République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université M'hamed Bougara de Boumerdes
Faculté des Sciences
Département des Mathématiques

Mémoire de fin de cycle

En vue de l'obtention du Diplôme de MASTER en Mathématique et Informatique

Option :

Recherche Opérationnelle, Optimisation et Management Stratégique

Thème

Détection Des Anomalies Dans Les Flux Des Données De Réseaux LTE

Réalisé par :
M^r Amarouche Elhadi
M^r Chérifi Walid

Soutenu le : 29/09/2020

Devant le jury composé de :

Président :	*M^r M.* Bezoui	M.C.B, à L'UMBB-Boumerdes
Promotrice :	*M^me W.* Drici	M.C.B, à L'UMBB-Boumerdes
Examinateurs :	*M^r F.* Cheurfa	M.A.A, à L'UMBB-Boumerdes

Année universitaire 2019/2020

Table des matières

1 Introduction **12**

1.1 Contexte et motivation 12

1.2 Objectif 12

1.3 Architecture et principes généraux 12

1.3.1 Présentation de l'eNodeBs 15

1.3.2 Système de surveillance et de mesure en LTE [8] 15

1.4 Analyse basée sur les KPIs 16

1.5 Les données kpi[8] 17

1.5.1 Obtention des données 17

1.5.2 Pré-traitement des données 17

1.5.3 Extraction des séries temporelles 17

1.6 Présentation du groupe Algérie Telécom 18

1.6.1 Historique du groupe AT 18

1.6.2 Missions et objectifs du groupe 19

1.6.3 Les domaines d'activité d'AT 19

2 Présentation de la Problématique **20**

2.1 Quelles sont les anomalies ? 22

2.2 Les Défis[3] 23

2.3 Techniques de détection d'anomalies[3] 23

2.3.1 Techniques de détection des anomalies basées sur la classification 23

2.3.2 Techniques de détection des anomalies basées sur le plus proche voisin 24

2.3.3 Techniques de détection des anomalies basées sur le clustering (ou groupage) 25

2.3.4 Techniques de détection des anomalies statistiques 26

2.3.5 Techniques de détection des anomalies basées sur la théorique de l'information 27

2.3.6 Techniques de détection des anomalies basées sur l'analyse spectrale 27

2.4 Apprentissage automatique 27

2.4.1 Un domaine pluri-disciplinaire 28

2.4.2 L'apprentissage automatique et matières connexes 28
2.4.3 Plusieurs types de problèmes en apprentissage automatique . . 29
2.4.4 Apprentissage supervisé 29
2.4.5 Notations 29
2.4.6 Formalisation du problème 30
2.4.7 Protocol expérimental en apprentissage supervisée 30
2.4.8 Validation croisée 31
2.4.9 Mesures d'évaluation pour le problème de régression 31

3 Modèles Linéaires multiple et régressions pénalisées 32
3.1 Notions 32
3.2 Régression linéaire multiple[15] 33
3.2.1 Formulation matricielle / vectorielle 35
3.2.2 Limitations de la régression linéaire 36
3.3 Régression linéaire pénalisée 36
3.3.1 Interprétation géométrique 37
3.3.2 Solution analytique de la régression linéaire pénalisée 38
3.3.3 Choix du paramètre de régularisation λ 39
3.4 Algorithme descente par coordonnée 40
3.4.1 Optimiser le problème des moindres carrés par une coordonnée à la fois 42
3.4.2 Optimiser le problème de ridge par une coordonnée à la fois . . 43
3.4.3 Optimiser le problème de lasso par une coordonnée à la fois . . 43
3.5 Application d'algorithme et résultats de l'analyse 44
3.6 conclusion 47

4 Implémentation et résultats 48
4.1 Outils et Environnement 48
4.1.1 Python 48
4.1.2 Bibliothèques utilisées 49
4.2 Les données 49
4.3 Implémentation et résultats 51
4.3.1 Visualisation de données 51
4.3.2 Data Preprocessing 51
4.3.3 Feature Engineering 52
4.3.4 L'algorithmeLassoCV 53
4.4 Présentation de l'interface graphique 53

A Quelques rappels de calcul différentiel, analyse et optimisation convexe et extremum **58**

A.1 Problèmes d'optimisations : 58

A.1.1 Extremum local et global : 58

A.1.2 Matrice (semi) défini positive,(semi) défini négative : 58

A.1.3 Optimisation convexe 59

A.1.4 Quelques Notations 60

Table des figures

1.1 Éléments du réseau EPS 13
1.2 Des équipements peuvent dialoguer entre eux même s'ils ne sont pas directement physiquement interconnectés par une liaison : dialogue via le réseau IP 14
1.3 eNodeBs 15

2.1 Le seuil critique supérieur a été défini à 95%, ce qui signifie que le processeur d'utilisation plus élevée, cela entraînerait un comportement imprévisible et pourrait entraîner une interruption des opérations de l'utilisateur final 20
2.2 Exemple des données d'un opérateur de télécommunications observé sur 4 jours, le première et le quatrième jour sont le profile normal de données, par contre en voix deux anomalies entouré par deux cercles bleu dans le deuxième et le troisième jour. 21
2.3 Série temporelle sans anomalies 22
2.4 Série temporelle avec anomalies 22
2.5 Utilisation de la classification pour la détection d'anomalies. 24
2.6 Exemple de centroïde d'un cluster 25
2.7 illustration de la validation croisée 31

3.1 Solution du Lasso 37
3.2 Solution du ridge 37
3.3 cette figure est une présentation graphique des 10 séries d'études, la série verte représente "E-RAB : Active time, all", la série bleue représente E-RAB :DR, les séries rouge sont les 8 autres séries donc le lasso est une méthode de détection efficace 46
3.4 Erreur quadratique moyenne du Lasso 46
3.5 Erreur quadratique moyenne du ridge 46
3.6 Chemin de régularisation pour la régression Lasso 46
3.7 Chemin de régularisation pour la régression ridge. 46

4.1 Logo anaconda 48
4.2 Logo python 49
4.3 Données csv 50
4.4 visualiser les données à l'aide d'un Graphique linéaire nous a permet de remarquer une fort corrélation entre chaque 24 heurs donc une corrélation faible indiquera un comportement inormal 51

4.5 dans notre analyse basé sur la corréllation les valeurs manquée sont éliminer des données 52
4.6 Les séries incohérents sont éliminer car ne sont pas des indicateurs clés de performance 52
4.7 illustration de la forte corrélattion entre fenêtre de 24 heurs d'une série temporelle 53
4.8 La fenêtre de bienvenue. 54
4.9 Interface du résultat de problème. 54
4.10 Resultats d'exécution 1. Le Lasso détecte que "E-RAB : Active time, all" affecte E-RAB :DR 55
4.11 Résultats d'exécution2, Les deux figures représentent le chemin de régularisation pour la régression Lasso (i.e comportement des paramètres estimé pour un échantillon des valeurs de lambdas) et le chemin de l'erreur relative pour les mêmes valeurs de lambdas ce qui preuve la convergence de l'algorithme descente par coordonnée et l'efficacité du lasso dans l'élimination des variables non pertinent 55
4.12 Contrairement à la régression Lasso, la régression Ridge rétrécit les coefficients , mais ne les réduit pas à zéro 56

Liste des tableaux

1.1 Données KPI avant le prétraitement 17
1.2 Séries temporelles pour kpiA de l'enb1 et cell1 18
1.3 Séries temporelles pour kpiB de l'enb1 et cell1 18
1.4 Séries temporelles pour kpiC de l'enb1 et cell2 18

3.1 Nom des kpis lié aux séries da la table 3.2 de 1 à 9 44
3.2 Nous essayons de découvrir lesquel des enregistrements 1 : 9 a un effet sur la série d'anomalies E-RAB : DR 45
3.3 Ces résultats sont obtenus en appliquant Lasso et Ridge sur les données de table3.2, Lasso détecte une influence entre E-RAB :DR et la série 5, $\beta_5 = 7.647342e-17$, $\lambda = 232.573375$, par conter Ridge n'a éliminé aucune des séries avec meilleur $\lambda = 232.573375$ 45

4.1 Séries temporelles avant agrégation 50
4.2 Séries temporelles après agrégation 50

Acronymes

AA Apprentissage Automatique. 28
API Application Programming Interface. 48

eNodeBs Evolved node base station. 4, 12–16
EPC Evolved packet core. 12, 13
EPS Evolved packet sustème. 4, 13

HSS Home subscription server. 13, 14

IP Internet Protocol. 4, 14

KPIs Key performance indicators. 12, 16, 17, 20, 21, 50, 55

Lasso Least Absolute Shrinkage and Selection Operator. 32, 40
LassoCV Least Absolute Shrinkage and Selection Operator Cross Validation. 2, 48, 53
LTE Long term evolution. 12, 13, 15, 20, 50, 51

ML Machine Learning. 52
MME Mobility management entity. 12–14

OSS Operations Support Systems. 50

P-GW Packet data network gateway. 12–14
PCRF Policy and charging resource function. 13

QoS Quality of service. 13, 16

RAN Radio access network. 12

S-GW Serving gateway. 12–14

UE User equipments. 12–15

Remerciements

Nous voudrions tout d'abord remercier ALLAH le Tout-Puissant et Miséricordieux, qui nous a donné le courage, la force et la patience pour la réalisation de ce mémoire.

Nous tenons à exprimer notre profonde gratitude à madame **W.Drici**, *notre promotrice de mémoire, pour la confiance qu'il nous a faite en acceptant de diriger nos recherches, et pour ses précieux conseils et orientations, ainsi que pour l'intérêt particulier qu'il a accordé à ce travail.Nous la remercions pour sa grande contribution à l'aboutissement de ce travail.*

Nos remerciements s'adressent également à l'ensemble des enseignants du Département des Mathématiques et spécialement spécialité Recherche Opérationnelle.

