Projet Modèle de durée

Enseignante : Dorothée Delaunay

Master 2 MAS - PPE

Problématique

Compréhension des facteurs clés influençant le risque de dysfonctionnement lié à l'ouverture d'une nouvelle ligne pour un opérateur téléphonique.

Léo DUTERTRE-LADUREE
leo.dutertre-laduree@etudiant.univ-rennes1.fr

Axel GARDAHAUT
axel.gardahaut@etudiant.univ-rennes1.fr

Guy TSANG
guy.tsang@etudiant.univ-rennes1.fr

Date limite de rendu
13 Mars 2020

UNIVERSITÉ DE RENNES 1

Table des matières

1 Introduction

Le respect des délais de mise en service est un levier majeur dans la satisfaction des clients souscrivant à un abonnement auprès d'un opérateur télécom. Ce dernier peut voir ses nouveaux abonnés partir chez un concurrent lorsque le délai d'attente est trop lent avant de pouvoir profiter des services habituels. En réalité, les conséquences peuvent aller plus loin : le nouvel abonné peut demander des dommages et intérêts pour le préjudice subit lorsque la mise en service d'une ligne n'est pas opérationnelle dans un délai d'un mois après souscription au contrat. Une ligne est dite opérationnelle lorsque le client peut profiter des services (téléphone, internet) sans dysfonctionnement.

Pour minimiser le risque de départ de nouveaux clients suite à un dysfonctionnement sur leur ligne internet, il est important pour les opérateurs d'identifier et comprendre les facteurs clés pouvant influencer le risque de dysfonctionnement lié à l'ouverture de la ligne. Cette problématique peut être traitée à l'aide de modèles de survie.

La durée de vie d'une ligne internet correspond au temps entre le moment de l'activation de la ligne et le moment où l'abonné déclare un dysfonctionnement sur la ligne. Lorsque la durée de vie est supérieure à 30 jours, le dysfonctionnement n'est plus lié à l'ouverture de la ligne et donc ne rentre pas dans le cadre de notre problématique. Ainsi, la variable de censure est placée 30 jours après activation de la ligne et prendra la valeur « 1 » lorsqu'un dysfonctionnement a été signalé dans un délai de 30 jours après activation de la ligne, « 0 » sinon.

Le but de cette étude est double : identifier et quantifier les leviers aggravant le risque de dysfonctionnement des lignes nouvellement activées afin que les opérateurs puissent s'y intéresser et estimer la durée avant la survenue d'un dysfonctionnement dans la période des 30 jours.

Une première partie de l'étude se concentrera sur les modèles non paramétriques afin d'estimer les courbes de survie et de risque et identifier les variables catégorielles qui semblent pertinentes à l'étude. Une seconde partie s'intéressera aux modèles paramétriques pour trouver l'ajustement paramétrique la plus adaptée aux données compte tenu des variables exogènes. Enfin, une dernière partie sera consacrée aux modèles semi-paramétriques qui permettront de quantifier les effets des leviers exogènes.

1.1 Champ d'étude

L'historique couvre les activations de ligne du mois d'Avril 2019 (table 1) pour les régions du Nord-Est, de l'Ouest et de l'Île-de-France. Au total, ce sont 42 807 ouvertures de lignes répertoriées, pour 13 740 signalements de dysfonctionnement. Une ligne sur 3 en moyenne subit un dysfonctionnement dans un délai de 30 jours après son activation. Ce taux semble relativement élevé et constitue une grande marge de manœuvre pour le fournisseur afin d'améliorer la satisfaction de ses nouveaux abonnés et éviter les contestations et départs prématurés.

Table 1 – Période d'étude

	Min	Max	N	NMiss
Date_activation	01APR19 :06 :06 :00	30APR19 :20 :30 :00	42807	0
Date_signal	01APR19 :09 :42 :00	29MAY19 :21 :22 :00	13740	29067

Cette période d'étude restreinte ne permet pas de capturer les phénomènes liés au temps. Par exemple, le nombre de dysfonctionnements pourrait fluctuer de manière importante selon la saison ou encore, selon la demande. Une période haute peut amener les techniciens à devoir installer et activer plus rapidement les nouvelles lignes, en prenant le risque de devoir revenir résoudre un dysfonctionnement.

Sur la carte suivante (figure 1), sont représentés les pourcentages de lignes dysfonctionnelles après activation récente pour chaque région couverte par les données. Les régions ayant moins de 5 lignes nouvellement activées dans la période d'étude ne sont pas représentées sur la carte ci-dessous à cause de la fiabilité des pourcentages sur de faibles effectifs.

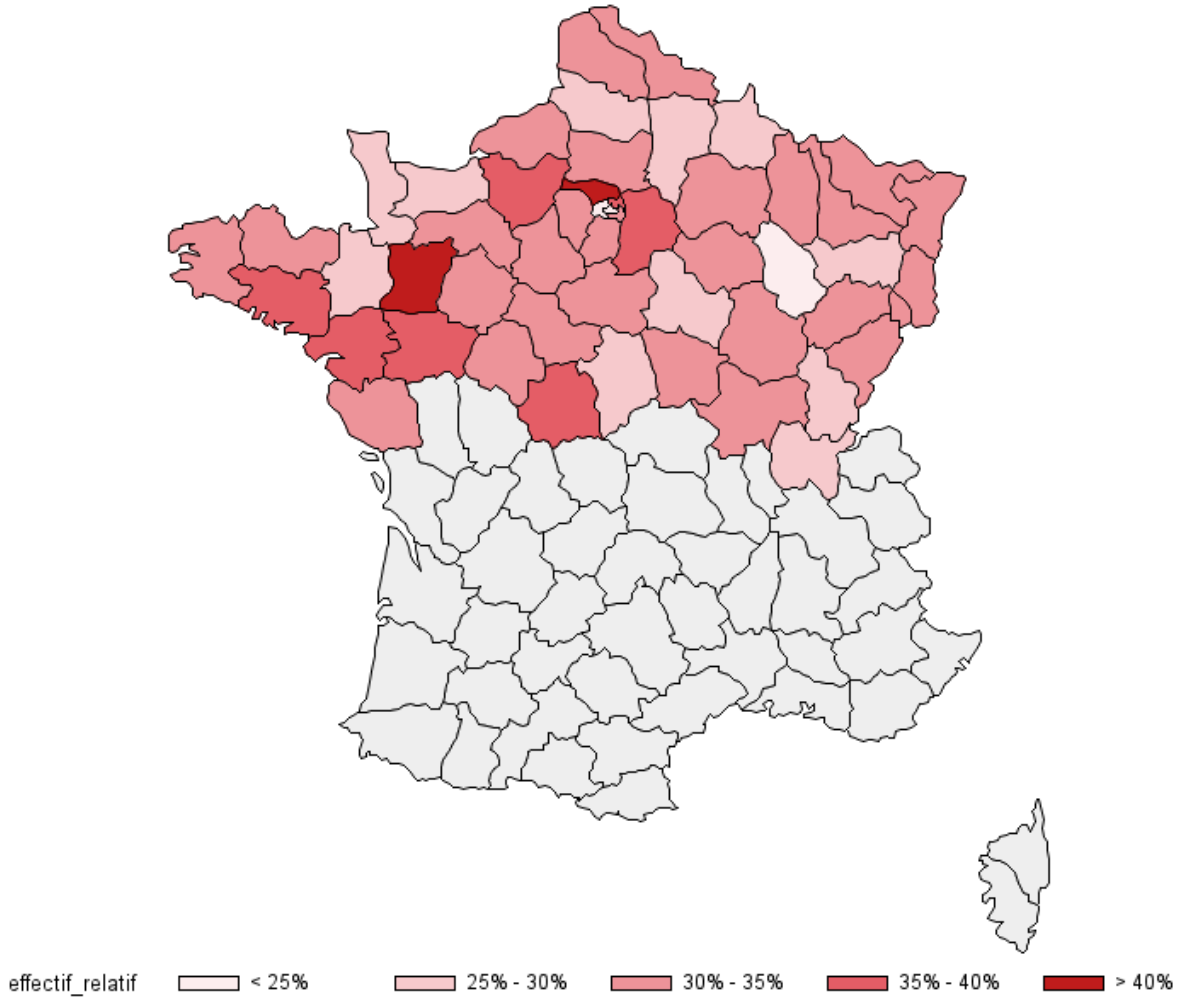

Figure 1 – Pourcentage de dysfonctionnement par région

On peut dans un premier temps noter que certaines régions (notamment en Mayenne) sont plus souvent assujetties aux dysfonctionnements liés à l'activation de lignes internet. De façon globale, les régions de l'Ouest et de l'Île-de-France semblent présenter des risques plus élevés de dysfonctionnement sur les lignes nouvellement activées que pour le Nord-Est. Ainsi, la considération de la localisation géographique à travers la région d'appartenance semble pertinente pour l'étude des facteurs exogènes à risque.
La table suivante (table 2) présente les durées de signalement minimales, moyennes, médianes et maximales. Lorsqu'il n'y a pas eu de signalement dans un délai de 30 jours, la variable de censure prend la valeur 0. Les dysfonctionnements représentent 32.10% des lignes. On note que la moitié des signalements sont effectués au plus tard 5 jours après l'activation de la ligne. La durée moyenne avant signalement est d'environ une semaine. Ainsi, l'apparition de dysfonctionnements est souvent très précoce (moins d'une semaine), ce qui laisse généralement 3 semaines pour planifier et effectuer une intervention afin de rendre la ligne opérationnelle avant le délai des 30 jours. Cependant, on note qu'un signalement a été effectué le 29ème jour, ce qui rend la situation délicate. Des clauses spécifiques peuvent exister dans les contrats pour ces cas de dysfonctionnements tardivement signalés. Il est également possible pour le fournisseur de savoir si ses services ont été régulièrement utilisés après activation de la ligne.

Table 2 – Statistiques descriptives des durées de signalement

Problème sur la ligne	N Obs	N Miss	Minimum	Mean	Median	Maximum
0	29067	0	30.00	30.00	30.00	30.00
1	13740	0	0	7.33	4.88	29.44

2 Approche non paramétrique

2.1 Estimation non paramétrique des fonctions de survie et de hazard

2.1.1 Méthode de Kaplan-Meier

La méthode de Kaplan-Meier estime de façon empirique le nombre de décès (dysfonctionnements) en se basant sur les données réelles. L'estimateur $\widehat{S}(t)$ de la fonction de survie est le produit de la probabilité d'avoir survécu jusqu'à présent et de la probabilité conditionnelle de survivre au prochain temps. De façon formelle,

$$\widehat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

où d_i est le nombre d'individus qui sont décédés à t_i et n_i le nombre d'observations à risque. De cette formule, on en tire trois propriétés de la fonction de survie :

- La survie est égale à 1 à l'étape initiale.
- La survie est estimée non-croissante pour tout t.
- La survie estimée est une fonction en marches d'escalier.

En appliquant cette méthode sur l'ensemble des données, sans stratification ni variable explicative, on obtient les courbes de survie et de hazard suivantes (figures 2 et 3).

