui correspond à modifier l'unité des données d'entrée de sorte que :

$$\begin{cases} T = 8760 \\ (w, k) = (240, 3) \\ \text{fréquence } (\lambda \text{ ou } p) = \frac{8}{8760} \end{cases}$$

Le nombre d'itérations nécessaires pour parcourir la période d'observation est $N = T = 8760$.

Nous avons :

$$M = \begin{bmatrix} 0.99909 & 0 & 8.3257 \times 10^{-3} & 0 & 0 & 0 & 0 \\ 9.1324 \times 10^{-4} & 0.99076 & 7.6104 \times 10^{-6} & 0 & 8.2563 \times 10^{-3} & 0 & 0 \\ 0 & 8.3257 \times 10^{-3} & 0.99076 & 0 & 6.9381 \times 10^{-5} & 1.6651 \times 10^{-2} & 0 \\ 0 & 9.0563 \times 10^{-4} & 0 & 0.98244 & 7.5469 \times 10^{-6} & 0 & 0 \\ 0 & 7.6104 \times 10^{-6} & 9.0563 \times 10^{-4} & 1.6651 \times 10^{-2} & 0.9825 & 1.5221 \times 10^{-5} & 0 \\ 0 & 0 & 0 & 0 & 8.2563 \times 10^{-3} & 0.98244 & 0 \\ 0 & 0 & 0 & 9.1324 \times 10^{-4} & 9.0563 \times 10^{-4} & 8.9802 \times 10^{-4} & 1 \end{bmatrix}$$

Résultats

Les probabilités de chaque état après $T = 8760$ itérations sont données par le vecteur

$$M^T \Pi = \begin{bmatrix} 0.71857 \\ 7.8251 \times 10^{-2} \\ 7.7374 \times 10^{-2} \\ 4.0422 \times 10^{-3} \\ 7.8936 \times 10^{-3} \\ 3.7142 \times 10^{-3} \\ 0.12964 \end{bmatrix}.$$

La probabilité que l'on ait un cluster de 3 accidents en 10 jours sur une année de 365 jours est donc de **0,12964**.

Conclusion

- Nous remarquons ici que discrétiser à l'heure apporte une nette amélioration, en ce sens que nous nous rapprochons du résultat obtenu par simulation : 0,1329 pour la simulation poissonnienne en temps continu.
- Le résultat obtenu avec le modèle à double-fenêtre est bon. Mais il n'est pas tellement meilleur que celui obtenu avec le modèle à simple fenêtre.

II.3. Modèle markovien complet

Nous considérons maintenant une chaîne de Markov discrète dont les états sont non seulement représentatifs du nombre d'événements dans la fenêtre de balayage mais aussi de leur position dans ladite fenêtre.

La fenêtre de longueur w démarrante en u , c'est-à-dire $[u, u+w[$, correspond à l'observation des variables aléatoires (Y_u, \dots, Y_{u+w-1}) . L'état correspondant à cette fenêtre est :

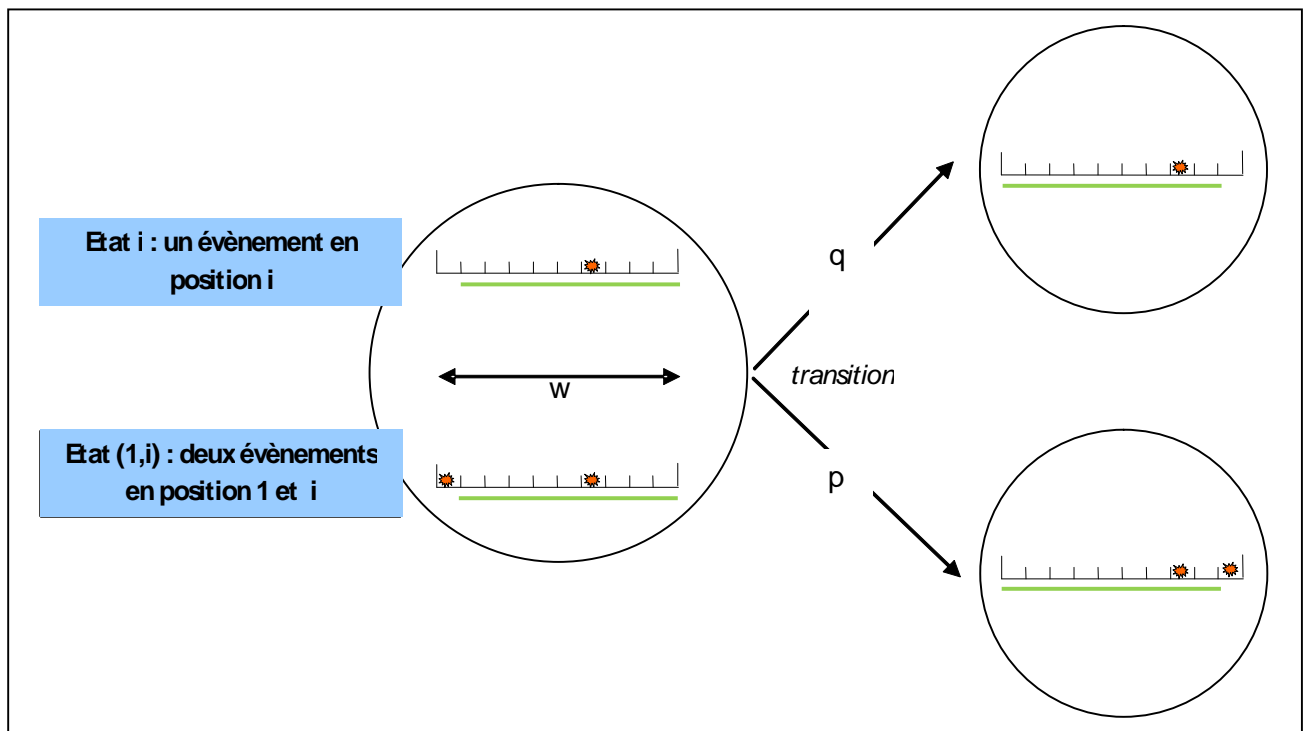
- Soit un état (y_1, y_2, \dots, y_w) codant une fenêtre qui ne comporte pas de cluster, c'est-à-dire telle que $y_1 + y_2 + \dots + y_w \leq k-1$.
- Soit l'état absorbant noté $A=E_r$, codant n'importe quelle fenêtre comportant un cluster, c'est-à-dire telle que $y_1 + y_2 + \dots + y_w \geq k$.

Remarquons dès à présent que l'observation des $(w-1)$ dernières composantes d'une fenêtre est suffisante puisque lors d'une transition, le premier élément du w -uplet, quel qu'il soit, est « perdu ». Prenons un exemple dans le cadre du modèle de Bernoulli : les états $(0, y_2, \dots, y_w)$ et $(1, y_2, \dots, y_w)$ sont équivalents pour tout $(y_2, \dots, y_w) \in \{0, 1\}^{w-1}$, en ce sens qu'ils emmènent tous deux le système :

- Soit vers l'état $(y_2, \dots, y_w, 0)$ si l'élément entrant est 0 ;
- Soit vers l'état $(y_2, \dots, y_w, 1)$ si l'élément entrant est 1 et si $y_2 + \dots + y_w + 1 \leq k-1$;
- Soit vers l'état absorbant A si l'élément entrant est 1 et si $y_2 + \dots + y_w + 1 = k$.

Ainsi, seuls les $(w-1)$ derniers éléments vont « caractériser » la transition, comme nous l'avons représenté sur la Figure 30.

Figure 30 : les états « i » et « $(1,i)$ » sont équivalents



II.3.a. Modèle de Bernoulli

Soit $p \in [0, 1]$ un réel. La famille (Y_0, \dots, Y_{w-2}) est une famille de variables aléatoires indépendantes et identiquement distribuées suivant une loi de Bernoulli de paramètre p :

$$P(Y_0 = 1) = p \text{ et } P(Y_0 = 0) = q = 1 - p.$$

La chaîne de Markov pour le modèle de Bernoulli est basée sur les éléments suivant :

- *Etats*

L'espace d'états est :

$$\mathcal{E}_{k,w} = \left\{ (y_0, \dots, y_{w-2}) \in \{0, 1\}^{w-1} / y_0 + \dots + y_{w-2} < k \right\} \cup A.$$

Le nombre d'états est :

$$r + 1 = 1 + \binom{w-1}{1} + \dots + \binom{w-1}{k-1} + 1.$$

Preuve

L'espace d'états contient l'état absorbant, et exactement $\sum_{i=0}^{k-1} \binom{w-1}{i}$ états transitoires, $\binom{w-1}{i}$ étant le nombre d'états codant des fenêtres qui comportent exactement i évènements.

CQFD

Notation. Pour faciliter la lecture, nous convenons d'adopter la notation suivante :

- L'état correspondant au cas de figure où la fenêtre courante ne comporte aucun accident sera noté état « 0 » ;
- L'état correspondant au cas de figure où la fenêtre courante comporte 1 accident exactement, cet accident étant localisé sur la $i^{\text{ème}}$ unité de temps de la fenêtre courante, sera noté état « i » ;
- L'état correspondant au cas de figure où la fenêtre courante comporte 2 accidents exactement, ces accidents étant localisés respectivement sur la $i^{\text{ème}}$ unité de temps et sur la $j^{\text{ème}}$ unité de temps de la fenêtre courante, sera noté état « (i,j) » ;
- ...
- L'état correspondant au cas de figure où la fenêtre courante comporte m accidents exactement, ces accidents étant localisés respectivement en position i_1, i_2, \dots, i_m , sera noté état « (i_1, i_2, \dots, i_m) » ;
- ...
- L'état correspondant au cas de figure où la fenêtre courante comporte $k-1$ accidents exactement, ces accidents étant localisés respectivement en position i_1, i_2, \dots, i_{k-1} , sera noté état « $(i_1, i_2, \dots, i_{k-1})$ ».

- *Condition initiale*

Le vecteur conditions initiales est le vecteur de taille $(r+1)$: $[1 \ 0 \ \dots \ 0]^t$.

- *Matrice de transition*

Le Tableau 7 explicite chaque transition (sans tenir compte de la simplification proposée c'est-à-dire lorsque l'on observe les w unités de temps) en fonction des valeurs prises par les variables aléatoire Y_u et Y_{u+w} . Le Tableau 8 explicite la probabilité de chacune de ces transitions. Le calcul des probabilités est explicité pour deux d'entre elles.

Tableau 7 : transitions réalisables

Etat de départ Etat d'arrivée	0	1	...	i $2 \leq i \leq w$...	$(1, i_2, \dots, i_m)$ $2 \leq i_2 < \dots < i_m \leq w$...	(i_1, i_2, \dots, i_m) $2 \leq i_1 < \dots < i_m \leq w$...	(i_1, \dots, i_{k-1}) $2 \leq i_1 < \dots < i_{k-1} \leq w$	A
0	(*) $\begin{cases} Y_u = 0 \\ Y_{u+w} = 0 \end{cases}$	$\begin{cases} Y_u = 1 \\ Y_{u+w} = 0 \end{cases}$...	-	...	-	...	-	...	-	-
...
i-1 $2 \leq i \leq w$	-	-	...	$\begin{cases} Y_u = 0 \\ Y_{u+w} = 0 \end{cases}$...	-	...	-	...	-	-
...
w	$\begin{cases} Y_u = 0 \\ Y_{u+w} = 1 \end{cases}$	$\begin{cases} Y_u = 1 \\ Y_{u+w} = 1 \end{cases}$...	-	...	-	...	-	...	-	-
...
(i-1, w) $2 \leq i \leq w$	-	-	...	$\begin{cases} Y_u = 0 \\ Y_{u+w} = 1 \end{cases}$...	-	...	-	...	-	-
...
(i_2-1, \dots, i_m-1) $2 \leq i_2 < \dots < i_m \leq w$	-	-	...	-	...	$\begin{cases} Y_u = 1 \\ Y_{u+w} = 0 \end{cases}$...	-	...	-	-
...
(i_2-1, \dots, i_m-1, w) $2 \leq i_2 < \dots < i_m \leq w$	-	-	...	-	...	$\begin{cases} Y_u = 1 \\ Y_{u+w} = 1 \end{cases}$...	-	...	-	-
...
$(i_1-1, i_2-1, \dots, i_m-1)$ $2 \leq i_1 < \dots < i_m \leq w$	-	-	...	-	...	-	...	$\begin{cases} Y_u = 0 \\ Y_{u+w} = 0 \end{cases}$...	-	-
...
$(i_1-1, i_2-1, \dots, i_m-1, w)$ $2 \leq i_1 < \dots < i_m \leq w$	-	-	...	-	...	-	...	$\begin{cases} Y_u = 0 \\ Y_{u+w} = 1 \end{cases}$...	-	-
...
$(i_1-1, \dots, i_{k-1}-1)$ $2 \leq i_1 < \dots < i_{k-1} \leq w$	-	-	...	-	...	-	...	-	...	$\begin{cases} Y_u = 0 \\ Y_{u+w} = 0 \end{cases}$	-
...
A	-	-	...	-	...	-	...	-	...	$\begin{cases} Y_u = 0 \\ Y_{u+w} = 1 \end{cases}$	1

(*) La valeur de Y_u est donnée par i_1 , elle est donc inutile à priori, nous l'avons précisée pour une meilleure lisibilité.

Tableau 8 : matrice de transition

	0	1	...	i $2 \leq i \leq w$...	$(1, i_2, \dots, i_m)$ $2 \leq i_2 < \dots < i_m \leq w$...	(i_1, i_2, \dots, i_m) $2 \leq i_1 < \dots < i_m \leq w$...	(i_1, \dots, i_{k-1}) $2 \leq i_1 < \dots < i_{k-1} \leq w$	A
0	q	q	...	-	...	-	...	-	...	-	-
...
$i-1$ $2 \leq i \leq w$	-	-	...	-	...	-	...	-	...	-	-
...
w	p	p	...	-	...	-	...	-	...	-	-
...
$(i-1, w)$ $2 \leq i \leq w$	-	-	...	q	...	-	...	-	...	-	-
...
(i_2-1, \dots, i_m-1) $2 \leq i_2 < \dots < i_m \leq w$	-	-	...	-	...	q	...	-	...	-	-
...
(i_2-1, \dots, i_m-1, w) $2 \leq i_2 < \dots < i_m \leq w$	-	-	...	-	...	p	...	-	...	-	-
...
$(i_1-1, i_2-1, \dots, i_m-1)$ $2 \leq i_1 < \dots < i_m \leq w$	-	-	...	-	...	-	...	q	...	-	-
...
$(i_1-1, i_2-1, \dots, i_m-1, w)$ $2 \leq i_1 < \dots < i_m \leq w$	-	-	...	-	...	-	...	p	...	-	-
...
$(i_1-1, \dots, i_{k-1}-1)$ $2 \leq i_1 < \dots < i_{k-1} \leq w$	-	-	...	-	...	-	...	-	...	q	-
...
A	-	-	...	-	...	-	...	-	...	p	1

Preuve : calcul des probabilités de transition

Sont réalisables :

- Les transitions d'un état $(y_0, y_1, \dots, y_{w-2})$ vers l'état $(y_1, \dots, y_{w-2}, 0)$. Elles ont lieu avec la probabilité $P(Y_{w-1} = 0) = q$.
- Les transitions d'un état $(y_0, y_1, \dots, y_{w-2})$ tel que $y_0 + y_1 + \dots + y_{w-2} \leq k-2$ vers l'état $(y_1, \dots, y_{w-2}, 1)$. Elles ont lieu avec la probabilité $P(Y_{w-1} = 1) = p$.
- Les transitions d'un état $(y_0, y_1, \dots, y_{w-2})$ tel que $y_0 + y_1 + \dots + y_{w-2} = k-1$ vers l'état A. Elles ont lieu avec la probabilité $P(Y_{w-1} = 1) = p$.
- Les boucles sur l'état A.

Nous introduisons donc l'application $\sigma : \mathcal{E}_{k,w} \times \{0,1\} \rightarrow \mathcal{E}_{k,w}$ telle que :

$$\sigma((y_0, \dots, y_{w-2}), 0) = \begin{cases} A & \text{si } (y_0, \dots, y_{w-2}) = A \\ (y_1, \dots, y_{w-2}, 0) & \text{sinon} \end{cases}$$

et

$$\sigma((y_0, \dots, y_{w-2}), 1) = \begin{cases} A & \text{si } y_0 + \dots + y_{w-2} = k-1 \\ (y_1, \dots, y_{w-2}, 1) & \text{sinon} \end{cases}$$

La matrice de transition correspondant à notre problème est la matrice $M = (m_{ij})_{1 \leq i, j \leq r}$, de dimensions $r \times r$ telle que :

$$m_{ij} = \begin{cases} p & \text{si } \sigma(e_j, 1) = e_i \\ q & \text{si } \sigma(e_j, 0) = e_i, 1 \leq i < r, 0 \leq j < r \end{cases}, m_{ir} = \begin{cases} 0 & \text{si } 1 \leq i < r \\ 1 & \text{si } i = r \end{cases} \text{ et } m_{ij} = 1 - \sum_{i=1}^{r-1} m_{ij}.$$

Application, transition de l'état 0 vers l'état w :

Une telle transition se produit lorsqu'un événement rentre dans la fenêtre courante, occupant ainsi la position w, ce qui arrive avec la probabilité $P(Y_{u+w-1} = 1) = p$.

Application, transition de l'état $(1, i_2, \dots, i_m)$, $2 \leq i_2 < \dots < i_m \leq w$, vers (i_2-1, \dots, i_m-1) , $2 \leq i_2 < \dots < i_m \leq w$:

Une telle transition se produit lorsque aucun événement ne rentre dans la fenêtre courante, ce qui arrive avec la probabilité $P(Y_{u+w-1} = 0) = q$.

CQFD

T itérations sont nécessaires pour couvrir la totalité de la fenêtre d'observation. Par conséquent, la probabilité d'observer au moins un (w,k)-cluster est donnée par le dernier terme du vecteur $M^T \Pi$.

II.3.b. Modèle de Bernoulli : application à l'exemple (T=365, w=10, k=3)

Pour cet exemple, nous traitons une matrice carrée de dimension $57 = 1 + \binom{10}{1} + \binom{10}{2} + 1$. Les états rencontrés par le système sont :

- l'état 0 : aucun accident dans la fenêtre courante ;
- les $n_i = \binom{10}{i}$ états notés i, $1 \leq i \leq 10$: un accident en position i dans la fenêtre courante ;

- les $n_2 = \binom{2}{10}$ états notés (i,j) , $1 \leq i < j \leq 10$: deux accidents en position i et j respectivement dans la fenêtre courante ;
- l'état absorbant A.

Nous reprenons le Tableau 9 pour construire la matrice de transition :

- De l'état 0, il est possible de rejoindre l'état 10 avec la probabilité p ou de rester sur l'état 0 avec la probabilité q ;
- De l'état 1, il est possible de rejoindre l'état 10 avec la probabilité p ou de rejoindre l'état 0 avec la probabilité q ;
- D'un état i , $2 \leq i \leq 10$, il est possible de rejoindre l'état $(i-1,10)$ avec la probabilité p ou l'état $i-1$ avec la probabilité q ;
- D'un état $(1, j)$, $2 \leq j \leq 10$, il est possible de rejoindre l'état $(j-1,10)$ avec la probabilité p ou l'état $j-1$ avec la probabilité q ;
- D'un état (i, j) , $2 \leq i < j \leq 10$, il est possible de rejoindre l'état absorbant A avec la probabilité p ou de rejoindre l'état $(i-1,j-1)$ avec la probabilité q ;
- Depuis l'état A, le seul mouvement possible est une boucle.

Tableau 9 : matrice de transition

Etats	0	1	2	3	4	5	6	7	8	9	10	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(1,7)	(1,8)	(1,9)	(1,10)	(2,3)	(2,4)	(2,5)	(2,6)	(2,7)	(2,8)	(2,9)	(2,10)	...	(i,j)	...	(8,9)	(8,10)	(9,10)	A			
0	q	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	q	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	q	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	q	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	q	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	0	0	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	p	p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(1,2)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(1,3)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(1,4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(1,5)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(1,6)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	0	0	0
(1,7)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	0	
(1,8)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	0	
(1,9)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	q	0	0	0	0	0	0	0	0	
(1,10)	0	0	p	0	0	0	0	0	0	0	0	p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,3)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,5)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,6)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,7)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,8)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,9)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,10)	0	0	0	p	0	0	0	0	0	0	0	0	p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
...	
(i-1,j-1)	
...	
(8,9)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	q	0		
(8,10)	0	0	0	0	0	0	0	0	0	0	p	0	0	0	0	0	0	0	0	p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(9,10)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p	0	0		
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p	p	p	p	p	p	p	p	p	p	p	p	p	p	p	1		

Le vecteur des conditions initiales est $\Pi = [1 \ 0 \ \dots \ 0]^t$.

Résultats :

La probabilité de l'état A après $T = 365$ itérations est donnée par le dernier terme du vecteur $M^T \Pi$ qui est égal à **0,1028**.

La probabilité que l'on ait un cluster de 3 accidents en 10 jours sur une année de 365 jours est donc de 0,1028.

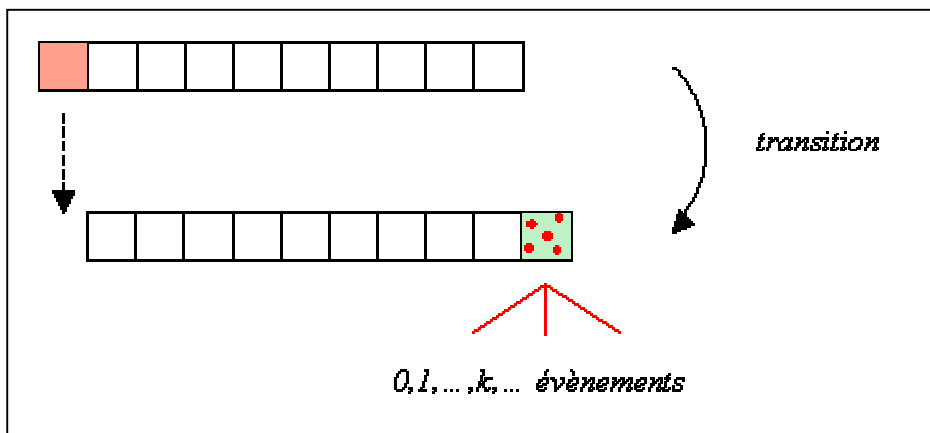
II.3.c. Modèle de Poisson discret

Nous présentons maintenant la théorie générale dans le cadre du modèle de Poisson discret. Ce modèle diffère du modèle de Bernoulli en ceci que le nombre d'événements par unité de temps n'est pas limité à 0 ou 1, mais peut prendre n'importe quelle valeur entière.

Le nombre d'événements susceptibles de se produire sur la $(u+1)^{i\text{ème}}$ unité de temps $[u, u+1[$ est décrit par la variable aléatoire $Y_u = N_1(u)$ qui suit une loi de Poisson de paramètre $\lambda \in [0, 1]$:

$$P\{Y_u = n\} = \frac{\lambda^n e^{-\lambda}}{n!}.$$

Figure 31 : transition dans le cadre du modèle poissonnien



La chaîne de Markov pour le modèle de Bernoulli est basée sur les éléments suivant :

- *Etats*

L'espace d'états est :

$$\mathcal{E}_{k,w} = \{(y_0, \dots, y_{w-2}) \in \{0, 1, \dots, k-1\}^{w-1} \mid y_0 + \dots + y_{w-2} < k\} \cup A,$$

Nous noterons encore $r+1$ la cardinalité de cet espace, que nous supposons ordonné selon l'ordre lexicographique.

- *Condition initiale*

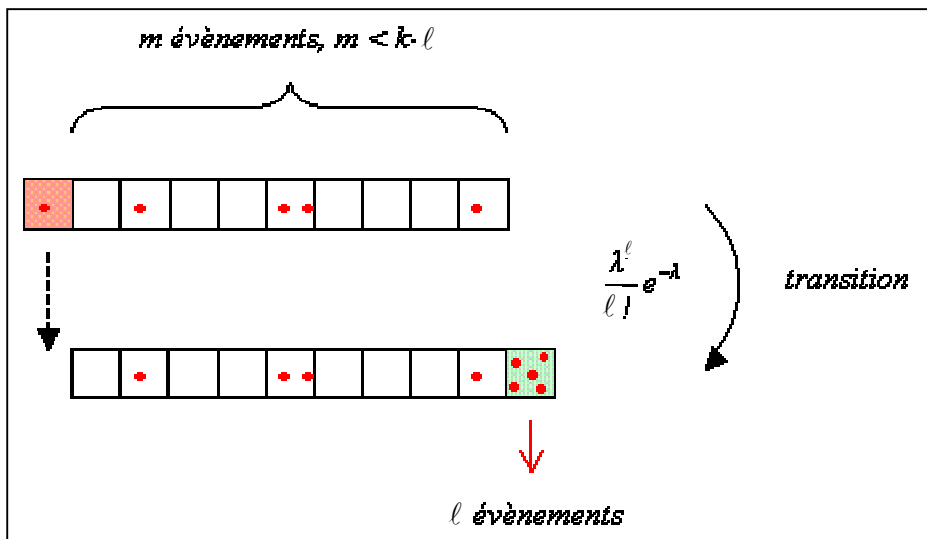
Le vecteur conditions initiales est le vecteur de taille $(r+1)$: $\Pi = [1 \ 0 \ \dots \ 0]^t$.

- *Matrice de transition*

Sont réalisables :

- Les transitions d'un état (y_0, \dots, y_{w-2}) tel que $y_0 + \dots + y_{w-2} < k - \ell$ vers l'état $(y_1, \dots, y_{w-2}, \ell)$ avec la probabilité $\frac{\lambda^\ell}{\ell!} e^{-\lambda}$ (cf. Figure 32), en particulier les transitions d'un état (y_0, \dots, y_{w-2}) vers l'état $(y_1, \dots, y_{w-2}, 0)$ avec la probabilité $e^{-\lambda}$.
- Les transitions d'un état (y_0, \dots, y_{w-2}) tel que $y_0 + y_1 + \dots + y_{w-2} = k - \ell$ vers l'état A avec la probabilité $1 - \sum_{i=0}^{\ell-1} \frac{\lambda^i}{i!} e^{-\lambda}$.
- Les boucles sur l'état A avec probabilité 1.