Notre reconnaissance s'adresse particulièrement à monsieur **S.Hafiane** *Ingénieur radio chez Algérie Telecom, de nous avoir accueillir et de nous avoir proposé ce thème.*

Un grand merci à nos trés chers parents, qui nous ont aidé à suivre nos études dans les meilleures conditions et qui nous ont toujours soutenues et encouragé sans limite.

Nous remercions nos frères et soeurs pour leur encouragement, anisi que toute la famille. Sans oublier de remercier aussi tous nos collègues, nos amies et tous ceux qui ont contribué de prés ou de loin à la réalisation de ce travail.

Dédicaces de Walid

Á les plus beaux créatures que Dieu a créées sur terre.
Á ces sources de tendresse, de patience et de générosité.
Á mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études.
Á ma grand soeur et son marie Mohamed et leur fils Anis.
Á mes petites soeurs.
Á mon cher ami et binôme Amarouche Elhadi et sa famille.
Á tous mes amis et collégues
Á tous les étudiants de la promotion 2019/2020,
Option :Recherche Opérationnelle Optimisation et Management Stratégique (ROOMS).
Á tous ceux qui, par un mot, m'ont donné la force de continuer.
Je dédie ce travail.

Dédicaces de Elhadi

Á les plus beaux créatures que Dieu a créées sur terre.
Á ces sources de tendresse, de patience et de générosité.
Á mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études.
Á mon grand frère.
Á ma soeur et son marie Mohamed et leur deux fils, souhaib et iyad.
Á mes soeurs.
Á mon cher ami et binôme Chérifi Walid et sa famille.
Á tous mes amis et collégues
Á tous les étudiants de la promotion 2019/2020,
Option :Recherche Opérationnelle Optimisation et Management Stratégique (ROOMS).
Á tous ceux qui, par un mot, m'ont donné la force de continuer.
Je dédie ce travail.

Résumé

Les développements récents des systèmes industriels fournissent une grande quantité des données chronologiques provenant des capteurs, journaux, paramètres du système et mesures physiques, etc. les données sont extrêmement précieuses pour fournir des informations sur les systèmes complexes et pourraient être utilisé pour détecter des anomalies à étapes préliminaires. Cependant, les caractéristiques particulières de ces données chronologiques, telles que les dimensions élevées et les dépendances complexes entre les variables, ainsi que leur volume massif, posent des grands défis aux algorithmes de détection d'anomalies existants. Dans ce mémoire, nous proposons des modèles de régression linéaire, comme une approche évolutive pour la détection des anomalies dont les résultats peuvent être facilement interpréter. Plus précisément, le modèle linéaire LASSO est un modèle qui exploite la dépendance entre variables en appliquant une régularisation l_1 pour apprendre la causalité. Notre objectif est de calculer efficacement un score robuste de corrélation entre anomalies pour chaque variable via un modèle linéaire qui peut fournir des informations sur les raisons possibles d'anomalies. Nous évaluons l'efficacité de nos algorithmes proposés à la fois sur l'ensemble des données applicatifs. Les résultats montrent que l'algorithme LASSO atteint performances nettement meilleures que les autres algorithmes et est évolutif pour les applications à grande échelle.

Chapitre 1

Introduction

1.1 Contexte et motivation

Les progrès continus des technologies des réseaux cellulaires ont fait de l'accès Internet haut débit mobile une norme. Cependant, les réseaux cellulaires sont vastes et complexes par nature, et donc les réseaux cellulaires souffrent souvent de performances (dégradations ou défaillances) pour diverses raisons, telles que les interférences, les pannes de courant, les dysfonctionnements des éléments du réseau et la déconnexion des câbles. Il est donc essentiel de détecter et répondre aux anomalies des réseaux cellulaires en temps réel , afin de maintenir la fiabilité du réseau et améliorer la qualité de service des abonnés . Pour identifier les problèmes de performances dans les réseaux cellulaires, une pratique courante adoptée par les administrateurs de réseau est de surveiller un ensemble diversifié d'indicateurs clés de performance KPIs, lesquels produisent des mesures de données sous forme des séries temporelles qui quantifient des ressources et des aspects de performance des éléments de réseau , et donc la tâche principale est d'identifier toutes anomalies qui se réfèrent à des modèles inattendus qui se produisent à un instant unique ou sur une période de temps prolongée. Des études récentes proposent d'utiliser l'apprentissage automatique pour la détection d'anomalies dans les réseaux cellulaires (par exemple [28],[1]).

1.2 Objectif

Dans ce travail, nous allons se focaliser sur l'utilisation des techniques d'analyse de données dans le but de consolider le processus de résolution de défaillances dans les réseaux. Pour ce faire, il faut définir deux objectifs principaux : la détection d'anomalies en temps réel dans le flux de donnée et le diagnostic des causes racines de ces anomalies.

1.3 Architecture et principes généraux

Selon [22] LTE est une technologie de communication mobile également connue sous le nom de 4G.Un réseau LTE comprend trois entités principales : UE, RAN, et l'EPC. Chaque UE fait référence à un appareil mobile d'utilisateur. Le RAN comprend plusieurs stations de base appelés eNodeBs, dont chacun gère les ressources radio des UE et fournit aux UE la connectivité sans fil. l'EPC comprend la MME, le S-GW, et le P-GW :le MME gère le plan de fonctions de UE (par exemple, authentification

de l'utilisateur, gestion de mobilité), tandis que S-GW et P-GW gèrent les plan de fonctions de données de UE (par exemple, routage des données). Pour envoyer ou recevoir des données via Internet, un UE établit d'abord une connexion radio avec un eNodeBs et un canal de signalisation avec le MME. Il définit ensuite une session de données avec l'EPC au sommet de la connexion radio, et utilise la session de données pour la transmission de données. Policy and charging resource function PCRF Il prend des décisions sur la manière de gérer les services en termes de QoS et fournit des informations au P-GW. Home subscription server HSS est un serveur de base de données des abonnés, Le rôle du HSS est de communiquer avec le réseau et de fournir le profil d'abonné et les informations d'authentification. La base de données stocke des informations sur les abonnés pour aider à l'autorisation, les détails des appareils, ainsi que l'emplacement de l'utilisateur et les informations de service.

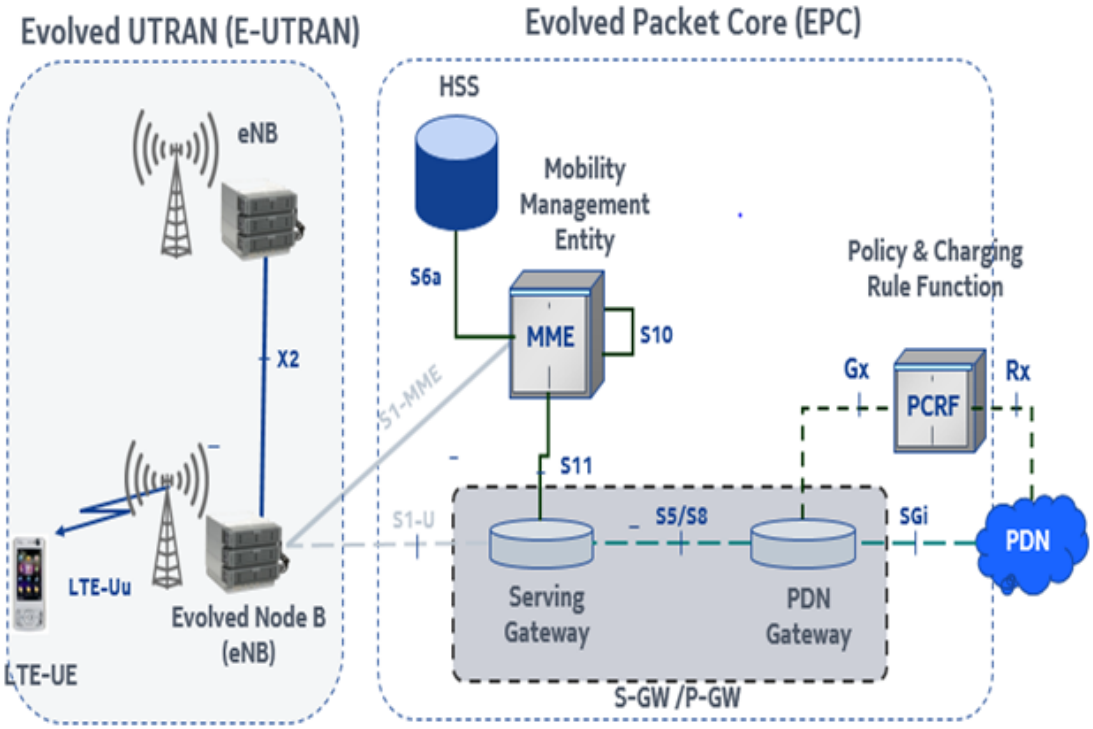


Figure 1.1 – Éléments du réseau EPS

Chaque eNodeBs dessert plusieurs zones géographiques appelées cellules, dont chacune couvre un certain nombre d'UE. La taille de chaque cellule dépend de la population d'utilisateurs locaux et du la couverture radio. Un réseau LTE couvre généralement des milliers de cellules.
Il existe plusieurs interfaces dans le réseau LTE, certaines des interfaces standard les plus connues sont répertoriées ci-dessous :