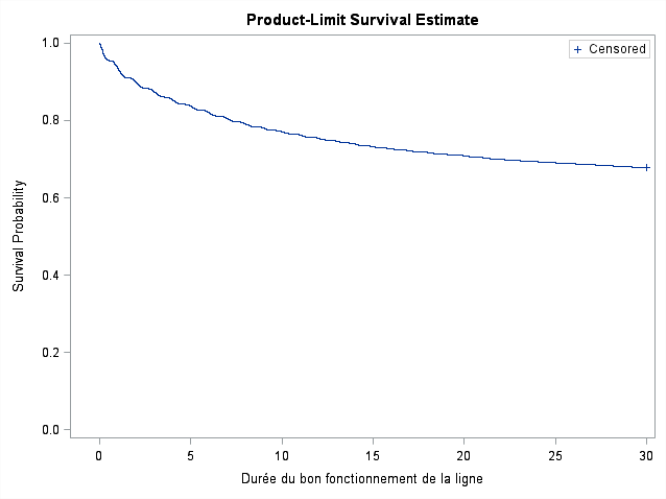

Figure 2 – Survie estimée par la méthode de Kaplan-Meier

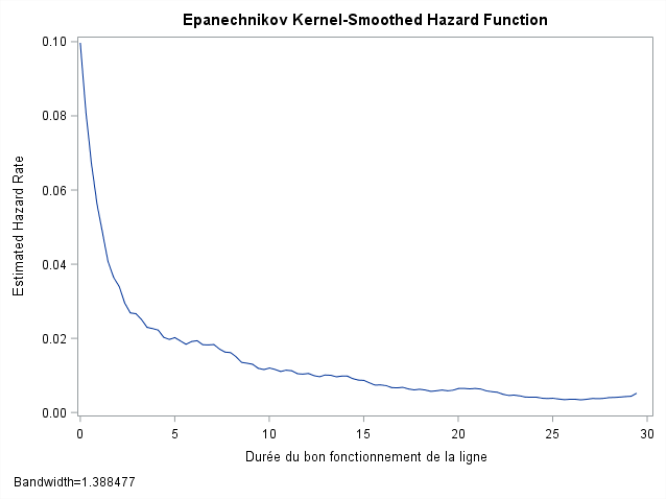

Figure 3 – Risque estimé par la méthode de Kaplan-Meier

On note une baisse régulière du nombre de lignes encore en vie au cours du temps, à l'exception des premiers jours où on peut remarquer une chute significative du nombre de lignes en vie. Ceci se traduit par une fonction de hazard ayant de fortes valeurs pour les premiers jours puis une stabilisation du risque au-delà d'une semaine.
Ces premiers résultats ne sont pas surprenants dans la mesure où si après activation de la ligne, celle-ci fonctionne et est dans un état stable (absence de perturbations), le risque de découvrir une anomalie est élevé au départ puis peu probable ensuite. Par ailleurs, l'activation des lignes se fait souvent à distance, avec l'aide de l'abonné sur place, guidé par un technicien. Si l'abonné échoue à réaliser certaines manipulations sans s'en rendre compte, ceci peut conduire à un comportement imprévu de la ligne et un dysfonctionnement sera signalé. Enfin, il se peut que l'abonné ne sache pas comment utiliser les services (internet) proposés par le fournisseur : un signalement sera malencontreusement effectué. Toutes ces pistes peuvent expliquer pourquoi le taux de risque est à 10% le premier jour.

Une autre façon d'estimer ces fonctions est la méthode actuarielle.

2.1.2 Méthode actuarielle

La méthode actuarielle découpe la période d'étude en intervalles afin de synthétiser l'information. Par conséquent, cette méthode est moins précise que la méthode de Kaplan-Meier mais est plus rapide puisqu'on travaille sur un ensemble réduit d'itérations. On a préféré couper la période en journées pour faciliter l'interprétation et pour conserver la période haute dans les premiers jours après activation de la ligne.

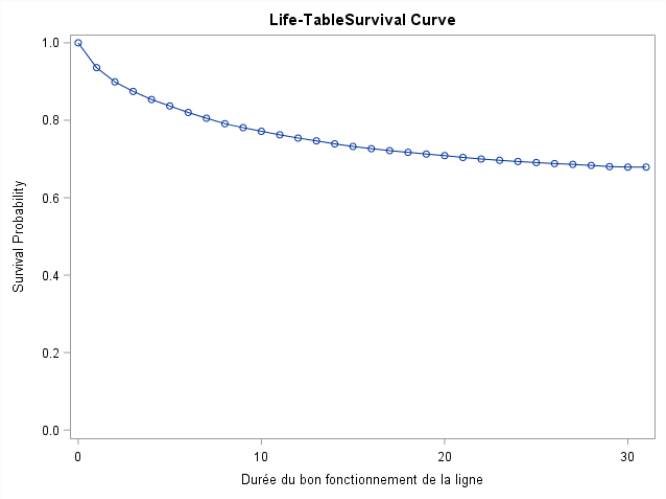

Figure 4 – Survie estimée par la méthode actuarielle

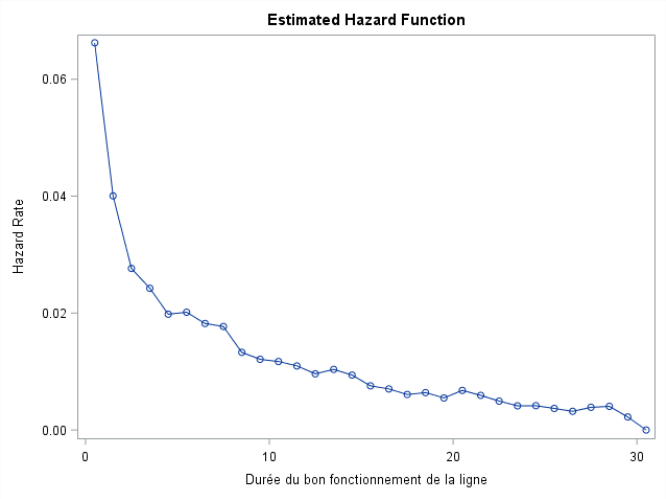

Figure 5 – Risque estimé par la méthode actuarielle

Cette méthode revient à approximer la méthode de Kaplan-Meier. Les conclusions qui en sortent sont les mêmes. Si ces méthodes d'estimation ne permettent pas de quantifier les effets de variables exogènes explicatives ou de stratification, elles peuvent néanmoins servir à identifier les variables de stratification permettant de différencier les courbes de survie en fonction de leurs caractéristiques.

2.2 Segmentation et comparaison des courbes de survie

D'après la carte des régions, on soupçonne que la localisation géographique peut influencer de manière significative la durée de vie des lignes. Parmi les variables disponibles, d'autres facteurs pourraient également y jouer un rôle. On peut alors estimer les fonctions de survie selon les différents segments possibles.

Pour quantifier la différenciation dûe à la stratification des individus, des tests des rangs exacts adaptés aux données censurées existent : le test du log-rank (Nathan Mantel) et le test de Gehan-Wilcoxon. On peut également passer par un test de ratio de log-vraisemblance.

L'hypothèse nulle du test du log-rank est la similitude entre la survie de la population entière et la survie pour chaque strate. Celle du test de Gehan-Wilcoxon est la similitude entre la survie de chaque strate.

La puissance de ces tests est influencée par la distribution des censures entre les strates et le croisement ou non des différentes courbes de survie. Pour chaque variable de stratification, on veillera alors à avoir une répartition équilibrée des censures et à l'absence de croisements entre les courbes de survie.

2.2.1 Statistiques descriptives des variables de stratification

Les variables sélectionnées parmi la liste initiale sont celles qui semblent présenter des classes pertinentes pour segmenter les nouvelles lignes activées. Parmi celles-ci, on trouve la distinction entre grandes régions, en cohérence avec ce qui a été dit précédemment. Par ailleurs, on a retenu le type de répartiteur : un répartiteur placé en hauteur (poteau électrique par exemple)

est plus sensible aux perturbations (vents, intempéries, etc.) qu'un répartiteur placé à l'intérieur d'un local. Aussi, le type de produit peut jouer un rôle : bien que le codage des modalités n'est pas explicité, on pourrait soupçonner des qualités de lignes différentes selon le type de produit souscrit. Enfin, on peut retenir le type d'opérateur, qui peut être un fournisseur ou un autre type d'opérateur.

* * *

Pour la stratification par grandes régions (table 3), on note un léger déséquilibre entre les effectifs mais sans plus. Chaque région a suffisamment d'observations (> 5% de la population disponible) et le nombre de classes est raisonnable pour considérer des strates. De plus, la répartition des censures est équitable entre les régions (entre 65% et 70%).

Table 3 – Fréquence et censures par grandes régions

Region				
Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
IDF	6471	15.12	6471	15.12
NORD-EST	15859	37.05	22330	52.16
OUEST	20477	47.84	42807	100.00

Summary of the Number of Censored and Uncensored Values					
Stratum	Region	Total	Failed	Censored	Percent Censored
1	IDF	5701	1958	3743	65.66
2	NORD−EST	14134	4215	9919	70.18
3	OUEST	17944	5674	12270	68.38
Total		37779	11847	25932	68.64

En ce qui concerne le type de centre de répartiteur (table 4), 70 lignes ne sont pas renseignées (INCONNU). Ces lignes seront donc retirées lors de l'étude des survies par strates pour cette variable. Les fréquences sont équitables, que ce soit en effectif des strates (entre 25% et 45%) ou en pourcentages de censures (toutes les strates sont à 68% environ).

Table 4 – Fréquence et censures par type de centre de répartiteur

PC				
PC	Frequency	Percent	Cumulative Frequency	Cumulative Percent
IMMEUBLE	13030	30.44	13030	30.44
INCONNU	70	0.16	13100	30.60
PC_HAUT	18866	44.07	31966	74.67
PC_SOL	10841	25.33	42807	100.00

Summary of the Number of Censored and Uncensored Values					
Stratum	PC	Total	Failed	Censored	Percent Censored
1	IMMEUBLE	11832	3707	8125	68.67
2	PC_HAUT	16263	5122	11141	68.51
3	PC_SOL	9644	3000	6644	68.89
Total		37739	11829	25910	68.66

Le type de produit (table 5) présente des catégories très peu représentées (ou rares) avec des effectifs marginaux de moins de 2%. La catégorie « P3 » ne représentant que 8% de la population disponible, les catégories rares seront fusionnées à celle-ci. Après fusion, la catégorie mixte « P1P2P3 » représentera un peu plus de 10% et les censures seront également bien distribuées (entre 66% et 77%).

Table 5 – Fréquence et censures par type de produit

Type_produit				
Type_produit	Frequency	Percent	Cumulative Frequency	Cumulative Percent
P1	670	1.57	670	1.57
P2	292	0.68	962	2.25
P3	3447	8.05	4409	10.30
P4	15237	35.59	19646	45.89
P5	23161	54.11	42807	100.00

Summary of the Number of Censored and Uncensored Values					
Stratum	type_produit	Total	Failed	Censored	Percent Censored
1	P1P2P3	3353	786	2567	76.56
2	P4	13626	4078	9548	70.07
3	P5	20800	6983	13817	66.43
Total		37779	11847	25932	68.64

Enfin, pour le type d'opérateur (table 6), il existe également quelques lignes non renseignées (28 lignes) qui seront retirées pour tester la stratification par cette variable catégorielle. Le pourcentage de censures est équivalent entre les strates (64.93% contre 69.06%). En revanche, les strates ne sont pas équilibrées en termes d'effectif (90% contre 10%), il n'y a cependant pas de classes rares (effectif inférieur à 5% de la population disponible).

Table 6 – Fréquence et censures par type d'opérateur

Operateur				
Operateur	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AUTRES	38557	90.07	38557	90.07
FOURNISSEUR	4222	9.86	42779	99.93
INCONNU	28	0.07	42807	100.00

Summary of the Number of Censored and Uncensored Values					
Stratum	Operateur	Total	Failed	Censored	Percent Censored
1	AUTRES	34090	10549	23541	69.06
2	FOURNISSEUR	3667	1286	2381	64.93
Total		37757	11835	25922	68.65

2.2.2 Tests et visualisations graphiques sur la pertinence des strates

L'ensemble des variables précédentes vont être utilisées une par une pour effectuer les tests des rangs exacts. Les résultats sont répertoriés dans les tables suivantes.

Table 7 – Tests d'égalité sur les strates pour REGION

Test of Equality over Strata (region)			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	39.2128	2	<.0001
Wilcoxon	37.5415	2	<.0001
-2Log(LR)	46.2680	2	<.0001

Table 8 – Tests d'égalité sur les strates pour PC

Test of Equality over Strata (pc)			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.5596	2	0.7559
Wilcoxon	0.8735	2	0.6461
-2Log(LR)	0.5802	2	0.7482

Table 9 – Tests d'égalité sur les strates pour TYPE_PRODUIT

Test of Equality over Strata (type_produit)			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	159.5157	2	<.0001
Wilcoxon	164.4483	2	<.0001
-2Log(LR)	197.5427	2	<.0001

Table 10 – Tests d'égalité sur les strates pour OPERATEUR

Test of Equality over Strata (operateur)			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	21.5215	1	<.0001
Wilcoxon	16.4896	1	<.0001
-2Log(LR)	27.1113	1	<.0001

Pour l'ensemble des tests et des variables à l'exception de PC, les résultats conduisent au rejet de l'hypothèse nulle d'équivalence des survies. Par conséquent, pour les variables REGION, TYPE_PRODUIT et OPERATEUR, les strates issues des différentes catégories de ces variables ont des survies significativement différentes. Ces résultats sont appuyés par les visualisations graphiques ci-dessous.

