Analyse de données de monitoring de performance dans les réseaux optiques cohérents

par

Banti Laure Mathilde YAMÉOGO

THÈSE PRÉSENTÉE À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DU
DOCTORAT
Ph. D.

MONTRÉAL, LE 23 JUILLET 2021

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

©Tous droits réservés, Banti Laure Mathilde Yaméogo, 2021

©Tous droits réservés
Cette licence signifie qu'il est interdit de reproduire, d'enregistrer ou de diffuser en tout ou en partie, le présent document. Le lecteur qui désire imprimer ou conserver sur un autre media une partie importante de ce document, doit obligatoirement en demander l'autorisation à l'auteur.

PRÉSENTATION DU JURY

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE :

Mme Christine Tremblay, directrice de thèse
Département de génie électrique à l'École de technologie supérieure

M. Maurice O'Sullivan, codirecteur de thèse
Ciena Corp.

M. Antoine Tahan, président du jury
Département de génie mécanique à l'École de technologie supérieure

M. François Blanchard, membre du jury
Département de génie électrique à l'École de technologie supérieure

M. David Plant, examinateur externe
Département de génie électrique et informatique à l'Université McGill

ELLE A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 14 JUILLET 2021

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je veux remercier ma directrice de thèse la professeure Christine Tremblay qui a cru en moi et qui m'a offert l'opportunité de mener cette thèse dans le domaine des réseaux de communication optiques et de l'analyse de données. Le domaine de l'optique a toujours été passionnant pour moi, et y étudier est une source d'épanouissement. Merci d'être une source d'inspiration et un exemple pour moi. Merci pour votre disponibilité, vos encouragements, vos conseils et pour m'avoir inculqué la discipline et la rigueur dans mon travail.

Je veux dire merci à mon co-directeur Maurice O'Sullivan pour ses précieux commentaires et remarques.

Je veux saluer la contribution inestimable de feu M. Michel Bélanger.

Je veux saluer toute l'équipe de Ciena et l'équipe de monitoring de l'ÉTS pour l'apport de leur expertise, disponibilité et leur patience. Un merci spécial à Christian Desrosiers, Doug Charlton et Dave Doucet.

Je veux remercier le Conseil de Recherches en Sciences Naturelles et en Génie (CRSNG) pour le soutien financier qui m'a permis d'entreprendre et de compléter ma thèse.

Je remercie le professeur Antoine Tahan pour mes premiers pas en statistiques.

Merci à mes parents Noélie et Jean-Paul pour m'avoir transmis le sens et la valeur des études.

Je dis un merci particulier à mon époux Pipi-Kassida Paul Kaboré pour sa patience, ses sacrifices consentis, sa confiance, et pour avoir cru en moi et qui a toujours trouvé les mots pour m'encourager et me relever.

À mes filles Celya & Delya qui ont embarqué avec moi dans cette aventure, je dis merci. Merci pour votre patience et les fins de semaines et vacances sacrifiées pour que maman travaille. Votre joie de vivre m'a toujours donné le goût d'aller de l'avant.

Je veux adresser un merci spécial à Tongré Wendé pour avoir toujours été là pour moi.

ANALYSE DE DONNÉES DE MONITORING DE PERFORMANCE DANS LES RÉSEAUX OPTIQUES COHÉRENTS

Banti Laure Mathilde YAMÉOGO

RÉSUMÉ

L'application conjointe de la détection cohérente, de formats de modulation en amplitude et phase, ainsi que de techniques avancées de traitement numérique du signal (*Digital Signal Processing, DSP*) et de correction des erreurs (*Forward Error Correction, FEC*), a permis d'augmenter le débit binaire par canal portant ainsi la capacité de transmission des systèmes optiques à plus de 10 Tb/s par lien. De tels systèmes permet aux opérateurs de réseaux de répondre aux besoins de plus en plus croissant et exigeant des utilisateurs. Cependant, il n'est pas suffisant d'offrir une connexion haut débit aux clients; les opérateurs doivent aussi s'assurer de leur garantir les accords de service (*Service Level Agreement, SLA*).

Dans ce contexte, les systèmes de contrôle du réseau (*Network Management System, NMS*) mesure en permanence la santé du réseau en recueillant des données à partir des dispositifs déployés dans le réseau. Ces données permettent d'analyser les causes de dégradation de performance dans le réseau et les défaillances des équipements, de gérer les pannes et les interventions de maintenance, et sont utilisées par les planificateurs de réseau pour planifier les déploiements futurs en fonction de la croissance du trafic. D'un autre côté, la disponibilité des données due au développement de l'internet des objets et de la quantité de sources de données connectées a donné naissance à la science des données déjà mise en application dans de nombreux domaines. Avec l'arrivée des systèmes cohérents qui amassent des données via les modules DSP et les moyens de stockage et de traitement de données massives, la science des données est apparue récemment dans le domaine des télécommunications optiques.

L'idée centrale de la thèse est de tirer le maximum d'informations et de connaissances en utilisant les techniques de l'analyse des données de performance de 155 connexions optiques collectées sur une période d'un an dans le réseau de production d'un opérateur dans le but de développer une meilleure compréhension des mécanismes de dégradation de performance et de migrer vers des approches proactives de gestion de réseau. La première partie de cette thèse consiste à caractériser les anomalies de performance détectées durant la période d'observation avec pour objectif d'établir des corrélations entre les paramètres. Un regroupement des 155 circuits basé sur leurs caractéristiques statistiques est ensuite fait en appliquant une méthode évolutive de regroupement de séries temporelles afin de découvrir des informations sous-jacentes aux données. Enfin, une méthode de détection de dégradation de perte optique dans une liaison optique, basée sur l'utilisation de trois méthodes statistiques complémentaires – *Seasonal-Trend decomposition using LOESS, STL*, test de Mann-Kendall et Sen's slope – est proposée et validée avec des données de terrain.

Mots-clés: analyse de tendance, décomposition de séries temporelles, détection d'anomalies, monitoring de réseaux optiques cohérents, regroupement de données

ANALYSIS OF PERFORMANCE MONITORING DATA IN COHERENT OPTICAL NETWORKS

Banti Laure Mathilde YAMÉOGO

ABSTRACT

The combined application of coherent detection, amplitude and phase modulation formats, as well as advanced Digital Signal Processing (DSP) and Forward Error Correction (FEC) techniques, has made it possible to increase the bit rate per channel, bringing the transmission capacity of optical transmission systems to more than 10 Tb/s per link. Such systems enable network operators to meet the growing and increasingly demanding needs of users. However, it is not enough to offer a high-speed connection to customers; operators must also ensure that service level agreements (SLAs) are met.

In this context, Network Management Systems (NMS) continuously monitor the health of the network by collecting data from the devices deployed in the network. This data is used to analyze the causes of performance degradation in the network and equipment failures, to manage outages and maintenance interventions, and is used by network planners to plan future deployments based on traffic growth.

On the other hand, the availability of data due to the development of the Internet of Things and the amount of connected data sources has given rise to data science which has already found applications in many areas. With the advent of coherent systems that collect data via DSP modules and massive data storage and processing capabilities, data science has recently emerged in the field of optical telecommunications.

The central idea of the thesis is to extract the maximum of information and knowledge by using the techniques of performance data analysis of 155 lightpaths collected over a period of one year in an operator's production network with the objective to better understand the root cause of performance degradations as a first step towards proactive network management approaches. The first part of this thesis consists in characterizing the performance anomalies detected during the observation period with the objective of establishing correlations between the monitored parameters. A clustering of the 155 lightpaths based on their statistical characteristics is then done by applying an evolutive method of time series clustering in order to discover information underlying the data. Finally, a method for detecting optical loss degradation in an optical link, based on the use of three complementary statistical methods – *Seasonal-Trend decomposition using LOESS*, STL, Mann-Kendall test and Sen's slope – is proposed and validated with field data.

Keywords: trend analysis, time series decomposition, anomaly detection, monitoring of coherent optical networks, data clustering

TABLE DES MATIÈRES

Page

INTRODUCTION 1

CHAPITRE 1 MONITORING DE PERFORMANCE DANS LES RÉSEAUX OPTIQUES COHÉRENTS 5
1.1 Introduction 5
1.2 Les transpondeurs cohérents 6
1.3 Surveillance de performance dans les réseaux optiques utilisant des transpondeurs cohérents 8
1.3.1 Les principaux défis de la surveillance de performance 8
1.3.2 Les paramètres réseaux nécessitant le monitoring 9
1.3.3 Les paramètres réseaux monitorés dans le cadre de ce projet 13
1.3.4 Réception des données et processus de pré-traitement 14
1.4 Les perspectives de la surveillance de performance 16
1.4.1 Les liens enfouis 16
1.4.2 Les liens aériens 17
1.4.3 Conclusion 17
1.5 Techniques d'analyse de données 18
1.5.1 Introduction 18
1.5.2 L'analyse statistique 19
1.5.3 L'analyse prédictive 20
1.5.4 L'apprentissage machine dans les réseaux optiques 22
1.6 Conclusion 26

CHAPITRE 2 MÉTHODES D'ANALYSE DESCRIPTIVE ET INFÉRENTIELLE DES SÉRIES TEMPORELLES 27
2.1 Introduction 27
2.2 Outils d'analyse statistique de base 27
2.2.1 Indicateur de tendance centrale 27
2.2.2 Indicateur de dispersion 28
2.2.3 Indicateur d'asymétrie 29
2.3 Outils d'analyse graphique de la densité 30
2.3.1 Histogramme 31
2.3.2 Boxplot 31
2.4 Outils d'analyse de la périodicité du signal 32
2.4.1 Fonction d'autocorrélation 32
2.5 Analyse de corrélation 33
2.5.1 Test de corrélation 33
2.5.2 Les types de coefficient de corrélation 34
2.5.3 Le test statistique de Shapiro – Wilk 35
2.6 Conclusion 36

CHAPITRE 3 CARACTÉRISATION DES ROUTES MONITORÉES CÔTÉ ANOMALIES 39
3.1 Introduction 39
3.2 Description des routes monitorées 40
3.3 Définition et détection des anomalies 41
3.3.1 Définition 41
3.3.2 Détection d'anomalies 41
3.4 Choix des circuits par catégorie 45
3.4.1 Catégorie 1 – Les anomalies affectent tous les canaux de la route 45
3.4.2 Catégorie 2 – Un seul canal de la route affecté par les anomalies 48
3.4.3 Catégorie 3 – Une direction de la route affectée par les anomalies 51
3.5 Analyse de corrélation entre les différents paramètres de performance 55
3.5.1 Catégorie 1 – Les anomalies affectent tous les canaux de la route 55
3.5.2 Catégorie 2 – Un seul canal de la route affectée par les anomalies 58
3.5.3 Catégorie 3 – Une direction de la route affectée par les anomalies 58
3.5.4 Amplitude des chutes de puissance optique et secondes erronées 58
3.5.5 Récapitulatif 59
3.6 Cause probable des anomalies détectées 60
3.6.1 Amplificateurs optiques 60
3.6.2 Catégorie 1 – Les anomalies affectent tous les canaux de la route 61
3.6.3 Catégorie 2 – Un seul canal de la route affectée par les anomalies 67
3.6.4 Catégorie 3 – Une direction de la route affectée par les anomalies 70
3.7 Discussion et conclusion 73

CHAPITRE 4 REGROUPEMENT DES 155 CIRCUITS BASÉ SUR LEURS CARACTÉRISTIQUES STATISTIQUES DE BASE 75
4.1 Introduction 75
4.2 Choix de l'algorithme de regroupement 76
4.2.1 Le regroupement hiérarchique 76
4.2.2 Le regroupement par partitionnement 77
4.2.3 Métriques de validation du regroupement 83
4.3 Regroupement des 155 circuits monitorés 86
4.3.1 Méthodologie 86
4.3.2 Transformation des données 88
4.3.3 Tendance des clusters dans les données 89
4.3.4 L'importance des attributs 92
4.3.5 Sélection des attributs dans les données de monitoring des 155 circuits 95
4.3.6 Application de l'algorithme k-moyennes 101
4.3.7 Résultats 101
4.3.8 Analyse des comportements sous-jacents dans les circuits 108
4.3.9 Récapitulatif 112
4.3.10 Conclusion 114

CHAPITRE 5 ANALYSE DES LIENS OPTIQUES.....117
5.1 Introduction.....117
5.2 Détection de tendance dans les spans de fibre optique.....117
5.2.1 Introduction..... 117
5.3 Méthodologie.....119
5.3.1 Système et prétraitement des données 119
5.3.2 Méthode de décomposition de série temporelle..... 123
5.3.3 Méthodes statistiques non paramétriques pour la quantification de la tendance..... 125
5.4 Validation de la tendance par régression linéaire 127
5.5 Résultats et discussion 128
5.6 Conclusion 136

CONCLUSION 139

RECOMMANDATIONS 143

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES.....146

LISTE DES TABLEAUX

Page

Tableau 2.1 Règles générales concernant la définition du coefficient de corrélation 34

Tableau 4.1 Synthèse des attributs sélectionnés par chaque technique 96

Tableau 4.2 Métriques de validation calculées pour chaque ensemble d'attributs sélectionné par chaque modèle 99

Tableau 4.3 Récapitulatif des caractéristiques des circuits par route ainsi que leur cluster d'appartenance 113

Tableau 5.1 Analyse statistique des données de perte de span 121

Tableau 5.2 Analyse statistique des données de perte par span 133

LISTE DES FIGURES

Page

Figure 0.1 Étapes principales du processus d'application de la science de données Tirée de Babu (2019) 2

Figure 1.1 Principe d'un système optique cohérent DP-QPSK Adaptée de Roberts et al. (2009b) 7

Figure 1.2 Paramètres monitorés dans la couche optique d'un système cohérent Adaptée de Campos, Jia et Wolcott (2018) 10

Figure 1.3 Architecture d'une liaison optique bidirectionnelle typique: ROADM: Reconfigurable Optical Add/Drop Multiplexer, WSS: Wavelength Selective Switch, OPT: Optical Power Transmitted, OPIN: Optical Power Input, OPOUT: Optical Power Output, OPR: Optical Power Received, OCH: Optical Channel, OTS: Optical Transmission Section, ES: Errored Second 13

Figure 1.4 Architecture de l'environnement de développement de la base de données ETL : Extract, Transform and Load, SSRS : Sql Server Reporting Services 15

Figure 3.1 Boîte à moustache (*Boxplot*) expliquant la méthode de détection d'anomalies basée sur l'IQR 42

Figure 3.2 Boîtes à moustache du SNR de trois (3) circuits bidirectionnels ainsi que les anomalies détectées durant une période d'observation de 12 mois 43

Figure 3.3 Anomalies détectées par la méthode IQR dans la série temporelle de SNR d'une connexion optique 44

Figure 3.4 Synoptique des circuits de la Catégorie 1 – Route 1 : CIRCUIT-1 et CIRCUIT-2, Route 2 : CIRCUIT-3 Direction A-B : aller (Site 1 à 2) et direction B-A : retour (Site 2 à 1) 46

Figure 3.5 Évolution temporelle du SNR et des secondes erronées pour les trois circuits bidirectionnels issus de deux routes présentant des canaux uniformément affectés par les anomalies : Catégorie 1 – Route 1 : CIRCUIT-1 et CIRCUIT-2, Route 2 : CIRCUIT-3 Direction : aller = A-B et retour = B-A 47

Figure 3.6 Évolution temporelle de la puissance reçue au récepteur et des secondes erronées pour les trois circuits bidirectionnels issus de deux routes présentant des canaux uniformément affectés par les anomalies : Catégorie 1 – Route 1 : CIRCUIT-1 et CIRCUIT-2, Route 2 : CIRCUIT-3 – Direction : aller = A-B et retour = B-A.....48

Figure 3.7 Synoptique des circuits de la route 3 : CIRCUIT-4, CIRCUIT-5 et CIRCUIT-6 Direction A-B : aller (Site 1 à 2) et direction B-A : retour (Site 2 à 1) 49

Figure 3.8 Évolution temporelle du SNR et des erreurs secondes pour les trois circuits bidirectionnels de la catégorie 2 – Un seul canal de la route affectée par les anomalies – Route 3 : CIRCUIT-4, CIRCUIT-5 et CIRCUIT-6 Direction : aller = A-B et retour = B-A.....50

Figure 3.9 Évolution temporelle de la puissance reçue au récepteur et des erreurs secondes pour les trois circuits bidirectionnels de la catégorie 2 – Un seul canal de la route affectée par les anomalies - Route 3 : CIRCUIT-4, CIRCUIT-5 et CIRCUIT-6 Direction : aller = A-B et retour = B-A.....51

Figure 3.10 Synoptique de la liaison optique de la route analysée dans la catégorie 3 : Un seul canal de la route affectée – Route 4 : CIRCUIT-7, CIRCUIT-8 et CIRCUIT-9 Direction : aller = A-B (site 2-3) et retour = B-A (site 3-2).....52

Figure 3.11 Évolution temporelle du SNR et des secondes erronées pour les trois circuits bidirectionnels de la même route présentant une direction affectée par les anomalies Route 4 : CIRCUIT-7, CIRCUIT-8 et CIRCUIT-9 Direction : aller = A-B et retour = B-A.....53

Figure 3.12 Évolution temporelle de la puissance reçue au récepteur et des secondes erronées pour les trois circuits bidirectionnels de la même route présentant une direction affectée par les anomalies – Route 4 : CIRCUIT-7, CIRCUIT-8 et CIRCUIT-9 Direction : aller = A-B et retour = B-A.....54

Figure 3.13 Matrice de corrélation du Rapport signal sur bruit (*SNR*), de la puissance reçue au récepteur (*OPRx*) et du nombre de secondes erronées (*ES*).....57

Figure 3.14 Représentation simplifiée de la liaison optique de la catégorie 1 Route 1 : CIRCUIT-1 (1050 km) : Site 1 à 3 et CIRCUIT-2 (700 km): Site 1 à 2 62

Figure 3.15 Évolution temporelle du gain des amplificateurs optiques et de la puissance reçue par circuit Catégorie 1 – Route 1 : CIRCUIT-1 et CIRCUIT-2, Route 2 : CIRCUIT-3 Direction : aller = A-B et retour = B-A 63

Figure 3.16 Évolution temporelle du SNR, de la puissance reçue au récepteur (*OPRx*) et du gain des amplificateurs variant dans la période du 24 au 25 mai (CIRCUIT-1).....65

Figure 3.17 Évolution temporelle du gain des amplificateurs variant dans la période du 24 au 25 mai (CIRCUIT-1).....66

Figure 3.18 Évolution temporelle du SNR, de la puissance reçue au récepteur (*OPRx*), du gain, de la puissance d'entrée (*OPIN*) et de la puissance de sortie (*OPOUT*) de l'amplificateur 11 présentant un pic pour la période du 6 au 7 juillet (CIRCUIT-2) 67

Figure 3.19 Représentation bidirectionnelle de la liaison optique de la route analysée dans la catégorie 2 – Un seul canal de la route affectée Route 3 : CIRCUIT-4, CIRCUIT-5 et CIRCUIT-6.....68

Figure 3.20 Évolution temporelle du gain des amplificateurs optiques et de leurs puissances d'entrée et de sortie de la catégorie 2 Route 3 : CIRCUIT-4 et CIRCUIT-5 et CIRCUIT-6 Direction : aller = A-B et retour = B-A.....69

Figure 3.21 Représentation bidirectionnelle de la liaison optique de la route analysée dans la catégorie 3 – Une direction de la route affectée Route 4 : CIRCUIT-7, CIRCUIT-8 et CIRCUIT-9.....70

Figure 3.22 Évolution temporelle des puissances d'entrée (OPIN) et de sortie (OPOUT) des amplificateurs dans la direction A-B des circuits de la catégorie 3 – Une direction de la route affectée : Route 4 71

Figure 3.23 Évolution temporelle du gain des amplificateurs optiques et de leurs puissances d'entrée et de sortie de la catégorie 3 – Route 4 : CIRCUIT-7 etCIRCUIT-8 et CIRCUIT-9 Direction : aller = A-B et retour = B-A.....71

Figure 3.24 Évolution temporelle du SNR, de la puissance reçue au récepteur (*OPRx*), du gain des deux amplificateurs

en faute pendant les deux évènements du 12 novembre (CIRCUIT-1 et CIRCUIT-2).....73

Figure 4.1 Méthodologie de la mise en cluster des 155 circuits.....88

Figure 4.2 Représentation pixellisée de la matrice de dissimilarité des 155 circuits.....92

Figure 4.3 Sélection des attributs : Technique ReliefF (*barplot*) et technique du gain d'information (*lollipop*) : Mean (valeur moyenne du SNR), Max (valeur maximale du SNR), Standard deviation (écart-type du SNR), IQR (écart-interquartile du SNR) et Diff_min (différence SNR_{min} et SNR_{seuil}).....96

Figure 4.4 Sélection des attributs (features) basée sur la technique de la corrélation : la corrélation de tous les attributs (a) et la corrélation des attributs sélectionnés (b) : Mean (valeur moyenne du SNR), Max (valeur maximale du SNR), Standard deviation (écart-type du SNR), IQR (écart-interquartile du SNR) et Diff_min (différence SNR_{min} et SNR_{seuil}).....96

Figure 4.5 Diagramme de fréquence en fonction du nombre de clusters pour : (a) le modèle 1 (tous les attributs) et (b) le modèle 2 (technique ReliefF) 98

Figure 4.6 Diagramme de fréquence en fonction du nombre de clusters pour : (a) le modèle 3 (technique du gain d'information) et (b) le modèle 4 à droite (technique basée sur la corrélation) 98

Figure 4.7 Graphique de la distribution des 155 circuits dans les trois clusters.....102

Figure 4.8 Graphe du coefficient de silhouette des 155 circuits 104

Figure 4.9 Graphe de coordonnées parallèles des 155 circuits.....105

Figure 4.10 Évolution temporelle du SNR des circuits appartenant au cluster 1 106

Figure 4.11 Évolution temporelle du SNR des circuits appartenant au cluster 2 107

Figure 4.12 Évolution temporelle du SNR des circuits appartenant au cluster 3 108

Figure 4.13 Synoptique des routes 5 et 6.....109

Figure 4.14 Synoptique des routes 1 et 2.....110

Figure 4.15 Synoptique des routes 7 et 8 111

Figure 4.16 Synoptique des routes 9 et 4 111

Figure 4.17 Synoptique des routes 10 et 11 112

Figure 5.1 Liaison optique de 205 Km à 4 spans 120

Figure 5.2 Différentes étapes de la méthodologie 120

Figure 5.3 Représentation de l'histogramme du span 1 (bleu), span 3 (rouge) Direction A – B (gauche), Direction B – A (droite) 122

Figure 5.4 Représentation de l'histogramme du span 2 (mauve), span4 (vert) : Direction A – B 122

Figure 5.5 Représentation de l'histogramme du span 2 (mauve), span4 (vert) : Direction B – A 123

Figure 5.6 Direction A - B : Composantes de la décomposition des séries temporelles pour les spans de 13 dB [span 1 (bleu), span 3 (rouge)] : données brutes (a) ; composante de tendance (b) ; composante saisonnière (c) ; résiduel (d) 129

Figure 5.7 Direction A - B : Composantes de la décomposition des séries temporelles pour les spans de 20 dB [span 2 (mauve), span 4 (vert)] : données brutes (a) ; composante de tendance (b) ; composante saisonnière (c) ; résiduel (d) 130

Figure 5.8 Direction A - B : Composantes de la tendance pour le span 1 (bleu), le span 3 (rouge), le span 2 (violet), le span 4 (vert) 130

Figure 5.9 Direction B - A : Composantes de la décomposition des séries temporelles pour les spans de 14 dB [span 1 (bleu), span 3 (rouge)] : données brutes (a) ; composante de tendance (b) ; composante saisonnière (c) ; résiduel (d) 131

Figure 5.10 Direction B - A : Composantes de la décomposition des séries temporelles pour les spans de 18 dB [span 2 (mauve), span 4 (vert)] : données brutes (a) ; composante de tendance (b) ; composante saisonnière (c) ; résiduel (d) 131

Figure 5.11 Direction B - A : Composantes de la tendance pour le span 1 (bleu), le span 3 (rouge), le span 2 (violet), le span 4 (vert) 132

Figure 5.12 Données de perte de span 1 : période de 12 premiers mois (*en noir*), données additionnelles de 12 mois (*en bleu*) – (Direction A-B gauche) et (Direction B-A droite) 134

Figure 5.13 Courbes de R^2 ajustées en fonction de l'horizon de temps pour la période d'observation supplémentaire de 12 mois : Direction A - B (noir), Direction B - A (rouge) 135

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ASE	Amplified Spontaneous Emission
BER	Bit Error Rate
CD	Chromatic Dispersion
DCF	Dispersion Compensating Fiber
DGD	Differential Group Delay
DMUX	DE multiplexer
DP-QPSK	Dual Polarization Quadrature Phase Shift Keying
DSF	Dispersion-Shifted Fiber
DSP	Digital Signal Processing
EDFA	Erbium Doped Fiber Amplifier
ES	Error Second
FEC	Forward Error Correction
GVD	Group Velocity Dispersion
HCCS	High Correction Count Second
I/Q	In phase/Quadrature
ITU	International Telecommunication Union
LED	Light Emitting Diode
MUX	Multiplexer
NDSF	Non-Dispersion-Shifted Fiber, NDSF
OADM	Optical Add-Drop Multiplexer
OIF	Optical Internetworking Forum
OPIN	Amplifier optical input power

OPM	Optical Performance Monitoring
OPOUT	Amplifier optical output power
OPRx	Optical receiver power
OPTx	Optical transmitter powe
OSNR	Optical Signal-to-Noise Ratio
PBC	Polarization Beam Combiners
PBS	Polarization Beam Splitter
PMD	Polarization Mode Dispersion
Rx / Tx	Récepteur / Transmetteur
SBS	Stimulated Brillouin Scattering
SER	Symbol Error Rate
SEFS	Severely Error Frame Seconds
SES	Severely Error Seconds
SLA	Service Level Agreement
SMF	Single Mode optical Fiber
SPM	Self-Phase Modulation
SRS	Stimulated Raman Scattering
WDM	Wavelength Division Multiplexing
XPM	Cross-Phase Modulation

LISTE DES SYMBOLES ET UNITÉS DE MESURE

UNITÉS DE BASE

m	mètre (unité de longueur)
s	seconde (unité de temps)

Aire

km^2	kilomètre carré (= 1 000 000 m^2)
m^2	mètre carré

UNITÉS GÉOMÉTRIQUES

m	mètre
km	kilomètre
µm	micromètre
nm	nanomètre

UNITÉS MÉCANIQUES

Vitesses

m/s	mètre par seconde

Vitesses de groupe

$ps^2.km^{-1}$	picoseconde par kilomètre

UNITÉS DE TEMPS

h	heure
min	minute
s	seconde
ms	milliseconde
µs	microseconde

Angle solide Fréquence

THz	terahertz
GHz	gigahertz
MHz	mégahertz
Hz	hertz

UNITÉS DE PUISSANCE

kW	kilowatt
W	watt
mW	milliwatt

UNITÉS D'INTENSITÉ DU SIGNAL

dB	décibel
dB/km	décibel par kilomètre

UNITÉS DE DÉBIT

Tbit/s	terabit par seconde
Mbit/s	megabit par seconde
Gbit/s	gigabit par seconde
Bit/s	bit par seconde

INTRODUCTION

Les communications à fibre optique connaissent une croissance alimentée principalement par une incessante et continue demande en bande passante. Cette demande est stimulée par le nombre d'utilisateurs, la diffusion de contenu à la demande, le *cloud computing*, la connectivité sans fil 4G, la connectivité des centres de données et les services d'entreprise Ethernet haute vitesse (Roberts et al., 2015) . Afin de répondre à cette demande croissante en bande passante, plusieurs techniques ont été déployées. L'une d'elles est la combinaison de la technologie cohérente couplée aux formats de modulation avancés et au traitement numérique de signal (*Digital Signal Processing, DSP*) qui permet d'atteindre un débit de 100 Gbit/s et plus par canal.

Ainsi, en plus d'offrir des connexions haut débit aux utilisateurs, il est important d'assurer l'intégrité des données qu'ils reçoivent. En effet, au regard de la quantité de données qui transite par chaque lien, les réseaux optiques sont plus vulnérables aux pannes dans le lien optique (Dong et al., 2016).

Dans ce sens, les opérateurs monitorent régulièrement les performances optiques de leurs réseaux pour garantir que les clients reçoivent leurs données de manière fiable (Woodward et al., 2008). Ainsi, les systèmes de contrôle des réseaux actuels permettent d'acquérir un certain nombre de paramètres en temps réel (comme les mesures de taux d'erreur binaire (*Bit Error Rate, BER*) à toutes les 15 minutes, par exemple). Ce qui fait défaut, c'est que ces paramètres ne sont pas utilisés ou mis à profit pour analyser les tendances en matière de performance. Autrement dit, les paramètres sont stockés à court terme et ne sont utilisés qu'en cas de panne, soit en mode réactif plutôt qu'en mode proactif. Cela introduit une grande contrainte dans la gestion des ressources et le maintien d'une meilleure qualité de service. Par conséquent, les réseaux souffrent d'une incapacité à évaluer à la fois l'état physique du réseau et la qualité du signal optique s'y propageant, conduisant au surdimensionnement des réseaux et une utilisation inefficace des ressources réseau.

D'autre part, la digitalisation de la société est à l'origine de l'augmentation considérable du volume des données disponibles (UCLouvain, 2020). Ces données d'une grande variété et quantité ont conduit à l'émergence d'une nouvelle discipline, la science des données ; changeant ainsi la façon de penser et de prendre des décisions des entreprises.

La science de données est un domaine pluridisciplinaire combinant les statistiques, les mathématiques et l'informatique. Elle est constituée de cinq étapes principales (Babu, 2019) :

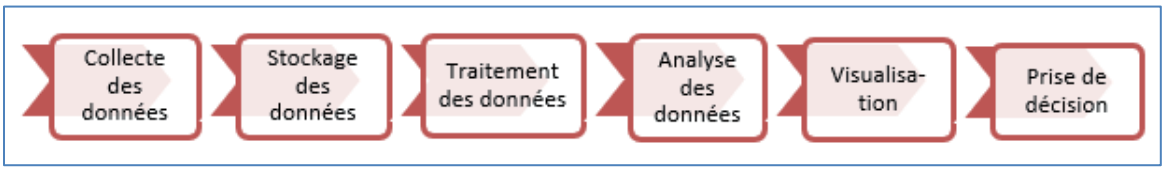

Figure 0.1 Étapes principales du processus d'application de la science de données
Tirée de Babu (2019)

La science des données, très rare dans le domaine des réseaux optiques est déjà présente dans plusieurs autres domaines tels que la finance, la météorologie, l'informatique et la démographie. Cette rareté est dû à la quasi-impossibilité d'avoir accès aux données de monitoring des réseaux optiques dans un contexte de compétition. Cependant l'application de l'analyse des données sur les données historiques serait d'un grand avantage pour les opérateurs qui pourront comprendre mieux la dynamique du réseau et la relation entre les différents paramètres.

Dans ce contexte, afin de permettre aux opérateurs une meilleure compréhension du comportement de son réseau à travers certains paramètres, l'objectif principal de cette thèse est de faire une analyse des données de performance réelles issues de la surveillance de plusieurs liens optiques. La particularité de cette thèse réside dans le fait que ces paramètres de performance sont des données réelles collectées sur plusieurs saisons dans divers types de liens optiques. L'analyse de données étant très vaste, il a été nécessaire de canaliser cette étude.

Cette thèse se subdivise en cinq parties :

La première partie traite de la revue de la littérature dans la surveillance des performances optiques et l'analyse des données.

La deuxième partie fait une description des méthodes d'analyse statistiques exploratoires et inférentielles utilisées dans cette thèse.

Dans la troisième partie, une caractérisation des données des circuits répartis par route est menée côté anomalie. L'objectif est d'étudier la relation entre les paramètres sous surveillance lorsque survient une anomalie.

Dans la quatrième partie, un regroupement des circuits sous analyse est effectué en utilisant les caractéristiques statistiques et l'algorithme de partitionnement des données k-moyennes (k-means). L'objectif est de découvrir des structures (groupes) naturels dans les circuits présentant des similarités.

Dans la dernière partie, une détection de tendance est menée en considérant la perte de la fibre optique. En effet, il est question de détecter les signes de vieillissement de la fibre optique pour les fibres présentant une certaine tendance croissante.

Les résultats de cette étude ont fait l'objet d'une publication dans le journal JLT (*Journal of Ligthwave Technology*) (Yameogo et al., 2020) et une distinction sur le site d'actualité scientifique et innovation, substance de l'ÉTS. Ils ont aussi fait l'objet d'une autre publication dans la détection d'anomalies.

CHAPITRE 1

MONITORING DE PERFORMANCE DANS LES RÉSEAUX OPTIQUES COHÉRENTS

1.1 Introduction

La demande croissante de capacités de transmission plus élevées a stimulé le développement de la transmission à grande vitesse, faisant passer le débit binaire par canal de 10 Gb/s à 100 Gb/s et bien au-delà. En raison de la largeur de bande limitée dans la bande de transmission optique, des formats de modulation d'ordre supérieur ayant une grande efficacité spectrale ont été introduits en combinaison avec la détection cohérente. Comme ces formats de modulation utilisent toutes les dimensions physiques pour encoder les données, la polarisation en particulier, un suivi actif et une compensation des effets de polarisation variant dans le temps sont nécessaires.

De plus, il est important de souligner que les réseaux de télécommunication optique actuels sont constitués de systèmes de différents formats de modulation à différents débits de données, de plusieurs composants et équipement optiques pouvant induire des distorsions du signal optique dans le temps (Chan, 2010). Tout cela offre une certaine flexibilité, dynamicité et une meilleure utilisation de la capacité de transmission disponible.

Au regard d'une telle complexité, il est donc d'une importance critique pour les fournisseurs de réseau de trouver le moyen de prévenir toutes défaillances du réseau afin d'éviter de grandes pertes de données et garantir une certaine qualité de service (*QoS*). Afin de réagir rapidement à toute influence néfaste sur la qualité du signal, il ne suffit pas de se fier au taux d'erreur sur les bits (*Bit Error Rate, BER*) avant la correction des erreurs par le FEC (*Forward Error Correction*) au récepteur pour garantir la qualité de service. Il est aussi nécessaire de surveiller l'état du réseau au travers de certains paramètres directement à partir de la couche optique, d'autant plus que le BER ne permet pas d'identifier l'origine de la défaillance (Hauske et al., 2009).

Cela se définit comme étant la surveillance de performance optique (*Optical Performance Monitoring, OPM*) et constitue une caractéristique essentielle et indispensable pour les réseaux optiques.

Ce chapitre présente donc la surveillance (*communément monitoring*) de performance optique dans les réseaux optiques de grande capacité en utilisant les transpondeurs cohérents. Une revue de la littérature des travaux déjà menés est également effectuée. Par la suite, des techniques d'analyse de données, qui seront utilisées pour analyser les données monitorées dans le cadre de ce projet sont présentées.

1.2 Les transpondeurs cohérents

L'évolution rapide des systèmes de communication optique longue distance est due à l'adoption progressive de formats de modulation à plusieurs niveaux à grande efficacité spectrale, combinés avec le multiplexage en polarisation (*Polarization Division Multiplexing, PDM*) et la détection cohérente assistée par le traitement numérique des signaux (*DSP, Digital Signal Processing*) (Roudas, 2012). Le transpondeur cohérent utilisé dans le cadre du projet fait partie de la plate-forme optique OM6500 de Ciena. Il s'agit d'un transceiver cohérent, dit OCLD (*Optical Channel Laser and Detector*), fonctionnant à 100 Gbit/s utilisant le format de modulation PDM-QPSK ou DP-QPSK (*Dual Polarization Quadrature Phase-Shift Keying*). Ce format de modulation est également le format recommandé par l'OIF (*Optical Internetworking Forum*) pour le débit de 100 Gbit/s (Pierre, 2014).

La Figure 1.1 illustre le fonctionnement d'un système optique cohérent, modulé en quadrature de phase et multiplexé en polarisation. À la transmission, les signaux clients à 100 Gbit/s, sont multiplexés et pré-codés avec un FEC ayant un en-tête de sept pour cent. Ces bits sont multiplexés à la valeur de 11.5 Gbaud ou 14 Gbaud, correspondant au nombre de symboles transmis par seconde. Quatre flux de données modulent alors la lumière d'un laser accordable aux travers des quatre modulateurs, codant les signaux en phase et en quadrature de phase dans les deux polarisations X et Y.

Ils sont par la suite combinés dans un seul signal optique puis transmis sur le lien optique. Ce lien est typiquement une cascade de segments de fibre optique et d'amplificateurs, de filtres et de ROADM (*Reconfigurable Optical Add Drop Multiplexers*) (Roberts et al., 2009b). À la réception, le signal optique est divisé suivant les deux polarisations orthogonales, par un diviseur de polarisation (*Polarization Beam Splitter, PBS*) qui sert de référence de polarisation locale, puis mixé avec un laser, servant d'oscillateur local. Chaque polarisation du signal est ensuite divisée et associée à chacune des deux phases orthogonales de la lumière polarisée de l'oscillateur local. Cela est effectué en utilisant des diviseurs optiques, des déphaseurs (*Phase Shift*) et des coupleurs, afin de fournir quatre signaux optiques orthogonaux aux quatre photodiodes. Le module *"Agile Engine"* permet de recouvrer les bits transmis à partir du signal reçu. Le module FEC effectue une correction d'erreurs. Une copie exacte des signaux clients originaux est par la suite retrouvée dans le module *"Tributary"* (Roberts et al., 2009a).

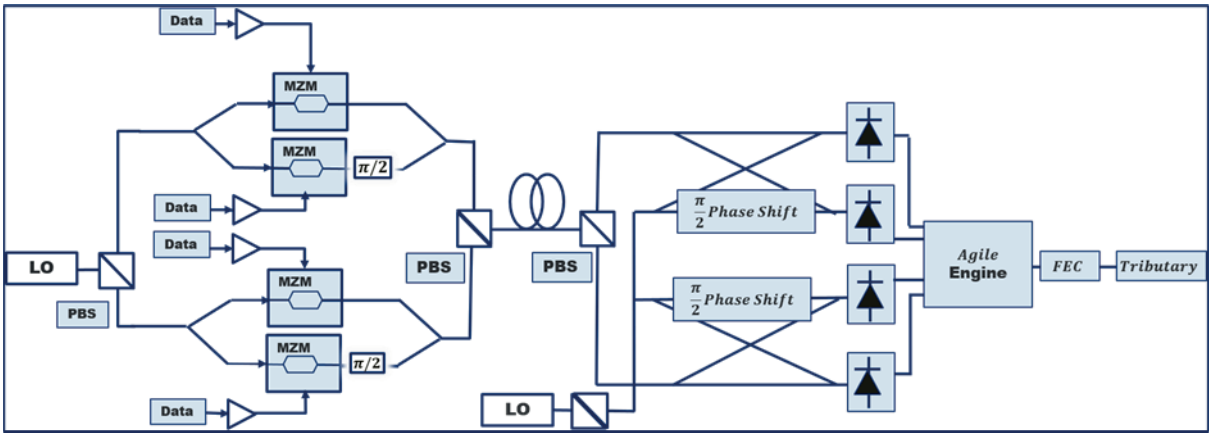

Figure 1.1 Principe d'un système optique cohérent DP-QPSK
Adaptée de Roberts et al. (2009b)

Contrairement à la détection directe, le récepteur cohérent à polarisation multiple préserve toutes les propriétés du signal et transfère linéairement le champ optique dans le domaine électrique.

Le récepteur cohérent numérique compense non seulement certaines dégradations physiques linéaires déterministes (dispersion chromatique, dispersion de mode de polarisation), mais permet également une estimation systématique des paramètres et un monitoring de

performance complet. En effet, dans les systèmes de détection cohérents, l'information complète du signal optique est disponible dans le domaine électrique, ce qui permet de mesurer la qualité du signal, l'extension de la gestion des fautes, et de juger la qualité de service (Hauske et al., 2009).

1.3 Surveillance de performance dans les réseaux optiques utilisant des transpondeurs cohérents

La surveillance de performances optiques consiste en l'estimation et l'acquisition de différents paramètres physiques des signaux transmis et des divers composants d'un réseau optique. Les fonctionnalités de l'OPM sont indispensables pour assurer un fonctionnement robuste du réseau et jouent un rôle clé car elles permettent la flexibilité et l'amélioration de l'efficacité globale du réseau. De plus, la surveillance permanente de l'état physique du réseau permet de réduire les coûts d'exploitation, d'assurer une utilisation optimale des ressources et de garantir un fonctionnement et une gestion adéquats des réseaux optiques dynamiques (Dong et al., 2016).

1.3.1 Les principaux défis de la surveillance de performance

L'objectif de la surveillance optique est de fournir des applications de gestion de la configuration, des défauts et des performances avec des informations fiables et périodiques sur l'état des entités de transport à chaque couche d'un réseau de transport optique (Chan, 2013). En d'autres termes, il vise à favoriser un fonctionnement autonome, robuste et rentable du réseau en permettant :

- le monitoring dynamiquement de l'état physique du réseau et de la qualité des signaux de données s'y propageant,
- le diagnostic et la réparation du réseau,
- la redirection du trafic et l'allocation dynamique des ressources, y compris la longueur d'onde, la puissance du signal, la compensation, l'égalisation accordable, le codage des données, la détermination du chemin et la largeur de bande du canal.

Cela représente un défi énorme dans les réseaux optiques de capacité de plus en plus élevée. En effet, ces réseaux sont caractérisés par une variété de formats de données et de débits, ainsi qu'une croissance et un changement permanent dans la structure du réseau (ajout de nouveaux équipements, de longueurs d'onde…). Dans ce contexte de trafics hétérogènes de plus en plus complexes, il devient difficile de prévoir et de gérer les dégradation des signaux de données, étant donné que les phénomènes de dégradations varient avec le temps (Zhongqi, Changyuan et Willner, 2010). Cela témoigne ainsi de la nature encore plus complexe de ces systèmes à grande capacité, surtout lorsque plusieurs canaux de longueurs d'onde différentes se propagent simultanément tout au long de la fibre optique (Chan, 2010).

1.3.2 Les paramètres réseaux nécessitant le monitoring

Outre la fibre optique, d'autres composants du réseau peuvent également être à l'origine de plusieurs dégradations. Les dégradations des réseaux optiques peuvent généralement être classées en deux catégories : les dégradations de nature « *catastrophique* » et les dégradations de nature « *non catastrophique* ». Les dégradations « *catastrophiques* » entraînent une perte de puissance optique et comprennent les ruptures de fibres, les défaillances des composants du réseau et les équipements de réseau mal installés, etc. Les dégradations non catastrophiques ne diminuent pas nécessairement la puissance du signal optique, mais elles peuvent gravement déformer le signal. Ces distorsions peuvent être de nature linéaire ou non linéaire et doivent être minimisées ou compensées de manière appropriée afin de garantir les performances souhaitées du réseau (Chan, 2013).

De ce fait, plusieurs paramètres méritent d'être surveiller de façon permanente pour garantir les accords de niveau de service (*Sevice Level Agreements, SLA*) avec le client.

De plus, ces réseaux optiques sont très sensibles aux problèmes de dégradation des signaux dits « non catastrophique » dus à la propagation dans le canal.

Cela signifie donc qu'il y a un signal de puissance suffisant mais les bits de données ne peuvent être recouverts à cause des phénomènes physiques. La Figure 1.2 montre une liste typique des paramètres monitorés pour une supervision de la qualité du signal dans la couche optique.

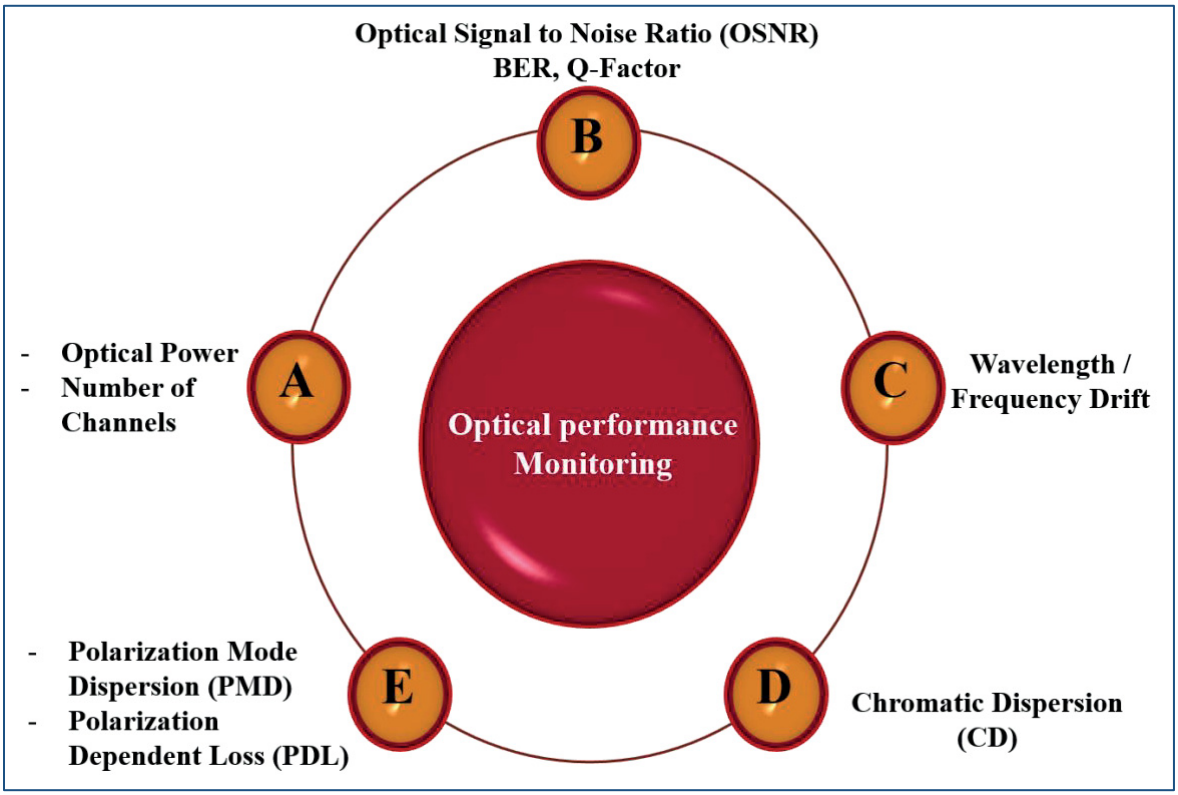

Figure 1.2 Paramètres monitorés dans la couche optique d'un système cohérent
Adaptée de Campos, Jia et Wolcott (2018)

La puissance optique du signal

Le paramètre le plus fondamental à surveiller dans un réseau optique pour mesurer la qualité d'un signal est la puissance.

La puissance optique peut diminuer considérablement en raison de l'atténuation des fibres et des pertes rencontrées au niveau des connecteurs, des épissures et des coupleurs ou des ruptures de fibres. Elle est généralement exprimée en échelle logarithmique (rapport de puissance en décibels entre la puissance mesurée et un milliwatt, dBm).

Dans les systèmes multiplexés par répartition en longueur d'onde (*Wavelength Division Multiplexing, WDM*), les informations sur la puissance de chaque canal sont nécessaires pour égaliser dynamiquement la puissance par un mécanisme de rétroaction, assurant ainsi une performance stable du système. La puissance du signal peut être mesurée dans diverses parties d'une liaison optique, notamment à travers des diviseurs de puissance avec un faible rapport (*taps*) qui se trouvent à divers endroits de la liaison, par exemple au niveau du transmetteur, du récepteur ou des amplificateurs (Brugiere, 2015; Dong et al., 2016).

Le rapport signal sur bruit optique

Les amplificateurs optiques comme les amplificateurs à fibres dopées à l'erbium (*Erbium Doped Fiber Amplifier, EDFA*) sont utilisés dans les réseaux optiques pour compenser les pertes de transmission sur de longues distances. Cependant, en plus de fournir un gain optique, les EDFA sont à l'origine de la principale source de bruit provenant de l'émission de photons issus du phénomène d'amplification spontanée (*Amplified Spontaneous Emission, ASE*). Ce bruit s'accumule après plusieurs cascades d'amplificateurs et causera des erreurs de détection du signal. Le rapport de la puissance du signal sur le bruit ASE est appelé le rapport signal sur bruit optique (*Optical Signal to Noise Ratio, OSNR*). L'OSNR est fréquemment cité comme une mesure majeure de la performance du système optique puisque l'ASE est la principale source de bruit (Chan, 2010). Il est généralement exprimé en décibel (dB).

Le taux d'erreur binaire et le facteur de qualité (*Q-Factor*)

Le taux d'erreurs binaire (*Bit Error Rate, BER*) et le facteur Q sont deux paramètres importants pour évaluer la performance globale d'un réseau optique.
Le monitoring du BER a traditionnellement été utilisé comme un outil préférable pour caractériser la qualité d'une liaison optique. Il représente la probabilité d'erreur binaire au récepteur. Le facteur Q est utilisé pour analyser la performance des systèmes de transmission pour lesquels la mesure directe du BER n'est pas pratique. Il est également un indicateur de la qualité du signal optique en raison de la forte corrélation entre le facteur Q et le BER.