Figure 32 : transition



Nous introduisons l'application $\sigma : \mathcal{E}_{k,w} \times \{0, 1, 2, \dots, k-1\} \rightarrow \mathcal{E}_{k,w}$ telle que :

$$\sigma((y_0, \dots, y_{w-2}), 0) = \begin{cases} A & \text{si } (y_0, \dots, y_{w-2}) = A \\ (y_1, \dots, y_{w-2}, 0) & \text{sinon} \end{cases}$$

et

$$\sigma((y_0, \dots, y_{w-2}), r) = \begin{cases} A & \text{si } y_0 + \dots + y_{w-2} \geq k - r \\ (y_1, \dots, y_{w-2}, r) & \text{sinon} \end{cases}$$

La matrice de transition $M = (m_{ij})_{0 \leq i, j \leq r}$, de dimensions $(r+1) \times (r+1)$, est telle que :

$$m_{ij} = \begin{cases} \frac{\lambda^\ell}{\ell!} e^{-\lambda} & \text{si } \sigma(e_j, \ell) = e_i \\ 0 & \text{sinon} \end{cases} \quad \text{et } m_{ir} = \begin{cases} 0 & \text{si } 1 \leq i < r \\ 1 & \text{si } i = r \end{cases} \quad \text{et } m_{rj} = 1 - \sum_{i=0}^{r-1} m_{ij}.$$

La probabilité que nous recherchons est alors donnée par le dernier terme du vecteur $M^T \Pi$.

II.3.d. Modèle de Poisson discret : application à l'exemple ($T=365, w=10, k=3$)

Tenant compte de la simplification proposée dans les paragraphes précédents, nous devons traiter une matrice de dimension $56 = 1 + \binom{9}{1} + \binom{9}{2} + 1$.

Les états rencontrés sont :

- l'état 0 : aucun accident dans la fenêtre courante,
- les $n_1 = \binom{9}{1}$ états notés i , $1 \leq i \leq 9$: un accident en position i dans la fenêtre courante,
- les $n_2 = \binom{9}{2}$ états notés i' , $1 \leq i \leq 9$: deux accidents sur la $i^{\text{ème}}$ unité de temps de la fenêtre courante,
- les $n_3 = \binom{9}{2}$ états notés (i,j) , $1 \leq i < j \leq 10$: deux accidents en position i et j respectivement dans la fenêtre courante,
- l'état absorbant noté A.

Le vecteur conditions initiales est le vecteur de taille 56 : $\Pi = (1, 0, \dots, 0)^t$.

La matrice de transition est la matrice $M = (m_{ij})_{1 \leq i, j \leq 56}$:

$$m_{ij} = \begin{cases} e^{-\lambda} & \text{si } \sigma(e_j, 0) = e_i \\ \lambda e^{-\lambda} & \text{si } \sigma(e_j, 1) = e_i \\ \frac{\lambda^2}{2} e^{-\lambda} & \text{si } \sigma(e_j, 2) = e_i \\ 0 & \text{sinon} \end{cases}, \text{ et } m_{i56} = \begin{cases} 0 & \text{si } 1 \leq i < 56 \\ 1 & \text{si } i = 56 \end{cases}, \text{ et } m_{56j} = 1 - \sum_{i=1}^{55} m_{ij}.$$

Cette matrice est représentée sur le Tableau 10.

Résultat :

La probabilité de l'état 3 après $T = 365$ itérations est donnée par le dernier terme du vecteur $M^N \Pi$ qui est égal à **0,1217**.

La probabilité que l'on ait un cluster de 3 accidents en 10 jours sur une année de 365 jours est donc de 0,1217.

Tableau 10 : matrice de transition

	0	1	2	3	...	i	...	9	1'	2'	3'	...	i'	...	9'	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(1,7)	(1,8)	(1,9)	(2,3)	(2,4)	(2,5)	(2,6)	(2,7)	(2,8)	(2,9)	...	(i,j)	...	(8,9)	3	
0	p_0	p_0	0	0	...	0	...	0	p_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	p_0	0	...	0	...	0	0	0	0	0	p_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	p_0	0	0	0	0	0	0	p_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	p_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
...	
i-1	p_0	
9	p_1	p_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1'	0	0	0	0	0	0	p_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2'	0	0	0	0	0	0	0	p_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3'	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
...
i'-1	p_0	
9'	p_2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(1,2)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p_0	0	0	0	0	0	0	0	0	0	
(1,3)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p_0	0	0	0	0	0	0	0	0	
(1,4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p_0	0	0	0	0	0	0	0	
(1,5)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p_0	0	0	0	0	0	0	
(1,6)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p_0	0	0	0	0	0
(1,7)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p_0	0	0	0	0
(1,8)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p_0	0	0	
(1,9)	0	0	p_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,3)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,5)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,6)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,7)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,8)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
(2,9)	0	0	0	p_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
...
(i-1,j-1)	p_0	
...
(i-1,9)	p_1
...
(8,9)	0	0	0	0	p_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	*	*	*	*	...	*	...	*	*	*	*	...	*	...	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	...	*	...	*	1

* complément à 1 sur chaque colonne

** état noté i' , $1 \leq i \leq 9$: deux accidents sur la $i^{\text{ème}}$ unité de temps de la fenêtre courante

III. Récapitulatif

Modèle	BERNOULLI		POISSON DISCRET		
	1 jour	1 heure	1 jour	1 heure	Temps continu
Simulation de Monte Carlo classique	0,1250	0,1310	0,1220	0,1301	0,1329
Simulation de Monte Carlo animée par un RdP	0,1225	0,1251	NaN*	NaN	0,1317
Premier modèle Markovien	0,0991	0,1274	0,1176	0,1280	NaN
Deuxième modèle Markovien (double fenêtre)	0,1014	0,1296	NaN	NaN	NaN
Troisième modèle Markovien (complet)	0,1028	NaN	0,1217	NaN	NaN

*NaN=Non available Number

ANNEXE : SIMULATION DE MONTE-CARLO SOUS VISUAL BASIC

Modèle de Bernoulli

```
Sub MacroResultats()
```

```
Nb_jours = 365  
Nb_simu = 1000  
Nb_acc = Cells(3, 2)  
agglo = Cells(1, 2)
```

```
For i = 0 To Nb_simu - 1
```

```
    cpt = 0  
    i_deb = 5  
    i_fin = 5 + Nb_acc - 1  
    cpt_acc = 0
```

```
    For j = 1 To Nb_jours  
        'jour_i est 1 si accident  
        alea_j = Rnd()  
        If alea_j <= Cells(2, 2) Then  
            jour_j = 1  
            cpt_acc = cpt_acc + 1  
            Cells(5 + i, 4 + cpt_acc) = j  
        End If  
    Next j  
    Cells(5 + i, 3) = cpt_acc
```

```
    While i_fin <= cpt_acc + 4
```

```
        dat_deb = Cells(5 + i, i_deb)  
        dat_fin = Cells(5 + i, i_fin)  
        ecart = dat_fin - dat_deb + 1
```

```
        If ecart <= agglo Then  
            Range(Cells(5 + i, i_deb), Cells(5 + i, i_fin)).Select  
            With Selection.Interior  
                .ColorIndex = 3  
            End With  
            cpt = cpt + 1  
            taille_agglo = Nb_acc  
            cpt_agglo = 0  
            dat_0 = dat_deb  
            dat_fin = Cells(5 + i, i_fin + 1)
```

```
            While dat_fin - dat_0 <= agglo And dat_fin <> vide  
                taille_agglo = taille_agglo + 1  
                cpt_agglo = cpt_agglo + 1  
                Cells(5 + i, i_fin + cpt_agglo).Select
```

```
With Selection.Interior
.ColorIndex = 3
End With
dat_fin = Cells(5 + i, i_fin + cpt_agglo + 1)
Wend
```

```
i_deb = i_deb + 1
i_fin = i_fin + cpt_agglo + 1
```

```
End If
```

```
If ecart > agglo Then
```

```
If i_fin - i_deb > Nb_acc - 1 Then
i_deb = i_deb + 1
Else
i_deb = i_deb + 1
i_fin = i_fin + 1
End If
```

```
dat_deb = Cells(5 + i, i_deb)
dat_fin = Cells(5 + i, i_fin)
```

```
End If
```

```
Wend
```

```
Cells(5 + i, 2) = cpt
If cpt > 0 Then
Cells(5 + i, 4) = 1
Else
Cells(5 + i, 4) = 0
End If
```

```
Next i
```

```
End Sub
```

Modèle Poissonien

```
Sub MacroResultats()
```

```
Nb_jours = 365
Nb_simu = 1000
Nb_acc = Cells(3, 2)
agglo = Cells(1, 2)
```

```
For i = 0 To Nb_simu - 1
```

```
cpt = 0
```

```

i_deb = 5
i_fin = 5 + Nb_acc - 1

dat_i = (1 / Cells(2, 2)) * (-Application.WorksheetFunction.Ln(Rnd()))
cpt_acc = 0

While dat_i <= Nb_jours
    cpt_acc = cpt_acc + 1
    Cells(5 + i, 4 + cpt_acc) = dat_i
    dat_i = dat_i + (1 / Cells(2, 2)) * (-Application.WorksheetFunction.Ln(Rnd()))

Wend

Cells(5 + i, 3) = cpt_acc

While i_fin <= cpt_acc + 4

    dat_deb = Cells(5 + i, i_deb)
    dat_fin = Cells(5 + i, i_fin)
    ecart = dat_fin - dat_deb

    If ecart <= agglo Then
        Range(Cells(5 + i, i_deb), Cells(5 + i, i_fin)).Select
        With Selection.Interior
            .ColorIndex = 3
        End With
        cpt = cpt + 1
        taille_agglo = Nb_acc
        cpt_agglo = 0
        dat_0 = dat_deb
        dat_fin = Cells(5 + i, i_fin + 1)

        While dat_fin - dat_0 <= agglo And dat_fin <> vide
            taille_agglo = taille_agglo + 1
            cpt_agglo = cpt_agglo + 1
            Cells(5 + i, i_fin + cpt_agglo).Select
            With Selection.Interior
                .ColorIndex = 3
            End With
            dat_fin = Cells(5 + i, i_fin + cpt_agglo + 1)
        Wend

        i_deb = i_deb + 1
        i_fin = i_fin + cpt_agglo + 1

    End If

    If ecart > agglo Then

        If i_fin - i_deb > Nb_acc - 1 Then

```

```
        i_deb = i_deb + 1
    Else
        i_deb = i_deb + 1
        i_fin = i_fin + 1
    End If

    dat_deb = Cells(5 + i, i_deb)
    dat_fin = Cells(5 + i, i_fin)

End If

Wend

Cells(5 + i, 2) = cpt
If cpt > 0 Then
    Cells(5 + i, 4) = 1
Else
    Cells(5 + i, 4) = 0
End If

Next i

End Sub
```


Chapitre 3

Définitions et état de l'art :

Fiabilité et Retour d'EXpérience

I. Fiabilité : la problématique industrielle

Pour évaluer la fiabilité de leurs matériels, les équipementiers ont recours à plusieurs techniques. Les trois démarches principales sont présentées ci-dessous.

- Une première démarche consiste à utiliser des banques de données de fiabilité de composants.

Dans le domaine de l'électronique, et plus particulièrement dans celui de l'électronique embarquée, la MIL-HDBK-217 F [MIL-HDBK-217F] est le recueil le plus couramment utilisé. Il propose des modèles empiriques de calcul de taux de défaillance pour la plupart des composants électroniques et pour quelques composants électromécaniques.

Ces modèles fournissent un taux de défaillance de base, taux de défaillance qui sera ensuite ajusté grâce à des coefficients multiplicateurs, ce en fonction de la qualité du composant et de l'utilisation particulière qu'en fera le client : température de fonctionnement, tension,

Malheureusement, les modèles de la MIL-HDBK-217 F n'ont pas été remis à jour depuis les années 90 et ne sont plus adaptés aux nouvelles technologies. Pour faire face à ce problème, un consortium d'industriels français des domaines de l'aéronautique et de la défense a créé, sous l'égide de la Délégation Générale pour l'Armement (DGA), une méthodologie d'ingénierie de la fiabilité en électronique [FIDES, 2004].

- Il est possible de réaliser des essais sur des échantillons de matériels.

L'objectif des essais de fiabilité [ASTE, 1993] [Charruau et al., 2006] est de vérifier le bon fonctionnement de produits soumis à des conditions de stress les plus proches possibles de la réalité.

Il faut bien-sûr être vigilant sur un certain nombre de critères : population homogène, échantillon aléatoire, conformité des contraintes appliquées aux contraintes réelles subies par les équipements livrables,

Si les produits testés sont particulièrement fiables, ce qui est le cas pour la majorité des applications électroniques, ces essais peuvent durer très longtemps... Lorsqu'il n'est pas envisageable, pour des contraintes de temps et d'argent (les deux étant liées...), de réaliser de longs essais, il est possible de recourir aux essais accélérés. Le principe de tels essais est d'amplifier les contraintes appliquées aux échantillons testés. L'application de niveaux de stress élevés accélère les mécanismes de dégradation et précipite l'arrivée des défaillances, réduisant la durée nécessaire pour estimer les caractéristiques comportementales du produit.

- Enfin, une autre démarche consiste à évaluer la fiabilité à partir des retours opérationnels, en faisant appel à des méthodes statistiques de traitement des données de vie, que nous développerons ultérieurement.

L'avantage de cette démarche est de ne point être empirique, mais de fournir au contraire une estimation de la fiabilité réellement observée, sur des matériels en exploitation. Son principal inconvénient est qu'il faut attendre quelques années de service du matériel considéré pour avoir un retour d'expérience suffisamment riche.

De plus, si l'on est certain que toute panne ayant lieu durant la période de garantie est connue de l'équipementier, les pannes se produisant au-delà de la période de garantie ne sont pas systématiquement reportées. L'étude est donc limitée dans le temps. Enfin, le statisticien se heurte aux problèmes courants de toute collecte de données : informations manquantes, données erronées,

La première démarche que nous avons citée, autrement dit le recours aux banques de données, est la plupart du temps utilisée en phase de conception ou de validation du produit, avant les premières livraisons. On parlera de *fiabilité prévisionnelle*. Le recours aux essais peut se situer à n'importe quel moment du cycle de vie du produit. La troisième démarche, quand à elle, n'est applicable qu'en période d'exploitation. On parlera de *fiabilité opérationnelle* et de Retour d'Expérience (REX).

II. Définitions

L'Union Technique de l'Electricité⁴ (UTE), sur recommandation de la Commission Electrotechnique Internationale, a proposé la définition suivante de la fiabilité [CEI 271, 1974] :

« *Aptitude d'un dispositif à accomplir une fonction requise, dans des conditions données et pour une période de temps donnée* ».

II.1. Fiabilité

La définition de la fiabilité, énoncée plus haut, est intuitive et qualitative. Par commodité, « fiabilité » et « mesure de la fiabilité » sont confondues. La définition probabiliste de ce concept ainsi que des différentes grandeurs qui lui sont inhérentes [Barlow et Proschan, 1996] [Faure et Laurière, 1974] [Pagès et Gondran, 1990], est donnée ci-dessous.

Supposons que nous observions le comportement d'un dispositif (matériel, équipement, composant, ...) en fonctionnement.

Nous introduisons la variable aléatoire T qui représente le temps écoulé depuis la mise en service du dispositif à l'instant $t = 0$ jusqu'à l'instant de sa première défaillance. La variable aléatoire T représente donc la durée de vie du dispositif ou, de manière équivalente, l'instant de sa défaillance.

La variable aléatoire T est positive, elle est de plus supposée absolument continue.

▪ Fiabilité

La fonction de fiabilité est définie par

$$R(t) = P(T \geq t), t \geq 0,$$

pour un t fixé. Elle représente la probabilité de bon fonctionnement du dispositif étudié sur l'intervalle de temps $[0, t]$.

La fiabilité est donc une fonction du temps, encore appelée fonction de survie.

⁴ Organisme français de normalisation électrotechnique

Remarque : $R(t)$ est une fonction monotone décroissante à valeurs dans $[0,1]$, telle que $R(0) = 1$ et $\lim_{t \rightarrow +\infty} R(t) = 0$.

Nous notons $F(t)$ la fonction de répartition de la variable aléatoire T . L'absolue continuité de T nous permet de définir sa densité de probabilité, notée $f(t)$.

- Fonction de répartition ou défiabilité

La fonction de répartition de la variable aléatoire T est

$$F(t) = P(T < t) = 1 - R(t),$$

pour un t fixé, et représente la probabilité de défaillance à un instant quelconque précédent l'instant t .

Par définition, on a $F(t) = 0$ pour $t < 0$.

- Densité de probabilité

La densité de probabilité est une fonction $f(t) \geq 0$ telle que pour tout $t \geq 0$,

$$F(t) = \int_0^t f(u) du.$$

En admettant que la fonction de répartition admette une dérivée au point t , nous pouvons écrire

$$f(t) = \lim_{dt \rightarrow 0} \frac{P(t \leq T < t + dt)}{dt} = F'(t) = -R'(t),$$

pour un t fixé, illustrant ainsi le fait que la quantité $f(t)dt$ est la probabilité de défaillance sur un petit intervalle de temps suivant l'instant t .

- Taux de défaillance

Le taux de défaillance $\lambda(t)$ est la fonction

$$\lambda(t) = \lim_{dt \rightarrow 0} \frac{P(t \leq T < t + dt | T > t)}{dt} = \frac{f(t)}{R(t)}.$$

Pour un t fixé, $\lambda(t)dt$ est la probabilité de défaillance du matériel sur un petit intervalle de temps suivant l'instant t , sachant qu'il a correctement fonctionné jusqu'à t .

Pour h suffisamment petit, nous avons

$$h \times \lambda(t) = P(t < T \leq t + h | T > t).$$

La relation suivante permet de relier le taux de défaillance et la fiabilité :

$$\lambda(t) = -\frac{R'(t)}{R(t)} = -\frac{d}{dt}[\ln R(t)].$$

Des exemples de chacune de ces fonctions seront données plus bas pour certaines lois de fiabilité usuelles.

II.2. Espérance et variance

- L'espérance de la variable aléatoire T,

$$E(T) = \int_0^{+\infty} tf(t)dt,$$

est, par définition de T, le temps moyen de bon fonctionnement avant la première défaillance. Cette quantité essentielle en fiabilité est couramment nommée MTTF ou Mean Time To Failure.

Le MTTF est relié à la fiabilité par :

$$MTTF = E(T) = \int_0^{+\infty} R(t)dt.$$

La preuve de cette relation est très classique, mais elle n'est pas toujours connue dans le milieu industriel. Nous la rappelons donc ci-dessous :

Preuve :

Un changement de variable conduit à :

$$E(T) = \int_0^{+\infty} tf(t)dt = [-tR(t)]_0^{+\infty} + \int_0^{+\infty} R(t)dt.$$

Le premier terme tend vers 0 à condition que la fonction de fiabilité R(t) décroisse plus vite que $t \rightarrow \frac{1}{t}$, ce que l'on admet en général.

CQFD

- La variance de la variable aléatoire T,

$$\text{Var}(T) = \int_0^{+\infty} t^2 f(t)dt - [E(T)]^2,$$

est reliée à la fiabilité par :

$$\text{Var}(T) = 2 \int_0^{+\infty} tR(t)dt - [E(T)]^2.$$

Preuve :

Un changement de variable conduit à :

$$\int_0^{+\infty} t^2 f(t) dt = \left[-t^2 R(t) \right]_0^{+\infty} + 2 \int_0^{+\infty} t R(t) dt.$$

Le premier terme tend vers 0 à condition que la fonction de fiabilité $R(t)$ décroisse plus vite que $t \rightarrow \frac{1}{t^2}$, ce que l'on admet en général.

CQFD

III. Lois de probabilité usuelles en fiabilité

Nous présentons dans ce paragraphe les lois de fiabilité les plus couramment utilisées.

III.1. La loi exponentielle

Cette loi est employée dans de très nombreuses applications. Elle est notamment prise comme hypothèse dans la plupart des banques de données de fiabilité, pour sa simplicité d'utilisation.

La distribution exponentielle est la seule distribution continue assurant un taux de défaillance constant.

- Densité de probabilité :

$$f(t) = \lambda \exp(-\lambda t).$$

- Fiabilité :

$$R(t) = \exp(-\lambda t).$$

- Fonction de répartition / défiabilité :

$$F(t) = 1 - \exp(-\lambda t).$$

- Taux de défaillance :

$$\lambda(t) = \frac{f(t)}{V(t)} = \lambda.$$

- MTTF :

$$MTTF = \int_0^{+\infty} R(t) dt = \frac{1}{\lambda}.$$

- Variance :

$$\text{Var}(T) = 2 \int_0^{+\infty} tR(t)dt - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Une propriété importante de la loi exponentielle est d'être « sans mémoire » :

Une variable aléatoire T à valeurs dans \mathbb{R}^+ suit une loi exponentielle si et seulement si
 $\forall t \geq 0, \forall h \geq 0, P(T \geq t+h | T \geq t) = P(T \geq h)$

La loi exponentielle est la seule loi continue vérifiant cette propriété.

Remarque : la loi géométrique est, quant à elle, la seule loi discrète « sans mémoire ».

III.2. Loi de Weibull

La loi de Weibull est une loi continue à trois paramètres :

- le paramètre de position γ qui représente le décalage pouvant exister entre le début de l'observation (date à laquelle on commence à observer un échantillon) et le début du processus que l'on observe (date à laquelle s'est manifesté pour la première fois le processus observé) ;
- le paramètre d'échelle η , qui, comme son nom l'indique, nous renseigne sur l'étendue de la distribution ;
- le paramètre de forme β , qui est associé à la cinétique du processus observé

- Densité de probabilité :

$$f(t) = \frac{\beta}{\eta} \left(\frac{t-\gamma}{\eta} \right)^{\beta-1} \exp\left(- \left(\frac{t-\gamma}{\eta} \right)^\beta \right).$$

Le paramètre de position γ étant souvent nul, on se ramène à

$$f(t) = \frac{\beta}{\eta} \left(\frac{t}{\eta} \right)^{\beta-1} \exp\left(- \left(\frac{t}{\eta} \right)^\beta \right).$$

- Fiabilité :

$$R(t) = \exp\left[- \left(\frac{t}{\eta} \right)^\beta \right].$$

- Fonction de répartition / défiabilité :

$$F(t) = 1 - \exp\left[- \left(\frac{t}{\eta} \right)^\beta \right].$$

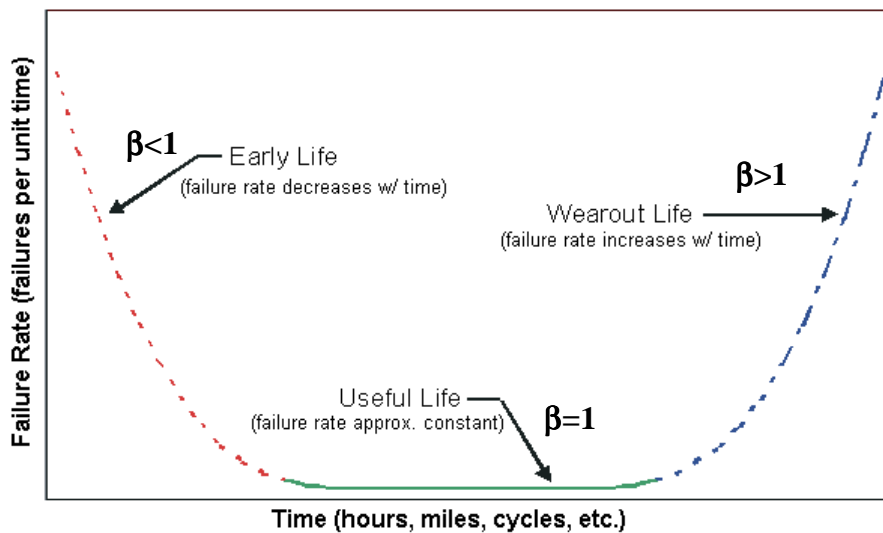
- Taux de défaillance :

$$\lambda(t) = \frac{\beta}{\eta} \left(\frac{t}{\eta} \right)^{\beta-1}.$$

Remarques :

- Nous constatons fréquemment que le taux de panne d'un équipement durant son cycle de vie suit une « courbe en baignoire », caractérisée par trois périodes :
 - La période de jeunesse, ou de mortalité infantile, présente un taux de défaillance relativement élevé mais décroissant.
 - La période de vie utile est caractérisée par un taux de défaillance constant et faible : le matériel est « mature » et présente peu de défauts.
 - La période de vieillissement ou d'usure présente un taux de défaillance croissant, du à la manifestation de phénomènes de dégradation (usure, fatigue,...).
- Chacune de ces trois plages, considérée individuellement, peut être modélisée par une loi de Weibull de paramètre $\beta < 1$, $\beta = 1$, $\beta > 1$ respectivement.
- La loi exponentielle est un cas particulier de la loi de Weibull correspondant au cas où le paramètre de forme vaut 1.

Figure 33: courbe en "baignoire"



Les graphes ci-dessous représentent les fonctions densité de probabilité (Figure 34), fiabilité (Figure 35) et taux de défaillance (Figure 36) pour différentes valeurs de β .