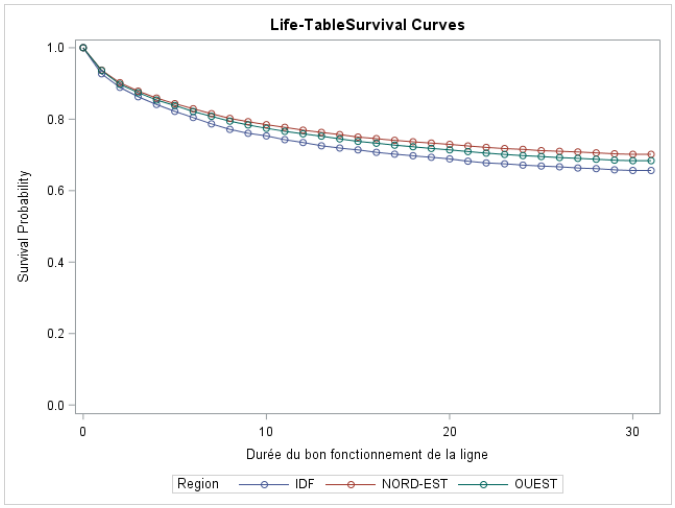

Figure 6 – Survie selon les strates de REGION

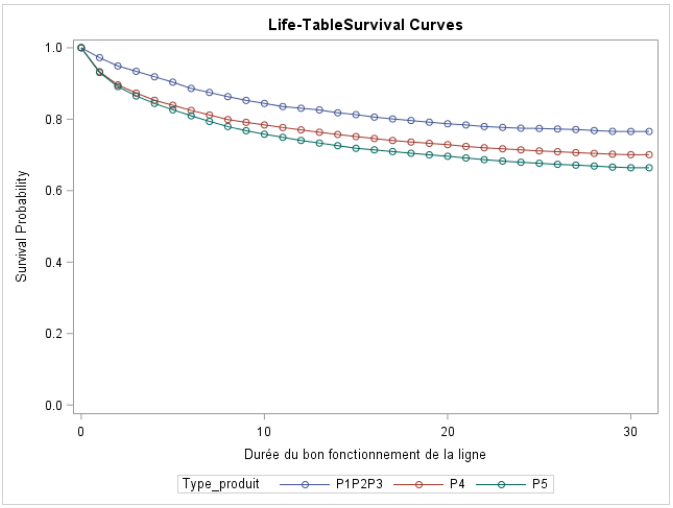

Figure 8 – Survie selon les strates de TYPE_PRODUIT

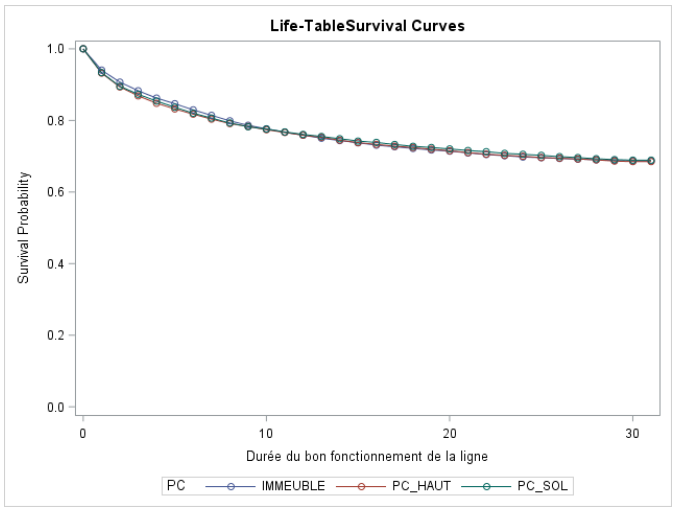

Figure 7 – Survie selon les strates de PC

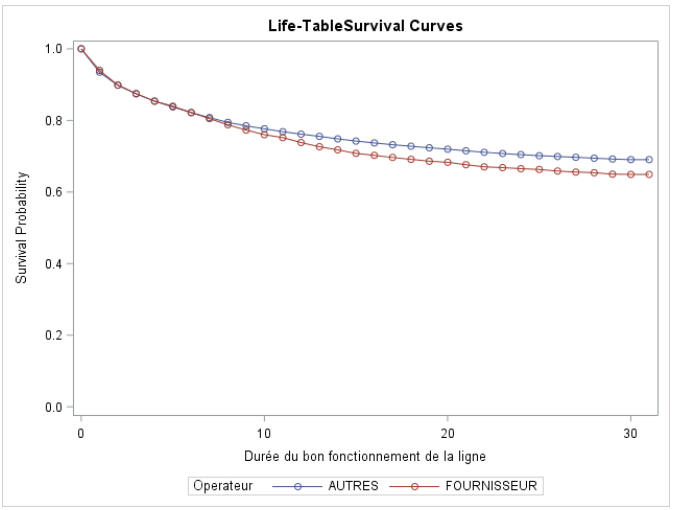

Figure 9 – Survie selon les strates de OPERATEUR

On note dans un premier temps que les survies ne se croisent pas, la puissance des tests précédents n'est donc pas affaiblie. Par ailleurs, on distingue bien la séparation entre les différentes courbes de survie pour toutes les variables à l'exception de PC (type de centre de répartiteur). Au final, les variables catégorielles REGION, TYPE_PRODUIT et OPERATEUR sont gardées en plus de la variable LONGUEUR_LIGNE (variable continue).

Ces variables permettront de construire les modèles paramétriques (section suivante). Selon l'information apportée, ces variables seront retenues ou non à partir des critères d'information comme l'AIC.

3 Approche paramétrique

Il est possible d'introduire des variables explicatives dans la modélisation du risque de survenue d'un dysfonctionnement sur la ligne des clients en supposant une distribution particulière des temps de survie des lignes en question. Par ailleurs, si la distribution privilégiée est correcte alors cela entraine le fait que les estimateurs obtenus sont plus efficients que les estimateurs obtenus par les méthodes non paramétriques. L'un des risques intrinsèque à l'utilisation des méthodes paramétriques correspond à un choix erroné de la distribution supposée, il est donc primordial de s'assurer de la pertinence du choix effectué, notamment par l'intermédiaire de tests statistiques et d'analyses graphiques. Les modèles paramétriques estimés lors de cette étude sont obtenus à partir de la procédure LIFEREG de SAS. Dans cette procédure, on suppose une hypothèse de temps de vie accéléré (Accelerated Failure Time (AFT)).

3.1 Recherche exhaustive des variables explicatives

Dans la classe des modèles AFT, les distributions exponentielle, Weibull, log-normale et log-logistique sont les plus couramment utilisées dans la PROC LIFEREG de SAS. Nous allons donc considérer ces quatre lois dans notre étude. Nous pourrions sélectionner la meilleure modélisation en prenant en compte la log-vraisemblance de chaque modèle. Cependant, il n'est pas optimal de comparer la log-vraisemblance pour des familles de loi différentes. Ainsi, nous allons uniquement nous servir de cet indicateur pour sélectionner les variables explicatives les plus pertinentes pour la construction du meilleur modèle. Plus précisemment, nous allons effectuer une recherche exhaustive sur les variables en comparant les AIC obtenus pour chaque loi étudiée.

3.1.1 Loi exponentielle

La loi exponentielle correspond à un risque instantané qui est constant. En d'autres termes, la probabilité de rencontrer un problème sur une ligne à l'instant t sachant que la ligne n'a pas subi de dysfonctionnement depuis son activation, est la même pour tout t. Les fonctions de densité *f(.)*, de risque *h(.)* et de survie *S(.)* associées à une modélisation pour une loi exponentielle sont respectivement telles que :

$$f(t) = \lambda \cdot \exp(-\lambda t)$$
$$h(t) = \lambda$$
$$S(t) = \exp(-\lambda t)$$

La loi exponentielle est une fonction à un paramètre (λ). L'étude non-paramétrique réalisée dans la première partie de l'étude a mis en évidence l'existence de certaines variables explicatives pertinentes pour expliquer la survenue de panne après l'ouverture d'une ligne. La construction d'un macro fonction SAS a permis de réaliser une analyse exhaustive de la pertinence de ces variables explicatives.

Table 11 – Résultats de la recherche exhaustive des variables explicatives pertinentes avec la loi Exponentielle

exogenes	aic	aicc	bic	loglik	lambda	gamma
region operateur type_produit longueur_ligne	97687.57	97687.57	97747.34	−48836.79	0.015486	.
longueur_ligne region type_produit	97711.26	97711.27	97762.50	−48849.63	0.013487	.
region operateur type_produit	97715.06	97715.06	97766.29	−48851.53	0.017436	.
longueur_ligne operateur type_produit	97737.55	97737.55	97780.24	−48863.77	0.015345	.
type_produit region	97738.70	97738.70	97781.39	−48864.35	0.015186	.
type_produit longueur_ligne	97758.76	97758.76	97792.92	−48875.38	0.013487	.
type_produit operateur	97759.69	97759.69	97793.84	−48875.84	0.017077	.
type_produit	97780.92	97780.93	97806.54	−48887.46	0.015009	.
longueur_ligne region operateur	97832.82	97832.82	97875.51	−48911.41	0.014469	.
longueur_ligne region	97858.65	97858.65	97892.80	−48925.32	0.012496	.
operateur longueur_ligne	97881.60	97881.60	97907.22	−48937.80	0.014208	.
operateur region	97904.09	97904.09	97938.25	−48948.05	0.016071	.
longueur_ligne	97905.08	97905.08	97922.16	−48950.54	0.012379	.
region	97931.51	97931.51	97957.13	−48962.76	0.013836	.
operateur	97949.12	97949.12	97966.20	−48972.56	0.015782	.
	97974.30	97974.30	97982.84	−48986.15	0.013703	.

Les résultats de cette analyse sont présentés dans la table précédente (table 11). On constate que le meilleur modèle au sens de la maximisation de l'AIC est celui qui est composé des variables suivantes : REGION, OPERATEUR, TYPE_PRODUIT et LONGUEUR_LIGNE. Ainsi, les quatre variables explicatives détectées lors de l'analyse non-paramétrique s'avèrent une nouvelle fois pertinente dans la construction d'un modèle paramétrique avec la loi exponentielle.

3.1.2 Loi de Weibull

La loi de Weibull est une généralisation de la loi exponentielle. Il s'agit d'une modélisation avec deux paramètres. Le premier paramètre correspond à l'échelle (λ) et le second paramètre correspond à la forme (γ). Les fonctions de densité *f(.)*, de risque *h(.)* et de survie *S(.)* associées à une modélisation pour une loi de Weibull sont respectivement telles que :

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1}\exp(-\lambda t^{\gamma})$$

$$h(t) = \lambda\gamma t^{\gamma-1}$$

$$S(t) = \exp(-\lambda t^{\gamma})$$

Table 12 – Résultats de la recherche exhaustive des variables explicatives pertinentes avec la loi de Weibull

exogenes	aic	aicc	bic	loglik	lambda	gamma
region operateur type_produit longueur_ligne	89167.66	89167.66	89235.97	−44575.83	0.079746	0.49322
longueur_ligne region type_produit	89186.62	89186.62	89246.39	−44586.31	0.070427	0.49313
region operateur type_produit	89191.06	89191.07	89250.84	−44588.53	0.089051	0.49311
type_produit region	89209.96	89209.96	89261.19	−44598.98	0.078648	0.49301
longueur_ligne operateur type_produit	89210.05	89210.05	89261.28	−44599.03	0.079120	0.49303
type_produit longueur_ligne	89227.00	89227.00	89269.69	−44608.50	0.070452	0.49294
type_produit operateur	89228.90	89228.90	89271.59	−44609.45	0.087409	0.49294
type_produit	89245.85	89245.85	89280.00	−44618.92	0.077832	0.49285
longueur_ligne region operateur	89296.63	89296.63	89347.86	−44642.32	0.074995	0.49274
longueur_ligne region	89317.44	89317.44	89360.13	−44653.72	0.065717	0.49263
operateur longueur_ligne	89338.02	89338.02	89372.18	−44665.01	0.073803	0.49255
longueur_ligne	89356.93	89356.93	89382.54	−44675.46	0.065184	0.49246
operateur region	89357.51	89357.51	89400.20	−44673.75	0.082725	0.49247
region	89379.59	89379.59	89413.75	−44685.80	0.072291	0.49236
operateur	89395.73	89395.73	89421.34	−44694.86	0.081415	0.49230
	89416.00	89416.00	89433.08	−44706.00	0.071691	0.49220

De la même manière que pour la loi exponentielle, on peut constater les résultats de la recherche exhaustive des variables explicatives pour la loi de Weibull. Ainsi, et à l'instar de la loi exponentielle, le meilleur modèle au sens de la maximisation de l'AIC est celui qui est composé des quatre variables explicatives retenues lors de l'étude non-paramétrique.