Avec l'introduction de codes correcteurs d'erreurs (*Forward Error Correction, FEC*) très performants, le BER maximum peut aller jusqu'à la valeur de 3.8×10^{-3} (Roberts et al., 2009b). Au titre du taux d'erreur binaire, le contenu des bits erronés dans le signal est aussi un paramètre clé qui doit être surveillé en permanence.

En effet, la responsabilité des systèmes et des réseaux de communication est de transporter en temps réel au terminal récepteur, les mêmes bits que le terminal émetteur a envoyés. Pour garantir cela, les normes définissent des paramètres de performance relatifs à la fréquence des erreurs de bits, tels que les secondes erronées (*Error Seconds, ES*), les secondes erronées graves (*Severely Error Seconds, SES*), (*Severely Errored Frame Second, SEFS*) et le nombre élevé de corrections en seconde (*High Correction Count Second, HCCS*), ainsi que la manière de mesurer et rendre compte de ces paramètres de performance (Kartalopoulos, 2001).

La dispersion chromatique (CD) et la dispersion de mode de polarisation (PMD)

En dehors des paramètres sus mentionnés qui offre un monitoring fiable de la qualité du signal optique, d'autres paramètres nécessite d'être monitorés : la CD (*Chromatic Dispersion*) et la PMD (*Polarization Mode Dispersion*). Ces deux paramètres appartiennent à la catégorie des dégradations dispersives et influencent directement la qualité des impulsions lumineuses transmises. Ce sont des paramètres essentiels à surveiller dans les réseaux optiques dynamiques. En effet, dans les réseaux optiques reconfigurables, les changements de longueur de chemin (distance) pour un canal donné (dus à la commutation) peuvent entraîner des quantités variables de CD accumulés. Il est donc essentiel de compenser la CD de manière adaptative avec d'étroites tolérances (Dong et al., 2016).
La PMD doit également être surveillée efficacement, car elle constitue une limitation majeure dans les réseaux de fibres optiques fonctionnant à des débits supérieurs à 40 Gb/s. Les effets de la PMD sont stochastiques, variables dans le temps et en fonction de la température et du débit de données (Chan, 2013).

En plus de tous ces facteurs, la longueur d'onde et le décalage de la fréquence (*Frequency drift*) sont aussi monitorés afin de vérifier que le signal optique est correctement placé dans son canal (Dong et al., 2016).

1.3.3 Les paramètres réseaux monitorés dans le cadre de ce projet

Les paramètres monitorés dans le cadre de ce projet sont représentés à la Figure 1.3 ainsi que l'emplacement du réseau où ils sont mesurés.

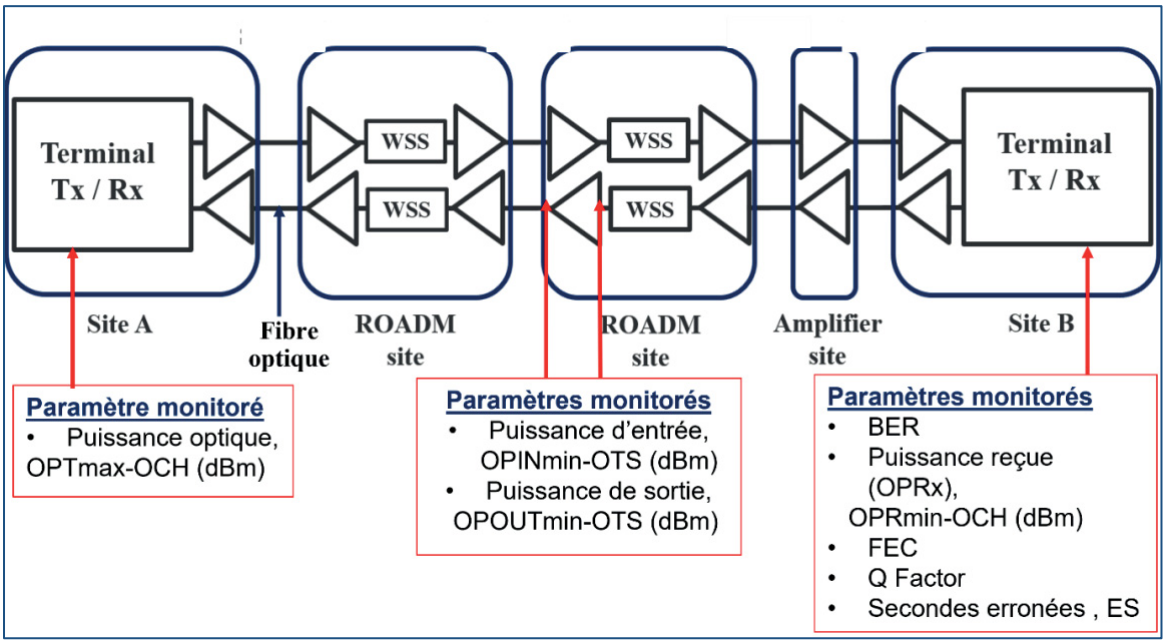

Figure 1.3 Architecture d'une liaison optique bidirectionnelle typique:
ROADM: Reconfigurable Optical Add/Drop Multiplexer, WSS: Wavelength Selective Switch, OPT: Optical Power Transmitted, OPIN: Optical Power Input, OPOUT: Optical Power Output, OPR: Optical Power Received, OCH: Optical Channel, OTS: Optical Transmission Section, ES: Errored Second

Plusieurs métriques de performance ont été monitorées dans la couche physique. Au niveau du transmetteur, la puissance optique est mesurée. Les puissances d'entrées et de sortie mesurées au niveau des amplificateurs ont permis le calcul des gains d'amplification et des pertes de spans optiques. La majorité des métriques est mesurée au niveau du récepteur. Il s'agit du BER, la puissance reçue au récepteur, le FEC et le facteur Q.

En plus de ces métriques de couche physique, au récepteur, des alarmes sont également monitorées. Dans ce cas-ci, c'est le système qui signale la présence d'erreurs. C'est le cas des secondes erronées qui représentent le nombre de secondes pendant lequel il y a une erreur.
Les termes OPIN, OPOUT et OPR sont les noms des paramètres tel que relevé par le système de contrôle.

1.3.4 Processus de prétraitement de données de monitoring

Les paramètres monitorés sur une période d'un an toutes les 15 min, sont reçus en format brut sous forme de fichiers *csv* (*comma separated values*), ce qui représente 20 Go de données. Une telle quantité de données mérite d'être placée sur une base de données pour une analyse effective. De ce fait, une étude minutieuse des données a été faite dans un premier temps dans le but de regrouper les paramètres par connexion et de définir un modèle qui permet de les enregistrer dans une base de données. Le modèle dimensionnel détaillé de la base de données est décrit en détail dans (Djeukui, 2017).

La solution proposée est représentée par l'architecture à la Figure 1.4 ci-dessous.

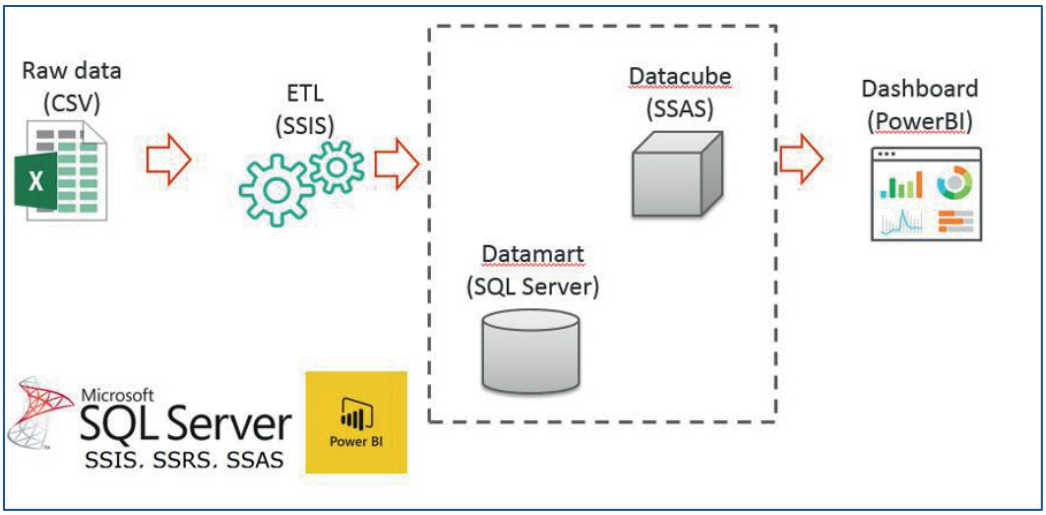

Figure 1.4 Architecture de l'environnement de développement de la base de données. ETL : Extract, Transform and Load,
SSRS : SQL Server Reporting Services

Les outils suivants ont été choisis pour le déploiement de l'environnement de développement:

- *SQL* (Structured Query Language) *Server 2014* pour la base de données,
- *SQL Server Management studio 2014 (SSMS) pour l'environnement de* développement de la base de données,
- *SQL Server Data Tools* pour le système de gestion de base de données,
- *Microsoft Visual Studio Ultimate 2013* pour l'environnement de programmation qui intègre SSIS (SQL Server Integration Services) pour l'intégration des données et SSAS (SQL Server Reporting Services) pour l'analyse et la création des data cubes,
- *NotePad* ++ pour la création des scripts SQL et,
- *Power BI Desktop* pour la création des rapports.

Afin de créer la base de données, un script SQL a été développé dans l'outil Notepad++ et exécuté dans SSMS pour pouvoir avoir la base de données au nom de l'opérateur. Une fois créée, la base de données est peuplée à l'aide des fichiers *csv* reçus en utilisant le flux de contrôle SSIS dans Visual Studio. Ces fichiers contiennent entre autres les informations du site, de la connexion, des transpondeurs et de la date complète.

Certains fichiers ont nécessité une transformation préalable avant d'être utilisables. En effet, les fichiers csv avaient des configurations différentes à la réception. Certains avaient des virgules (,) comme délimiteur et d'autres des points-virgules (;). Dans le même ordre d'idées, certains fichiers avaient une colonne supplémentaire à la fin et d'autres non. Il a donc fallu transformer ces fichiers pour pouvoir les uniformiser. Le délimiteur utilisé est la virgule et tous les fichiers ont une colonne vide à la fin suivie d'une virgule.

Un tableau de bord (Dashboard) a été construit pour visualiser les connexions sur une carte géographique et d'afficher l'évolution temporelle des différents paramètres monitorés.

1.4 Les perspectives de la surveillance de performance

Le monitoring de performance dans sa définition même permet de rendre le réseau optique intelligent. Dans ce sens qu'il rendra le réseau autonome et dynamique. Pour y parvenir, il est impératif d'avoir une connaissance approfondie du fonctionnement du réseau, de son comportement vis-à-vis de tous les facteurs susceptibles de détériorer ses performances. Cela permettra de pouvoir quantifier les fluctuations de la performance et même de le prédire. Plusieurs travaux sur le monitoring de performance ont été menés et deux groupes basés sur le type de déploiement de la liaison se distinguent dans: Les liaisons enfouies et les liaisons aériennes.

1.4.1 Les liens enfouis

Dans son étude, (Michel, 2016) démontre que l'analyse de la variation temporelle des paramètres de performance tel que le BER laisse paraître un comportement saisonnier. Ce qui signifie que le BER est plus haut en hiver qu'en automne ou en été. L'évolution temporelle de la PDL (*Polarization Dependent loss*) montre des pics les après-midis et les jours de semaines contrairement aux fins de semaine. Ce qui laisse penser qu'il y a une corrélation entre l'activité humaine et la PDL.

Par ailleurs, même si certaines corrélations entre la température et les paramètres de performance ont pu être établies, le lien enfoui démontre des fluctuations mineures en termes de performance et d'effets de polarisation par rapport aux liens aériens. Cela confirme les travaux de (Haro et Horche, 2008) et de (Woodward et al., 2014).

1.4.2 Les liens aériens

Contrairement aux liens enfouis, les liens aériens montrent une variation rapide de l'état de la polarisation. (Ednay, 2010) a étudié la dynamique des effets de polarisation dans une variété d'environnement : lors du passage de train, autour d'un fil de terre d'une ligne de distribution d'électricité de 132 KV. Il ressort que ces différents environnements ont un effet significatif sur l'état de la polarisation.

Une seconde étude un peu similaire (câble aérien + câble à côté d'une voie ferrée), mais sur le plus long terme a été menée par (Brugiere, 2015). Il montre une corrélation entre la température, l'environnement de déploiement, la vitesse du vent, ou l'activité humaine et la performance du réseau optique.

1.4.3 Conclusion

L'un des objectifs à long terme du monitoring est de rendre le réseau proactif. Jusque-là, les études se sont limitées à montrer l'impact des facteurs externes tels que la température, la vitesse du vent, l'environnement de déploiement sur la polarisation. Ces études ont été menées sur des réseaux optiques expérimentaux (boucle de recirculation optique). Les distances déployées sont de 24 km et 80 km dans le cas de (Ednay, 2008) sur une période de 4 jours. (Michel, 2016) a utilisé un transpondeur cohérent à 100 Gb/s sur une distance de 243 km dans un anneau métropolitain sur 6 mois. (Brugière, 2015) quant à lui, a fait son étude sur 2 ans mais dans une fibre aérienne de 750 km +250 km enfouis à côté d'une voie ferrée avec un transpondeur à 40 Gb/s. Pratiquement toutes ces études s'intéressent à la polarisation sans la lier forcément à la performance optique de la liaison.

Le manque de données réelles et de connaissance de la topologie physique explique l'absence d'études des effets des facteurs internes sur la performance du réseau. En effet, les opérateurs monitorent régulièrement leur réseau mais ces données pour des raisons de sécurité et de confidentialité sont tenues secrètes. De plus, ces données ne sont pas exploitées mais sont utilisées seulement en cas de panne. Cela explique le fait qu'à date aucune étude complète de la fluctuation de performance n'ait été menée.

L'avantage de cette thèse est que l'étude se base sur une grande capacité de données réelles monitorées sur une période d'une année, issues d'une variété de type de liens. Dans le prochain paragraphe, une exploration des techniques d'analyse de donnée sera menée.

1.5 Techniques d'analyse de données

1.5.1 Introduction

La révolution du big data (entendre par là l'explosion de la disponibilité des données) promet de transformer la façon de vivre, travailler et penser, en permettant l'optimisation des processus, en favorisant l'extraction d'informations, et en améliorant la prise de décision des organisations. La réalisation de ce grand potentiel repose sur la capacité d'extraire de la valeur de ces données historiques massives grâce à l'analyse des données (L'heureux et al., 2017).

L'objectif de cette seconde section est de faire une revue de la littérature des techniques d'analyse de données qui pourraient s'appliquer à ce projet, et un récapitulatif des travaux effectués dans le domaine optique. Plusieurs techniques d'analyse de données sont citées dans la littérature mais elles sont regroupées principalement en deux catégories à savoir:

- l'analyse statistique et ,
- l'analyse prédictive basée sur les méthodes d'apprentissage machine.

1.5.2 L'analyse statistique

L'analyse statistique dont il est question se résume en l'analyse descriptive et en l'analyse exploratoire des données. L'analyse descriptive est la forme la plus courante de l'analyse des données car son objectif premier est de présenter les données observées sous une forme qui facilite la prise de connaissance sans autre interprétation. L'analyse exploratoire des données va plus loin et s'appuie sur l'analyse descriptive en recherchant des découvertes, des tendances, des corrélations ou des relations entre les mesures pour générer des idées ou des hypothèses (Leek et Peng, 2015).

Les données ou variables utilisées dans l'analyse descriptive et exploratoire peuvent être de trois natures distinctes : quantitative, qualitative et temporelle.

- les variables quantitatives sont des données qui peuvent être mesurées, écrites avec des nombres et qui peuvent se prêter à une opération mathématique simple telle une addition. Elles peuvent être de nature discrète représentée par des nombres entiers, ou de nature continue, représentée par des nombres fractionnaires. Pour une meilleure compréhension, le nombre d'amplificateurs dans le lien optique est une variable discrète tandis que le SNR est une variable continue (Schyns, 2011).
- les variables qualitatives ou catégoriques quant à elles, représentent une qualité et ne sont pas traduit par un nombre. Elles se décomposent en trois familles (Komorowski et al., 2016) :
  - les variables qualitatives nominales qui ne peuvent être classées : le type de liens optiques (enfoui ou aérienne).
  - les variables qualitatives ordinales qui induisent une notion de classement : classement de l'activité du SNR dans les liens optiques (faible activité du SNR, activité modérée du SNR, et activité élevée du SNR).
  - les variables qualitatives binaires qui ne peuvent pas être hiérarchisées et qui possèdent deux classes possibles : (anomalies = oui si la donnée en est une, et non sinon).

- les variables temporelles peuvent être de deux types : date (19-01-2017) et horaire (19 : 30) et peuvent être exploitées mathématiquement (Dérobert, 2020).

Dépendamment de leur nature, les données peuvent être résumées de plusieurs manières :

- les fréquences, les distributions entre autres sont utilisées pour examiner le nombre d'occurrences de valeurs dans un groupe ou un intervalle particulier.
- le calcul de la tendance centrale : la moyenne, la médiane et le mode des données, résume les données en une seule valeur qui est typique ou représentative de toutes les valeurs de l'ensemble de données.
- l'évaluation de la dispersion : la plage, les quartiles, la variance ou l'écart type, montre à quel point les valeurs sont dispersées et à quel point elles diffèrent de la valeur moyenne.
- la visualisation ou la présentation de ces données et analyses est un moyen important de transmettre des informations. Bien qu'il existe diverses façons de le faire sous forme graphique, les visualisations à l'aide de graphiques à barres, de boîte à moustaches (Boxplot) et de nuage de points sont des approches courantes (Medecine, 2017).

1.5.3 L'analyse prédictive

L'analyse prédictive va encore plus loin que l'analyse descriptive et exploratoire en extrayant des informations de l'ensemble de données afin de déterminer des structures (*patterns),* prédire des résultats ou prédire la tendance (Medecine, 2017). L'objectif principal est donc d'apprendre à partir des données. Pour ce faire, l'outil d'analyse prédictive retenu dans ce travail est l'apprentissage machine. L'apprentissage machine (*Machine learning, ML*) est un terme qui a été donné au domaine d'étude, qui attribue aux ordinateurs la capacité d'apprendre sans être explicitement programmé. En d'autres termes, L'apprentissage machine développe des techniques par lesquels l'ordinateur peut reconnaître des modèles complexes et prendre des décisions intelligentes directement à partir des données (Han, Kamber et Pei, 2012a; Ratner, 2017).

Ces techniques se répartissent en trois grandes classes à savoir les techniques supervisées, les techniques non supervisées et les techniques de renforcement. Seules les deux premières techniques seront détaillées dans cette thèse.

Les techniques supervisées

Étant donné un ensemble d'apprentissage (*training set*) de N pairs d'exemples entrée / sortie : (x_1, y_1), (x_2, y_2), … (x_N, y_N), où chaque étiquette y_j (*label*) a été généré par une fonction inconnue $y - f(x)$, le principe général de l'apprentissage supervisé est de trouver une fonction h qui approxime la véritable fonction f (Russell et Norvig, 2002). Les variables x_1 et y_1 ne sont pas forcément numériques. La fonction h est une hypothèse. L'apprentissage consiste à chercher parmi tant d'autres hypothèses, celle qui performe le mieux, même avec de nouvelles entrées. Pour mesurer la précision d'une hypothèse, un ensemble de données de test (*test set*) distincts de l'ensemble d'apprentissage (*training test*) sera implémenté.

L'hypothèse est dite bien se généraliser, si elle prédit correctement la valeur de y pour les données de test (Russell et Norvig, 2002).

Si la variable de sortie y est une variable qualitative (*class label*) alors, il s'agit d'un problème de classification. Elle est appelée classification booléenne ou binaire si elle a seulement deux valeurs. Cependant, si y est quantitative, le problème d'apprentissage est une régression.

Les algorithmes d'apprentissage machine supervisés les plus courants sont : l'algorithme du plus proche voisin (*k-nearest neighbour, K-NN*), la régression logistique et l'algorithme machine à vecteurs de support (*Support Vector Machine, SVM*) souvent appelé séparateur à vaste marge (Besse, 2005; Dey, 2016), au titre de la classification. Pour ce qui est de la régression, la régression linéaire et la régression polynomiale peuvent être citées. Certains algorithmes tels que les réseaux de neurones peuvent être utilisés tant en classification qu'en régression (L'heureux et al., 2017).

Les techniques non supervisées

Contrairement à l'apprentissage supervisé, l'apprentissage non supervisé traite le cas où on dispose seulement des entrées (x) sans avoir au préalable les sorties.
Trois types d'apprentissage non supervisé existent à savoir : l'association, La réduction de dimensionnalité et le regroupement.

- l'association est utilisée pour découvrir une co-occurrence entre deux variables.
- la réduction de dimensionnalité sert à réduire le nombre de variables d'un ensemble de données (*feature selection*), tout en garantissant que les informations importantes sont toujours transmises. La méthode populaire pour la réduction de dimensionnalité est l'analyse en composante principale (*principal component analysis, PCA*).
- le regroupement vise à construire des groupes (clusters) d'objets similaires à partir d'un ensemble hétérogène d'éléments.

Chaque sous-ensemble regroupe des éléments ayant des caractéristiques communes qui correspondent à des critères de proximité. Le but des algorithmes de regroupement est donc de minimiser la distance intra-classe (grappes d'éléments homogènes) et de maximiser la distance inter-classe afin d'obtenir des sous-ensembles le plus distincts possible. La mesure des distances (encore appelées métriques) est un élément prépondérant pour la qualité de l'algorithme de regroupement et il est définit par l'utilisateur (Calas, 2006; Mahammed, 2018). Par ailleurs, le regroupement fait appel à des méthodes de classification ascendante hiérarchique ou à des algorithmes de réallocation dynamique k-moyenne (*k-means en anglais*).

1.5.4 L'apprentissage machine dans les réseaux optiques

L'explosion de la quantité des données a atteint un niveau sans précédent à cause des développements dans les technologies web, les réseaux sociaux, les appareils mobiles.

Ces mégadonnées offrent un potentiel énorme en termes de valeur commerciale dans une variété de domaines tels que les soins de santé, la biologie, les transports, la publicité en ligne,

la gestion de l'énergie et les services financiers. Ce potentiel est rendu exploitable au travers de techniques d'analyse de données basées sur l'apprentissage machine (L'heureux et al., 2017).

Par ailleurs, la révolution des mégadonnées implique des réseaux capables de délivrer de telle bande passante : les réseaux optiques. Afin de satisfaire à cette demande incessante, les réseaux optiques de plus en plus hétérogènes, ont atteint un niveau de complexité tant dans la conception des composants que dans la gestion elle-même.

Face à une telle complexité il est nécessaire de penser à une gestion plus « intelligente » (*cognitive management*) et à des méthodes de monitoring de performance plus modernes afin d'affronter les différents défis qui s'y présentent. Déjà connu dans plusieurs domaines, et encore plus proche dans les réseaux sans fil, l'apprentissage machine a aussi séduit le monde des réseaux optiques de par ses avantages. Il a donc été appliqué pour la résolution de problématiques liées à ces réseaux dans les dernières années (Tremblay, 2020).

Dans ce sens, plusieurs travaux portant entre autres sur l'estimation de la qualité de transmission (Qualité of transmission, *QoT*), l'évaluation de la performance de transmission, le contrôle des amplificateurs optiques, la prédiction de pannes (*Fault prediction*) (Musumeci et al., 2018) et la détection d'anomalies.

Monitoring de performance et estimation de la qualité de transmission

Dans les réseaux de transport optique, la qualité de la transmission (QoT) est estimée avant d'approvisionner de nouvelles connexions ou de mettre à niveau les connexions existantes. Traditionnellement, un modèle de couche physique *(physical layer mode, PLM*) était utilisé pour l'estimation de la QoT, avec des marges élevées pour tenir compte de l'inexactitude du modèle et de l'incertitude quant à la fluctuation des conditions de la couche physique.
La réduction des marges augmente l'efficacité du réseau mais nécessite une estimation précise de la qualité de service (Sartzetakis, Christodoulopoulos et Varvarigos, 2019). Dans cette

optique, plusieurs travaux ont été menés à date (Gu, Yang et Ji, 2020; Mata et al., 2018; Musumeci et al., 2018).

Rottondi et al. (2018) proposent une méthode de prédiction de la QoT d'une liaison optique avant son déploiement. Sur la base de caractéristiques de la liaison : la longueur totale du lien, la longueur du plus long lien et le nombre de liens optiques; du format de modulation et le débit, les auteurs prédisent si le BER de la nouvelle liaison excèdera le seuil. L'algorithme de classification Random Forest est utilisé.

À quelque paramètres près et avec plus de techniques d'apprentissage machine que l'étude précédente, Aladin et Tremblay (2018) ont proposé un outil cognitif qui prend en compte les effets de dégradation linéaire et non linéaire en utilisant des données de BER synthétiques. Elles ont appliqué trois techniques de classification d'apprentissage machine, la forêt aléatoire (*Random Forest, RF*), le SVM et le KNN, ainsi que les paramètres du signal tels que la longueur totale du lien, la longueur du span, la puissance du canal, le format de modulation et le débit. Les résultats démontrent que le SVM performe le mieux et pourrait être utilisé pour la classification de la liaison optique.

L'algorithme SVM a été aussi testé sur une base de données plus grande donnant une précision de classification plus grande et un taux de faux positif plus bas (Tremblay, Allogba et Aladin, 2019). Parallèlement un type d'algorithme de réseaux de neurones récurrents LSTM (*Long Short Term Memory*) est utilisé pour la prédiction du SNR sur un horizon de 24 heures sur la base de données historiques réelles d'un lien optique, ce qui constitue une première dans le domaine.

Allogba et Tremblay (2018) proposent une classification du BER basée sur l'algorithme KNN, en utilisant des données réelles de BER (valeur moyenne et maximale), la température externe, la période du jour. Dépendamment de la combinaison des attributs, la précision du modèle peut atteindre 97.8%.

Techniques de gestion des défaillances (*Failure*)

Les défaillances des réseaux optiques peuvent entraîner une dégradation des performances du réseau et même une énorme perte de données. Il existe principalement deux types de méthodes de gestion des défaillances: les méthodes réactives et les méthodes proactives.
Dans les méthodes réactives, les opérateurs de réseau prennent des mesures lorsque des alarmes se produisent. Cependant, la perte de données s'est déjà produite et le fonctionnement manuel peut ne pas faire face à des alarmes massives et des défaillances invisibles dans les aspects de flexibilité et de rapidité.

Ainsi, l'introduction des techniques d'apprentissage peut rendre le système proactif, capable de détecter les défaillances potentielles et donner suffisamment de temps pour les actions de restauration (Gu, Yang et Ji, 2020).

(Shahkarami et al., 2018) ont développé et testé plusieurs méthodes d'apprentissage machine pour la détection et l'identification de défaillances des équipements dans les réseaux optiques. À partir des données de BER monitorées sur 24 heures, les auteurs ont extrait certaines caractéristiques statistiques telles que les valeurs minimum, maximum et moyenne, l'écart-type ainsi que les composantes spectrales en appliquant la transformée de Fourier. Deux modules ont été entrainés : le module de détection de défaillance et le module d'identification de défaillance. Le premier utilise différentes méthodes de classification pour la détection des anomalies, à savoir SVM, forêt aléatoire (*Random Forest, RF*) et un réseau de neurones à une couche cachée. Dans le second cas, un réseau de neurones à deux couches est appliqué avec 98 % de précision contre 100% dans le premier.

Dans la même lancée, (Wang et al., 2017) ont aussi prédit la défaillance des équipements dans un réseau optique mais en utilisant une méthode de prédiction combinant le SVM et le double lissage exponentiel (*double exponential smoothing, DES*). Les auteurs considèrent que la défaillance des équipements de réseaux optiques peut être quantifiée par les métriques recueillis par les opérateurs de réseaux pendant leurs périodes d'utilisation.

Ces métriques sont la puissance optique, l'intensité du laser, la température ambiante, la température interne du module, la longueur d'onde centrale, l'OSNR, la consommation électrique et les autres paramètres que les opérateurs de réseau enregistrent quotidiennement et constituent les attributs du modèle. Le DES est utilisé dans un premier temps pour effectuer la prédiction à $t + T$ pour chaque métrique et à partir de là, le SVM sur la base de cette valeur juge si l'équipement est en état de défaillance ou non.

Dans une autre étude, (Vela et al., 2017) ont étudié les effets de plusieurs défaillances de la couche optique dues au filtrage et à l'interférence inter-canal, sur la qualité de la transmission, à partir de la puissance reçue et le pre-FEC BER obtenue expérimentalement. Ils ont proposé un algorithme de localisation de défaillance probabiliste basé sur les réseaux bayésiens (*Bayesian Networks, BN*) et implémenté avec les caractéristiques statistiques des données telles que le minimum, le maximum, la moyenne, la tendance, la taille et la présence du niveau de changement du signal. Lorsqu'une erreur excessive est détectée, l'algorithme localise et identifie la cause la plus probable de l'erreur dans la couche optique.

Contrairement aux deux premières analyses, (Panayiotou, Chatzis et Ellinas, 2018) utilisent des méthodes d'apprentissage statistiques pour localiser des défaillance dans un seul lien dans le cas d'un réseau optique transparent. En effet, un classificateur de processus gaussien est entrainé sur des données historiques, dans le but de modéliser et de prédire la probabilité de défaillance de chaque lien qui s'y trouve.

1.6 Conclusion

L'apprentissage machine a atteint un niveau de maturité sans précédent et son efficacité n'est plus à démontrer à en juger par le nombre de publications. Les réseaux optiques ne sont pas restés en marge. Son adoption permettra d'améliorer leur survie, leur performance en les rendant plus dynamique avec une gestion plus proactive. En combinant les avantages du monitoring de performance et des techniques d'analyse de données, il serait possible d'apprendre plus des données historiques. Avant de mener toute analyse, il est nécessaire de décrire les différentes méthodes statistiques qui seront utilisées.

CHAPITRE 2

MÉTHODES D'ANALYSE DESCRIPTIVE ET INFÉRENTIELLE DES SÉRIES TEMPORELLES

2.1 Introduction

Une fois les données collectées, il est nécessaire d'avoir un aperçu de ces données tant du point de vue de ses caractéristiques intrinsèques que graphique. Pour ce faire, les méthodes d'analyses exploratoire sont abordées dans ce chapitre. Comme l'a si bien décrit Tukey dans ses mots:

« L'analyse exploratoire des données (*Exploratory data analysis, EDA*) est un travail de détective ; soit un travail de détective numérique ou de comptage ; soit un travail de détective graphique. ... [Il s'agit] d'examiner des données pour voir ce qu'elles semblent dire. Elle se concentre sur des opérations arithmétiques simples et des images. Elle considère toutes les apparences identifiées comme des descriptions partielles, et essaie de regarder en profondeur pour trouver de nouvelles informations. » (Beyer, 1981).

De ce fait, un ensemble de méthodes d'analyse et de graphiques sont nécessaires. L'objectif de ce chapitre est donc de décrire les outils d'analyses qui seront utilisés au travers des chapitres suivants. La première partie de ce chapitre présente les outils d'analyse statistique descriptive, la deuxième partie porte sur les outils d'analyse inférentielle.

2.2 Outils d'analyse statistique de base

2.2.1 Indicateur de tendance centrale

Un indicateur de tendance se défini comme la mesure statistique qui identifie une valeur unique comme représentative d'une distribution entière.

Elle vise à fournir une description précise de l'ensemble des données collectées (Manikandan, 2011b). Les trois indicateurs de tendance centrale des données couramment utilisés sont : la moyenne, la médiane et le mode.

- la moyenne (moyenne arithmétique) représente la somme de toutes les valeurs dans le jeu de données divisée par le nombre total d'observations. Le désavantage principal de la moyenne reste sa sensibilité aux valeurs extrêmes.
- la médiane est la valeur du milieu lorsque les données sont classées en ordre ascendant ou descendant. Elle divise la distribution de fréquence en deux. Cinquante pour cent des observations d'une distribution ont un score égal ou inférieur à la médiane. La médiane est donc le 50ème percentile (Manikandan, 2011a). Cet indicateur n'est pas sensible aux valeurs extrêmes.
- le mode désigne la valeur qui apparait le plus fréquemment dans les données. Certains ensembles de données n'ont pas de mode car chaque valeur ne se produit qu'une seule fois. D'autre part, certains ensembles de données peuvent avoir plus d'un mode. Les données avec un, deux ou trois modes sont respectivement appelées unimodal, bimodal et trimodal. En général, les données avec deux ou plusieurs modes sont multimodal. Le mode est rarement utilisé comme statistique sommaire comme les deux premiers indicateurs, sauf pour décrire une distribution multimodale (Manikandan, 2011a).

2.2.2 Indicateur de dispersion

Plusieurs indicateurs sont utilisés pour mesurer l'étalement de la distribution des données dont la variance, l'écart-type et l'écart interquartile.

La variance est la mesure standard de la dispersion. Ils indiquent le degré de dispersion d'une distribution de données. La variance de N observations x_1, x_2, …, x_n, pour une variable quantitative est (Han, Kamber et Pei, 2012b) :

$$\sigma^2 = \frac{1}{N-1}\sum_{i}^{N}(x_i - \bar{x})^2 \tag{2.1}$$

Où $\bar{x}$ représente la valeur moyenne des observations.

L'écart-type σdes observations est la racine carrée de la variance σ^2. Son avantage par rapport à la variance est qu'il a les mêmes unités que la variable originale, ce qui le rend plus interprétable.

La troisième mesure de la dispersion est l'écart interquartile. Afin de le décrire, il est nécessaire de définir les concepts de quartiles. En effet, les quartiles d'une variable sont les trois valeurs qui divisent les données en quatre parties consécutives. Ainsi, un quart des données (25%) se situe en dessous du premier quartile (Q_1), la moitié des données se situe en dessous du deuxième quartile (Q_2) et les trois quarts (75%) se situent en dessous du troisième quartile (Q_3). Ces trois quartiles sont appelés respectivement 25ème, 50ème et 75ème percentile (Han, Kamber et Pei, 2012b).
La différence entre le troisième et le premier quartile représente l'écart interquartile (*interquartile range, IQR*) :

$$IQR = \hat{Q}_3 - \hat{Q}_1 \tag{2.2}$$

2.2.3 Indicateur d'asymétrie

Les indicateurs de tendance centrale et de dispersion discutés précédemment ne révèlent pas toutes les caractéristiques des données au regard de la distribution de fréquence. En effet, deux distributions peuvent avoir des moyennes et écart-type identiques, et différer dans la forme de la distribution, d'où la nécessité du coefficient d'asymétrie (*skewness*) et du coefficient d'aplatissement (*kurtosis*) (University, 2018-2019).

- Le coefficient d'asymétrie mesure l'absence de symétrie de la distribution et vaut :

$$skewness = \frac{1}{N}\sum_{i}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)^3 \tag{2.3}$$

Il est égal à 0 si la série de données est symétrique, positive si elle est allongée vers la droite (queue de distribution vers la droite), et négative si elle est allongée vers la droite (queue de distribution vers la droite).

- Quant au coefficient d'aplatissement, il est égal à :

$$kurtosis = \frac{1}{N}\sum_{i}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)^4 \tag{2.4}$$

Il est égal à 3 lorsque la série a le même aplatissement (ou la même concentration) que celle de Gauss (distribution « «mésokurtique»), supérieur à 3 si elle est plus concentrée que celle de Gauss (distribution « «leptokurtique»), et inférieur à 3 si elle est plus aplatie que celle de Gauss (distribution « «platikurtique») (Tufféry, 2007).

2.3 Outils d'analyse graphique de la densité

Les méthodes non graphiques évoqués dans le paragraphe précédent, bien qu'elles soient quantitatives et objectives ne donnent pas une image complète des données; par conséquent, des méthodes graphiques, plus qualitatives et impliquant un certain degré d'analyse subjective, sont également nécessaires et complémentaires (Scholar, 2020). En effet, la visualisation ou la présentation de ces données et analyses est un moyen important de transmettre des informations.

Le premier outil le plus utilisé et efficace pour montrer la tendance dans les données est la représentation temporelle des variables. En plus de cette représentation de base s'ajoutent deux autres méthodes particulièrement utiles tel que l'histogramme et la boîte à moustache communément appelée *Boxplot* qui sont utilisées dans cette thèse.

2.3.1 Histogramme

L'histogramme figure parmi les outils d'analyse graphique le plus utilisé pour visualiser la distribution en fréquence des données, ainsi que les indicateurs de tendance centrale, de dispersion, la modalité et les valeurs aberrantes. C'est un diagramme à barre dans lequel chaque barre représente la fréquence (nombre, "*count*") ou la proportion (nombre / nombre total). La plage de données de chaque barre est appelée "*bin*". Le nombre ou la largeur de "*bins*" influence énormément l'aspect final de l'histogramme (Komorowski et al., 2016). Ce qui implique une certaine précaution dans le choix de valeur. La largeur des "*bins*" (*Bin size*) peut être déterminée selon la formule de (Scott, 1979):

$$Bin\ size = \frac{3.5\sigma}{(Nombre\ d'observations\ dans\ la\ variable)^{-1/3}} \tag{2.5}$$

qui vise à faire un compromis entre l'erreur d'échantillonnage et la résolution.

Par ailleurs, afin d'avoir un meilleur aperçu de la forme de la distribution, il est possible de représenter la fonction de densité de probabilité (*Probability density Function, PDF*) superposée à l'histogramme.

2.3.2 Boxplot

L'histogramme tel qu'évoqué plus haut présente la distribution en fréquence des données dans son ensemble. Quant au boxplot, il est un outil graphique qui présente de façon synthétique, les caractéristiques essentielles de la distribution des données. Ces caractéristiques sont :

- la médiane (Q_2) représentée par un trait à l'intérieur de la boîte,
- la dispersion (IQR = $\hat{Q}_3 - \hat{Q}_1$) qui est la hauteur de la boîte,
- les valeurs minimum et maximum sont les deux lignes (appelées moustaches) en dehors de la boîte.

Les données qui se trouvent au -delà des valeurs minimum et maximum sont définies comme des données aberrantes (*outliers*). De façon générale, elles sont situées au-delà de la valeur minimum - *3IQR* et de la valeur maximum - *3IQR.*

Le Boxplot est une excellente technique d'analyse parce qu'elle repose sur deux statistiques robustes tel la médiane et l'écart interquartile par contre elle ne fournit pas d'information sur la modalité de données.

2.4 Outils d'analyse de la périodicité du signal

L'étude de la périodicité contrairement aux deux premiers points concernent uniquement les séries temporelles dans lesquelles le temps est un facteur primordial. En effet il est important de savoir si la série sous-étude présente une dépendance vis – à – vis du temps c'est - à - dire une certaine périodicité.

2.4.1 Fonction d'autocorrélation

L'autocorrélation (ou autocovariance) d'une série fait référence au fait que dans une série temporelle, la mesure d'un phénomène à un instant t peut être corrélée aux mesures précédentes (au temps t − 1, t − 2, t − 3, etc.) ou aux mesures suivantes (à t + 1, t + 2, t + 3, ...). Une série auto corrélée est ainsi corrélée à elle-même, avec un décalage (*lag*) donné (Lise, 2015).

On appelle coefficient d'autocorrélation d'ordre 1, le coefficient de corrélation linéaire simple calculée entre la série chronologique et cette même série décalée d'une période de temps. Les coefficients d'autocorrélation sont calculés pour des ordres allant de 0 à K, K étant le décalage maximum admissible. On choisit en général $\frac{n}{6} \leq K \leq \frac{n}{3}$ mais aussi K égal à 2 ou 3 fois la période de la saisonnalité, $K = \frac{n}{5}$ si n ≥ 150 ou encore comme certains auteurs le suggèrent $K = \sqrt{n}$ (Bourbonnais et Terraza, 2016).

La représentation graphique de la fonction d'autocorrélation (*Autocorrélation Function, ACF*) c'est-à-dire r_k en fonction du lag k est appelée corrélogramme. Le coefficient d'autocorrélation d'ordre k est donné par :

$$r_k = \frac{\sum_{t=k+1}^{n}(x_t - \overline{x_1})\,(x_{t-k} - \overline{x_2})}{\sqrt{\sum_{t=k+1}^{n}(x_t - \overline{x_1})^2\ \sum_{t=k+1}^{n}(x_{t-k} - \overline{x_2})^2}} \tag{2.6}$$

Avec

$$\overline{x_1} = \frac{1}{n-k}\sum_{t=k+1}^{n} x_t \text{ et } \overline{x_2} = \frac{1}{n-k}\sum_{t=k+1}^{n} x_{t-k} \tag{2.7}$$

De façon plus simplifiée :

$$r_k = \frac{R_k}{R_0} \text{ Avec } R_k = \frac{1}{n-k}\sum_{t=k+1}^{n}(x_t - \bar{x})\ (x_{t-k} - \bar{x}) \text{ et } R_0 = Var(x_t) \tag{2.8}$$

La valeur de r_k au lag 0 est toujours égal à 1 et est montré dans le corrélogramme.

2.5 Analyse de corrélation

2.5.1 Test de corrélation

La corrélation est une méthode d'analyse inférentielle permettant d'évaluer une éventuelle association linéaire à double sens entre deux variables continues. La corrélation est mesurée par une statistique appelée coefficient de corrélation *r*, qui représente la force de l'association linéaire supposée entre les variables en question. Il s'agit d'une quantité sans dimension qui prend une valeur comprise entre -1 et +1 (Mukaka, 2012). Le tableau ci-dessous permet d'interpréter la valeur de la corrélation entre les variables (Twaissi, 2008).

Tableau 2.1 Règles générales concernant la définition du coefficient de corrélation

Valeur de la corrélation	**Force de l'association**
0.91 à 1.00 (−0.91 to −1.00)	Corrélation positive (négative) très élevée
0.71 à 0.90 (−0.71 to −0.90)	Corrélation positive (négative) élevée
0.41 à 0.70 (−0.41 to −0.70)	Corrélation positive (négative) modérée
0.21 à 0.40 (−0.21 to −0.40)	Corrélation positive (négative) faible mais bien définie
0.10 à 0.20 (−0.10 a −0.20)	Corrélation faible et plutôt négligeable

2.5.2 Les types de coefficient de corrélation

Deux principaux types de coefficients de corrélation existent: le coefficient de corrélation du moment du produit de Pearson qui est une méthode paramétrique et le coefficient de corrélation du rang de Spearman qui est une méthode non paramétrique. L'utilisation correcte du type de coefficient de corrélation dépend des types de variables étudiées. Pour des variables dont la distribution est normale, le coefficient de Pearson s'applique, et la méthode non paramétrique est suggérée pour les données avec des valeurs extrêmes. La méthode de Pearson est également recommandée pour les mesures prises à partir d'une échelle d'intervalle et pour montrer la relation linéaire entre deux variables (Hauke et Kossowski, 2011). De ce fait, pour une corrélation entre deux variables x et y données, la formule de calcul du coefficient de corrélation de Pearson est :

$$\hat{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^{n}(x_i - \bar{x})^2] \cdot [\sum_{i=1}^{n}(y_i - \bar{y})^2]}} \tag{2.9}$$

Avec :

- n = la taille de l'échantillon,
- x_i, y_i = les différents points de données indexés de i,
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, la moyenne.

Pour valider l'existence d'une relation entre deux variables, une régression linéaire simple, voire une régression non linéaire est appliquée et la qualité de la relation supposée est mesurée par le coefficient de corrélation dit «*de Pearson* » vu plus haut.

Cependant, il existe des situations pour lesquelles une mesure de la corrélation sur les valeurs est inadaptée. Si les variables sont ordinales, discrètes, ou si des données aberrantes risquent de biaiser les résultats, ou enfin qu'elles ne suivent pas une loi normale, la corrélation des rangs est recommandée (Baudot, 2020). Le coefficient de corrélation de Spearman appelé ρ (rhô) ou r_s est calculé comme suit :

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{2.10}$$

Avec d_i étant la différence entre le rang de l'observation i et celui de sa valeur. Cette formule est issue du coefficient de corrélation linéaire de Pearson, mais appliqué aux rangs.

Afin de choisir lequel des deux modèles utiliser, il est recommandé de déterminer la normalité oui ou non de la distribution.

Pour une représentation graphique de de la dépendance entre plusieurs variables en même temps, la matrice de corrélation est utilisée. C'est une table contenant les coefficients de corrélation entre chaque variable et les autres.

2.5.3 Le test statistique de Shapiro – Wilk

Le test de Shapiro – Wilk est un test d'hypothèse de la normalité permettant de savoir si une série de données suit une loi normale. Il est basé sur la statistique W qui est le coefficient de détermination (le carré du coefficient de corrélation) entre la série des quantiles générées à partir de la loi normale et les quantiles empiriques obtenues à partir des données. Plus W est élevé, plus la compatibilité avec la loi normale est crédible (Rakotomalala, 2011).

Interprétation du test

L'hypothèse de validation de la normalité de la distribution se formule comme suite (Berg, 2020; Wikipedia, 2020) :

Hypothèse nulle : les données sont normalement distribuées
Hypothèse alternative : les données ne sont pas normalement distribuées.

Soit p-value la probabilité utilisée pour quantifier la signification statistique du résultat dans le cadre de l'hypothèse nulle :

- si la p-value est inférieure à un niveau α choisi (par exemple 0.05), alors l'hypothèse nulle est rejetée ce qui signifie qu'il est improbable d'obtenir de telles données en supposant qu'elles soient normalement distribuées.
- si la p-value est supérieure au niveau alpha choisi, alors on ne doit pas rejeter l'hypothèse nulle. La valeur de la p-value alors obtenue ne présuppose en rien de la nature de la distribution des données.

2.6 Conclusion

Les méthodes statistiques sont une bonne façon de caractériser les données. Elles seront utilisées dans les chapitres suivants pour caractériser les métriques optiques de performance et dresser un portrait des données. Deux méthodes d'analyse statistiques ont été abordées dans ce chapitre : la méthode descriptive et la méthode inférentielle.

La méthode descriptive consiste à dresser un portrait des données en utilisant les indicateurs de tendance centrale, de dispersion, d'asymétrie mais en utilisant aussi les méthodes graphiques pour un visuel. Ces méthodes statistiques, quoique élémentaires à calculer sont importantes car elles aident à comprendre les données d'une vue d'ensemble et dans les plus petits détails.

Elles aident à déterminer également les techniques inférentielles qui seront utilisées car le choix de ces techniques dépend souvent de l'asymétrie des données.

Les techniques inférentielles peuvent consister en une analyse de la relation entre des variables ou en une régression.

CHAPITRE 3

CARACTÉRISATION DES ROUTES MONITORÉES CÔTÉ ANOMALIES

3.1 Introduction

Dans ce chapitre, une caractérisation des mesures de performance des 155 canaux collectées durant 12 mois et stockées dans la base de données SQL Server 2014, est faite afin d'approfondir la connaissance de l'ensemble de données, de découvrir les structures sous-jacentes et d'extraire les variables importantes (Filliben et Heckert, 2013). Dans cette caractérisation, une analyse des deux paramètres de performance, le rapport signal sur bruit (*signal-to-noise ratio, SNR*) et la puissance optique reçue au récepteur est faite par rapport aux autres métriques monitorées. Une attention particulière est notamment portée sur les "*anomalies*" observées dans les séries temporelles.

Par anomalies, on fait référence aux chutes brusques de SNR et de puissance se soldant par un disfonctionnement ou une perte du signal. L'objectif de cette étude est d'avoir une meilleure compréhension des anomalies et des erreurs observées dans les données de performance des circuits dans un contexte de réseaux longue distance, et d'établir de possible relations entre les paramètres de performance monitorés. L'idée est de repérer dans la mesure du possible les causes des anomalies dans le SNR et la puissance reçue au récepteur et identifier la cause possible. Pour ce faire, cette analyse se découpe en quatre parties :

- dans la première partie du chapitre, une description des routes monitorées est présentée ainsi que la méthode de détection d'anomalie utilisée.
- dans la deuxième partie, une analyse de corrélation est effectuée entre le SNR, la puissance reçue au récepteur et les erreurs. La question qui se pose est : les anomalies détectées dans ces deux métriques génèrent-elles forcément des erreurs? Cette analyse permettra d'établir ou non des liens entre les différentes métriques de performance monitorées.

- dans la troisième partie, une analyse de l'évolution temporelle du gain et de la puissance des amplificateurs optiques est faite avec pour objectif de rechercher de possibles cause des anomalies et des erreurs.

3.2 Description des routes monitorées

Les connexions optiques à l'étude sont déployées sur douze (12) routes du réseau de production d'un grand opérateur de réseau aux États-Unis. Il s'agit d'un réseau de production en ce sens qu'il transporte du trafic réel sur des connexions optiques actives (*lightpaths*) ou circuits actifs entre les sites (nœuds du réseau). Chaque route regroupe plusieurs circuits actifs. Ces routes sont composées d'une variété de types de fibre optique monomode de diverses générations. Le type le plus courant est la fibre optique à dispersion non décalée ITU-T G652 (*Non Dispersion Shifted Fiber, NDSF*). Un autre type de fibre optique est la fibre optique à dispersion décalée G.653 (*Dispersion Shifted Fiber, DSF*). Enfin, différentes fibres optiques G.655 à dispersion décalée non nulle (*Non-Zero Dispersion-Shifted Fiber, NZDSF*) font partie d'une autre famille de fibres optiques et incluent des variantes commerciales comme la *LEAF (Large effective area Fiber*) de Corning, différentes variantes de TrueWave d'OFS et la fibre optique LS (*Long Span*).