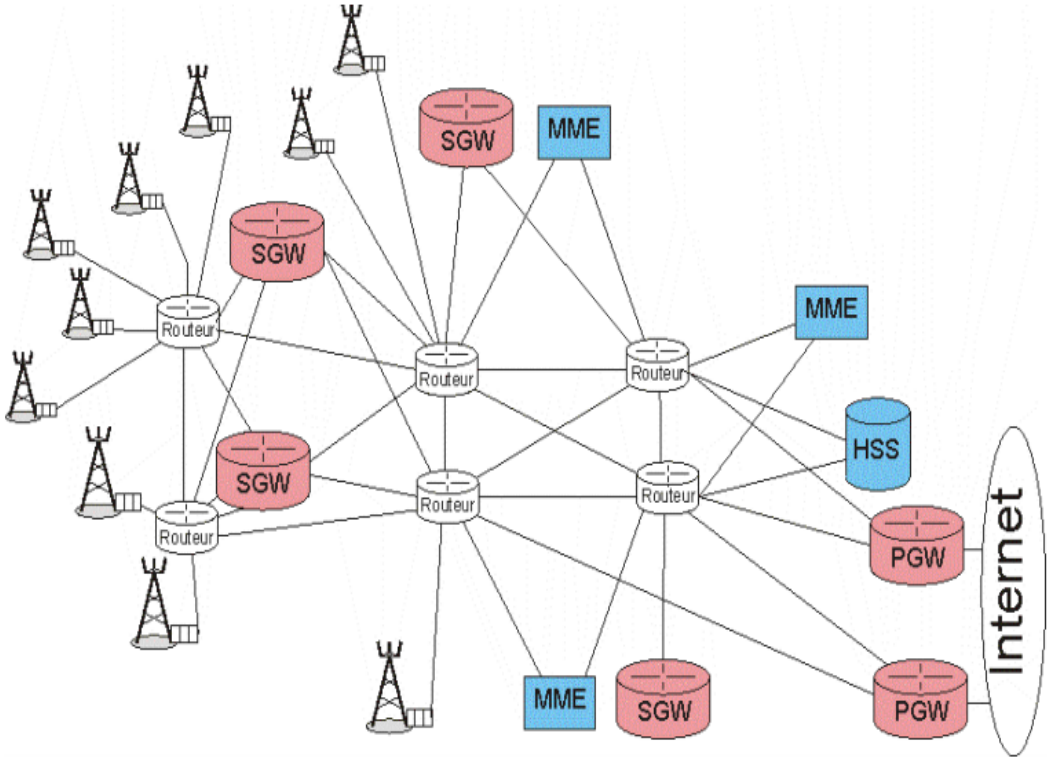


FIGURE 1.2 – Des équipements peuvent dialoguer entre eux même s'ils ne sont pas directement physiquement interconnectés par une liaison : dialogue via le réseau IP

- Interface SGi : entre le P-GW et le réseau IP externe (Internet)
- Interface S5 : entre le S-GW et le P-GW (d'un même réseau), Transport des données utilisateurs + quelques messages de signalisation
- Interface S11 : entre le S-GW et le MME, Transport des messages de signalisation
- Interface S6a : entre le MME et le HSS, Transport des messages de signalisation
- Interface S1-MME : entre l'eNodeB et le MME, Transport des messages de signalisation
- Interface S1-U : entre l'eNodeB et le S-GW, Transport des données utilisateurs, pas d'échange de signalisation
- Interface X2 : entre 2 eNodeBs, Transport des données utilisateurs et des messages de signalisation
- Interface Uu ou interface radio : entre le terminal UE et l'eNodeBs, Transport des données utilisateurs et des messages de signalisation
- Interface S8 : entre le S-GW et le P-GW d'un autre réseau

1.3.1 Présentation de l'eNodeBs

Figure 1.3 – eNodeBs

Un eNodeBs se compose d'une tour et de plusieurs antennes attachées à la tour. Dans de nombreux cas, trois antennes sont utilisées et elles sont séparées de 120 degrés. Dans les sites plus denses, plus d'antennes pourraient être utilisées pour augmenter la capacité et être en mesure de servir plus d'utilisateurs. Les antennes montées sur la tour recherchent en continu les UE à desservir. L'eNodeBs et l'UE sont appelés E-UTRAN .
Un eNodeBs fournit les deux fonctions principales suivantes :

— Envoie et reçoit une transmission radio vers tous les UE connectés en utilisant les fonctions de traitement du signal analogique et numérique de l'interface aérienne LTE

— Contrôle les opérations de niveau bas de tous les UE connectés en leur envoyant des messages de signalisation, tels que des commandes de transfert.

voire [8].

1.3.2 Système de surveillance et de mesure en LTE [8]

La surveillance et la mesure des performances facilitent grandement la tâche de l'opérateur pour atteindre ces objectifs et soutiennent l'opérateur dans de nombreuses tâches et processus connexes. La surveillance des performances permet de collecter des informations sur les éléments suivants :

— Intensité du trafic réseau

— Evénements se produisant à certains endroits du réseau (et à quelle fréquence se produisent-ils)

— Efficacité de la planification (c'est-à-dire si les instructions sont remplies ou lorsque des modifications supplémentaires sont nécessaires)

— Lieux où des pannes fréquentes sont signalées
— Comportement de l'abonné (s'il correspond au modèle supposé)

Grâce à la surveillance des performances, l'opérateur peut avoir une vision claire des performances, de la capacité et de la qualité du réseau et, espérons-le, améliorer la satisfaction et la fidélité de ses utilisateurs.
Domaines clés où la surveillance des performances peut être utilisée est :

— Planification du réseau
— Acceptation et vérification
— Vérification du modèle de trafic
— Analyse comparative
— Dépannage
— Surveillance de la QoS
— Optimisation du réseau

1.4 Analyse basée sur les KPIs

Dans cette section, l'analyse actuelle effectuée sur eNodeBs sur la base des données de compteur et de KPIs est expliquée [8].

Un opérateur de télécommunications pourrait facilement avoir plus de 10 000 éléments de réseau de stations de base radio (eNodeB). Une énorme quantité de données est collectée à partir de ces éléments de réseau. Les données collectées vont des journaux d'alarmes, des journaux d'événements, des fichiers de configuration et divers fichiers de compteur de surveillance des performances. Les outils existants sont utilisés pour analyser les fichiers bruts et le résultat de l'analyse est présenté à l'opérateur sous la forme d'un rapport basé sur le Web. L'énorme quantité de rapports et la quantité énorme d'informations dans chacun d'entre eux, rendent pratiquement impossible pour les opérateurs de trouver des problèmes dans leur réseau radio. Cela conduit à de nombreux cas où les problèmes ne sont pas détectés, ce qui pourrait entraîner une dégradation des services et éventuellement une perte de revenus.
Les éléments du réseau radio produisent une large gamme de données de compteurs qui pourraient être utilisées pour surveiller ses performances et sa qualité de service. Les compteurs sont regroupés en entités administratives appelées mesures. Les compteurs sont les éléments constitutifs des principaux indicateurs de performance (KPIs). Un est essentiellement une formule composée d'un ou de plusieurs compteurs. Les formules sous-jacentes pour le calcul des sont confidentielles et ne sont pas accessibles au public. Les KPIs sont utilisés pour créer des rapports de niveau supérieur, qui indiquent les performances et les fonctionnalités du réseau.
Les compteurs sont généralement collectés à des intervalles prédéfinis de 15, 30 ou 60 minutes et sur une longue période de temps (séries temporelles). Les experts définissent généralement des seuils pour les compteurs et les KPIs qui pourraient être utilisés pour déclencher des alarmes. Cependant, cette méthode n'est pas optimale car elle repose sur des valeurs fixes qui pourraient devenir non pertinentes avec une nouvelle tendance dans le trafic ou un changement dans la configuration du réseau.
Une alternative à l'analyse basée sur les seuils consiste à utiliser l'apprentissage automatique et en particulier les méthodes de détection des anomalies pour trouver les moments dans le temps où un compteur ou une valeur s'écarte de sa plage normale. La détection d'anomalie peut être appliquée à un compteur à la fois (analyse univariée) ou à plusieurs compteurs à la fois (analyse multivariée). Ce dernier fournira très

probablement des capacités de prédiction plus élevées car il examine la combinaison de nombreux aspects de l'élément de réseau en même temps.

1.5 Les données kpi[8]

1.5.1 Obtention des données

La collecte des données se fait au niveau de l'élément de réseau. Les compteurs sont enregistrés et agrégés à des intervalles prédéfinis (15 min, 60 min ou moins souvent). Il s'agit généralement d'un paramètre configurable qui peut varier d'un opérateur à l'autre. Chaque package quotidien pour un eNodeB contient plusieurs fichiers par période d'agrégation. Par exemple, il y aura 24 fichiers si la période d'agrégation est de 60 minutes.
Les fichiers sont traités dans un format tabulaire qui comprend les informations suivantes :

— La date et l'heur
— Nom et valeur du compteur
— L'ID de cellule dans l'eNodeB.

L'identifiant de cellule correspond à un module radio dans l'eNodeB qui est connecté à une antenne physique montée dans une tour. Après l'analyse, les données du compteur sont utilisées pour calculer un ensemble de définis à l'aide de formules définies par des experts technologiques pour mesurer certains aspects des performances et de la qualité de service de l'eNodeB.

1.5.2 Pré-traitement des données

Les données KPIs sont extraites de la base de données et plusieurs étapes de prétraitement sont appliquées aux données avant que l'analyse proprement dite ne soit effectuée. L'objectif du prétraitement est d'obtenir des données de séries temporelles pour chaque combinaison d'un KPI, d'un eNodeB et d'un ID de cellule agrégés à la même fréquence.