Figure 34: densité de probabilité

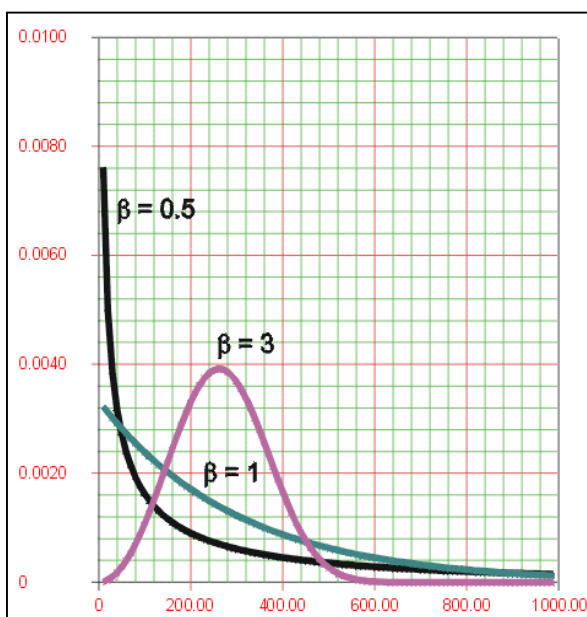


Figure 35: fiabilité

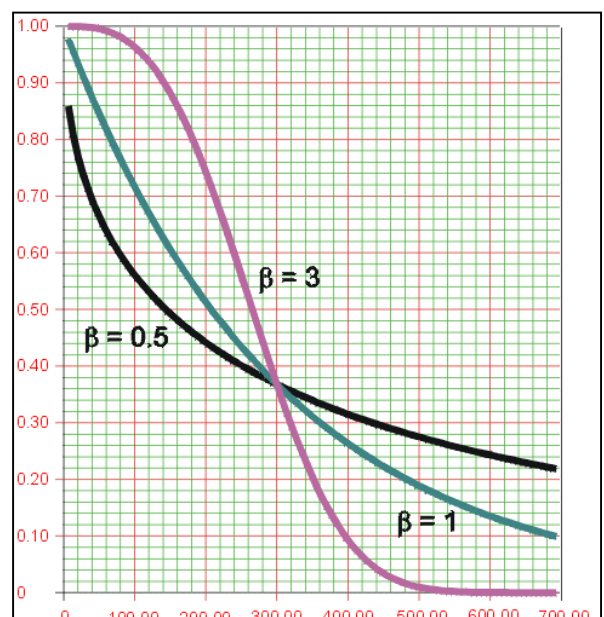
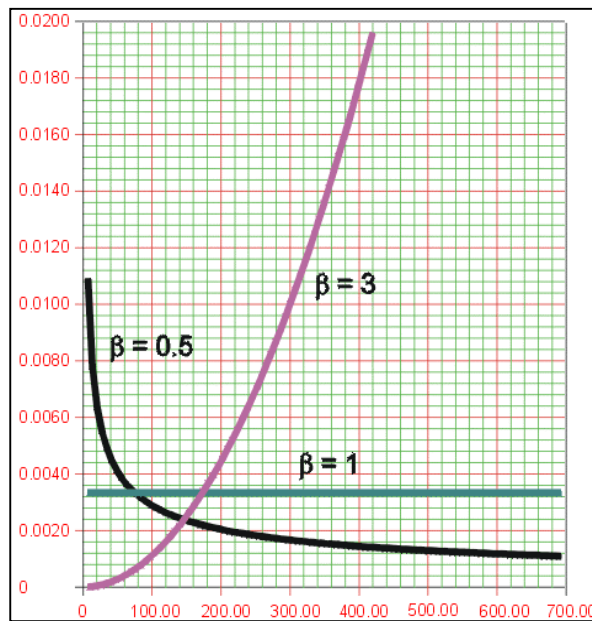


Figure 36: taux de défaillance



IV. Les méthodes statistiques d'exploitation du Retour d'EXpérience

Dans cette partie, nous nous intéressons aux techniques statistiques de traitement des données de fiabilité. Il s'agit d'estimer les paramètres de fiabilité d'un individu (fiabilité, taux de défaillance, MTTF) grâce aux données obtenues des retours clients.

Un certain nombre de méthodes présentées dans la littérature [Bacha et al., 1998] [Hoang, 2003] [Lannoy et Procaccia, 1994] [Procaccia et Piepszownik, 1992] permettent d'estimer la loi de survie intrinsèque d'un individu à partir de l'étude d'un échantillon, notamment lorsque les données sont fortement censurées. Nous présentons ici trois de ces méthodes :

- La méthode de Kaplan-Meier,
- La méthode des rangs corrigés de Johnson,
- La méthode du maximum de vraisemblance.

Les deux premières sont non paramétriques tandis que la troisième est paramétrique.

IV.1. Le problème des données censurées

Rappelons que nous souhaitons estimer la fonction de fiabilité $R(t) = P(T > t)$ d'une population donnée, ou du moins une quantité qui lui est reliée (taux de défaillance, défiabilité, ...), T étant la variable aléatoire mesurant le temps écoulé depuis la mise en service jusqu'à la première défaillance.

Si l'on souhaite faire appel aux statistiques, le bon sens suggère de prélever un échantillon aléatoire de matériels dans la population qui nous intéresse, et de relever le temps de survie de chaque individu constituant l'échantillon.

Le terme « temps de survie » est utilisé ici pour désigner la durée de bon fonctionnement du matériel jusqu'à la défaillance.

Mais les données de survie que l'on collecte sont généralement censurées, c'est-à-dire que les temps de survie effectifs sont inconnus pour un certain nombre d'individus de l'échantillon.

Il y a plusieurs types de censure, correspondant chacun à une situation bien particulière. Nous listons plus bas les quatre plus fréquents :

- Lorsque l'analyse de fiabilité est réalisée avant que tous les matériels ne soient défectueux, certains individus sont « survivants » à l'issue de l'observation. Pour ces derniers, on ne dispose pas du temps de survie mais simplement d'une borne inférieure. Les données correspondantes sont *censurées à droite*.
On parle de *censure de type I* (time censoring).
- Lorsqu'il est décidé par avance de poursuivre l'observation jusqu'à ce qu'un nombre suffisant de défaillances, disons r , soit observé, on se retrouve comme précédemment dans le cas de *données censurées à droite*.
On parle de *censure de type II* (failure censoring).
- Lorsque les défaillances ne sont découvertes qu'aux instants d'inspection, aucun temps de survie n'est connu de manière exacte. Au lieu de cela, on dispose d'une borne inférieure et d'une borne supérieure sur chacun d'entre eux.
Si un matériel est constaté défectueux lors de la première inspection, on obtient une borne supérieure sur le temps de survie et la donnée correspondante est *censurée à gauche*. Si un matériel est survivant lors de la dernière inspection, on obtient une borne inférieure sur le temps de survie (la borne supérieure étant $+\infty$) et la donnée est *censurée à droite*. Enfin, si un matériel est constaté défectueux lors d'une inspection quelconque, on obtient un intervalle de censure.
Les données sont dites *censurées par intervalles*.
- Si les matériels sont mis en exploitation à des dates différentes et si, de plus, l'analyse est menée avant que tous ne soient défectueux, on parlera de *censures multiples à droite*. Notons que, dans ce cas, il est possible de rencontrer des durées de bon fonctionnement supérieures à des durées censurées. C'est le cas dans l'exemple 2 présenté plus bas.

Ainsi, quel que soit le type de censure, nous récupérerons deux types de données :

- Des temps de défaillance pour les équipements qui ont été défectueux sur la période d'observation,
- Des temps de censure pour les équipements qui n'ont pas encore été défectueux à la fin de la période d'observation (censures à droite) ou pour ceux qui sont tombés en panne avant une date d'inspection (censures à gauche).

Les différents mécanismes de censures de censure sont expliqués dans de nombreux ouvrages, notamment [Meeker et Escobar, 1998].

Nous présentons ci-dessous un exemple de censures de type I et un exemple de censures multiples à droite.

Exemple 1 : L'exemple ci-dessous présente les données de survie de $n = 48$ cartes électroniques sur lesquelles était suspecté un problème de corrosion du circuit imprimé. Des défaillances ont été relevées sur 5 des matériels en service, entre 4500 et 7300 heures de fonctionnement effectif (sous tension). Les 43 autres matériels ont eux fonctionné 7400 heures sans défaillance. Nous avons donc 5 données de défaillance et 43 données censurées à droite, et il s'agit de données censurées de type I.

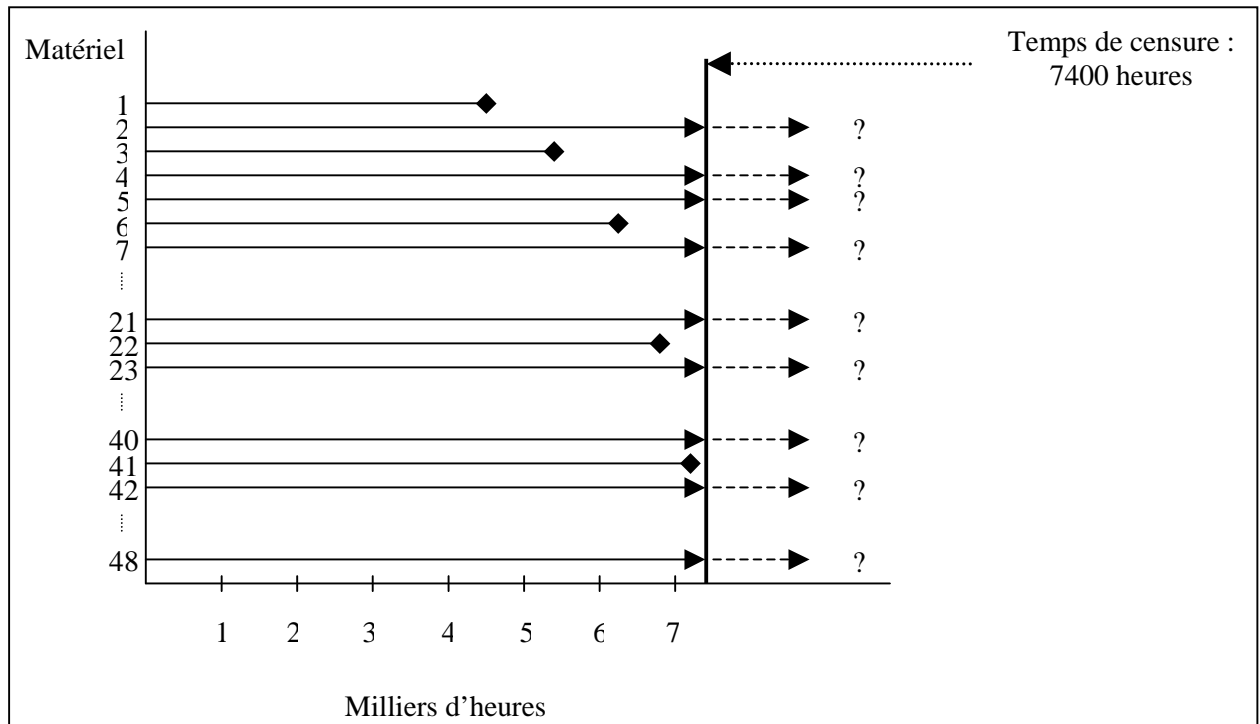
Les données relatives à cet exemple sont présentées sous deux formes différentes, traditionnellement rencontrées dans la littérature. Le tableau 11 liste les dates de défaillance observées ; le nombre de données censurées est donné à titre informatif en commentaire. La figure 37 illustre la double présence de données de défaillance et de données censurées.

Tableau 11: Temps de défaillance de cartes électroniques en heures de fonctionnement

4500	5400	6250	6800	7300
------	------	------	------	------

Après 7400 heures, 43 cartes étaient toujours en fonctionnement.

Figure 37 : Comportement général des cartes électroniques. Les données de défaillance et les données censurées sont visibles.



Exemple 2 : L'exemple 2 présente les données de survie d'un échantillon de visualisations mis en service de manière échelonnée au long de l'année 2004. Cet échantillon est observé à la fin 2005.

Des pannes ont été reportées pour 76 des matériels étudiés, 341 étant encore en fonctionnement au moment de l'étude.

On est donc dans le cas de censures multiples, avec des temps de défaillance s'échelonnant entre 1 et 24 mois et des temps de censure courant de 13 à 24 mois.

Les données relatives à cet exemple sont présentées sous forme d'histogramme (figure 38) puis de graphiques (figures 39 et 40).

Figure 38: Diagramme en bâtons représentant les temps de défaillance et les temps de censure pour la population de l'exemple 2

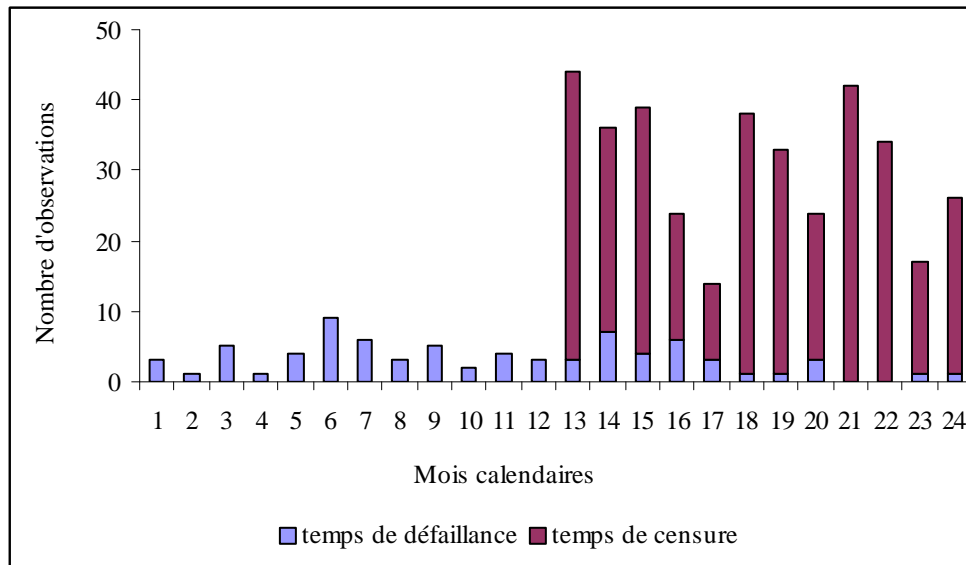
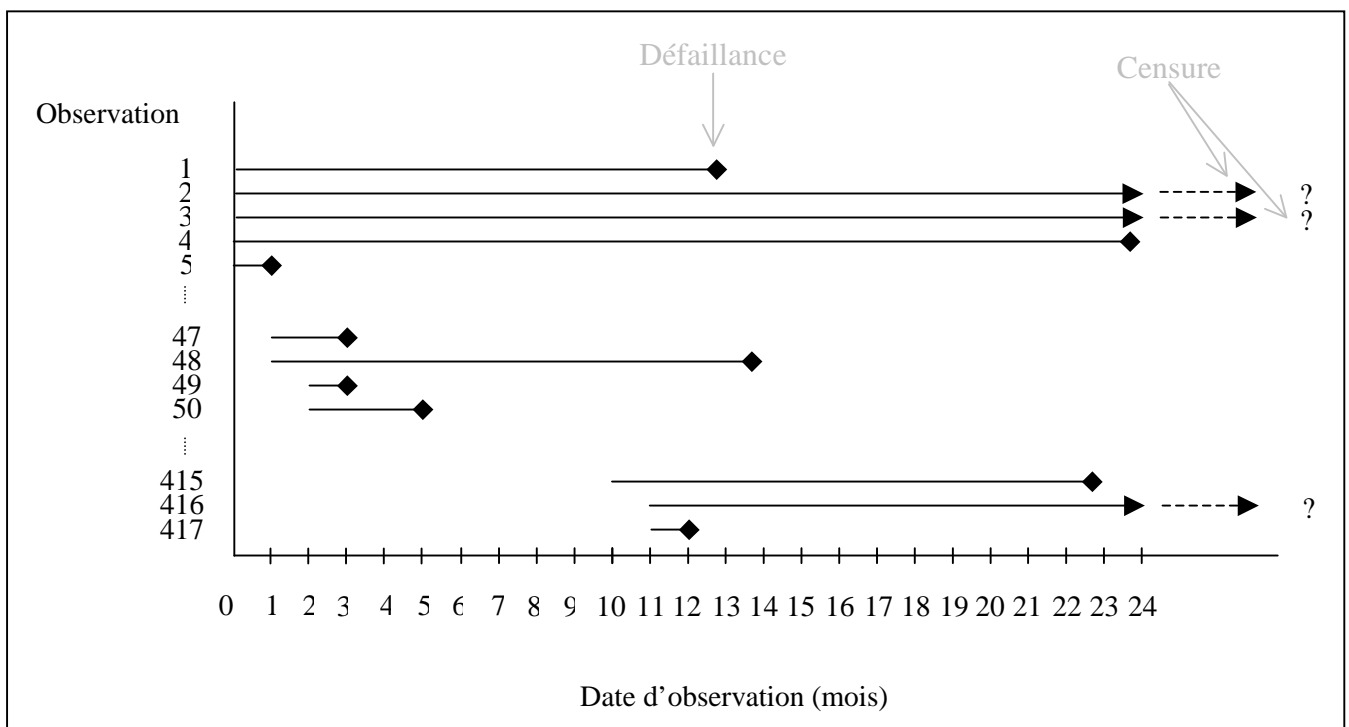
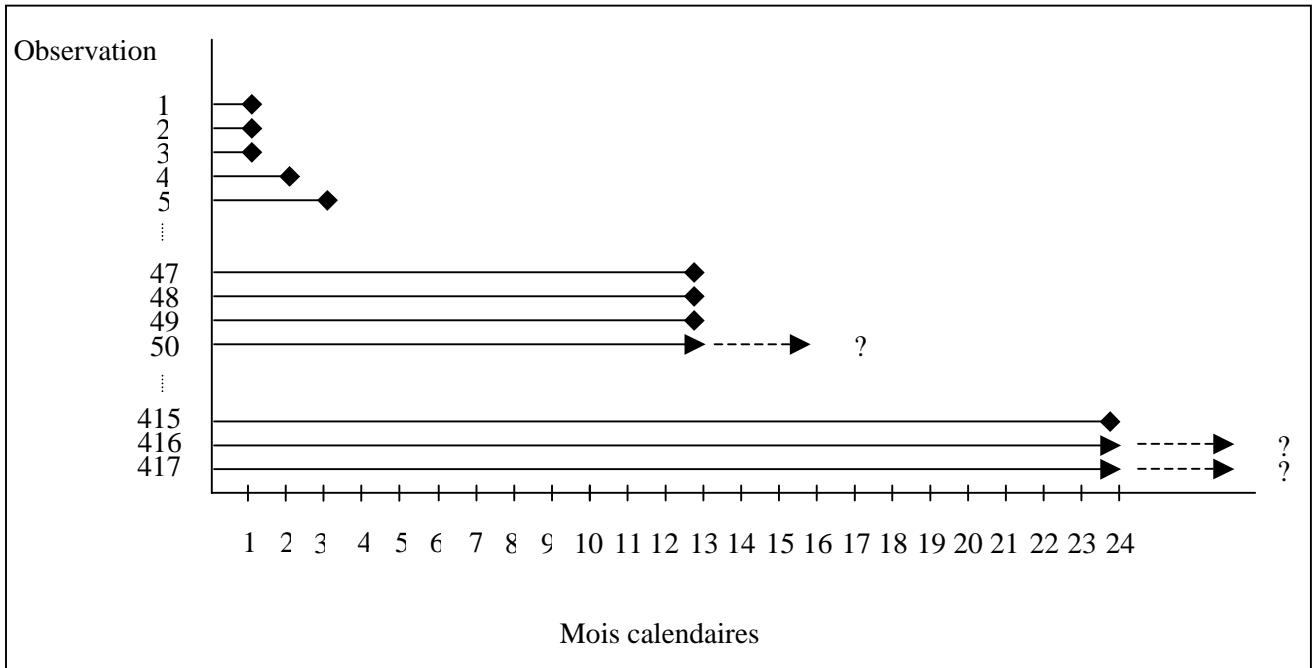


Figure 39 : Comportement général des visualisations, illustrant les mises en service échelonnées ainsi que les défaillances et les censures



Pour utiliser la méthode de Kaplan-Meier, il est nécessaire de translater les temps de défaillances et de censures afin de les ramener à une même origine. La figure ci-dessous présente les données translattées.

Figure 40 : Données translâtées pour un recalage en zéro.



Regardons l'**exemple 2**. Considérons uniquement les temps de défaillance. Nous allons vérifier que la loi de Weibull s'ajuste correctement à ces temps.

- *Droite de Weibull :*

Dans le cas où la loi de survie de la population considérée est une loi de Weibull, les points de coordonnées $\left(\ln(t), \ln \left[\ln \left(\frac{1}{1-F(t)} \right) \right] \right)$ sont alignés. La droite d'alignement a pour coefficient directeur le paramètre β de la loi, et pour ordonnée à l'origine la quantité $-\beta \ln(\eta)$.

Notons t_i , $1 \leq i \leq 76$, les temps de défaillance des visualisations étudiées dans l'exemple 2, ordonnés de manière croissante.

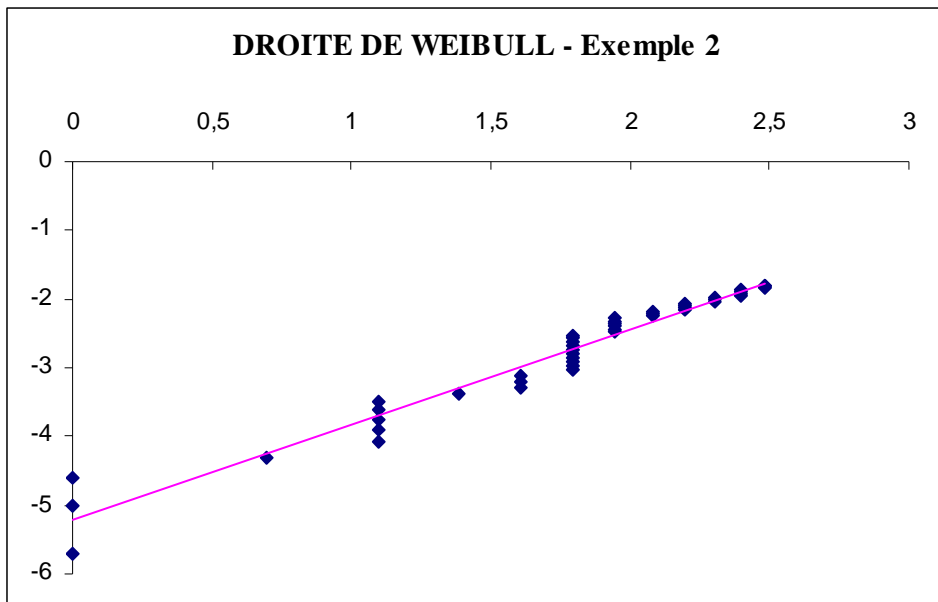
Les points de coordonnées $\left(\ln(t_i), \ln \left[\ln \left(\frac{1}{1-F(t_i)} \right) \right] \right)$ sont représentés sur le graphe de

Weibull (figure 41). La droite passant au plus près du nuage de points, obtenue par régression linéaire, est la droite de Weibull.

Le coefficient directeur et l'ordonnée à l'origine de la droite de Weibull nous permettent d'évaluer les paramètres β et η de la loi. Nous obtenons $\beta = 1,39$ et $\eta = 43$.

La loi de Weibull semble être un modèle adéquat puisque les points sont approximativement alignés.

Figure 41 : Droite de Weibull



Mais les estimations ainsi obtenues sont forcément erronées : en effet, nous n'avons pas tenu compte des données censurées. Les éliminer entraîne forcément une perte d'information importante, puisqu'on ne tient pas compte des matériels qui ont justement une durée de vie plus longue.

La présence de données censurées impose un traitement statistique particulier. Les méthodes adéquates sont présentées dans les paragraphes qui suivent.

IV.2. Méthodes tenant compte des données censurées

- *Méthode du maximum de vraisemblance*

C'est une méthode paramétrique fréquemment utilisée pour inférer les paramètres de la distribution de probabilité d'un échantillon donné.

Son estimateur est convergent, exhaustif, asymptotiquement sans biais et efficace.

La vraisemblance d'un échantillon constitué de N observations indépendantes s'écrit

$$L(\theta) = \prod_{i=1}^N L_i(\theta, \text{data}_i),$$

où θ est le vecteur des paramètres à estimer, data_i la donnée correspondant à la $i^{\text{ème}}$ observation et $L_i(\theta, \text{data}_i)$ la vraisemblance de l'observation i .

Pour estimer θ à partir de l'échantillon observé, il suffit de chercher la valeur de θ qui maximise la vraisemblance.

La quantité $L_i(\theta, \text{data}_i)$ s'écrit différemment selon que l'observation i est exacte, censurée à gauche, censurée à droite ou censurée par un intervalle.

Supposons les N observations constituées de m instants de défaillance distincts et de r instants de censures distincts, chacun de ces instants pouvant donner lieu à des ex-æquo. Le tableau ci-dessous donne la contribution de chaque type de donnée à la fonction de vraisemblance.

Tableau 12 : Contributions à la vraisemblance selon le type de donnée

Données	Observation	Vraisemblance
n_i défaillances en t_i	$T = t_i$	$[f(t_i)]^{n_i}$
d_i censures dans l'intervalle $(t_{i-1}, t_i]$	$t_{i-1} < T \leq t_i$	$[F(t_i) - F(t_{i-1})]^{d_i}$
l_i censures à gauche en t_i	$T \leq t_i$	$[F(t_i)]^{l_i}$
r_i censures à droite en t_i	$T > t_i$	$[1 - F(t_i)]^{r_i}$

Le nombre de défaillances est $\sum_{i=1}^m n_i$ et le nombre de censures est $\sum_{i=1}^r d_i + l_i + r_i$.

Ainsi, dans le cas d'un échantillon de N données de survie non censurées, la fonction de vraisemblance est

$$L(\theta) = \prod_{i=1}^m f(t_i, \theta)^{n_i} ;$$

Dans ce cas, le nombre N d'instant d'observation vérifie $\sum_{i=1}^m n_i = N$.