Les câbles optiques sont déployés dans divers environnements tels qu'en proximité des autoroutes, des voies ferrées, des passages de métro. Leur installation peut être aérienne, enfouie dans le sol ou dans des conduits. Selon la distance, les liaisons optiques sont métropolitaines (< 300 km), régionales (< 1000 km) ou longue distance (> 1000 km). En fonction de leur localisation, elles traversent diverses zones géographiques et climatiques (désertique, nordique et sismique).

L'intérêt de ce chapitre étant porté sur les anomalies, un ensemble de circuits présentant ainsi des anomalies a été choisi pour l'analyse. Les résultats d'analyse des autres circuits ont été placés en annexe afin de permettre d'avoir une vue d'ensemble des données.

La section suivante présente une description des critères de sélection des circuits et une description des caractéristiques des circuits sélectionnés.

3.3 Définition et détection des anomalies

3.3.1 Définition

La détection d'anomalies est un processus d'exploration de données utilisé pour identifier les types de comportement anormal trouvés dans un ensemble de données et en déterminer les caractéristiques (Provotar, Linder et Veres, 2019). Aussi connue sous le nom de valeur aberrante (*outlier*), une anomalie se définit comme toute déviation du comportement normal d'un point de données pouvant durer un certain temps (Chandola, Banerjee et Kumar, 2009). Le comportement normal dans ce cas-ci est défini comme étant l'intervalle où la majorité des valeurs de la métrique de performance se situe.

3.3.2 Détection d'anomalies

La méthode de détection des anomalies utilisées est celle basée sur la médiane et l'écart interquartile (*Interquartile Range, IQR*) décrits au paragraphe 2.2.2 et présentée à la Figure 3.1. Le choix de cette méthode repose sur le fait que les données de performance sous analyse ne suivent pas la fonction de probabilité d'une gaussienne (Saradjian et Akhoondzadeh, 2011). L'IQR est une méthode de dispersion statistique qui est égale à la différence entre le 25ème et le 75ème centile (respectivement quartiles inférieur et supérieur). Cette méthode établit donc une distribution de base autour de la médiane des données représentant le comportement normal. Le seuil de détection à partir duquel tout point de données est considéré comme une anomalie est fixé à *3IQR* en dessous de la limite inférieure de l'écart interquartile (25%) afin de capturer uniquement les évènements extrêmes (Dancho et Vaughan, 2018; Sseguya, 2020). De ce fait :

$$IQR = \hat{Q}_3 - \hat{Q}_1 \tag{3.1}$$

Avec Q_3 le troisième quartile ou 75% et Q_1 le premier quartile ou 25% et,

$$anomalie \leq Q_1 - 3IQR \tag{3.2}$$

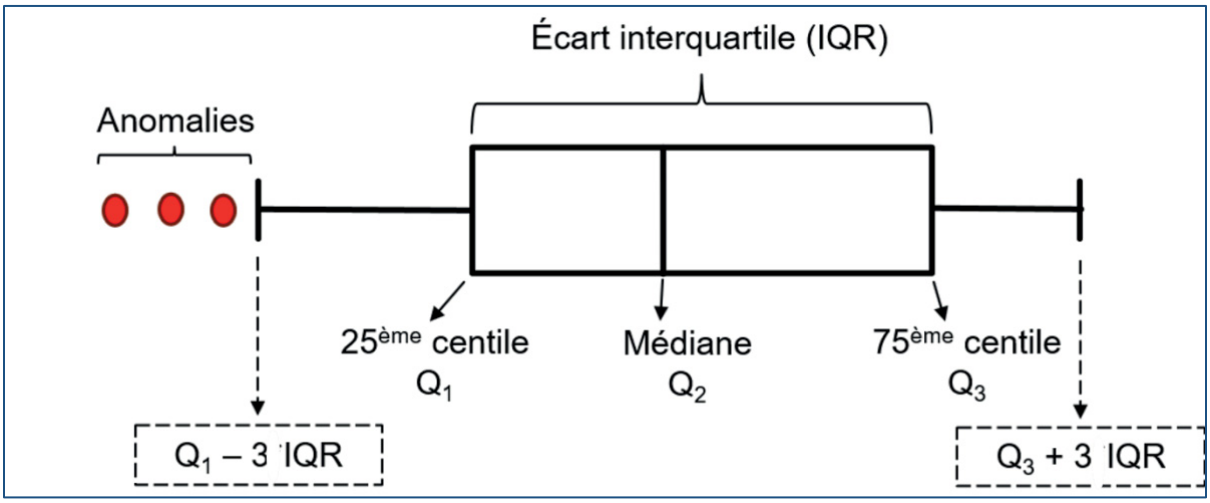

Figure 3.1 Boîte à moustache (*Boxplot*) expliquant la méthode de détection d'anomalies basée sur l'IQR

La Figure 3.2 montre les boîtes à moustache du SNR de trois (3) circuits bidirectionnels (direction A et B, respectivement) ainsi que les points de données (en rouge) considérés comme des anomalies. Ces circuits appartiennent à des routes différentes (ou à la même route) et sont choisis expressément à raison de leur nombre élevé d'anomalies. En effet, la théorie suppose que les valeurs de SNR ne fluctuent pas avec le temps mais restent stable autour d'une valeur moyenne. Les valeurs s'écartant donc de ces moyennes sont considérées comme des anomalies.

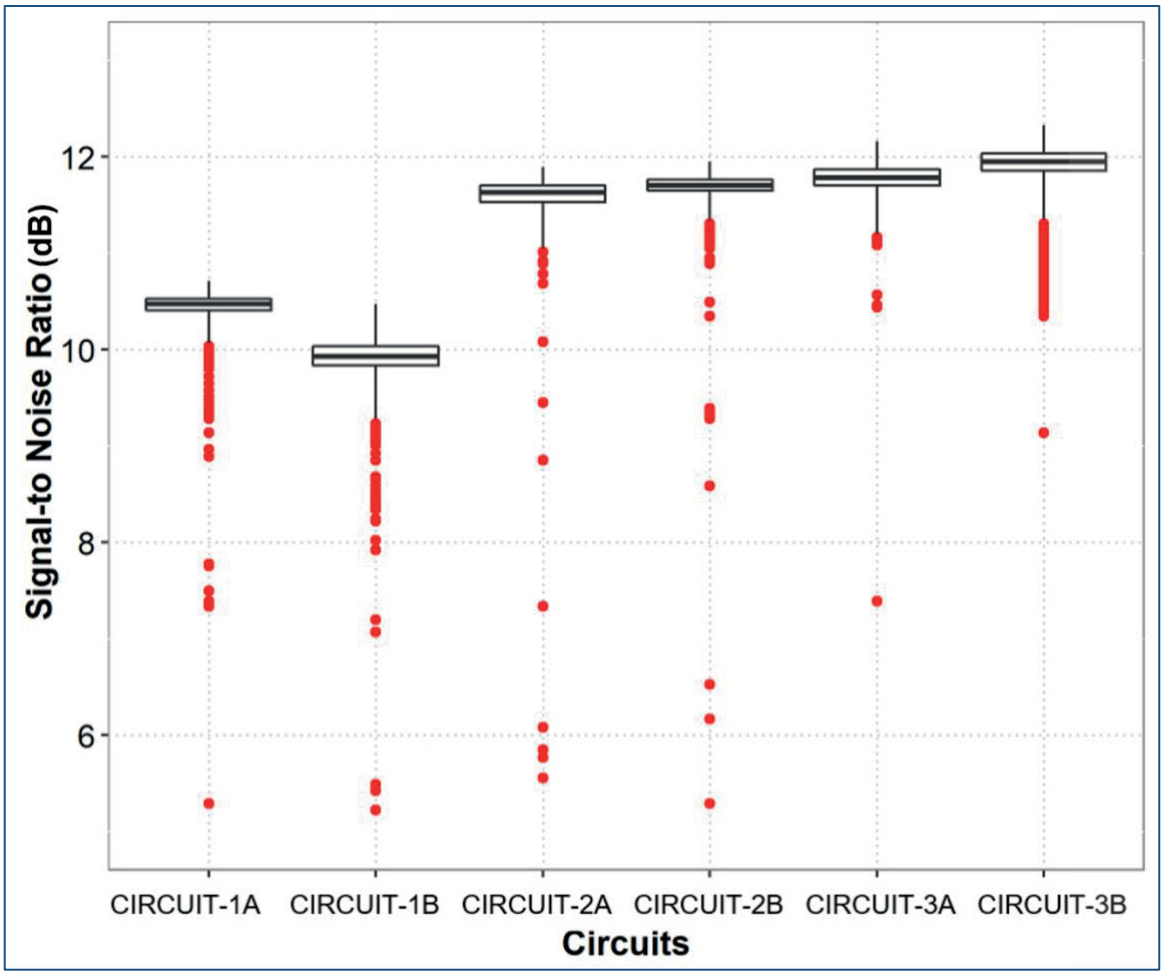

Figure 3.2 Boîtes à moustache du SNR de trois (3) circuits bidirectionnels ainsi que les anomalies détectées durant une période d'observation de 12 mois

La Figure 3.3 présente l'application de la méthode de détection d'anomalies basée sur l'écart interquartile et le seuil de détection d'anomalies représenté par la bande grise autour des points de données pour un circuit donné en considérant la représentation temporelle du rapport signal sur bruit (*signal-to-noise ratio, SNR*) du CIRCUIT-3A.

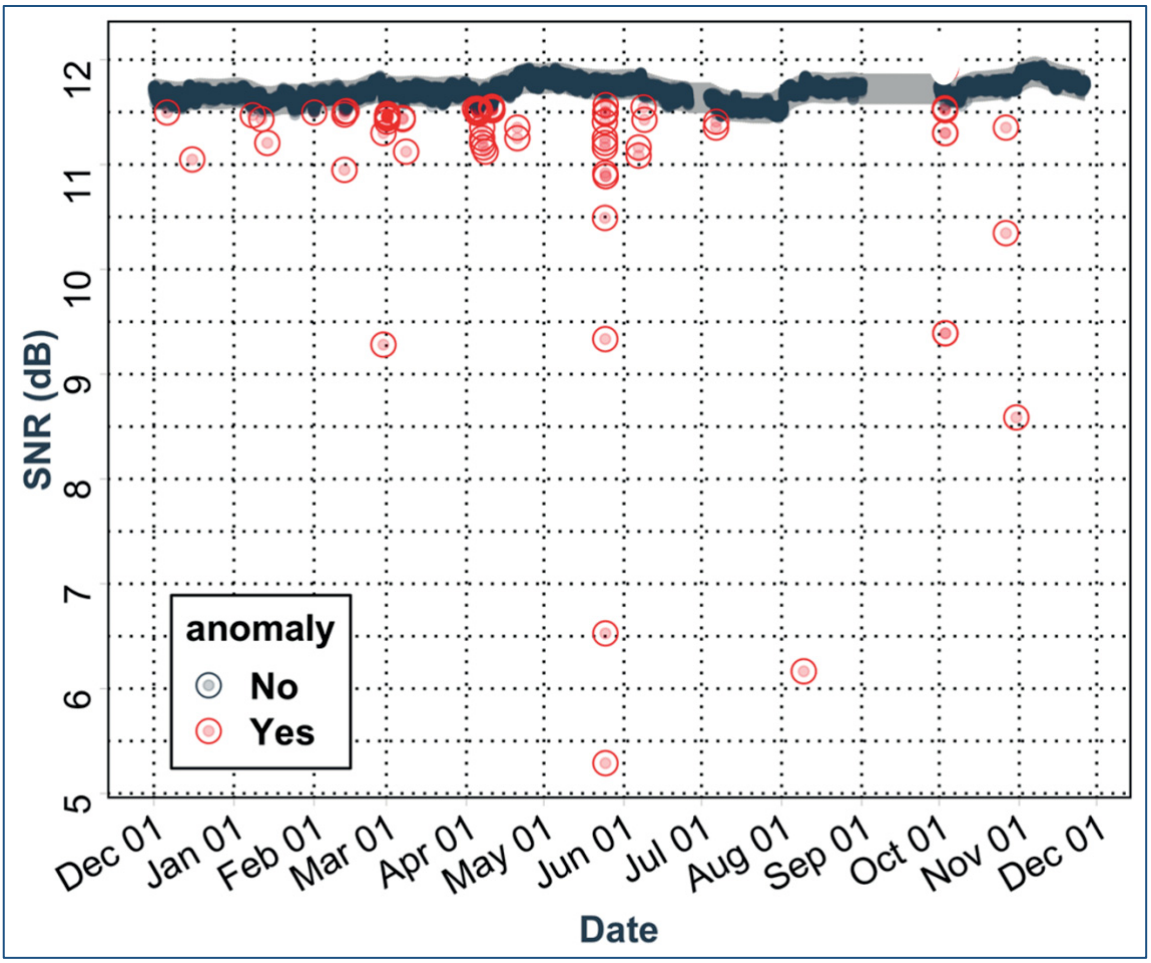

Figure 3.3 Anomalies détectées par la méthode IQR dans la série temporelle de SNR d'une connexion optique

Une fois les anomalies détectées, la prochaine étape consiste à vérifier des liens possibles avec d'autres métriques de performance. Les erreurs définies comme étant le nombre de secondes durant lesquelles une erreur est enregistrée au récepteur, constituent la première métrique d'intérêt. L'objectif de l'analyse est de déterminer s'il y a une relation entre les anomalies détectées dans le SNR et la puissance reçue au récepteur, et les secondes erronées.

Une autre métrique d'intérêt est la relation entre les chutes de puissance au récepteur et les anomalies. Il sera intéressant de vérifier si l'amplitude de la chute de puissance est liée au niveau d'erreurs générées. La troisième métrique à considérer est le gain des amplificateurs déterminés à partir de leurs puissances d'entrée et de sortie, l'objectif étant d'essayer de

localiser la cause possible des anomalies détectées dans le SNR et la puissance reçue au récepteur.

Trois catégories d'anomalies sont considérées dans cette analyse en se basant sur la manière dont les anomalies affectent les circuits :

- **catégorie 1 :** La première catégorie fait référence aux cas où les anomalies affectent tous les canaux de la même route.
- **catégorie 2 :** La seconde catégorie est le cas où un seul canal de la route est affecté.
- **catégorie 3 :** La troisième catégorie concerne les cas où les canaux d'une seule direction sont affectés.

Pour les besoins de l'analyse, différentes routes et circuits seront choisis sur la base de ces trois catégories.

3.4 Choix des circuits par catégorie

Ce sont au total 18 circuits qui ont été retenus pour les trois catégories soit 3 circuits bidirectionnels appartenant à des routes distinctes. Une description physique de ces circuits est menée au travers des 3 paragraphes ci-dessous.

3.4.1 Catégorie 1 – Les anomalies affectent tous les canaux de la route

Deux routes **1** et **2** dont le synoptique est représentée à la Figure 3.4 ont été choisies pour analyser cette catégorie d'anomalies.

La route **1** est d'une longueur de 1400 km et comprend neuf (9) circuits bidirectionnels déployés entre 5 sites intermédiaires. La route **1** fait partie des routes qui enregistrent le plus de circuits ayant un nombre élevé d'occurrence d'anomalies et d'erreurs dans le SNR et dans la puissance reçue au récepteur.

La route **2** est d'une longueur de 600 km et est composée de 14 segments communément appelés *spans* de 47 km environ. Elle compte un ensemble de trois (3) circuits bidirectionnels de bout en bout.

Pour la route **1**, les circuits d'un même tronçon de route exhibent le même comportement. Ainsi, deux circuits bidirectionnels (CIRCUIT-1 et CIRCUIT-2) appartenant à deux tronçons de 1050 km et 700 km, partageant une bonne partie de la route et présentant un grand nombre d'anomalies et d'erreurs sont choisis.

Tous les circuits sur la route **2** présentent le même comportement, de ce fait, un seul circuit bidirectionnel (CIRCUIT-3) est donc retenu pour l'analyse.

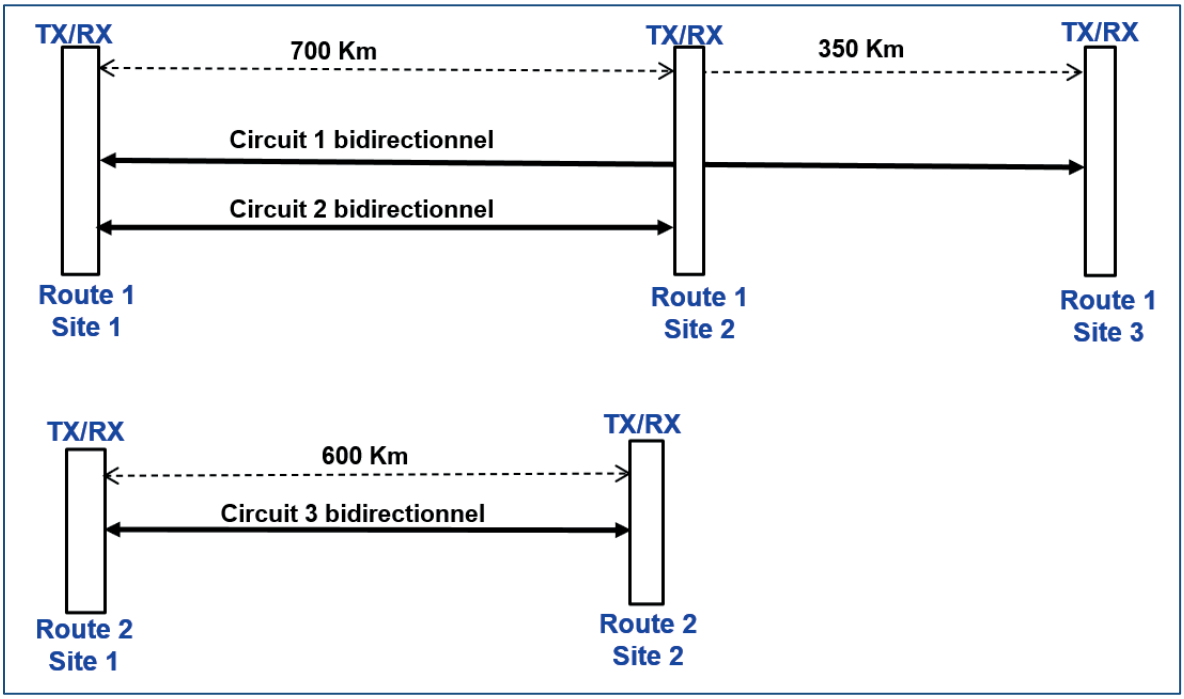

Figure 3.4 Synoptique des circuits de la Catégorie 1 –
Route 1 : CIRCUIT-1 et CIRCUIT-2, Route 2 : CIRCUIT-3
Direction A-B : aller (Site 1 à 2) et direction B-A : retour (Site 2 à 1)

Les Figure 3.5 et Figure 3.6 montrent l'évolution temporelle du SNR, de la puissance reçue au récepteur, des anomalies détectées ainsi que des secondes erronées pour les trois circuits bidirectionnels des route 1 (CIRCUIT-1 et CIRCUIT-2) et route 2 (CIRCUIT-3).

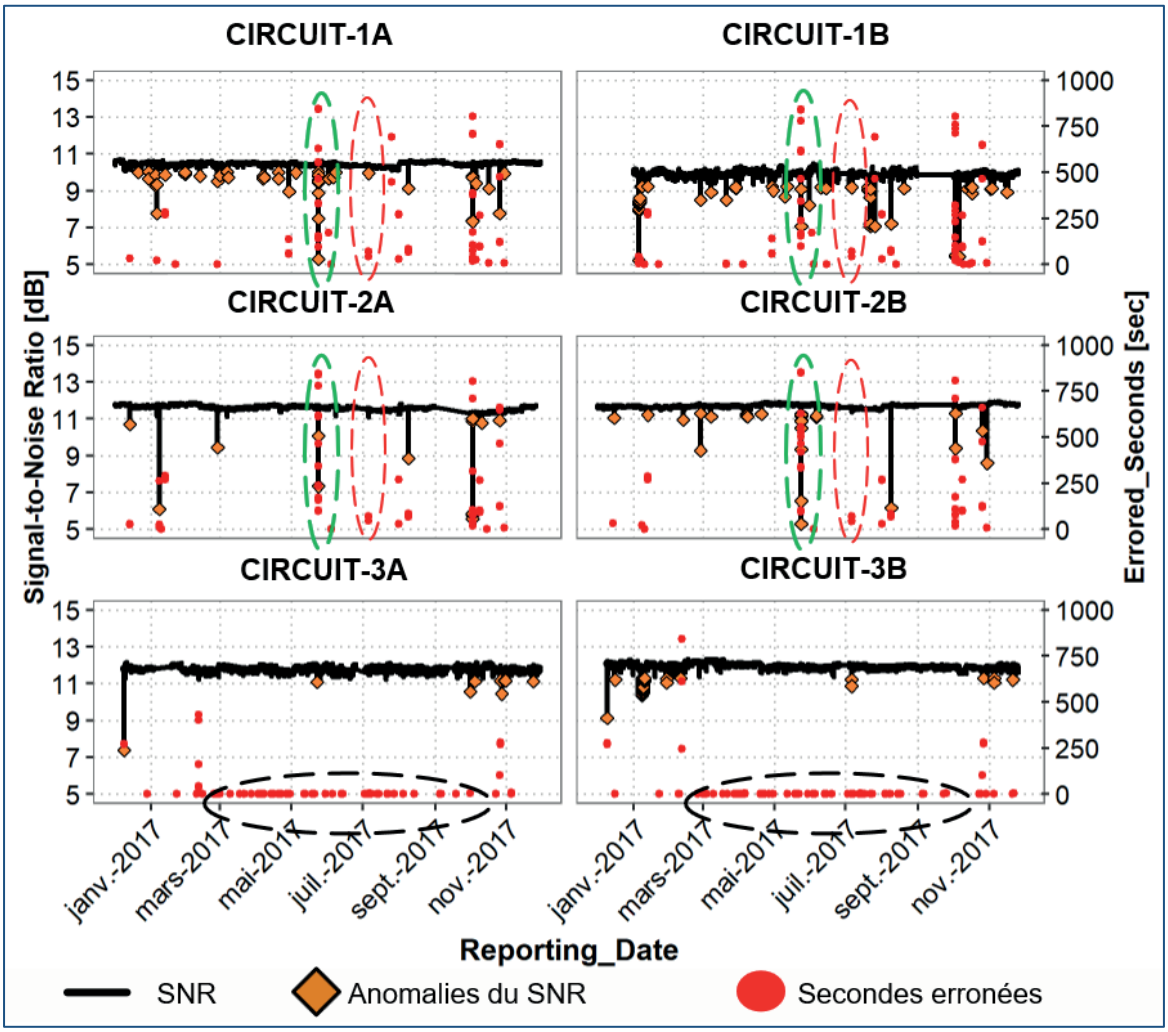

Figure 3.5 Évolution temporelle du SNR et des secondes erronées pour les trois circuits bidirectionnels issus de deux routes présentant des canaux uniformément affectés par les anomalies : Catégorie 1 – Route 1 : CIRCUIT-1 et CIRCUIT-2, Route 2 : CIRCUIT-3 Direction : aller = A-B et retour = B-A

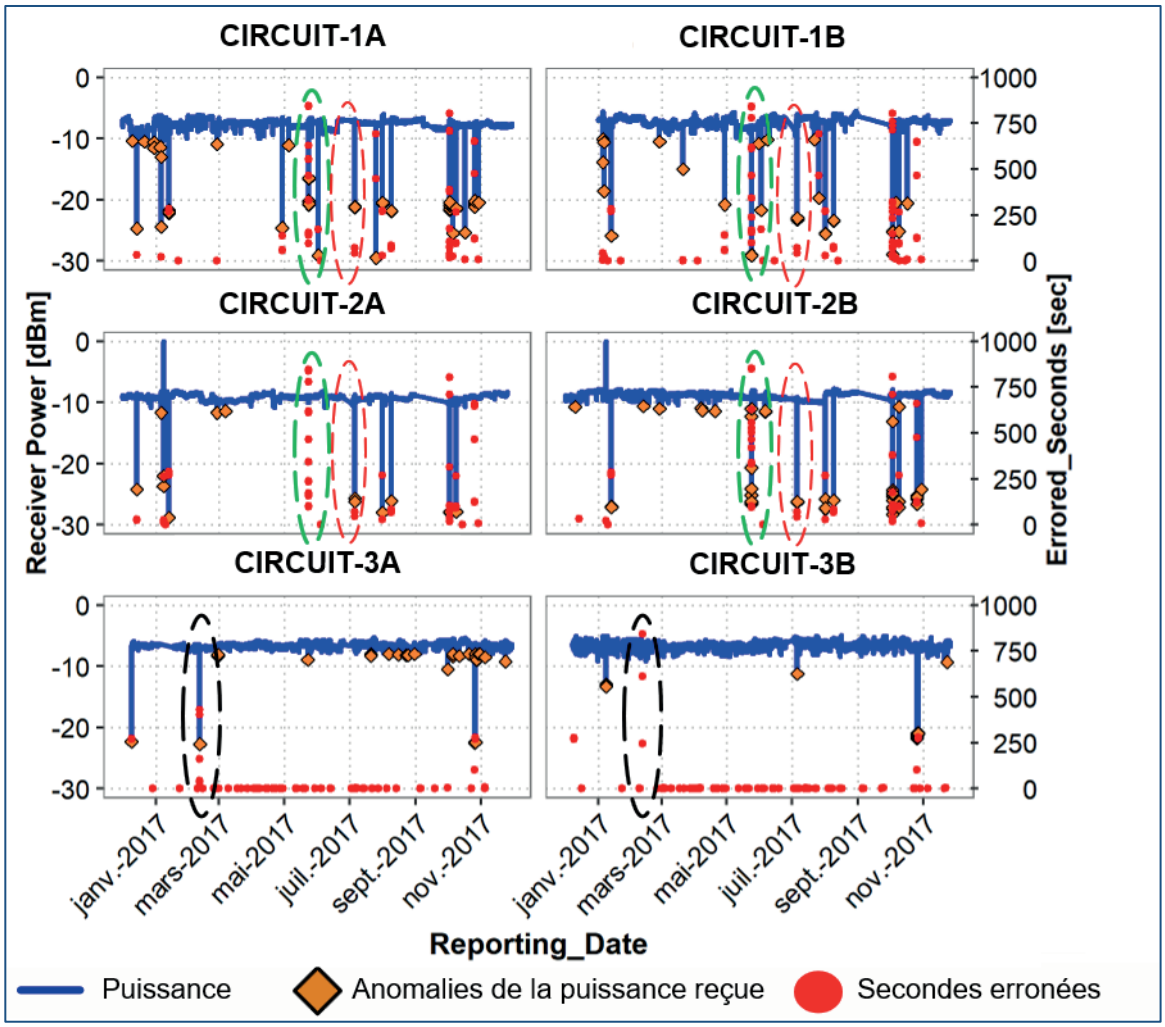

Figure 3.6 Évolution temporelle de la puissance reçue au récepteur et des secondes erronées pour les trois circuits bidirectionnels issus de deux routes présentant des canaux uniformément affectés par les anomalies : Catégorie 1 – Route 1 : CIRCUIT-1 et CIRCUIT-2, Route 2 : CIRCUIT-3 – Direction : aller = A-B et retour = B-A

3.4.2 Catégorie 2 – Un seul canal de la route affecté par les anomalies

La Figure 3.7 représente la seule route monitorée où un seul canal de la route est affecté. Cette route est une liaison déployée de bout en bout de 400 km constituée de 3 circuits bidirectionnels (CIRCUIT-4, CIRCUIT-5 et CIRCUIT-6) et de 3 amplificateurs optiques.

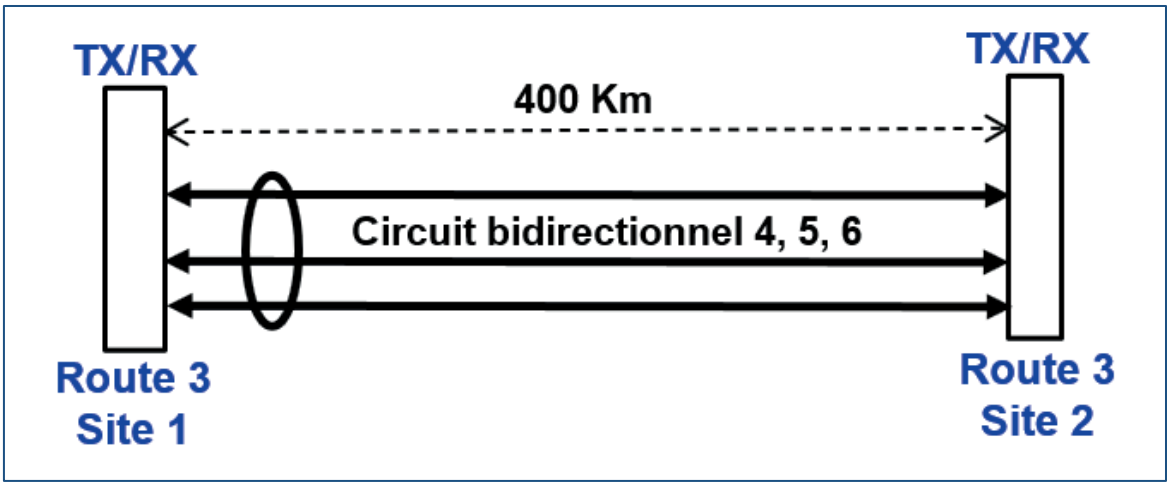

Figure 3.7 Synoptique des circuits de la route 3 : CIRCUIT-4, CIRCUIT-5 et CIRCUIT-6
Direction A-B : aller (Site 1 à 2) et direction B-A : retour (Site 2 à 1)

Les Figure 3.8 et Figure 3.9 montrent l'évolution temporelle du SNR, de la puissance reçue au récepteur, des anomalies détectées ainsi que des secondes erronées pour les trois circuits bidirectionnels de la route 3 (CIRCUIT-4, CIRCUIT-5 et CIRCUIT-6).

Seul le SNR du CIRCUIT-4 est affecté par une anomalie en même temps dans les deux directions A et B. Le rapport signal sur bruit de ce circuit présente une anomalie avant le mois de novembre encadrée en rouge à la Figure 3.8 accompagnées d'une génération de secondes erronées. La puissance reçue au récepteur à la Figure 3.9 ne présente pas d'anomalies pour la même période.

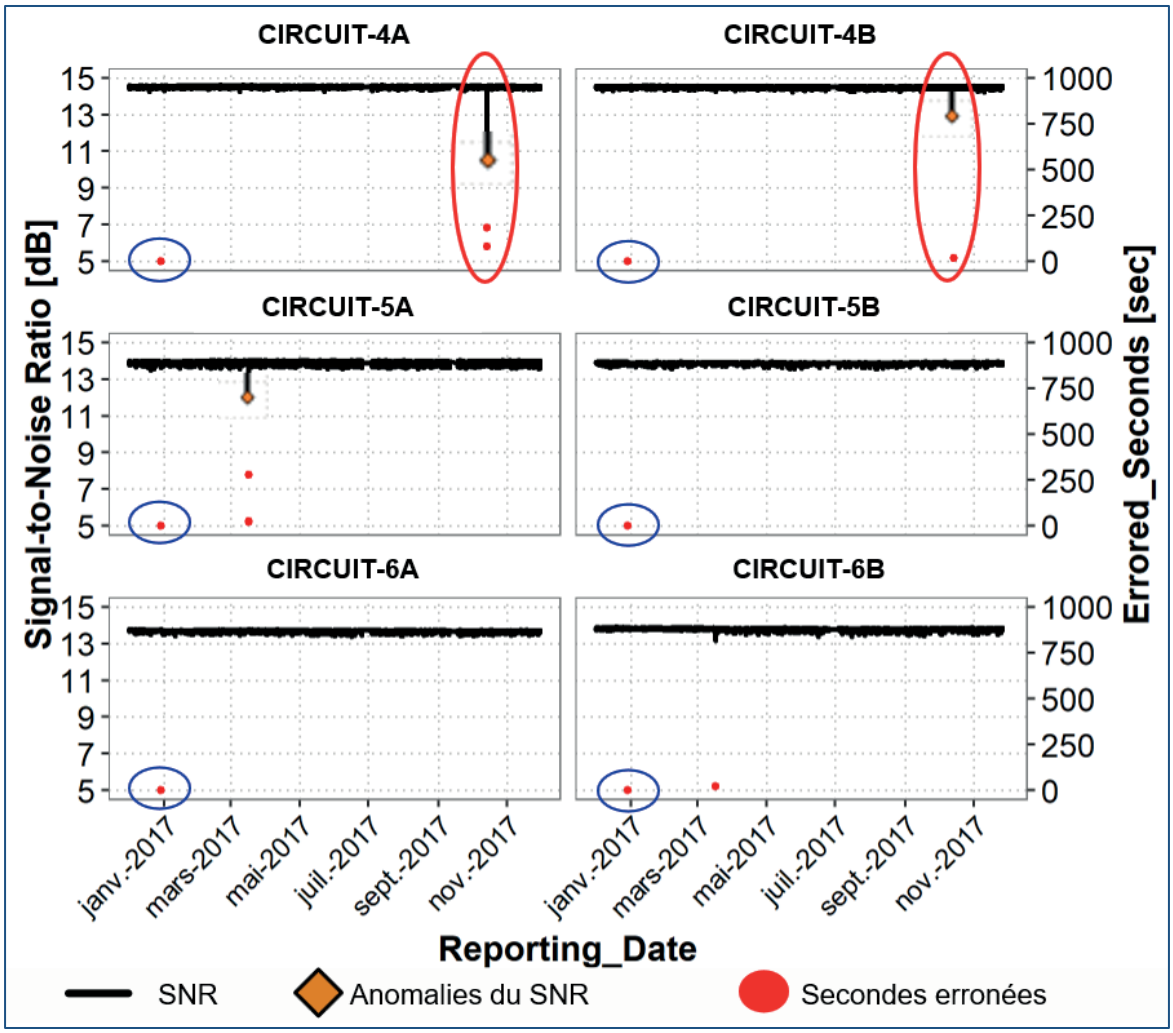

Figure 3.8 Évolution temporelle du SNR et des erreurs secondes pour les trois circuits bidirectionnels de la catégorie 2 – Un seul canal de la route affectée par les anomalies – Route 3 : CIRCUIT-4, CIRCUIT-5 et CIRCUIT-6
Direction : aller = A-B et retour = B-A

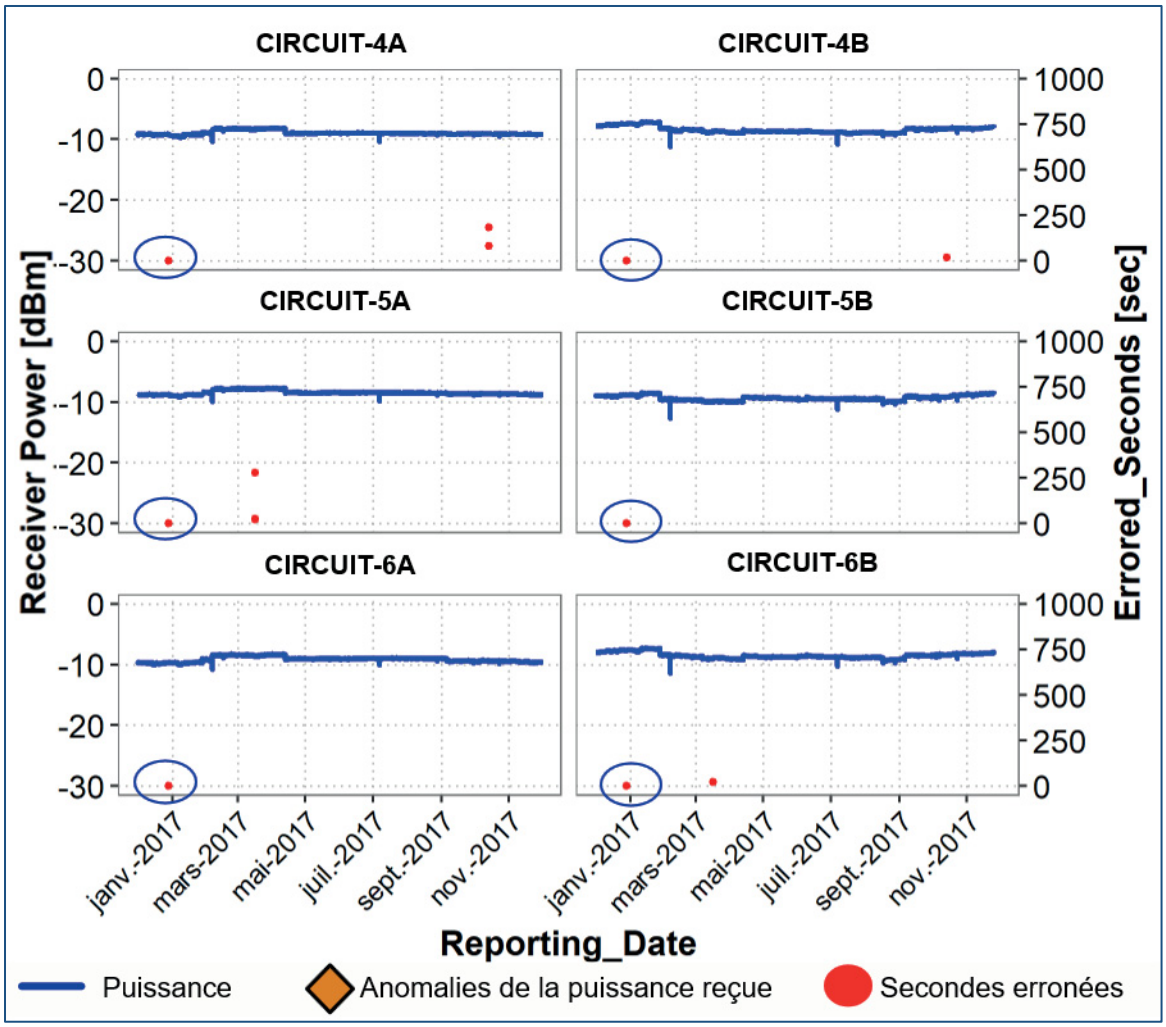

Figure 3.9 Évolution temporelle de la puissance reçue au récepteur et des erreurs secondes pour les trois circuits bidirectionnels de la catégorie 2 – Un seul canal de la route affectée par les anomalies - Route 3 : CIRCUIT-4, CIRCUIT-5 et CIRCUIT-6 Direction : aller = A-B et retour = B-A

3.4.3 Catégorie 3 – Une direction de la route affectée par les anomalies

Les circuits analysés dans cette section font partie d'une route constituée de neuf circuits bidirectionnels monitorés, de longueur égale à 340 km. La Figure 3.10 montre le synoptique de la route avec les circuits choisis représentés. Ces circuits situés entre les sites 2 et 3 de longueur 205 km sont celles qui présentent le cas de figure où une seule direction est affectée par les anomalies et inclue 3 sites d'amplification.

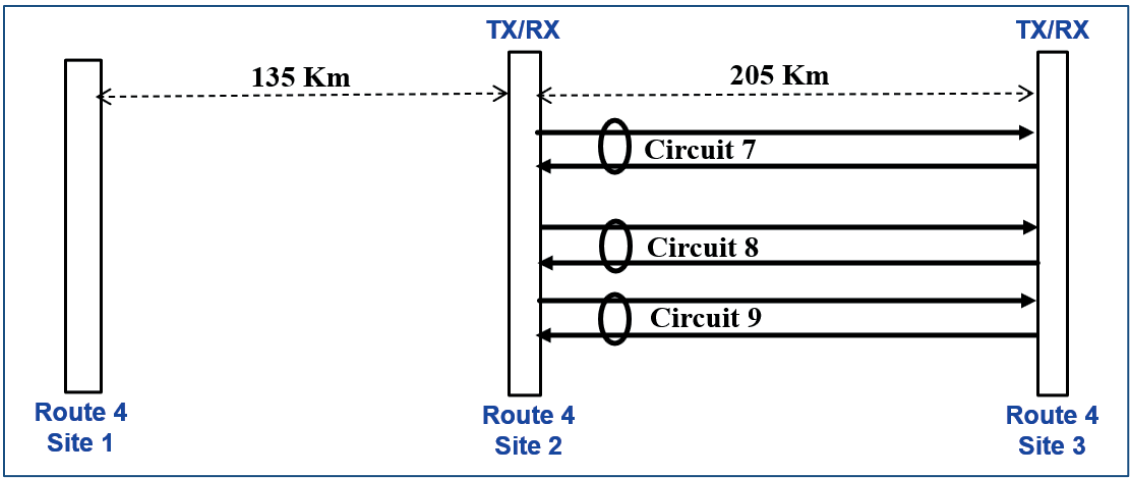

Figure 3.10 Synoptique de la liaison optique de la route analysée dans la catégorie 3 : Un seul canal de la route affectée – Route 4 : CIRCUIT-7, CIRCUIT-8 et CIRCUIT-9 Direction : aller = A-B (site 2-3) et retour = B-A (site 3-2)

L'évolution temporelle du SNR et de la puissance reçue au récepteur est présentée aux Figure 3.11 et Figure 3.12. Les circuits *7A*, *8A* et *9A* sont les circuits qui vont du site 2 au site 3 et les circuits *7B*, *8B* et *9B* les circuits déployés dans la direction site 3 vers site 2 respectivement direction A-B et B-A. Il y a une très forte dynamicité et beaucoup plus d'anomalies dans la direction A-B que dans la direction opposée B-A. En revanche, il y a très peu de secondes erronées enregistrées dans la période d'observation.

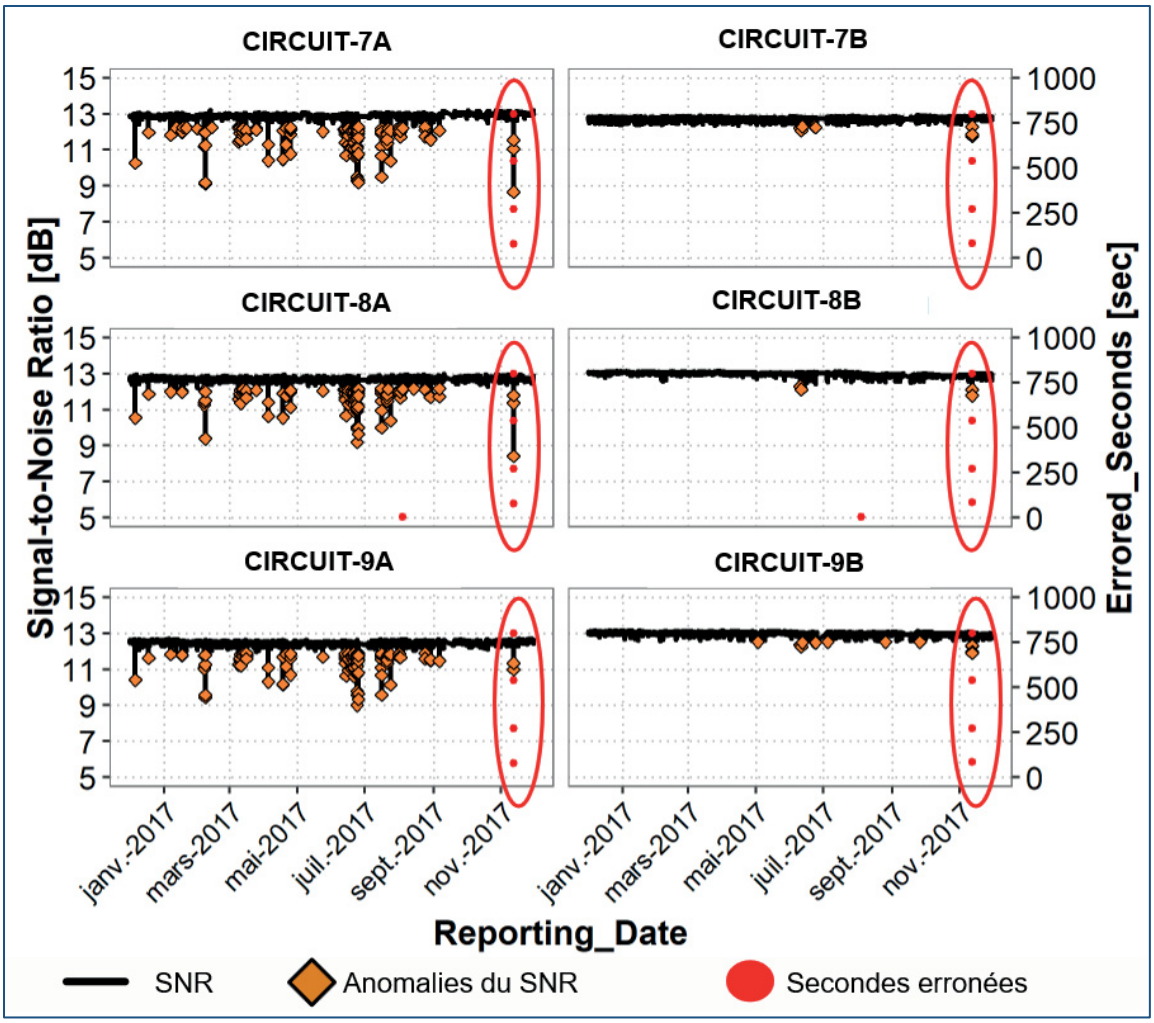

Figure 3.11 Évolution temporelle du SNR et des secondes erronées pour les trois circuits bidirectionnels de la même route présentant une direction affectée par les anomalies
Route 4 : CIRCUIT-7, CIRCUIT-8 et CIRCUIT-9
Direction : aller = A-B et retour = B-A

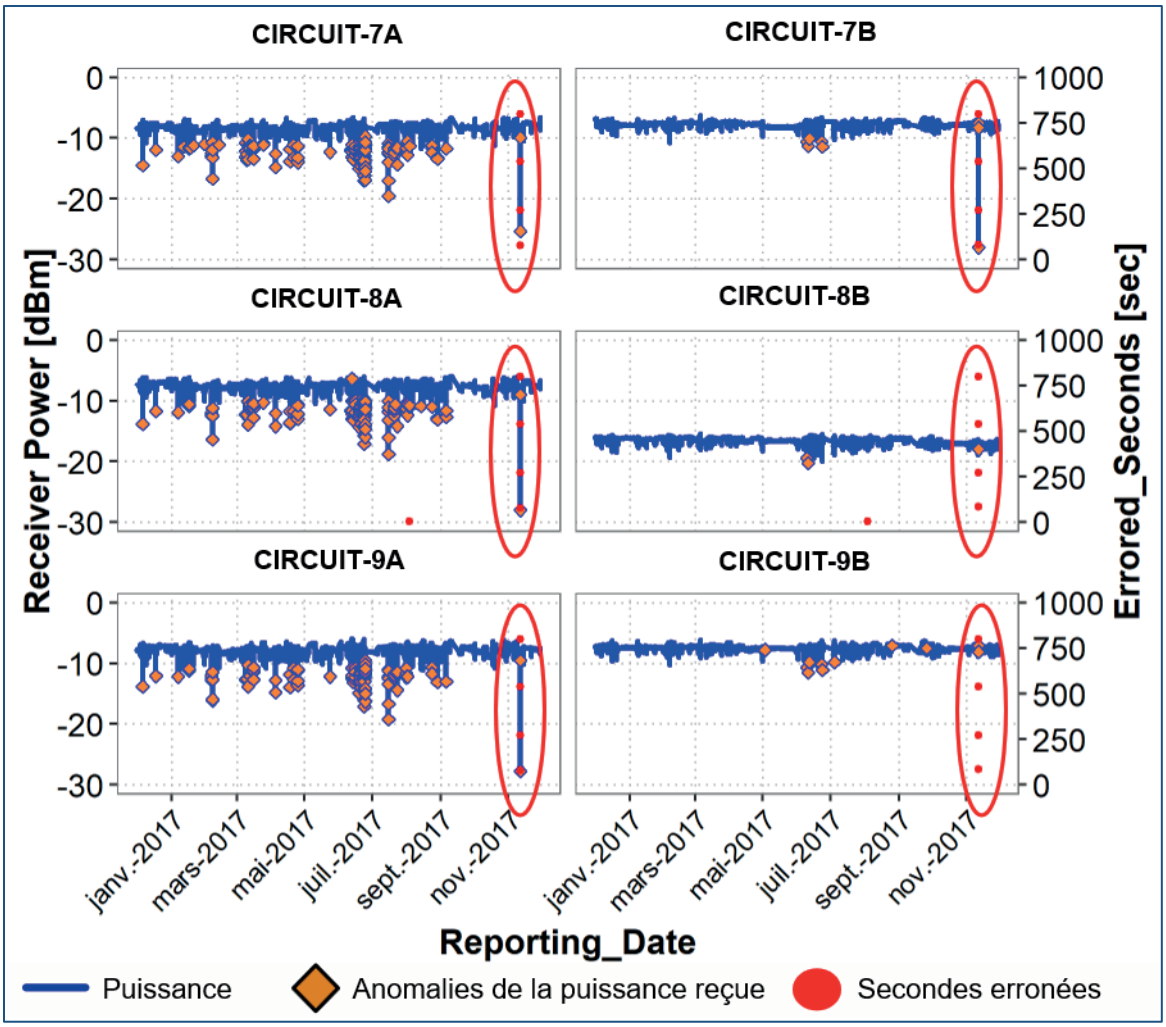

Figure 3.12 Évolution temporelle de la puissance reçue au récepteur et des secondes erronées pour les trois circuits bidirectionnels de la même route présentant une direction affectée par les anomalies – Route 4 : CIRCUIT-7, CIRCUIT-8 et CIRCUIT-9
Direction : aller = A-B et retour = B-A

Afin d'avoir une meilleure compréhension des anomalies et des erreurs observées dans les données de performance, une analyse de corrélation est menée en considérant les circuits précédemment décrits. La question de recherche qui se pose est : les secondes erronées sont-elles causées par les anomalies?