1.5.3 Extraction des séries temporelles

Les données KPI sont stockées dans un format tabulaire, chaque rangée présentant la valeur d'un KPI spécifique pour une cellule spécifique à un horodatage spécifique. Dans cette étape, on fait extraire toutes les valeurs qui appartiennent à la même cellule et le KPI dans un tableau séparé. Un exemple des données avant et après le prétraitement est présenté dans les tableaux suivants :

Timestamp	eNodeB	KPI ID	Cell ID	Value
T1	enb1	kpiA	cell1	1
T1	enb1	kpiB	cell1	2
T2	enb1	kpiA	cell1	1
T2	enb1	kpiB	cell1	2
T2	enb1	kpiC	cell2	3
T3	enb1	kpiA	cell1	1.1
T3	enb1	kpiB	cell1	2.1
T3	enb1	kpiC	cell2	3.1

Table 1.1 – Données KPI avant le prétraitement

Timestamp	Value
T1	1
T2	1
T3	1.1

Table 1.2 – Séries temporelles pour kpiA de l'enb1 et cell1

Timestamp	Value
T1	2
T2	2
T3	2.1

Table 1.3 – Séries temporelles pour kpiB de l'enb1 et cell1

Timestamp	Value
T2	3
T3	3.1

Table 1.4 – Séries temporelles pour kpiC de l'enb1 et cell2

1.6 Présentation du groupe Algérie Telécom

1.6.1 Historique du groupe AT

ALGERIE TELECOM est une société par actions au capital de 61 275 180 000 DA opérant sur le marché des réseaux et services de communications électroniques. Sa naissance a été consacrée par la loi 2000/03 du 5 août 2000, relative à la restructuration du secteur des Postes et Télécommunications, qui sépare notamment les activités Postales de celles des Télécommunications.

ALGERIE TELECOM est donc régie par cette loi qui lui confère le statut d'une entreprise publique économique sous la forme juridique d'une société par actions SPA, Entrée officiellement en activité à partir du 1er janvier 2003, elle s'engage dans le monde des Technologies de l'Information et de la Communication avec trois objectifs :

— Rentabilité .

— Efficacité .

— Qualité de service.

1.6.2 Missions et objectifs du groupe

L'ambition d'Algérie Telécom est d'avoir un niveau élevé de performances techniques, économiques et sociales pour se maintenir durablement comme leader dans son domaine, dans un environnement devenu concurrentiel. Son souci consiste, aussi, à préserver et développer sa dimension internationale et participer à la promotion de la société de l'information en Algérie L'activité majeure d'Algérie Télécom est de :

- Fournir des services de télécommunications permettant le transport et l'échange de la voix, des messages écrits, des données numériques, d'informations audiovisuelles,..
- Développer, exploiter et gérer les réseaux publics et privés de telecommunications.
- Etablir, exploiter et gérer les interconnexions avec tous les opérateurs des réseaux.

Algérie Télécom est engagée dans le monde des TIC(Technologies de l'Information et de la Communication) avec les objectifs suivants :

- Accroître l'offre de services téléphoniques et faciliter l'accés aux services de télécommunications au plus grand nombre d'usagers, en particulier en zones rurales.
- Accroître la qualité de services offerts et la gamme de prestations rendues et rendre plus compétitifs les services de télécommunications.
- Développer un réseau national de télécommunications fiable et connecté aux autoroutes de l'information.

Les responsabilités d'AT s'exercent dans les trois domaines suivants :

- Les actionnaires : AT doit mériter leurs soutiens en valorisant leurs patrimoines ;
- Les clients : AT doit anticiper leurs besoins en leur fournissant des produits et des services de qualité afin de gagner et de conserver leurs confiances ;
- Le personnel : AT doit satisfaire ses attentes en organisant les conditions de l'épanouissement.

1.6.3 Les domaines d'activité d'AT

La société AT est l'acteur majeur des télécommunications en Algérie avec cinq domaines d'activités :

- **Téléphonie fixe** : avec deux millions de lignes en service et un réseau WLL en plein expansion.
- **Téléphonie mobile** : activité au travers d'une filiale Mobilis, qui détient une part de marché de 13%
- **Transmission de données** : une activité de réseaux de données pour les entreprises (X25...)
- **Accés Internet à travers** : DJAWEB, FAWRI ADSL et derniérement EASY ADSL.
- **Réseau satellitaire** : des services de télécommunications s'appuyant sur VSAT, Inmarsat le réseau Thuraya.

Chapitre 2

Présentation de la Problématique

Ce chapitre traite le problème général de la détection d'anomalies, les défis et les techniques communément connues, et enfin une section qui sera discutée sur l'apprentissage automatique, et comme un cas particulier l'apprentissage supervisé qui est couvert dans ce mémoire sera présenter pour un problème de régression.Quel algorithme de régression ? et comment l'utiliser il faut voir le chapitre théorique 3.

Comme vu dans le chapitre 1 la détection des anomalies dans le réseau LTE est basé sur la surveillance d'un ensembles d'indicateurs clés de performance . Les experts ont spécifié des seuils pour le surveiller des KPIs du processeur d'utilisation. Si le processeur d'utilisation dépasse l'une des valeurs de seuil, une action de l'équipe de maintenance doit être entreprise.

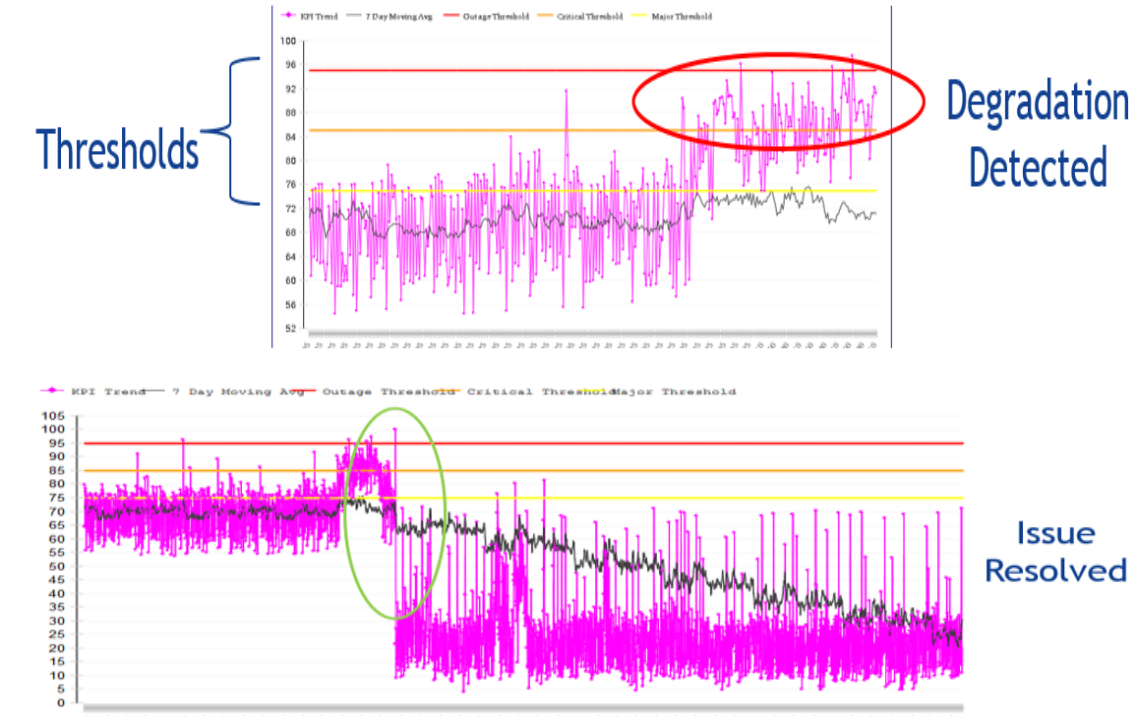


FIGURE 2.1 – Le seuil critique supérieur a été défini à 95%, ce qui signifie que le processeur d'utilisation plus élevée, cela entraînerait un comportement imprévisible et pourrait entraîner une interruption des opérations de l'utilisateur final

Comme le montrent les cas d'utilisation en direct, l'analyse basée sur des seuils a

été très efficace pour détecter un problème réel dans le réseau. La réaction rapide de l'équipe de maintenance et le déploiement d'un correctif rapide pourraient résoudre les problèmes et garantir que l'impact est minimisé et que les opérations normales sont reprises. Cependant, définir des seuils pour les indicateurs clés de performance n'est généralement pas un travail trivial et nécessite une connaissance approfondie de la technologie en général et de la configuration spécifique à l'opérateur. Les seuils sont définis au niveau des KPIs et les problèmes qui pourraient être causés par des changements dans plusieurs KPIs ne sont pas détectés, comme les performances de plusieurs cellules dans différentes métriques KPIs ne respectent pas le seuil d'exigence fixe et qu'il est pratiquement difficile de fixer simultanément toutes les cellules de toutes les métriques, la détection manuelle et subjective rencontre des difficultés pour hiérarchiser les cellules pour les actions correctives. Ainsi, les anomalies restent non corrigées plus longtemps et deviennent une cause d'anomalies de plus grande ampleur et de pannes complètes du réseau. Le processus de détection manuelle est sujet aux erreurs, ainsi les anomalies détectées soulèvent des questions de crédibilité dans différentes régions de l'entreprise et deviennent parfois source de conflits entre departments. En résumé, les principaux problèmes sont :

— Le seuil dur et fixe ne tient pas compte de la nature dynamique des cellules car il est fixe.

— La détection manuelle effectuée sur la base des valeurs de performance rencontre des difficultés pour inclure les valeurs de performances de l'historique.

— La méthode de détection manuelle des anomalies est sujette à des erreurs, elle peut donc conduire à une analyse erronée et laisser les anomalies ne pas être corrigées.