Dans le cas où il y a des observations censurées à droite, la vraisemblance s'écrit

$$L(\theta) = \prod_{i=1}^m f(t_i, \theta)^{n_i} \prod_{j=1}^r (1 - F(t_j^*, \theta))^{r_j},$$

où les t_i sont les temps de défaillance et les t_j^* les temps de censure.

Dans ce cas, le nombre N d'instant d'observation vérifie $\sum_{i=1}^m n_i + \sum_{i=1}^r r_i = N$.

Remarque : $f(t_i, \theta)$ est la densité de probabilité évaluée à l'instant t_i et $1 - F(t_i^*, \theta)$ représente la probabilité de survie au-delà du temps t_i^* . Chercher l'estimateur du maximum de vraisemblance revient à chercher la valeur du paramètre θ qui rend l'échantillon le plus « probable » possible.

Comme maximiser $L(\theta)$ est équivalent à maximiser $\ln[L(\theta)]$ nous posons en pratique $l(\theta) = \ln[L(\theta)]$ et nous cherchons à maximiser $l(\theta)$.

Cas où l'échantillon suit une distribution de Weibull

Nous nous plaçons dans le cas de données censurées à droite, comme dans l'exemple 2, et nous supposons que l'échantillon suit une loi de Weibull à deux paramètres β et η que nous cherchons à déterminer.

Nous avons alors :

$$l(\beta, \eta) = \sum_{i=1}^m n_i \ln \left[\frac{\beta}{\eta} \left(\frac{t_i}{\eta} \right)^{\beta-1} \exp \left(- \left(\frac{t_i}{\eta} \right)^\beta \right) \right] + \sum_{j=1}^r r_j \ln \left[\exp \left(- \left(\frac{t_j^*}{\eta} \right)^\beta \right) \right] \quad (1).$$

Soit, en posant $n = \sum_{i=1}^m n_i$:

$$l(\beta, \eta) = n \ln(\beta) - n \beta \ln(\eta) + (\beta - 1) \sum_{i=1}^m n_i \ln(t_i) - \frac{1}{\eta^\beta} \left[\sum_{i=1}^m n_i t_i^\beta + \sum_{j=1}^r r_j (t_j^*)^\beta \right].$$

Pour la concision de l'écriture, posons $\sigma = \frac{1}{\beta}$ et $\mu = \ln(\eta)$, $x_i = \ln(t_i)$ et $x_i^* = \ln(t_i^*)$.

$$\text{On obtient } l(\mu, \sigma) = -n \ln \sigma - \frac{n\mu}{\sigma} + \frac{1-\sigma}{\sigma} \sum_{i=1}^m n_i x_i - \sum_{i=1}^m n_i \exp \left(\frac{x_i - \mu}{\sigma} \right) - \sum_{j=1}^r r_j \exp \left(\frac{x_j^* - \mu}{\sigma} \right)$$

Les dérivées partielles sont :

$$\begin{cases} \frac{\partial l}{\partial \mu} l(\mu, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma} \sum_{k=1}^m n_k \exp \left(\frac{x_k - \mu}{\sigma} \right) + \frac{1}{\sigma} \sum_{j=1}^r r_j \exp \left(\frac{x_j^* - \mu}{\sigma} \right), \\ \frac{\partial l}{\partial \sigma} l(\mu, \sigma) = -\frac{n}{\sigma} + \frac{n\mu}{\sigma^2} - \frac{1}{\sigma^2} \sum_{i=1}^m n_i x_i + \frac{1}{\sigma^2} \sum_{i=1}^m n_i (x_i - \mu) \exp \left(\frac{x_i - \mu}{\sigma} \right) + \frac{1}{\sigma^2} \sum_{j=1}^r r_j (x_j^* - \mu) \exp \left(\frac{x_j^* - \mu}{\sigma} \right) \end{cases}$$

Les valeurs de μ et σ qui maximisent la fonction $l(\mu, \sigma)$ sont les valeurs qui annulent le système formé des deux dérivées partielles.

On en déduit que les estimateurs $\hat{\mu}$ et $\hat{\sigma}$ vérifient les deux équations suivantes:

$$\begin{cases} \hat{\mu} = \hat{\sigma} \ln \left(\frac{1}{n} \sum_{i=1}^m n_i \exp \left(\frac{x_i}{\hat{\sigma}} \right) + \frac{1}{n} \sum_{j=1}^r r_j \exp \left(\frac{x_j^*}{\hat{\sigma}} \right) \right), \\ \frac{1}{n} \sum_{i=1}^m n_i x_i + \hat{\sigma} - \frac{\sum_{i=1}^m n_i x_i \exp \left(\frac{x_i}{\hat{\sigma}} \right) + \sum_{j=1}^r r_j x_j \exp \left(\frac{x_j^*}{\hat{\sigma}} \right)}{\sum_{i=1}^m n_i \exp \left(\frac{x_i}{\hat{\sigma}} \right) + \sum_{j=1}^r r_j \exp \left(\frac{x_j^*}{\hat{\sigma}} \right)} = 0. \end{cases}$$

Ces équations non linéaires sont difficiles à résoudre. En principe, on les résout par des algorithmes numériques tels que ceux du gradient, de Newton ou de Newton-Raphson.

Remarques :

➤ Excel dispose d'un solveur qui permet de trouver directement les valeurs de β et η qui maximisent le log de la vraisemblance, autrement dit (1). Ce solveur utilise la méthode du gradient ou celle de Newton.

➤ La loi exponentielle étant un cas particulier de la loi de Weibull, avec $\beta=1$ et $\frac{1}{\eta} = \lambda$, le

log de la vraisemblance (1) devient $l(\lambda) = n \ln(\lambda) - \lambda \left(\sum_{i=1}^m t_i + \sum_{j=1}^r r_j t_j^* \right) - \lambda \sum_{i=1}^N t_i$ et on

$$\text{obtient } \hat{\lambda} = \frac{n}{\sum_{i=1}^m n_i t_i + \sum_{j=1}^r r_j t_j^*}.$$

Application à l'exemple 2

La vraisemblance s'écrit

$$L(\beta, \eta) = \prod_{i=1}^{24} f(t_i, \beta, \eta)^{n_i} (1 - F(t_i^*, \beta, \eta))^{r_i},$$

où n_i , respectivement r_i , est le nombre de défaillances, respectivement de censures, à l'instant t_i .

Le tableau 13 récapitule les données, translatées pour un recalage en zéro.

Les paramètres β et η sont obtenus par maximisation de la vraisemblance à l'aide du solveur d'Excel : $\beta = 1,35$ et $\eta = 61$.

Remarque : dans le cas où l'on ne tenait pas compte des données censurées (méthode de la droite de Weibull), nous avons trouvé $\beta = 1,39$ et $\eta = 43$. L'estimation du paramètre d'échelle était donc très différente.

Tableau 13

t_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
n_i	3	1	5	1	4	9	6	3	5	2	4	3	3	7	4	6	3	1	1	3	0	0	1	1
r_i	0	0	0	0	0	0	0	0	0	0	0	0	41	29	35	18	11	37	32	21	42	34	16	25

La période d'observation s'étend sur 24 mois. Les premières censures arrivent après 12 mois.

▪ Méthode de Kaplan-Meier

Nous montrons dans ce paragraphe comment construire un estimateur non paramétrique de la fonction de fiabilité lorsque l'on est en présence de données censurées. L'estimateur de Kaplan-Meier est l'estimateur non paramétrique le plus direct de la fonction de survie.

Soit un échantillon de taille initiale N constitué de matériels qui ont été mis en service à des dates distinctes. Notons :

- n_j le nombre de matériels qui ont été défaillants sur le $j^{\text{ème}}$ intervalle de temps $(t_{j-1}, t_j]$;

- u_j le nombre de matériels « à risque » à la date t_j , c'est-à-dire qui n'ont été ni défectueux ni censurés avant la date t_j .

L'estimateur de Kaplan-Meier calcule la fiabilité à la date t_i :

$$R(t_i) = \prod_{j: t_j \leq t_i} \left(1 - \frac{n_j}{u_j} \right)$$

Remarque : on se contente souvent d'évaluer la fonction de fiabilité aux instants correspondant à des dates de défaillance. L'estimateur de Kaplan-Meier est donc constant entre ces dates.

Application à l'exemple 2

Tableau 14 : estimateur de Kaplan-Meier

Temps en mois	1	2	3	4	5	6	7	8	9	10	11	12
Nombre de défaillances n_j	3	1	5	1	4	9	6	3	5	2	4	3
Nombre de censures r_i	0	0	0	0	0	0	0	0	0	0	0	0
Nombre de matériels à risque u_i	417	414	413	408	407	403	394	388	385	380	378	374
Estimateur de Kaplan-Meier	0.993	0.990	0.978	0.976	0.967	0.945	0.930	0.923	0.911	0.906	0.897	0.890

13	14	15	16	17	18	19	20	21	22	23	24
3	7	4	6	3	1	1	3	0	0	1	1
41	29	35	18	11	37	32	21	42	34	16	25
371	327	291	252	228	214	176	143	119	77	43	26
0.882	0.864	0.852	0.831	0.820	0.817	0.812	0.795	0.795	0.795	0.777	0.747

Population initiale de N=417 individus

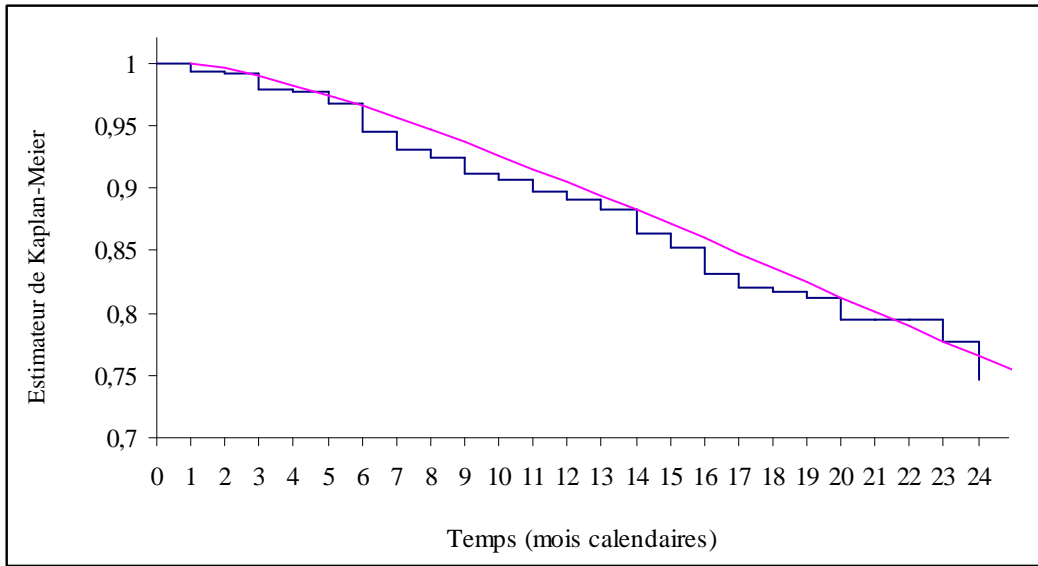
L'estimateur de Kaplan-Meier de la fiabilité est présenté sur la figure ci-dessous.

Nous avons par ailleurs estimé, par la méthode des moindres carrés, les paramètres de la loi de Weibull qui s'ajuste au mieux aux valeurs $R(t_i)$ données par la méthode de Kaplan-Meier. Autrement dit, nous avons estimé les valeurs de β et η qui minimisent la quantité

$$\sum_{i=1}^{24} \left[R(t_i) - \exp \left(- \left(\frac{t_i}{\eta} \right)^\beta \right) \right]^2.$$

Ces valeurs sont $\beta = 1,32$ et $\eta = 62$.

Figure 42 : Graphe de l'estimateur de Kaplan-Meier de la fiabilité R(t) pour les données de l'exemple 2, et loi de Weibull obtenue par moindres carrés.



▪ *Méthode de Johnson ou des rangs médians*

La méthode de Johnson permet de corriger l'ordre d'arrivée des défaillances en tenant compte de la probabilité pour qu'un individu censuré soit défaillant après la date à laquelle il a été censuré.

En pratique, on ajoute à cet ordre un « incrément » qui est fonction du nombre de censures entre deux défaillances.

Soit N le nombre initial d'individus dans l'échantillon observé, θ_i le rang corrigé de la $i^{\text{ème}}$ défaillance et s_j le nombre de systèmes qui, juste avant l'instant t_j de la $j^{\text{ème}}$ défaillance, n'ont été ni défaillants ni censurés. Alors l'incrément I_i de la $i^{\text{ème}}$ défaillance observée est donné par :

$$I_i = \frac{N+1-\theta_{i-1}}{1+s_i},$$

avec

$$\theta_0 = 0, \theta_i = \theta_{i-1} + I_i.$$

L'estimateur de Johnson de la fiabilité est évalué par :

$$R(t_i) = 1 - \frac{\theta_i - 0.3}{N + 0.4}.$$

Remarques :

- θ_i est le « rang médian » d'ordre i.
- $R(t_i)$ n'est évalué qu'aux temps où des défaillances ont été observées.

Nous expliquons ci-dessous pourquoi cette méthode fournit une estimation de la fiabilité.

Les défaillances sont ordonnées par âge (c'est-à-dire que t_i désigne un temps de défaillance) et nous supposons que les temps de censure coïncident avec les temps de défaillance.

Considérons un intervalle $[t_{i-1}, t_i[$ entre deux défaillances et un matériel censuré à t_{i-1} . S'il avait continué son service, ce matériel aurait pu être défaillant sur $[t_{i-1}, t_i[$, ou bien sur n'importe quel intervalle $[t_j, t_{j+1}[$ avec $i \leq j \leq n-1$, ou encore sur l'intervalle $[t_n, t[$ suivant la dernière défaillance.

Il y a donc $s_i + 1$ intervalles possibles et nous dirons qu'en moyenne, les « chances » de tomber en panne sur chacun de ces intervalles sont égales. En particulier, la probabilité de défaillance sur $[t_{i-1}, t_i[$ est de $\frac{1}{s_i + 1}$.

Les candidats à la défaillance sur $[t_{i-1}, t_i[$ sont :

- ✓ Les $N - s_{i-1} - \theta_{i-1} + 1$ matériels survivants à t_{i-1} ($N - s_{i-1}$ étant le nombre d'évènements, défaillances et censures indifféremment, avant t_{i-1} , et θ_{i-1} le rang corrigé de la $(i-1)^{\text{ième}}$ défaillance) ;
- ✓ Les $(N - s_i) - (N - s_{i-1}) - 1$ censures survenues à t_{i-1} .

Il y a donc $A_i + B_i = N - s_i - \theta_{i-1}$ candidats à la défaillance sur l'intervalle $[t_{i-1}, t_i[$, assortis d'une probabilité de $\frac{1}{s_i + 1}$. L'incrément au rang i est donc :

$$1 + \frac{N - s_i - \theta_{i-1}}{1 + s_i} = \frac{N + 1 - \theta_{i-1}}{1 + s_i}.$$

Les calculs détaillés pour l'estimateur de Johnson sont présentés dans le tableau 15.

Le tableau 16 et la figure 43 comparent les estimateurs de Kaplan-Meier et de Johnson. Nous pouvons constater que les deux méthodes donnent des estimations particulièrement proches.

Tableau 15 : Estimateur de Johnson de la fiabilité

<i>Ordre de la défaillance</i>	<i>Temps de défaillance t_i</i>	<i>Nombre de matériels à risque r_i</i>	<i>Incrément I_i</i>	<i>Ordre corrigé de la défaillance θ_i</i>	<i>Estimateur de Johnson de la fiabilité</i>
0	0	0	0	0	1
1	1	417	1,000	1,000	0,998
2	1	416	1,000	2,000	0,996
3	1	415	1,000	3,000	0,994
4	2	414	1,000	4,000	0,991
5	3	413	1,000	5,000	0,989
6	3	412	1,000	6,000	0,986
7	3	411	1,000	7,000	0,984
8	3	410	1,000	8,000	0,982
9	3	409	1,000	9,000	0,979
10	4	408	1,000	10,000	0,977
11	5	407	1,000	11,000	0,974
12	5	406	1,000	12,000	0,972
13	5	405	1,000	13,000	0,970
14	5	404	1,000	14,000	0,967
15	6	403	1,000	15,000	0,965
16	6	402	1,000	16,000	0,962
17	6	401	1,000	17,000	0,960
18	6	400	1,000	18,000	0,958
19	6	399	1,000	19,000	0,955
20	6	398	1,000	20,000	0,953
21	6	397	1,000	21,000	0,950
22	6	396	1,000	22,000	0,948
23	6	395	1,000	23,000	0,946
24	7	394	1,000	24,000	0,943
25	7	393	1,000	25,000	0,941
26	7	392	1,000	26,000	0,938
27	7	391	1,000	27,000	0,936
28	7	390	1,000	28,000	0,934
29	7	389	1,000	29,000	0,931
30	8	388	1,000	30,000	0,929
31	8	387	1,000	31,000	0,926
32	8	386	1,000	32,000	0,924
33	9	385	1,000	33,000	0,922
34	9	384	1,000	34,000	0,919
35	9	383	1,000	35,000	0,917
36	9	382	1,000	36,000	0,914
37	9	381	1,000	37,000	0,912
38	10	380	1,000	38,000	0,910
39	10	379	1,000	39,000	0,907
40	11	378	1,000	40,000	0,905
41	11	377	1,000	41,000	0,902
42	11	376	1,000	42,000	0,900
43	11	375	1,000	43,000	0,898
44	12	374	1,000	44,000	0,895
45	12	373	1,000	45,000	0,893
46	12	372	1,000	46,000	0,891
47	13	371	1,000	47,000	0,888
48	13	370	1,000	48,000	0,886
49	13	369	1,000	49,000	0,883
50	14	327	1,125	50,125	0,881
51	14	326	1,125	51,250	0,878
52	14	325	1,126	52,376	0,875
53	14	324	1,126	53,502	0,873
54	14	323	1,127	54,629	0,870
55	14	322	1,127	55,756	0,867
56	14	321	1,127	56,883	0,864
57	15	291	1,240	58,123	0,861
58	15	290	1,241	59,363	0,858
59	15	289	1,241	60,605	0,856
60	15	288	1,242	61,847	0,853
61	16	252	1,415	63,262	0,849
62	16	251	1,417	64,679	0,846
63	16	250	1,418	66,097	0,842
64	16	249	1,420	67,517	0,839
65	16	248	1,422	68,939	0,836
66	16	247	1,423	70,362	0,832
67	17	228	1,537	71,899	0,828
68	17	227	1,539	73,439	0,825
69	17	226	1,542	74,981	0,821
70	18	214	1,623	76,604	0,817
71	19	176	1,966	78,570	0,812
72	20	143	2,410	80,980	0,807
73	20	142	2,420	83,399	0,801
74	20	141	2,430	85,829	0,795
75	23	43	7,818	93,647	0,776
76	24	26	12,704	106,351	0,746

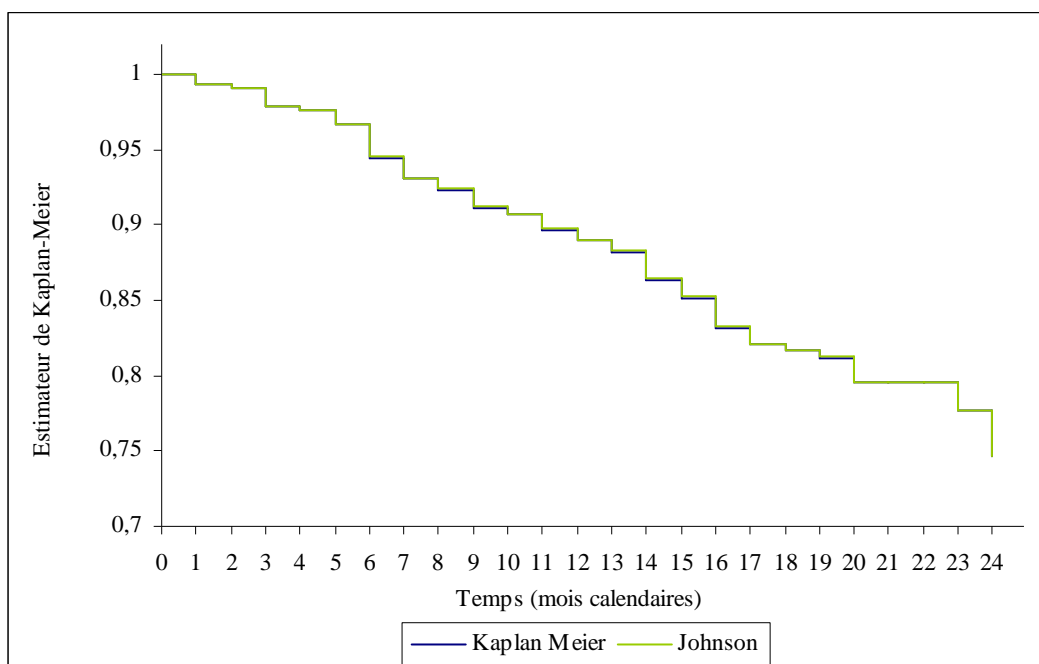
Tableau 16 : Comparaisons des estimateurs de Kaplan-Meier et de Johnson

Temps en mois	1	2	3	4	5	6	7	8	9	10	11	12
Estimateur de Kaplan-Meier	0.993	0.990	0.978	0.976	0.967	0.945	0.930	0.923	0.911	0.906	0.897	0.890
Estimateur de Johnson	0.993	0.991	0.979	0.978	0.967	0.946	0.931	0.924	0.912	0.907	0.898	0.890

	13	14	15	16	17	18	19	20	21	22	23	24
	0.882	0.864	0.852	0.831	0.820	0.817	0.812	0.795	0.795	0.795	0.777	0.747
	0.883	0.864	0.852	0.832	0.821	0.817	0.812	0.795	0.795	0.795	0.776	0.747

Population initiale de N=417 individus

Figure 43 : Estimateur de Kaplan-Meier vs estimateur de Johnson



V. Les limites des méthodes statistiques

Dans les deux exemples que nous avons traités ci-dessus, toutes les données étaient parfaitement connues, que ce soient les temps de défaillance ou les temps de censure. Malheureusement ce n'est pas toujours le cas. Comme nous l'avons précisé en introduisant cette partie, les données manquantes, erronées ou partielles constituent un réel problème pour les statisticiens. Il est même d'usage de dire que la collecte, le tri et la vérification des données prennent bien plus de temps que leur traitement statistique à proprement parler.

Dans notre cas, la situation est encore plus délicate ; en effet, nous ne disposons pas toujours des temps de défaillance, et encore moins des temps de censure.

La capture d'écran ci-dessous montre le type de données à disposition après élimination d'un certain nombre de renseignements superflus.