3.1.3 Loi log-logistique

La modélisation avec la loi log-logistique est, comme la loi de Weibull, un modèle avec deux paramètres. De la même manière, il y a un paramètre d'échelle (λ) et un paramètre de forme (γ). Les fonctions de densité *f(.)*, de risque *h(.)* et de survie *S(.)* associées à une modélisation pour une loi log-logistque sont respectivement telles que :

$$f(t) = \frac{\lambda\gamma\exp(\gamma-1)}{(1+\lambda t\gamma)^2}$$

$$h(t) = \frac{\lambda\gamma\exp(\gamma-1)}{1+\lambda t\gamma}$$

$$S(t) = \frac{1}{1+\lambda t\gamma}$$

Table 13 – Résultats de la recherche exhaustive des variables explicatives pertinentes avec la loi log-logistique

exogenes	aic	aicc	bic	loglik	lambda	gamma
region operateur type_produit longueur_ligne	88828.94	88828.94	88897.25	−44406.47	0.082238	0.54609
longueur_ligne region type_produit	88842.98	88842.98	88902.75	−44414.49	0.072268	0.54595
region operateur type_produit	88852.34	88852.34	88912.11	−44419.17	0.093774	0.54585
type_produit region	88866.34	88866.35	88917.58	−44427.17	0.082411	0.54571
longueur_ligne operateur type_produit	88869.37	88869.37	88920.60	−44428.68	0.081638	0.54571
type_produit longueur_ligne	88881.91	88881.91	88924.60	−44435.95	0.072355	0.54558
type_produit operateur	88888.51	88888.51	88931.20	−44439.25	0.091996	0.54551
type_produit	88901.07	88901.07	88935.22	−44446.53	0.081534	0.54538
longueur_ligne region operateur	88967.38	88967.38	89018.61	−44477.69	0.076762	0.54496
longueur_ligne region	88983.20	88983.20	89025.89	−44486.60	0.066841	0.54481
operateur longueur_ligne	89006.71	89006.71	89040.86	−44499.35	0.075546	0.54460
longueur_ligne	89021.11	89021.11	89046.72	−44507.55	0.066319	0.54446
operateur region	89026.16	89026.16	89068.86	−44508.08	0.086092	0.54442
region	89043.03	89043.03	89077.19	−44517.52	0.074740	0.54426
operateur	89062.71	89062.71	89088.33	−44528.36	0.084759	0.54408
	89078.24	89078.24	89095.32	−44537.12	0.074162	0.54393

La recherche du meilleur modèle au sens de la maximisation de l'AIC pour la loi log-logistique est le même que celui retenu pour les lois exponentielles et de Weibull. En effet, on constate dans la tableau ci-dessus (table 13) que le modèle retenu pour la loi log-logistique est celui qui est construit avec les variables explicatives suivantes : REGION, OPERATEUR, TYPE_PRODUIT et LONGUEUR_LIGNE.

3.1.4 Loi log-normale

La dernière loi considérée dans cette étude du risque de dysfonctionnement d'une ligne internet est la loi log-normale. Il s'agit une nouvelle fois d'une loi composée de deux paramètres. Le paramètre d'échelle (λ) correspond au paramètre de moyenne (μ) et le paramètre de forme (γ) correspond au paramètre de variance (σ) de la loi. Les fonctions de densité *f(.)*, de risque *h(.)* et de survie *S(.)* associées à une modélisation pour une loi log-normale sont respectivement telles que :

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp(-\frac{1}{t\sigma^2}(\ln(t) - \mu)^2)$$

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

$$S(t) = 1 - \Phi\left(\frac{\ln(t) - \mu}{\sigma}\right)$$

avec Φ la fonction de répartition de la loi normale centrée-réduite : $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$ pour tout $x \in \mathbb{R}$.

La recherche du meilleur modèle concernant la loi log-normale nous amène à considérer les quatre variables retenue lors de l'étude non-paramétrique. En effet, on constate sur la table ci-après (table 14) que le modèle composé des variables explicatives REGION, OPERATEUR, TYPE_PRODUIT et LONGUEUR_LIGNE est celui qui maximise l'AIC.

Table 14 – Résultats de la recherche exhaustive des variables explicatives pertinentes avec la loi log-normale

exogenes	aic	aicc	bic	loglik	lambda	gamma
region operateur type_produit longueur_ligne	88296.60	88296.61	88364.91	−44140.30	4.84028	3.35122
longueur_ligne region type_produit	88303.81	88303.81	88363.58	−44144.90	5.03320	3.35105
region operateur type_produit	88317.96	88317.96	88377.73	−44151.98	4.59948	3.35266
type_produit region	88325.04	88325.04	88376.27	−44156.52	4.79181	3.35249
longueur_ligne operateur type_produit	88332.81	88332.81	88384.04	−44160.40	4.85362	3.35338
type_produit longueur_ligne	88339.00	88339.00	88381.70	−44164.50	5.03205	3.35320
type_produit operateur	88350.44	88350.44	88393.13	−44170.22	4.63271	3.35460
type_produit	88356.56	88356.56	88390.71	−44174.28	4.81081	3.35442
longueur_ligne region operateur	88442.74	88442.74	88493.97	−44215.37	4.97353	3.36185
longueur_ligne region	88451.20	88451.20	88493.90	−44220.60	5.18300	3.36170
operateur longueur_ligne	88477.85	88477.85	88512.00	−44234.92	5.00094	3.36389
longueur_ligne	88485.37	88485.37	88510.99	−44239.69	5.19689	3.36374
operateur region	88492.66	88492.66	88535.35	−44241.33	4.77384	3.36434
region	88501.64	88501.64	88535.79	−44246.82	4.98757	3.36421
operateur	88525.67	88525.67	88551.28	−44259.83	4.79937	3.36630
	88533.77	88533.77	88550.84	−44264.88	5.00018	3.36616

* * *

On peut voir que les résultats concernant la pertinence de ces quatre variables explicatives sont robustes dans la mesure où elles semblent apporter des informations significatives concernant le potentiel dysfonctionnement d'une ligne internet pour les différentes lois retenues lors de l'étude. Par ailleurs, on peut voir qu'une hiérarchisation de l'importance des variables dans la construction du modèle par la loi log-normale est envisageable. Le deuxième meilleur modèle correspond au meilleur modèle auquel on aurait retiré la variable OPERATEUR. Le troisième meilleur modèle correspond au meilleur modèle amputé de la variable LONGUEUR_LIGNE. Cela peut donc nous amener à penser que les deux variables les plus importantes sont les variables REGION et TYPE_PRODUIT, suivies respectivement des variables OPERATEUR et LONGUEUR_LIGNE.

Maintenant qu'on dispose de covariates pertinentes à l'étude, il reste à identifier la loi la plus adéquate afin d'estimer le modèle final pour cette approche paramétrique.

3.2 Hazard plotting : sélection de la loi adéquate

Lorsqu'on compare des lois de familles de lois différentes, la méthode du maximum de vraisemblance n'est plus valable. Pour contourner ce problème, l'utilisation de la méthode graphique de hazard plotting est adéquate. Chaque loi paramétrique a une fonction de survie qui s'exprime en fonction du temps :

$$S(t) = F(\theta, t)$$

Afin d'effectuer le hazard plotting, il est essentiel de modifier la fonction de survie de façon à obtenir une équation de type :

$$g(S(t)) = h(\theta) \cdot \ln(t) + s(\theta)$$

Dès lors, on estime h et s par une régression linéaire puis on calcule la statistique du R^2 associée qui correspond au coefficient de détermination. De cette manière, on peut sélectionner la loi qui maximise ce coefficient de détermination. Par ailleurs, graphiquement, on sélectionne la loi qui possède la fonction de survie estimée qui est la plus proche de la fonction de survie empirique (figure 10).

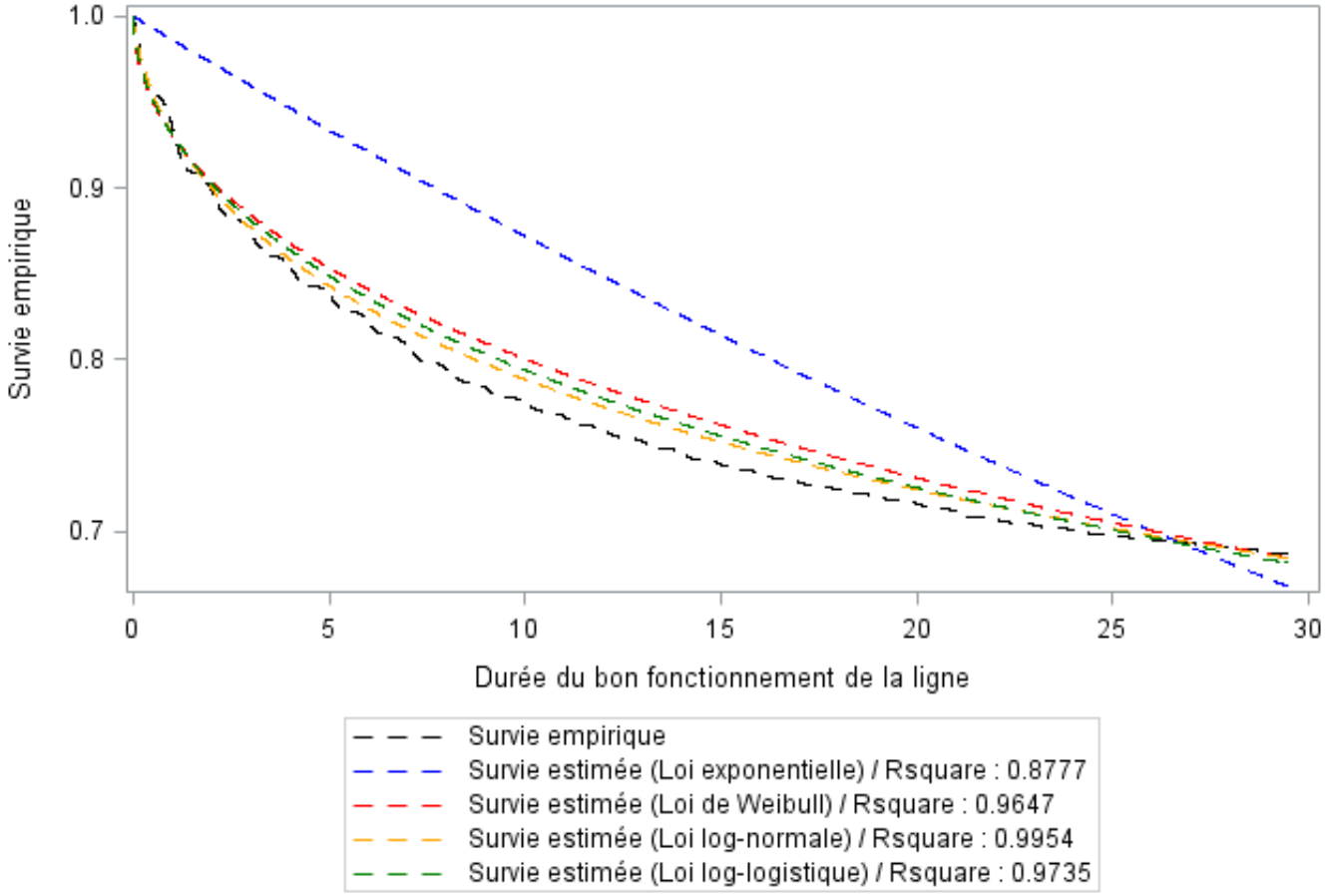

Figure 10 – Méthode du hazard plotting pour la sélection de la loi

La méthode du hazard plotting nous amène à considérer la loi log-normale. En effet, graphiquement, on constate que la fonction de survie estimée pour la loi log-normale correspond à celle qui est la proche de la fonction de survie empirique. En outre, on peut voir que la maximisation du coefficient de détermination (R^2) est réalisée avec la loi log-normale. Avant la censure des 30 jours, on peut voir que 30% des lignes internet étudiées ont rencontré un dysfonctionnement. On peut constater que le phénomène de dysfonctionnement d'une ligne internet est relativement récent après l'ouverture de celle-ci. En effet, environ 15% des lignées étudiées ont rencontré un problème durant les cinq premiers jours après l'ouverture.