3.5 Analyse de corrélation entre les différents paramètres de performance

Dans cette section, une analyse de corrélation est menée entre les anomalies détectées dans le SNR ou la puissance optique reçue au récepteur et les secondes erronées et ce, pour les différentes catégories. La question est de savoir si ces anomalies sont à l'origine des secondes erronées enregistrées. Le principe de l'analyse de corrélation a été détaillé au paragraphe 2.5 du chapitre précédent. L'occurrence des anomalies et des secondes erronées est très élevée dans chacun des circuits de la catégorie 1 (CIRCUIT-1, CIRCUIT-2 et CIRCUIT-3) sous analyse, et cela se remarque également pour tous les autres circuits qui partagent la même route.

En observant les représentations temporelles des circuits de la catégorie 2 et 3, respectivement – un seul canal de la route affecté et une direction de la route affectée, il y a très peu de secondes erronées enregistrées dans la période d'observation de 12 mois en comparaison de la catégorie 1.

3.5.1 Catégorie 1 – Les anomalies affectent tous les canaux de la route

Les résultats de l'analyse de corrélation sont présentés à la Figure 3.13. Les principales observations tirées de cette analyse sont les suivantes :

Le CIRCUIT-1 montre une forte corrélation positive de coefficient de corrélation 0.65 entre la puissance reçue au récepteur (*output power at receiver, OPRx*) et les secondes erronées (ES) dans les deux directions A-B et B-A. La corrélation entre le SNR et les secondes erronées (ES) est plutôt négligeable de l'ordre de 0.12 dans la direction B-A et bien définie dans la direction opposée A-B avec un coefficient de corrélation de 0.35.

Le CIRCUIT-2 montre également une corrélation positive modérée entre la puissance reçue au récepteur (*OPRx*) et les secondes erronées (ES) dans les deux directions A-B et B-A soient 0.69 et 0.63. Le SNR est corrélé avec les secondes erronées dans les deux directions A-B et B-A de 0.32 et 0.45.

Ces deux circuits sont issus de la route 1 mais de tronçons de route différents. Il partage le même site 2 dans la Figure 3.4. Les autres circuits du même tronçon présentent le même comportement.

Dans le CIRCUIT-3, le coefficient de corrélation de la puissance optique reçue et des secondes erronées est plus élevé dans la direction B-A (0.62) que dans la direction A-B (0.39). Le SNR présente une corrélation négligeable avec les secondes erronées de 0.29 pour la direction A-B et 0.14 pour la direction B-A. La puissance optique reçue au récepteur est corrélée de façon modérée au SNR dans la plupart des cas, pouvant atteindre un coefficient de corrélation de 0.63. Les événements d'erreurs avec un faible nombre de secondes erronées (de l'ordre de 1 à 4 ES) ne sont pas liés aux variations des paramètres de surveillance étudiés (Voir encadré pointillé noir à la Figure 3.5). Ce qui suggère que leur cause n'est pas le signal lui-même.

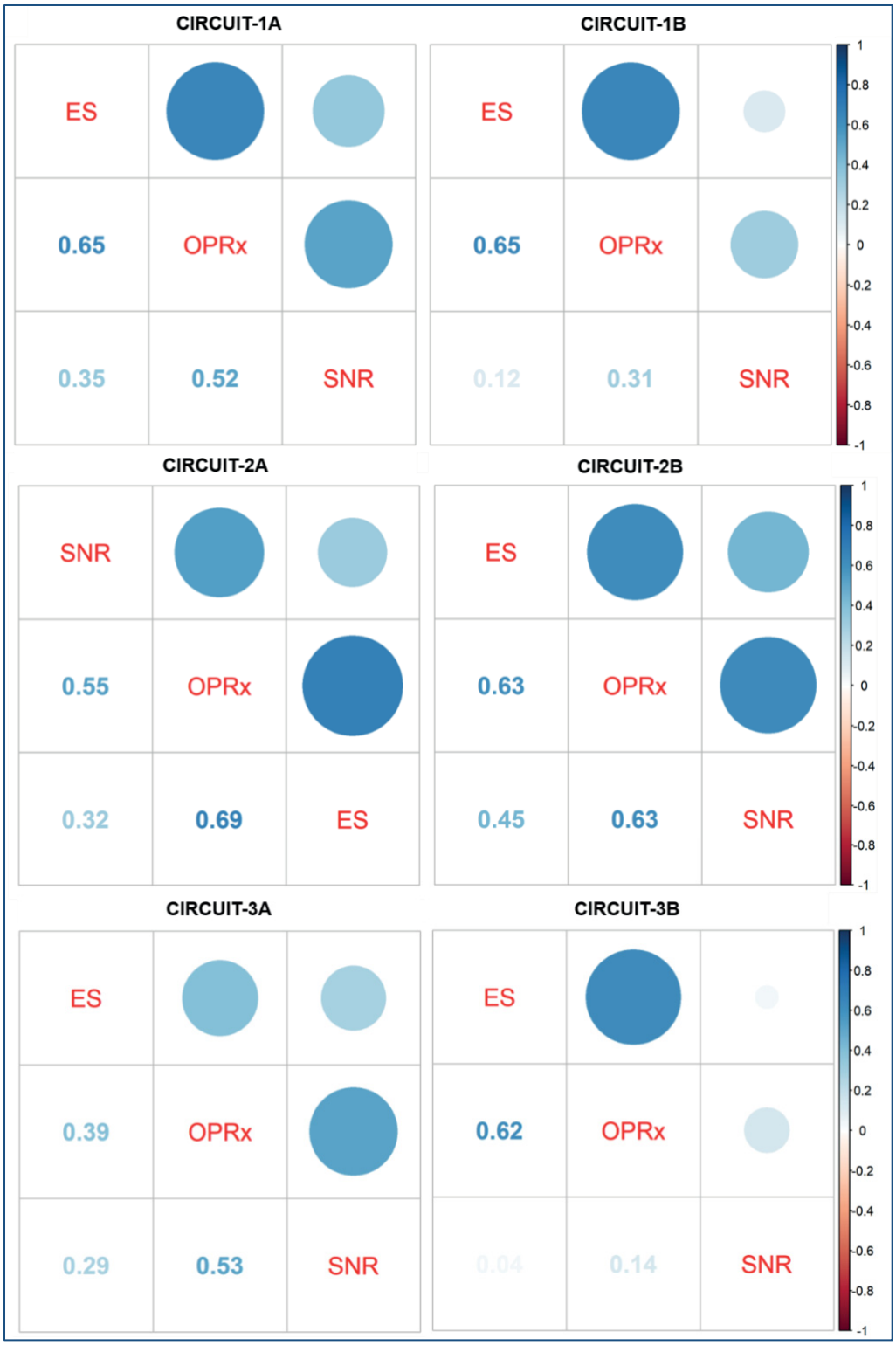

Figure 3.13 Matrice de corrélation du Rapport signal sur bruit (*SNR*), de la puissance reçue au récepteur (*OPRx*) et du nombre de secondes erronées (*ES*)

Contrairement aux circuits de la première catégorie où le nombre d'anomalies et de secondes erronées est en grand nombre, les deux autres catégories présentent quelques évènements qui seront analysés par observation.

3.5.2 Catégorie 2 – Un seul canal de la route affectée par les anomalies

Deux évènements majeurs caractérisent ces 3 circuits bidirectionnels à la Figure 3.8 et Figure 3.9. Chaque pic encadré en rouge dans le SNR est accompagné d'une génération de secondes erronées d'une durée de 4 minutes mais la puissance reçue au récepteur reste stable pendant ce moment. Un autre évènement qui survient le 31 décembre, est marquée par la génération d'une seconde erronée encadré en bleu mais n'est accompagnée d'aucune variation dans le SNR et la puissance reçue au récepteur. Une analyse plus approfondie de ces évènements sera menée dans le paragraphe 3.6.3.

3.5.3 Catégorie 3 – Une direction de la route affectée par les anomalies

La génération des secondes erronées dans un intervalle de 30 mn le 12 novembre dans tous les circuits, survient au moment de la chute du SNR de 4 dB dans les circuits de la direction A-B, et de la puissance reçue de 15 dBm dans la même direction, repérée en rouge dans la Figure 3.11 et Figure 3.12. Dans ce cas-ci, la chute abrute du SNR et de la puissance ont conduit à la génération de secondes erronées. Cet évènement sera analysé plus en détail dans l'analyse des causes probables des anomalies au paragraphe 3.6.4.

3.5.4 Amplitude des chutes de puissance optique et secondes erronées

Une autre question qui pourrait se poser est celle de savoir si l'amplitude des chutes de puissance optique reçue au récepteur est liée aux secondes erronées générées. L'amplitude des anomalies est déterminée en considérant une fenêtre coulissante de 24 heures. L'amplitude de baisse de l'anomalie à l'instant t est égale à la différence entre la moyenne de la fenêtre à l'instant t-1 et le point de données considéré (x_t).

$$Amplitude_{anomalie(x_t)} = Moyenne_{fenêtre(t-1)} - x_t \tag{3.3}$$

Une analyse de corrélation entre l'amplitude de ces anomalies et le nombre de secondes erronées révèle un coefficient de corrélation positif et faible de 0.3 en considérant les 18 circuits sous analyse. Dans certains circuits un coefficient de corrélation positif élevé de 0.71 est atteint.

3.5.5 Récapitulatif

Il ressort de l'analyse de corrélation que les secondes erronées sont plus corrélées avec les anomalies dans la puissance reçue au récepteur qu'avec le SNR, tel qu'on peut voir sur les Figure 3.5, Figure 3.6 et Figure 3.13. Plusieurs circuits appartenant à d'autres routes présentent le même comportement (voir **Error! Reference source not found.** à **Error! Reference source not found.** pour les détails).

Au regard des valeurs de coefficients de corrélation calculés, certains évènements de secondes erronées enregistrés semblent liés à certaines anomalies. Cependant cela n'est pas toujours vérifié comme en témoigne le CIRCUIT-3. Le pic dans la puissance reçue du CIRCUIT-3 encadré en pointillé noir à la Figure 3.6 est corrélé à un certain nombre de secondes erronées de valeur cumulative de 1705 ES soit 29 minutes dans la direction A-B. En rappel, les secondes erronées représentent le nombre de secondes en erreur.

Une observation étrange qui a pu se faire est la présence de secondes erronées de 3647 ES environ le double, soit 1 heure dans la direction opposée B-A du CIRCUIT-3 alors que l'anomalie se trouve dans une seule direction à savoir la direction A-B.

Le même constat se fait dans la puissance reçue au récepteur du CIRCUIT-2. Les erreurs sont corrélées avec un pic qui survient entre le mois de mai et de juin (encadré en pointillé vert) dans la direction B-A et sont aussi enregistrées dans la direction A-B même en l'absence d'anomalies dans la puissance dans cette direction.

3.6 Cause probable des anomalies détectées

3.6.1 Amplificateurs optiques

Les systèmes de transmission optique longue distance sont caractérisés par la présence d'amplificateurs optiques déployés afin de compenser l'atténuation dans les fibres optiques et les pertes d'insertion des composants déployés dans les liaisons et dans chacun des sites. Les amplificateurs les plus couramment utilisés dans la transmission longue distance sont les amplificateurs à fibres dopées à l'erbium (*Erbium Doped Fiber Amplifier, EDFA*). À noter que des amplificateurs optiques de type Raman peuvent également être déployés dans les liaisons optiques longue distance, en complément des EDFA. Les liaisons à l'étude ici ne contiennent pas d'amplificateurs Raman.
L'EDFA est un composant essentiel qui permet d'amplifier simultanément la puissance de tous les canaux WDM dans la bande C, peu importe le type de fibre optique, le débit et format de modulation des canaux.

La Figure 3.14 présente la représentation simplifiée (i.e. faisant abstraction des autres composants) de la topologie physique de deux circuits bidirectionnels à longue portée simple avec des amplificateurs à fibres optiques montés en cascade et déployés à tous les 50-100 km de façon typique. Les amplificateurs optiques sont généralement classés selon leurs fonctions.

- le premier type d'amplificateur (Figure 3.14, site terminal 1) est un amplificateur de puissance (*booster)* ou post-amplificateur et qui est utilisé pour compenser les pertes d'insertion des composants optiques (par exemple les multiplexeurs, démultiplexeurs, etc.).
- le second type d'amplificateur optique, utilisé devant un récepteur pour améliorer sa sensibilité, est appelé préamplificateur.
- le troisième type d'amplificateur optique est l'amplificateur de ligne (*optical line amplifier, OLA*). Ils sont montés en cascade le long de la liaison pour compenser les pertes dans chaque segment (*span*) de la liaison.

Le gain des amplificateurs optiques est fourni par des pompes optiques opérant aux longueurs d'onde de 1490 et/ou 980 nm.

De manière générale, des amplificateurs de type préamplificateur/post-amplificateur sont déployés et incluent des atténuateurs optiques variables (*VOA*) afin d'ajuster la puissance de sortie à l'entrée des préamplificateurs. La liaison optique amplifiée est gérée en mode « transparent » afin de maintenir les niveaux de puissance par canal injectés dans chaque segment constant, avec pour résultat que le gain de chaque amplificateur placé à la fin d'un segment est réglé pour être égal à la perte optique du segment précédent (Chan, 2010). La longueur optimale des segments est de l'ordre de 70-90 km.

Il est important de noter que, sur le terrain, les segments sont de longueur très variable selon la topologie physique et les contraintes géographiques. De plus, les types d'équipements varient d'un circuit à l'autre selon les époques de déploiement.

Deux métriques sont monitorées au niveau des amplificateurs dans le cadre de ce projet : la puissance d'entrée (*optical input power, OPIN*) et la puissance de sortie (*optical output power, OPOUT*). Le calcul du gain d'un amplificateur donné est obtenu en faisant la différence entre la puissance de sortie et la puissance d'entrée mesurée au niveau de cet amplificateur.

$$Gain\ (dB) = OPOUT(dBm) - OPIN(dBm) \tag{3.4}$$

3.6.2 Catégorie 1 – Les anomalies affectent tous les canaux de la route

La Figure 3.14 présente la représentation simplifiée de la topologie physique des deux circuits bidirectionnels de la route 1. Le CIRCUIT-1 va du site 1 au site 3, et le CIRCUIT-2 du site 1 au site 2.

Afin d'observer l'impact du comportement des amplificateurs durant les anomalies, une analyse de l'évolution du gain est faite par rapport à la puissance optique au récepteur et au SNR.

Dans cette catégorie de circuits, c'est-à-dire les circuits où les anomalies affectent tous les canaux de la route, il est question de valider si oui ou non ces anomalies qui y surviennent sont dues aux amplificateurs ou du moins si des anomalies de performance sont observées dans le comportement (gain, puissance) des amplificateurs au moment des anomalies. La question qui se pose donc est de savoir si les anomalies observées dans le SNR ou dans la puissance reçue au récepteur peuvent être associées à des événements au niveau des amplificateurs optiques.

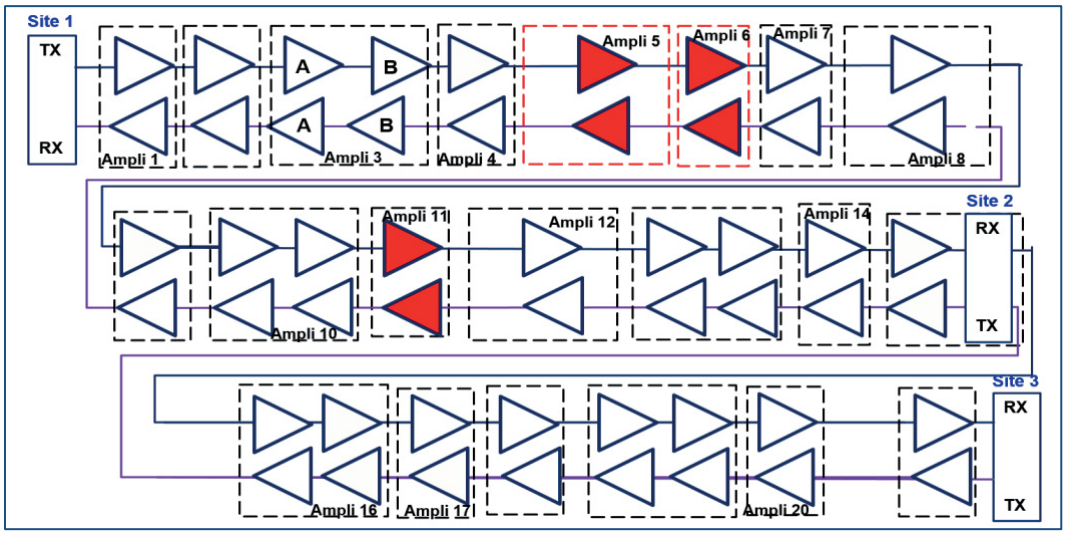

Figure 3.14 Représentation simplifiée de la liaison optique de la catégorie 1
Route 1 : CIRCUIT-1 (1050 km) : Site 1 à 3 et CIRCUIT-2 (700 km): Site 1 à 2

La Figure 3.15 présente les évolutions temporelles du gain des amplificateurs pour chacun des circuits considérés. Les courbes semblent aussi caractérisées par une grande variabilité comme dans le SNR et la puissance reçue.

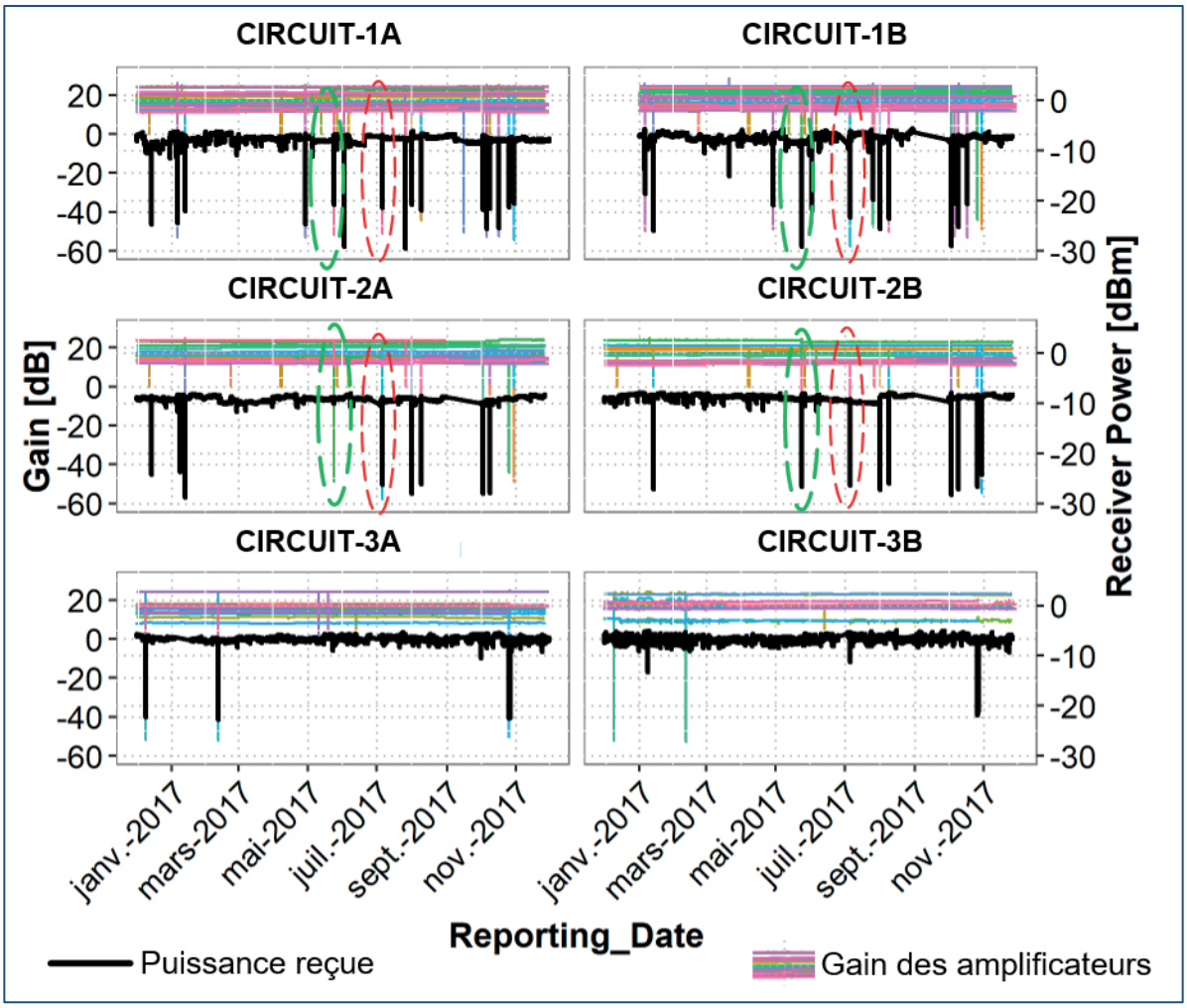

Figure 3.15 Évolution temporelle du gain des amplificateurs optiques et de la puissance reçue par circuit
Catégorie 1 – Route 1 : CIRCUIT-1 et CIRCUIT-2, Route 2 : CIRCUIT-3
Direction : aller = A-B et retour = B-A

Les séries temporelles de puissance reçue au récepteur et du gain des amplificateurs dans le même circuit laisse voir que la majorité des anomalies dans la puissance survient au même moment que ceux dans le gain des amplificateurs optiques.

Cependant deux autres cas de figures se présentent également : le cas où les anomalies sont présentes dans le gain et pas dans la puissance reçue et le cas où les anomalies sont présentes dans la puissance reçue et pas dans le gain des différents amplificateurs.

Ces trois cas sont représentatifs de ce qui se passe en réalité dans l'ensemble des circuits monitorés sous analyse. Autrement dit, certaines anomalies sont associées à des anomalies dans les amplificateurs optiques et d'autres non. Pour mieux illustrer cela, une analyse ciblée sur certaines périodes du signal temporel est menée en considérant chacun des circuits à la Figure 3.15.

Les circuits bidirectionnels 1 et 2 qui partagent une bonne partie de la route (environ 700 km) montrent de nombreuses similitudes dans les anomalies. Un ensemble d'anomalies communes à tous les circuits de ces deux sous-routes survient dans la période du 24 mai dans la puissance reçue au récepteur (entouré en vert dans la Figure 3.15). À la même date, le gain de certains amplificateurs et le SNR présentent également une chute dans la valeur et une grande occurrence de secondes erronées (7 occurrences de valeurs comprises entre 1 ES à 843 ES soit une durée approximative de 51 mn). Une analyse ciblée de la période montre que 3 amplificateurs sont à l'origine de ces anomalies.

La Figure 3.16 présente l'évolution temporelle du SNR, de la puissance reçue au récepteur et des gains des amplificateurs qui montre un changement d'état pendant la période cible.

La Figure 3.17 montre le zoom de l'évolution temporelle du gain de ces amplificateurs. Le gain du premier amplificateur en faute *ampli 5* chute de 14.9 dB à -50 dB entre 15h00 et 18h15 et le gain reste plafonné pendant 2h30 à -51.6 dB. Pendant ce temps, le gain de l'*ampli 6* chute de 23.8 dB à 9.9 dB. Le gain de l'*ampli 7* augmente de 11.4 dB à 13.2 dB soit de 1.8 dB. Le gain de l'*ampli 8* augmente de 19.5 dB à 20 dB. Au cours de la même période de temps, le gain de l'*ampli 11* chute de 18.5 dB à 15.1 dB soit 3.4 dB et le gain de l'*ampli 12* monte de 11.4 dB à 13.1 dB soit 1.7 dB. Le premier amplificateur en faute dans la chaine est l'*ampli 5* qui présente une puissance d'entrée stable de 1.2 dBm.
Par la suite, l'*ampli 6* et l'*ampli 11* sont également touchés (Voir amplificateurs en rouge dans la Figure 3.14). La puissance reçue au récepteur et le SNR montre une perte de signal pendant cette période.

Par la suite, le gain de l'*ampli 5* revient à 14.9 dB pendant 1h de temps pour chuter à nouveau à -51.6 dB, remonter à 1.2 dB pendant 1h15. Il chute à -51.6 dB pendant 45 mn avant de repartir à 13.3 dB.

Tous les autres amplificateurs de la chaine non mentionnés présentent une valeur stable durant la période sélectionnée.

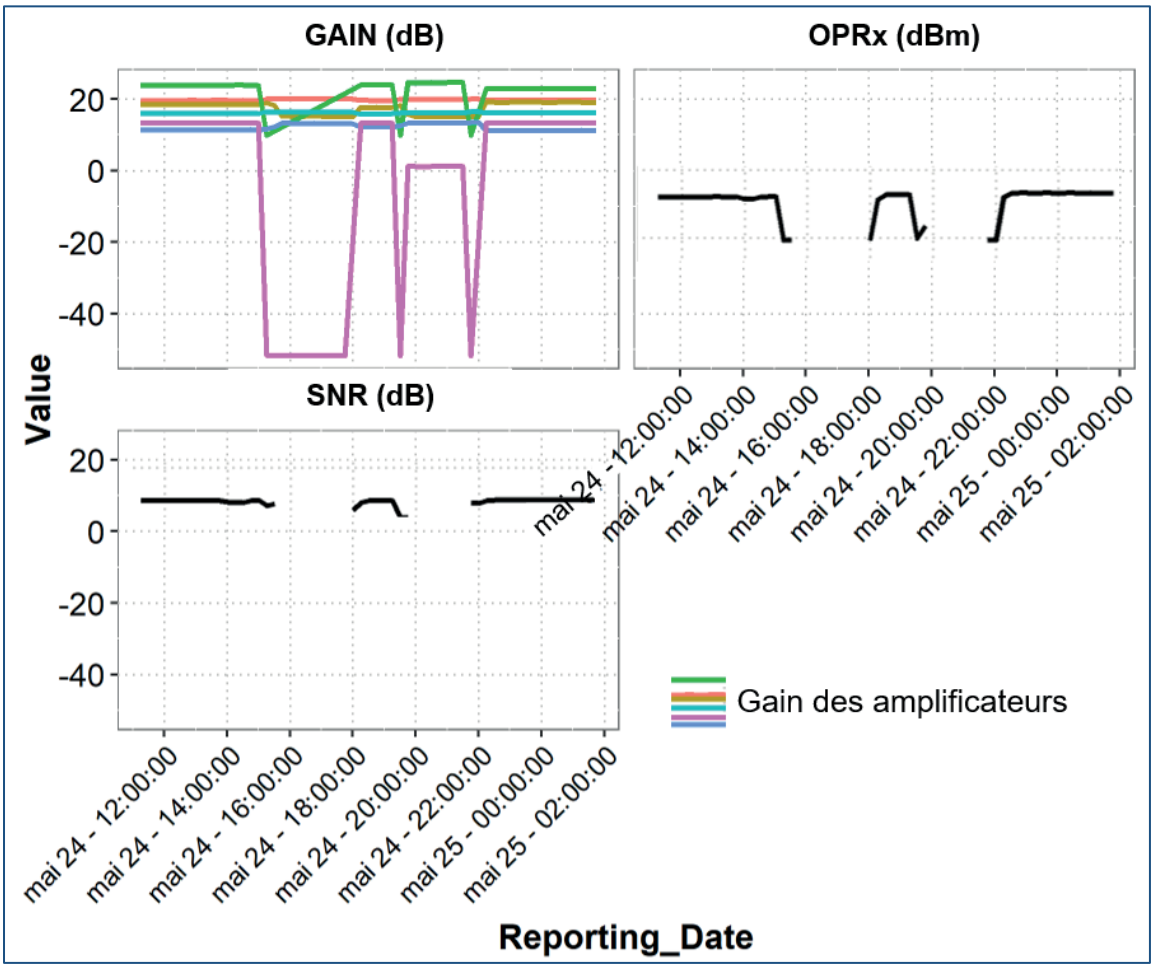

Figure 3.16 Évolution temporelle du SNR, de la puissance reçue au récepteur (*OPRx*) et du gain des amplificateurs variant dans la période du 24 au 25 mai (CIRCUIT-1)

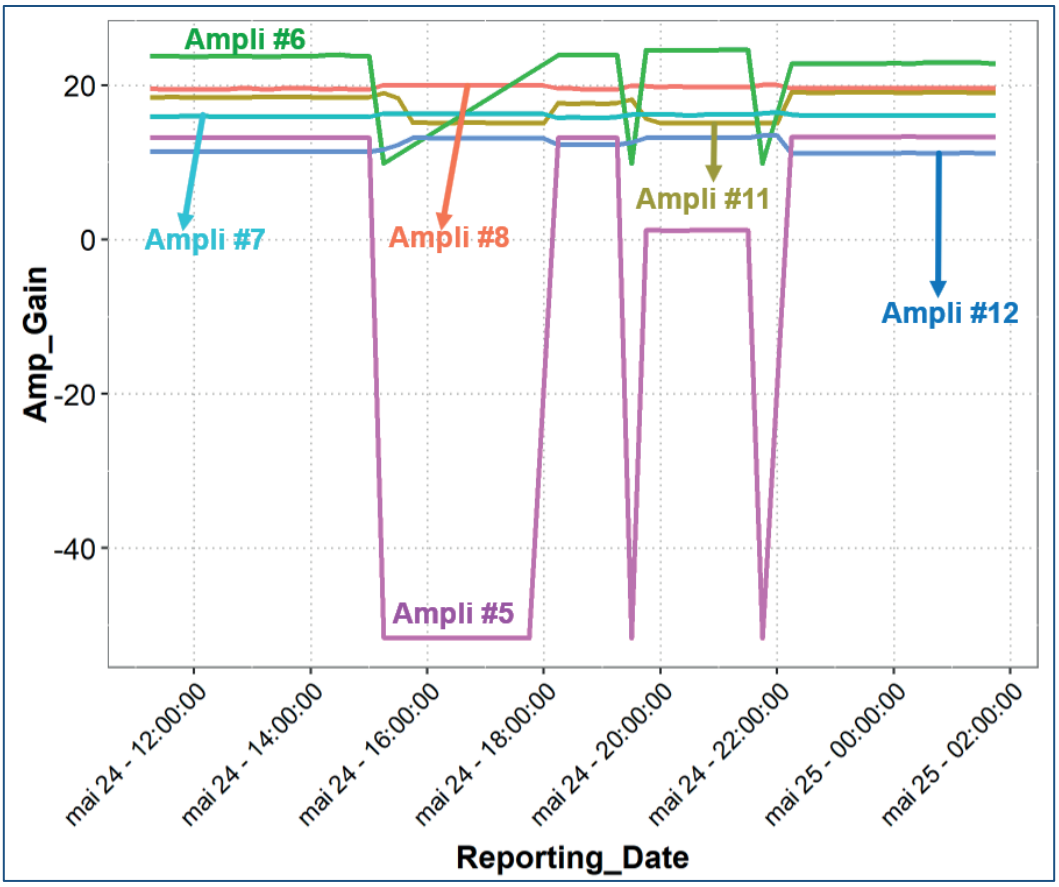

Figure 3.17 Évolution temporelle du gain des amplificateurs variant dans la période du 24 au 25 mai (CIRCUIT-1)

Une autre anomalie commune à tous les circuits de ces deux sous-routes survient dans la période du 6 juillet dans la puissance reçue au récepteur (repérée en rouge dans la Figure 3.15). En regardant de plus près, un amplificateur situé au site d'amplification *ampli 11* de la Figure 3.14 semble en être la cause.

À ce propos, un zoom sur la période du 6 juillet montre une valeur stable du rapport signal sur bruit et de la puissance d'entrée de l'ampli 11 (OPIN) autour de 0 dBm. La puissance de sortie (OPOUT) chute considérablement à -50 dBm à deux reprises dans un intervalle de 2 heures. Le même phénomène s'observe dans le gain, et la puissance reçue au récepteur chute à -20 dBm. Dans ce cas précis, la panne de l'amplificateur a un impact sur la puissance reçue et non sur le SNR.

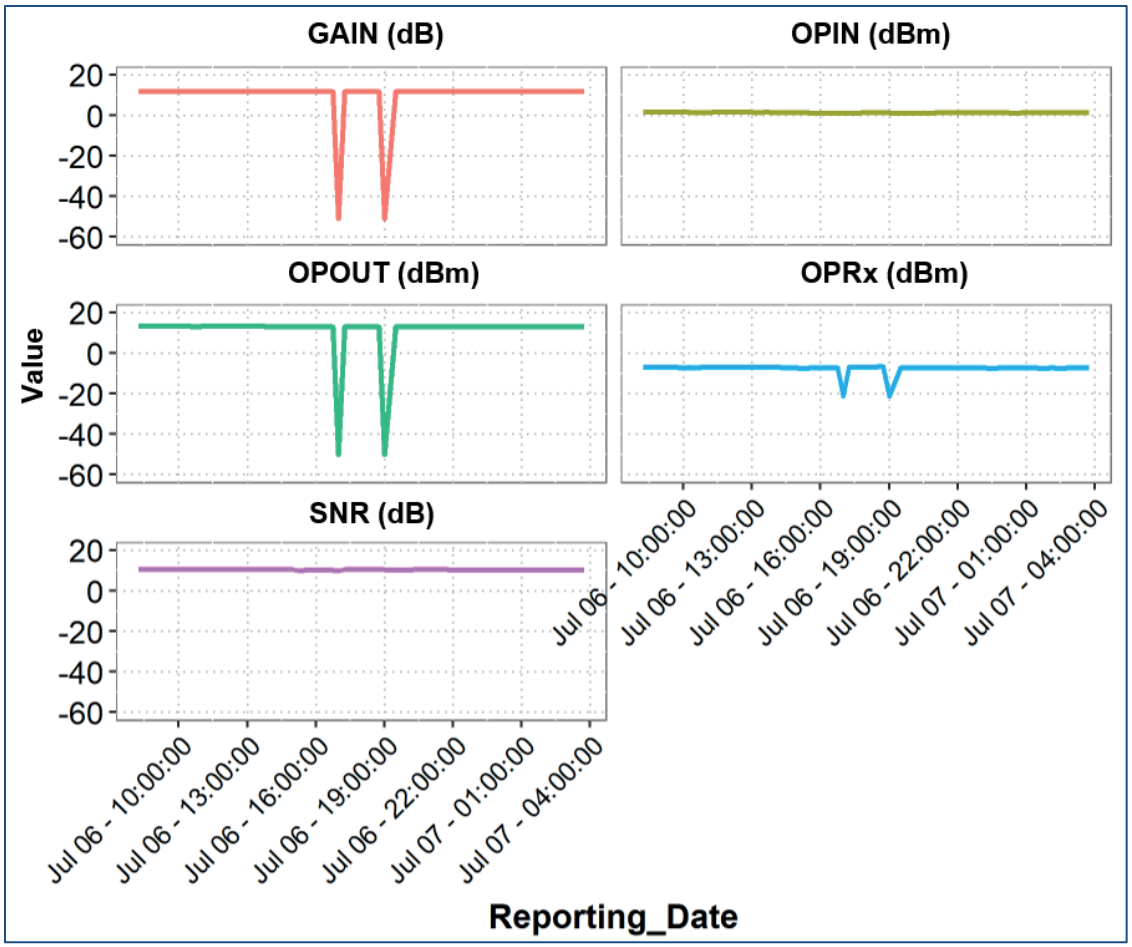

Figure 3.18 Évolution temporelle du SNR, de la puissance reçue au récepteur (*OPRx*), du gain, de la puissance d'entrée (*OPIN*) et de la puissance de sortie (*OPOUT*) de l'amplificateur 11 présentant un pic pour la période du 6 au 7 juillet (CIRCUIT-2)

3.6.3 Catégorie 2 – Un seul canal de la route affectée par les anomalies

La Figure 3.19 présente la représentation simplifiée de la topologie physique des circuits bidirectionnels de la route 3.

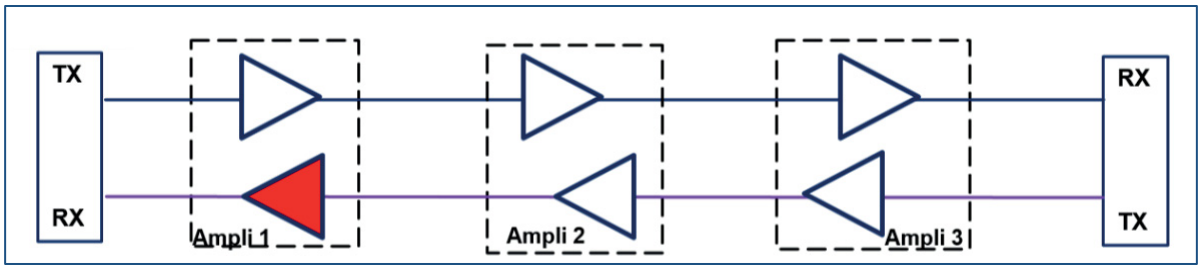

Figure 3.19 Représentation bidirectionnelle de la liaison optique de la route analysée dans la catégorie 2 – Un seul canal de la route affectée Route 3 : CIRCUIT-4, CIRCUIT-5 et CIRCUIT-6

À la Figure 3.8, dans la zone repérée en rouge, le circuit 4 présente une chute de SNR de 4 dB et 2 dB dans la direction A-B et B-A. Le gain des 3 amplificateurs est stable mais la puissance d'entrée et de sortie des amplificateurs présente une chute de 4 dB dans la direction A-B tel que présenté à la Figure 3.20 encerclé en rouge.

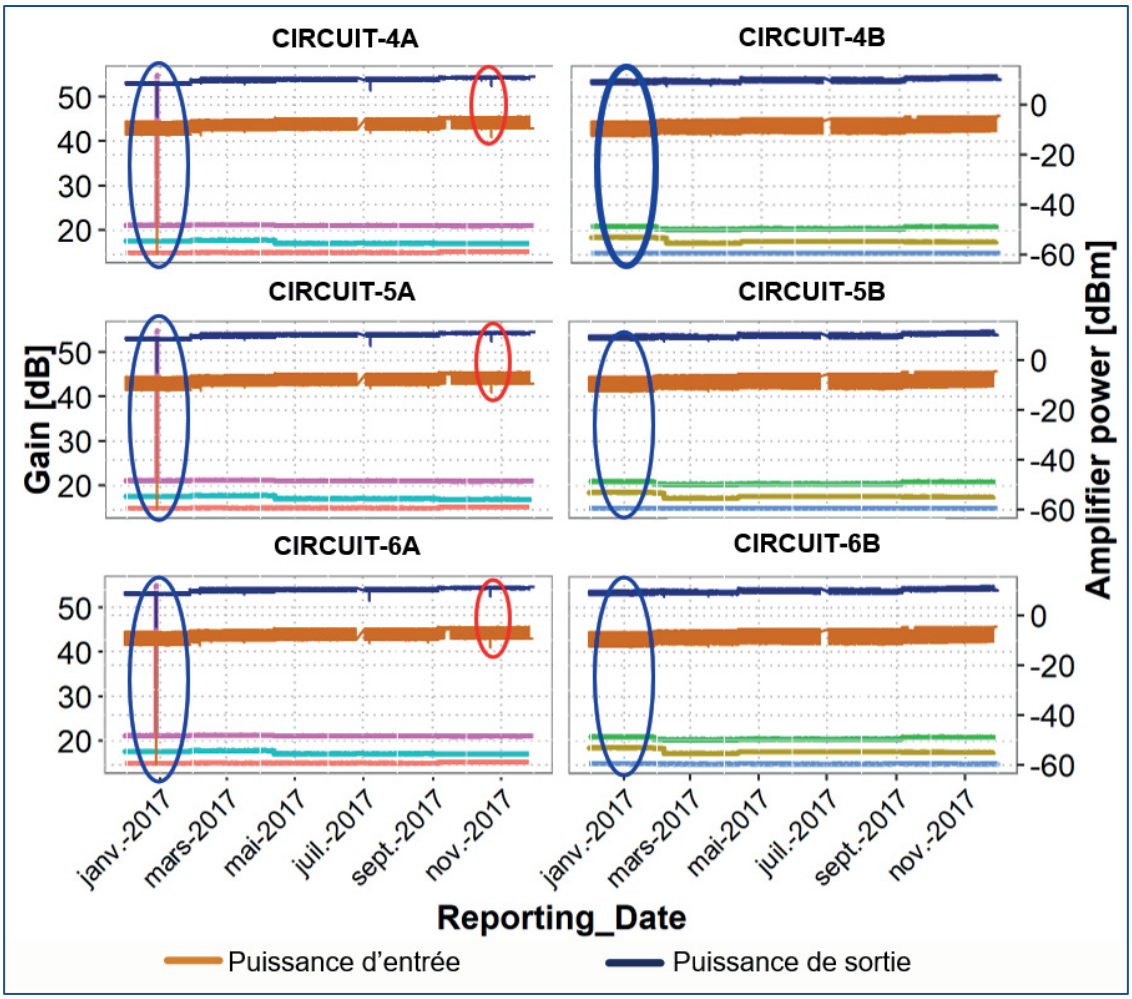

Figure 3.20 Évolution temporelle du gain des amplificateurs optiques et de leurs puissances d'entrée et de sortie de la catégorie 2
Route 3 : CIRCUIT-4 et CIRCUIT-5 et CIRCUIT-6
Direction : aller = A-B et retour = B-A

En observant les puissances d'entrée et de sortie des amplificateurs, il apparait que la seconde erronée enregistrée en janvier (encerclée en bleu dans les Figure 3.9 et Figure 3.20) bien que très peu élevée (1 ES) est liée en réalité à la chute de la puissance d'entrée d'une trentaine de décibels du dernier amplificateur dans la direction B-A (*ampli 1* en rouge à la Figure 3.19). Cette chute de la puissance d'entrée est accompagnée également d'une chute de la puissance de sortie et d'une valeur de gain qui passe de 20 dB à environ 55 dB. Bien que cet évènement ne soit arrivé que dans une seule direction, la seconde erronée est aussi présente dans les 3 circuits de la direction opposée.

3.6.4 Catégorie 3 – Une direction de la route affectée par les anomalies

La Figure 3.21 présente la représentation simplifiée de la topologie physique des circuits bidirectionnels de la route 4.

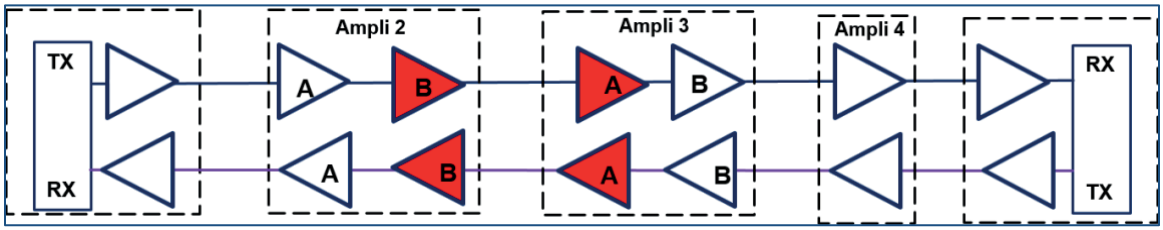

Figure 3.21 Représentation bidirectionnelle de la liaison optique de la route analysée dans la catégorie 3 – Une direction de la route affectée Route 4 : CIRCUIT-7, CIRCUIT-8 et CIRCUIT-9

L'évolution temporelle du SNR et de la puissance reçue au récepteur est présentée aux Figure 3.11 et Figure 3.12. Les circuits *7A*, *8A* et *9A* sont les circuits qui vont du site 2 au site 3 et les circuits *7B*, *8B* et *9B* sont les circuits déployés dans la direction site 3 vers site 2 (voir synoptique à la Figure 3.10).

Tel que relevé plus haut dans le paragraphe 3.4.3, il y a une très forte dynamicité et beaucoup plus d'anomalies dans les circuits de la direction A-B (CIRCUIT-7A, CIRCUIT-8A et CIRCUIT-9A) que dans les circuits de la direction opposée B-A à savoir CIRCUIT-7B, CIRCUIT-8B et CIRCUIT-9B. En revanche, il y a très peu de secondes erronées enregistrées dans la période d'observation soit l'équivalent de 30 mn. Cette dynamicité est également observée dans les puissances d'entrée et de sortie des amplificateurs tel qu'illustré à la Figure 3.22. Cela expliquerait peut-être cette grande variabilité.

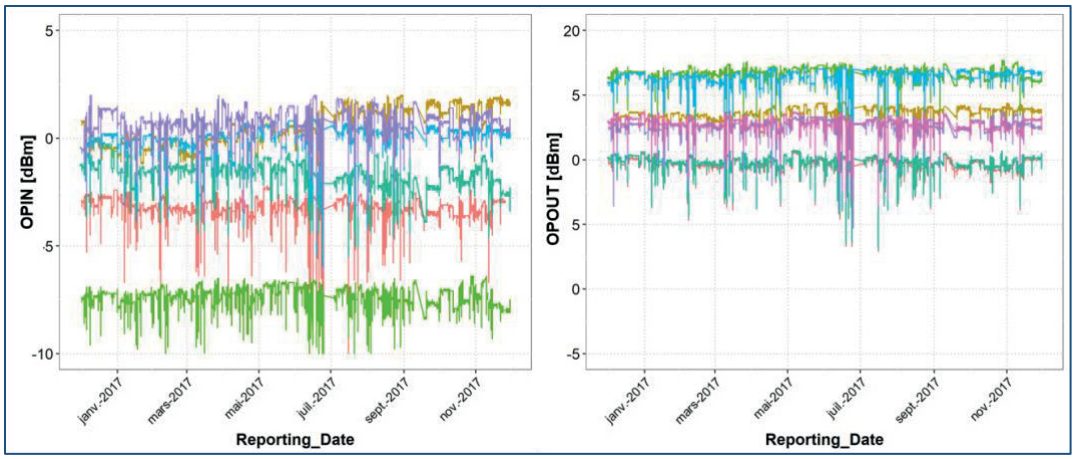

Figure 3.22 Évolution temporelle des puissances d'entrée (OPIN) et de sortie (OPOUT) des amplificateurs dans la direction A-B des circuits de la catégorie 3 – Une direction de la route affectée : Route 4

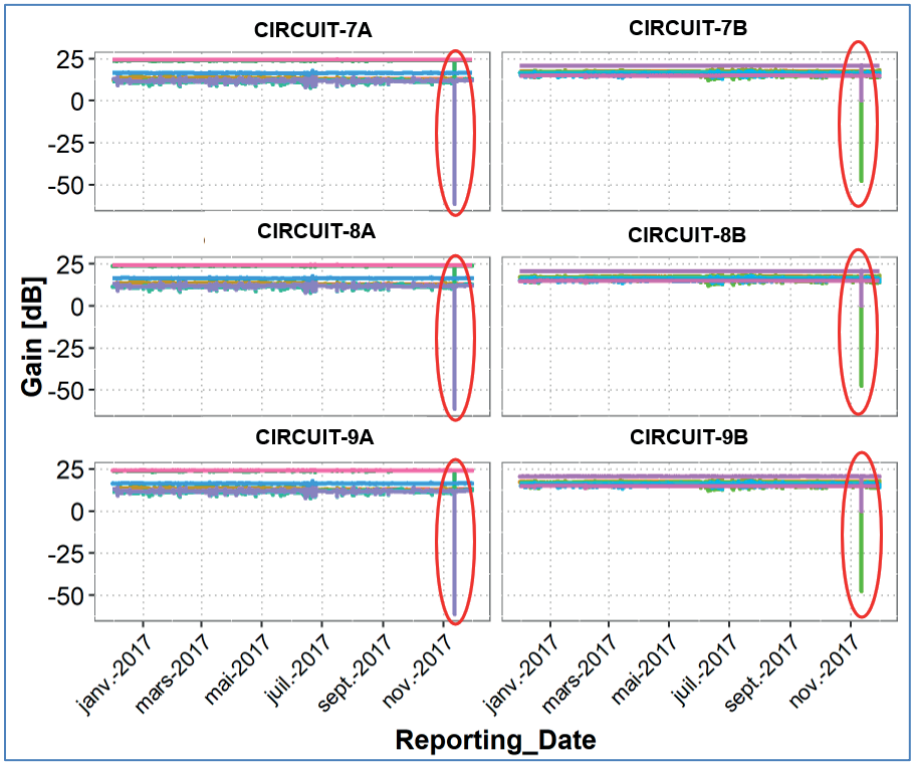

Figure 3.23 Évolution temporelle du gain des amplificateurs optiques et de leurs puissances d'entrée et de sortie de la catégorie 3 – Route 4 : CIRCUIT-7 etCIRCUIT-8 et CIRCUIT-9
Direction : aller = A-B et retour = B-A

La date du 12 novembre est marquée par une chute dans le SNR et la puissance reçue dans la direction A-B et du gain encerclé en rouge dans les Figure 3.11, Figure 3.12 et Figure 3.23. Cette chute est accompagnée de la plus grande occurrence de secondes erronées dans la période monitorée soit une valeur cumulative de 1700 ES, l'équivalent de 30 mn environ. En effectuant un zoom de ces paramètres dans cette journée, deux évènements majeurs sont à signaler tel que cela est présenté à la Figure 3.24 :

- durant le premier évènement, le SNR chute de 1.87 dB dans un intervalle de 45 mn soit entre 5h30 et 6h15. Les gains des amplificateurs restent stables et la puissance reçue chute de 15 dB environ et garde un plateau à -25 dBm pendant 15 mn.
- durant l'évènement 2 qui a lieu entre 7h15 et 7h45, le SNR chute de nouveau de 4 dB et la puissance reçue de 15 dB avec un plateau à -25 dBm pendant 30 mn. Les gains des deux amplificateurs *Ampli 2B* et *Ampli 3A* repérés par les flèches à la Figure 3.24, montrent de grandes chutes. L'amplificateur *Ampli 2B* chute de 24 dB à 0 dB et l'ampli 3A chute de 7.5 dB.