Figure 44 : Base de données des pannes

* * * * *

Avis	Zone de tri	Date de début de panne	SN IN	SN OUT	TSN	TSO	Raison de dépose	Actions correctives	Textes internes	Aircraft	Aircraft SN	Poste de travail	Date début garantie équipement	Avionnage
3067253	RBUS FR	15/04/2005	C19298002710	C19298002710			DURING AIRCRAF	INSPECTION CHE	MO 12H 600E	A319	2452	TOU-00	23/06/2005	A
3067256	RBUS FR	22/04/2005	C19366000215	C19366000215			APPLICATION DU	ICM board S/N E24		A340	0360	TOU-00	0	D
3067260	RBUS FR	22/04/2005	C19366000253	C19366000253			APPLICATION DU	SB C19366A-31-006		A340	0360	CLI-DS	0	D
3067263	RBUS FR	22/04/2005	C19366000222	C19366000222			APPLICATION DU	PSM board modifi		A340	0360	TOU-00	0	D
3067267	RBUS FR	22/04/2005	C19298000230	C19298000230			APPLICATION DU	PSM board modifi		A340	0360	CLI-DS	0	D
3067269	RBUS FR	22/04/2005	C19366000212	C19366000212			APPLICATION DU	SB C19366A-31-006		A340	0360	TOU-00	0	D
3067940	LHT	15/04/2005	C19298001320	C19298001320	2794	2794	FREE OF CHARGE	replace three SRAM		A319	2222	CLI-DS	13/05/2004	
3068021	AGASIN A	23/04/2005	C19298002025	C19298002025	171		SVC LOAN S/N C18	modify PPM board		A330	0632	CLI-DS	24/03/2005	
3068291	AFR	20/06/2005	C19298000380	C19298000380			INVALIDE DATA U	standard exchange		A320	1961	CLI-DS	03/04/2003	
3068688	LHT	30/04/2005	C19366000246	C19366000246	15280	15280	FREE OF CHARGE	standard exchange		A340	0540	CLI-DS	0	
3068765	LHT	03/05/2005	C19366000380	C19366000380	4664	4664	CPT PFD SHOWS	Repair : SB C19366	panne nkinit lors de	A340	0543	CLI-DS	06/02/2004	
3068923	RBUS FR	28/04/2005	C19298002569	C19298002569			INVALID DISPLAY	INSPECTION CHE	mo 30h 1500e	A330	0665	TOU-00	25/08/2005	A
3068924	RBUS FR	28/04/2005	C19298002534	C19298002534			FD : SUR FD /O LI	INSPECTION TES	MO:60h, soit 3000 h	A330	0663	TOU-00	05/10/2005	A
3068929	RBUS FR	28/04/2005	C19366000606	C19366000606			FD : PAR 2 FOIS LI	INSPECTION TES	MO:30h, soit 1500 E	A340	0630	TOU-00	25/07/2005	A
3069004	RTECHNIC	21/04/2005	C19298000676	C19298000676	4846		UNSCHEMULED 21	replace C187, 158, 15		A319	2037	CLI-DS	25/09/2003	
3069123	RTECHNIC	01/06/2005	C19298000627	C19298000627	7511		MAINT REF. SEQ #	standard exchange		A340	0556	CLI-DS	08/11/2003	
3069127	RTECHNIC	31/05/2005	C19298000611	C19298000611	7557		IN CRUISE INVALID	no fault found after		A340	0556	CLI-DS	08/11/2003	
3069132	RTECHNIC	01/06/2005	C19298001428	C19298001428	2850		REF ADDI-49, DU F	standard exchange		A319	2245	CLI-DS	11/06/2004	
3069226	RBUS FR	11/05/2005	C19366000621	C19366000621			LCDU NE DEMARRI	INSPECTION CHE	MO 10H 500E	A340	0681	TOU-00	09/11/2005	A
3069260	SAA	25/04/2005	C19366000440	C19366000440	6749	6749	PLEASE REPAIR I	replace 3 SRAM UI		A340	0534	CLI-DS	18/11/2003	
3069321	AGASIN A	18/05/2005	C19298001232	C19298001232			RESTITUTION PAR	no fault found, mod		A320	NC	CLI-DS	0	
3069548	RBUS FR	03/05/2005	C19298001743	C19298001743			THE DIGITAL CLOI	INSPECTION CHE	MO 60H 3000E	A319	2483	TOU-00	19/09/2005	A
3069692	RBUS FR	13/05/2005	C19298002718	C19298002718			FD : FOURD FAILE	INSPECTION CHE	MO 8H 400E	A320	2423	TOU-00	08/11/2005	A
3069791	AGASIN A	09/06/2005	C19298001235	C19298001235			RESTITUTION PAR	no fault found, mod		A320	NC	CLI-DS	0	
3069908	IBE	08/05/2005	C19366000533	C19366000533	2319	2319	INOPERATIVE	Replace UI8,UI2 ar		A340	0606	CLI-DS	24/11/2004	
3069977	RBUS FR	11/05/2005	C19298002914	C19298002914			PFDI DU SHOWS II	INSPECTION CHE	MO 20H 1000E	A319	2467	TOU-00	04/08/2005	A
3070340	AFR	23/05/2005	C19298000457	C19298000457	8106	8106	BLANKING DISPL	modify BLM board		A321	1972	CLI-DS	15/05/2003	
3070355	SAA	06/05/2005	C19298002528	C19298002528	366	366	SHOWS INVALID U	Standard exchange		747	28030	CLI-DS	10/03/2005	
3070509	UAE	14/05/2005	C19298001672	C19298001672	2471	2471	FLASHING INVALL	replace 3 SRAM UI		A340	0608	CLI-DS	30/11/2004	
3070660	RBUS FR	12/05/2005	C19298002806	C19298002806			AIRBUS STATEME	TEST MD FAULT F	MO:20H, SOIT 1000	A319	2467	TOU-00	22/09/2005	A
3071182	IBE	17/05/2005	C19366000520	C19366000520	2536	2536	INOPERATIVE	replace 3 SRAM UI		A340	0604	CLI-DS	09/11/2004	
3072153	RTECHNIC	20/05/2005	C19298001229	C19298001229	3731		FAILED (VAS OFIK	replace 3 SRAM UI		A340	NC	CLI-DS	01/04/2004	
3072212	AFR	13/06/2005	C19298000837	C19298000837			INOP(VOIR PFR J)	replace 3 SRAM int		A318	2071	CLI-DS	21/11/2003	
3072387	SAA	15/05/2005	C19298000671	C19298000671	50	50	UNIT U/S ERROR R	standard exchange		A320	NC	CLI-DS	03/07/2003	
3072630	RBUS FR	06/06/2005	C19366000670	C19366000670			FD : LEFUSURE D	INSPECTION TES	MO 4H 1000E	A320	0670	TOU-00	25/07/2005	A

Zone de tri : Compagnie aérienne utilisatrice, ou avionneur.

Date de début de panne : Date à laquelle la panne a été reportée.

S/N IN : Numéro de série entré en réparation.

S/N OUT : Numéro de série en sortie si différent.

TSN : Time Since New, temps de fonctionnement en heures de vol depuis la mise en service.

TSO : Time Since Overhaul, temps de fonctionnement en heures de vol depuis la défaillance précédente.

Date début garantie équipement : Date à laquelle prend effet la garantie pour la compagnie aérienne.

Avionnage : Panne constatée durant la période d'avionnage, ou non.

Les difficultés que nous rencontrons sont les suivantes :

- Idéalement, le temps de fonctionnement d'un équipement aéronautique devrait être mesuré en heures de vol, le temps « calendaire » ayant beaucoup moins de sens. En effet, un équipement n'est pas toujours sous tension. De plus, l'utilisation peut fortement varier d'une compagnie aérienne à une autre donc le temps calendaire perd de son sens : par exemple, une année de fonctionnement peut correspondre à 1000 heures de vol pour une compagnie à faible activité et à 5000 heures de vol pour une compagnie à forte activité.

Malheureusement les heures de vol sont rarement renseignées, cela pour deux raisons. D'une part, parce que le nombre d'heures de vol est inscrit dans la mémoire de l'équipement ; sa récupération nécessite une manipulation logicielle qui ne fait pas (encore) partie du processus obligatoire ; ce nombre est par ailleurs systématiquement remis à zéro lors de la recette. D'autre part, parce que les garanties sont exprimées en années

- Le temps de fonctionnement, en unité calendaire, est obtenu en faisant la différence entre la date à laquelle le matériel a été déposé et sa date de sa mise en service :

Temps de fonctionnement = Date de début de panne - Date début garantie équipement,

or il arrive fréquemment que les dates de début de garantie des équipements ne soient pas renseignées ou soient erronées. Lorsque ces dates ne sont pas renseignées, il est toujours possible de récupérer les dates de livraison à l'avionneur, mais il y a alors une incertitude sur la durée d'avionnage avant livraison à la compagnie aérienne.

- La date de début de garantie ne coïncide pas toujours exactement avec la mise en service.

Pour ces raisons, l'utilisation de méthodes statistiques, telles que celles que nous avons données en exemple plus haut, est laborieuse.

Ces difficultés sont à l'origine de la méthode que nous développons dans le chapitre suivant. Cette méthode, nous le verrons en détail, permet de s'affranchir des données de durées de vie et de temps de censure.

Chapitre 4

Traitement des retours opérationnels :
la « méthode convolutive »

I. Problématique générale

Dans le contexte concurrentiel international qu'ils rencontrent aujourd'hui, les équipementiers industriels doivent faire de gros efforts pour maintenir une offre compétitive. La fiabilité est devenue un domaine crucial et deux enjeux majeurs émergent :

- Respecter les clauses contractuelles de fiabilité avec leurs clients,
- Etre en mesure de répondre aux exigences de fiabilité présentes dans les appels d'offres qui sont portés à leur connaissance.

Ces exigences concernent essentiellement deux des grandeurs fiabilistes que nous avons présentées dans le chapitre précédent et qui sont étroitement liées : le taux de défaillance et le MTBF.

Comme nous l'avons vu auparavant, la fiabilité est estimée en phase de développement, bien souvent sur la base de formules issues de recueils normatifs. D'une part, ces recueils sont aujourd'hui obsolètes. D'autre part, la plupart des standards supposent que le taux de défaillance est constant, hypothèse qui n'est pas toujours vérifiée, et cela pour deux raisons : bien que le déverminage élimine la plupart des défauts de jeunesse, les matériels électroniques peuvent présenter un taux de défaillance légèrement décroissant du à la mortalité infantile ; des phénomènes d'usure peuvent aussi être observés sur certains sous-ensembles complexes même si cela est plus rare, les matériels étant en principe retirés de l'exploitation avant que l'usure ne se manifeste.

Les calculs de fiabilité prévisionnelle sont donc erronés. Ce biais nuit non seulement à l'équipementier qui se positionne comme plus mauvais qu'il n'est en réalité, mais aussi au client qui n'est pas à même de prévoir la politique de maintenance optimale ou de dimensionner correctement ses stocks.

Il est donc nécessaire de vérifier et, le cas échéant, de corriger ces valeurs prévisionnelles. Pour ce faire, une solution consiste à suivre les équipements en exploitation afin d'estimer leur fiabilité opérationnelle c'est-à-dire la fiabilité réellement observée.

Ce retour d'expérience aura un double impact :

- Il permettra de recalibrer les estimations prévisionnelles,
- Il améliorera la connaissance des performances réelles des matériels et permettra ainsi à l'équipementier de s'améliorer et de se positionner favorablement sur les marchés futurs.

Malheureusement, nous rencontrons là plusieurs obstacles... Tout d'abord, les résultats d'exploitation sont bien souvent « globaux », en ce sens qu'ils peuvent être affectés aux compagnies aériennes qui ont bien voulu les communiquer, mais en aucun cas à tel ou tel équipement individuellement identifiable. La traçabilité des données n'est donc pas, à notre connaissance, vérifiable. Ensuite, la date de livraison d'un équipement ne coïncide pas systématiquement avec sa date de mise en service puisque la période d'exploitation en compagnie est précédée d'une période d'avionnage dont la durée varie selon le porteur. Enfin, à ces difficultés spécifiques s'ajoutent les problèmes plus classiques comme les données indisponibles, manquantes ou erronées.

Les méthodes statistiques classiques, qu'elles soient paramétriques, telle que la méthode du maximum de vraisemblance, ou non paramétriques, telle que celle de Kaplan-Meier, sont souvent mises en défaut : elles nécessitent la connaissance des temps à la défaillance et des temps de censure.

Nous présentons dans ce chapitre une méthodologie permettant d'estimer le taux de défaillance d'un équipement [Berthon et al., 2006a] [Berthon et al., 2006b]. Ce modèle

analytique s'utilise dans des conditions peu usuelles puisque les seules données d'entrée nécessaires concernent le nombre d'équipements livrés par unité de temps (année, mois, quinzaine, ..., selon l'unité choisie) et le nombre de défaillances constatées par unité de temps.

L'originalité du travail réside dans le fait de déterminer, à partir de ces deux seules données globales, l'expression du taux de défaillance propre à chacun des équipements, que nous considérons comme ayant le même profil de mission (c'est-à-dire exploités dans des conditions environnementales similaires) et la même loi de survie (population isonome).

Quant à son intérêt, il réside principalement dans la simplicité des données nécessaires : la méthode proposée constitue une alternative efficace aux méthodes statistiques traditionnelles qui posent de fréquents problèmes d'utilisation du fait de la méconnaissance des temps de survie ou de censure.

II. Modélisation des retours opérationnels

II.1. Contexte et hypothèses

Ce paragraphe précise le contexte nécessaire à l'application de la méthode proposée.

Nous considérons une population d'équipements sur lesquels nous pouvons faire les hypothèses suivantes :

- Ils sont identiques : ils ont été fabriqués selon un design et un process similaires.
- Ils ont le même profil de mission : ils sont exploités dans des conditions similaires, qu'il s'agisse des conditions environnementales (climat) ou des conditions d'utilisation (profil de vol, type de porteur,...).
- Ils ont été mis en service à des dates distinctes.
- Chacun d'eux fonctionne jusqu'à défaillance et est alors retiré de l'exploitation sans être réparé.
- La loi de survie est une loi de Weibull.

A une date t donnée, nous observons la population.

Les deux premiers points nous assurent que tous les équipements possèdent la même loi de survie. Le troisième point implique que la population étudiée est constituée d'individus d'âge différent, si bien que nous avons affaire à des censures multiples à droite.

II.2. Terminologie

Afin d'assurer une certaine autonomie à ce chapitre, nous rappelons brièvement la terminologie usuelle qui est décrite plus en détail dans le chapitre 2.

La variable aléatoire T représente le temps de bon fonctionnement, c'est-à-dire le temps écoulé avant la défaillance d'un équipement. C'est, par hypothèse, une variable aléatoire à densité.

La fonction de répartition associée à un nombre réel positif t la probabilité qu'un équipement donné, dont la durée de vie est décrite par la variable aléatoire T et qui a été mis en service à la date 0, soit défaillant à une date inférieure ou égale à t :

$$F(t) = P(T \leq t).$$

La fonction de survie ou fiabilité, notée R , associe à t la probabilité que l'équipement fonctionne encore correctement à l'âge t :

$$R(t) = 1 - F(t).$$

La densité de probabilité (pdf) est la fonction $t \in \mathbb{R} \rightarrow f(t) \in \mathbb{R}^+$ telle que :

$$F(t) = \int_0^t f(u) du,$$

Enfin, le taux de défaillance est relié à la probabilité que l'équipement soit défaillant dans un petit intervalle de temps suivant l'instant t , sachant qu'il a correctement fonctionné jusqu'à l'instant t :

$$\lambda(t) = \lim_{h \rightarrow 0} \frac{P(t < T \leq t + h | T > t)}{h} = \frac{f(t)}{1 - F(t)}.$$

La relation suivante permet de relier le taux de défaillance et la fiabilité :

$$\lambda(t) = -\frac{R'(t)}{R(t)} = -\frac{d}{dt} [\ln R(t)].$$

II.3. Notations

Nous considérons une collection finie d'équipements identiques e_1, e_2, \dots, e_n qui ont été mis en service respectivement aux dates $t_1 = t(e_1) \leq t_2 = t(e_2) \leq \dots \leq t_n = t(e_n)$.

Nous notons $Q(t)$ le nombre d'équipements mis en service avant la date t , de sorte que Q est la fonction en escalier continue à droite décrite par

$$Q(t) = \text{Card}\{j \in \{1, \dots, n\}; t_j \leq t\} = \sum_{j=1}^n 1_{\{t_j \leq t\}}.$$

Nous définissons de la même manière la fonction en escalier continue à droite $t \rightarrow N(t)$ qui représente le nombre d'équipements déposés (défaillants) avant la date t . Si nous notons x_i la durée de bon fonctionnement du $i^{\text{ème}}$ équipement, nous pouvons écrire

$$N(t) = \sum_{j=1}^n 1_{\{t_j + x_j \leq t\}}$$

Nous insistons sur le fait que bien souvent, les seules données dont l'équipementier peut disposer de manière aisée sont les fonctions Q et N , voire une approximation discrète de ces fonctions lorsque les données sont agrégées par périodes. D'autre part c'est la fiabilité R d'un équipement (ou de manière équivalente la fonction de répartition F des défaillances) qui intéresse ce même équipementier.

Figure 45 : Flux cumulé de livraison

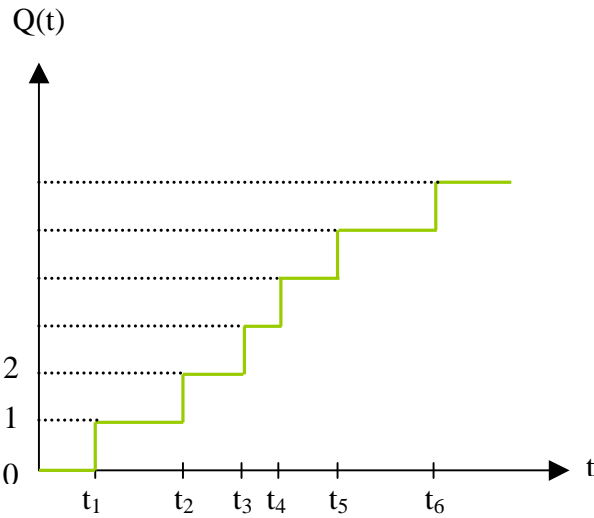


Figure 46 : Flux cumulé de défaillances

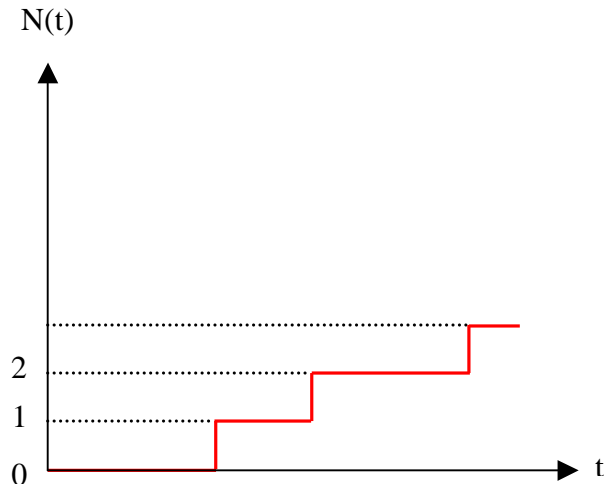
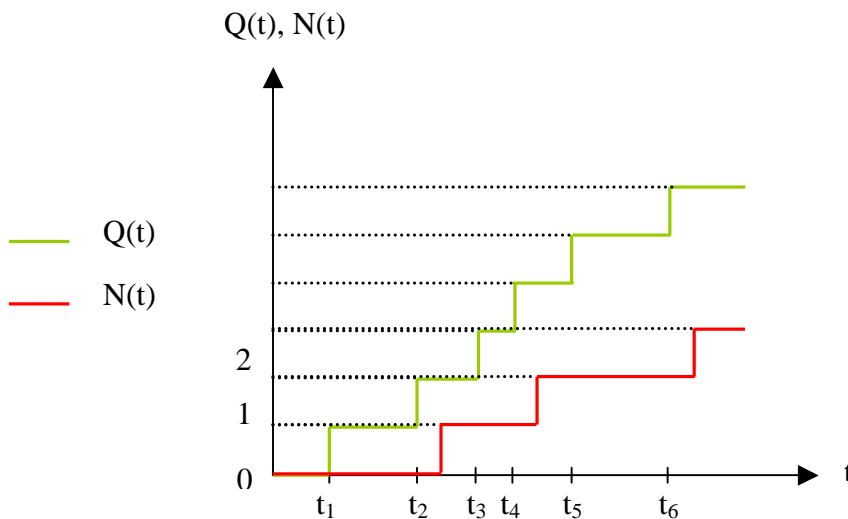


Figure 47 : Flux cumulés de livraison et de défaillances



Pour terminer, le comportement de la population peut être caractérisé par la fonction de survie W , qui associe à une date t la probabilité qu'aucun des équipements n'ait été défaillant avant cette date et ce, quelle que soit sa date de mise en service. Autrement dit :

$$W(t) = P(N(t) = 0).$$

Remarque : nous voyons immédiatement que la fonction W prend très rapidement de très petites valeurs. Ajoutons qu'elle ne peut être directement observée. Mais cette fonction a le mérite de caractériser la population d'équipements dans son ensemble. De plus, nous verrons dans le paragraphe qui suit qu'elle présente un second avantage : son logarithme $\ln W$ peut s'exprimer comme un produit de convolution impliquant les fonctions Q et R , Q étant connue et R étant la fonction qui nous intéresse.

II.4. Relation entre W, Q et R

Au vu des hypothèses que nous avons faites dans le paragraphe II.1. nous pouvons affirmer que tous les équipements ont la même fonction de fiabilité R. Si bien qu'en regardant l'ensemble des équipements constituant la population observée à l'instant t, nous exprimons la probabilité qu'il n'y ait aucune défaillance sur $[t_1, t]$:

$$W(t) = \prod_{j=1}^n R(t-t_j) = \prod_{j, t_j \leq t} R(t-t_j), \quad [1]$$

t_1 désignant la date à laquelle le premier équipement a été mis en service.

L'équation [1] mène trivialement à

$$\ln W(t) = \sum_{j, t_j \leq t} \ln(R(t-t_j)),$$

que nous pouvons réécrire [Charruau, 1990]

$$\boxed{-\ln W(t) = Q(t) * \lambda(t)}. \quad [2]$$

Preuve de l'équation [2] :

Introduisons l'instant initial $t_0=0$ vérifiant $Q(t_0)=0$ et l'instant final $t_{n+1}=t$ vérifiant $Q(t_{n+1})=Q(t_n)=Q(t)$.

Par définition de la fonction Q, nous avons $Q(t_{j+1})-Q(t_j)=1$ pour tout $1 \leq j \leq n-1$, et $Q(t_{n+1})-Q(t_n)=0$, si bien que nous pouvons réécrire :

$$\ln W(t) = \sum_{j=1}^{n+1} \left\{ \ln(R(t-t_j)) \times (Q(t_j) - Q(t_{j-1})) \right\},$$

ce qui nous permet de reconnaître dans la partie droite de l'équation l'intégrale de Stieltjes

$$\int_{t_1}^t \ln(R(t-u)) dQ(u),$$

ou encore

$$\int_0^t \ln(R(t-u)) dQ(u),$$

car la fonction Q est nulle entre 0 et t_1 .

Puisque $R(0) = 1$, $\ln(R(t-u))$ est nulle en t. De plus $Q(t_0)=0$. Une intégration par parties nous mène donc à

$$\ln W(t) = - \int_0^t Q(u) d(\ln R(t-u)). \quad [3]$$

Introduisant la quantité $\lambda(u) = -\frac{d}{du} \ln(R(u))$, soit $-\frac{d}{du} \ln(R(t-u)) = -\lambda(t-u)$. Nous avons :

$$\ln W(t) = -\int_0^t \lambda(t-u)Q(u)du. \quad [4]$$

Effectuons maintenant le changement de variable $v = t - u$:

$$\ln W(t) = +\int_t^0 \lambda(v)Q(t-v)dv = -\int_0^t \lambda(v)Q(t-v)dv. \quad [5]$$

Enfin, nous reconnaissons dans l'équation [5] le produit de convolution des fonctions Q et λ que nous réécrivons :

$$-\ln W(t) = Q(t) * \lambda(t),$$

où l'opérateur $*$ désigne le produit de convolution.

CQFD

Remarque : le produit de convolution est bien défini car Q et λ sont à support positif.

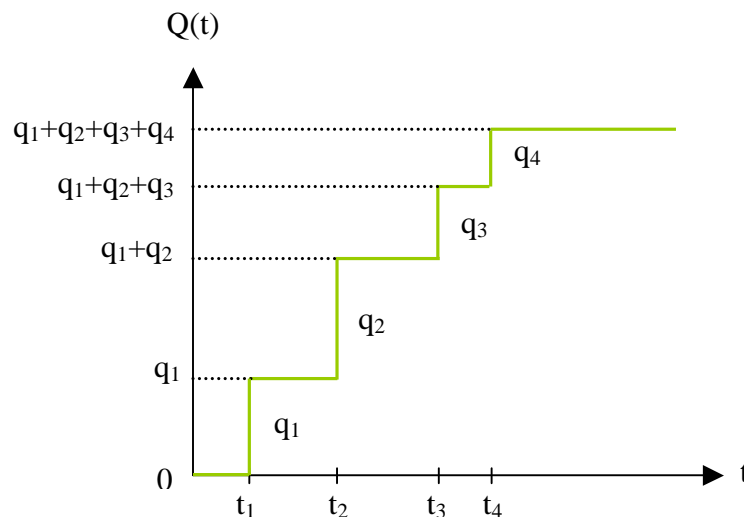
II.5. Un cas particulier : les données groupées

A l'instant t , nous observons une population d'équipements qui ont été mis en service de manière groupée : q_1 à la date t_1 , q_2 à la date t_2 , ..., q_n à la date t_n .

La fonction Q est constante et égale à $Q(t_i)$ sur tout l'intervalle $[t_i, t_{i+1}[$ pour $1 \leq i \leq n-1$. Nous pouvons donc reformuler l'équation [4] :

$$\ln W(t) = -\sum_{i=1}^{n-1} Q(t_i) \int_{t_i}^{t_{i+1}} \lambda(t-u)du. \quad [6]$$

Figure 48 : Données groupées



Si les intervalles $[t_i, t_{i+1}[$ sont suffisamment petits, la quantité $\lambda(t-u)$ est quasi constante et peut-être approchée par $\lambda(t-t_i)$: il suffit en effet d'écrire le développement de Taylor-Young d'ordre 0 en t_i : $\lambda(t-u) = \lambda(t-t_i) + g(u)$ où $\lim_{u \rightarrow t_i} g(u) = 0$.

Nous pouvons donc approcher $\ln W(t)$ par la somme ci-dessous :

$$\ln W(t) \approx - \sum_{i=1}^{n-1} Q(t_i) \lambda(t - t_i) (t_{i+1} - t_i). \quad [7]$$

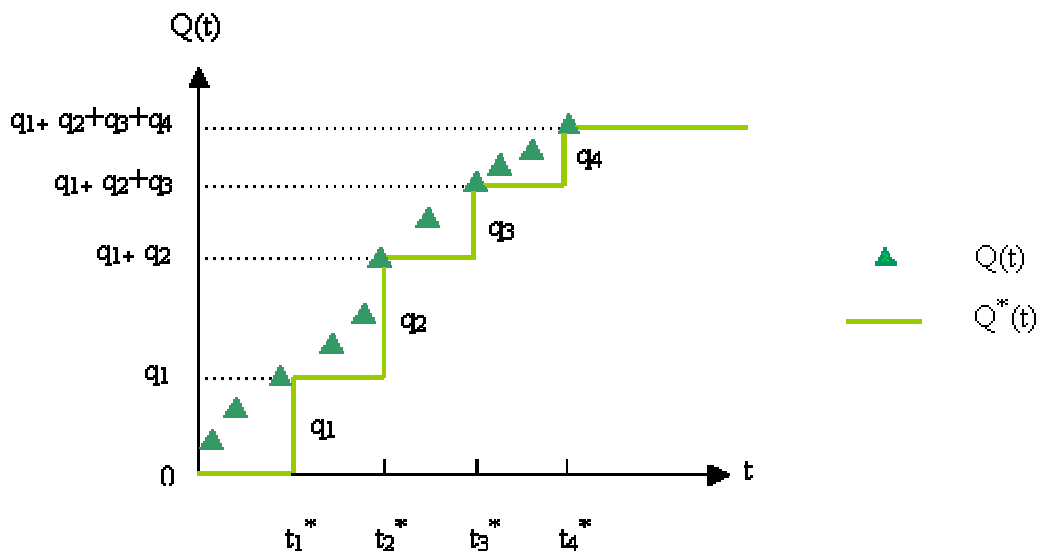
Remarque : l'expression ci-dessus n'est rien d'autre qu'un produit de convolution discret.