3.3 Estimation paramétrique (AFT) du modèle retenu

Les résultats de l'estimation paramétrique (AFT) du modèle composé des variables explicatives REGION, TYPE_PRODUIT, OPERATEUR et LONGUEUR_LIGNE par la loi log-normale sont présentés dans le tableau ci-après (table 15).

Lorsqu'on s'intéresse à la significativité individuelle de chacune des variables explicatives du modèle, on constate que les variables REGION, LONGUEUR_LIGNE, TYPE_PRODUIT et OPERATEUR sont toutes les quatre statistiquement pertinentes pour un risque de première espèce de 5%.

Nous pouvons donc procéder à l'interprétation des paramètres estimés. Dans le cas d'une variable explicative qualitative, l'interprétation est réalisée par rapport à une modalité de référence. La variable explicative LONGUEUR_LIGNE est continue et n'a donc pas de modalité de référence. En revanche, la variable OPERATEUR est qualitative et sa modalité de référence est « Fournisseur ». De la même manière, la variable REGION est une variable explicative qualitative qui a pour référence la modalité « Ouest » et la variable TYPE_PRODUIT a pour modalité de référence « P1P2P3 ».

Table 15 – Estimation du meilleur modèle retenu avec la loi log-normale

Analysis of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Std Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	6.2013	0.0944	6.0164	6.3863	4318.39	<.0001
Region	IDF	1	−0.3989	0.0639	−0.5240	−0.2737	39.02	<.0001
Region	OUEST	1	−0.1664	0.0465	−0.2575	−0.0754	12.83	0.0003
Region	NORD−EST	0	0.0000	.	.	.	.	.
Longueur_ligne		1	−0.0001	0.0000	−0.0001	−0.0001	23.41	<.0001
type_produit	P4	1	−0.9044	0.0875	−1.0758	−0.7329	106.89	<.0001
type_produit	P5	1	−0.9816	0.0824	−1.1432	−0.8200	141.76	<.0001
type_produit	P1P2P3	0	0.0000	.	.	.	.	.
Operateur	FOURNISSEUR	1	−0.2130	0.0701	−0.3503	−0.0756	9.24	0.0024
Operateur	AUTRES	0	0.0000	.	.	.	.	.
Scale		1	3.3512	0.0254	3.3019	3.4013		

Variable REGION :

On peut voir que le signe de la modalité « IDF » de la variable REGION est négatif, cela implique que le risque de dysfonctionnement d'une ligne internet est plus élevé pour un client résidant en Ile-de-France par rapport à un client qui réside dans le Nord-Est de la France. En effet, la durée de vie moyenne d'une ligne internet pour un client résidant en Ile-de-France serait environ 32,9% plus courte par rapport à un client possédant une ligne dans le Nord-Est de la France. De même, on remarque que la modalité « Ouest » possède un coefficient estimé négatif. Ainsi, toutes choses égales par ailleurs, le fait d'habiter à l'Ouest de la France implique un risque de dysfonctionnement d'une ligne internet plus faible que pour un client qui réside dans le Nord-Est de la France. En effet, la durée de vie moyenne d'une ligne internet pour un client résident dans l'Ouest de la France serait environ 15,3% plus courte que celle d'une ligne internet pour un client résidant dans le le Nord-Est de la France.

Variable LONGUEUR_LIGNE :

Le coefficient estimé associé à la variable continue LONGUEUR_LIGNE est négatif. Cela implique que, *ceteris paribus*, plus la longueur de la ligne internet va être longue et plus le temps de survenue d'un dysfonctionnement sera court. En d'autres termes, plus la longueur de la ligne d'internet d'un client sera grande et plus le risque de survenue d'un problème de fonctionnement sera élevé. En effet, après l'ouverture d'une ligne internet, toutes choses égales par ailleurs, si la ligne était plus longue d'un kilomètre, alors le temps jusqu'au dysfonctionnement de la ligne diminue de 9,5%.

Variable TYPE_PRODUIT :

Les coefficients estimés pour les types « P4 » et « P5 » sont tous les deux négatifs. Par conséquent, un nouvel abonné qui souscrit à l'un de ces deux types de produits aura une ligne internet ou téléphonique de durée de vie plus courte que s'il avait souscrit à un produit de type « P1 », « P2 » ou « P3 ». On aurait respectivement une réduction de la durée de vie de 59,5% et 62.5% pour les types « P4 » et « P5 » par rapport à un type « P1P2P3 », toutes choses égales par ailleurs.

Variable OPERATEUR :

Le coefficient estimé associé à la modalité « Autres » de la variable qualitative OPERATEUR est négatif. Ainsi, toutes choses égales par ailleurs, si l'opérateur n'est pas un fournisseur, cela implique que le risque de dysfonctionnement est plus élevé. En effet, la durée de vie moyenne d'une ligne internet pour un client possédant un opérateur qui est un fournisseur serait environ 19,2% plus courte que celle d'une ligne internet pour un client possédant un opérateur qui n'est pas un fournisseur.

Finalement, l'estimation paramétrique nous amène à considérer le risque de dysfonctionnement d'une ligne internet comme étant d'autant plus grand que la ligne internet est longue et qu'elle appartienne à un client d'Ile-de-France possédant un opérateur qui est un fournisseur.

4 Approche semi-paramétrique

On utilise le modèle de Cox qui permet d'évaluer l'effet de covariables sur la durée de vie sans émettre d'hypothèses sur la distribution des temps de survie.

On estime la fonction de risque comme le produit du risque instantané « de base » identique pour tous les individus (fonction uniquement du temps) et d'une fonction qui ne dépend que des caractéristiques individuelles.

Une fois les estimateurs des coefficients obtenus, on détermine un estimateur du risque de base par une approche non-paramétrique (Kaplan-Meier ici). La constante est incluse dans cette fonction de risque de base.

Le modèle de Cox repose sur 2 hypothèses :

1. Risques proportionnels (PH)
2. Log-linéarité entre la fonction de risque et les covariables

Après estimation du modèle final, on s'assurera que ces deux hypothèses soient vérifiées.

4.1 Modèles avec une seule covariable à la fois

Les variables retenues dans la section précédente (estimation paramétrique) sont : la région, le type de produit, la longueur de la ligne et l'opérateur. Afin de s'assurer de la significativité de chacune de ces variables, des modèles avec l'utilisation d'une seule covariable à la fois seront construits et étudiés.

4.1.1 Région

On estime le modèle de Cox pour variable explicative seule la variable REGION avec pour modalité de référence la région Nord-Est.

Table 16 – Tests sur le modèle avec REGION

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	245284.32	245245.39
AIC	245284.32	245249.39
SBC	245284.32	245264.14

Testing Global Null Hypothesis : BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	38.9294	2	<.0001
Score	39.5546	2	<.0001
Wald	39.4981	2	<.0001

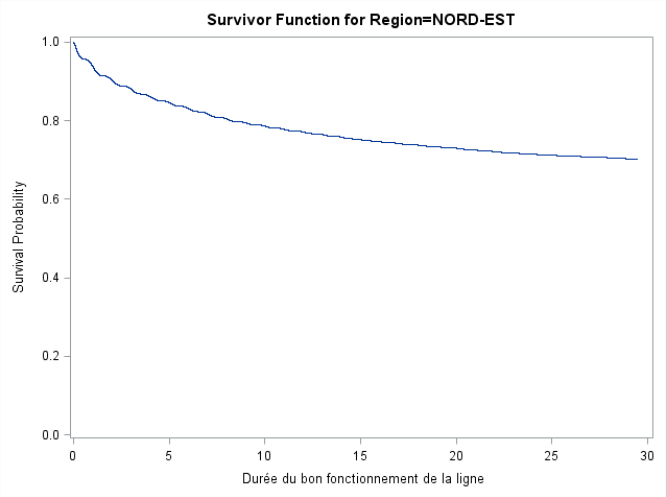

Figure 11 – Fonction de survie pour l'individu de référence (Region = NORD-EST)

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	
Region	IDF	1	0.17079	0.02736	38.9710	<.0001	1.186	
Region	OUEST	1	0.06760	0.02035	11.0366	0.0009	1.070	

Le modèle est globalement satisfaisant, ainsi au moins un des coefficients est significativement différent de 0. Par ailleurs, les tests individuels sont également significatifs, on peut alors interpréter les rapports de risques.
Ici, on peut noter qu'une ligne activée en Île-de-France a 18.60% plus de risque de présenter un dysfonctionnement dans un délai de 30 jours maximum qu'une ligne activée dans la région du Nord-Est. Ce pourcentage est de 7% pour une ligne activée dans la région de l'Ouest.

Ceci confirme l'intuition de départ (introduction), la localisation géographique est un facteur jouant sur le risque de dysfonctionnement et ce risque est plus élevé en Île-de-France et à l'Ouest qu'au Nord-Est.

4.1.2 Type de produit

On estime le modèle de Cox pour variable explicative seule la variable type_produit avec pour modalité de référence la modalité regroupant les types P1, P2 et P3.

Table 17 – Tests sur le modèle avec TYPE_PRODUIT

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	245284.32	245113.83
AIC	245284.32	245117.83
SBC	245284.32	245132.58

Testing Global Null Hypothesis : BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	170.4897	2	<.0001
Score	159.7399	2	<.0001
Wald	157.4502	2	<.0001

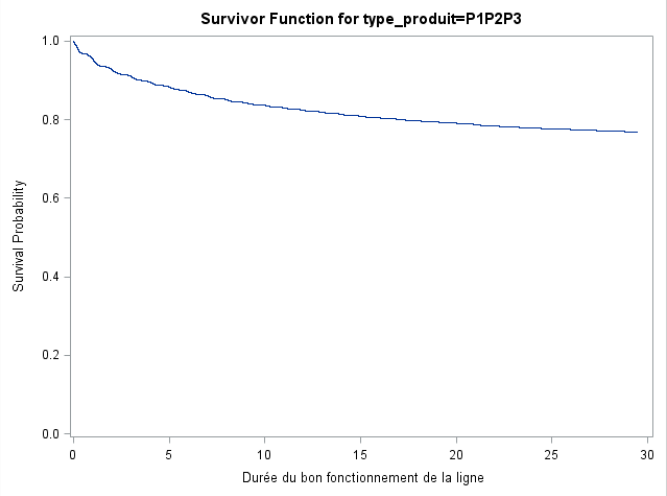

Figure 12 – Fonction de survie pour l'individu de référence (type_produit = P1P2P3)

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	
type_produit	P4	1	0.30230	0.03896	60.1940	<.0001	1.353	
type_produit	P5	1	0.43768	0.03763	135.3008	<.0001	1.549	

À nouveau, le modèle est globalement satisfaisant. Ainsi au moins un des coefficients est significativement différent de 0. De plus, les tests individuels sont significatifs, ainsi on peut interpréter les rapports de risques.