Donc le principal défi pour résoudre le problème de la dégradation cellulaire est de crée une méthode robuste pour la modélisation de comportement cellulaire normal. Cette approche utilise les indicateurs clés de performances (KPIs), qui sont des mesures collectées sous forme de séquencesde valeurs ordonnées d'une variable à intervalles de temps également espacés, ils constituentune série temporelle et peut être analysé avec des méthodes et algorithmes d'apprentissage automatique connues pour l'analyse des séries temporelle[14] .Cependant, la détection d'anomalies basée sur l'apprentissage automatique est soumise à plusieurs défis bien connus [3],[16].
Une anomalie dans une série temporelle peut être soit une seule observation ou une sous-séquence d'une série temporelle par rapport à une série temporelle normale.

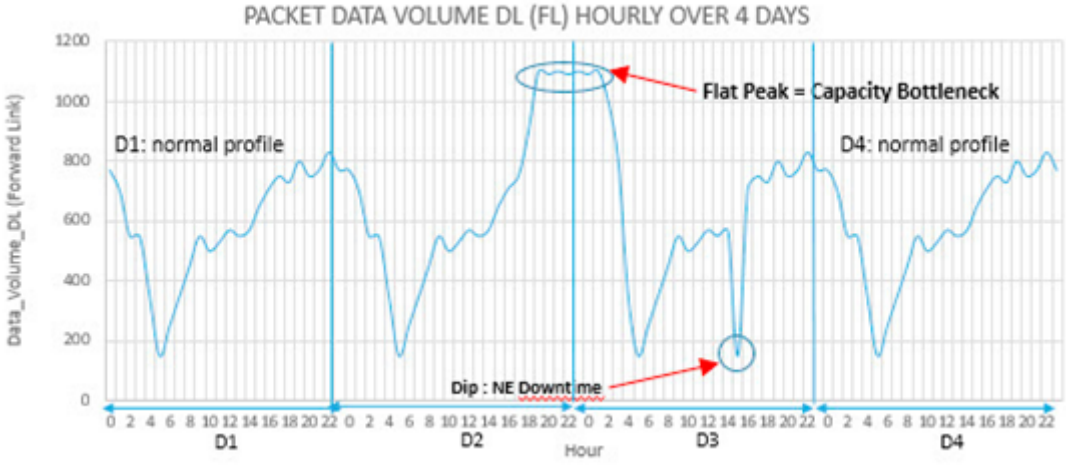


Figure 2.2 – Exemple des données d'un opérateur de télécommunications observé sur 4 jours, le première et le quatrième jour sont le profile normal de données, par contre en voix deux anomalies entouré par deux cercles bleu dans le deuxième et le troisième jour.

Donc ont peut définir la détection des anomalies comme un processus d'identification des éléments ou événements inattendus dans les ensembles de données, qui repose sur deux hypothèses de base :

— Les anomalies ne se produisent que très rarement dans les données.

— caractéristiques diffèrent considérablement des instances normales.

2.1 Quelles sont les anomalies ?

Prenons cet exemple de données sous forme d'une série temporelle, nous remarquons qu'il y a une tendance qui monte et descend.

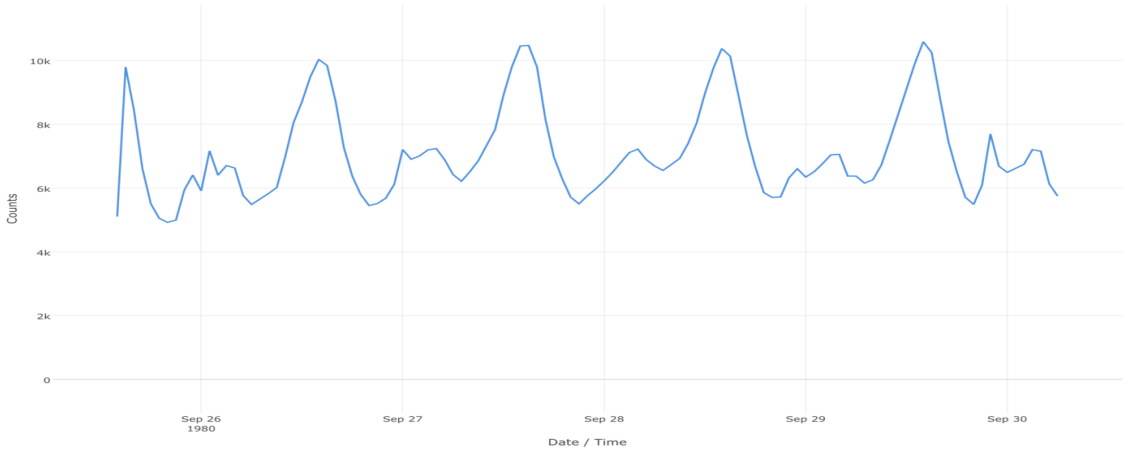


Figure 2.3 – Série temporelle sans anomalies

Mais alors, nous pourrions voir de gros sauts ou des baisses qui sont inhabituels de temps en temps, comme ceux avec le point rouge ci-dessous.

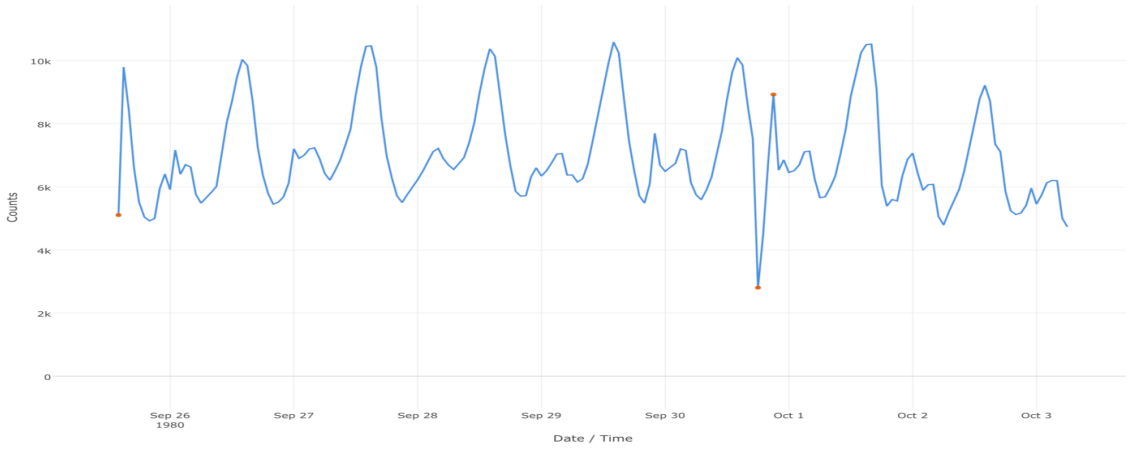


Figure 2.4 – Série temporelle avec anomalies

La chose délicate ici est que ce type de données avec tendance monte et descend généralement chaque jour, semaine ou toute période. Par exemple, disons que nous avons eu une énorme visite de pages à la fin du mois dernier. Mais si nous avons de tels sauts à chaque fin de mois, alors ce n'est qu'une tendance mensuelle et c'est en fait une chose normale. Ou, si vous avez une entreprise de vente au détail, votre site Web s'attendra probablement à un trafic plus important pendant la période des fêtes, ce que vous attendez chaque année et ne le considérerez pas comme une anomalie à moins qu'il

soit beaucoup plus important que les autres saisons de vacances dans le passé. . Vous souhaitez donc prendre en compte le schéma général de la tendance sous-jacente ainsi que la «saisonnalité» avant de déterminer si le trafic plus important que d'habitude est vraiment une anomalie ou non.

2.2 Les Défis[3]

À un niveau abstrait, une anomalie est définie comme un motif non conforme à un comportement normal attendu. Une approche simple de détection des anomalies, par conséquent, est de définir une région représentant un comportement normal et de déclarer toute observation dans les données qui n'appartiennent pas à cette région normale comme une anomalie. Mais plusieurs facteurs font de cette approche apparemment simple très difficile :

— Définir une région normale qui englobe tous les comportements normaux possibles est très difficile. De plus, la frontière entre comportement normal et comportement anormal est souvent pas précis. Ainsi, une observation anormale située près de la frontière peut en fait être normal, et vice versa.

— Dans de nombreux domaines, le comportement normal continue d'évoluer et une notion actuelle de normal le comportement pourrait ne pas être suffisamment représentatif à l'avenir.

— La disponibilité de données étiquetées pour l'entrainement / validation des modèles utilisés par la détection d'anomalies les techniques sont généralement un problème majeur.

— Souvent, les données contiennent du bruit qui tend à être similaire aux anomalies réelles et il est donc difficile de distinguer et de supprimer.

En raison de ces défis, le problème de détection des anomalies, dans sa forme la plus générale, n'est pas facile à résoudre. En fait, la plupart des techniques de détection d'anomalies existantes résolvent une formulation spécifique du problème. La formulation est induite par divers facteurs tels que la nature des données, la disponibilité des données, le type d'anomalies à détecter, etc. Souvent, ces facteurs sont déterminés par le domaine d'application dans lequel les anomalies doivent être détectées. Les chercheurs ont adopté des concepts de diverses disciplines comme les statistiques, l'apprentissage automatique, l'exploration de données, la théorie de l'information, la théorie de spectre, et ils les ont appliqués à des formulations de problèmes spécifiques.