II.6. Cas où les données sont agrégées

En pratique, la situation est quelque peu différente. En effet, la date de mise en service d'un équipement n'est pas toujours connue de manière exacte. Les rapports qui sont à la disposition des fiabilistes ne donnent pas le nombre d'équipements qui *sont mis en service* aux temps t_1, t_2, \dots, t_n mais le nombre d'équipements qui *sont en service* à certains temps $t_1^*, t_2^*, \dots, t_m^*$.

Ainsi nous ne connaissons que les valeurs prises par la fonction Q aux dates $t_1^*, t_2^*, \dots, t_m^*$.

Figure 49 : données agrégées



Exemple : Les dates de livraison d'un matériel A sur l'année 2007 sont reportées dans le tableau 17 ci-dessous :

Tableau 17 : flux de livraison d'un matériel A sur l'année 2007

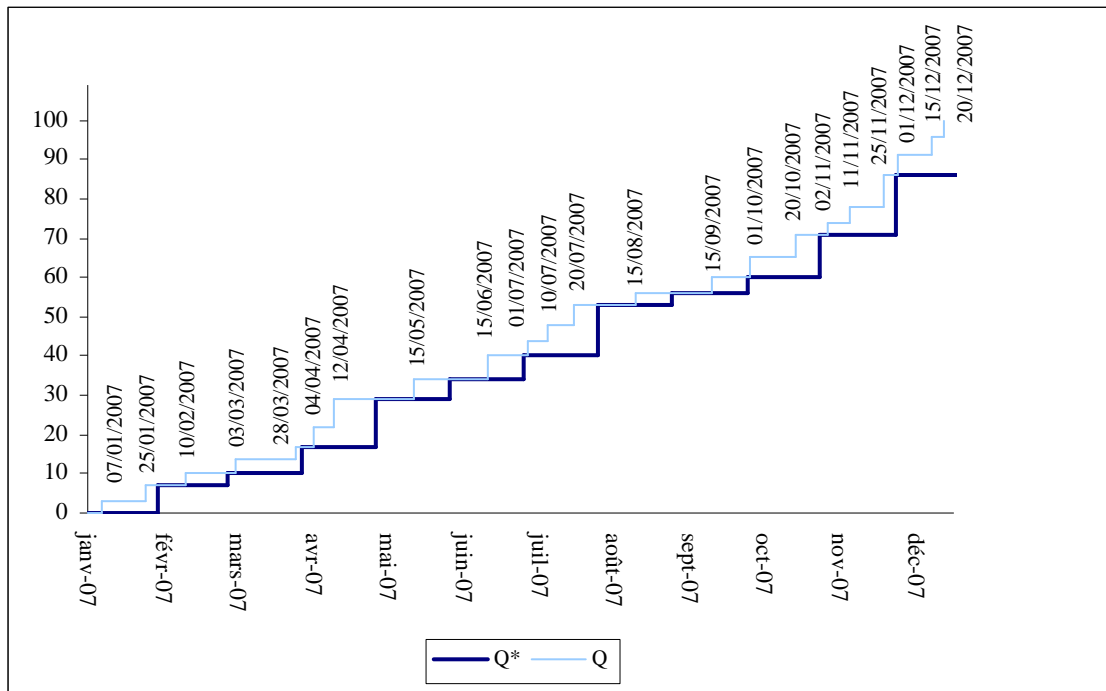
date	07/01/07	25/01/07	10/02/07	03/03/07	28/03/07	04/04/07	12/04/07	15/05/07	15/06/07	01/07/07
Livraisons	3	4	3	4	3	5	7	5	6	4
	10/07/07	20/07/07	15/08/07	15/09/07	01/10/07	20/10/07	02/11/07	11/11/07	25/11/07	01/12/07
	4	5	3	4	5	6	3	4	8	5
	15/12/07	20/12/07								
	5	4								

Le tableau 18 présente le flux de livraison de ce matériel agrégé au mois. La figure 50 montre le flux de livraison « exact » et le flux de livraison agrégé au mois.

Tableau 18 : Flux de livraison agrégé au mois

Mois	Quantité livrée
Janv-07	7
Févr-07	3
Mars-07	7
Avr-07	12
Mai-07	5
Juin-07	6
Juil-07	13
Août-07	3
Sept-07	4
Oct-07	11
Nov-07	15
Déc-07	14

Figure 50 : données agrégées pour l'exemple considéré



Remarque : dans cet exemple, les données sont agrégées au mois, mais elles peuvent aussi bien l'être au semestre, à l'année, ..., suivant le type de données considérées, la taille des populations étudiées, etc....

Nous pouvons dire de Q :

- Que c'est une fonction croissante ;
- Que $Q(t_{i+1}) - Q(t_i)$ est petit⁵ (en moyenne) devant $Q(t_i)$.

⁵ Le fait que $Q(t_{i+1}) - Q(t_i)$ est petit en moyenne devant $Q(t_i)$, soit $\frac{1}{n-1} \sum_{i=1}^{n-1} \frac{Q(t_{i+1}) - Q(t_i)}{Q(t_i)} \ll 1$, est

vérifiable, et vérifié en pratique, dans les cas de figure que nous avons rencontrés, car le flux cumulé de production croît rapidement.

Nous pouvons donc introduire la fonction en escalier continue à droite Q^* telle que :

- Q^* est continue en dehors des t_i^* ;
- Q^* coïncide avec Q aux points t_i^* .

En reprenant l'équation [4], nous approchons alors $\ln W(t)$ par

$$-\int_{t_1}^t Q^*(u)\lambda(t-u)du .$$

En d'autres termes, nous modélisons notre situation par le cas particulier présenté précédemment : sous l'hypothèse que $t_{i+1}^* - t_i^*$ est petit⁶, nous avons encore l'approximation :

$$\ln W(t) \approx -\sum_{i=1}^{n-1} Q^*(t_i^*)\lambda(t-t_i^*)(t_{i+1}^* - t_i^*) .$$

Or, par définition de Q^* , nous avons $Q^*(t_i^*) = Q(t_i^*)$, d'où

$$\ln W(t) \approx -\sum_{i=1}^{n-1} Q(t_i^*)\lambda(t-t_i^*)(t_{i+1}^* - t_i^*) . \quad [7]$$

Remarque : de nouveau, nous obtenons un produit de convolution discret.

III. Vérification de la formule de convolution sur deux exemples élémentaires

L'objet de la présente section est de valider l'équation de convolution, que nous rappelons ci-dessous, sur des exemples « cas d'école » :

$$-\ln W(t) = Q(t) * \lambda(t) .$$

Nous procédons de la manière suivante : la loi de survie est supposée connue, de sorte que nous savons expliciter le taux de défaillance $\lambda(t)$ mais aussi la fonction $W(t)$; nous nous donnons un « plan de production » $Q(t)$. A partir des données de $Q(t)$ et $W(t)$, nous récupérons le taux de défaillance $\lambda(t)$ puis nous vérifions qu'il est cohérent avec la loi de survie choisie.

Les deux cas traités ci-après à titre d'exemples impliquent les lois de survie exponentielle et de Weibull et des plans de production très simples.

III.1. Exemple 1 : loi de Weibull, flux de production constant

Dans cet exemple, nous supposons :

⁶ Le fait que $t_{i+1} - t_i$ est petit est cohérent avec le fait que $Q(t_{i+1}) - Q(t_i)$ est petit en moyenne devant $Q(t_i)$.

- o Que la loi de survie est une loi de Weibull. La fiabilité s'écrit donc

$$R(t) = \exp\left[-\left(\frac{t}{\eta}\right)^\beta\right] \text{ et le taux de défaillance est de la forme } \lambda(t) = \frac{\beta}{\eta^\beta} t^{\beta-1}.$$

- o Que sur chacune des quatre périodes $[i, i+1[$ constituant l'intervalle d'étude, un nombre identique q de matériels est mis en service.

La fonction $Q(t)$, quantité cumulée de mises en service, s'exprime donc au moyen de la fonction d'Heaviside :

$$Q(t) = q[\gamma(t) + \gamma(t-1) + \gamma(t-2) + \gamma(t-3)],$$

ce qui donne bien :

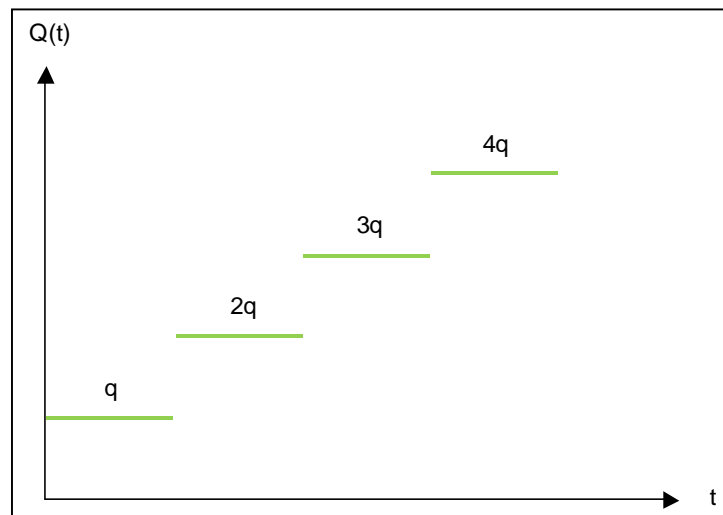
Première période $0 \leq t < 1$ et $Q(t) = q$,

Deuxième période $1 \leq t < 2$ et $Q(t) = 2q$,

Troisième période $2 \leq t < 3$ et $Q(t) = 3q$,

Quatrième période $3 \leq t < 4$ et $Q(t) = 4q$.

Figure 51 : Flux de production cumulé



Nous observons le comportement de la population à la date $t=4$.

Pour la transformée de Laplace du flux de production, il vient

$$LQ(t) = qL\gamma(t)(1 + e^{-s} + e^{-2s} + e^{-3s}),$$

soit

$$LQ(t) = \frac{q}{s}(1 + e^{-s} + e^{-2s} + e^{-3s}).$$

Puisque

$$\ln W(t) = \sum_{i=1}^4 q_i \ln R(t - t_i),$$

il vient pour la transformée de Laplace de la fonction $t \rightarrow \ln W(t)$:

$$L[\ln W(t)] = qL[\ln R(t)](1 + e^{-s} + e^{-2s} + e^{-3s}).$$

Or

$$L[\ln R(t)] = L\left(-\frac{t^\beta}{\eta^\beta}\right) = -\frac{\Gamma(\beta+1)}{\eta^\beta s^{\beta+1}}$$

Donc :

$$L[\ln W(t)] = -q \frac{\Gamma(\beta+1)}{\eta^\beta s^{\beta+1}} (1 + e^{-s} + e^{-2s} + e^{-3s}).$$

Finalement

$$L\lambda(t) = \frac{L[-\ln W(t)]}{L[Q(t)]} = \frac{\Gamma(\beta+1)}{\eta^\beta s^\beta} = \beta \frac{\Gamma(\beta)}{\eta^\beta s^\beta},$$

et nous récupérons bien par transformée inverse le taux de défaillance correspondant à une loi de Weibull :

$$\lambda(t) = \frac{\beta}{\eta^\beta} t^{\beta-1}.$$

III.2. Exemple 2 : loi exponentielle et production linéairement croissante

Dans cet exemple, nous supposons :

- Que la loi de survie est une loi exponentielle si bien que la fiabilité s'écrit :

$$R(t) = \exp[-\lambda t],$$

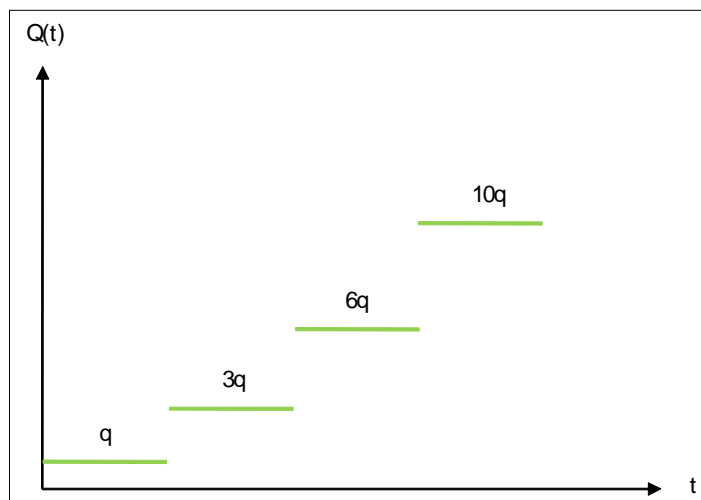
Et que le taux de défaillance est constant :

$$\lambda(t) = \lambda.$$

- Que la quantité d'équipements mise en service sur une période $[i, i+1[$ est supérieure de q unités à celle mise en service durant la précédente période : en d'autres termes $q_1=q$, $q_2=2q$, etc.

Nous observons le comportement de la population à l'issue des quatre premières périodes.

Figure 52 : Flux de production cumulé



Procédant de la même manière que dans l'exemple 1, nous déterminons successivement les transformées de Laplace de $Q(t)$, $\ln W(t)$ pour obtenir celle de $\lambda(t)$.

$Q(t)$ est la fonction en escalier :

$$Q(t) = q\gamma(t) + 2q\gamma(t-1) + 3q\gamma(t-2) + 4q\gamma(t-3).$$

Sa transformée de Laplace est

$$LQ(t) = \frac{q}{s}(1 + 2e^{-s} + 3e^{-2s} + 4e^{-3s}).$$

La transformée de Laplace de $\ln W(t)$ s'écrit

$$L[\ln W(t)] = qL[\ln R(t)](1 + 2e^{-s} + 3e^{-2s} + 4e^{-3s}).$$

Puisque

$$L[\ln R(t)] = L(-\lambda t) = -\frac{\lambda}{s^2},$$

il vient

$$L[\ln W(t)] = -\frac{q\lambda}{s^2}(1 + 2e^{-s} + 3e^{-2s} + 4e^{-3s}).$$

Finalement nous récupérons

$$L\lambda(t) = \frac{\lambda}{s},$$

de sorte que le taux de défaillance est bien constant égal à λ .

IV. Détermination du taux de défaillance

Nous avons montré dans les sections précédentes l'existence d'une relation de convolution entre les fonctions W , Q , et λ , relation que nous écrivons :

$$-\ln W(t) = Q(t) * \lambda(t)$$

équation [2] - cas continu,

et

$$\ln W(t) \approx -\sum_{i=1}^{n-1} Q(t_i)\lambda(t-t_i)(t_{i+1}-t_i)$$

équation [7] - cas discret.

Ces équations montrent qu'il est possible de déterminer le taux de défaillance $\lambda(t)$ si les fonctions W et Q sont connues.

Néanmoins, rappelons que $W(t)$, qui est la probabilité qu'aucun des équipements en service n'ait été défaillant avant la date t , nous est inconnue et ne peut être obtenue directement. C'est la raison pour laquelle nous avons essayé d'estimer $W(t)$ à l'aide de $N(t)$.

IV.1. Estimation de $W(t)$ à l'aide de $N(t)$

IV.1.a. Cas particulier

La fonction $-\ln W(t)$ peut être estimée à l'aide de $N(t)$ sous l'hypothèse ci-dessous :

$$\forall i \leq n, F(t - t_i) \ll 1, \text{ où } F \text{ est la défiabilité.}$$

Preuve :

Nous observons, à l'instant t , une famille d'équipements qui ont été mis en service aux dates t_1, t_2, \dots, t_n en quantités respectives q_1, q_2, \dots, q_n . Le nombre de défaillances observées avant l'instant t parmi les q_i équipements mis en service à la date t_i est noté D_i .

D_i est une variable aléatoire suivant une loi binomiale de paramètres q_i (nombre de mises en service à l'instant t_i) et $F(t - t_i)$ (probabilité de défaillance avant l'instant t pour un matériel mis en service à la date t_i).

Nous avons donc $P(D_i = k) = C_{q_i}^k F(t - t_i)^k R(t - t_i)^{q_i - k}$ et $ED_i = q_i F(t - t_i)$.

Il vient alors pour $N(t)$:

$$EN(t) = \sum_{i=1}^n ED_i = \sum_{i=1}^n q_i F(t - t_i).$$

En approchant $EN(t)$ par $N(t)$, il vient :

$$N(t) \approx \sum_{i=1}^n q_i F(t - t_i) = \sum_{i=1}^n q_i (1 - R(t - t_i)).$$

Considérons maintenant la fonction $t \rightarrow -\ln W(t)$; les équipements ayant été mis en service par quantités groupées, la fonction $W(t)$ s'écrit :

$$W(t) = \prod_{i=1}^n [1 - F(t - t_i)]^{q_i}.$$

Il vient

$$-\ln W(t) = -\sum_{i=1}^n q_i \ln[1 - F(t - t_i)].$$

Notons qu'en posant $q_i=1$ pour tout i , nous retrouvons l'équation [1].

Sous l'hypothèse que $F(t - t_i) \ll 1$, nous pouvons écrire :

$$\ln[1 - F(t - t_i)] \approx -F(t - t_i) = -(1 - R(t - t_i)).$$

En effet, le développement limité au voisinage de 0 et à l'ordre 1 de la fonction $u \rightarrow \ln(1 - u)$ s'écrit $\ln(1 - u) = -u + o(u)$.

Cela nous mène à :

$$-\ln W(t) \approx \sum_{i=1}^n q_i \times (1 - R(t - t_i)).$$

Finalemment :

$$-\ln W(t) \approx N(t). \quad [8]$$

CQFD

Remarque : l'hypothèse $F(t-t_i) \ll 1$ sous-entend que la probabilité de défaillance est très faible quel que soit l'âge de l'équipement. Autrement dit, cette hypothèse est valable si les matériels considérés sont d'une part « très fiables », et d'autre part s'ils sont retirés de l'exploitation avant la phase d'usure. Ces deux conditions garantiront l'inégalité $F(t-t_i) \ll 1$, y compris pour les grandes valeurs de t .

IV.1.b. Cas général

Dans la majorité des cas rencontrés, nous ne pouvons garantir que les hypothèses autorisant l'approximation ci-dessus sont vérifiées. Ainsi, dans le cas plus général où l'on ne peut s'assurer que $F(t-t_i) \ll 1$, il est nécessaire de trouver une autre relation entre les fonctions W et N .

Or une bonne approximation de la fonction $W(t)$ est

$$W(t) \approx \left(1 - \frac{N(t)}{Q(t)}\right)^{Q(t)}. \quad [9]$$

Preuve :

A l'instant t , nous observons une population d'équipements :

- De taille $M = Q(t)$;
- Parmi lesquels $m = N(t)$ présentent le caractère « défaillance ».

La probabilité qu'un équipement donné soit défaillant peut être approchée par :

$$\frac{m}{M} = \frac{N(t)}{Q(t)}.$$

La probabilité qu'un équipement donné ne soit pas défaillant peut être approchée par :

$$1 - \frac{m}{M} = 1 - \frac{N(t)}{Q(t)}.$$

La probabilité qu'il n'y ait aucune défaillance dans la population peut donc être approchée par :

$$\left[1 - \frac{m}{M}\right]^M = \left[1 - \frac{N(t)}{Q(t)}\right]^{Q(t)}.$$

Par conséquent, la fonction $t \rightarrow -\ln W(t)$ peut être approchée par :

$$-Q(t) \ln \left(1 - \frac{N(t)}{Q(t)}\right).$$

CQFD

Remarque : ces approximations nous ont posé quelques problèmes. En effet, supposons que l'on veuille estimer la probabilité pour un individu quelconque d'une population de présenter un caractère « A » ; la théorie de l'estimation ponctuelle nous assure que, si la population est homogène vis-à-vis du caractère étudié, une bonne estimation de cette probabilité est

$$p = \frac{n_A}{N},$$

où n_A est le nombre d'individus présentant le caractère dans un échantillon de taille N . Or la population que nous considérons, c'est-à-dire l'ensemble des matériels en service à l'instant t , ne semble pas homogène : les matériels ayant été mis en service à des dates distinctes, le « risque » de défaillance à l'instant t n'est pas le même pour chacun.

IV.2. Méthodologie de résolution

Résumons ce que nous avons démontré dans la section précédente :

Le taux de défaillance $\lambda(t)$ vérifie la relation :

$$X(t) \approx Q(t) * \lambda(t) \quad [10]$$

où :

- Q est le flux cumulé de production

$$- X(t) = \begin{cases} N(t) & \text{si } F(t - t_i) \ll 1 \text{ pour tout } i \\ -Q(t) \ln \left[1 - \frac{N(t)}{Q(t)} \right] & \text{sinon.} \end{cases}$$

Il s'agit donc de résoudre l'équation de convolution ci dessus. Dans les deux cas, nous faisons uniquement appel aux fonctions $Q(t)$ et $N(t)$ qui sont connues.

Pour résoudre l'équation, il est nécessaire de procéder à une « déconvolution » et plusieurs méthodes peuvent être envisagées [Commenges, 1980].

La manière la plus naturelle de résoudre l'équation [9] consiste à appliquer la transformée de Laplace [Demengel, 2002]. En effet, la transformée de Laplace transforme le produit de convolution en un produit direct. Il vient

$$LX(t) \approx LQ(t)L\lambda(t),$$

la fonction $LQ(t)$ étant connue, puisque $Q(t)$ est connue, et la fonction $LX(t)$ pouvant être approchée, puisque l'on connaît une approximation de la fonction $X(t)$.

Nous recouvrons ainsi la quantité $L\lambda(t)$. Puis par application de la transformée de Laplace inverse, nous récupérons enfin le taux de défaillance $\lambda(t)$:

$$\lambda(t) \approx L^{-1} \left(\frac{LX(t)}{LQ(t)} \right) \quad [11]$$

Rappel : Le signe \approx provient du fait que $X(t)$ est une approximation de la fonction $t \rightarrow -\ln W(t)$. L'équation exacte serait $\lambda(t) = L^{-1} \left(\frac{L[-\ln W(t)]}{LQ(t)} \right)$.

Toutefois, nous devons nous assurer que la transformée de Laplace des fonctions précitées est bien définie.

Remarque : la transformée de Laplace d'une fonction f définie sur $]0, +\infty[$ existe si :

- f est continue par morceaux ;
- Il existe $0 < \alpha < 1$ tel que $\lim_{t \rightarrow 0} t^\alpha |f(t)| = 0$;
- f est d'ordre exponentiel c'est-à-dire qu'il existe $t^* > 0, \beta > 0$ et $M > 0$ tels que $\forall t > t^*, |f(t)| < M \exp[\beta t]$.

Les fonctions Q et N sont continues par morceaux. Par conséquent le premier point ne pose pas de problème. Elles sont de plus bornées car positives et majorées par la valeur maximale prise par Q (c'est à dire le flux de production cumulé à la date d'observation).

Ainsi les deux dernières conditions sont aussi vérifiées et les transformées de Laplace de Q et N sont bien définies.

Quand à la transformée de Laplace de λ , on supposera qu'elle est définie, comme c'est le cas pour n'importe quel taux de défaillance usuel.

IV.3. Hypothèse et choix des fonctions

Les données qui nous sont fournies étant agrégées nous avons choisi, pour résoudre l'équation [10], d'approcher les paires $(t, Q(t))$ et $(t, X(t))$ par des fonctions :

- Qui assurent la meilleure approximation possible au sens des moindres carrés ;
- Dont les transformées de Laplace directes et inverses sont « tabulées ».

Nous verrons que des fonctions particulièrement adéquates sont les fonctions polynomiales.

Par ailleurs, nous avons du faire un certain nombre d'hypothèses sur :

- La nature de la loi de survie et par conséquent la forme du taux de défaillance ;
- L'allure du flux de production.

Ces hypothèses sont présentées dans le paragraphe qui suit.

Nous faisons les hypothèses suivantes :

Pour le flux de production cumulé $Q(t)$:

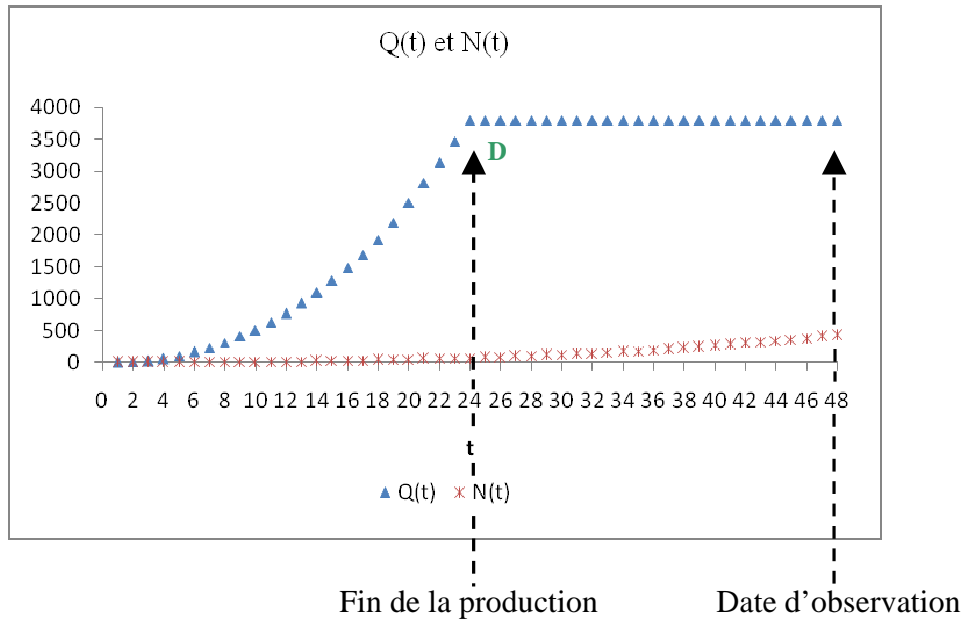
1. La production cesse après une période de longueur D ; $Q(t)$ est donc une fonction croissante sur l'intervalle $[0, D]$ puis constante sur $]D, +\infty[$;
2. La date d'observation peut être située en-deçà ou au-delà de D ;
3. Le flux de production cumulé est approché par une fonction polynomiale.