La modalité de référence est la catégorie « P1P2P3 » qui regroupe les trois premiers types de produits (« P1 », « P2 » et « P3 »). Ainsi, une ligne activée associée à un produit de type « P4 » est 35.3% plus risquée qu'une ligne associée à un produit du type « P1 » ou « P2 » ou « P3 ». Ce rapport de risque est encore plus important (54.9%) pour les produits de type « P5 ».

Ces différences de risques peuvent s'expliquer par la nature du produit. Certains types de produits sont peut être plus risqués ou instables que d'autres (extension d'une ligne nécessaire ? différences entre une ligne internet et une ligne téléphonique ? débit fourni ?). Sans la signification des modalités codées, ceci n'est que hypothèses.

4.1.3 Opérateur

On estime le modèle de Cox pour variable explicative seule la variable operateur avec pour modalité de référence la modalité « autres fournisseurs ».

Cette variable est intéressante pour le fournisseur. En effet, selon ce que désigne la variable OPERATEUR, l'étude des coefficients estimés permettra de savoir l'effet d'acteurs tiers sur le risque de dysfonctionnement (la variable désigne-t-elle l'entité qui exploite les lignes ? S'agit-il de fournisseurs de l'entreprise ? Sous-traitance de l'installation des lignes ?).

Table 18 – Tests sur le modèle avec OPERATEUR

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	245284.32	245263.58
AIC	245284.32	245265.58
SBC	245284.32	245272.96

Testing Global Null Hypothesis : BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	20.7343	1	<.0001
Score	21.5212	1	<.0001
Wald	21.4954	1	<.0001

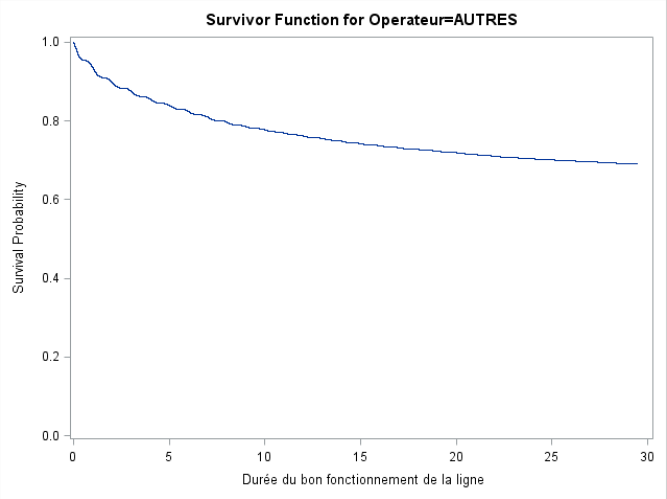

Figure 13 – Fonction de survie pour l'individu de référence (operateur = AUTRES)

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	
Operateur	FOURNISSEUR	1	0.13694	0.02954	21.4954	<.0001	1.147	

Les tests globaux montrent qu'au moins un des coefficients est signficativement différent de 0. Puisqu'on ne dispose que d'un seul paramètre, le test individuel correspond aux tests globaux.
Lorsque la variable OPERATEUR prend la valeur « autres », le risque n'est pas plus élevé. Au contraire, s'il s'agit du fournisseur, la ligne a 14.7% plus de risques d'être dysfonctionnel dans un délai de 30 jours.

4.1.4 Longueur de la ligne

On estime le modèle de Cox pour variable explicative seule la variable indiquant la longueur de la ligne.

Table 19 – Tests sur le modèle avec LONGUEUR_LIGNE

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	245284.32	245225.14
AIC	245284.32	245227.14
SBC	245284.32	245234.52

Testing Global Null Hypothesis : BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	59.1752	1	<.0001
Score	60.7948	1	<.0001
Wald	60.7515	1	<.0001

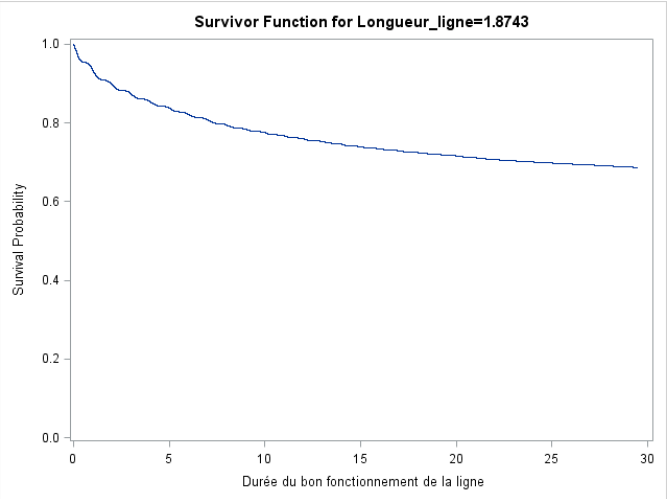

Figure 14 – Fonction de survie pour l'individu de référence (longueur_ligne = 1.8743 km)

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	
Longueur_ligne	1	0.04869	0.00625	60.7515	<.0001	1.050	

Plus la longueur de la ligne est élevée et plus le risque de dysfonctionnement augmente. Pour donner une idée, une ligne longue de 2.87 km a 5% plus de risques de présenter un dysfonctionnement dans un délai de 30 jours par rapport à une ligne de longueur 1.87 km (valeur de référence).

Il reste cependant à vérifier si cet accroissement est proportionnel, cest-à-dire qu'il est identique pour toute longueur de départ.

4.2 Modèles avec l'ensemble des variables retenues

Le modèle aura pour covariables : la région, le type de produit, l'opérateur et la longueur de la ligne.

Table 20 – Tests sur le modèle avec toutes les covariables

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	245284.32	245031.17
AIC	245284.32	245043.17
SBC	245284.32	245087.44

Testing Global Null Hypothesis : BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	253.1474	6	<.0001
Score	244.0370	6	<.0001
Wald	241.7184	6	<.0001

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Region	2	45.1613	<.0001
Longueur_ligne	1	25.2628	<.0001
type_produit	2	116.4588	<.0001
Operateur	1	20.1059	<.0001

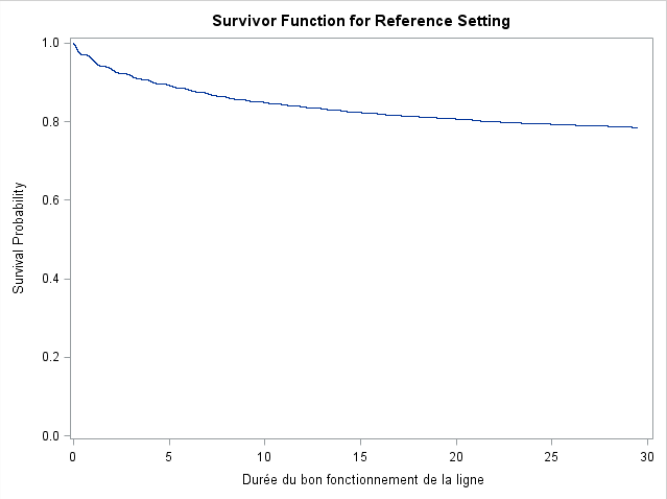

Figure 15 – Fonction de survie pour l'individu de référence (region = "NORD-EST", longueur_ligne = 1.8743 km, type_produit = "P1P2P3", operateur = "AUTRES")

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Param. Estimate	Std. Error	Chi-Square	Pr > ChiSq	Hazard Ratio	
Region	IDF	1	0.18117	0.02748	43.4615	<.0001	1.199	
Region	OUEST	1	0.08146	0.02040	15.9502	<.0001	1.085	
Longueur_ligne		1	0.04111	0.00818	25.2628	<.0001	1.042	
type_produit	P4	1	0.36147	0.04046	79.8224	<.0001	1.435	
type_produit	P5	1	0.40949	0.03803	115.9493	<.0001	1.506	
Operateur	FOURNISSEUR	1	0.13271	0.02960	20.1059	<.0001	1.142	

Dans un premier temps, on observe une baisse du critère d'information d'Akaike (AIC) pour le modèle global par rapport aux modèles avec les variables seules, on en déduit que cette approche est cohérente.
Dans un second temps, le modèle est globalement satisfaisant et les coefficients seuls sont tous significatifs d'après les tests de significativité globale et individuelle.

Il reste cependant à vérifier les deux hypothèses du modèle de Cox (log-linéarité et proportionnalité des risques) avant de pourvoir interpréter les coefficients et les rapports de risques.
Sous les hypothèses du modèle de Cox, une augmentation de la longueur de la ligne de 1 kilomètre entraîne une augmentation du risque de 4.2% et cette hausse du risque instantané est identique de 1 à 2 km comme de 4 à 5 km, ce qui sera testé par la suite. La courbe de référence s'applique pour une longueur de ligne de 1.87 km.

4.3 Vérification des hypothèses du modèle de Cox

4.3.1 Test de log-linéarité des covariables : résidus de martingale

Le modèle retenu ne contient qu'une seule variable continue donc la question de la forme fonctionnelle ne se pose que pour la variable longueur de la ligne.

Pour cela nous allons observer les résidus de martingales. L'idée étant d'observer le nombre d'évenements prédits par le modéle de Cox sans la variable continue à étudier par rapport à la réalité.

La forme fonctionnelle qui lie les résidus de martingale obtenus et la covariable est celle qui devrait être utilisée dans le modèle.

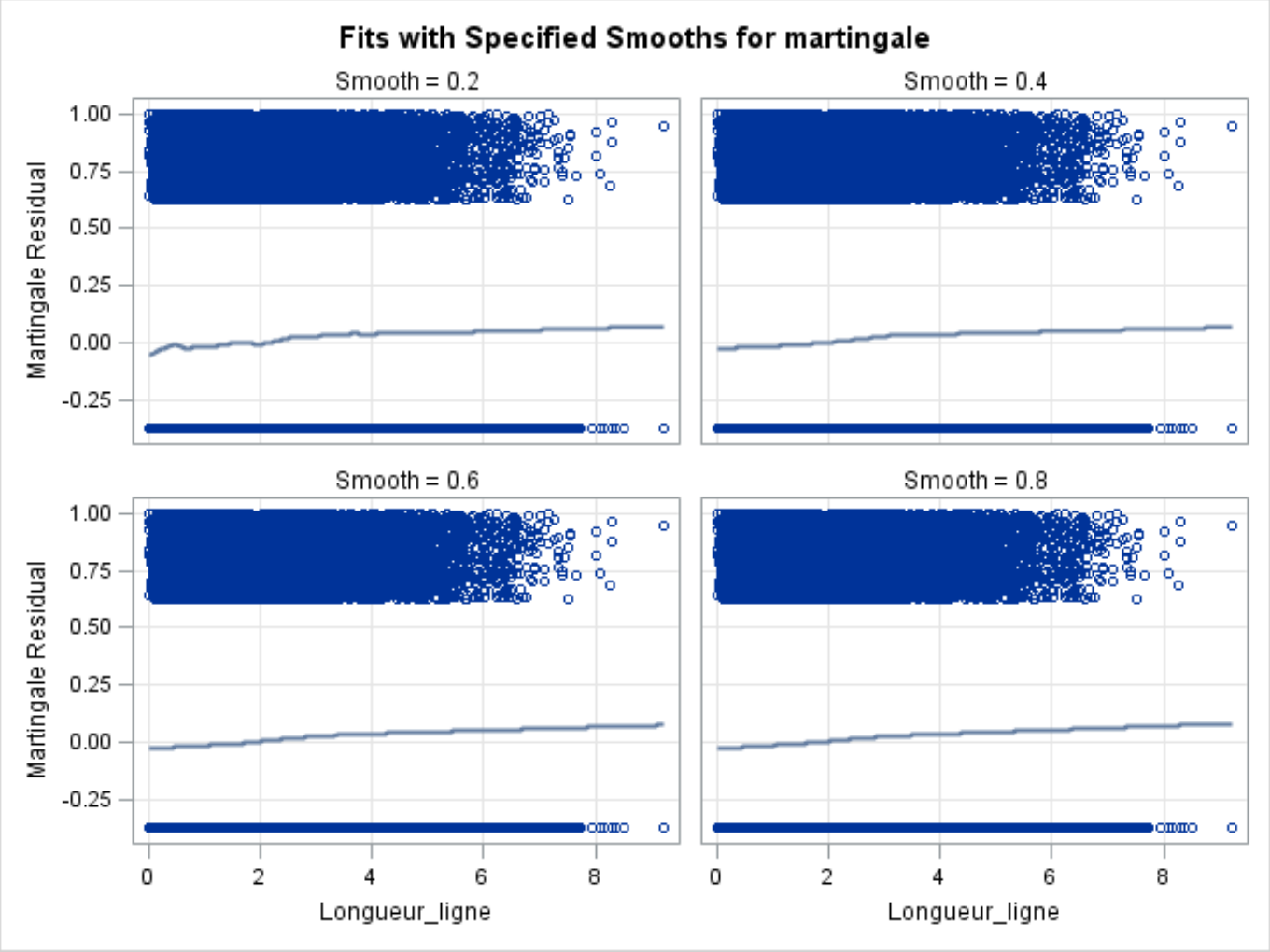

Figure 16 – Résidus de martingale par rapport à la longueur de la ligne

La forme obtenue de la courbe de régression s'approche d'une droite. Ainsi, une modélisation linéaire semble adaptée bien qu'on puisse tester une transformation en logarithme.