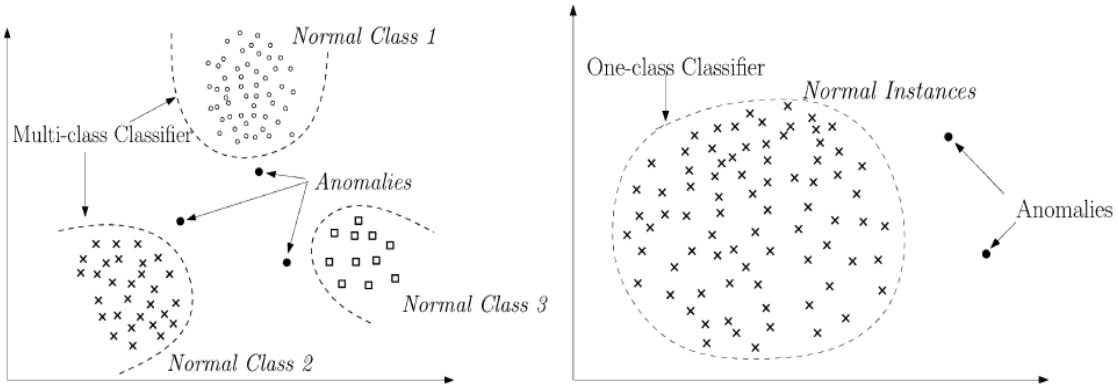
2.3 Techniques de détection d'anomalies[3]

2.3.1 Techniques de détection des anomalies basées sur la classification

Cette approche suppose que les points de données peuvent être regroupés en classes. Il y a deux possibilités de visualiser les anomalies :

— soit comme des points dispersés loin d'un centre dense de groupe de points de données normales (classification en une classe) ;

— ou sous forme de groupes de données denses loin des groupes de données normales (classification multi-classes).



(a) Multi-class Anomaly Detection (b) One-class Anomaly Detection

FIGURE 2.5 – Utilisation de la classification pour la détection d'anomalies.

Les techniques de classification supposent que nous avons un ensemble de données d'entrainement. Dans cette approche, le processus de détection d'anomalies peut être divisé en les étapes suivantes : Premièrement, classer les données d'entrainement et identifier les attributs de classification. Ensuite, nous apprenons un modèle en utilisant les données d'entrainement. Enfin, nous pouvons classer de nouvelles données en utilisant le modèle d'apprentissage. Pour détecter des anomalies, de nombreux algorithmes de classification ont été utilisés :

— Le SVM(Support Vector Machines)[17]
— les réseaux de neurones[18]
— Les réseaux bayésiens[26]
— l'apprentissage des règles qui capturent le comportement normal d'un système[25]

2.3.2 Techniques de détection des anomalies basées sur le plus proche voisin

Le concept d'analyse du plus proche voisin a été utilisé dans plusieurs techniques de détections d'anomalies[5]. Ces techniques sont basées sur l'hypothèse clé suivante : '' Des instances de données normales se produisent dans des quartiers denses, tandis que des anomalies se produisent loin de leurs voisins les plus proches."
Les techniques de détection des anomalies basées sur le voisin le plus proche nécessitent une distance ou une mesure similaire définie entre deux instances de données. La distance (ou similitude) entre deux instances de données peuvent être calculées de différentes manières :

— Pour les attributs continue, la distance euclidienne est un choix populaire, mais d'autres mesures peuvent être utilisées
— Pour les attributs catégoriels, un simple coefficient d'appariement est des mesures de distance souvent utilisées mais plus complexes peuvent également être utilisées
— Pour les instances de données multivariées, la distance ou la similitude est généralement calculé pour chaque attribut, puis combiné

Les techniques de détection des anomalies basées sur le voisin le plus proche peuvent être regroupées deux catégories :

1. Techniques qui utilisent la distance d'une instance de données à son k ième plus proche voisin comme le score d'anomalie ;

2. techniques qui calculent la densité relative de chaque instance de données pour calculer le score d'anomalie.

2.3.3 Techniques de détection des anomalies basées sur le clustering (ou groupage)

Le clustering est utilisé pour regrouper des instances de données similaires en cluster. Le clustering est principalement une technique non supervisée bien que semi-supervisée, Le clustering et la détection des anomalies semblent être fondamentalement différentes les unes des autres, plusieurs techniques de détection d'anomalies basées sur le clustering ont été développées. Les techniques de détection des anomalies le clustering peuvent être regroupées en trois catégories :

La première catégorie

repose sur l'hypothèse suivante : '' Les instances de données normales appartiennent à un cluster , tandis que les anomalies n'appartiennent à aucun cluster."
Les techniques basées sur cette hypothèse appliquent des algorithmes de clustering pour la reconnaissance des données qui déclarer toute instance de données qui n'appartient à aucun cluster comme anormal, tel que DBSCAN[4], ROCK et le SNN[24]. L'algorithme Find Out est une extension de l'algorithme Wave Cluster dans laquelle les clusters détectés sont supprimés des données et les instances résiduelles sont déclarées comme anomalies.
Un inconvénient de ces techniques est qu'elles ne sont pas optimisées pour trouver des anomalies, car l'objectif principal de l'algorithme de clustering est de trouver des clusters.

La deuxième catégorie

repose sur l'hypothèse suivants :" Les données normales se trouvent près du centroïde de leur cluster, tandis que les anomalies sont loin de centroïde ."
On appelle centroïde d'un cluster le barycentre des points de ce cluster :$\mu_k = \frac{1}{C_k} \sum_{C_k} x$

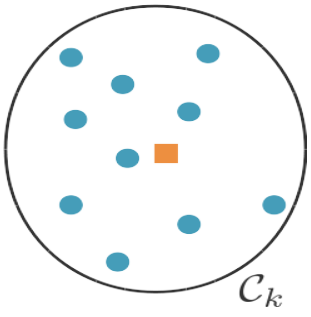


Figure 2.6 – Exemple de centroïde d'un cluster

Les techniques basées sur cette hypothèse consistent en deux étapes. Dans un premier temps, les données sont regroupé en utilisant un algorithme de regroupement. Dans la deuxième étape, pour chaque instance de données, la distance à son centroïde le plus proche est calculé pour faire la différence entre les données normale et l'anomalie. Un certain nombre de techniques de détection d'anomalies qui suivent cette approche en deux étapes ont été proposé en utilisant différents algorithmes de clustering : le clustering K-means et la maximisation des attentes (EM : Expectation Maximization) pour

regrouper les données d'apprentissage, puis utiliser les clusters pour classer les données de test.
Notez que si les anomalies dans les groupes de données forment des clusters par elles-mêmes, ces techniques ne sera pas capable de détecter de telles anomalies. Pour résoudre ce problème, une troisième catégorie des techniques basées sur le clustering ont été proposées.

La troisième catégorie

repose sur l'hypothèse suivante :" Les instances de données normales appartiennent à des clusters larges et denses, tandis que les anomalies appartiennent à des clusters petites ou clairsemée.
Les techniques basées sur cette hypothèse déclarent des instances appartenant à des clusters dont la taille et / ou la densité qu'est inférieure à un seuil, comme une anomalie. Plusieurs variantes techniques ont été proposées : le CBLOF (Cluster-Based Local Outlier Factor) , qui capture la taille du cluster auquel appartient l'instance de données, ainsi que la distance de l'instance de données au centroïde de son cluster.

2.3.4 Techniques de détection des anomalies statistiques

Le principe des technique de détection d'anomalies statistiques est : «Une anomalie est une observation qui est soupçonnée d'être partiellement ou totalement hors de propos parce qu'il n'est pas généré par le modèle stochastique supposè. Les techniques de détection d'anomalies statistiques sont basées sur l'hypothèse clé suivante :" les instances de données normales se produisent dans des régions à forte probabilité d'un modèle stochastique, tandis que les anomalies se produisent dans les régions à faible probabilité du modèle stochastique.
Les techniques statistiques adaptent un modèle statistique pour le comportement normal des données et ensuite appliquer un test de statistique inférence pour déterminer si une instance appartient à ce modèle ou non. Les instances dont la probabilité de génération est faible dans le modèle appris, sont déclarés comme des anomalies.
Des techniques paramétriques et non paramétriques ont été appliquées pour s'adapter à un modèle. Alors que les techniques paramétriques supposent la connaissance de la distribution et estime les paramètres à partir des données , les techniques non paramétrique ne supposent généralement pas la connaissance de la distribution sous-jacente.
Dans les deux sous-sections suivantes, nous discuterons des techniques paramétriques et non paramétriques de détection d'anomalies.

Les techniques paramétriques

Modèle gaussien :
Ces techniques supposent que les paramètres ont une distribués guassien. Dans ce type de techniques, les paramètres sont estimés par le maximum de vraisemblance estimé (MLE : Maximum Likelihood Eestimated). Une déviation est appliqué aux scores d'anomalies pour classer les anomalies.
Modèle de régression :
La détection d'anomalies à l'aide de la régression a été appliquée à de nombreux aspects tels que les modèles de régression linéaire, la détection de valeurs aberrantes à haute dimension, les données catégoriques ou mixtes et les données de séries chronologiques.
La technique de détection des anomalies basée sur un modèle de régression de base

comprend deux étapes. La première phase consiste à apprendre un modèle de régression à l'aide des données. Dans la deuxième étape, le résidu, la partie non expliquée par le modèle de régression, pour chaque instance de test est utilisé pour déterminer le score d'anomalie. Avec une certaine confiance, la quantité de résidu peut être utilisée comme score d'anomalie pour les données de test.

Techniques non paramétriques

L'histogramme :
La technique statistique non paramétrique la plus simple consiste à utiliser histogrammes pour maintenir un profil des données normales. Ces techniques sont également mentionnées comme basé sur la fréquence ou basé sur le comptage.
Fonction noyau(ou Kernel Function)
Une technique non paramétrique pour l'estimation de la densité de probabilité. Cela implique l'utilisation des fonctions du noyau pour approximer la densité réelle. Techniques de détection d'anomalies basées sur la fonction noyau est similaires aux méthodes paramétriques décrites précédemment. La seule différence est la technique d'estimation de densité utilisée parmi les :ACM Computing.