Pour le taux de défaillance $\lambda(t)$:

4. Nous supposons que la loi de survie est une loi de Weibull⁷ de paramètres β et η . Par conséquent, le taux de défaillance s'écrit $\lambda(t) = \frac{\beta}{\eta^\beta} t^{\beta-1}$.

⁷ Rappelons que la loi de Weibull est massivement utilisée en fiabilité puisqu'elle permet aussi bien de modéliser une « loi de jeunesse » (taux de défaillance décroissant correspondant à $\beta < 1$), une loi exponentielle (taux de défaillance constant correspondant à $\beta = 1$), et une « loi d'usure » (taux de défaillance croissant correspondant à $\beta > 1$).

Figure 53 : flux de production / flux de défaillances cumulés



Au vu des hypothèses faites sur la fonction \$Q(t)\$ et sur le taux de défaillance \$\lambda(t)\$, la fonction \$X(t)\$ peut s'écrire :

$$X(t) = \frac{\beta}{\eta^\beta} \Gamma(\beta) \sum_{i=0}^m \frac{P^{(i)}(0)t^{\beta+i}}{\Gamma(\beta+1+i)} - \frac{\beta}{\eta^\beta} \Gamma(\beta) \sum_{i=1}^m \frac{P^{(i)}(D)\gamma(t-D)(t-D)^{\beta+i}}{\Gamma(\beta+1+i)}.$$

Preuve :

➤ Première étape : flux de production

Les hypothèses 1 et 3 nous assurent que \$Q(t)\$ s'écrit

$$Q(t) = P(t)[\gamma(t) - \gamma(t-D)] + P(D)\gamma(t-D),$$

où \$P\$ est un polynôme de degré \$m\$:

$$P(t) = \sum_{i=0}^m a_i t^i,$$

\$\gamma\$ est la fonction d'Heaviside vérifiant :

$$\gamma(t) = \begin{cases} 1 & \text{si } t \geq 0 \\ 0 & \text{sinon} \end{cases},$$

ce qui donne :

$$\gamma(t-D) = \begin{cases} 1 & \text{si } t \geq D \\ 0 & \text{sinon} \end{cases} \quad \text{et} \quad \gamma(t) - \gamma(t-D) = \begin{cases} 1 & \text{si } t < D \\ 0 & \text{si } t \geq D \end{cases}.$$

Il vient pour la transformée de Laplace :

$$LQ(t) = L[P(t)\gamma(t)] - L[P(t)\gamma(t-D)] + L[P(D)\gamma(t-D)].$$

Le tableau ci-dessous regroupe les transformées de Laplace que nous allons utiliser :

Tableau 19 : Quelques transformées de Laplace utiles

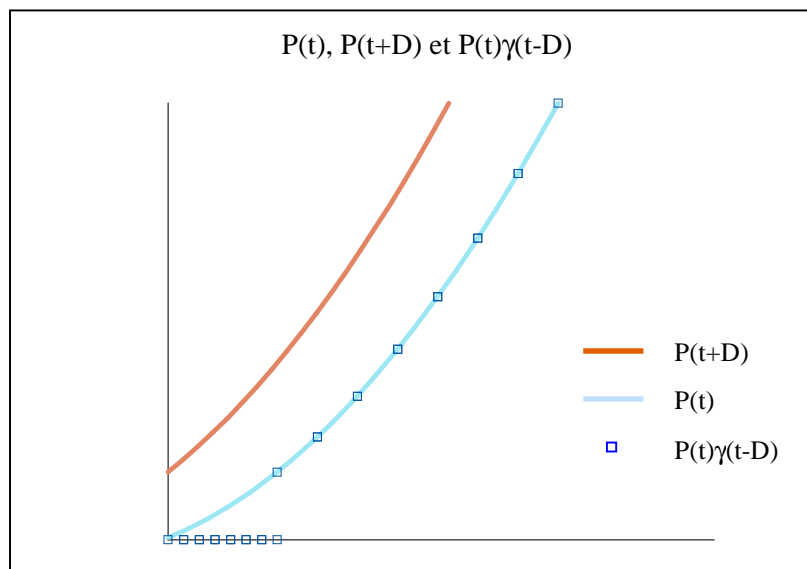
Fonction	$f(t)$	$\hat{f}(s)$
Heaviside (échelon unité)	$\gamma(t)$	$\frac{1}{s}$
Heaviside retardée (échelon unité retardé)	$\gamma(t - a), a > 0$	$\frac{e^{-as}}{s}$
Puissance entière	$\frac{t^n}{n!} \gamma(t), n \in \mathbb{N}$	$\frac{1}{s^{n+1}}$
Puissance non entière	$\frac{t^a}{\Gamma(a+1)} \gamma(t), a \in \mathbb{R}$	$\frac{1}{s^{a+1}}$
Translatée d'une fonction g	$g(t-a)$	$e^{-as} \hat{g}(s)$

Il vient immédiatement pour la transformée de Laplace d'un polynôme $P(t) = \sum_{i=0}^m a_i t^i$:

$$LP(t) = \sum_{i=0}^m i! a_i \frac{1}{s^{i+1}} . \quad [12]$$

Remarquons maintenant que la fonction $t \rightarrow P(t)\gamma(t-D)$ est une translatée à droite de la fonction $t \rightarrow P(t+D)\gamma(t)$, comme le montre la figure 54.

Figure 54 : $P(t)$, $P(t+D)$ et $P(t)\gamma(t-D)$



Il vient donc

$$L[P(t)\gamma(t-D)] = e^{-sD} L[P(t+D)\gamma(t)].$$

Ecrivons

$$P(t + D) = \sum_{i=0}^m a_i (t + D)^i = \sum_{i=0}^m a_i \sum_{j=0}^i C_i^j t^j D^{i-j} ;$$

En utilisant la propriété de linéarité de la transformée de Laplace, il vient :

$$L[P(t + D)] = \sum_{i=0}^m a_i \sum_{j=0}^i C_i^j D^{i-j} L[t^j] = \sum_{i=0}^m a_i \sum_{j=0}^i C_i^j D^{i-j} \frac{j!}{s^{j+1}} .$$

Or $C_i^j \times j! = A_i^j$, d'où :

$$L[P(t + D)] = \sum_{i=0}^m a_i \sum_{j=0}^i A_i^j D^{i-j} \frac{1}{s^{j+1}} ,$$

que l'on peut aussi bien écrire

$$L[P(t + D)] = \sum_{j=0}^m \frac{1}{s^{j+1}} \sum_{i=j}^m A_i^j a_i D^{i-j} ,$$

puisque les signes sommes sont interchangeables et que :

$$\begin{cases} 0 \leq i \leq m \\ 0 \leq j \leq i \end{cases} \Leftrightarrow \begin{cases} 0 \leq j \leq m \\ j \leq i \leq m \end{cases} .$$

Remarquons maintenant que pour tout polynôme $P(t) = \sum_{i=0}^m a_i t^i$:

- La dérivée $k^{\text{ième}}$, pour $k \leq m$, s'écrit $P^{(k)}(t) = \sum_{i=k}^m A_i^k a_i t^{i-k}$;
- La dérivée $k^{\text{ième}}$ en zéro s'écrit $P^{(k)}(0) = k! a_k$.

Par conséquent nous avons :

$$L[P(t + D)] = \sum_{j=0}^m P^{(j)}(D) \frac{1}{s^{j+1}} . \quad [13]$$

Pour finir, regardons la transformée de Laplace de la fonction $t \rightarrow P(D)\gamma(t - D)$. La formule de translation à droite nous dit qu'elle s'écrit

$$L[P(D)\gamma(t - D)] = P(D)e^{-sD}L[\gamma(t)] = P(D) \frac{e^{-sD}}{s} . \quad [14]$$

Finalement, en sommant [12], [13] et [14], nous obtenons :

$$LQ(t) = \sum_{i=0}^m P^{(i)}(0) \frac{1}{s^{i+1}} - \sum_{i=1}^m P^{(i)}(D) e^{-sD} \frac{1}{s^{i+1}} . \quad [15]$$

➤ Deuxième étape : taux de défaillance

L'hypothèse 4 nous assure que la transformée de Laplace du taux de défaillance s'écrit :

$$L(\lambda(t)) = \frac{\beta}{\eta^\beta} \Gamma(\beta) \frac{1}{s^\beta}. \quad [16]$$

➤ Troisième étape : transformée de Laplace de la fonction $X(t)$

La relation de convolution [10], transposée dans le domaine fréquentiel nous donne

$$LX(t) = LQ(t)L\lambda(t),$$

et nous assure que la transformée de Laplace de la fonction $X(t)$ s'écrit de la manière suivante :

$$L(X(t)) = \sum_{i=0}^m \frac{\frac{\beta}{\eta^\beta} \Gamma(\beta) P^{(i)}(0)}{s^{\beta+i+1}} - \sum_{i=1}^m \frac{\frac{\beta}{\eta^\beta} \Gamma(\beta) P^{(i)}(T) e^{-sD}}{s^{\beta+i+1}}. \quad [17]$$

➤ Quatrième étape : originale de la fonction $X(t)$

En utilisant les transformées présentées dans le tableau 19, nous pouvons obtenir l'expression de la fonction $X(t)$:

$$X(t) = \frac{\beta}{\eta^\beta} \Gamma(\beta) \sum_{i=0}^m \frac{P^{(i)}(0)}{\Gamma(\beta+i+1)} t^{\beta+i} - \frac{\beta}{\eta^\beta} \Gamma(\beta) \sum_{i=1}^m \frac{P^{(i)}(D)}{\Gamma(\beta+i+1)} \gamma(t-D)(t-D)^{\beta+i}, \quad [18]$$

ce qui marque la fin de la preuve.

CQFD

Rappelons que la fonction $X(t)$ est connue puisqu'elle s'écrit :

$$X(t) = \begin{cases} N(t) & \text{si } F(t-t_i) \ll 1 \text{ pour tout } i \\ -Q(t) \ln \left[1 - \frac{N(t)}{Q(t)} \right] & \text{sinon.} \end{cases}$$

Nous venons d'en donner une seconde écriture, faisant intervenir β et η , qui sont justement les paramètres de la loi de Weibull qui nous intéresse.

Dans la pratique, nous disposons des couples $(t_i, N(t_i))$ et $(t_i, Q(t_i))$, et donc des couples $(t_i, X(t_i))$, pour $1 \leq i \leq n$.

Ainsi, à ce stade de l'étude, nous formulons le problème de la manière suivante : nous devons trouver les valeurs de β et de η qui minimisent l'expression $\sum_{i=1}^n (X_{Q,N}(t_i) - X_{\beta,\eta}(t_i))^2$, où $X_{Q,N}$ est l'écriture de X en fonction de Q et N , et où $X_{\beta,\eta}$ est l'écriture de X en fonction de β et η .

V. Vérification de la méthode sur des jeux de données simulées

Pour des raisons de confidentialité, les résultats que nous avons obtenus sur des équipements aéronautiques actuellement en service ne peuvent être présentés. Néanmoins, nous avons appliqué la méthodologie proposée à des jeux de données simulées. De plus, et ce

afin de valider notre méthode, nous avons comparé les résultats qu'elle fournit à ceux obtenus par la méthode statistique standard de Kaplan-Meier. Ces résultats sont présentés dans les paragraphes qui suivent.

V.1. Simulation des données

Considérons l'exemple basique suivant :

- ❖ La production s'étale sur une période de $D=24$ mois. Chaque mois, nous connaissons la quantité d'équipements qui ont été mis en service.
- ❖ Sur cette période, le flux cumulé de production croît comme la fonction $t \rightarrow 5t^2$.

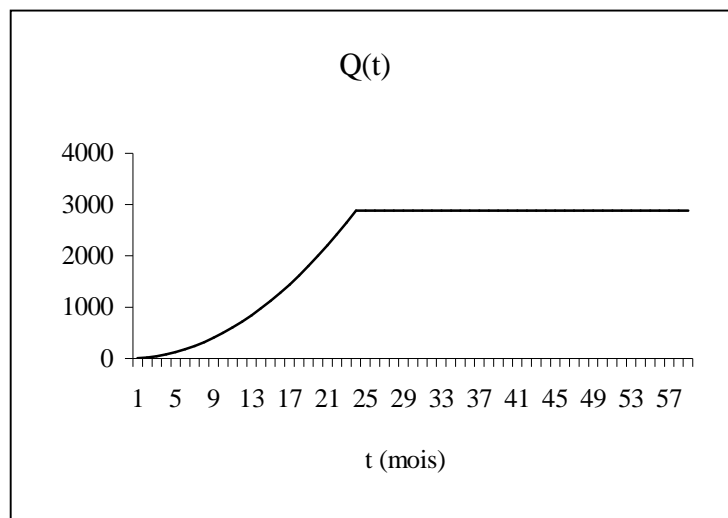
Le flux de production cumulé s'écrit alors:

$$\begin{cases} Q(t) = 5t^2 & \text{si } t \leq D \\ Q(t) = 5D^2 & \text{si } t > D \end{cases}$$

Notons que cela correspond à une croissance linéaire du nombre d'équipements mis en service.

Ce flux est représenté sur la figure 55.

Figure 55 : Flux de production cumulé



- ❖ La loi de survie est une loi de Weibull.
- ❖ La population est observée sur une durée de 59 mois. L'observation démarre dès le premier mois de production et se termine à la fin du 59^{ième} mois. Par conséquent, lorsque la date de défaillance générée est supérieure à 60 mois, l'observation correspondante est censurée à droite.

V.2. Application de la méthode

Nous allons appliquer la méthodologie proposée à différents jeux de données, différant par la loi de survie des matériels considérés.

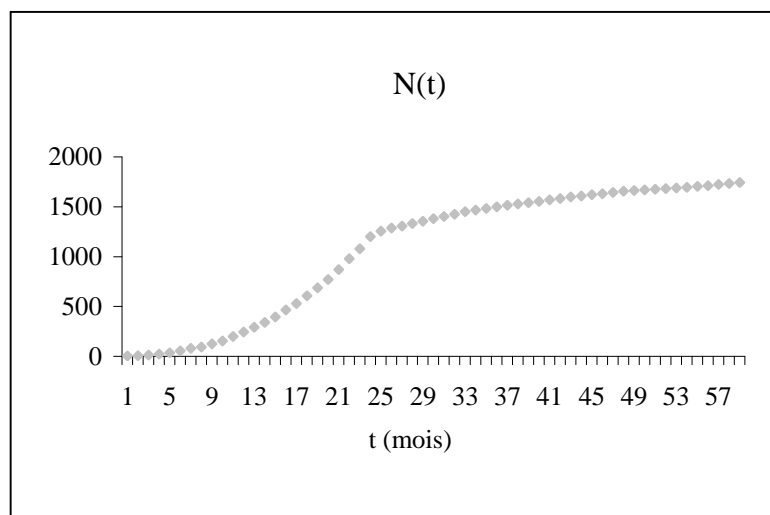
Sur ces données, nous allons vérifier que la méthode présentée dans les paragraphes précédents et qui rappelle le, exprime le taux de défaillance en fonction des seules fonctions $Q(t)$ et $N(t)$, permet de retrouver les paramètres de la loi de Weibull spécifiée.

V.2.a. Loi de Weibull de paramètres $\beta=0.3$ et $\eta=60$

Nous commençons avec une loi de Weibull de paramètres $\beta=0.3$ et $\eta=60$. Le paramètre de forme $\beta < 1$ nous assure un taux de défaillance décroissant.

Pour chaque équipement mis en service, une date de défaillance est générée aléatoirement suivant la loi de Weibull spécifiée. Nous obtenons de cette manière le flux de défaillances $N(t)$ qui est représenté sur la figure 56.

Figure 56 : Flux cumulé des défaillances



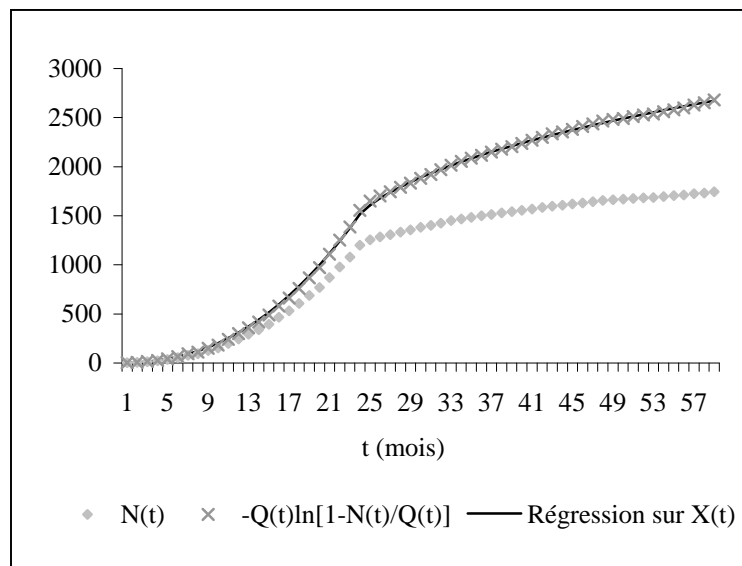
Au vu des paramètres β et η que nous avons choisis, l'approximation de la fonction $t \rightarrow -\ln W(t)$ par le flux de défaillances $N(t)$ risque d'être insuffisante.

En effet, rappelons que le paramètre η représente physiquement le temps au bout duquel environ 63.2% des équipements sont défaillants. Nous pouvons donc nous attendre à ce qu'une grande partie de la population soit défaillante après 60 mois d'observation. La population ne peut donc être qualifiée de « très fiable » et nous nous tournons vers la seconde approximation proposée pour la fonction $X(t)$:

$$X(t) \approx -Q(t) \ln \left[1 - \frac{N(t)}{Q(t)} \right].$$

Nous pouvons d'ailleurs comparer les deux approximations sur la figure 57 et constater que c'est la seconde qu'il faut choisir.

Figure 57 : Comparaison des fonctions $t \rightarrow N(t)$ et $t \rightarrow -Q(t) \ln \left(1 - \frac{N(t)}{Q(t)} \right)$



Le tableau qui suit montre les résultats que nous avons obtenus pour dix simulations. Ces résultats semblent assez bons. Sur ces dix observations, la valeur moyenne du paramètre β est de 0.287 avec un écart-type de 0.013 et la valeur moyenne du paramètre η est de 62.7 avec un écart-type de 5.35. Ces valeurs sont proches de celles que nous avons initialement choisies.

Tableau 20 : Résultats obtenus avec la méthode proposée

Numéro de la simulation	Valeur obtenue pour β	Erreur commise sur β en %	Valeur obtenue pour η	Erreur commise sur η en %
1	0.31	+3.3	57	-5
2	0.30	0	57	-5
3	0.29	-3.3	63	+5
4	0.28	-6.7	63	+5
5	0.28	-6.7	65	+8.3
6	0.29	-3.3	68	+13.3
7	0.29	-3.3	63	+5
8	0.28	-6.7	55	-8.3
9	0.26	-13.3	74	+23.3
10	0.29	-3.3	62	+3.3

Application de la méthode de Kaplan-Meier

Nous appliquons maintenant la méthode de Kaplan-Meier à notre problème. Cette méthode est la plus couramment utilisée dans le cas de données censurées. Comme nous l'avons expliqué dans le chapitre 2, les données (temps de défaillance et temps de censure) sont ordonnées de manière croissante. Notons qu'il n'y a aucune censure pour les premières 36 classes d'âge puisque chacun des équipements a été observé sur une période d'au moins 36 mois. Les premières censures arrivent pour la 37^{ième} classe d'âge.

Figure 58 : Histogramme des défaillances pour les 36 premières classes d'âge

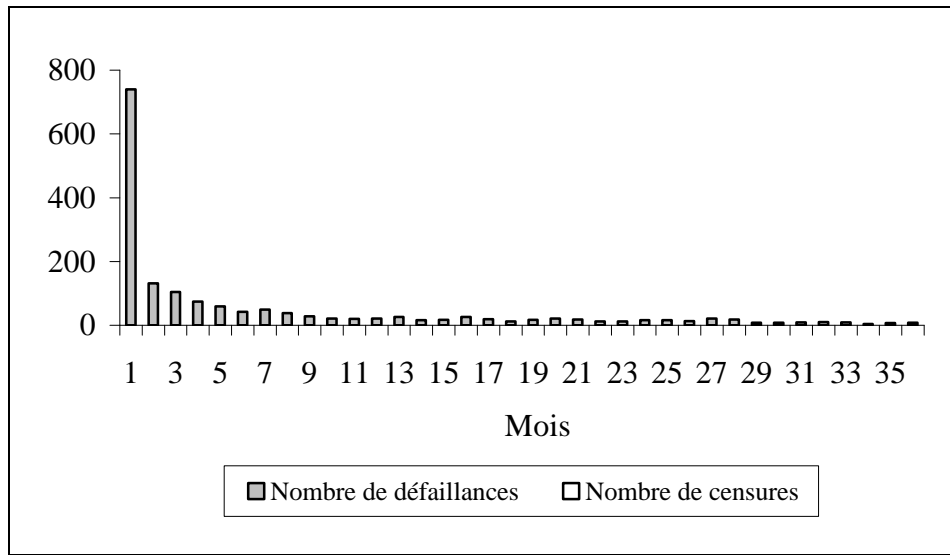
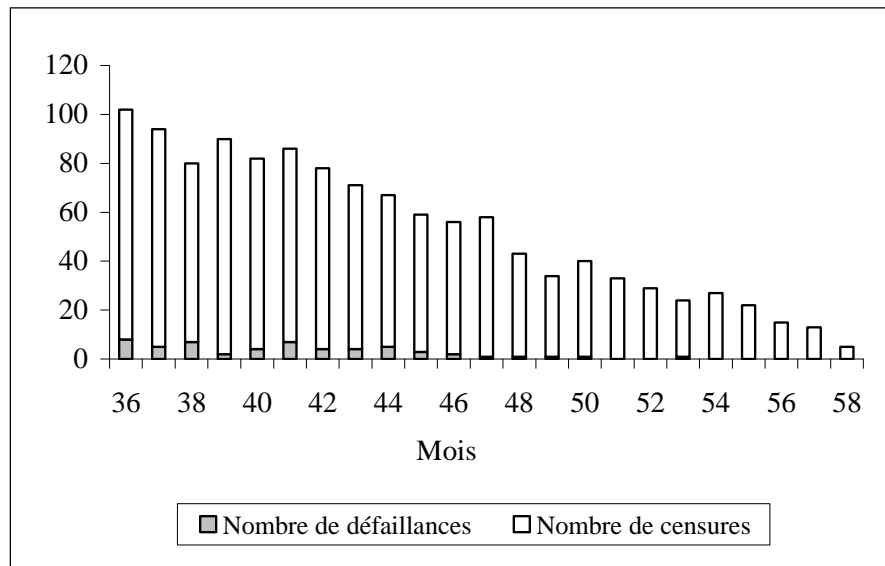


Figure 59 : Histogramme du nombre de défaillances sur les derniers mois d'observations



Remarque : Sur la figure 58, nous ne voyons que des défaillances. En effet, aucun matériel n'est censuré dans les premiers 36 mois puisque tous ont été en service pendant au moins 36 mois. Les premières censures apparaissent sur la figure 59.

Par ailleurs, on constate que le nombre de censures diminue fortement. Ceci est normal : le flux de production est croissant, et les matériels ayant les temps de censure les plus grands sont ceux qui ont été mis en service les premiers, et ce en faibles quantités.

Le tableau 21 présente les résultats obtenus pour dix simulations. La valeur moyenne observée du paramètre β est de 0.292 avec un écart-type de 0.01 et la valeur moyenne du paramètre η est de 64 avec un écart-type de 4.4.

Tableau 21 : Résultats obtenus avec la méthode de Kaplan-Meier - 1

Numéro de la simulation	Valeur obtenue pour β	Erreur commise sur β en %	Valeur obtenue pour η	Erreur commise sur η en %
1	0.3	0	65	+7.7
2	0.29	-3.3	64	+6.25
3	0.28	-6.7	66	+9
4	0.29	-3.3	57	-5.3
5	0.29	-3.3	65	+7.7
6	0.30	0	65	+7.7
7	0.31	+3.3	72	+16.7
8	0.28	-6.7	68	+11.8
9	0.30	0	57	-5.3
10	0.28	-6.7	61	+1.6

Les résultats obtenus avec la méthode de Kaplan-Meier sont assez proches de ceux obtenus avec notre méthode, ce qui semble confirmer le bien-fondé de cette dernière.

Rappel : les paramètres β et η ont été obtenus par minimisation de la quantité

$$\sum_{i=1}^{59} \left[\hat{R}(t_i) - \exp\left(-\left(\frac{t_i}{\eta}\right)^\beta\right) \right]^2, \text{ où } \hat{R}(t_i) \text{ est l'estimateur de Kaplan-Meier.}$$

V.2.b. Loi de Weibull de paramètres $\beta=1$ et $\eta=120$

Nous nous contenterons ici de présenter le tableau comparatif des résultats obtenus avec la méthode convolutive et la méthode de Kaplan-Meier sur 10 simulation successives.

Tableau 22 : Résultats obtenus, $\beta=1$ et $\eta=120$

Numéro de la simulation	β		η	
	Méthode Convolutive	Kaplan-Meier	Méthode Convolutive	Kaplan-Meier
1	0.87	0.88	151	149
2	0.97	0.89	123	158
3	1.05	0.93	117	137
4	0.95	0.95	129	135
5	0.91	0.90	142	135
6	0.98	0.97	130	135
7	0.96	0.95	128	135
8	0.95	0.86	137	153
9	0.97	0.94	132	135
10	0.96	0.90	129	140
Moyenne	0.957	0.917	132	141
Ecart-type	0.05	0.04	10	9

Remarque : le taux de censure dans cet exemple est de 70%.

V.2.c. Loi de Weibull de paramètres $\beta=2$ et $\eta=36$

Nous nous contenterons encore de présenter le tableau comparatif des résultats obtenus avec la méthode convolutive et la méthode de Kaplan-Meier sur 10 simulation successives.