On peut confirmer notre analyse en observant les résidus de martingale cumulatifs par rapport à la covariable longueur de la ligne (figure 17).

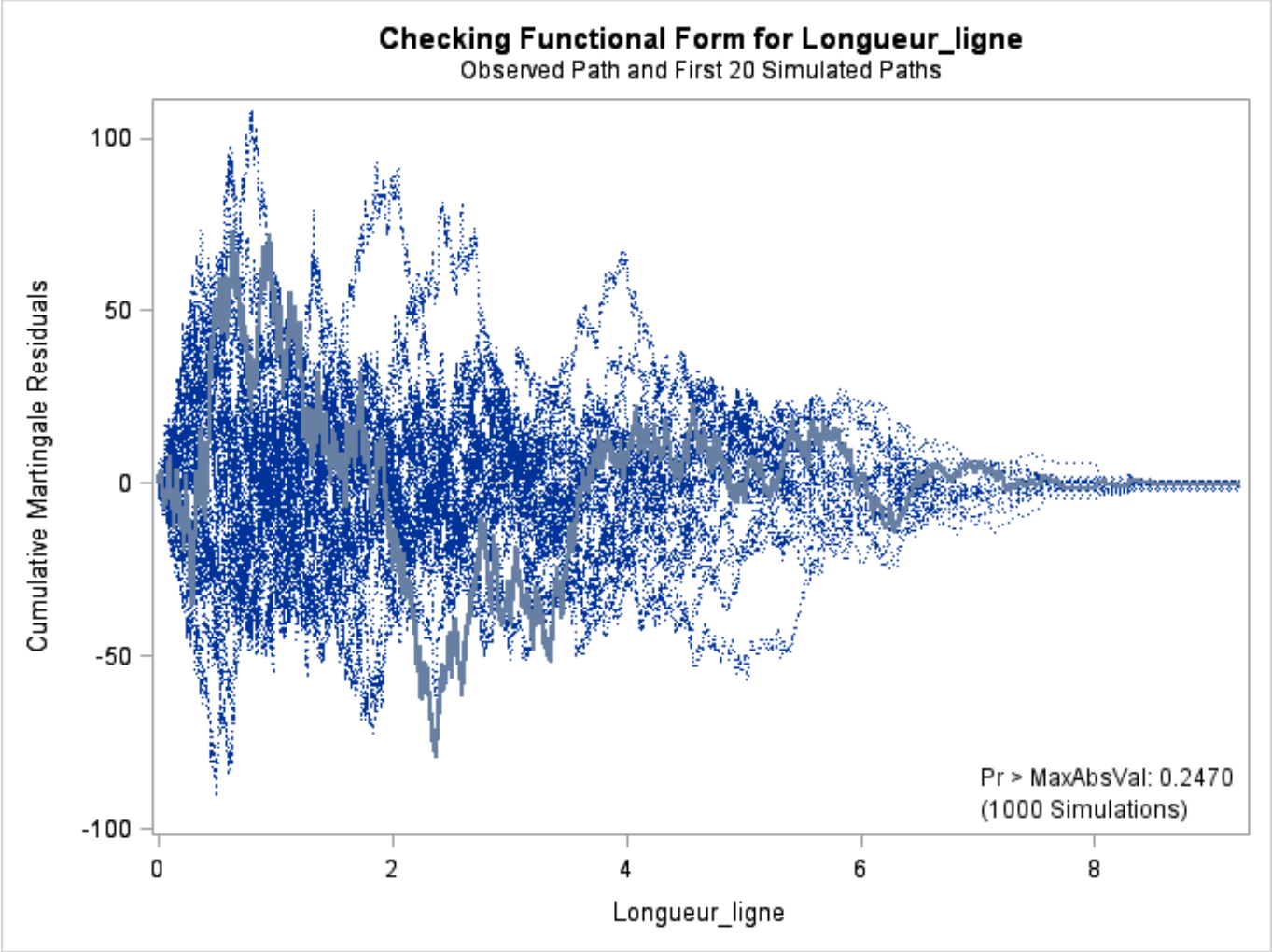

Figure 17 – Simulations des résidus de martingale cumulatifs

Supremum Test for Functional Form				
Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
Longueur_ligne	79.3929	1000	26271121	0.2470

Les résidus de martigale cumulatifs restent dans les limites établies par les 20 lignes pointilées qui représentent les résidus simulés attendus sous l'hypothèse nulle.
Par conséquent, Les résidus ne diffèrent pas significativement des simulations comme en atteste la p-value du test de supremum de Kolmogorov de 0.247, on ne rejette pas l'hypothèse nulle.
L'hypothèse de log-linéarité est vérifiée.

4.3.2 Test de proportionnalité des risques

Pour les variables catégorielles, l'hypothèse de proportionnalité des risques est vérifiée lorsque les courbes de survie estimées par la méthode de Kaplan-Meier sont parallèles. Cette hypothèse a déjà été vérifiée pour les variables catégorielles (figures 6, 8 et 9).

Pour vérifier cette hypothèse pour la variable longueur de la ligne, nous allons utiliser les résidus de Schoenfeld.
Le résidu de Schoenfeld d'un individu i pour la covariable x_k est la différence entre la valeur prise par la covariable k pour l'individu i à la date t_i et la moyenne des valeurs prises par les individus encore à risque à cette même durée.
L'hypothèse nulle est la constance des coefficients dans le temps.

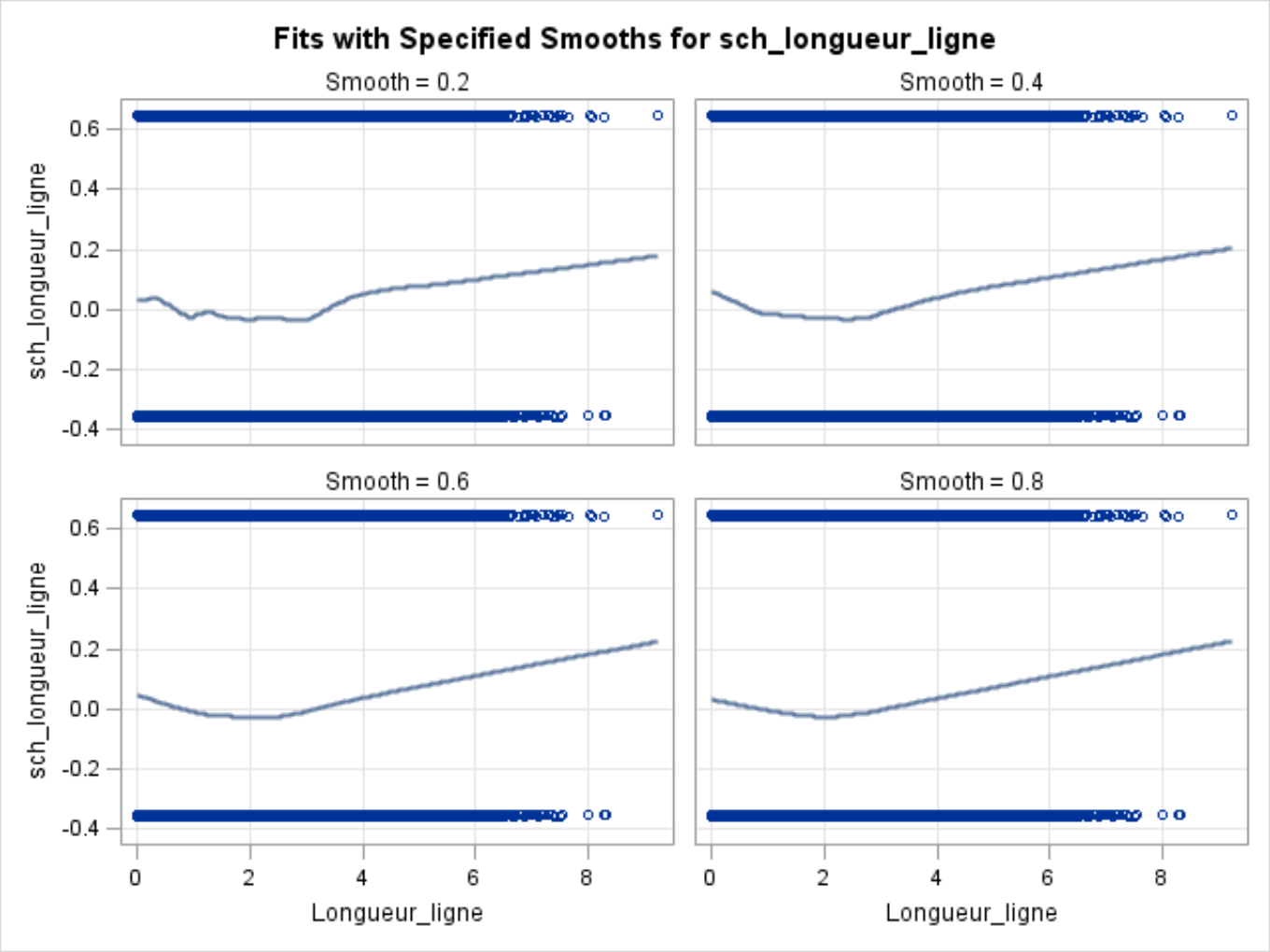

Figure 18 – Résidus de Schoenfeld

On observe que les résidus présentent une rupture à 2000 m de longueur avec une tendance croissante, ce qui suggère un effet non proportionnel.

Pour contourner ce problème, nous avons choisi de discrétiser la variable en 3 classes :

1. longueur de la ligne < 2000 m
2. 2000 m $\leq$ longueur de la ligne < 4500 m
3. longueur de la ligne ≥ 4500 m.

Ces classes ont été choisies afin d'obtenir des classes relativement équilibrées et en prenant compte de la rupture à 2000 mètres. On vérifie à nouveau l'hypothèse de proportionnalité des risques en utilisant cette fois-ci la représentation graphique des fonctions de survie estimées par la méthode de Kaplan-Meier puisque la variable est dorénavant catégorielle.