Les secondes erronées apparaissent également dans les circuits de la direction B-A (CIRCUIT-7B, CIRCUIT-8B et CIRCUIT-9B) même si le SNR dans cette période chute de 1 dB et la puissance reçue dans deux des circuits ne présente aucune anomalie.

En observant le gain des amplificateurs dans la direction opposée, il apparaît que les amplificateurs de même niveau que ceux affectés dans la direction A-B soit *Ampli 2A* et *Ampli 3B* sont tout autant affecté, tel que présenté en rouge dans le synoptique de la liaison à la Figure 3.24.

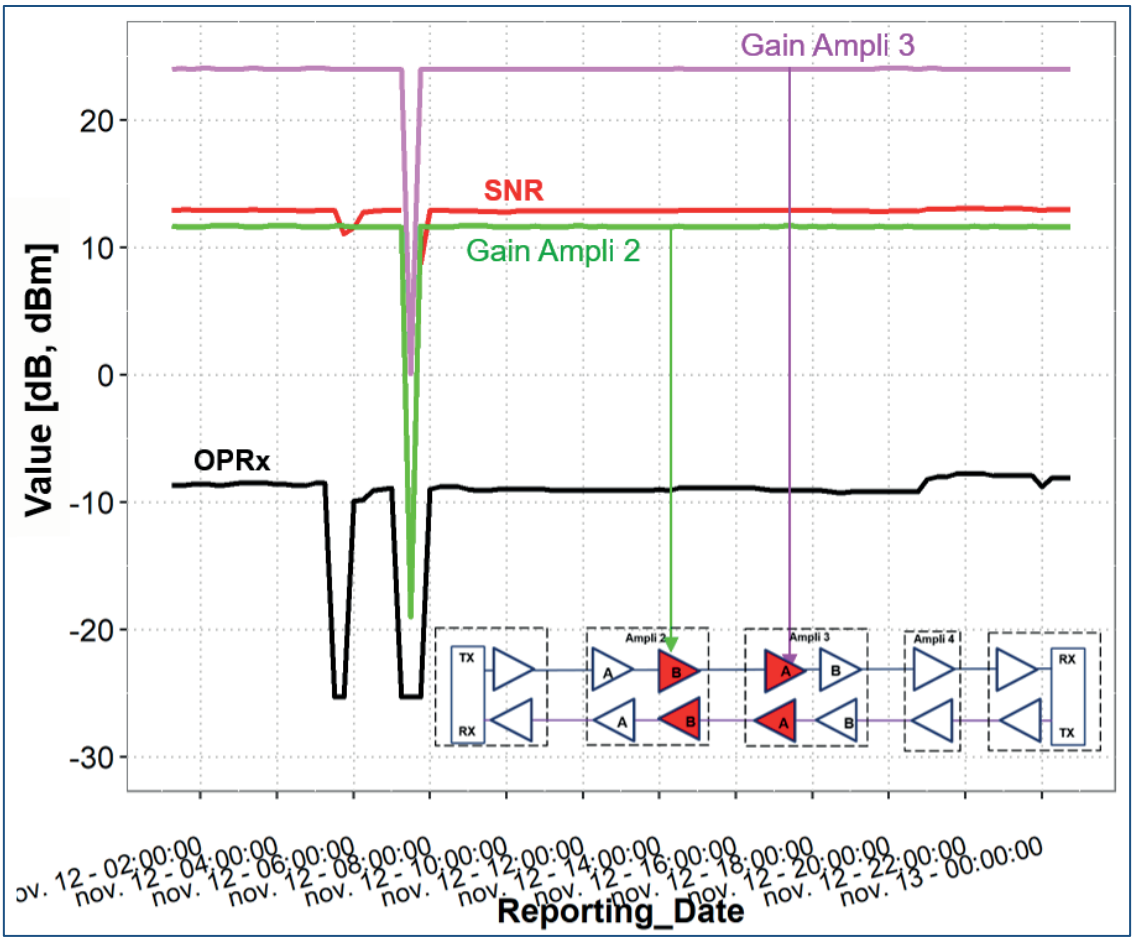

Figure 3.24 Évolution temporelle du SNR, de la puissance reçue au récepteur (*OPRx*), du gain des deux amplificateurs en faute pendant les deux évènements du 12 novembre (CIRCUIT-1 et CIRCUIT-2)

3.7 Discussion et conclusion

L'objectif de l'analyse menée dans ce chapitre consiste à caractériser les connexions optiques du point de vue des anomalies. La technique de détection des anomalies basée sur l'écart interquartile a été choisi pour identifier les anomalies.

Ainsi, trois catégories d'anomalie ont été définies selon que les anomalies affectent tous les canaux, un seul canal ou une seule direction. Trois questions de recherche sont posées :

Les secondes erronées enregistrées sont-elles toujours associées à des anomalies observées dans le SNR, la puissance reçue au récepteur ou le gain, les puissances d'entrée et de sortie des amplificateurs optiques?

L'analyse de corrélation montre une forte corrélation entre la puissance reçue et les secondes erronées et une corrélation négligeable dans le SNR des circuits bidirectionnels choisis dans cette catégorie. Un grand nombre d'évènement de secondes erronées enregistrées semblent plus lié aux anomalies dans la puissance reçue, cela même en considérant l'ensemble des routes monitorées en annexe. Une analyse des gains et puissances d'entrée et sortie des amplificateurs ont permis de voir qu'un grand nombre d'anomalies dans le SNR et la puissance reçue au récepteur peut être associé à des anomalies au niveau des amplificateurs optiques.

L'analyse du CIRCUIT-2 de la catégorie 2 montre une propagation bidirectionnelle des secondes erronées. Bien que l'anomalie ait lieu dans une seule direction, les secondes erronées sont présentes dans les deux directions. Cela est également le cas dans le CIRCUIT-3 de la catégorie 1 et dans les circuits de la catégorie 3. Les circuits de la catégorie 3 bien que dynamique dans le SNR et la puissance reçue au récepteur, c'est-à-dire présentant une grande variabilité n'est pas caractérisé par un grand nombre de secondes erronées enregistrées. Les puissances d'entrée et de sortie des amplificateurs montrent aussi une très grande variabilité mais que peut en être la cause?

À la suite de cette analyse, il ressort qu'il est possible d'identifier différentes catégories d'anomalies. Il demeure que la cause des anomalies n'a pas pu être établie. Des données additionnelles seraient nécessaires afin d'identifier la cause probable des erreurs observées et des anomalies. Parmi les informations additionnelles qui seraient utiles, les relevés d'intervention humaine sur le terrain (entraînant des chutes de puissance au récepteur par manipulation de fibre optique par exemple), des pannes observées au niveau des éléments de réseau et des ajouts et extractions de canaux qui causent des fluctuations de puissance au niveau des amplificateurs.

CHAPITRE 4

REGROUPEMENT DES 155 CIRCUITS BASÉ SUR LEURS CARACTÉRISTIQUES STATISTIQUES DE BASE

4.1 Introduction

Le partitionnement ou regroupement des données, communément appelé clustering, est un processus de regroupement naturel d'un ensemble de données d'observations en plusieurs groupes ou clusters afin que les données d'observations d'un même cluster aient une grande similitude (basée sur une certaine mesure de similarité), mais qu'elles soient très différentes des données d'autres clusters. C'est une technique d'apprentissage machine non supervisé qui est beaucoup utilisée dans de nombreux domaines d'application tels que la biologie, la sécurité, la reconnaissance de forme, l'intelligence économique et la recherche Web (Han, Kamber et Pei, 2011).

Le regroupement est une technique très utile en ce sens qu'il peut conduire à la découverte de groupes auparavant inconnus dans les données. Bien qu'elle soit également une méthode de classification, le terme clustering est utilisé dans ce chapitre pour éviter toute confusion avec la classification, communément utilisée pour désigner la méthode d'apprentissage supervisée.

Au titre des séries temporelles, plusieurs méthodes de regroupement existent dépendamment du type de données à segmenter. Wang et al. (2004) propose deux catégories principales: le regroupement global (''*whole clustering''*) et le regroupement par sous-séquence (''*subsequence clustering''*). Le regroupement global est une méthode appliquée sur plusieurs séries temporelles afin de regrouper celles qui présentent des similitudes. Le regroupement par sous-séquence vise quant à lui à trouver des similitudes et des différences entre différentes fenêtres temporelles coulissantes d'une série temporelle. Au regard du nombre de séries temporelles sous-étude, cette méthode ne convient pas.

Ce chapitre se focalise donc sur la méthode de regroupement global en utilisant une variété de mesures statistiques pour capturer les caractéristiques globales des séries temporelles.

L'objectif est de regrouper les circuits analysés dans le chapitre précédent en considérant un certain nombre d'attributs (*attributs*) extraits à partir des différents paramètres monitorés. Plus précisément, il s'agit d'appliquer sur ces circuits une méthode évolutive de regroupement de séries temporelles applicable aux séries temporelles de longueur variable et insensible aux données manquantes telle que proposée par (Wang et al., 2004; Wang, Smith et Hyndman, 2006).

Contrairement aux techniques d'apprentissage supervisé telles que la classification et la régression, l'analyse par regroupement n'utilise pas de données étiquetées, mais utilise plutôt la similarité entre les caractéristiques des données pour les regrouper en clusters souvent en prélude à une modélisation prédictive.

4.2 Choix de l'algorithme de regroupement

Les deux méthodes de regroupement les plus populaires sont le regroupement hiérarchique et le regroupement par partitionnement (Ganguly, 2017). Encore connues sous le nom de méthodes basées sur la distance, elles doivent leur popularité au fait qu'elles peuvent être utilisées avec presque tous les types de données, à condition qu'une fonction distance appropriée puisse être créée pour ce type de données. De ce fait, le problème du regroupement peut être réduit au problème de trouver une fonction distance (Aggarwal et Reddy, 2014).

4.2.1 Le regroupement hiérarchique

Comme son nom le mentionne, le regroupement hiérarchique crée une hiérarchie à plusieurs niveaux de clusters en créant une arborescence de clusters. Il s'agit d'une approche plus complexe du regroupement de données.

Fondamentalement, chaque élément reçoit son propre cluster. Une paire de clusters est jointe sur la base de similitudes, ce qui donne un cluster de moins. Ce processus est répété jusqu'à ce que tous les éléments soient regroupés.

Le dendrogramme est le graphique qui montre les clusters hiérarchiques. Cet algorithme est limité aux petits ensembles de données en raison de sa complexité de calcul quadratique (Wang et al., 2004).

4.2.2 Le regroupement par partitionnement

Le regroupement par partitionnement est une division de l'ensemble des objets de données en sous-ensembles non superposés (clusters) de telle sorte que chaque objet de données soit dans un sous-ensemble distinct. Étant donné un ensemble de n objets, cette méthode construit k partitions de données, où chaque partition représente un cluster avec $k \leq n$. Les données sont donc divisées en k groupes de sorte que chaque groupe doit contenir au moins un objet.

Le regroupement par partitionnement est retenu pour la suite de l'analyse au regard de la quantité des données et aussi dû au fait que, en plus de sa durée d'exécution mentionnée plus haut, le regroupement hiérarchique n'est pas évolutif. Une fois l'arborescence construite, elle ne peut plus être modifiée (Aggarwal et Reddy, 2014; Zolhavarieh, Aghabozorgi et Teh, 2014). Le regroupement par partitionnement comprend plusieurs algorithmes à savoir : l'algorithme k-moyennes et ses variantes.

L'algorithme des k-moyennes

L'algorithme k-moyennes est l'algorithme par partitionnement le plus largement utilisé dans le l'exploration de données en raison de sa simplicité et de son évolutivité (Critical Data, 2016). Il a été conçu pour regrouper des données numériques en k clusters dont le centre est la moyenne des mesures dans le groupe. Conceptuellement, le centroïde d'un cluster est son point central. L'exécution de l'algorithme k-moyennes peut se résumer en deux phases : la phase d'initialisation et la phase d'itération.

Dans la phase d'initialisation, l'algorithme assigne aléatoirement les points de données dans les *k* clusters. Dans la phase d'itération, l'algorithme calcule la distance entre chaque point et chaque cluster et assigne le point au cluster le plus proche. Ainsi, la composition des clusters et la position des centroïdes changent de positions à chaque fois jusqu'à la convergence.
En d'autres termes jusqu'à ce que la position ne change plus de manière significative ou que la composition des clusters ne soit plus modifiée (Aggarwal et Reddy, 2014; Gan, Ma et Wu, 2007).

Plusieurs mesures de proximité peuvent être choisies lors du calcul du centroïde le plus proche. Ce choix peut avoir une incidence significative sur l'affectation des centroïdes et la performance de l'algorithme. Les différents types de mesures qui peuvent être utilisés sont : la distance de Manhattan (*norme L1*), la distance euclidienne (*norme L2*) et la similarité cosinus. En général, la métrique de la distance euclidienne est le choix le plus populaire pour le k-moyennes. Par ailleurs, l'algorithme k-moyennes utilise une fonction objective appelée la somme des erreurs quadratiques (*Sum of Squared Errors, SSE*) ou somme résiduelle quadratique (*Residual Sum of Squared Errors, RSS*) pour le calcul du regroupement (Aggarwal et Reddy, 2014). Le principe étant de minimiser la somme des distances quadratiques (la variance) entre chaque point de données et le plus proche centroïde (Fränti et Sieranoja, 2019), la *SSE* diminue de façon exponentielle à chaque itération et finit par converger vers un minimum local. La formulation mathématique de la fonction objective se décrit comme suite.

Soit un ensemble de données $D = [x_1, x_2, \ldots, x_n]$, contenant n points de données, et le regroupement $C = [C_1, C_2, \ldots, C_k$ les k clusters distincts], la fonction objective pour ce regroupement est :

$$SSE(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - c_k\|^2, \tag{4.1}$$

Et

$$c_k = \sum_{\frac{x_i \in C_k}{|C_k|}}^{K} x_i \tag{4.2}$$

Où c_k est le centroïde du cluster C_k. Cette fonction objective essaie de rendre les k clusters résultants aussi compacts et aussi séparés que possible les uns des autres. Ce qui signifie que la fonction objective vise une forte similitude intra-cluster et une faible similitude inter-cluster.

Facteurs affectant l'algorithme k-moyennes

Deux facteurs majeurs impactent la performance de l'algorithme k-moyennes. Ce sont le choix des centroïdes initiaux et l'estimation du nombre de clusters. La méthode la plus utilisée pour le choix des centroïdes est la méthode aléatoire telle que décrite dans le paragraphe ci-dessus. Dans cette méthode l'algorithme initialise k centroïdes de façon aléatoire. D'autres méthodes tels que le concept de la densité du plus proche voisin ou méthode *Hartigan* et l'algorithme k-moyennes++ sont proposées afin d'améliorer les performances du k-moyennes (Aggarwal et Reddy, 2014).

Le concept de la densité du plus proche voisin suggère que les points qui sont très séparés et qui ont un grand nombre de points dans leur sphère multidimensionnelle environnante peuvent être de bons candidats pour les points initiaux. Les points suivants sont choisis dans l'ordre de leur densité décroissante tout en respectant simultanément la distance de séparation euclidienne moyenne de tous les points précédents.

L'algorithme k-moyennes++ sélectionne soigneusement les centroïdes initiaux pour le regroupement. Il suit une approche basée sur les probabilités, où le premier centroïde est initialement sélectionné au hasard. Le centroïde suivant sélectionné est celui qui est le plus éloigné du centroïde actuellement sélectionné.
Cette sélection est décidée sur la base d'un score de probabilité pondéré. La sélection se poursuit jusqu'à l'obtention de *k* centroïdes, puis le regroupement de k-moyennes est effectué à l'aide de ces centroïdes.

Estimation du nombre de clusters k

Le problème de l'estimation correcte du nombre de clusters k est l'un des défis majeurs pour le regroupement k-moyennes.

Plusieurs auteurs ont proposé de nouvelles méthodes pour relever ce défi dans la littérature. Quelques-unes des méthodes les plus importantes sont décrites dans la suite du paragraphe (Aggarwal et Reddy, 2014).

- **l'indice Calinski–Harabasz :** il est défini par l'équation (4.3) :

$$CH(K) = \frac{\frac{B(K)}{(K-1)}}{\frac{W(K)}{N-1}} \tag{4.3}$$

Où *N* représente le nombre de points de données. Le nombre de clusters est choisi de sorte à maximiser la fonction *CH(K)*. *B(K)* et *W(K)* sont respectivement la somme de carré inter-cluster et intra-cluster avec K clusters.

- **Le "Gap Statistic" :** Cette méthode permet de produire *B* ensembles de données différents ayant chacun les mêmes plages de données que les données d'origine. La somme des carrés à l'intérieur d'un cluster est calculée pour chacun d'entre eux avec un nombre différent de clusters. $W_b(K)$ est la somme des carrés intra- cluster pour l'ensemble de données.

$$Gap(K) = \frac{1}{B} \times \sum \log(W_b(K)) - \log(W(K)) \tag{4.4}$$

Le nombre de clusters choisi est la plus petite valeur de K qui satisfait à :

$$Gap(K) \geq Gap(K+1) - S_{k+1} \tag{4.5}$$

Où S_{k+1} représente l'estimation de la déviation standard de $\log(W_b(K+1))$

- **Le coefficient de silhouette :** Cette méthode tient compte des distances intra- cluster et inter- cluster. Pour un point x_i donné, la moyenne des distances par rapport à tous les points d'un même cluster est déterminée et la valeur est paramétrée à a_i. Ensuite, pour chaque cluster qui ne contient pas x_i, la distance moyenne de x_i à tous les points de données dans chaque cluster est calculée. Cette valeur est paramétrée à b_i. En utilisant ces deux valeurs, le coefficient de silhouette du point est ainsi estimé. La moyenne de tous les coefficients de silhouette de l'ensemble de données est appelée la largeur moyenne de silhouettes (*average silhouettes width*) pour tous les points dans l'ensemble de données. Pour évaluer la qualité d'un regroupement, le coefficient moyen de silhouette de tous les points peut être calculé.

$$Si = \frac{\sum_{i=1}^{N} \frac{b_i - a_i}{\max(a_i, b_i)}}{N} \tag{4.6}$$

  Une valeur de silhouette *Si* élevée (proche de 1) suggère que les données correspondent bien aux clusters désignés. Une valeur de silhouette faible voire négative (-1), indique une mauvaise correspondance des données au cluster alloué. Un mauvais regroupement peut signifier que la configuration de l'algorithme a trop ou trop peu de clusters (Rousseeuw, 1987).

Les variantes du k-moyennes

En raison de sa simplicité, de son efficacité et de sa large applicabilité, l'algorithme k-moyennes est devenu l'un des algorithmes de regroupement les plus populaires. Cependant, en raison de l'utilisation de la valeur moyenne dans le calcul du centroïde du cluster, le k-moyennes est très sensible aux valeurs aberrantes (Ganguly, 2017).

D’autres variantes plus robustes sont ainsi proposées.

Certaines sont basées sur le choix de différents prototypes représentatifs des clusters (*k-médoïdes, k-médianes*), le choix de meilleures estimations initiales des centroïdes (*k-means intelligent, k-means génétique, Fuzzy k-means*), et sur l'application d'une sorte de technique de transformation des caractéristiques (*k-means pondérée, k-means du kernel*).

- **l’algorithme k-médoïdes :** c’un algorithme de regroupement qui est plus résistant au bruit et aux valeurs aberrantes que k-moyennes. Comme ce dernier, le but de l’algorithme k-médoïdes est de trouver une solution de regroupement qui minimise une fonction objective prédéfinie. Contrairement au k-moyennes où le cluster est représenté par son centroïde, le cluster dans le k-médoïdes le cluster est représenté par un de ses objets prédominants, le médoïde. L'algorithme de k-médoïdes vise à minimiser le critère d'erreur absolue (*absolute error criterion*) qui est une fonction objective qui définit la distance moyenne entre un objet et le médoïde. Afin d’améliorer la qualité du regroupement de l’algorithme k-médoïdes de nombreuses variantes existent.

- **l’algorithme PAM (*Partitioning Around Medoids*):** l’idée de l’algorithme PAM consiste à commencer avec un ensemble *k* de médoïdes puis échanger le rôle entre un objet médoïde et un non-médoïde si cela permet de minimiser la fonction objective (Aggarwal et Reddy, 2014). Son inconvénient majeur est son coût de calcul, ce qui le rend non adaptable pour une population importante de données (Nabila, 2011). Pour pallier à cela, l’algorithme CLARA (*Clustering LARge Application*) qui est une combinaison du PAM avec une méthode d'échantillonnage a été proposé.

CLARA considère de nombreux échantillons issus des données et applique l’algorithme PAM sur chacun d'entre eux pour finalement retourner l'ensemble des médoïdes optimaux (Ganguly, 2017).

- **l’algorithme k-médianes :** il calcule la médiane de chaque cluster par opposition au calcul de la moyenne du cluster dans k-moyennes.

L'approche k-médianes est également plus stable au bruit et aux valeurs aberrantes, car la médiane d'un ensemble de valeurs est généralement moins sensible aux valeurs extrêmes des données. Cet algorithme choisit k centres de clusters qui visent à minimiser la somme d'une mesure de distance entre chaque point et le centre du cluster le plus proche. La mesure de proximité utilisée dans l'algorithme des k-médianes est la distance Manhattan par opposition au carré de la distance euclidienne utilisée dans l'algorithme k-moyennes. La fonction de référence pour l'algorithme des k-médianes est défini comme suit :

$$S = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left| x_{ij} - med_{kj} \right| \tag{4.7}$$

Où x_{ij} représente le *j-ème* attribut de l'instance x_i et med_{kj} représente la médiane du *j-ème* attribut dans le *k-ème* cluster C_k. Pour un ensemble de cluster *k* allant de *1 à K*, pour chaque objet x_i dans le cluster C_k, la valeur absolue de la distance de cet objet est déterminée par rapport à la médiane.

L'objectif du regroupement k-médianes est de déterminer les sous-ensembles de points médians qui minimisent le coût d'assignation des points de données aux médianes les plus proches (Aggarwal et Reddy, 2014).

4.2.3 Métriques de validation du regroupement

Le regroupement est une méthode d'apprentissage machine non supervisée visant à partitionner un ensemble de données en un ensemble de groupes ou clusters.
Une question importante qui se pose lors de l'évaluation des clusters est de savoir si un cluster donné est "réel". Autrement dit, le cluster représente-t-il une structure réelle dans les données, ou est-il un artefact de l'algorithme de mise en clusters (Zumel, Mount et Porzak, 2014)?

Dans ce sens, différentes métriques de validation sont à considérer lors de l'évaluation finale de l'algorithme de regroupement : les métriques de validation externe, la validation relative et

les métriques de validation interne. La différence principale entre ces deux métriques réside dans le fait que des informations externes soient utilisées ou non pour la validation du regroupement.

Contrairement aux métriques de validation externe, qui utilisent des informations externes non présentes dans les données, les mesures de validation interne reposent uniquement sur les informations contenues dans les données. Les métriques de validation interne évaluent la validité du regroupement sans tenir compte des informations externes. Puisque les métriques de validation externes connaissent à l'avance le nombre réel de clusters (*ground truth*) et la structure sous-jacente des données, elles sont principalement utilisées pour choisir un meilleur algorithme de regroupement sur un ensemble de données spécifique. Les métriques de validation internes peuvent aussi être utilisées pour choisir le meilleur algorithme de regroupement et le plus approprié ainsi que le nombre optimal de clusters sans aucune donnée externe.

En pratique, les informations externes telles que les étiquettes de classe ne sont souvent pas disponibles dans de nombreux scénarios d'application. Par conséquent, dans le cas où aucune information externe n'est disponible comme c’est le cas dans cette étude, les métriques de validation internes sont la seule option pour la validation des clusters (Alkhayrat, Aljnidi et Aljoumaa, 2020; Liu et al., 2010).

Métriques de validation internes des clusters

L'objectif du regroupement étant de faire en sorte que les objets d'un même cluster soient similaires et que les objets de différents clusters soient distincts, les métriques de validation internes sont souvent basées sur les deux critères suivants (Liu et al., 2010):

- **la compacité :** elle mesure le degré de cohésion entre les objets d'un cluster. Ces métriques de validation internes évaluent la compacité des clusters basée sur la variance. Une variance plus faible indique une meilleure cohésion.

- **la séparation :** Elle mesure le degré de distinction ou de séparation d'un cluster par rapport aux autres clusters.

Plusieurs métriques de validation internes sont répertoriées dans la littérature. Les plus couramment utilisés pour mesurer la qualité du regroupement sont:

- le R-carré (*R-Squared, RS*) est le rapport entre la somme des carrés entre les clusters et la somme totale des carrés de l'ensemble des données. Il mesure le degré de différence entre les clusters c'est-à-dire qu'il indique dans quelle mesure les groupes sont différents les uns des autres.
- la statistique modifiée de *Hubert Γ* (Γ) évalue la différence entre les clusters en comptant les désaccords de paires d'objets de données dans deux partitions.
- l'indice Calinski-Harabasz (*CH*) évalue la validité des clusters en se basant sur la somme moyenne des carrés inter-cluster et intra-cluster tel que décrit plus haut dans l'estimation du nombre de *k*.
- l'indice I (*I*) mesure la séparation sur la base de la distance maximale entre les centres des clusters, et mesure la compacité en se basant sur la somme des distances entre les objets et le centre de leur cluster.
- l'indice de Dunn (D) utilise la distance minimale par paire entre les objets de différents clusters comme distance inter-cluster et le diamètre maximal entre tous les clusters comme compacité intra-cluster.

La valeur de l'indice de Dunn devrait être importante si les clusters de l'ensemble des données sont bien séparés. Si l'ensemble de données comporte des clusters compactes et bien séparées, la distance entre les clusters devrait être plus grande et le diamètre des clusters devrait être plus petit. Les clusters sont compacts et bien séparés en maximisant la distance inter-cluster tout en minimisant la distance intra-cluster.

Le nombre optimal de clusters est déterminé en maximisant la valeur de ces indices.

- l'indice de silhouette (*Si*) valide la performance de regroupement en fonction de la différence par paire des distances intra – cluster et inter – cluster. En outre, le nombre optimal de clusters est déterminé en maximisant la valeur de cet indice.
- l'indice Davies-Bouldin (*DB*) mesure la similarité moyenne entre chaque cluster et son plus proche voisin. L'objectif de cet indice est d'obtenir une variance minimale au sein des clusters et une variance maximale entre les séparations de clusters. Plus l'indice est petit, meilleur est le résultat du regroupement (Nawrin, Rahman et Akhter, 2017).
- l'indice de Xie-Beni (*XB*) définit la séparation inter-cluster comme la distance carrée minimale entre les centres des clusters, et la compacité intra-cluster comme la distance carrée moyenne entre chaque objet de données et le centre de son cluster d'appartenance. Le nombre optimal de clusters est atteint lorsque le minimum de *XB* est trouvé.
- l'indice SD (*SD*) dont l'idée est basée sur les concepts de la dispersion moyenne et de la séparation totale des clusters. Le premier terme évalue la compacité sur la base des variances des objets des clusters, et le second terme évalue la différence de séparation sur la base des distances entre les centres des clusters. L'indice SD est la somme de ces deux termes, et le nombre optimal de clusters peut être obtenu en minimisant la valeur de SD.
- l'indice *S_Dbw* prend en compte la densité pour mesurer la séparation entre les clusters. L'idée de base est que pour chaque paire de centres de clusters, au moins une de leurs densités doit être supérieure à la densité de leur point médian. La compacité intra-cluster est la même qu'avec l'indice *SD*. De même une valeur minimale de *S_Dbw* indique un nombre optimal de clusters (Halkidi et Vazirgiannis, 2001).

4.3 Regroupement des 155 circuits monitorés

4.3.1 Méthodologie

L'objectif principal du présent chapitre est de regrouper les 155 circuits en groupes selon leur similitude en considérant un certain nombre de caractéristiques du SNR. L'objectif spécifique revient à tenter de répondre à la question suivante : Ces clusters révèlent-ils d'autres informations sous-jacentes selon la situation géographique? La longueur du lien?

Pour ce faire, un algorithme de regroupement est appliqué aux données des 155 circuits monitorés dans le réseau de production de l'opérateur. Les données monitorées sont collectées toutes les 15 minutes à partir du système de gestion du réseau (*Network Management System, NMS*) pendant une période de 12 mois. Cette portion de réseau sous analyse comprend douze (12) circuits de production réseau avec différents sites transportant du trafic actif.

À partir des séries de données temporelles constituées de valeurs du rapport signal sur bruit (SNR) des 155 circuits (canaux WDM) calculées à partir des mesures de taux d'erreur binaire, des métriques statistiques sont dérivées pour les valeurs de SNR sur chaque circuit : les valeur maximale (*Max*), minimale (*Min*) et moyenne (*Mean*), l'écart interquartile (*IQR*) et l'écart type (*standard deviation*). Une métrique supplémentaire est introduite; une valeur minimale du SNR requis par canal (ou valeur seuil pour éviter des erreurs) est définie. Ces métriques sélectionnées sont appelées attributs ou caractéristiques. La Figure 4.1 décrit les étapes dans l'application du regroupement:

- **le traitement de données :** une fois les données collectées, elles sont transformées et la tendance de regroupement vérifiée. Cette étape va donc permettre de définir si les données sous analyse contiennent des regroupements naturels. La transformation des données permet de mettre à l'échelle les variables stockées dans divers formats ou unités dans les données brutes en formats ou unités plus uniformes.
- une fois l'ensemble de données intégré et transformé, la sélection des attributs permet de supprimer les variables redondantes, ainsi que de réorganiser les données de manière efficace et "ordonnée" pour l'analyse (Critical Data, 2016).
- les attributs sélectionnés sont soumis à l'algorithme de regroupement et les clusters correspondants sont produits testés et validés à l'aide des métriques de validation.
- les circuits correspondant aux clusters définis sont alors analysés afin de découvrir des caractéristiques (*patterns*) cachés.

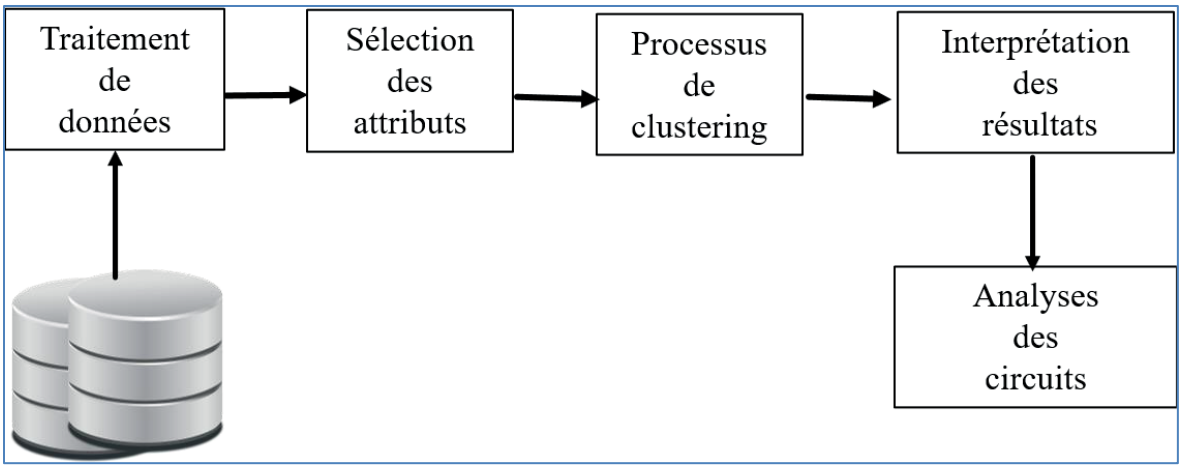

Figure 4.1 Méthodologie de la mise en cluster des 155 circuits

4.3.2 Transformation des données

La normalisation est l'une des étapes clés du prétraitement des données dans un contexte où les valeurs de SNR (et donc de marge) pour les 155 circuits sont différentes. L'application de l'étape de normalisation permet d'obtenir des résultats de regroupement de meilleure qualité, plus efficaces et plus précis (Alkhayrat, Aljnidi et Aljoumaa, 2020).

Il existe de nombreuses méthodes de normalisation des données entre autres la normalisation *min-max*, la normalisation z-*score* et la normalisation par *échelle décimale*.

Mohamad et Usman (2013) ont analysé les performances de ces trois méthodes sur l'algorithme k-moyennes et le résultat obtenu par la méthode de normalisation du z-score est plus efficace et plus performant que les méthodes de normalisation à échelle min-max et décimale. Ce dernier sera donc retenu pour la normalisation des données dans ce chapitre.

Dans la normalisation z-score, les valeurs d'un attribut A, sont normalisées sur la base de la moyenne et de l'écart type de A ou de l'écart absolu de A, plus robuste que celle de l'écart type. Cette méthode préserve la plage des données (maximum et minimum) et tient compte de

la dispersion de données (écart-type, variance) (Han, Kamber et Pei, 2011). De ce fait, une valeur v_i de l'attribut A est normalisée en v_i' en calculant :

$$v_i' = \frac{v_{i-\bar{A}}}{\sigma_A} \tag{4.8}$$

Où $\bar{A}$ et σ_A sont respectivement la moyenne et l'écart type de l'attribut A.

À la suite de la normalisation des données, une autre étape importante du prétraitement des données est la réduction de la dimensionnalité ou l'importance des attributs.

4.3.3 Tendance des clusters dans les données

Le regroupement se résume à classifier un ensemble d'objets en groupes aux caractéristiques similaires. Comme il a été mentionné plus haut, l'algorithme a tendance à regrouper les données dans des groupes même en l'absence de clusters dans les données. Avant d'appliquer une méthode de regroupement aux données, il est donc important d'évaluer si l'ensemble de données sous analyse contient effectivement des clusters significatifs (c'est-à-dire des structures non aléatoires) ou non.

En d'autres termes il s'agit de savoir si l'ensemble de données présente une prédisposition à se regrouper en clusters naturels sans forcément identifier les groupes eux-mêmes.
Ce processus est appelé évaluation de la tendance du regroupement ou la faisabilité du regroupement.

Deux méthodes d'évaluation existent soit la méthode statistique : la *statistique de Hopkins* et la méthode visuelle, la sériation ou VAT (*Visual Assessment of cluster Tendency*) (Bezdek, Hathaway et Huband, 2007; Kassambara, 2020).

La statistique de Hopkins

Cette statistique est utilisée pour évaluer la tendance au regroupement d'un ensemble de données en mesurant la probabilité qu'un ensemble de données considéré soit généré par une distribution uniforme des données.

Soient un ensemble de données D et *m* échantillons uniformes $(p_1, p_2, \dots, p_m)$ issus de D. Soit $w_i = dist(p_i, p_j)$ la distance de chaque point p_i à son plus proche voisin p_j et $u_i = dist(q_i, q_j)$ la distance de chaque point q_i tiré d'un ensemble de points aléatoires à son plus proche voisin q_j dans D. La statistique de Hopkins (*H*) est (Banerjee et Dave, 2004; Kassambara, 2020; Lawson et Jurs, 1990) :

$$H = \frac{\sum_{j=1}^{m} u_i}{\sum_{j=1}^{m} u_i + \sum_{j=1}^{m} w_i} \tag{4.9}$$

En résumé, la statistique de Hopkins est la distance moyenne du plus proche voisin dans l'ensemble de données aléatoires divisée par la somme des distances moyennes du plus proche voisin dans l'ensemble de données réelles et dans l'ensemble de données aléatoires simulées.

Ainsi :

- **pour H = 0** : $u_i << w_i$ et cela signifie qu'il n'y a pas de structure dans les données.
- **pour H ~ 0.5** : $u_i \sim w_i$ ou $u_i << w_i$, cela veut dire qu'il n'y a pas de distribution régulière des objets plus précisément il n'y a pas de clusters.
- **pour 0.5 < H ≤ 1** : $u_i >> w_i$ alors une structure en clusters existe dans les données.

La méthode visuelle

La sériation est le second moyen de vérifier la tendance au regroupement. C'est un outil graphique pour l'évaluation visuelle de la tendance des données à former des clusters, basée sur la matrice de dissimilarité. Dans la matrice de dissimilarité, chaque cellule fait référence à une dissimilitude entre une paire d'objets. Les petites dissimilitudes sont représentées par des nuances sombres et les grandes dissimilitudes sont représentées par des nuances claires. Dans le graphique, les dissimilitudes sont réorganisées de telle manière que, grosso modo, les blocs diagonaux (ombrés en foncé) correspondent à des clusters dans les données. Par conséquent, k blocs sombres le long de la diagonale principale suggèrent que les données contiennent k clusters (non encore trouvés) et la taille de chaque bloc représente la taille approximative du cluster (Bezdek et Hathaway, 2002).

Résultats de la tendance des clusters sur les données des 155 circuits

La statistique de Hopkins H mesure la probabilité qu'un ensemble de données déterminé soit généré par une distribution uniforme des données. Ainsi, les hypothèses nulles et alternatives sont définies comme suit :

- **hypothèse nulle :** l'ensemble de données est uniformément distribué c'est-à-dire qu'il n'y a pas de groupes significatifs.
- **hypothèse alternative :** l'ensemble de données D n'est pas uniformément distribué c'est-à-dire qu'il contient des clusters.

La statistique de Hopkins obtenue avec la fonction *get_clust_tendency ()* de la librairie *clustertend*, donne une valeur de 0.83 à un niveau de confiance de 90 %, démontrant ainsi qu'il y a une structure dans les données monitorées.

Afin de le confirmer graphiquement, la méthode virtuelle est représentée à la Figure 4.2 par une carte de chaleur (heat map). Le long de la diagonale, des blocs représentant de probables clusters sont visibles et identifiés 1, 2 et 3. Ce qui signifie que les données peuvent être regroupées en clusters. La taille de chaque bloc représente la taille approximative du cluster

tel que mentionné plus haut. Par ailleurs, la présence des clusters 4 et 5 de même couleur que les clusters 1 et 2 situés le long de la diagonale montre que les deux clusters 1 et 2 ne seront pas clairement distincts et qu'il y aura possiblement une répartition imparfaite des objets entre ces deux clusters. Ce qui n'est pas le cas pour le troisième cluster.

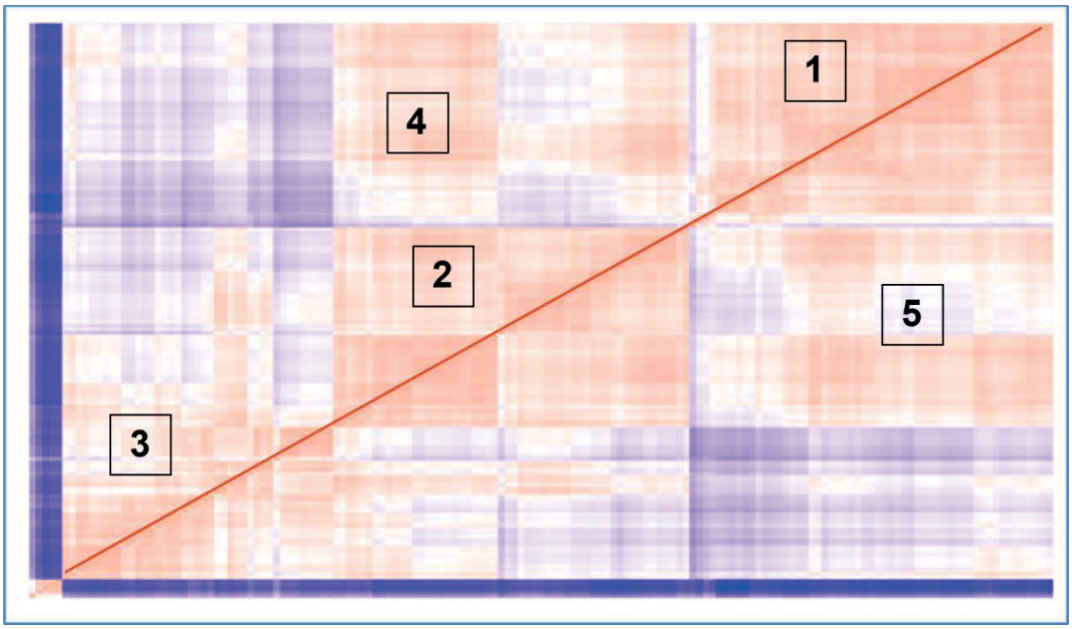

Figure 4.2 Représentation pixellisée de la matrice de dissimilarité des 155 circuits

Les données ayant révélé des clusters significatifs, la prochaine étape est de déterminer l'importance des différents attributs et procéder à une sélection.

4.3.4 L'importance des attributs

L'objectif de la sélection des attributs pour l'apprentissage non supervisé est de trouver le plus petit sous-ensemble de caractéristiques qui permet de découvrir au mieux les groupes "naturels intéressants" (clusters) à partir des données selon le critère choisi (Dy et Brodley, 2004). En effet, elle vise à filtrer les attributs non pertinents, redondants et les échantillons de données bruyants. Elles sont donc appliquées notamment pour:

- La simplification des modèles de données pour en faciliter l'interprétation.

- La réduction du temps de traitement des données.
- La réduction du sur-apprentissage.

Par ailleurs, les attributs doivent être sélectionnés de manière à ce que la précision et les performances ne soient pas affectées (Alkhayrat, Aljnidi et Aljoumaa, 2020; Aparna et Paul, 2016). La sélection de attributs repose sur deux approches principales à savoir: les approches par filtrage et les approches embarquées.

Dans les approches par filtrage, les attributs sont d'abord sélectionnés sans mention de l'algorithme, puis ce sous-ensemble de attributs est utilisé pour exécuter l'algorithme de regroupement. Elles sont basées sur l'application de métriques dont les plus courantes sont : les métriques de dépendance telle que la corrélation, les métriques d'information (le gain d'information) et les métriques de distance.

Contrairement aux approches par filtrage, les approches imbriquées (*wrappers approach*) utilisent la performance en aval pour sélectionner les attributs. En effet, un algorithme de regroupement est appliqué à l'ensemble de données brutes afin d'identifier les attributs les plus pertinents (Saeys, Inza et Larrañaga, 2007).

La technique ReliefF

ReliefF est un algorithme de sélection des caractéristiques basé sur le filtrage développé par Kira et Rendell (1992) qui élimine la redondance dans les attributs. Il s'appuie notamment sur la mesure des similarités et dissimilarités dans l'ensemble des cas analysés, et permet ainsi d'estimer la pertinence (ou la non-pertinence) des différentes caractéristiques à l'aide d'un score global.

Pour évaluer la pertinence d'une valeur, l'algorithme va en quelque sorte chercher à savoir de quelle façon une caractéristique évolue au sein d'une même classe et à l'extérieur de cette classe. Plus une valeur varie au sein d'un cluster, moins celle-ci est importante. À l'opposé,

plus elle évolue entre différentes classes, plus son score d'importance augmente. Relief choisit donc une instance X parmi l'ensemble de données et ses deux plus proches voisins : un de la même classe appelée "*near-hit*" et l'autre d'une classe opposée appelée "*near-miss*". Ensuite, il met à jour le vecteur de pondération W pour chaque triplet et l'utilise pour déterminer le vecteur de caractéristiques de pertinence moyenne (Wosiak et Zakrzewska, 2018).

La technique du gain d'information (*Information Gain, IG*)

L'algorithme de gain d'information est une méthode de filtrage utilisée pour le classement des attributs (*feature ranking*). Le principe du *feature ranking* est de noter les variables selon un algorithme donné et de retourner les attributs les plus significatifs. De ce fait, l'algorithme *IG* classe des sous-ensembles de caractéristiques en mesurant la diminution de l'entropie (Alhaj et al., 2016). L'entropie peut être considérée comme une mesure de l'incertitude, car elle permet de savoir à quel point une variable donnée est pure ou impure (Phyu et Oo, 2016). Par ailleurs, le principe de l'algorithme du gain d'information est de chercher un sous-ensemble minimal d'attributs maintenant la cohérence. Une incohérence survenant lorsque deux éléments ont la même description mais sont de classes différentes (Pacheco, 2015). Les variables qui maximisent le gain d'information sont celle retenues par cette technique.

La technique de la corrélation

La méthode de corrélation de Pearson, accompagnée d'un *test t*, est utilisée pour trouver l'association entre les différentes caractéristiques. Le coefficient de corrélation du *test t* vérifie si l'association entre les caractéristiques est statistiquement significative. Le principe est le suivant : les caractéristiques ayant une corrélation élevée sont linéairement dépendants et ont donc presque le même effet sur la variable dépendante.

Ainsi, lorsque deux caractéristiques présentent une corrélation élevée, l'une des deux est utilisée (Blessie et Karthikeyan, 2012; Pacheco, 2015). La valeur seuil du coefficient de corrélation à partir duquel la corrélation est élevée est de 0.70 en considérant le Tableau 2.1.

4.3.5 Sélection des attributs dans les données de monitoring des 155 circuits

Les attributs considérés dans cette analyse sont déterminés à la suite de l'analyse des routes présentées dans le CHAPITRE 3, l'idée étant de trouver un regroupement des circuits selon un certain nombre de caractéristiques. Ces attributs sont : la valeur moyenne du SNR (*Mean*), la valeur maximale du SNR (*Max*), l'écart-type du SNR (*Standard deviation*), l'écart-interquartile du SNR (*IQR*) et la métrique *Diff_min* qui est dérivée à partir de l'équation :

$$Diff_{min} = SNR_{min} - SNR_{seuil} \tag{4.10}$$

Où SNR_{min} représente la valeur minimum du SNR, et SNR_{seuil} , le seuil de SNR (8.65 dB) correspondant à une valeur seuil de taux d'erreur binaire de 3.8×10^{-3}défini pour les systèmes cohérents à 100 Gbit/s selon Sun, Wu et Roberts (2008).

Méthode de filtrage

Les résultats de sélection de ces attributs en appliquant les trois techniques décrites plus haut, à savoir la technique ReliefF, la technique du gain d'information et la technique de corrélation, sont représentés dans les figures ci-dessous. De ce fait, la Figure 4.3 présente le classement des attributs pour la technique ReliefF et les attributs sélectionnés dans le cas de la technique du gain d'information.

La Figure 4.4 quant à elle, présente la matrice de corrélation de tous les attributs (a) et la corrélation des attributs sélectionnés issus de la technique basée sur la corrélation (b). Le récapitulatif des différents choix des différents algorithmes est représenté dans le Tableau 4.1.