2.3.5 Techniques de détection des anomalies basées sur la théorique de l'information

Les techniques de la théorie de l'information analysent le contenu informationnel d'un ensemble de données en utilisant différentes mesures théoriques de l'information telles que la complexité de Kolomogorov, l'entropie, entropie relative, etc. Ces techniques sont basées sur l'hypothèse clé suivante :" Les anomalies dans les données induisent des irrégularités dans le contenu des informations de l'ensemble de données"[13].

2.3.6 Techniques de détection des anomalies basées sur l'analyse spectrale

Les techniques spectrales essaient de trouver une approximation des données en utilisant une combinaison des attributs qui captent la majeure partie de la variabilité des données. Ces techniques sont sur la base de l'hypothèse clé suivante :" Les données peuvent être intégrées dans un sous-espace de dimension faible dans lequel les cas normaux et les anomalies semblent sensiblement différents".Parmi ces techniques : Principal Component Analysis (PCA)[12], Fast Fourier Transform (FFT)[23], The Wavelet transform et Hough transform

2.4 Apprentissage automatique

L'apprentissage automatique (en anglais : machine learning), apprentissage artificielle ou apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'apprendre à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes.

L' apprentissage automatique a quatre classes d'applications communes : la classification, la prévision de la valeur suivante, la détection d'anomalies et la découverte de la structure. Parmi eux, la détection des anomalies fait référence au problème de trouver des modèles dans les données qui ne sont pas conformes au comportement attendu. Ces modèles non conformes sont souvent appelés anomalies, valeurs aberrantes, observations discordantes, exceptions, aberrations, surprises, particularités ou contaminants dans différents domaines d'application. Parmi ceux-ci, les anomalies et les valeurs aberrantes sont deux termes utilisés le plus souvent dans le contexte de la détection d'anomalies ; parfois de manière interchangeable. La détection d'anomalies trouve une large utilisation dans une grande variété d'applications comme la détection des fraudes pour les cartes de crédit, les assurances ou les soins de santé, la détection des intrusions pour la cybersécurité, la détection de pannes dans les systèmes critiques pour la sécurité et la surveillance militaire pour les activités ennemies.

En quoi consiste l'apprentissage automatique ?

L'apprentissage automatique consiste alors à programmer des algorithmes permettant d'apprendre automatiquement de données et d'expériences passées, un algorithme cherchant à résoudre au mieux un problème considéré.

2.4.1 Un domaine pluri-disciplinaire

L'apprentissage automatique (AA) est à la croisée de plusieurs disciplines :

- Les statistiques : pour l'inférence de modèles à partir de données.
- Les probabilités : pour modéliser l'aspect aléatoire inhérent aux données et au problème d'apprentissage.
- L'intelligence artificielle : pour étudier les tâches simples de reconnaissance de formes que font les humains (comme la reconnaissance de chiffres par exemple), et parce qu'elle fonde une branche de l'AA dite symbolique qui repose sur la logique et la représentation des connaissances.
- L'optimisation : pour optimiser un critère de performance afin, soit d'estimer des paramètres d'un modèle, soit de déterminer la meilleure décision à prendre étant donné une instance d'un problème.
- L'informatique : puisqu'il s'agit de programmer des algorithmes et qu'en AA ceux-ci peuvent être de grande complexité et gourmands en termes de ressources de calcul et de mémoire.

2.4.2 L'apprentissage automatique et matières connexes

Quelques références et domaines d'application faisant intervenir l'AA :

- Les statistiques ("Statistical Machine Learning") : modèles d'AA traités sous l'angle des statistiques [20].
- L'intelligence artificielle ("Artificial Intelligence") : modèles d'AA mettant l'accent sur le raisonnement, l'inférence et la représentation des connaissances.
- La fouille de données ("Data Mining") : lorsque les objets étudiés sont stockés dans des bases de données volumineuses.
- La reconnaissance de formes ("Pattern Recognition") : lorsque les objets concernés sont de type "signal" comme les images, les vidéos ou le son.

— Le traitement automatique du langage ("Natural Langage Processing" : NLP) : lorsque les problèmes concernent l'analyse linguistique de textes.

Plus récemment :

— La science des données ("Data science") : approche pluri-disciplinaire pour l'extraction de connaissances à partir de données hétérogènes.

— Les données massives ("Big data") : mettant l'accent sur les problématiques (volume, variété, vélocité, véracité) et des éléments de solutions issus du stockage/calcul distribué.

2.4.3 Plusieurs types de problèmes en apprentissage automatique

Apprentissage automatique supervisé :

On dispose d'un ensemble d'objets et pour chaque objet une valeur cible associée, il faut apprendre un modèle capable de prédire la bonne valeur cible d'un objet nouveau.

Apprentissage automatique non supervisé :

On dispose d'un ensemble d'objets sans aucune valeur cible associée, il faut apprendre un modèle capable d'extraire les régularités présentes au sein des objets pour mieux visualiser ou appréhender la structure de l'ensemble des données.

Dans le cadre de ce mémoire, nous étudierons les problèmes d'apprentissage supervisé : il s'agit donc de définir et d'estimer des modèles de prédiction étant donné un ensemble d'objets et leurs valeurs cibles respectives. On parle également d'algorithmes d'apprentissage supervisé dans le cadre d'un problème de régression.

2.4.4 Apprentissage supervisé

Il existe deux types de sous-problèmes en apprentissage supervisé :

— **Régression :** lorsque la valeur cible à prédire est continue.

— **classification ou catégorisation :** lorsque la valeur cible à prédire est discrète.

Par ailleurs nous supposerons également que les objets étudiés qui peuvent être complexes à l'origine (comme des données mutimédia) sont représentés dans un format numérique structuré. En d'autres termes :

— On représente un objet X_i par un vecteur noté x_i défini dans un espace de description composé de plusieurs variables.

— A chaque x_i on lui associe une valeur cible notée y_i .

2.4.5 Notations

Comme données à notre disposition nous supposerons que nous avons une table X avec n lignes et p colonnes et un vecteur colonne (variable cible) y de n éléments.

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

— La ligne i de X est associée à l'objet X_i et l'ensemble des objets $\{X_1, ..., X_n\}$ sera noté $\mathbb{O}$.

— La colonne j de X est associée à la variable ou attribut X^j et l'ensemble des variables $\{X^1, ..., X^p\}$ sera noté $\mathbb{A}$.

— x_{ij} terme général de X est la valeur de la variable X_j pour l'objet X_i .

— A chaque objet X_i est associé une valeur y_i de la variable $Y \in \mathbb{Y}$ où Y est l'ensemble des valeurs que peut prendre Y.

— Chaque objet X_i est associé à un vecteur numérique x_i appartenant à un espace de description $\mathbb{X}$.

2.4.6 Formalisation du problème

— Etant donné un ensemble d'entraînement $\mathbb{E}$, on cherche à déterminer $f : \mathbb{X} \to \mathbb{Y}$ une fonction modélisant la relation entre les X décrits dans l'espace de représentation $\mathbb{X}$ et la variable cible Y :

$$f(X) = Y$$

— En revanche, ne connaissant pas la vraie nature de la relation entre X et Y et les données observées en $\{X^1, ..., X^p\}$ étant soit bruitées, soit incomplètes, il n'est pas raisonnable de supposer une relation déterministe. Aussi, il est davantage raisonnable de poser le problème en les termes suivants :

$$f(X) = Y + \epsilon$$

où ϵ est l'erreur ou le résidu.

— Autrement dit, il s'agit d'approximer f en commettant le moins d'erreurs possibles sur $\mathbb{E}$ tout en faisant de bonnes prédictions pour des valeurs de $\mathbb{X}$ non encore observées.

2.4.7 Protocol expérimental en apprentissage supervisée

Etant donné une tâche d'apprentissage supervisé, le but est donc d'estimer plusieurs modèles afin de prédire au mieux la variable cible pour des données futures. Pour sélectionner le modèle, il faut procéder en distinguant au moins deux ensembles de données :

1. Un ensemble des données d'apprentissage ou d'entraînement $\mathbb{E}$ à partir duquel on estime une ou plusieurs fonctions de prédiction appartenant à un ou plusieurs espaces d'hypothèses.
2. Un ensemble de données de validation noté $\mathbb{V}$ qui n'est pas utilisé lors de l'estimation des modèles et qui sert à mesurer l'erreur de prédiction des differents modèles appris.

C'est l'erreur de prédiction mesurée sur $\mathbb{V}$ qui permet en pratique de sélectionner le meilleur modèle $\tilde{f}^*$.

— En revanche, si l'on souhaite avoir une estimation de l'erreur en généralisation de $\tilde{f}^*$ alors on ne peut pas utiliser celle mesurée à l'aide de $\mathbb{V}$. On a recourt à un troisième jeu de données appelé ensemble de données de test et noté $\mathbb{T}$.

2.4.8 Validation croisée

Précédemment on a supposé les données annotées séparées en $\mathbb{E}$ et $\mathbb{T}$. Mais l'estimation de l'erreur de prédiction est plus précise si on avait à disposition plusieurs ensembles $\mathbb{E}$ et $\mathbb{T}$.
La validation croisée consiste à :

— Séparer aléatoirement l'ensemble des données annotées en k sous-ensembles.

— Utiliser un sous-ensemble comme ensemble de test $\mathbb{T}$.

— Utiliser l'union des $k-1$ sous-ensembles restants comme ensemble d'entraînement $\mathbb{E}$.