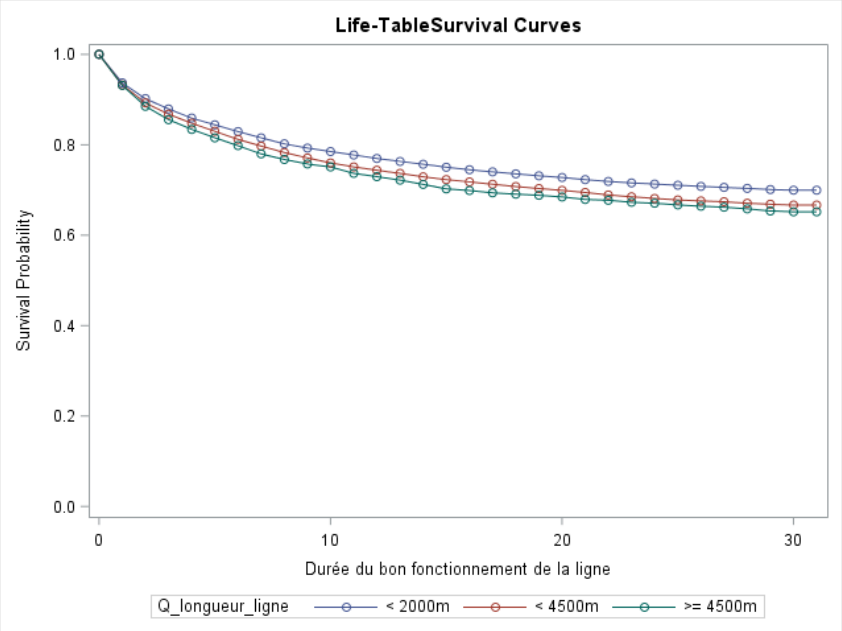

Figure 19 – Stratification à partir de la longueur de la ligne (méthode KM)

Le modèle obtenu vérifie l'hypothèse de proportionnalité des risques pour la nouvelle variable discrétisée, les différentes courbes étant parallèles (figure 19).
L'hypothèse de proportionnalité des risques est vérifiée pour l'ensemble des covariables.

4.4 Modèle final

On peut donc interpréter les rapports de risque par rapport au site de référence à savoir la ligne située en région Nord-Est, concernant un produit P1, P2 ou P3 avec un autre opérateur que l'entreprise et ayant une longueur de moins de 2 000 mètres.

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Param. Estimate	Std Error	Chi-Square	Pr > ChiSq	Hazard Ratio	
Region	IDF	1	0.17923	0.02750	42.4736	<.0001	1.196	
Region	OUEST	1	0.08070	0.02040	15.6526	<.0001	1.084	
q_longueur_ligne	< 4500m	1	0.07895	0.02433	10.5283	0.0012	1.082	
q_longueur_ligne	>= 4500m	1	0.15777	0.03878	16.5527	<.0001	1.171	
type_produit	P4	1	0.34451	0.04001	74.1378	<.0001	1.411	
type_produit	P5	1	0.41589	0.03805	119.4935	<.0001	1.516	
Operateur	FOURNISSEUR	1	0.13220	0.02960	19.9512	<.0001	1.141	

On a respectivement 19.6% plus de risques et 8.4% plus de risques d'avoir un défaut lorsqu'on est situé en Ile-de-France et dans l'Ouest que dans le Nord-Est, les autres covariables étant inchangées.
L'Ile-de-France représente seulement 15% des observations tandis que les 2 autres régions couvrent de manière équitable le reste des observations.
Il peut donc y avoir un léger biais. Cependant d'autres variables explicatives non observées telles que la qualité du réseau local et des détériorations matérielles par exemple peuvent expliquer ces écarts importants.
On a respectivement 51.6% plus de risques et 41.1% plus de risques d'avoir un défaut pour un produit de type P5 et P4 que pour les autres produits (P1, P2 ou P3), bien plus communs.
On peut donc supposer qu'il s'agit des caractéristiques particulières de ces produits qui expliquent ces écarts.
Un site installé par le fournisseur a 14.1% plus de risques de défaut qu'un site installé par un autre fournisseur.
Pour une longueur de ligne entre 2 000 et 4 500 mètres, le taux de risque augmente de 8.2% et pour une longueur de plus de 4 500 mètres, ce taux monte à 17.1%, par rapport à une ligne de moins de 2 000 mètres, tout en gardant les autres covariables identiques à l'individu de référence.

5 Conclusion

L'étude de l'historique des activations et signalements de lignes téléphoniques et internet du fournisseur à l'aide de modèles de durée a permis de répondre à la problématique de départ. À savoir identifier et comprendre les facteurs clés pouvant influencer le risque de dysfonctionnement 30 jours après l'activation de la ligne. Cependant, cette étude reste partielle dans le temps et dans l'espace. En effet, l'historique mis à disposition ne concerne que les activations de ligne ayant eu lieu pendant le mois d'Avril 2019 et dans le Nord de la France (régions Ouest, Île-de-France et Nord-Est).

Ainsi, les enseignements qui sont tirés des résultats obtenus peuvent ne pas se généraliser sur l'ensemble de la France et à différentes périodes de l'année. Des spécificités géographiques et temporelles peuvent intervenir, préconisant de contrôler l'inférence sur d'autres zones ou périodes d'étude. Par exemple, l'idéal serait d'utiliser l'historique (plus large dans la période couverte) de chaque région afin de refaire des études par secteur géographique. Les effets peuvent changer selon la région, à cause de dépendances entre les variables à travers des effets non observés tels que la qualité du réseau par exemple.

L'approche non paramétrique a permis de montré, à travers une considération empirique des données observées que la probabilité d'un signalement pour dysfonctionnement sur la ligne est très élevée peu après l'activation de la ligne. Ce risque est de 10% le premier jour pour l'ensemble de la population. Environ un tiers des activations mènent à un dysfonctionnement signalé dans les 30 jours à venir et la moitié d'entre eux sont effectués dans les 5 jours à venir. La durée moyenne avant un signalement est d'une semaine environ. Le risque de signalement est stable et faible ensuite.
Cette approche permet également d'identifier les covariables pouvant segmenter les lignes afin d'affiner l'étude. Les tests statistiques et les analyses graphiques ont permis d'en identifier trois : la région, le type de produit et l'opérateur. Si l'analyse graphique permet de comparer les strates (position des courbes entre-elles), cette approche ne permet pas de quantifier l'effet.

L'approche paramétrique permet de prédire le risque instané à chaque temps et pour chaque cas. Pour cela, il était nécessaire d'identifier la loi la plus adaptée aux données (celle qui s'ajuste au mieux). Le choix de la loi a été fait à l'aide du hazard plotting. En transformant la fonction de survie de façon à ce qu'elle soit fonction linéaire du temps (logarithme du temps) pour chaque loi considérée (exponentielle, Weibull, log-normale, log-logistique), on peut effectuer des régression linéaires. La loi ayant la fonction de survie estimée qui se rapproche le plus de la survie empirique (à travers le coefficient de détermination R^2) sera celle choisie. Dans notre cas, il s'agit de la loi log-normale.

La loi ayant été identifiée, il reste à intégrer les covariables de segmentation retenues dans l'étude non paramétrique. On soupçonne également que la variable indiquant la longueur de la ligne a une influence significative sur la durée de vie de ces lignes. Pour cela, on effectue une recherche exhaustive sur ces variables de façon à identifier les variables qui minimisent le critère d'information d'Akaike (AIC). En résultat, les variables candidates peuvent toutes êtres intégrées dans le modèle final puisque l'AIC est minimal en les considérant toutes, et ce, pour les quatre lois étudiées.

L'estimation se fait à l'aide d'un modèle à temps de vie accéléré (AFT). Les résultats qui en sortent s'interprètent comme des différences de durées de vie. Ceci permet de quantifier d'une première façon les effets de ces covariables. De plus, il est possible de connaître le risque instantané (donc le risque de survenue d'un dysfonctionnement sachant que la ligne n'en a pas subi depuis son activation) pour un temps donné pour chaque ligne, à l'aide de la macro fonction PREDICT de SAS.

Enfin, l'approche semi-paramétrique permet une deuxième quantification des effets liés aux facteurs considérés. Pour cela, le modèle estimé doit vérifier les deux hypothèses d'un modèle de Cox : la log-linéarité des variables continues et la proportionnalité des risques pour l'ensemble des variables. Le modèle retenu ne respectant pas la deuxième hypothèse pour la variable sur la longueur de la ligne, cette dernière est discrétisée en trois classes. Le modèle final en considérant cette variable synthétique vérifie l'ensemble des hypothèses et les tests de significativité individuelle. Les estimations sont donc interprétables. Ceci passe par l'analyse des rapports de risques (hazard ratio) qui se lisent comment le pourcentage de risque en plus ou en moins de subir un dysfonctionnement par rapport à une observation de référence.

De manière générale et pour les trois approches, on note que le risque est plus élevé en région Ouest qu'en région Nord-Est et encore plus pour l'Île-de-France. Les produits de type « P5 » sont plus risqués que les produits « P4 » qui sont eux-mêmes plus risqués que les produits « P1 », « P2 » ou « P3 ». Par ailleurs, lorsque l'opérateur est un fournisseur, le risque de dysfonctionnement est plus élevé. Enfin, plus la longueur de la ligne est élevée et plus le risque est important. Ces covariables sont donc des facteurs clés où l'opérateur peut agir pour réduire le taux de dysfonctionnement ou pour mieux calibrer l'appel à ses techniciens dans le délai imparti.

Annexes

Annexe 1 – Macro fonction recherche_exhaustive_aft : documentation

%recherche_exhaustive_aft(database, out, var_cat_x, var_cont_x, var_y, censor, censor_value)

Inputs :
- **database** : table SAS contenant les données individuelles à utiliser pour estimer les modèles paramétriques.
- **out** : table SAS qui contiendra en sortie de la fonction les critères de qualité d'estimation de tous les modèles (AIC, AICC, BIC, log-vraisemblance, paramètres d'échelle et de position).
- **var_cat_x** : chaîne de caractères listant les variables catégorielles exogènes à intégrer dans les modèles. Les variables sont séparées par un espace uniquement.
- **var_cont_x** : chaîne de caractères listant les variables continues exogènes à intégrer dans les modèles. Les variables sont séparées par un espace uniquement.
- **var_y** : chaîne de caractères indiquant la variable de durée de vie.
- **censor** : chaîne de caractères indiquant la variable de censure.
- **censor_value** : chaîne de caractères indiquant la valeur prise par la variable de censure (**censor**) lorsqu'il y a censure.

Outputs :
- **out** : table SAS qui contiendra en sortie de la fonction les critères de qualité d'estimation de tous les modèles (AIC, AICC, BIC, log-vraisemblance, paramètres d'échelle et de position).
- **lambda_exponential, gamma_exponential, lambda_weibull, gamma_weibull, lambda_lognormal, gamma_lognormal, lambda_llogistic, gamma_llogistic** : macro variables (scope global) contenant les paramètres d'échelle et de position des lois testées qui serviront pour le hazard plotting.
- **sortie affichée** : 4 tableaux (une par loi) contenant les modèles testés triés par ordre croissant de l'AIC.

Exemple d'utilisation :

```
%recherche\_exhaustive\_aft(database = telco_clean,
    out          = recap,
    var_cat_x    = region type_produit operateur,
    var_cont_x   = longueur_ligne,
    var_y        = dv_ligne,
    censor       = pbm,
    censor_value = 0);
```

Description :
La macro fonction **recherche_exhaustive_aft** sert à tester de façon exhaustive toutes les combinaisons de variables afin d'en extraire des critères de sélection de modèles. Cette recherche est faite pour les 4 lois suivantes : exponentielle, weibull, log-normale, log-logistique.
L'utilisation de la log-vraisemblance ne permet pas de comparer des lois appartenant à différentes familles. Plutôt que de s'en servir pour choisir la loi la plus adéquate, on va s'en servir (avec une pénalisation : AIC) afin de sélectionner les variables. Le hazard plotting servira à décider de la loi par la suite.

Fonctionnement :
Le premier tiers de la macro fonction sert à préparer la liste exhaustive des modèles possibles (stockée dans une table SAS). La recherche exhaustive est effectuée à l'aide de deux boucles imbriquées, une sur les lois à tester et l'autre (imbriquée) sur les modèles à estimer. À chaque itération, une estimation à l'aide de la procédure LIFEREG est effectuée en s'assurant que des variables continues ne soient pas incluses parmis les variables de stratification (statement CLASS de la procédure).
Les critères d'information, la log-vraisemblance et les estimations des paramètres du modèles sont extraites et sauvegardées dans des tables SAS. La suite du programme sert alors à synthétiser l'ensemble des éléments intéressants dans une même table SAS afin de s'en servir comme outil d'aide à la décision.