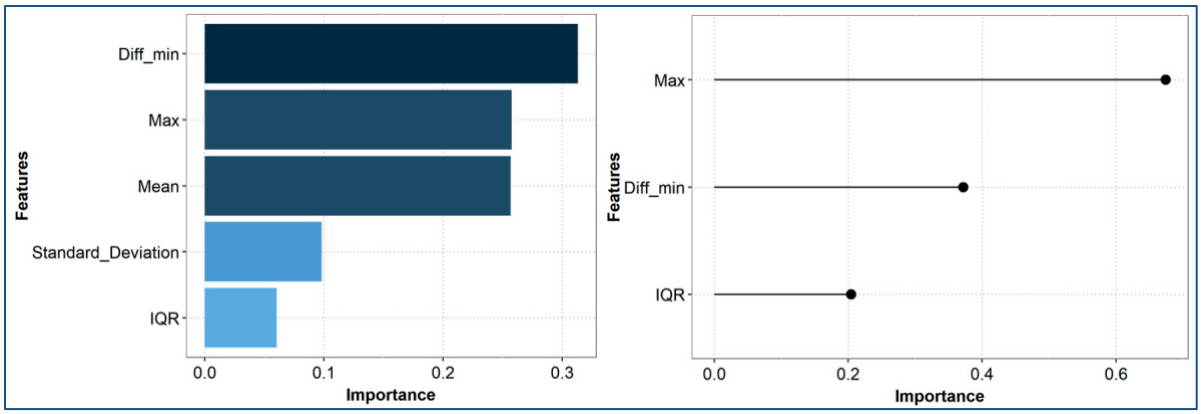

Figure 4.3 Sélection des attributs : Technique ReliefF (*barplot*) et technique du gain d'information (*lollipop*) : Mean (valeur moyenne du SNR), Max (valeur maximale du SNR), Standard deviation (écart-type du SNR), IQR (écart-interquartile du SNR) et Diff_min (différence SNR_{min} et SNR_{seuil})

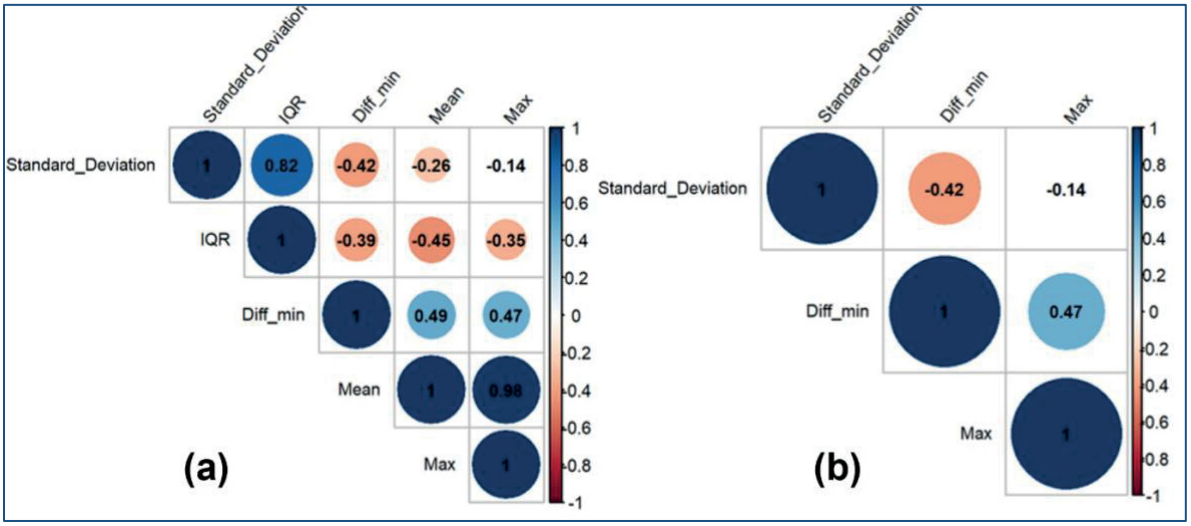

Figure 4.4 Sélection des attributs (features) basée sur la technique de la corrélation : la corrélation de tous les attributs (a) et la corrélation des attributs sélectionnés (b) : Mean (valeur moyenne du SNR), Max (valeur maximale du SNR), Standard deviation (écart-type du SNR), IQR (écart-interquartile du SNR) et Diff_min (différence SNR_{min} et SNR_{seuil})

Tableau 4.1 Synthèse des attributs sélectionnés par chaque technique

Modèle	**Technique**	**Attributs sélectionnés**
Modèle 1	Modèle de base	Max, Mean, Standard_Deviation, IQR et Diff_min

Modèle 2	ReliefF	Diff_min, Max et Mean
Modèle 3	Gain d'information	Max, Diff_min et IQR
Modèle 4	Corrélation	Standard_Deviation, Diff_min et Max

Le modèle 1 représente le modèle de base, c'est-à-dire qu'il regroupe tous les attributs considérés dans l'analyse par défaut. Il s'agit des valeurs maximale, moyenne, de l'écart-type, de l'écart interquartile et de Diff_min qui représente la marge. La marge étant la différence entre le seuil de 8.53 dB en dessous duquel on observe des erreurs.

À ce stade de la sélection des attributs les plus importants, les attributs sélectionnés varient d'une technique à l'autre. Une méthode imbriquée est donc utilisée pour analyser la validité des combinaisons d'attributs proposées.
Contrairement aux méthodes de filtrage utilisées dans les modèles 2, 3 et 4 qui ne prennent pas en compte l'algorithme de regroupement, les approches imbriquées quant à elles, appliquent l'algorithme de regroupement afin d'évaluer la validité des combinaisons d'attributs.

Méthode imbriquée

Le principe de la méthode imbriquée consister à appliquer l'algorithme k-moyennes sur chacune des combinaisons d'attributs (modèles 1, 2, 3 et 4). L'idée est de choisir le meilleur modèle pour un regroupement optimum des circuits. Par ailleurs, avant d'appliquer l'algorithme de regroupement, il est important de déterminer le nombre de clusters.

Sélection du nombre optimal de clusters

L'estimation du nombre optimal de clusters *k* est l'un des principaux défis à relever pour le regroupement du k-moyennes. De ce fait, une grande variété d'indices a été proposée dans la littérature, afin de définir ce nombre optimal de clusters.
Pour une estimation pertinente de *k*, la fonction R *NbClust* (Charrad et al., 2014) qui comprend 30 indices de validation de regroupement par partitionnement et hiérarchique confondues, est utilisée. L'idée est de permettre la détermination du nombre de clusters le plus approprié pour

l'ensemble de données d'intérêt. Et cela en faisant une évaluation simultanée de plusieurs schémas de regroupement tout en faisant varier le nombre de clusters, les mesures de distance (euclidienne, Manhattan, …) et les méthodes de regroupement (k-moyennes, k-médiane, …).

Les résultats de l'estimation du nombre optimal de clusters sont représentés aux Figure 4.5 et Figure 4.6. Le nombre optimal de clusters est fixé à 5 pour le modèle 1 comprenant tous les attributs. Le modèle 2 qui représente les attributs sélectionnés par la technique *ReliefF* a un nombre optimal de 3 clusters. Un nombre optimal de 4 clusters est désigné par un ensemble de huit indices de validation pour les modèles 3 et 4 qui regroupent respectivement les attributs sélectionnés par les techniques du gain d'information et de corrélation.

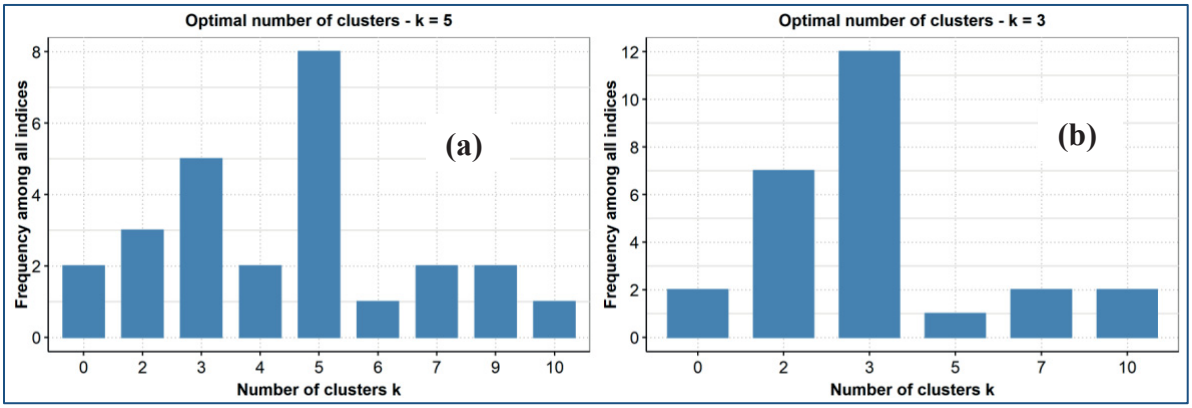

Figure 4.5 Diagramme de fréquence en fonction du nombre de clusters pour :
(a) le modèle 1 (tous les attributs) et (b) le modèle 2 (technique ReliefF)

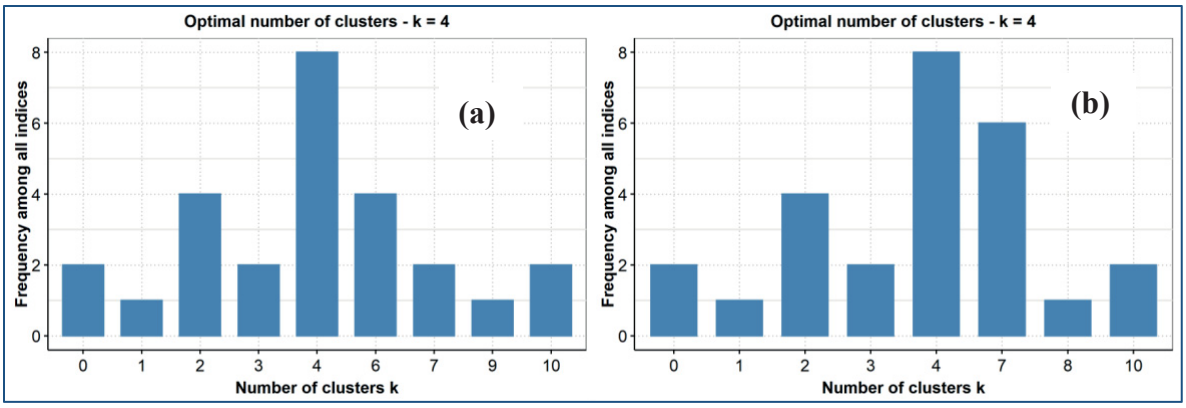

Figure 4.6 Diagramme de fréquence en fonction du nombre de clusters pour :
(a) le modèle 3 (technique du gain d'information) et
(b) le modèle 4 à droite (technique basée sur la corrélation)

Une fois le nombre optimal de clusters identifié pour chacun des modèles, la prochaine étape est de les intégrer dans l'algorithme k-moyennes afin de choisir le meilleur modèle (Halkidi, Batistakis et Vazirgiannis, 2002). En rappel, l'objectif du k-moyennes est de minimiser l'écart intra-cluster (*Within Cluster Sum of Square Error, WCSS*) et de maximiser l'écart inter-cluster. En d'autres termes la compacité à l'intérieur des clusters et la séparation d'avec les autres clusters sont désirées.

Le Tableau 4.2 présente les résultats obtenus en appliquant les quatre modèles considérant certaines métriques. La valeur optimale représente la valeur optimale désirable pour chaque métrique validant le choix du modèle. Les métriques sont sans unité et le modèle dont la valeur de métrique est meilleure est coloré en vert. Le modèle validant le plus de métriques est celui qui est retenu.

Tableau 4.2 Métriques de validation calculées pour chaque ensemble d'attributs sélectionné par chaque modèle

Métriques	**Modèle 1**	**Modèle 2**	**Modèle 3**	**Modèle 4**	**Valeur optimale**
	Tous les attributs	**ReliefF**	**Gain d'Information**	**Corrélation**	
WCSS	130.10	113.19	**95.31**	128	Minimale
Coefficient de silhouette	0.47	**0.65**	0.47	0.44	Maximale
Calinski – Harabasz(CH)	194	**234**	194	173	Maximale
Indice David Boulder (*DB*)	0.8	**0.7**	**0.7**	**0.7**	Minimale
Indice *Dunn*	0.06	0.07	0.07	**0.08**	Maximale
Indice *I*	**0.36**	0.31	0.28	0.32	Maximale
R – Squared	**0.82**	0.76	0.56	0.72	Maximale

Métriques	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Valeur optimale
	Tous les attributs	ReliefF	Gain d'Information	Corrélation	
S_D*bw*	0.09	**0.05**	0.07	0.09	Minimale
Indice *SD*	3.35	2.98	2.74	**2.69**	Minimale
Pearson gamma (Γ)	0.49	**0.55**	0.54	0.54	Maximale

Le modèle 2 (technique *ReliefF*) est retenu pour le regroupement des différents circuits. Il présente le plus grand nombre d'indices qui respecte les conditions de validation de la valeur optimale. Il présente la valeur optimale du coefficient de Silhouette.

Le *coefficient de Silhouette* mesure la qualité de la mise en clusters des attributs et estime la distance moyenne entre les clusters en mesurant la similarité d'un objet avec son propre cluster par rapport aux autres clusters. Une valeur de **0.65** indique que l'objet est bien apparié à son propre groupe et mal apparié aux autres clusters.

L'indice *Calinski-Harabasz (CH)* évalue la validité des clusters sur la base de la somme moyenne des carrés entre les clusters et à l'intérieur des clusters. Le modèle 2 présente la valeur maximale de *CH*.

Une valeur minimale de l'indice *David Boulder (DB)* montre que les clusters sont distincts les uns des autres, et permet donc d'obtenir le meilleur partitionnement.

L'indice de validation *S_Dbw* utilise la densité pour mesurer la séparation inter-cluster. La valeur optimale est la plus petite, soit celle de la technique *ReliefF,* qui est de **0.05**.

L'indice *Pearson gamma*, encore appelé *statistique modifiée d'Hubert Γ* (Γ) évalue la différence entre les clusters en comptant les désaccords des paires d'objets de données dans

deux partitions. Elle est comprise entre 0 et 1. La valeur optimale relevée pour les quatre méthodes est de **0.65** (Xiong et Li, 2013).

À l'issue des deux méthodes de sélection d'attributs à savoir les approches par filtrage et les approches embarquées, le modèle 2 basée sur la technique *ReliefF* avec un nombre optimal de trois clusters est choisi pour le regroupement des données de monitoring des 155 circuits sous analyse, selon la valeur de SNR. L'utilisation de la technique embarquée tout en permettant de comparer les différents ensembles de attributs entre eux, a aussi permis d'effectuer la validation des clusters en utilisant les métriques de validation interne dans le Tableau 4.2.

4.3.6 Application de l'algorithme k-moyennes

L'application de k-moyennes utilisant l'algorithme *Hartingan - Wong* est utilisée pour répartir les 155 circuits dans des clusters selon leur similarité. Cette méthode permet d'améliorer les performances du k-moyennes en se basant sur le concept de la densité du plus proche voisin pour le choix des centroïdes initiaux, contrairement au choix aléatoire qui se fait dans le k-moyennes de base (voir paragraphe 4.2.2). L'objectif étant de maximiser l'écart inter-clusters.

4.3.7 Résultats

La Figure 4.7 présente les trois clusters et les circuits contenus dans chacun d'eux. Les clusters sont représentés sur un plan, c'est-à-dire selon deux dimensions (*Dim1, Dim2*), en utilisant la fonction *fviz_cluster ()* dans *R*.
Le plan est celui d'une analyse en composante principale (ACP) qui est réalisée en amont de la représentation.

L'ACP ou Principal Component Analysis (PCA) en anglais, permet d'analyser et de visualiser un ensemble de données contenant des individus décrits par plusieurs variables quantitatives en perdant le moins possible d'information.

Les trois attributs retenus (*Mean, Diff_min* et *Max*) expliquent 99 % de la variance dans les données soit la somme des deux dimensions 1 et 2 (*Dim1[75.7%], Dim2[23.5%]*). Les valeurs moyennes de chaque attribut dans chaque cluster sont également présentées à la Figure 4.7.

D'une manière générale, les différents clusters sont bien séparés les uns des autres.

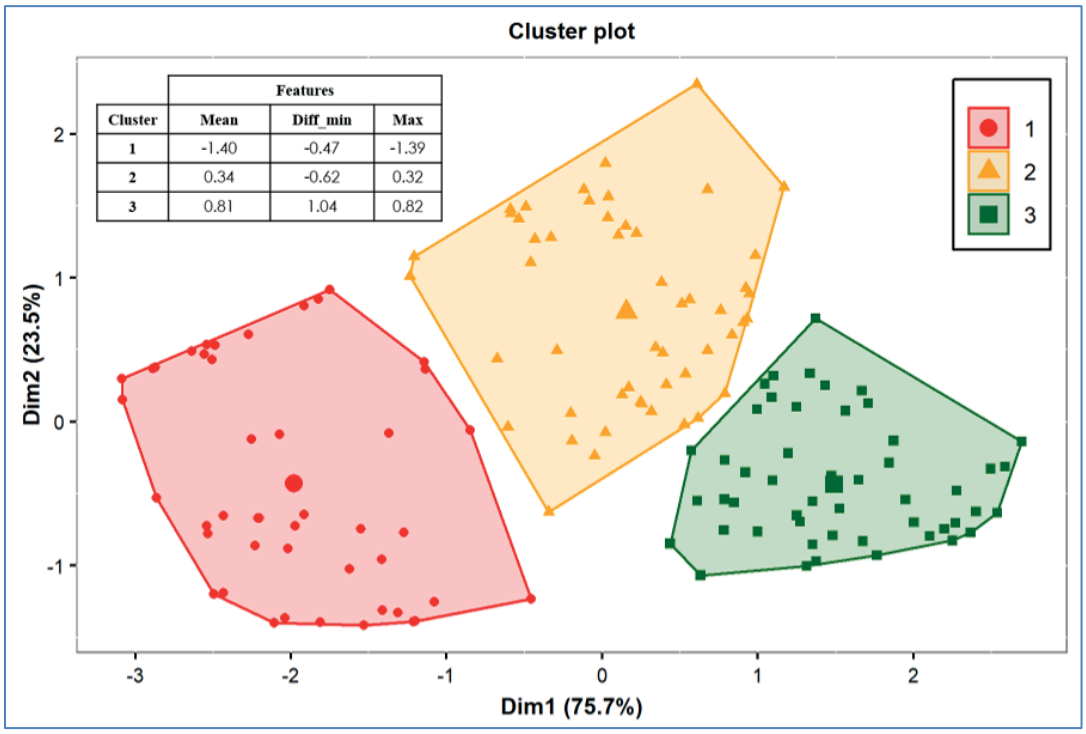

Figure 4.7 Graphique de la distribution des 155 circuits dans les trois clusters

Graphe du coefficient de silhouette

Une autre manière visuelle d'évaluer les clusters est la représentation graphique du coefficient de silhouette. Le coefficient de silhouette est une mesure de la compacité et de la séparation des clusters.

Le coefficient de silhouette est plus élevé pour les clusters compacts qui sont bien séparés et plus faible pour les clusters qui se chevauchent. Il varie de -1 à +1, et plus la valeur est élevée,

mieux c'est. Le coefficient de silhouette est calculé par instance et est déterminé par l'équation suivante (Dangeti, 2017) :

$$Si = \frac{b \times a}{\max(a, b)} \tag{4.11}$$

La variable a est la distance moyenne entre les instances du cluster, b est la distance moyenne entre l'instance et les instances du groupe suivant le plus proche.

La Figure 4.8 présente le graphique de silhouette, la taille (nombre d'objets par cluster) ainsi que leur coefficient de silhouette. Pour chaque point représentant un circuit, le coefficient de silhouette est la différence entre la distance moyenne avec les points du même cluster que lui (cohésion) et la distance moyenne avec les points des autres groupes voisins (séparation). Si cette différence est positive, le point est en moyenne plus proche de son groupe que du groupe voisin : il est donc bien classé. Par contre, si cette différence est négative, le point est en moyenne plus proche du groupe voisin que du sien : il est donc mal classé (Rousseeuw, 1987). C'est le cas pour certains circuits du cluster 1 et 2.

Le modèle 2 basé sur la technique ReliefF a été considéré dans le paragraphe 4.3.5 précédent comme étant la combinaison optimale des attributs avec la valeur moyenne de coefficient de silhouette la plus élevée, soit 0.65.

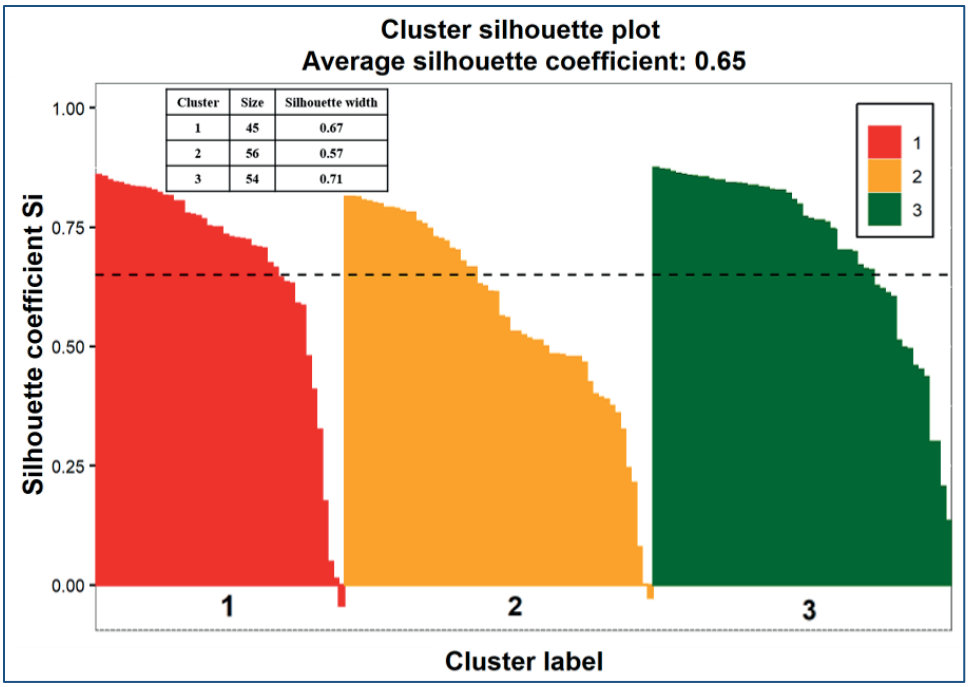

Figure 4.8 Graphe du coefficient de silhouette des 155 circuits

Graphe de coordonnées parallèles

Afin de comparer la caractéristique de plusieurs observations individuelles (*circuits*) sur un ensemble de variables numériques (*Attributs*), une méthode graphique nommée *graphe parallèle* ou *graphe de coordonnées parallèles* (*parallel coordinates plot*) est utilisée. Chaque barre verticale représente une variable dont la couleur représente son cluster d'appartenance. Comme les trois variables ont des plages très différentes, ce graphique a été réalisé avec des valeurs standardisées, où chaque variable a été normalisée pour avoir une moyenne nulle et une variance unitaire. Chaque observation (circuit) consiste en des mesures sur trois variables, et chaque mesure est représentée comme la hauteur à laquelle la ligne correspondante croise chaque axe de coordonnées (Jonietz et al., 2018; Schonlau, 2004).

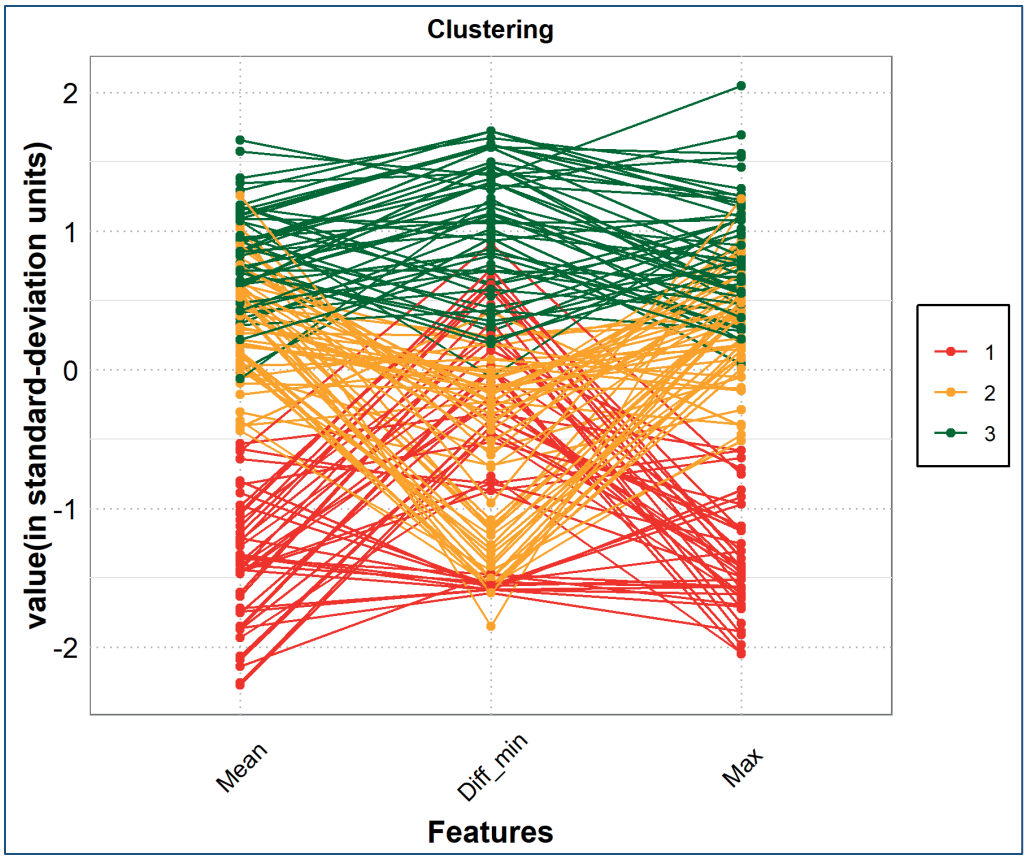

Figure 4.9 Graphe de coordonnées parallèles des 155 circuits

- Les circuits du cluster 1 ont les valeurs moyennes et maximales les plus petites. Par contre la plupart des circuits ont une valeur minimale en bas du seuil de 8.53 dB indiquant des valeurs de *Diff_min* négatives.
- Les circuits dans le cluster 2 possèdent des valeurs moyennes et maximales plus élevées que dans le cluster 1. La majorité des circuits dans ce cluster ont des valeurs de *Diff_min* négatives témoignant du fait que la valeur du SNR de ces circuits va en deçà du seuil de 8.53 dB.
- Les circuits dans le cluster 3, contrairement aux deux premiers, a des valeurs de *Diff_min* positives d'où le fait que la majorité des circuits ont une valeur de SNR loin du seuil fixé. Ce cluster possède également les circuits ayant les valeurs moyennes et maximales les plus élevées.

Afin de mieux comprendre le regroupement des circuits, la représentation temporelle ainsi que la localisation géographique de chacun des 155 circuits par route et selon leur cluster d'appartenance est présentée dans les Figure 4.10, Figure 4.11 et Figure 4.12.

- Le **cluster 1** (Figure 4.10) regroupe les circuits ayant une faible marge, une grande fluctuation dans les valeurs du SNR et présentant des erreurs et des anomalies.

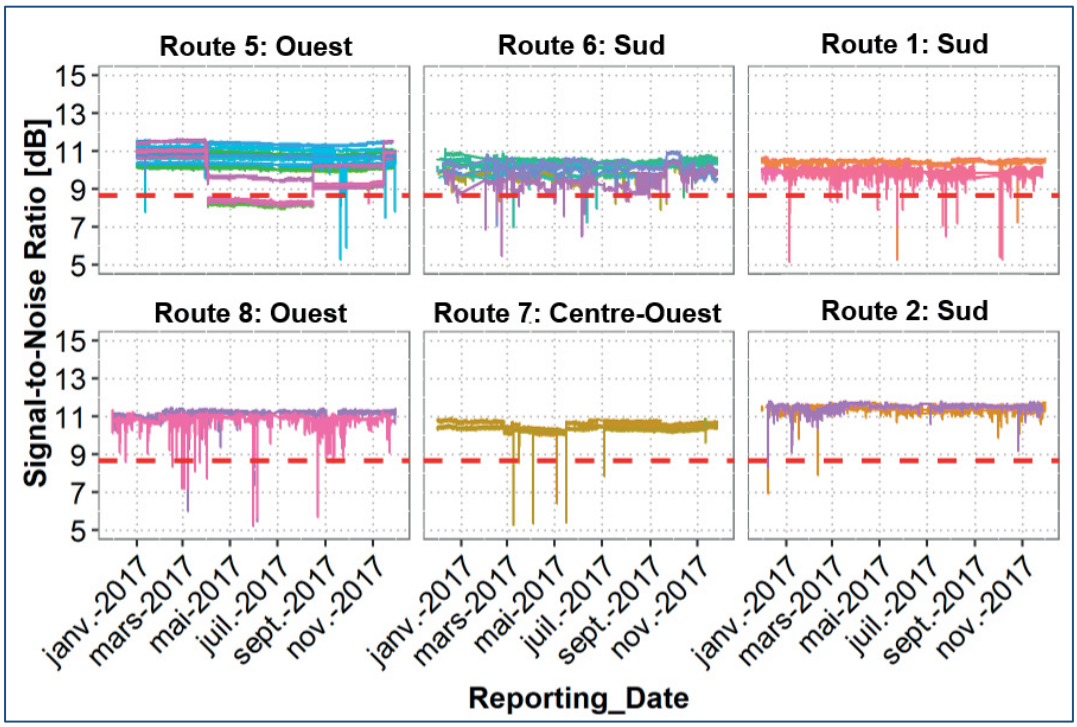

Figure 4.10 Évolution temporelle du SNR des circuits appartenant au cluster 1

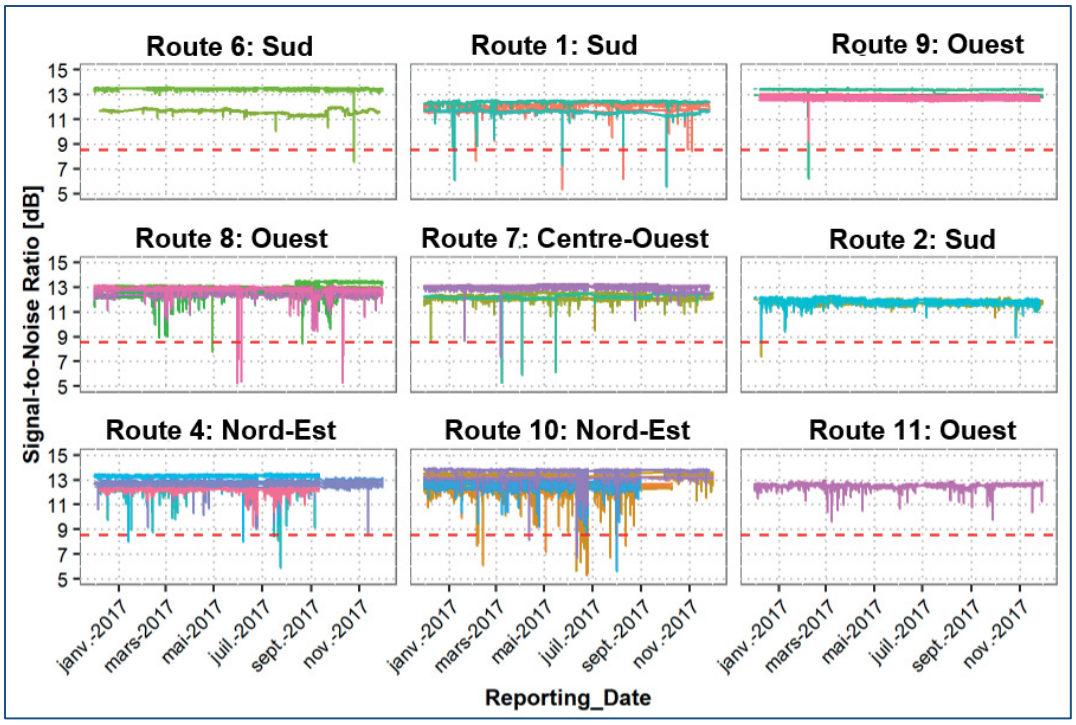

Figure 4.11 Évolution temporelle du SNR des circuits appartenant au cluster 2

- Les circuits du **cluster 2** à la Figure 4.11 présentent de grandes valeurs maximales du SNR et une certaine dynamicité dans leur comportement. Quelques circuits dans ce cluster présentent une seule erreur pendant la période de monitoring de 12 mois.

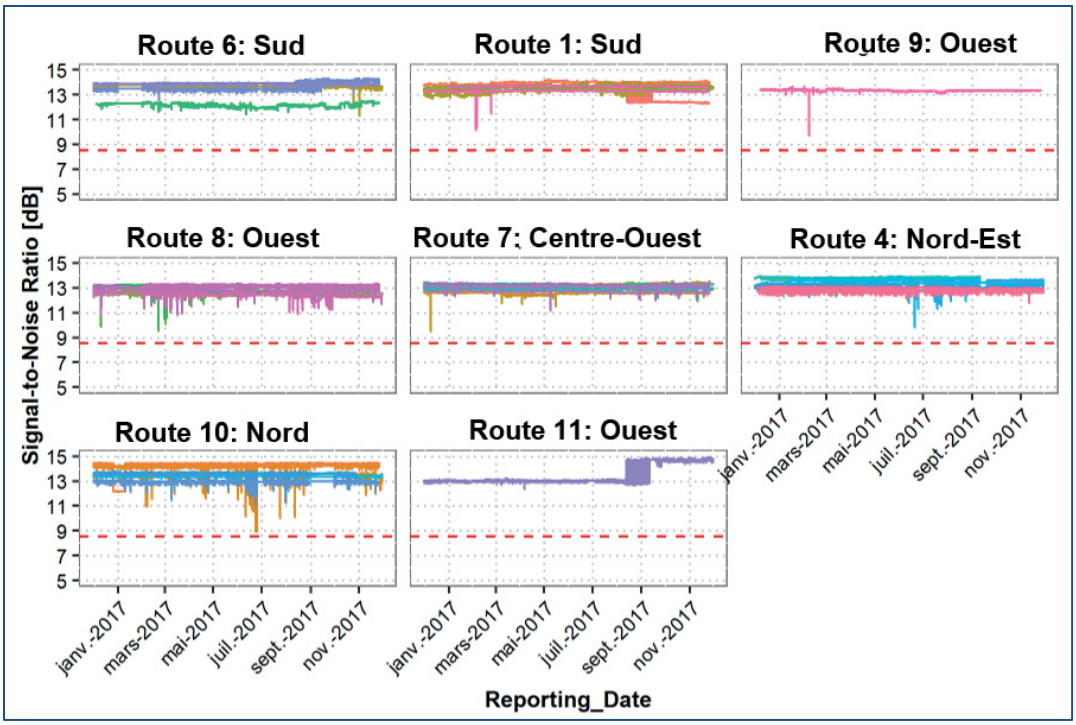

Figure 4.12 Évolution temporelle du SNR des circuits appartenant au cluster 3

- Les circuits appartenant au **cluster 3** à la Figure 4.12 sont les circuits ayant une grande marge et une faible activité du SNR.

Les circuits de la route 10 dans le **cluster 3** semblent avoir les mêmes caractéristiques que les circuits du **cluster 2** et le circuit de la route 11 dans le **cluster 2** présente le même comportement que les autres circuits du **cluster 3**.

4.3.8 Analyse des comportements sous-jacents dans les circuits

L'idée qui prévaut dans ce paragraphe est de déterminer s'il est possible à partir des trois groupes de circuits d'observer des comportements sous-jacents. Une représentation de la topologie physique des routes montrant les sites extrêmes et intermédiaires ainsi que la distance inter-sites est effectuée ci-dessous. Chaque circuit est coloré selon la couleur de son cluster d'appartenance.

Une grande majorité (2/3) des circuits du **cluster 1** appartiennent à deux routes (l'une aérienne et l'autre enfouie) considérées dynamiques par l'opérateur dont les caractéristiques sont présentées dans le Tableau 4.3. Une observation de ces circuits sur la carte, montrent qu'ils sont situés dans la zone des montagnes Rocheuses (*Rocky Mountains*) qui s'étendent sur plus de 3000 km depuis le Nouveau-Mexique jusqu'en Colombie-Britannique. La situation géographique semble être le facteur commun entre ces deux routes. La longueur de ces 30 circuits varie de 300 km à 900 km environ. Les autres circuits appartenant à ce même cluster 1 mais situé dans deux autres routes sont de type régional (distance inférieure à 1000 km), et longue portée (distance supérieure à 1000 km) (Figure 4.13).

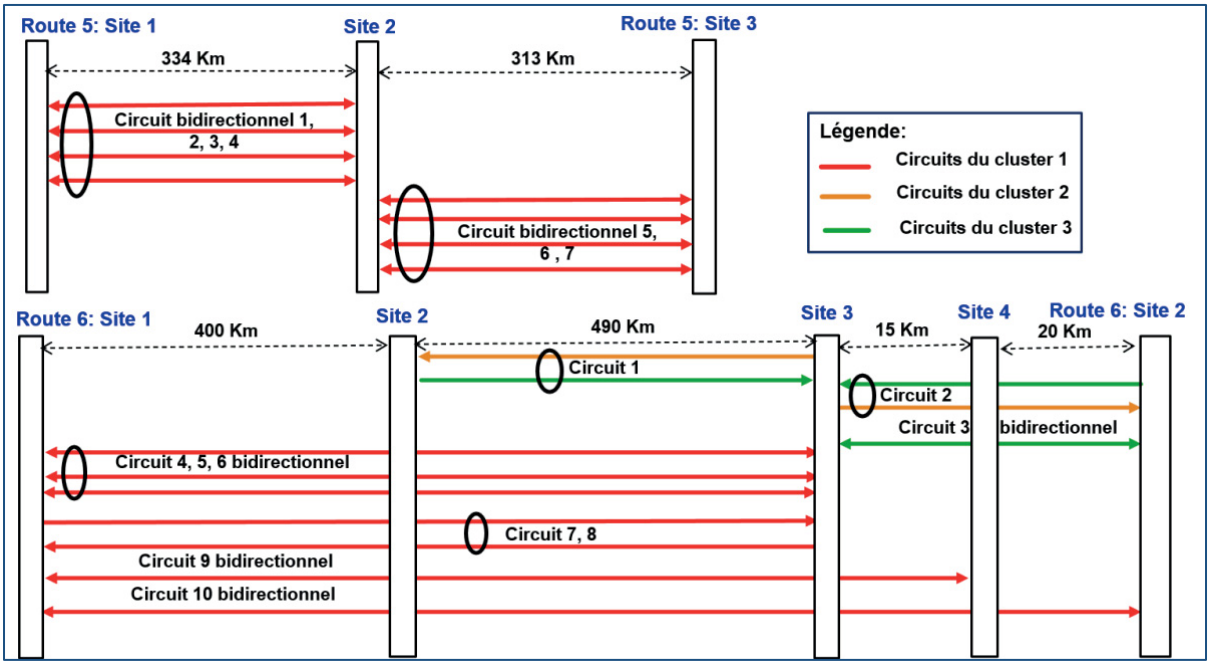

Figure 4.13 Synoptique des routes 5 et 6

La route 6 est considérée dynamique par l'opérateur. Les circuits appartenant au **cluster 1** sont l'ensemble des circuits dont la marge est faible et qui présente le plus d'anomalies. C'est le cas des circuits de la route 5. En observant la représentation temporelle du SNR des circuits de la route 5, il semble qu'un certain nombre de changements étaient en cours dans certains circuits

pendant la période d'observation. Cela doit être confirmé pour expliquer le nombre important de changements survenus sur cette liaison.

La route 6 montre de grandes fenêtres (environ 1 mois) de données manquantes dans la puissance au transmetteur, récepteur et dans le SNR.

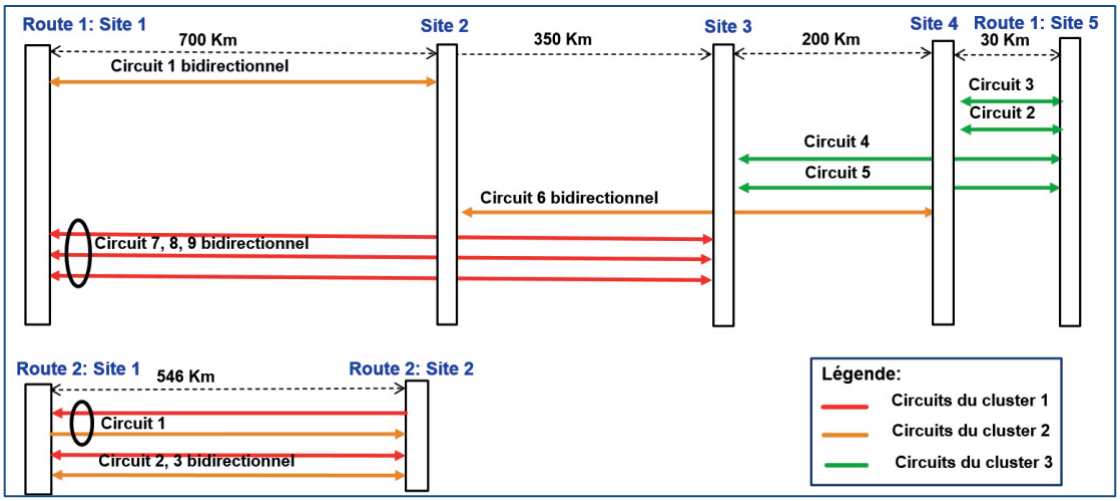

Figure 4.14 Synoptique des routes 1 et 2

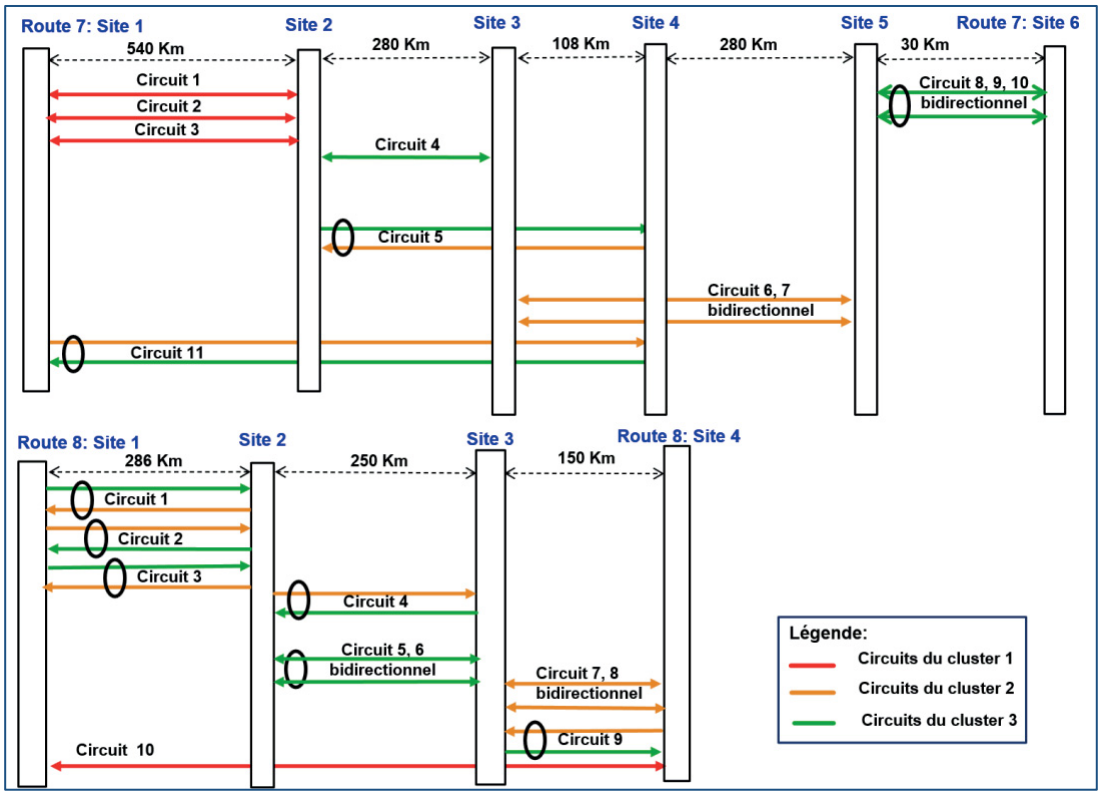

Figure 4.15 Synoptique des routes 7 et 8

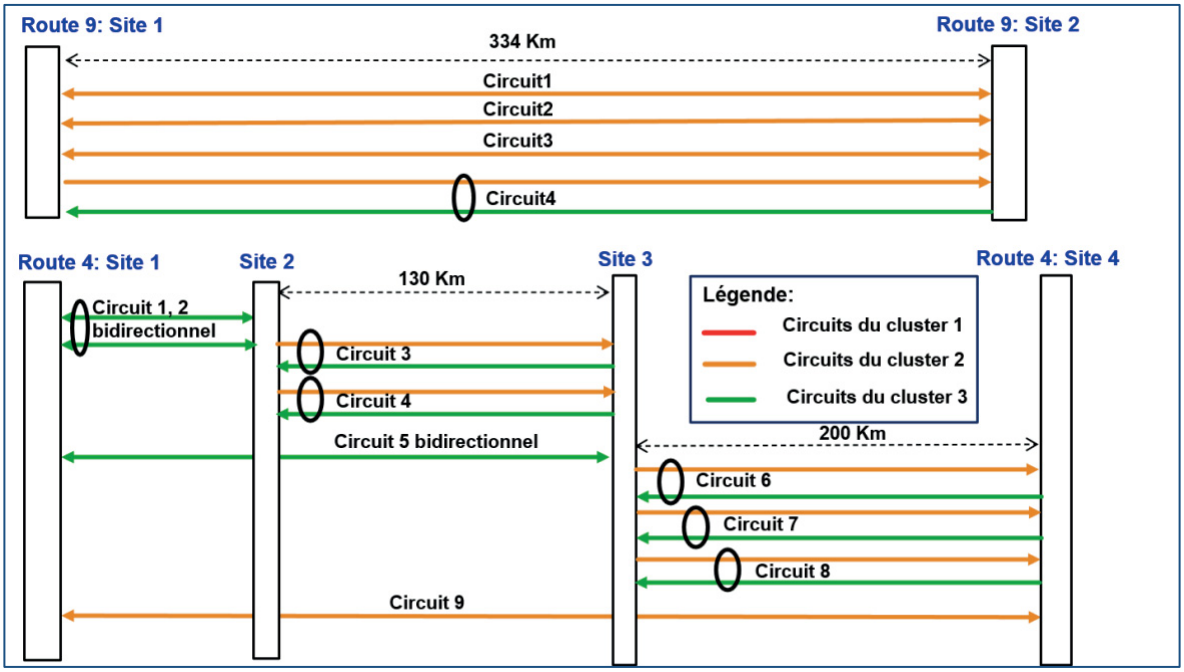

Figure 4.16 Synoptique des routes 9 et 4

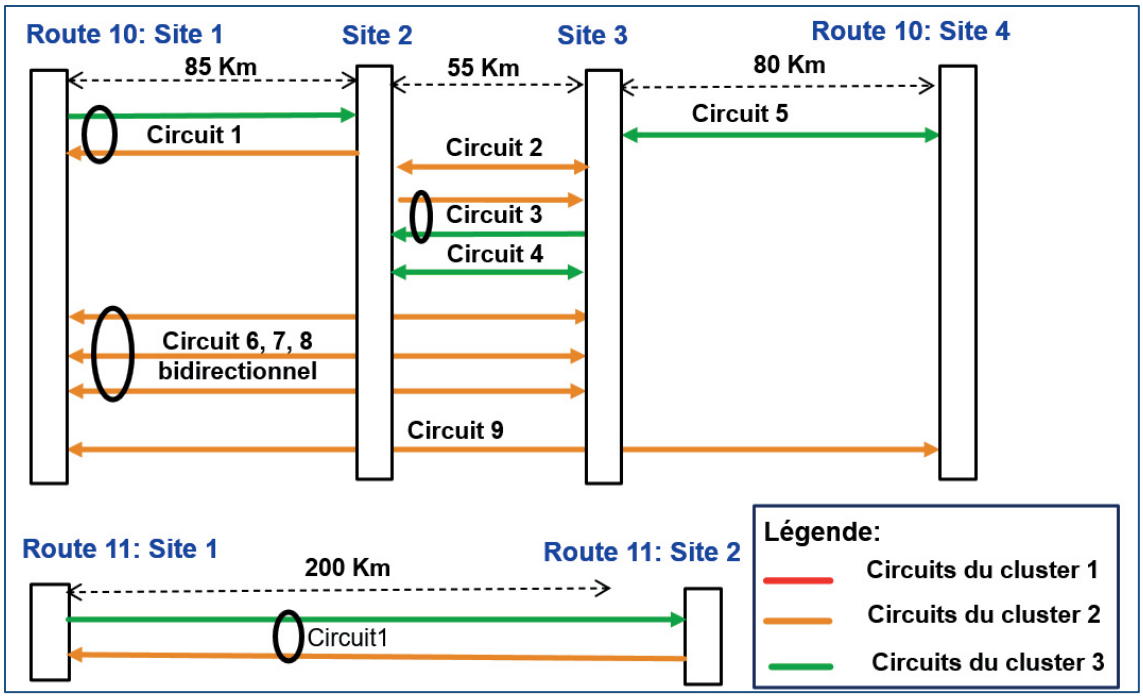

Figure 4.17 Synoptique des routes 10 et 11

Les circuits dans le **cluster 3** sont de type métropolitain c'est-à-dire que leur distance est inférieure à 300 km à l'exception de deux circuits bidirectionnels mesurant 400 km et 650 km appartenant à une même route située dans le Centre-Ouest du pays. Les circuits métropolitains dont la distance est inférieure à 300 km sont situés dans les **clusters 2** et **3**.

Les routes 4 et 10, situées dans le Nord-Est de type métropolitaines sont constituées principalement des circuits du **cluster 2** et **3**. Les fibres optiques les constituant sont de type *LS*. L'évolution temporelle du SNR à la Figure 4.11 montre de grandes activités du SNR.

4.3.9 Récapitulatif

Le Tableau 4.3 dresse un récapitulatif des caractéristiques des routes : la longueur bout en bout, la longueur par tronçon intermédiaire ainsi que les circuits bidirectionnels notés (**x 2**), le cluster d'appartenance, le type de déploiement et de fibres optiques, ainsi que la zone géographique.