Tableau 23 : Résultats obtenus, $\beta=2$ et $\eta=36$

Numéro de la simulation	β		η	
	Méthode Convolutive	Kaplan-Meier	Méthode Convolutive	Kaplan-Meier
1	1.97	1.78	37	37
2	2.00	1.83	37	37
3	1.85	1.87	38	37
4	1.94	1.80	36	37
5	1.98	1.78	37	37
6	1.89	1.81	37	37
7	1.87	1.82	37	37
8	1.97	1.82	37	38
9	1.95	1.88	37	36
10	1.88	1.87	37	37
Moyenne	1.93	1.83	37	37
Ecart-type	0.05	0.05	0.5	0.5

Remarque : le taux de censure dans cet exemple est de 25%.

VI. Application à des données industrielles

Nous proposons maintenant d'appliquer la méthode à un jeu de données réelles. Les deux types d'information dont nous disposons sont :

- Le flux des mises en service, qui est donné dans le Tableau 24 ; le flux cumulé des mises en service est représenté graphiquement sur la Figure 60.
- Le flux des pannes, que nous avons renseigné dans le Tableau 25 ; ce flux est construit à partir de la base de données des déposes (Figure 61) ; le flux cumulé correspondant est représenté graphiquement sur la Figure 62.

Tableau 24 : Flux de livraison

Mois	livraisons	Cumul livraisons
Mars-03	12	12
Avril-03	12	24
Mai-03	12	36
Juin-03	0	36
Juillet-03	18	54
Aout-03	12	66
Sept-03	48	114
Oct-03	66	180
Nov-03	60	240
Déc-03	60	300

Figure 60 : Flux cumulé de livraison

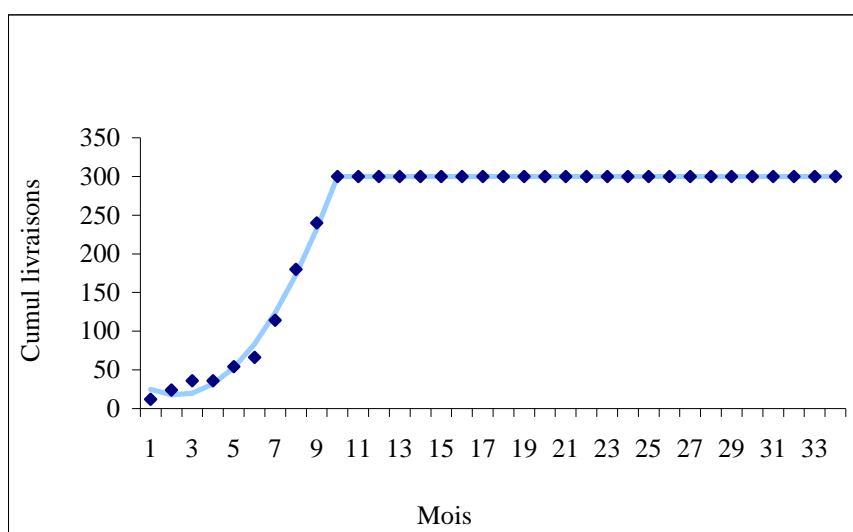


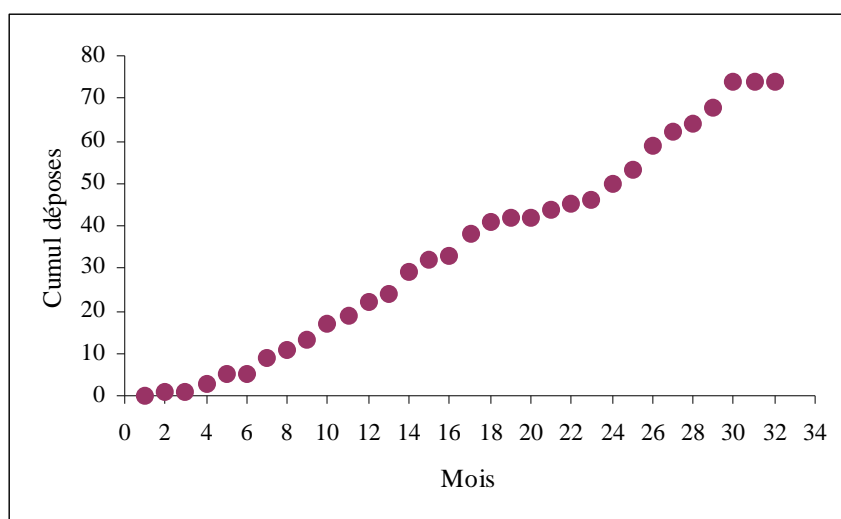
Figure 61 : Dépotes observées en clientèle

prog	a/c	Mfr P/N	Mfr S/N	Removal date	ICAO cd	a/c SN	Removal Typ	conf	Comp	mois de dépose
SA	A319	C19298AF05	C19298001527	29/12/2005	EZY	H2249	U	Y	BITE EEPROM	déc-05
SA	A320	C19298AF05	C19298001324	27/12/2005	EIN	D2294	U	Y	UND	déc-05
SA	A320	C19298AF05	C19298001372	26/12/2005	AMC	D2189	U	Y	BLM CALIBRATION	déc-05
SA	A319	C19298AF05	C19298002073	25/12/2005	AFR	H2213	U	Y	UND	déc-05
SA	A320	C19298AF05	C19298001010	15/12/2005	ACI	D2152	U	Y	LCD PANEL	déc-05
SA	A320	C19298AF05	C19298001860	10/12/2005	ANZ	D2297	U	Y	LOCKER SPRING	déc-05
SA	A319	C19298AF05	C19298001249	09/12/2005	AMC	H2186	U	Y	BLM CALIBRATION	déc-05
SA	A320	C19298AF05	C19298001328	21/11/2005	QTR	D1648	U	Y	UND	nov-05
SA	A319	C19298AF05	C19298001832	21/11/2005	EZY	H2319	U	Y	SRAM EXCHANGE	nov-05
SA	A318	C19298AF05	C19298001992	21/11/2005	FFT	UNK	U	Y	UND	nov-05
SA	A320	C19298AF05	C19298001491	16/11/2005	ANZ	D2090	U	Y	BLM CALIBRATION	nov-05
SA	A320	C19298AF05	C19298003421	13/11/2005	XXX	UNK	U	Y	LOCKER SPRING	nov-05
SA	A320	C19298AF05	C19298001402	10/11/2005	EIN	UNK	U	Y	N CRACKED OR SCRA	nov-05
SA	A319	C19298AF05	C19298001209	08/11/2005	BMA	H2188	U	Y	UND	nov-05
SA	A319	C19298AF05	C19298001596	08/11/2005	FIN	H2266	U	Y	SRAM EXCHANGE	nov-05
SA	A318	C19298AF05	C19298001653	08/11/2005	FFT	P2276	U	Y	SRAM EXCHANGE	nov-05
SA	A320	C19298AF05	C19298000844	06/11/2005	QTR	D1566	U	Y	BLM CALIBRATION	nov-05
SA	A321	C19298AF05	C19298001271	06/11/2005	QTR	G2107	U	Y	BLM CALIBRATION	nov-05
SA	A321	C19298AF05	C19298001289	06/11/2005	QTR	G2107	U	Y	BLM CALIBRATION	nov-05
SA	A321	C19298AF05	C19298001294	06/11/2005	QTR	G2107	U	Y	BLM CALIBRATION	nov-05
SA	A321	C19298AF05	C19298001290	04/11/2005	QTR	G2107	U	Y	BLM CALIBRATION	nov-05
SA	A319	C19298AF05	C19298001128	02/11/2005	EZY	H2129	U	Y	SRAM EXCHANGE	nov-05
SA	A321	C19298AF05	C19298000385	30/10/2005	QTR	G1928	U	Y	LCD PANEL	oct-05
SA	A321	C19298AF05	C19298000660	30/10/2005	QTR	G1928	U	Y	BLM CALIBRATION	oct-05
SA	A320	C19298AF05	C19298000874	28/10/2005	AFL	D2133	U	Y	UND	oct-05
SA	A319	C19298AF05	C19298001724	28/10/2005	EZY	H2283	U	Y	LOCKER SPRING	oct-05
SA	A320	C19298AF05	C19298001202	24/10/2005	QTR	D1895	U	Y	BLM CALIBRATION	oct-05

Tableau 25 : Flux de déposes

Mois	Déposes	Cumul déposes
Mars-03	0	0
Avril-03	0	0
Mai-03	0	0
Juin-03	1	1
Juillet-03	0	1
Aout-03	2	3
Sept-03	2	5
Oct-03	0	5
Nov-03	4	9
Déc-03	2	11
Janv-04	2	13
Févr-04	4	17
Mars-04	2	19
Avril-04	3	22
Mai-04	2	24
Juin-04	5	29
Juillet-04	3	32
Aout-04	1	33
Sept-04	5	38
Oct-04	3	41
Nov-04	1	42
Déc-04	0	42
Janv-05	2	44
Févr-05	1	45
Mars-05	1	46
Avril-05	4	50
Mai-05	3	53
Juin-05	6	59
Juillet-05	3	62
Aout-05	2	64
Sept-05	4	68
Oct-05	6	74

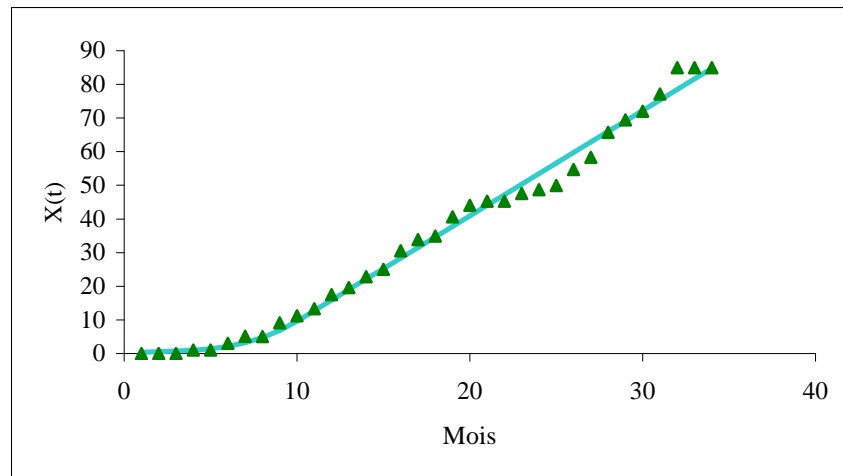
Figure 62 : Flux cumulé des déposes



Regardons maintenant la fonction $t \rightarrow X(t) = -Q(t) \ln \left[1 - \frac{N(t)}{Q(t)} \right]$; nous avons vu que, lorsque la loi de survie était une loi de Weibull de paramètres β et η , $X(t)$ pouvait s'écrire sous la forme [18] :

$$\frac{\beta}{\eta^\beta} \Gamma(\beta) \sum_{i=0}^m \frac{P^{(i)}(0)t^{\beta+i}}{\Gamma(\beta+1+i)} - \frac{\beta}{\eta^\beta} \Gamma(\beta) \sum_{i=1}^m \frac{P^{(i)}(D)\gamma(t-D)(t-D)^{\beta+i}}{\Gamma(\beta+1+i)} \quad (\text{cf Figure 63}).$$

Figure 63 : X(t) et régression



Les paramètres β et η sont obtenus par régression, comme indiqué précédemment. Nous obtenons :

$$\begin{cases} \beta = 1,00 \\ \eta = 96 \end{cases}$$

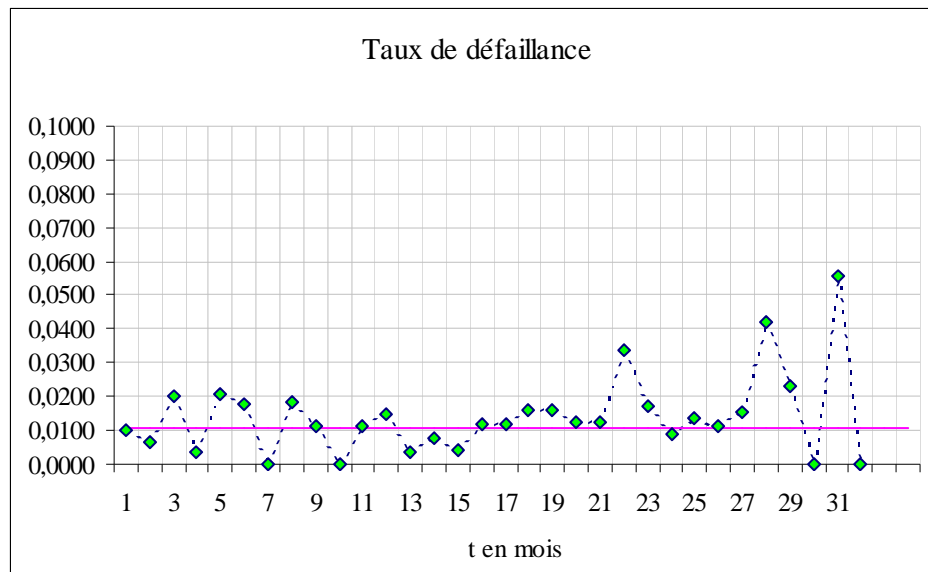
Remarque : quelques mois après avoir fait ces calculs, nous avons pu disposer de données supplémentaires. Les dates de mise en service de chaque matériel nous ont été fournies, si bien que nous avons pu évaluer les temps de survie et les temps de censures.

Le taux de défaillance obtenu par la méthode de Kaplan-Meier est représenté sur la Figure 64. Les derniers valeurs sont peu représentatives car elles correspondent aux premiers mois de mise en service, pour lesquels les flux étaient très faibles (une dizaine de matériels). Sur les 26 premiers mois, le taux de défaillance est quasi constant, ce qui est cohérent avec la valeur $\beta = 1$ trouvée précédemment.

La méthode du maximum de vraisemblance nous a donné les valeurs :

$$\begin{cases} \beta = 1,07 \\ \eta = 90 \end{cases} .$$

Figure 64 : taux de défaillance obtenu par la méthode de Kaplan-Meier



VII. Conclusion

Nous avons présenté dans ce chapitre une méthodologie permettant d'estimer la fiabilité intrinsèque d'une population d'équipements, dans des conditions peu usuelles où les seules informations disponibles sont le flux de production et le flux de défaillances. C'est à notre avis précisément dans la simplicité des données d'entrée que réside l'intérêt de la méthode. La collecte des données est en effet un problème récurrent pour les fiabilistes. Les équipements aéronautiques sur lesquels nous avons travaillé sont produits en grand nombre. Les clients sont très nombreux et il nous a été difficile d'obtenir un bon « état des lieux » de la population. En particulier, il était très difficile de connaître les âges de défaillance ou de censure de chacun des équipements mis en service afin d'employer les méthodes statistiques classiques : Kaplan-Meier, Johnson, etc...

En ce qui concerne le traitement de l'équation de convolution, le choix de la transformation de Laplace peut être discuté. Nous avons envisagé plusieurs méthodes de « déconvolution ». Puisque les données dont nous disposions étaient discrètes, nous avons expérimenté la déconvolution discrète [Commenges, 1980] et la transformation de Fourier discrète [Brigham, 1974] [Dutuit, 1979]. Ces méthodes n'ont rien donné.

Quoi qu'il en soit, la méthode proposée nous a semblé être une bonne alternative puisqu'elle permet, malgré le peu d'informations disponibles, d'estimer de manière précise la loi de survie de l'équipement considéré.

CONCLUSION ET PERSPECTIVES

Si on commence avec des certitudes, on finit avec des doutes. Si on commence avec des doutes, on finit avec des certitudes.

Francis Bacon.

J'ai entamé ma thèse avec beaucoup de doutes... Le sujet était vaste, ses limites me semblaient bien incertaines, comment allais-je l'aborder ?

Puis les problématiques se sont dégagées, l'une après l'autre, chacune répondant à un besoin bien spécifique...

Nous sommes ainsi partis de deux problèmes concrets, posés par l'entreprise. Autour de ces problèmes, nous avons construit des solutions adaptées, répondant à deux critères essentiels : tout d'abord, elles étaient fondées sur des théories qui, nous semblait-il, pouvaient être scientifiquement valorisées ; ensuite, elles se voulaient adaptées au monde industriel.

Pour mémoire, nous rappelons ci-dessous ces deux problématiques.

La première est centrée sur l'analyse des clusters d'évènements. Elle fait appel à une notion statistique peu connue (du moins ne la connaissions-nous pas avant de rencontrer notre problème...) : la statistique de balayage.

Avant d'attaquer le fond de la question, il nous a semblé essentiel de consacrer quelques lignes à son fondement. Le premier chapitre de ce manuscrit constitue donc une introduction à l'analyse des « clusters » d'évènements. Afin de couper court à tout problème de notation ou de compréhension des énoncés, nous avons proposé très tôt une formulation mathématique du problème. Ensuite, nous avons jugé bon de mettre en avant sa particularité et les difficultés qui en découlaient, sur un exemple élémentaire : un « jeu » de pile ou face.

Les différentes méthodes que nous avons proposées sont, elles, présentées dans le deuxième chapitre du manuscrit. Deux d'entre elles font appel à une simulation de Monte-Carlo, les autres recourent à la théorie des chaînes de Markov.

Un problème sans solution est un problème mal posé.

Albert Einstein.

La deuxième est focalisée sur l'analyse du Retour d'EXpérience. Concernant cette thématique, qui nous a posé de nombreux problèmes, la question était très précise : était-il possible d'estimer la loi de survie (ou, de manière équivalente, le taux de défaillance) d'un matériel lorsque les données de durées de vie n'étaient pas disponibles ? En d'autres termes, pouvait-on s'affranchir de méthodes statistiques telles que celle de Kaplan-Meier, qui nécessitent la connaissance des temps de survie ou des temps de censure de chaque matériel inclus dans l'étude ?

Après quelques rappels de fiabilité, nécessaires pour appréhender la suite, le chapitre 3 présente quelques unes des méthodes statistiques en question. Le chapitre 4, quant à lui, apporte une réponse positive grâce à la méthode dite « convolutive ». Cette dernière s'est révélée être une alternative efficace aux méthodes statistiques traditionnelles que nous n'étions pas toujours en mesure d'appliquer sur nos jeux de données.

Les méthodes proposées dans ce manuscrit sont, à notre connaissance, originales : nous n'avons pas eu vent, au cours de nos recherches, de l'utilisation d'un réseau de Petri ou d'une chaîne de Markov pour traiter la statistique de balayage. De même, la manière bien

particulière dont la méthode convolutive appréhende le Retour d'EXpérience semble nouvelle.

Ainsi, notre étude ne se prétend pas être une synthèse exhaustive sur ces deux thématiques. Nous avons plutôt fait le choix de l'originalité.

Je termine ces trois années d'étude avec deux certitudes. La première est que la fiabilité est une discipline très ouverte et très étendue, qui soulève constamment de nombreuses questions. N'est-ce pas d'ailleurs ce qui fait tout son intérêt ? La seconde est que certaines thématiques, semblant à première vue très théoriques, peuvent parfaitement trouver leur écho dans le monde industriel.

Les perspectives sont nombreuses. La méthode convolutive peut certainement s'enrichir de concepts tels que la considération de délais de mise en service après livraison, la prise en compte des réparations, ou encore l'application de lois de survie plus complexes que la loi de Weibull.

*Un problème prévu est un problème en moins.
Maurice Dantec, extrait de La sirène rouge.*

Concernant les statistiques de balayage, nous avons observé certains phénomènes qui mériteraient d'être approfondis. Par exemple, le fait que les approches markoviennes fournissent une approximation de la statistique de balayage n'est pas démontré à ce jour. D'un point de vue plus industriel, il est prévu de développer un outil spécifique qui permettrait de déterminer le nombre de défauts acceptables pour certains types de composants, et permettrait ainsi de définir des « seuils d'alertes ».

BIBLIOGRAPHIE

- [ASTE, 1993] ASTE. Le rôle des essais dans la maîtrise de la fiabilité. Editions ASTE.
- [Bacha et al., 1998] Mostafa Bacha, Gilles Celeux, Edwige Idée, André Lannoy et Dominique Vasseur. Estimation de modèles de durées de vie fortement censurées. Collection de la Direction des Etudes et Recherches d'Electricité de France. Eyrolles.
- [Barlow et Proschan, 1996] R.E. Barlow and and F. Proschan. Mathematical Theory of Reliability, 12th. Edition. New-York : J.Wiley.
- [Berthon et al., 2006a] J. Berthon, S. Charruau, J. Antoch, J.M. Deshouillers et Y. Dutuit. Une approche analytique permettant d'obtenir la fiabilité opérationnelle d'équipements embarqués. Actes de la conférence Lambda Mu 15. Lille, France.
- [Berthon et al., 2006b] J. Berthon, S. Charruau, J. Antoch, J.M. Deshouillers et Y. Dutuit. An approach to assess the operational reliability of embedded equipments. Proceedings of the Conference ESREL. Estoril, Portugal.
- [Berthon et al., 2008a] J. Berthon, J.M. Deshouillers, F. Innal et Y. Dutuit. Scan statistics : Looking for clusters of undesirable events. Proceedings of the Conference PSAM 9. Hong-Kong, China.
- [Berthon et al., 2008b] J. Berthon, J.M. Deshouillers, F. Innal et Y. Dutuit. Evaluation de la probabilité d'occurrence d'agrégats d'accidents au moyen des statistiques de balayage. Actes de la conférence Lambda Mu 16. Avignon, France.
- [Billingsley, 1986] P. Billingsley. Probability and measure. Second edition. John Wiley and Sons, New York. Chapitre 23.
- [Bryant, 1986] R.Bryant. Graph-Based Algorithms for Boole and Functions Manipulation. IEEE Transactions on Computers. 35, n° 8, pp. 677-692.
- [Brigham, 1974] E. Oran Brigham. The Fast Fourier Transform. Prentice-Hall, Inc.
- [CEI 271, 1974] Comission Electrotechnique Internationale. Liste des termes de base, définitions et mathématiques, application à la fiabilité. Publication n°271. 2^{ème} édition.
- [Charruau, 1990] Cours de qualité et de fiabilité en électronique. Communication personnelle.
- [Charruau et al., 2006] S. Charruau, F. Guerin, J. Hernandez Dominguez and J. Berthon. Reliability Estimation of Aeronautic Component by Accelerated Tests. Proceedings of the Conference ALT'06. Angers, France.
- [Commenges, 1980]. D. Commenges. Etude théorique et mise en œuvre d'algorithmes de déconvolution - Application à la détermination de la distribution des temps de transit pulmonaires. Thèse soutenue à l'Université Bordeaux I.

- [Demengel, 2002] G. Demengel. Transformation de Laplace. Théorie et illustrations par les exemples. Paris, Ellipses.
- [Dupin, 2007] G. Dupin. A non-existent series of railways accidents. RTS (Recherche Transports Sécurité), n° 96, pp. 255-268.
- [Dutuit et al., 1997] Y. Dutuit, E. Châtelet, J.P. Signoret and P. Thomas. Dependability modelling and evaluation by using stochastic Petri nets : application to two test cases. Reliability Engineering & System Safety, vol.55, n°2, pp.117-124.
- [Dutuit et Rauzy, 2002] Y. Dutuit et A. Rauzy. Approche analytique évènementielle : l'arbre des défaillances. Maîtrise des risques et sûreté de fonctionnement des systèmes de production. Hermès-Lavoisier. Paris.
- [Dutuit, 1979] Y. Dutuit : La transformation de Fourier discrète en spectroscopie temporelle. Revue de Physique appliquée, novembre 1979.
- [Faure et Laurière, 1974] R. Faure et J.L. Laurière. Fiabilité et renouvellement des équipements. Bordas.
- [FIDES, 2004] Guide FIDES Edition A. Méthodologie de fiabilité pour les systèmes électroniques. DGA – DM/STTC/CO/477-A.
- [Foata et Fuchs, 2003] D. Foata et A. Fuchs. Calcul des probabilités : Cours, exercices et problèmes corrigés. Dunod.
- [Foata et Fuchs, 2004] D. Foata et A. Fuchs. Processus stochastiques : processus de Poisson, chaînes de Markov et martingales. Dunod.
- [Glaz et al., 2001] J. Glaz, J. Nauz and S. Wallenstein. Scan Statistics. Springer Series in Statistics. Springer-Verlag New-York, Inc.
- [Hoang, 2003] Hoang Phan. Handbook of reliability engineering. Chapitre 26 : Statistical Methods for Reliability Data Analysis. Ed. Springer.
- [Lannoy et Procaccia, 1994] André Lannoy et Henri Procaccia. Méthodes avancées d'analyse des bases de données du retour d'expérience industriel. Collection de la Direction des Etudes et Recherches d'Electricité de France. Eyrolles.
- [Limnios, 2005] N. Limnios. Arbres de défaillance, Hermès-Lavoisier. Paris
- [Malhotra et Trivedi, 1995] M. Malhotra and K.S. Trivedi. Dependability modelling using Petri Nets. IEEE Trans. Reliability, vol.44, n°3, pp.428-440.
- [Meeker et Escobar, 1998] William Q. Meeker and Luis A. Escobar. Statistical Methods for Reliability Data. Wiley series in probability and statistics.
- [MIL-HDBK-217F, 1991] MIL-HDBK-217F. Military Handbook. Reliability prediction of electronic equipment. Department Of Defense. Washington DC.

[Pagès et Gondran, 1990] Alain Pagès et Michel Gondran. Fiabilité des systèmes. Collection de la Direction des Etudes et Recherches d'Electricité de France. Eyrolles.

[Proccacia et Piepszownik, 1992] Henri Procaccia et Laurent Piepszownik. Fiabilité des équipements et Théorie de la décision statistique fréquentielle et bayésienne. Collection de la Direction des Etudes et Recherches d'Electricité de France. Eyrolles.

[Rauzy, 1996] A.Rauzy. A brief introduction to Binary Decision Diagrams. Binary Decision Diagrams and Reliability. RAIRO-APII-JESA. 30, n°8, pp. 1033-1050.

[Saporta, 2006] G. Saporta. Probabilités, analyse des données et statistique. Technip.