Tableau 4.3 Récapitulatif des caractéristiques des circuits par route ainsi que leur cluster d'appartenance

Longueur totale de la route (km)	**Longueur par tronçon, km Et (nombre de circuits)**	**Cluster**	**Type de déploiement (Fibre utilisée)**	**Zone géographique**
Route 5 (16 circuits)				
590	334 (4 circuits x 2)	C1	Enfoui (DS)	Ouest (Montagnes Rocheuses)
	313 (4 circuits x 2)	C1		
Route 6 (18 circuits)				
1000	800 (6 x 2)	C1	Aérien (LS / NDSF)	Sud
	490 (1 x 2)	C2 + C3		
	40 (2 x 2)	C3		
Route 1 (16 circuits)				
1400	700 (1 x 2)	C2	Non spécifié (NDSF, LEAF)	Côte-Est
	1050 (3 x 2)	C1		
	500 (1 x 2)	C2		
	230 (2 x 2)	C3		
	30 (1 x 2)	C3		
Route 8 (20 circuits)				
600	600 (1 x 2)	C1	Enfoui (Mixed fiber)	Côte-Ouest
	286 (3 x 2)	C2 + C3		
	250 (1 x 2)	C2 + C3		
	250 (2 x 2)	C3		
	150 (2 x 2)	C2		
	150 (1 x 2)	C2 + C3		
Route 7 (16 circuits)				
1200	540 (3 x 2)	C1	Enfoui (NDSF)	Centre-Ouest
	650 (1 x 2)	C2 + C3		
	280 (1 x 2)	C3		
	400 (1 x 2)	C2 + C3		
	400 (1 x 2)	C2		
	30 (3 x 2)	C3		
Route 2 (6 circuits)				
545	545 (1 x 2)	C1	Aérien (*Non spécifié*)	Sud
	545 (2 x 2)	C2		

Tableau 4.3 Récapitulatif des caractéristiques des circuits par route ainsi que leur cluster d'appartenance (suite)

Longueur totale de la route (km)	**Longueur par tronçon, km Et (nombre de circuits)**	**Cluster**	**Type de déploiement (Fibre utilisée)**	**Zone géographique**
Route 9 (8 circuits)				
350	350 (3 x 2)	C2	Aérien (LS & True Wave)	Ouest
	350 (1 x 2)	C2 + C3		
Route 4 (18 circuits)				
365	10 (2 x 2)	C2	Enfoui, segment subway (LS)	Côte-Est
	130 (2 x 2)	C2 + C3		
	140 (1 x 2)	C3		
	200 (3 x 2)	C2 + C3		
	365 (1 x 2)	C2		
Route 10 (16 circuits)				
220	85 (1 x 2)	C2 + C3	*Non spécifié* (LS)	Nord-Est
	55 (2 x 2)	C2 + C3		
	80 (1 x 2)	C3		
	140 (3 x 2)	C2		
	220 (1 x 2)	C2		
Route 11 (2 circuits)				
200	200 (1 x 2)	C3	Enfoui (TW Classic)	Ouest
	200 (1 x 2)	C2		

4.3.10 Conclusion

Les méthodes non supervisées trouvent une structure dans les données, souvent en prélude à une modélisation prédictive. L'objectif du regroupement est de découvrir ou de faire ressortir les similitudes entre les sous-ensembles de vos données. Au titre des séries temporelles deux méthodes de regroupement se distinguent : le regroupement total (''*whole clustering*'') et le regroupement par sous-séquence (''*subsequence clustering*'').

Le regroupement total est une méthode effectuée sur plusieurs séries temporelles afin de regrouper celles qui présentent des similitudes. Le regroupement par sous-séquence quant à

elle, est basée sur l'extraction de fenêtre coulissante sur une série temporelle et vise à trouver des similitudes et des différences entre les différentes fenêtres temporelles. Vu le nombre de séries temporelles sous étude, le regroupement total a été appliqué. La première étape a consisté à extraire les caractéristiques statistiques de chaque série, les transformer et à les utiliser ensuite pour catégoriser les circuits en utilisant l'algorithme k-moyennes.

Auparavant, les caractéristiques qui permettent un meilleur regroupement ont été choisi au travers de deux méthodes : les méthodes de filtrage et les méthodes embarquées.

Trois clusters ont été trouvés dans les données de 155 circuits. Un premier cluster composé de 45 circuits très dynamiques avec une faible marge, un second cluster avec 56 circuits possédant une grande marge et de nature dynamiques. Ces deux premiers clusters présentent des anomalies et des erreurs dans le rapport signal sur bruit optique. Le cluster 1 est constitué des circuits présentant le plus d'anomalies et de secondes erronées.

Le troisième cluster de 54 circuits se compose des circuits calmes et des circuits ayant une activité modérée dans les fluctuations du SNR.

CHAPITRE 5

ANALYSE DES LIENS OPTIQUES

5.1 Introduction

Parmi les paramètres monitorés dans un réseau optique, la puissance optique est d'une importance capitale pour attester de la qualité du signal. Cette puissance peut être affectée par plusieurs facteurs dont la majorité est liée à l'état du lien optique. Le lien optique se constitue principalement des éléments d'amplification, de filtrage, de redirection du signal et de la fibre optique déployée par segment communément appelé *span optique*. Chacun de ces éléments peuvent donc contribuer à la dégradation du signal. Le réseau sous étude comporte environ 900 spans optiques toutes routes confondues en considérant la longueur de chaque circuit optique et une longueur moyenne de span de 50 km.

L'objectif du monitoring de performance à travers l'analyse des données historiques, est de fournir à l'opérateur réseau une vue d'ensemble des paramètres de performance, qui lui permettront entre autres de détecter certaines tendances de dégradation afin de prévoir des opérations de maintenance préventive ou de remplacement.

Ce chapitre a fait l'objet d'un article qui a été publié dans le journal JLT (Journal of ligthwave technology) sous le titre de ''*Trends in Optical Span Loss Detected Using the Time Series Decomposition Method*''.

5.2 Détection de tendance dans les spans de fibre optique

5.2.1 Introduction

Les câbles à fibres optiques des réseaux longue distance, métropolitains et d'accès sont composés de spans de fibres de longueurs et de nombres différents. La perte optique de span doit toujours être gérée avec soin pour une transmission de données sans erreur (Xu et al., 2015). Plusieurs générations de câbles à fibre optique existent dans les réseaux actuels.

La qualité de transmission des signaux optiques transmis dans les liaisons par fibre optique change au fil du temps en raison des interactions linéaires et non linéaires entre les canaux, des opérations de maintenance dans l'installation extérieure et des facteurs environnementaux, ainsi que la dérive du calibrage des composants à long terme et les effets du vieillissement (Willner, Pan et Yu, 2008).

Les opérateurs de réseaux monitorent continuellement les performances optiques de leurs réseaux pour une transmission fiable des données. Les premiers systèmes de contrôle des réseaux permettent un monitoring en temps réel des paramètres du système et des liaisons, tels que les niveaux de puissance optique aux sites terminaux et d'amplification, ainsi que le taux d'erreur binaire (*Bit Error Rate, BER*) au récepteur. Ces paramètres de performance sont stockés dans les bases de données, et sont utilisées pour l'analyse des causes profondes en cas de défaillance de composants ou du lien. Avec le vieillissement des installations de fibre optique, les opérateurs de réseau cherchent à quantifier les effets à long terme du vieillissement du réseau à fibres optiques.

La caractérisation des performances à long terme des câbles à fibres optiques déployés depuis plus de 30 ans reste un problème ouvert. Beaucoup de travaux ont été consacrés à ce problème jusqu'à présent, mais le processus de vieillissement de la fibre optique n'est pas encore totalement compris (Poulain et Severin, 2007). Divers modèles théoriques ont été appliqués pour la caractérisation mécanique des fibres optiques. Le plus courant est basé sur les statistiques de Weibull. La loi de Weibull exprime la probabilité de défaillance d'une fibre optique soumise à une contrainte appliquée (Matthewson, 1993).

D'autres techniques physiques ont été développées, telles que la mesure de la résistance à la traction ou de l'élongation à la rupture du matériau d'isolation ; toutefois, ces méthodes sont intrinsèquement destructrices et nécessitent un échantillon du câble vieilli pour les essais (Severin et al., 2005). Un câble et une méthode de monitoring de son vieillissement ont été proposés, mais cette approche ne peut pas être appliquée aux réseaux de fibres déjà déployés (Gazdzinski, 1999).

Plus récemment, l'utilisation d'un réflectomètre optique dans le domaine temporel (*Optical Time Domain Reflectometer, OTDR*) et des mesures de puissance effectuées sur un câble de 24 fibres optiques déployé depuis 16 ans sur une route de 34 km, ont montré une augmentation moyenne de 20% de la perte moyenne d'atténuation. Cela a été attribué à la présence d'eau dans la gaine protectrice de la fibre optique dans ce cas spécifique (Maslo et al., 2020).

Dans ce chapitre, une méthode de décomposition des séries temporelles pour la détection des tendances à long terme de la perte de span des fibres optiques est proposée. Cette méthode est testée avec des données réelles recueillies dans une liaison optique bidirectionnelle à 4 span d'un réseau de production, en utilisant trois approches statistiques complémentaires, largement utilisées dans le domaine de la météorologie, des finances et du monitoring de la qualité de l'air et de l'eau : la régression à pondération locale (*Locally weighted regression, Loess*) (STL), le test de Mann-Kendall, et le test de *Sen's slope* (Aguilera et al., 2015).

Le chapitre s'organise comme suit : dans la section II, les données des séries temporelles et la méthode de décomposition de série temporelle sont décrites. En outre, les tests statistiques et la méthode de régression linéaire sont également couvertes. Les résultats sont présentés dans la section III et les conclusions sont tirées dans la section IV.

5.3 Méthodologie

5.3.1 Système et prétraitement des données

Les données de monitoring de performances (PM) utilisées sont collectées à des intervalles de 15 minutes sur une période de 12 mois sur une modulation en quadrature de phase multiplexée en polarisation (*Polarization Multiplexed Quadrature Phase Shift Keying, PM-QPSK*) à 100 Gbit/s déployé dans une liaison optique amplifiée de 205 km (Figure 5.1).

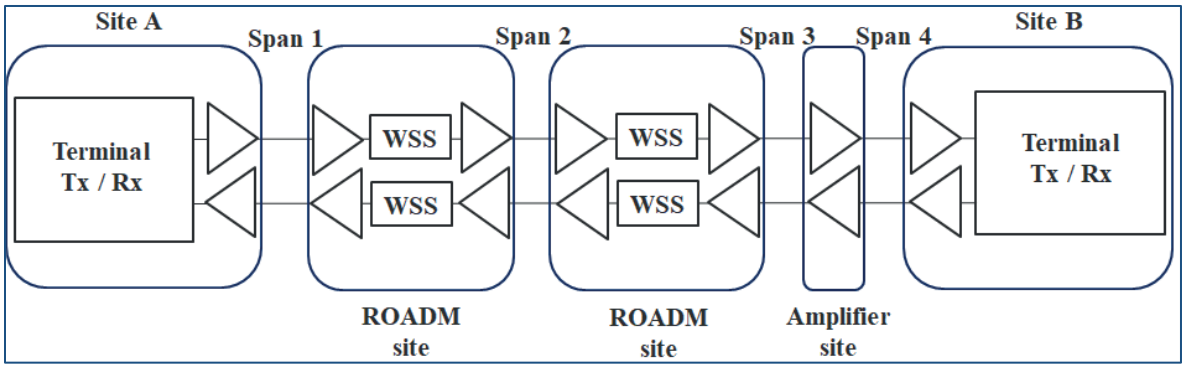

Figure 5.1 Liaison optique de 205 Km à 4 spans

La Figure 5.2 décrit la méthodologie suivie pour réaliser l'analyse. Dans un premier temps, les séries temporelles sont constituées de données de perte de span calculées à partir des puissances d'entrée et de sortie mesurés au niveau des amplificateurs à fibre dopée à l'erbium (*Erbium-Doped Fiber Amplifier, EDFA*). Étant donné que les pertes de span ne varient pas beaucoup en une journée, les valeurs ont été moyennées journalièrement.

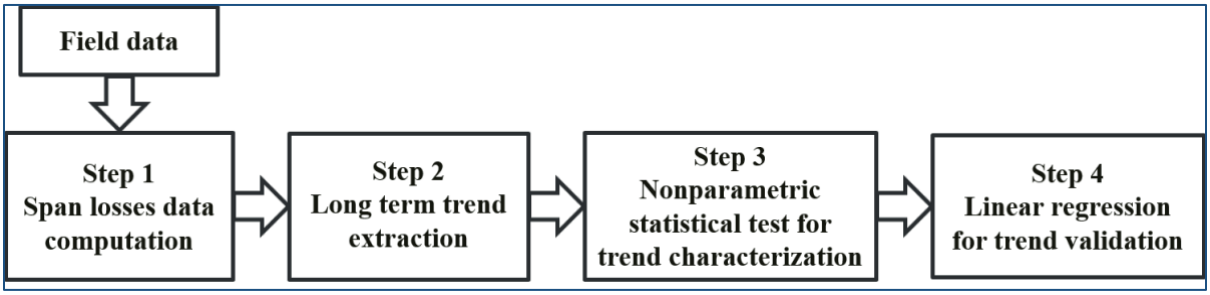

Figure 5.2 Différentes étapes de la méthodologie

La seconde étape de la méthodologie consiste à extraire la tendance à long terme de la série en effectuant une décomposition de la série temporelle en composantes saisonnière, tendance et bruit. Dans la troisième étape, les tests non paramétriques de Mann-Kendall et Sen's slope sont appliqués pour caractériser la composante tendance. Par la suite, une régression linéaire est appliquée pour la validation de la tendance.

Avant d'effectuer les différentes analyses, il est important de disposer d'un profil des données. Par conséquent, un aperçu des caractéristiques statistiques de base des données de perte de

span est présenté dans le Tableau 5.1. Aux fins de l'analyse, les spans 1 et 3, et les spans 2 et 4, ont été regroupés car elles présentent une perte médiane comparable : environ 13 dB pour les plages 1 et 3, environ 20 dB pour les plages 2 et 4.

Tableau 5.1 Caractéristiques statistiques des données de perte de span

	Span	**Médiane**	**Min**	**Max**	**IQR**	**CV**	γ_1
		(dB)				**(%)**	**-**
Direction A - B	1	13.5	13.3	13.6	0.1	0.5	- 0.6
	2	20.0	19.9	20.6	0.1	0.7	2.6
	3	13.1	13.0	13.2	0	0.3	- 0.3
	4	21.1	21	21.1	0	0.1	- 0.8
Direction B - A	1	13.7	13.6	13.8	0.1	0.3	-0.5
	2	20.3	20.1	20.7	0.1	0.5	2.7
	3	14.2	14.1	14.2	0.1	0.2	-0.7
	4	18.6	18.5	18.7	0	0.1	-1.5

Aucune variation significative de la perte de span n'a été observée au cours de la période de 12 mois, tel que confirmé par des valeurs de coefficients de variation (CV) inférieurs à 1 %. Les histogrammes de perte par span Figure 5.3 à Figure 5.5 sont relativement étroits, avec un écart - interquartile (IQR) inférieur à 0.2 dB.

Afin de vérifier si les différentes données relatives aux pertes de span suivent une distribution normale, le coefficient d'asymétrie (γ_1) est calculé. C'est une mesure de symétrie ou, plus précisément, du manque de symétrie dans une distribution. L'asymétrie pour une distribution normale est de zéro. Les spans 1, 3 et 4 sont étalés vers la gauche et le span 2 est étalé vers la droite.

La nature bimodale des distributions des données n'est pas due à des effets saisonniers, mais plutôt à des changements brusques : augmentation ou diminution de la perte de span pendant la période d'observation. L'ampleur de ces changements est de l'ordre de quelques dixièmes de

dB et résulte probablement d'actions de maintenance (telles que l'épissure des fibres optiques) sur les fibres.

Comme les données ne suivent pas une distribution normale, des méthodes de test statistique non paramétriques sont utilisées dans les étapes suivantes (Ali et al., 2019; Hipel et McLeod, 1994).

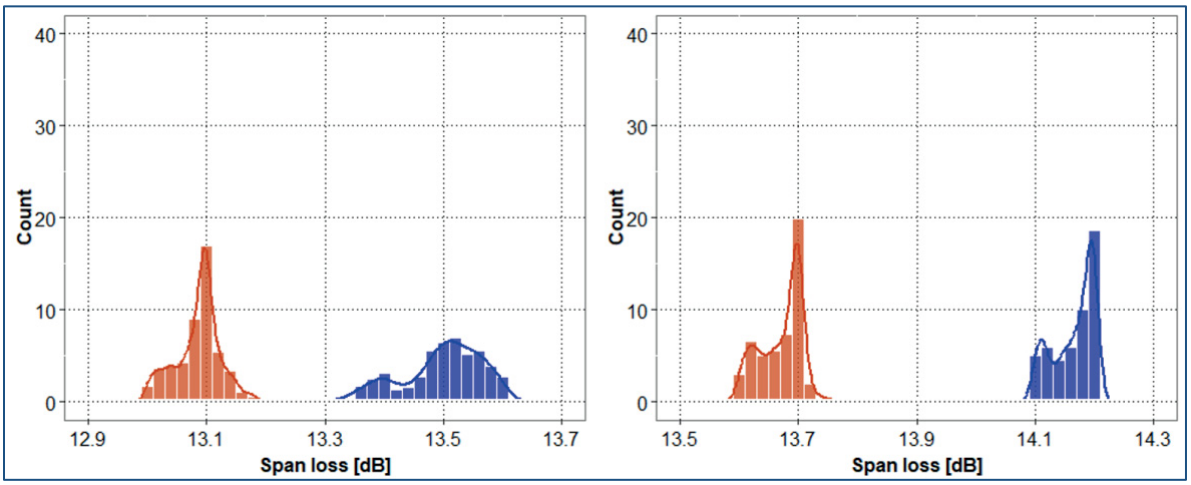

Figure 5.3 Représentation de l'histogramme du span 1 (bleu), span 3 (rouge)
Direction A – B (gauche), Direction B – A (droite)

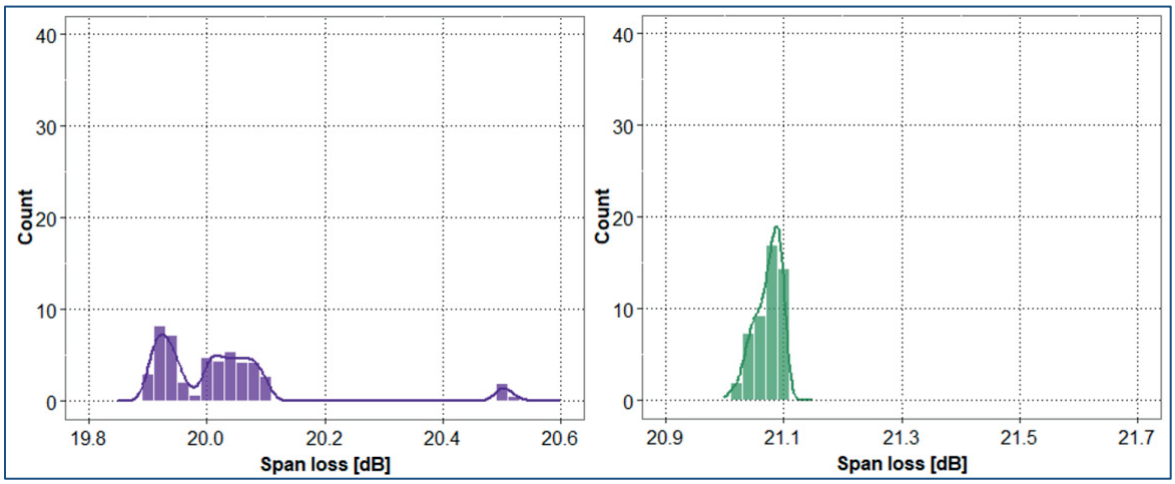

Figure 5.4 Représentation de l'histogramme du span 2 (mauve), span4 (vert) :
Direction A – B

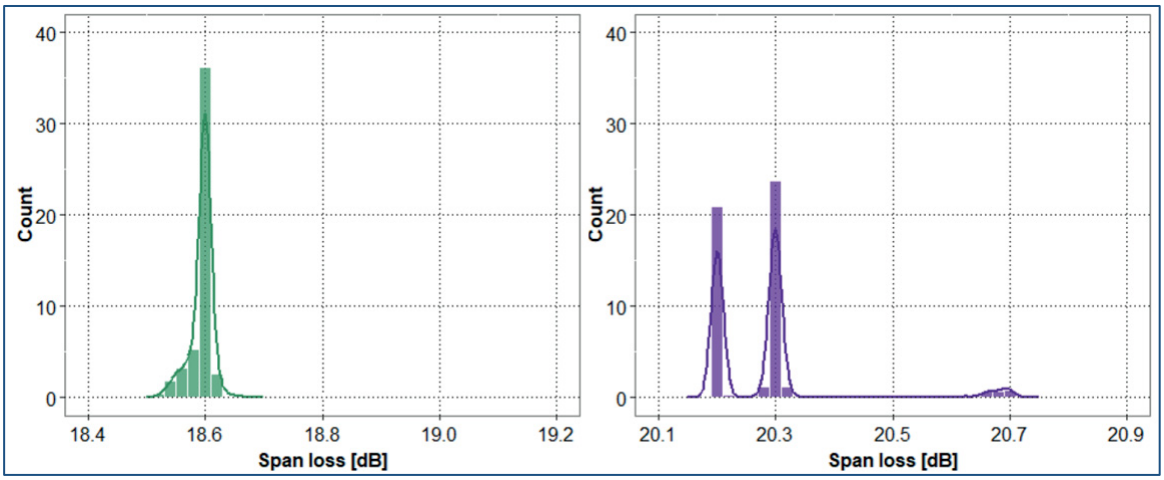

Figure 5.5 Représentation de l'histogramme du span 2 (mauve), span4 (vert) : Direction B – A

5.3.2 Méthode de décomposition de série temporelle

La deuxième étape de l'analyse consiste à utiliser la méthode de décomposition des séries temporelles pour effectuer la détection de la tendance à long terme. Cette méthode sépare une série en trois composantes sous-jacentes qui se combinent pour produire les données sous étude : la tendance à long terme, la saisonnalité (variation saisonnière) et un terme de bruit (reste). Le modèle structurel de base des séries temporelles est (Harvey et Peters, 1990):

$$Y_t = U_t + S_t + I_t \, , t = 1, \dots T \tag{5.1}$$

Où :

- U_t la composante tendance au temps *t* reflétant la progression à long terme de la série temporelle.
- S_t est la composante saisonnière au temps *t* qui reflète la saisonnalité et
- I_t est la composante irrégulière (ou "bruit") au temps *t*, qui représente les résidus ou le reste de la série temporelle après que les autres composantes aient été supprimées. Elle suit en théorie une distribution normale avec une moyenne nulle et une variance σ_I^2.

- T représente le nombre total de points de données.

Le processus générant la tendance peut être considéré comme l'approximation locale d'une tendance linéaire, c'est-à-dire :

$$U_t = U_{t-1} + \beta_{t-1} + \eta_t \tag{5.2}$$

Où U_{t-1} est la composante stochastique et β_{t-1} le paramètre de la pente générée au temps *t-1*. La composante η_t représente le terme de perturbation avec une moyenne nulle et une variance σ_η^2 (Harvey et Peters, 1990).

La procédure de décomposition est basée sur la méthode Loess (*locally weighted regression smoothing*) implémenté au travers de la librairie *STLplus* de l'environnement RStudio version 3.4.3 (Hafen, 2016). STL est l'acronyme de "*Seasonal and Trend decomposition using Loess*", une méthode polyvalente et robuste de décomposition des séries temporelles, et Loess est une méthode d'estimation des relations non linéaires (Cleveland et al., 1990). L'algorithme fonctionne en exécutant deux boucles :

- une boucle externe où les coefficients de robustesse sont calculés afin de minimiser l'effet des valeurs aberrantes.
- une boucle interne où la composante tendance et la composante saisonnière sont mises à jour de manière itérative en utilisant le lissage Loess.

L'algorithme peut être configuré pour fonctionner pendant un nombre fixe d'itérations pour chaque boucle, ou il peut être configuré pour fonctionner jusqu'à ce qu'un critère de convergence spécifique soit rempli. Le paramètre de lissage de la composante saisonnière est un paramètre important à définir. Il est représenté par le paramètre de fréquence qui est le nombre d'observations par cycle. Pour les données quotidiennes, Cleveland et al. (1990) recommande d'utiliser une fenêtre d'au moins 7 jours (semaine). Les valeurs manquantes dans la tendance et les composantes saisonnières sont traitées par lissage (Hafen, 2016).

5.3.3 Méthodes statistiques non paramétriques pour la quantification de la tendance

Deux tests statistiques non paramétriques sont appliqués sur les données pour l'estimation de la tendance, à savoir les tests de Mann-Kendall et de Sen's slope.

Test de Mann- Kendall

Le test de Mann-Kendall analyse les différences entre les points de données antérieurs et postérieurs. Il s'agit d'un test basé sur le classement de rang et ne nécessite pas le respect de la distribution normale, largement adopté dans l'analyse des séries temporelles pour détecter la présence de tendances monotones croissante ou décroissante. L'hypothèse nulle (H0) de ce test non paramétrique indique l'absence d'une tendance monotone, et l'hypothèse alternative (H1) indique l'existence d'une tendance positive, négative ou non nulle. La valeur de la statistique principale *S* du test de Mann-Kendall est calculée à l'aide des équations (5.3)-(5.5) (Balasubramanian et al., 2019).

$$S = \sum_{a-1}^{z-1} \sum_{b=a+1}^{z} sgn(x_b - x_a) \tag{5.3}$$

Et

$$sgn(x_b - x_a) = \begin{cases} 1 & if\ x_b - x_a > 0 \\ 0 & if\ x_b - x_a = 0 \\ -1 & if\ x_b - x_a < 0 \end{cases} \tag{5.4}$$

Où z est le nombre total de points de données, et x_a et x_b sont des points de données individuels avec *b* supérieur à *a*.

$$variance(S) = \frac{z(z-1)(2z+5) - \sum_t t(t-1)(2t+5)}{18} \tag{5.5}$$

Ici, *t* indique l'intervalle entre les points d'échantillonnage et $\sum_t$ la somme des *t*. Le test normal standard τ pour estimer la présence d'une tendance statistiquement significative est calculé en utilisant (Jaiswal, Samuel et Kadabgaon, 2018) :

$$\tau = \begin{cases} \dfrac{S-1}{\sqrt{variance(S)}} & if\ S > 0 \\ 0 & if\ S = 0 \\ \dfrac{S+1}{\sqrt{variance(S)}} & if\ S < 0 \end{cases} \tag{5.6}$$

Une valeur négative ou positive de τ correspond respectivement à une tendance décroissante ou croissante avec un niveau de signification α = 0.05 (c'est-à-dire à 95 % de niveau de confiance). Le test de Mann-Kendall suppose qu'une valeur peut toujours être déclaré inférieur, supérieur ou égal à une autre valeur ; que les données sont indépendantes ; et que la distribution des données reste constante dans les unités d'origine ou les unités transformées (Helsel et Hirsch, 2002).

En plus de ne pas avoir à se conformer à une distribution particulière, ce test est également résistant aux valeurs aberrantes et manquantes (Meals et al., 2011). Par conséquent, la méthode est appliquée sans transformation préalable des données.

Test de Sen's slope

Le test de Sen's slope ou estimateur de *Theil-Sen* est un test statistique largement utilisé pour les données non paramétriques afin d'estimer la valeur de la tendance détectée par le test de Mann-Kendall (Caloiero et al., 2017). Les valeurs positives et négatives de la pente de Sen indiquent respectivement des tendances à la hausse et à la baisse. Dans l'hypothèse d'une tendance linéaire, la pente de Sen's est la médiane de toutes les pentes :

$$\hat{Q} = \frac{y_j - y_k}{t_j - t_k}, \qquad j > k \tag{5.7}$$

entre les points de données y_j et y_k pendant l'intervalle de temps $t_j - t_k$ (Balasubramanian et al., 2019; Olofintoye, Adeyemo et Otieno, 2012). Ce qui correspond dans ce cas-ci à un changement de perte de span par unité de temps.

5.4 Validation de la tendance par régression linéaire

Une fois la tendance estimée, sa validation est testée, comme indiqué dans la Figure 5.2. Pour ce faire, la méthode de régression linéaire de *Theil-Sen* est utilisée. Cette méthode utilise la pente médiane de toutes les lignes par des paires de points d'échantillonnage bidimensionnels (y_j , y_k) pour la régression (Conover, 1999; Helsel et Hirsch, 2002). L'équation est obtenue comme suit :

$$y = \hat{\beta}_0 + \hat{\beta}_1 t \tag{5.8}$$

Le paramètre $\hat{\beta}_1$ est l'estimation non paramétrique de la pente. Il s'agit de la médiane de toutes les valeurs de *Q* dans l'équation 5.7 pendant la période d'observation. $\hat{\beta}_0$ représente l'intercept et *t* est le temps.

$$\hat{\beta}_0 = y_{MED} - \hat{\beta}_1 \times t_{MED} \tag{5.9}$$

Avec :

- y_{MED}= médiane de *n* points de données
- t_{MED} = médiane de *n* points de temps t_1 , t_2 , …, t_n

À partir des équations 5.8 et 5.9, la droite linéaire estimée *y* peut être réécrite comme :

$$y = y_{MED} + \hat{\beta}_1(t - t_{MED}) \tag{5.10}$$

5.5 Résultats et discussion

Les données brutes de perte par span, ainsi que les résultats de la décomposition des séries temporelle en trois composantes constitutives (saisonnière, tendance et bruit) sont représentées pour les deux directions dans les figures 6 et 9, pour les spans 1, 3 et, dans les figures 7 et 10, pour les spans 2, 4.

La composante saisonnière (**c**) est très faible (< 0,01 dB). Ceci est normal car il s'agit d'un lien enfoui. Ainsi, la composante saisonnière existe mais elle semble indiquer que la température n'a pas d'impact sur la variabilité de la liaison optique, puisqu'elle n'est pas directement exposée aux aléas de saisons.

Les anomalies (ou pics) et le bruit dans les courbes de perte de span brute se retrouvent dans la composante de décomposition résiduelle (**d**). La composante de tendance qui en résulte est une courbe lissée. L'asymétrie de la composante résiduelle (**d**) se situe entre -0,5 et 0,5, ce qui confirme que la composante résiduelle suit la distribution sigma-carrée décrit dans le paragraphe 5.3.2.

En dehors des patterns ou événements non systématiques dans les données, le résiduel pourrait fournir une indication de l'ajustement approprié des autres composantes.

Une corrélation élevée dans cette composante indique que certains patterns liés à d'autres composantes n’ont pas été bien filtrée en raison d'une estimation inexacte (Krispin, 2019) (voir le graphe de la composante résiduelle (**d**) du span 2 pour la direction B vers A, dans la Figure 5.10).

Les composantes de tendance pour les 8 spans de fibres optiques sont indiquées dans les Figure 5.8 et Figure 5.11. Une tendance légèrement à la hausse, à peine visible dans les données brutes, peut être observée un peu mieux dans la composante de tendance pour le span 1 dans la direction A-B. Cette tendance est également observée dans la direction opposée, comme

prévu. Elle semble refléter une dégradation de la fibre, dont l'évolution sera analysée dans la prochaine partie du chapitre. Plusieurs facteurs peuvent influencer la dégradation, à savoir, entre autres, l'âge, les conditions environnementales, etc.

La composante de tendance pour le span 2 dans les deux directions (Figure 5.7 et Figure 5.10 respectivement) montre des changements soudains dans la perte de span, ce qui peut introduire une certaine incertitude dans la détermination de la valeur de $Q_{médiane}$. Dans ce cas, l'extraction de la tendance a permis de capturer ces événements en forme d'escalier probablement causés par le ré-épissure de la fibre ou la reconfiguration du câble. Les composantes de tendance pour les spans 3 et 4 ne montrent pas de changements significatifs pendant la période d'observation. Malgré les différences mises en évidence dans chaque span, les tests statistiques ont été appliqués aux quatre spans.

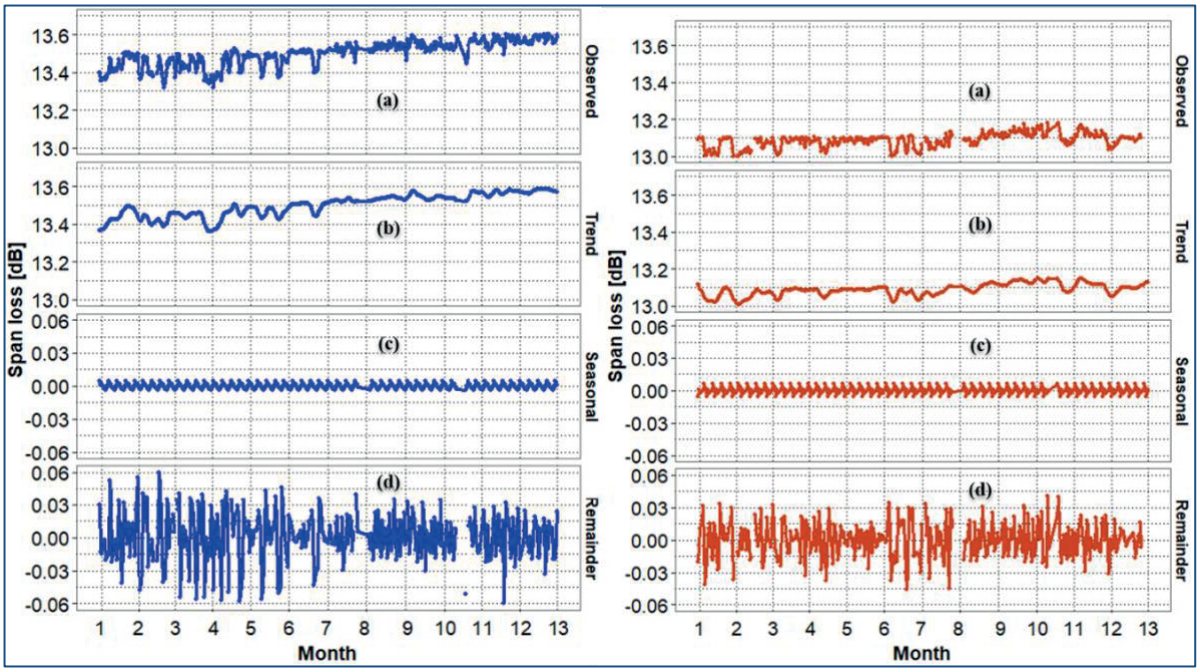

Figure 5.6 Direction A - B : Composantes de la décomposition des séries temporelles pour les spans de 13 dB [span 1 (bleu), span 3 (rouge)] : données brutes (a) ; composante de tendance (b) ; composante saisonnière (c) ; résiduel (d)

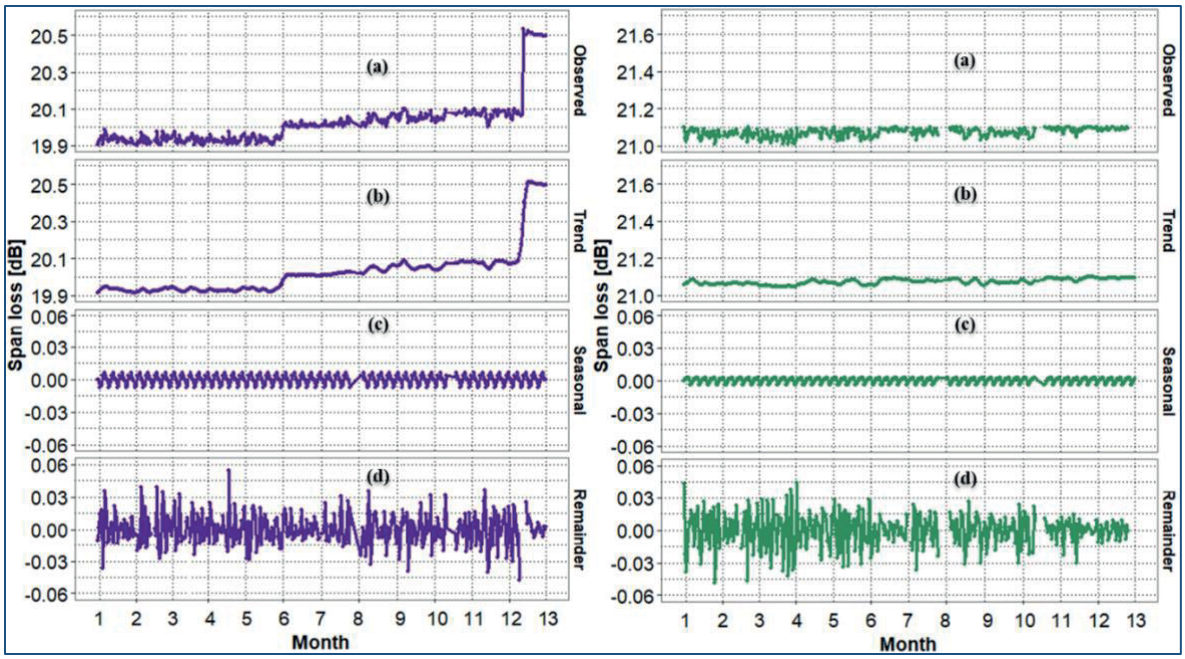

Figure 5.7 Direction A - B : Composantes de la décomposition des séries temporelles pour les spans de 20 dB [span 2 (mauve), span 4 (vert)] :
données brutes (a) ; composante de tendance (b) ; composante saisonnière (c) ; résiduel (d)

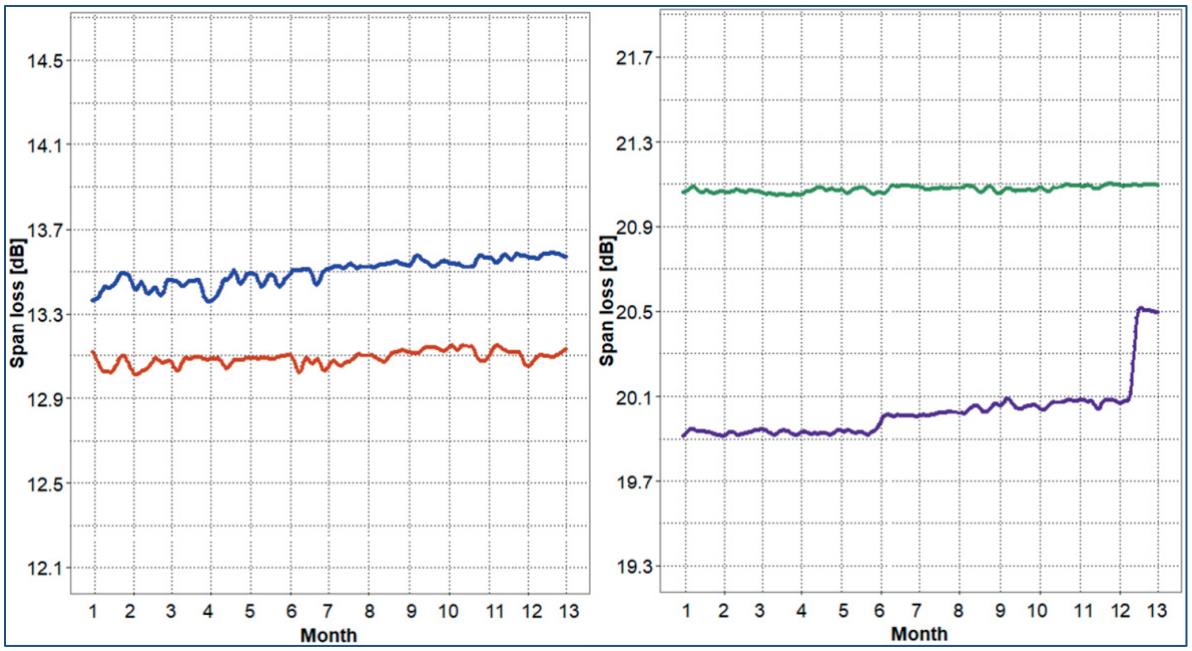

Figure 5.8 Direction A - B : Composantes de la tendance pour le span 1 (bleu), le span 3 (rouge), le span 2 (violet), le span 4 (vert)

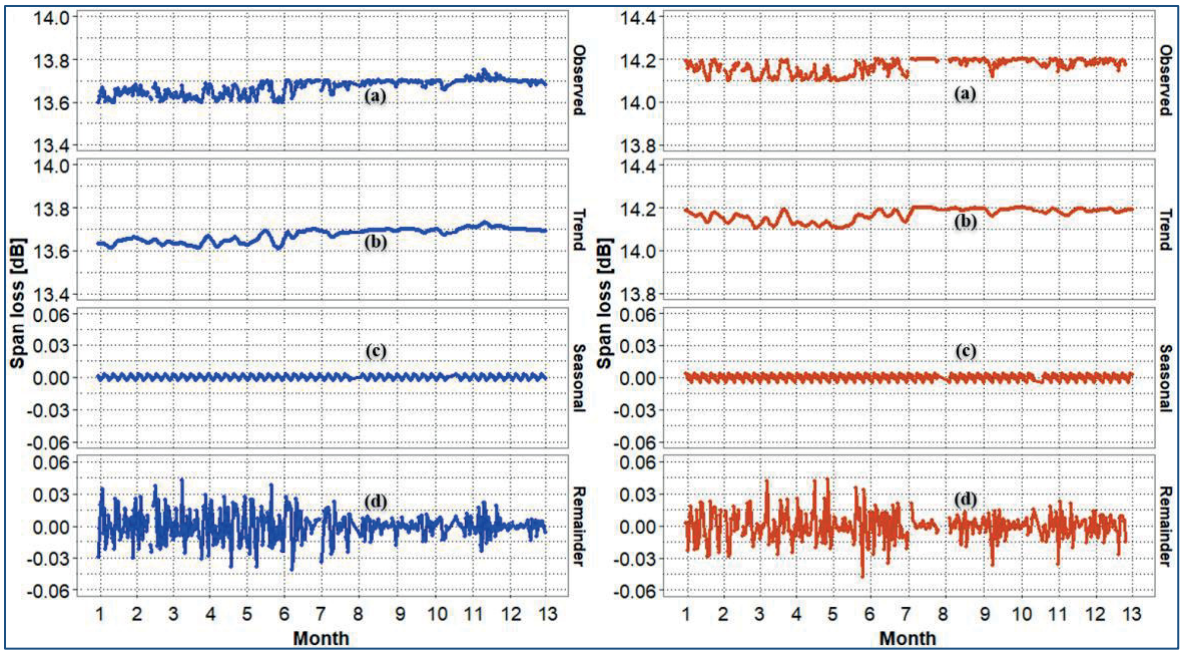

Figure 5.9 Direction B - A : Composantes de la décomposition des séries temporelles pour les spans de 14 dB [span 1 (bleu), span 3 (rouge)] : données brutes (a) ; composante de tendance (b) ; composante saisonnière (c) ; résiduel (d)

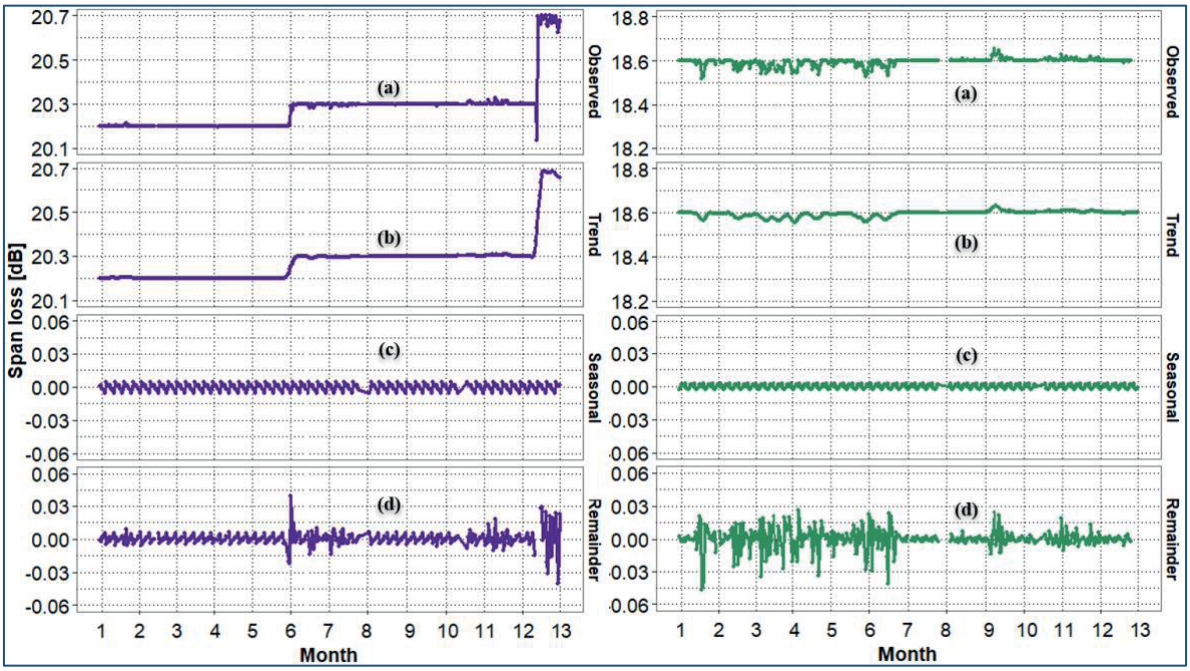

Figure 5.10 Direction B - A : Composantes de la décomposition des séries temporelles pour les spans de 18 dB [span 2 (mauve), span 4 (vert)] : données brutes (a) ; composante de tendance (b) ; composante saisonnière (c) ; résiduel (d)

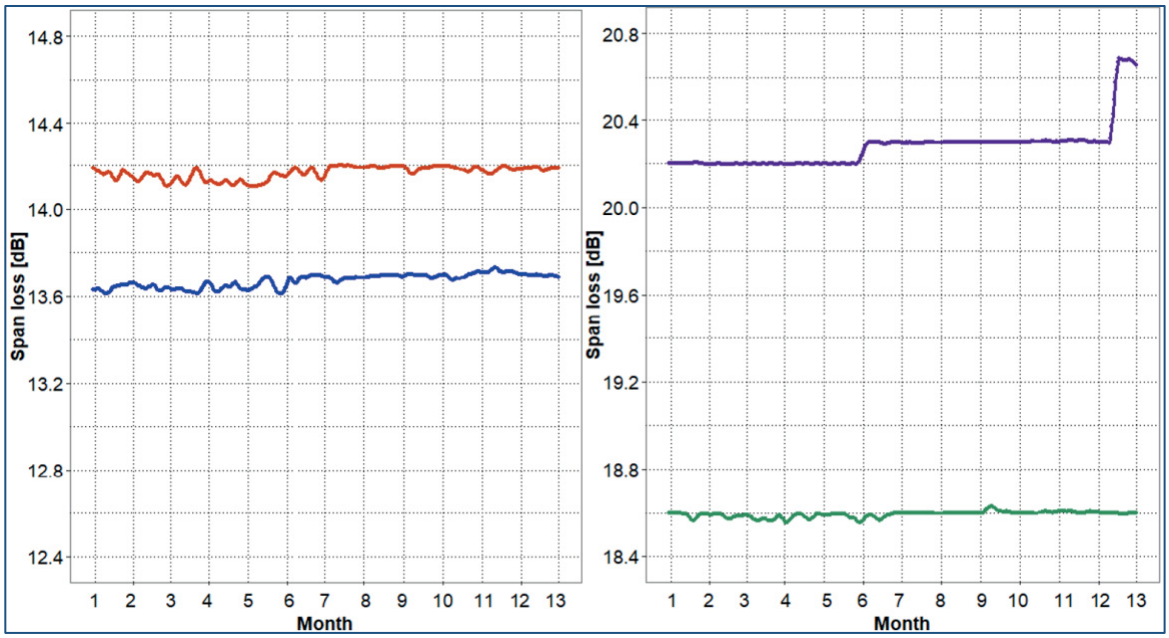

Figure 5.11 Direction B - A : Composantes de la tendance pour le span 1 (bleu), le span 3 (rouge), le span 2 (violet), le span 4 (vert)

L'étape suivante consiste à déterminer la pente de la composante de tendance à l'aide des tests statistiques de Mann-Kendall et Sen's slope. La librairie "*modifiedmk*" de RStudio est utilisées pour effectuer les deux tests avec un niveau de confiance de 0,95.

Les résultats sont présentés dans le Tableau 5.2. Le span 1 montre une variation de 0,2 dB/an dans un sens (A-B) et de 0,08 dB/an dans le sens opposé (B-A), avec une moyenne de 0,15 dB/an. La plage 2 montre une variation de 0,22 dB/an et de 0,15 dB/an. Les valeurs positives du paramètre τ et de la valeur S de Kendall indiquent la présence d'une tendance positive.

Le pourcentage de changement (*Pc*) de la composante de tendance est calculé en supposant une tendance linéaire. Ainsi, le pourcentage de variation est égal à la pente médiane (*Sen's slope*) multiplié par la longueur de la période (nombre de points de données dont la moyenne est calculée) divisé par la valeur moyenne (Taxak, Murumkar et Arya, 2014):

$$Pc = \frac{(Sen's\ slope\ \times num.data\ points)}{(mean\ over\ year)} \times 100\ \% \tag{5.11}$$

Tableau 5.2 Analyse statistique des données de perte par span

	Span	**τ**	**S**	**Q_{median}**	**Change**	**P_c**
	-	**-**	**($x10^4$)**	**($x10^{-4}$)**	**(dB/year)**	**(%)**
Direction A - B	1	0.7	4.7	5.0	0.20	1.3
	2	0.7	4.6	6.2	0.22	1.1
	3	0.4	2.6	1.9	0.07	0.5
	4	0.4	2.9	0.9	0.03	0.2
Direction B - A	1	0.6	4.0	2.3	0.08	0.6
	2	0.7	4.3	4.2	0.15	0.7
	3	0.4	2.3	1.5	0.05	0.4
	4	0.5	3.0	0.6	0.02	0.1

Afin de valider les résultats des tests, c'est-à-dire vérifier l'évolution de la perte par span en dB/an sur l'année suivante, le span 1 a été choisie étant donné qu'il montre une tendance positive progressive dans la perte. La Figure 5.12 montre les données de perte de span pour les 12 premiers mois (*en noir*) et les 12 mois supplémentaires (*en bleu*) pour les deux directions. La perte de span augmente dans les premiers mois de la deuxième année, comme prévu, mais semble se saturer à 13,7 dB. Le même comportement de saturation est observé dans la direction opposée.

Des mesures indépendantes de perte de span obtenues via le canal de supervision optique (*Optical Supervisory Channel, OSC*), bien que moins précis, a confirmé que la tendance observée est due à l'installation de fibres elle-même.

Une équation linéaire basée sur le premier ensemble de données de 12 mois (*en noir*) a été estimée et ajustée sur une période supplémentaire de 12 mois (*en bleu*). Le modèle d'ajustement linéaire basé sur la médiane *Theil-Sen*, est appliqué sur les douze premiers mois.

La Figure 5.12 indique l'ajustement linéaire (*en rouge*) appliquée aux données de perte par span actuelles (*en noir*), et supplémentaires (*en bleu*).

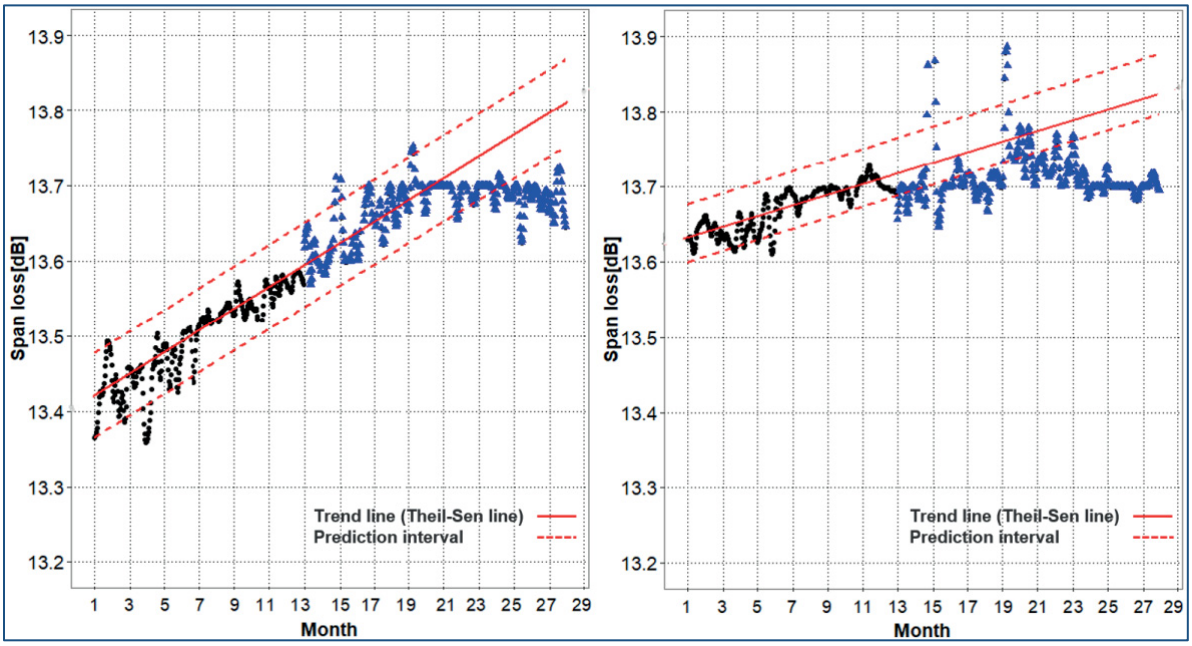

Figure 5.12 Données de perte de span 1 : période de 12 premiers mois (*en noir*), données additionnelles de 12 mois (*en bleu*) – (Direction A-B gauche) et (Direction B-A droite)

L'intercept et la pente de ces droites sont présentées dans le Tableau 5.3. Le modèle de régression s'avère significatif à un taux de signification de 5% avec une faible erreur-type résiduelle (Tableau 5.3). L'intervalle de prédiction représenté par la ligne pointillée, est la région qui contient environ 95 % des mesures. Cela signifie que, si une autre mesure est effectuée, il y a 95% de chance qu'elle se situe dans la bande de prédiction pointillée.

Tableau 5.3 Paramètres des droites ajustées et l'erreur type résiduelle

	Intercept (dB)	**Slope** (10^{-4})	***s* (dB)**
Span 1 (A-B)	5.31	4.73	0.028
Span 1 (B-A)	9.47	2.43	0.019

La qualité de l'ajustement de la régression est estimée en déterminant R^2 (Grekousis, 2020) :

$$R^2{}_{adj} = 1 - \frac{(n-1)}{(n-p)} \times \frac{SSE}{SST} \tag{5.12}$$

Où *SSE* est la somme des erreurs au carré, *SST* est la somme totale des carrés, *n* est le nombre d'observations et p est le nombre de variables indépendantes.

La Figure 5.13 montre les courbes de R^2 ajustées en fonction du temps dans le jeu de données supplémentaire de 12 mois (*en bleu*). Dans la direction (A - B), la valeur de R^2 augmente pendant un horizon de temps de 8 mois, ce qui valide les tendances observées pendant les 12 premiers mois d'observation, et diminue au fur et à mesure que la tendance se stabilise. Dans la direction (B - A), les valeurs de R^2 ajustées sont négatives car l'erreur dépasse la variance pour tous les horizons de temps.

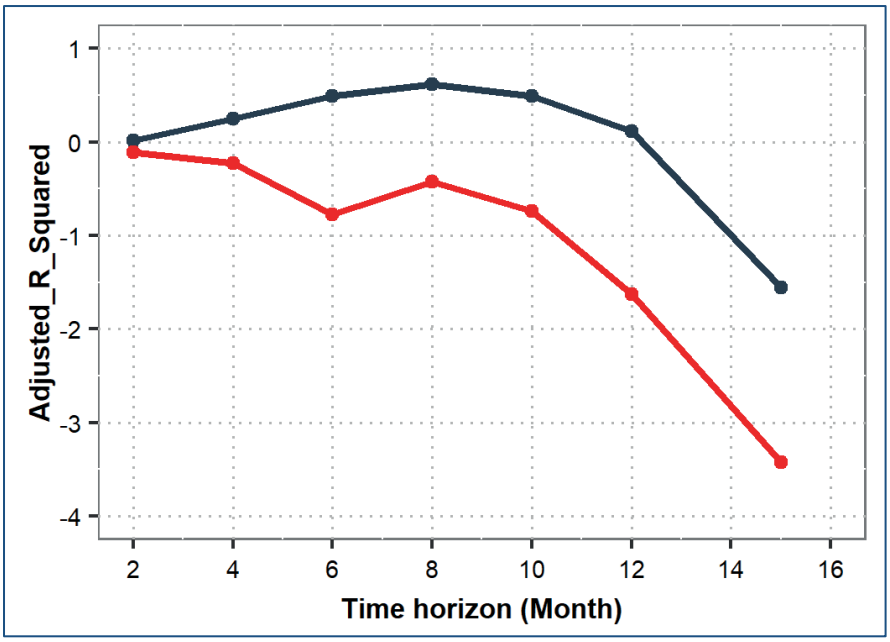

Figure 5.13 Courbes de R^2 ajustées en fonction de l'horizon de temps pour la période d'observation supplémentaire de 12 mois : Direction A - B (noir), Direction B - A (rouge)

5.6 Conclusion

Dans ce chapitre, la méthode de décomposition des séries temporelles a été appliquée aux données de perte de span réelles collectées dans une liaison optique amplifiée de 205 km d'un réseau de production sur une période de 12 mois, avec pour objectif de détecter une éventuelle dégradation à long terme de la perte de fibre dans les 8 spans étudiés. Les tests statistiques de Mann-Kendall et Sen's slope ont été utilisés pour estimer la variation de la perte de span. Les résultats montrent une variation en pourcentage de 1.3 ± 0.002% sur la période d'observation de 12 mois pour le span1 ou d'environ 0.2 dB.

L'augmentation de la perte de span détectée est très faible, comme prévu étant donné la période d'observation relativement courte. L'augmentation détectée de la perte dans le span1 a été validée en appliquant une régression linéaire sur un ensemble de données de 12 mois supplémentaires. Cela a confirmé une augmentation de 0.3 dB dans la tendance sur une période de 19 mois (direction A à B). Une saturation de la perte du span à 13.7 dB a été observée dans les deux sens.

Tel que confirmé par des mesures indépendantes des pertes par le biais de l'*OSC*, l'augmentation de la tendance et la saturation de la perte du span à 13.7 dB sont due à l'installation de fibre elle-même, et non pas à la mesure ou à une erreur d'instrumentation. Des investigations additionnelles avec des données de perte de span sur des périodes d'observation plus longues seront nécessaire pour évaluer la précision de la méthode de prédiction proposée et pour obtenir une meilleure compréhension de la dynamique de perte de span dans les réseaux de câbles de fibres optiques installées.

L'objectif de cette étude était de montrer comment la méthode de décomposition des séries temporelles combinée à l'analyse statistique pourrait être utilisée pour détecter les tendances de performance à long terme des réseaux optiques (Wen et al., 2018).

La méthode a été utilisée pour l'analyse des données historiques de perte de span collectées dans un réseau de production sur une période d'observation de plusieurs mois. Ces données de perte de span sont facilement accessibles aux opérateurs de réseau et ne nécessitent pas l'utilisation d'instruments supplémentaires tels que les réflectomètres temporels (*Optical time-domain reflectometer, OTDR*).

Cet outil serait utile pour l'analyse des données de perte de span, notamment pour localiser les endroits où des événements "imprévisibles" (ou points de rupture) et des changements de tendance se sont produits. Si elle est automatisée, cette méthode pourrait être utilisés par les opérateurs de réseau pour le monitoring des fibres optiques, pour une détection anticipée de la dégradation de l'installation à fibres optiques et la planification proactive du remplacement des fibres à l'aide des données historiques.

La méthode serait également utile pour quantifier la composante saisonnière de la perte de span des fibres ou d'autres métriques de performance telles que le taux d'erreur binaire.

Finalement, la méthode pourrait être appliquée pour identifier et localiser les événements imprévisibles des données d'observations brutes pour une analyse plus approfondie de la composante résiduelle.

CONCLUSION

Les développements actuels des réseaux de transport optique sont une réponse à la demande sans cesse croissante de capacité de réseau. Parmi les tendances récentes qui contribuent à cette demande, figurent le streaming vidéo, le déploiement de la téléphonie mobile 5G, l'interconnexion des centres de données, l'Internet des objets et d'autres applications (Boertjes, Reimer et Côté, 2019). Cela implique donc une quantité astronomique de données qui transite au travers de ces réseaux et une nécessité pour les opérateurs de surveiller la qualité des données. En effet il n'est pas suffisant de fournir des connexions haut débit aux utilisateurs mais il est important d'assurer l'intégrité des données qu'ils reçoivent dans un environnement réseau aussi complexe.

Dans ce sens donc, les opérateurs monitorent régulièrement les performances optiques de leurs réseaux pour garantir que les clients reçoivent leurs données de manière fiable. Les systèmes de contrôle du réseau mesurent en permanence la santé du réseau en collectant les paramètres de performances optiques toutes les 15 minutes.

Dans ce projet, 155 circuits de production ont été monitorés toutes les 15 minutes sur une période d'un an en utilisant des transpondeurs cohérents à 100 G du système OME6500 de Ciena. Les données ainsi collectées ont été stockées dans une base de données MySQL.

L'objectif principal de cette thèse est de faire une analyse des données de performance réelles issues de la surveillance de plusieurs liens optiques. La particularité de cette thèse réside dans le fait que ces paramètres de performance sont des données réelles collectées sur plusieurs saisons dans divers types de liens optiques.

Les deux premières parties du document font une revue de la littérature et une description des techniques statistiques qui sont utilisées dans l'analyse.

Dans le troisième chapitre, deux questions de recherches se sont posées au travers de la caractérisation des routes monitorées côté anomalies. D'abord trois catégories de circuits sont définies : la première catégorie où les anomalies affectent tous les canaux, la deuxième où un seul canal de la route est affecté et la troisième où une seule direction est affectée. L'idée est d'avoir une meilleure compréhension des paramètres monitorés en considérant les anomalies.

La première question qui se pose est : les anomalies détectées dans le rapport signal sur bruit et la puissance optique reçue au récepteur génèrent-elles forcément des secondes erronées? L'amplitude des pics d'anomalies détectés dans la puissance reçue est-elle proportionnelle aux secondes erronées générées? Cette analyse permettra d'établir ou non des liens entre les différentes métriques de performance monitorées dans ce projet.

Une analyse de corrélation menée sur certains circuits montre qu'un grand nombre d'évènements de secondes erronées enregistrées semblent plus lié aux anomalies dans la puissance reçue, cela même en considérant l'ensemble des routes monitorées en annexe.

La seconde question est relative aux recherches de causes probables de ces anomalies détectées. L'analyse de l'amplitude des pics et des secondes erronées montre dans certains cas une très forte corrélation entre les deux, mais pour l'ensemble des circuits la corrélation est faible. Une analyse des gains et puissances d'entrée et sortie des amplificateurs ont permis d'observer qu'un grand nombre d'anomalies dans le SNR et la puissance reçue au récepteur est lié à un disfonctionnement des amplificateurs. Cependant d'autres n'ont pu être établis.

Les secondes erronées montrent une propagation bidirectionnelle dans certains circuits c'est-à-dire que l'anomalie se trouve dans une direction donnée mais les secondes erronées sont présentes dans les deux directions.

Dans la quatrième partie, un regroupement des circuits sous analyse est effectué en utilisant les caractéristiques statistiques et l'algorithme de partitionnement k-moyennes (k-means).

L’objectif est de découvrir des groupes naturels présentant des similarités dans les circuits. Trois clusters ont été trouvés dans les données des 155 circuits.

Un premier cluster composé de 45 circuits très dynamiques avec une faible marge, un second cluster avec 56 circuits possédant une grande marge et de nature dynamiques. Ces deux premiers clusters présentent des anomalies et des erreurs dans le rapport signal sur bruit optique. Le cluster 1 est constitué des circuits présentant le plus d’anomalies et de secondes erronées. Le troisième cluster de 54 circuits se compose des circuits calmes et des circuits ayant une activité modérée dans les fluctuations du SNR. Le facteur géographique semble être le facteur commun entre les circuits du cluster 1 pour les 2/3 des circuits et le reste correspondant à des circuits régionaux (distance inférieure à 1000 km) et longue portée (distance supérieure à 1000 km).

Dans la dernière partie, une détection de tendance est menée en considérant la perte de la fibre optique. En effet, il est question de détecter les signes de vieillissement de la fibre optique pour les fibres présentant une certaine tendance croissante. Trois tests statistiques complémentaires sont appliqués sur des données de perte optique dans un segment (span) de fibre optique afin de détecter une éventuelle dégradation à long terme. L'augmentation de la perte détectée est très faible, comme prévu étant donné que la période d'observation est relativement courte.

Cependant, des mesures indépendantes ont confirmé que cette augmentation est due à l’installation de fibre elle-même, et non pas à la mesure ou à une erreur d'instrumentation.
Cet outil serait utile pour l'analyse des données de perte de span, notamment pour localiser les endroits où des événements "imprévisibles" (ou points de rupture) et des changements de tendance se sont produits. Si elle est automatisée, cette méthode pourrait être utilisés par les opérateurs de réseau pour la surveillance des fibres optiques, pour une détection anticipée de la dégradation de l’installation à fibres optiques et la planification proactive du remplacement des fibres à l'aide des données historiques.

Tous les circuits sélectionnés montrent une perte du signal autour de la fin juin début juillet due à une absence du signal au niveau du transmetteur bien qu'ils soient déployés dans des zones géographiques différentes. Le même phénomène est aussi remarqué dans tous les circuits sous étude en **Error! Reference source not found.** à **Error! Reference source not found.**.

Certains circuits présentent une périodicité de 24 h, c'est le cas de la route 2, de certains circuits de la route 6 et 8. L'amplitude est toutefois faible de l'ordre de 0.09 dB, obtenue en analysant la composante saisonnière issue de la décomposition de la série temporelle du SNR et en observant les fonctions d'autocorrélation.

RECOMMANDATIONS

Les données soumises à notre analyse sont une véritable mine d'or et n'ont pas fini de révéler leurs secrets. La présente étude est une première étape en ce sens qui a permis de dresser un portrait des données et d'extraire des caractéristiques communes ou particulières aux différents circuits optiques.

Dans le futur, il serait intéressant d'effectuer une prédiction d'anomalies à partir du regroupement des circuits et de l'analyse des données historiques d'anomalies. La méthode proposée par Sseguya (2020) semble une bonne piste. L'idée générale serait de faire une prédiction de performance (BER ou SNR, par exemple) et d'appliquer une technique de détection et de classification des anomalies sur la valeur prédite. Un tel modèle pourrait être utilisé comme système d'alerte par les opérateurs de réseau.

Les données que nous avons analysées ne sont pas complètes car des informations sur le type de fibre optique, de déploiement étaient inconnues pour plusieurs circuits. Ce qui n'a pas permis une validation rigoureuse de la segmentation effectuée au CHAPITRE 4. Dans ce sens, il serait intéressant de compléter ces informations afin d'avoir une meilleure analyse des groupes formés par l'algorithme k-moyennes.

Une autre recommandation serait d'obtenir la topologie exacte et plus complète incluant les caractéristiques des ROADM, des amplificateurs Raman et un inventaire précis et complet de tous les équipements déployés dans le réseau. Une telle connaissance permettrait de faire des analyses beaucoup plus détaillées sur l'influence des différents paramètres sur la dynamicité des connexions, le nombre et la cause des anomalies observées dans les liens optiques.

Il serait aussi important de voir l'évolution du nombre de connexions par lien en fonction du temps pour voir l'impact sur la performance des connexions, lors de l'injection et l'extraction de canaux par exemple.

Les travaux ont également montré qu'il serait important d'avoir une cartographie du réseau montrant la marge des circuits, de manière à permettre à l'opérateur de surveiller les circuits à faible marge. C'est le cas des circuits regroupés dans le groupe 1 qui présentent un grand nombre d'anomalies et de secondes erronées.

La règle intuitive au début du projet à l'effet que les circuits déployés dans des liaisons aériennes à très longue distance (plusieurs centaines de kilomètres) seraient nécessairement plus dynamiques que les circuits déployés dans des liens enfouis très courts ne s'est pas avérée réelle dans les faits. En effet, on a pu observer que des liens enfouis relativement courts présentaient un grand nombre d'anomalies. Il serait intéressant de pousser plus loin l'investigation, avec un peu plus d'informations sur le type de déploiement, le type et l'âge des segments de fibre optique, ainsi que les caractéristiques des câbles optiques.

L'étude réalisée au chapitre 5 a révélé une augmentation des pertes optiques dans un span de fibre optique. Cette liaison mériterait d'être surveillée par l'opérateur au-delà de la période d'observation afin de vérifier si la tendance se confirme et prendre des actions de maintenance préventive (remplacement de câbles) le cas échéant. De manière générale, la méthode de détection de dégradation de performance de fibre optique qui a été proposée aurait avantage à être utilisée par les opérateurs de réseau pour surveiller les câbles optiques estimés problématiques ou les liens d'un certain âge. Cette méthode serait simple à utiliser à partir des données de puissance des amplificateurs optiques et présente l'avantage de ne pas requérir d'instrumentation de test supplémentaire. Ainsi une collecte spéciale des données des liens problématiques par les opérateurs, suivie de la mise en application de la technique de décomposition de séries temporelles, permettrait de surveiller les liens en continu et de générer une alarme en cas de dégradation dépassant un certain seuil.

La source des anomalies de performance observées pour un grand nombre de connexions n'étant pas totalement comprise, et sachant par ailleurs que des erreurs sont souvent associées aux anomalies, il est recommandé aux opérateurs de réseau de porter attention à l'activité humaine autour des équipements et câbles optiques.

Afin de détecter la cause de toutes les anomalies observées, il serait important dans le futur d'ajouter aux données de monitoring des informations additionnelles telles que les relevés d'intervention humaine sur le terrain, ce qui pourrait entraîner des chutes de puissance au récepteur par manipulation de fibre optique par exemple. Si l'impact de l'activité humaine pouvait être établi, il faudrait alors privilégier des pratiques de manipulations qui minimisent les anomalies. D'autres informations à ajouter dans ce sens seraient les relevés des pannes observées au niveau des éléments de réseau et les relevés des ajouts et extractions de canaux qui causent des fluctuations de puissance au niveau des amplificateurs optiques.

Il a été déterminé que les événements caractérisés par un faible nombre de secondes erronées (de l'ordre de 1 à 4) dans une des routes aériennes d'intérêt (route 2) pour l'opérateur réseau n'étaient pas liés aux fluctuations des métriques de performance disponibles (voir encadré pointillé noir à la Figure 3.5). Une investigation plus poussée s'impose pour en comprendre l'origine. Une piste serait d'observer les fluctuations d'état de polarisation et de la température.

De manière générale, il y aurait nettement intérêt pour l'opérateur de poursuivre l'analyse des routes et connexions au-delà de la période de 12 mois étudiée dans le cadre de cette thèse.

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES

Aguilera, Fátima, Fabio Orlandi, Luis Ruiz-Valenzuela, Monji Msallem et Marco Fornaciari. 2015. « Analysis and interpretation of long temporal trends in cumulative temperatures and olive reproductive features using a seasonal trend decomposition procedure ». *Agricultural and Forest Meteorology,* vol. 203, p. 208-216.

Aladin, Sandra, et Christine Tremblay. 2018. « Cognitive tool for estimating the QoT of new lightpaths ». In *Optical Fiber Communication Conference*. p. M3A. 3. Optical Society of America.

Alhaj, Taqwa Ahmed, Maheyzah Md Siraj, Anazida Zainal, Huwaida Tagelsir Elshoush et Fatin Elhaj. 2016. « Feature selection using information gain for improved structural-based alert correlation ». *PloS one,* vol. 11, nº 11, p. e0166017.

Ali, Rawshan, Alban Kuriqi, Shadan Abubaker et Ozgur Kisi. 2019. « Long-Term Trends and Seasonality Detection of the Observed Flow inYangtze River Using Mann-Kendall and Sen's Innovative Trend Method ». *Water*.

Alkhayrat, Maha, Mohamad Aljnidi et Kadan Aljoumaa. 2020. « A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA ». *Journal of Big Data,* vol. 7, nº 1, p. 9.

Allogba, Stéphanie, et Christine Tremblay. 2018. « K-Nearest neighbors classifier for field bit error rate data ». In *2018 Asia Communications and Photonics Conference (ACP)*. p. 1-3. IEEE.

Aparna, UR, et Shaiju Paul. 2016. « Feature selection and extraction in data mining ». In *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*. p. 1-3. IEEE.

Babu, Ugendra. 2019. «What is difference between Data Science and Big Data?». < https://www.quora.com/What-is-difference-between-Data-Science-and-Big-Data >. Consulté le 24 juin.

Balasubramanian, Anand, Karunanidhi Duraisamy, Subramani T, Srinivasamoorthy Krishnaraj et M. Suresh. 2019. « Long-term trend detection and spatiotemporal analysis of groundwater levels using GIS techniques in Lower Bhavani River basin, Tamil Nadu, India ». *Environment, Development and Sustainability*.

Banerjee, Amit, et Rajesh N Dave. 2004. « Validating clusters using the Hopkins statistic ». In *2004 IEEE International conference on fuzzy systems (IEEE Cat. No. 04CH37542)*. Vol. 1, p. 149-153. IEEE.

Baudot, JY. 2020. « Les corrélations des rangs ». In *Techniques et concepts de l'entreprise, de la finance et de l'économie (et fondements mathématiques)*. < http://www.jybaudot.fr/Correl_regress/rangs.html >. Consulté le 21 septembre 2020.

Berg, Ruben Geert van den. 2020. « Shapiro-Wilk Test - What is It? ». < https://www.spss-tutorials.com/spss-shapiro-wilk-test-for-normality/#shapiro-wilk-test-null-hypothesis >. Consulté le 2020-09-18.

Besse, Philippe. 2005. « DATA Mining II. Modélisation Statistique & Apprentissage ». *Université Pde Toulouse*.

Beyer, H. 1981. « Tukey, John W.: Exploratory Data Analysis. Addison-Wesley Publishing Company Reading, Mass. — Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario, Sydney 1977, XVI, 688 S ». *Biometrical Journal,* vol. 23, nº 4, p. 413-414.

Bezdek, James C, et Richard J Hathaway. 2002. « VAT: A tool for visual assessment of (cluster) tendency ». In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*. Vol. 3, p. 2225-2230. IEEE.

Bezdek, James C, Richard J Hathaway et Jacalyn M Huband. 2007. « Visual assessment of clustering tendency for rectangular dissimilarity matrices ». *IEEE Transactions on fuzzy systems,* vol. 15, nº 5, p. 890-903.

Blessie, E Chandra, et E Karthikeyan. 2012. « Sigmis: A feature selection algorithm using correlation based method ». *Journal of Algorithms & Computational Technology,* vol. 6, nº 3, p. 385-394.

Boertjes, David W, Michael Reimer et David Côté. 2019. « Practical considerations for near-zero margin network design and deployment ». *Journal of Optical Communications and Networking,* vol. 11, nº 9, p. C25-C34.

Bourbonnais, R., et M. Terraza. 2016. *Analyse des séries temporelles - 4e éd.: Cours et exercices corrigés - Applications à l'économie et à la gestion*. Dunod.

Brugiere, Thomas. 2015. « Monitoring à long terme d'une liaison optique aérienne cohérente à 40 Gbit/s ». Montréal, ÉTS, 292 p.

Calas, Guillaume. 2006. « Etudes des principaux algorithmes de data mining ». *Article de recherche cole ingnieurs en informatique EPITF, France*.

Caloiero, T., R. Coscarelli, E. Ferrari et B. Sirangelo. 2017. « Trend analysis of monthly mean values and extreme indices of daily temperature in a region of southern Italy ». *International Journal of Climatology,* vol. 37, nº S1, p. 284-297.

Campos, L Alberto, Z Jia et L Wolcott. 2018. « Proactive Network Maintenance Evolution to the Optical Domain in Coherent Optics ». *SCTE/ISBE Cable-Tec Expo'18*.

Chan, C. C. K. 2013. « Advanced optical performance monitoring for next generation access networks ». In *2013 18th OptoElectronics and Communications Conference held jointly with 2013 International Conference on Photonics in Switching (OECC/PS 2013), 30 June-4 July 2013*. (Piscataway, NJ, USA), p. 2 pp. Coll. « 2013 18th OptoElectronics and Communications Conference held jointly with 2013 International Conference on Photonics in Switching (OECC/PS 2013) »: IEEE.

Chan, Calvin CK. 2010. *Optical performance monitoring: advanced techniques for next-generation photonic networks*. Academic Press.

Chandola, Varun, Arindam Banerjee et Vipin Kumar. 2009. « Anomaly detection: A survey ». *ACM computing surveys (CSUR),* vol. 41, nº 3, p. 1-58.

Charrad, Malika, Nadia Ghazzali, Veronique Boiteau et Azam Niknafs. 2014. « Determining the best number of clusters in a data set ». *J Stat Softw.*

Cleveland, Robert B., William S. Cleveland, Jean E. McRae et and Irma Terpenning. 1990. « STL: A Seasonal-Trend Decomposition Procedure Based on Loess ». *Journal of Official Statistics,* vol. Vol. 6. No. I. , p. pp. 3-73.

Conover, W.J. 1999. *Practical nonparametric statistics*. Wiley.

Critical Data, MIT. 2016. *Secondary analysis of electronic health records*. Springer Nature.

Dancho, Matt, et Davis Vaughan. 2018. « Package 'anomalize' ».

Dangeti, Pratap. 2017. *Statistics for machine learning*. Packt Publishing Ltd.

Dérobert, Nicolas. 2020. « Les variables statistiques ». < https://commentprogresser.com/statistique-variables.html >. Consulté le 29 juin.

Dey, Ayon. 2016. « Machine learning algorithms: a review ». *International Journal of Computer Science and Information Technologies,* vol. 7, nº 3, p. 1174-1179.

Dong, Z., F. N. Khan, Q. Sui, K. Zhong, C. Lu et A. P. T. Lau. 2016. « Optical Performance Monitoring: A Review of Current and Future Technologies ». *Journal of Lightwave Technology,* vol. 34, nº 2, p. 525-543.

Dy, Jennifer G, et Carla E Brodley. 2004. « Feature selection for unsupervised learning ». *Journal of machine learning research,* vol. 5, nº Aug, p. 845-889.

Ednay, Richard. 2010. « Field trial of pmd test methods and investigation into the dynamics of polarisation effects in a variety of installed cable environments ». In *Terena Networking Conference*.

Filliben, James J., et Alan Heckert. 2013. « NIST/SEMATECH e-Handbook of Statistical Methods ». < https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm >. Consulté le 10 juin.

Fränti, Pasi, et Sami Sieranoja. 2019. « How much can k-means be improved by using better initialization and repeats? ». *Pattern Recognition,* vol. 93, p. 95-112.

Gan, Guojun, Chaoqun Ma et Jianhong Wu (162). 2007. *Data clustering: theory, algorithms, and applications*. SIAM, 489 p.

Ganguly, Kuntal. 2017. *R Data Analysis Cookbook*. Packt Publishing Ltd.

Veuillez sélectionner un type de document autre que « Generic » afin de faire afficher la référence bibliographique.

Grekousis, George (370). 2020. *Spatial Analysis Theory and Practice: Describe - Explore - Explain through GIS* (2020-04-29). United Kingdom: Cambridge University Press.

Gu, Rentao, Zeyuan Yang et Yuefeng Ji. 2020. « Machine learning for intelligent optical networks: A comprehensive survey ». *Journal of Network and Computer Applications,* vol. 157, p. 102576.

Hafen, Ryan. 2016. « Enhanced Seasonal Decomposition of Time Series by Loess ». < https://cran.r-project.org/web/packages/stlplus/stlplus.pdf >.

Halkidi, Maria, Yannis Batistakis et Michalis Vazirgiannis. 2002. « Clustering validity checking methods: part II ». *ACM Sigmod Record,* vol. 31, nº 3, p. 19-27.

Halkidi, Maria, et Michalis Vazirgiannis. 2001. « Clustering validity assessment: Finding the optimal partitioning of a data set ». In *Proceedings 2001 IEEE International Conference on Data Mining*. p. 187-194. IEEE.

Han, Jiawei, Micheline Kamber et Jian Pei. 2011. « Data mining concepts and techniques third edition ». *The Morgan Kaufmann Series in Data Management Systems*, p. 83-124.

Han, Jiawei, Micheline Kamber et Jian Pei. 2012a. « 1 - Introduction ». In *Data Mining (Third Edition)*, sous la dir. de Han, Jiawei, Micheline Kamber et Jian Pei. p. 1-38. Boston: Morgan Kaufmann. < http://www.sciencedirect.com/science/article/pii/B9780123814791000010 >.

Han, Jiawei, Micheline Kamber et Jian Pei. 2012b. « 2 - Getting to Know Your Data ». In *Data Mining (Third Edition)*, sous la dir. de Han, Jiawei, Micheline Kamber et Jian Pei. p. 39-82. Boston: Morgan Kaufmann. < http://www.sciencedirect.com/science/article/pii/B9780123814791000022 >.

Haro, Justino, et Paloma R. Horche. 2008. « Evolution of PMD with the temperature on installed fiber ». *Optical Fiber Technology,* vol. 14, nº 3, p. 203-213.

Harvey, Andrew C, et Simon Peters. 1990. « Estimation procedures for structural time series models ». *Journal of Forecasting,* vol. 9, nº 2, p. 89-108.

Hauke, Jan, et Tomasz Kossowski. 2011. « Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data ». *Quaestiones geographicae,* vol. 30, nº 2, p. 87-93.

Hauske, F. N., M. Kuschnerov, B. Spinnler et B. Lankl. 2009. « Optical Performance Monitoring in Digital Coherent Receivers ». *Journal of Lightwave Technology,* vol. 27, nº 16, p. 3623-3631.

Helsel, Dennis, et Robert Hirsch. 2002. *Statistical methods in water resources*. Coll. « Book 4 : Hydrologic Analysis and Interpretation »: USGS, 524 p. < https://pubs.usgs.gov/twri/twri4a3/twri4a3.pdf >. Consulté le 09 september 2019.

Hipel, Keith W., et A. Ian McLeod. 1994. « Chapter 23 Nonparametric Tests for Trend Detection ». In *Developments in Water Science*, sous la dir. de Hipel, Keith W., et A. Ian McLeod. Vol. 45, p. 853-938. Elsevier. < http://www.sciencedirect.com/science/article/pii/S0167564808706889 >.

Jaiswal, Anand, Cherian Samuel et V. M. Kadabgaon. 2018. « Statistical trend analysis and forecast modeling of air pollutants ». *Global Journal of Environmental Science and Management,* vol. 4, p. 427-438.

Jonietz, David, Dominik Bucher, Henry Martin et Martin Raubal. 2018. « Identifying and interpreting clusters of persons with similar mobility behaviour change processes ». In *The Annual International Conference on Geographic Information Science*. p. 291-307. Springer.

Kartalopoulos, Stamatios V. 2001. *Fault Detectability in DWDM: Toward Higher Signal Quality and System Reliability*. Wiley-IEEE Press.

Kassambara, Alboukadel. 2020. « Assessing Clustering Tendency ». <https://www.datanovia.com/en/lessons/assessing-clustering-tendency/#why-assessing-clustering-tendency >. Consulté le 16 octobre 2020.

Kira, Kenji, et Larry A Rendell. 1992. « The feature selection problem: Traditional methods and a new algorithm ». In *Aaai*. Vol. 2, p. 129-134.

Komorowski, Matthieu, Dominic C. Marshall, Justin D. Salciccioli et and Yves Crutain. 2016. « Exploratory Data Analysis ». In *Secondary Analysis of Electronic Health Records*, sous la dir. de Open, Springer. p. 436. MIT Critical Data. Consulté le 20200616.

Krispin, Rami (153). 2019. *Hands-On Time Series Analysis with R: Perform time series analysis and forecasting using R* (May 2019). Birmingham: Packt Publishing Ltd., 418 p.

L'heureux, Alexandra, Katarina Grolinger, Hany F Elyamany et Miriam AM Capretz. 2017. « Machine learning with big data: Challenges and approaches ». *IEEE Access,* vol. 5, p. 7776-7797.

Lawson, Richard G., et Peter C. Jurs. 1990. « New index for clustering tendency and its application to chemical problems ». *Journal of Chemical Information and Computer Sciences,* vol. 30, nº 1, p. 36-41.

Lise, Vaudor. 2015. « Autocorrélation de séries temporelles ou spatiales ». In *R-atique: Analyse de données avec R*. < http://perso.ens-lyon.fr/lise.vaudor/autocorrelation-de-series-temporelles-ou-spatiales/ >. Consulté le 17 juin.

Liu, Yanchi, Zhongmou Li, Hui Xiong, Xuedong Gao et Junjie Wu. 2010. « Understanding of internal clustering validation measures ». In *2010 IEEE International Conference on Data Mining*. p. 911-916. IEEE.

Mahammed, Fatima Ait. 2018. « Approches d'apprentissage automatique pour la détection du spam web : exploration de diverses caractéristiques ». Master. Université du Québec à Montréal, 122 p. Consulté le 20200630.

Manikandan, S. 2011a. « Measures of central tendency: Median and mode ». *Journal of pharmacology and pharmacotherapeutics,* vol. 2, nº 3, p. 214.

Manikandan, S. 2011b. « Measures of central tendency: The mean ». *Journal of Pharmacology and Pharmacotherapeutics,* vol. 2, nº 2, p. 140.

Maslo, Anis, Mujo Hodzic, Edvin Skaljo et Aljo Mujcic. 2020. « Aging and Degradation of Optical Fiber Parameters in a 16-Year-Long Period of Usage ». *Fiber and Integrated Optics,* vol. 39, p. 39.

Mata, Javier, Ignacio de Miguel, Ramon J Duran, Noemi Merayo, Sandeep Kumar Singh, Admela Jukan et Mohit Chamania. 2018. « Artificial intelligence (AI) methods in optical networks: A comprehensive survey ». *Optical switching and networking,* vol. 28, p. 43-57.

Matthewson, M John. 1993. « Optical fiber reliability models ». In *Fiber Optics Reliability and Testing: A Critical Review*. Vol. 10272, p. 1027203. International Society for Optics and Photonics.

Meals, Donald W., Jean Spooner, Steven A. Dressing et Jon B. Harcum. 2011. « Statistical Analysis for Monotonic Trends ». < https://www.epa.gov/sites/production/files/2016-05/documents/tech_notes_6_dec2013_trend.pdf >. Consulté le 2020-04-21.

Medecine, National Academies of Sciences Engineering. 2017. *Applied and Theretical Statistics*. Coll. « Strengthening Data Science Methods for Department of Defense personnel and readiness missions. ». Washington, 165 p. < www.nap.edu >.

Michel, Annie. 2016. « Monitoring of performance parameters in a metro ring using a 100G coherent transponder ». ÉTS, 120 p.

Mohamad, Ismail Bin, et Dauda Usman. 2013. « Standardization and its effects on K-means clustering algorithm ». *Research Journal of Applied Sciences, Engineering and Technology,* vol. 6, nº 17, p. 3299-3303.

Mukaka, Mavuto M. 2012. « A guide to appropriate use of correlation coefficient in medical research ». *Malawi medical journal,* vol. 24, nº 3, p. 69-71.

Musumeci, Francesco, Cristina Rottondi, Avishek Nag, Irene Macaluso, Darko Zibar, Marco Ruffini et Massimo Tornatore. 2018. « An overview on application of machine learning techniques in optical networks ». *IEEE Communications Surveys & Tutorials,* vol. 21, nº 2, p. 1383-1408.

Nabila, Kateb. 2011. « Une approche multi agents pour le Data Mining ». Informatique. Algérie, Larbi Ben M'hidi Oum El Bouaghi, 126 p.

Nawrin, Sadia, Md Rahatur Rahman et Shamim Akhter. 2017. « Exploreing k-means with internal validity indexes for data clustering in traffic management system ». *International Journal of Advanced Computer Science and Applications,* vol. 8, nº 3, p. 264-272.

Olofintoye, Oluwatosin, Josiah Adeyemo et Fred Otieno. 2012. « Impact of Regional Climate Change on Freshwater Resources and Operation of the Vanderkloof Dam System in South Africa ». In *Global warming , impacts and future perspective* sous la dir. de Singh, Bharat Raj. p. 364. Croatia: Janeza Trdine 9, 51000 Rijeka

Pacheco, Erik Rodriguez. 2015. *Unsupervised Learning with R*. Packt Publishing Ltd.

Panayiotou, Tania, Sotirios P Chatzis et Georgios Ellinas. 2018. « Leveraging statistical machine learning to address failure localization in optical networks ». *IEEE/OSA Journal of Optical Communications and Networking,* vol. 10, nº 3, p. 162-173.

Phyu, Thu Zar, et Nyein Nyein Oo. 2016. « Performance comparison of feature selection methods ». In *MATEC Web of Conferences*. Vol. 42, p. 06002. EDP Sciences.

Veuillez sélectionner un type de document autre que « Generic » afin de faire afficher la référence bibliographique.

Poulain, Marcel, et Irina Severin. 2007. *Aging and reliability of single-mode silica optical fibers in Optical Fibers Research Advances*.

Provotar, Oleksandr I, Yaroslav M Linder et Maksym M Veres. 2019. « Unsupervised anomaly detection in time series using lstm-based autoencoders ». In *2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT)*. p. 513-517. IEEE.

Rakotomalala, Ricco. 2011. *Tests de normalité: Techniques empiriques et tests statistiques*. Université Lumière Lyon 2, 59 p.

Ratner, Bruce. 2017. *Statistical and Machine-Learning Data Mining, Third Edition: Techniques for Better Predictive Modeling and Analysis of Big Data, Third Edition*. Chapman & Hall/CRC.

Roberts, K., S. H. Foo, M. Moyer, M. Hubbard, A. Sinclair, J. Gaudette et C. Laperle. 2015. « High Capacity Transport—100G and Beyond ». *Journal of Lightwave Technology,* vol. 33, nº 3, p. 563-578.

Roberts, K., M. O'Sullivan, Wu Kuang-Tsan, Sun Han, A. Awadalla, D. J. Krause et C. Laperle. 2009a. « Performance of Dual-Polarization QPSK for Optical Transport Systems ». *Lightwave Technology, Journal of,* vol. 27, nº 16, p. 3546-3559.

Roberts, Kim, Maurice O'Sullivan, Kuang-Tsan Wu, Han Sun, Ahmed Awadalla, David J Krause et Charles Laperle. 2009b. « Performance of dual-polarization QPSK for optical transport systems ». *Journal of lightwave technology,* vol. 27, nº 16, p. 3546-3559.

Rottondi, Cristina, Luca Barletta, Alessandro Giusti et Massimo Tornatore. 2018. « Machine-learning method for quality of transmission prediction of unestablished lightpaths ». *Journal of Optical Communications and Networking,* vol. 10, nº 2, p. A286-A297.

Roudas, Ioannis. 2012. « Coherent optical communication systems ». In *WDM Systems and Networks*. p. 373-417. Springer.

Rousseeuw, Peter J. 1987. « Silhouettes: a graphical aid to the interpretation and validation of cluster analysis ». *Journal of computational and applied mathematics,* vol. 20, p. 53-65.

Russell, Stuart, et Peter Norvig. 2002. « Artificial intelligence: a modern approach ».

Saeys, Yvan, Iñaki Inza et Pedro Larrañaga. 2007. « A review of feature selection techniques in bioinformatics ». *bioinformatics,* vol. 23, nº 19, p. 2507-2517.

Saradjian, MR, et M Akhoondzadeh. 2011. « Thermal anomalies detection before strong earthquakes (M> 6.0) using interquartile, wavelet and Kalman filter methods ». *Natural Hazards and Earth System Sciences,* vol. 11, nº 4, p. 1099-1108.

Sartzetakis, Ippokratis, Konstantinos Kostas Christodoulopoulos et Emmanouel Manos Varvarigos. 2019. « Accurate quality of transmission estimation with machine learning ». *Journal of Optical Communications and Networking,* vol. 11, nº 3, p. 140-150.

Scholar, Semantic. 2020. « Chapter 4 Exploratory Data Analysis ». In. <https://www.semanticscholar.org/paper/Chapter-4-Exploratory-Data Analysis/d1103d06edb64bae58b0ba30ce1a8901c2ef2a7e >.

Schonlau, Matthias. 2004. « Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams ». *Computational Statistics,* vol. 19, nº 1, p. 95-111.

Schyns, H. 2011. « Cours de STATISTIQUE - Statistique descriptive ». < http://notesdecours.drivehq.com/courspdf/StatDescr.pdf >. Consulté le 29 juin.

Scott, David W. 1979. « On optimal and data-based histograms ». *Biometrika,* vol. 66, nº 3, p. 605-610.

Severin, I, R El Abdia, M Poulaina et G Amza. 2005. « Fatigue testing procedures of silica optical fibres ». *Journal of Optoelectronics and Advanced Materials,* vol. 7, nº 3, p. 1581-1587.

Shahkarami, Shahin, Francesco Musumeci, Filippo Cugini et Massimo Tornatore. 2018. « Machine-learning-based soft-failure detection and identification in optical networks ». In *2018 Optical Fiber Communications Conference and Exposition (OFC)*. p. 1-3. IEEE.

Sun, Han, Kuang-Tsan Wu et Kim Roberts. 2008. « Real-time measurements of a 40 Gb/s coherent system ». *Optics Express,* vol. 16, nº 2, p. 873-879.

Taxak, Arun Kumar, A. R. Murumkar et D. S. Arya. 2014. « Long term spatial and temporal rainfall trends and homogeneity analysis in Wainganga basin, Central India ». *Weather and Climate Extremes,* vol. 4, p. 50-61.

Tremblay, Christine. 2020. *Towards Cognitive Management and Performance Monitoring in Coherent Optical Networks*. École de technologie supérieure, 2 p.

Tremblay, Christine, Stéphanie Allogba et Sandra Aladin. 2019. « Quality of transmission estimation and performance prediction of lightpaths using machine learning ».

Tufféry, Stéphane. 2007. *Data mining et statistique décisionnelle: l'intelligence des données*. Paris, 529 p.

Twaissi, Naseem M. 2008. « An evaluation of the implementation of Total Quality Management (TQM) within the Information and Communications Technology (ICT) sector in Jordan »

UCLouvain. 2020. « Science des données ». < https://uclouvain.be/fr/facultes/epl/sciences-des-donnees.html >. Consulté le 24 juin.

University, Strathmore. 2018-2019. « Math- Statistics- Skewness, Moments And Kurtosis Notes ». < https://www.studocu.com/row/document/strathmore-university/business-in-information-technology/lecture-notes/math-statistics-skewness-moments-and-kurtosis-notes/2650413/view >. Consulté le 11-06.

Vela, Alba P, Marc Ruiz, Francesco Fresi, Nicola Sambo, Filippo Cugini, Gianluca Meloni, Luca Potì, Luis Velasco et Piero Castoldi. 2017. « BER degradation detection and failure identification in elastic optical networks ». *Journal of Lightwave Technology,* vol. 35, nº 21, p. 4595-4604.

Wang, Xiaozhe, Kate A Smith, Rob Hyndman et Damminda Alahakoon. 2004. « A scalable method for time series clustering ». In *Proc. Unrefereed Res. Papers*. Vol. 1.

Wang, Xiaozhe, Kate Smith et Rob Hyndman. 2006. « Characteristic-based clustering for time series data ». *Data mining and knowledge Discovery,* vol. 13, nº 3, p. 335-364.

Wang, Zhilong, Min Zhang, Danshi Wang, Chuang Song, Min Liu, Jin Li, Liqi Lou et Zhuo Liu. 2017. « Failure prediction using machine learning and time series in optical network ». *Optics Express,* vol. 25, nº 16, p. 18553-18565.

Wen, Qingsong, Jingkun Gao, Xiaomin Song, Liang Sun, Huan Xu et Shenghuo Zhu. 2018. « RobustSTL: A Robust Seasonal-Trend Decomposition Algorithm for Long Time Series ». *Association for the Advancement of Artificial Intelligence* vol. AAAI-19, p. 5409-5416.

Wikipedia. 2020. « Test de Shapiro - Wilk ». < https://fr.wikipedia.org/wiki/Test_de_Shapiro-Wilk >. Consulté le 18 septembre 2020.

Willner, Alan E., Zhongqi Pan et Changyuan Yu. 2008. « Optical performance monitoring ». In., p. 233-292. Coll. « Optical Fiber Telecommunications V1B »: Elsevier Inc. < http://dx.doi.org/10.1016/B978-0-12-374172-1.00007-2 >.

Woodward, S. L., L. E. Nelson, C. R. Schneider, L. A. Knox, M. O' Sullivan, C. Laperle, M. Moyer et S. Foo. 2014. « Long-Term Observation of PMD and SOP on Installed Fiber Routes ». *IEEE Photonics Technology Letters,* vol. 26, nº 3, p. 213-216.

Wosiak, Agnieszka, et Danuta Zakrzewska. 2018. « Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis ». *Complexity,* vol. 2018.

Xu, Tianhua, Jie Li, Gunnar Jacobsen, Sergei Popov, Anders Djupsjöbacka, Richard Schatz, Yimo Zhang et Polina Bayvel. 2015. « Field trial over 820km installed SSMF and its potential Terabit/s superchannel application with up to 57.5-Gbaud DP-QPSK transmission ». *Optics Communications,* vol. 353, p. 133-138.

Yameogo, B. L. M., D. W. Charlton, D. Doucet, C. Desrosiers, M. O' Sullivan et C. Tremblay. 2020. « Trends in Optical Span Loss Detected Using the Time Series Decomposition Method ». *Journal of Lightwave Technology*, p. 1-1.

Zhongqi, Pan, Yu Changyuan et A. E. Willner. 2010. « Optical performance monitoring for the next generation optical communication networks ». *Optical Fiber Technology: Materials, Devices and Systems,* vol. 16, nº 1, p. 20-45.

Zolhavarieh, Seyedjamal, Saeed Aghabozorgi et Ying Wah Teh. 2014. « A review of subsequence time series clustering ». *The Scientific World Journal,* vol. 2014.

Zumel, Nina, John Mount et Jim Porzak. 2014. *Practical data science with R*. Manning Shelter Island, NY.