En changeant chaque fois l'ensemble de validation, on voit qu'une k validation croisée permet d'avoir k paires d'échantillons $(\mathbb{E}, \mathbb{T})$ et ainsi k estimations de l'erreur de prédiction. On moyenne l'ensemble des k mesures d'erreurs afin d'avoir une estimation plus robuste de l'erreur de prédiction.

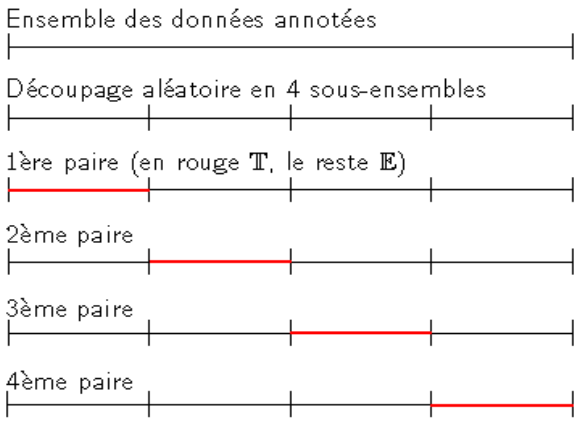


Figure 2.7 – illustration de la validation croisée

2.4.9 Mesures d'évaluation pour le problème de régression

— La Somme des carrés des résidus ou les Moindres Carrés Ordinaires ("Residual Sum of Square") :

$$scr(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

— La Moyennes des carrés des résidus ("Mean Squared Error") :

$$mse = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

— Contrairement au scr, le mse permet de comparer les erreurs de prédiction mesurés sur des ensembles de données de tailles différentes.

— La moyenne des résidus en valeurs absolues ("Mean Absolute Error") :

$$mae = \frac{1}{n} \sum_{i=1}^{n} |y_i - f(x_i)|$$

Chapitre 3

Modèles Linéaires multiple et régressions pénalisées

En statistique, l'un des principaux objectifs est de construire un modèle qui représente mieux un jeu de données, ce processus inclut la tâche de sélection des caractéristiques. Le seul objectif du chercheur est de construire un modèle qui décrit une variable d'intérêt, pour ce faire, l'une des premières questions que le chercheur devrait être en mesure de répondre : quelles caractéristiques / variables dois-je prendre en considération ? ou Quels sont les attributs les plus importants pour décrire la variable d'intérêt. ?.
Ce chapitre de recherche vise à répondre à ces questions montrant le processus de sélection des variables explicatives et description de l'une des méthodes possibles pour accomplir cette tâche. En particulier, l'accent est mis sur la sélection des fonctionnalités à l'aide du la méthode de pénalisation Lasso . Le but de ce chapitre est de décrire la méthode Lasso. La méthode Lasso sera analysée pour les modèles linéaires. De plus, afin de tester l'efficacité de Lasso, la méthode va être appliqués à des données réelles, et les résultats seront analysés et décrits mais avant de commencer, il est nécessaire d'introduire quelques notions.

3.1 Notions

Sélection des variables explicatives

La sélection des variables explicatives[15] est le processus de choisir un nombre réduit de ces variable pour décrire une variable d'intérêt. Où les principales raisons pour lesquelles la sélection est utilisée sont :

- rendre le modèle plus facile à interpréter, en supprimant les variables qui sont redondantes et n'ajoutent aucune information.
- réduire la taille du problème pour permettre aux algorithmes de fonctionner plus rapidement, permettant de manipuler des données de grande dimension.
- réduire le surajustement.

La sélection de variables est encore plus importante pour les ensembles de données de grande dimension, ici le nombre de traits est très élevé, parfois supérieur au nombre d'observations. Dans ces situations, il est difficile de dire facilement lequel des les variables sont pertinentes et celles qui ne le sont pas, et d'autre part, il est difficile, en raison de problèmes de dimensionnalité, de construire et d'interpréter un modèle qui prend en considération toutes les variables. Pour ces raisons, la sélection des variables

explicatives est une tâche importante.
Dans la littérature, il existe plusieurs types de méthodes pour compléter la tâche de sélection des variables explicatives[9], [29], [30].

Inférence causale

L'inférence causale désigne le processus par lequel on peut établir une relation de causalité entre un élément et ses effets. C'est un champ de recherche à la croisée des statistiques, et de l'intelligence artificielle.
Lorsque les humains rationalisent le monde, nous pensons souvent en termes de cause à effet - si nous comprenons pourquoi quelque chose s'est passé, nous pouvons changer notre comportement pour améliorer les résultats futurs.
Disons que nous examinons les données d'un réseau de serveurs. Nous souhaitons comprendre comment les modifications apportées à nos paramètres réseau affectent la latence. Nous utilisons donc l'inférence causale pour choisir de manière proactive nos paramètres en fonction de ces connaissances.

Cela pourrait nous aider à comprendre la cause profonde d'un problème ou à créer des modèles d'apprentissage automatique plus robustes. L'inférence causale nous donne des outils pour comprendre ce que cela signifie que certaines variables affectent d'autres.

Parcimonie

La parcimonie [2] est un principe consistant à n'utiliser que le minimum de causes élémentaires pour expliquer un phénomène.

Homoscédasticité

L'homoscédasticité est une propriété fondamentale du modèle de la régression linéaire générale et fait partie de ses hypothèses de base. On parle d'homoscédasticité lorsque la variance des erreurs stochastiques de la régression est la même pour chaque observation i (de 1 à n observations).

Hétéroscédasticité

En statistique, l'on parle d'hétéroscédasticité lorsque les variances des résidus des variables examinées sont différentes.

3.2 Régression linéaire multiple[15]

Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en œuvre pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple.
Dans ce qui suit, nous allons montrer les problèmes rencontrés par la méthode des moindres carrés.

— Si la véritable relation entre la variable d'intérêt et les variables explicatives sont sensiblement linéaire, les estimateurs issus de la méthode des moindres carrés auront un biais faible.

— Si le nombre n d'observations est beaucoup plus grand que le nombre p de variables ($n >> p$), alors les estimateurs de la méthode des moindres carrés ont tendance à avoir également une petite variance.

— Si $p > n$, alors les estimateurs de la méthode des moindres carrés ne sont pas uniques et leur variance a tendance à être très grande, bien que leur biais reste petit . Ainsi, on ne peut pas utiliser la méthode en présence de tous les prédicteurs. Les méthodes de rétrécissement offrent une réduction de cette variance au prix d 'une augmentation du biais. Ceci se fait en ajoutant des pénalités à la fonction de perte de la méthode des moindres carrés. De telles pénalités forcent les coefficients à être petits en les rétrécissant. Ainsi, cela permet de réduire la variance des estimateurs, même si cela introduit une petite augmentation du biais.

— Interprétation du modèle : souvent, certains ou plusieurs des variables explicatives utilisés dans un modèle de régression multiple ne sont pas associés à la variable d'intérêt y. Les prédicteurs non pertinents compliquent la résolution du modèle résultant . En éliminant ces variables (c.-à-d. en forçant les estimateurs des coefficients correspondants à être nuls), nous pouvons obtenir un modèle qui est plus facile à interpréter. Cependant, il est peu probable que la méthode des moindres carrés nous donne des estimateurs de coefficients qui sont exactement des zéros.
  Dans cette section, nous allons voir quelques méthodes qui permettent la résolution de ce problème et qui peuvent fournir des estimations nulles des coefficients de variables qui s'apparentent à des bruits.

— Le rétrécissement : cette approche consiste à ajuster un modèle impliquant tous les p variables d'intérêt. Toutefois, les coefficients estimés sont rétrécis vers zéro par rapport aux estimateurs de la méthode des moindres carrés. Ce rétrécissement (également connu sous le nom de régularisation) a pour effet de réduire la variance. Selon la méthode de régularisation effectuée, certains coefficients peuvent être estimé exactement par zéro. Par conséquent, ces méthodes peuvent également effectuer la sélection des variables importantes pour la variable d'intérêt.

Définition 1. *Soit une variable d'intérêt y à prédire et $x_1, x_2, ..., x_p$ p variables explicatives. Le modèle linéaire général suppose :*

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \quad \forall i = 1, ..., n$$

Ou sous forme matricielle :

$$y = X\beta + \epsilon$$

Où $\beta = (\beta_0, \beta_1, ..., \beta_p)$ n vecteur de paramètres à estimer. Nous supposons également que les erreurs (les ϵ) sont gaussiennes, centrées, homoscédastiques et non corrélées, c'est-à-dire :

$$\begin{cases} E[\epsilon_i] = 0 & \forall i = 1, ..., n \\ Var[\epsilon_i] = \sigma^2 & \forall i = 1, ..., n \\ Cov(\epsilon_i, \epsilon_j) = 0 & \forall i \neq j \end{cases}$$

3.2.1 Formulation matricielle / vectorielle

Soit le système matricielle suivant :

$$y = X\beta + \epsilon \tag{3.1}$$

où :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix}$$

Un tel système n'a généralement pas de solution exacte, le but est donc de trouver les coefficients $\beta_0, \beta_1, ..., \beta_p$, la méthode des moindres carrés est la plus utilisée dans ce genre de résolutions de problème. :

$$\tilde{\beta} = argmin_{\beta \in \mathbb{R}^{p+1}} SCR(\beta) = \{\beta \in \mathbb{R}^{p+1} | \forall \beta^{'} \in \mathbb{R}^{p+1}, SCR(\beta^{'}) \geq SCR(\beta)\} \tag{3.2}$$

où la fonction objectif SCR est la somme des carrés des résidus définie par :

$$SCR(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right) \right)^2 = \|y - X\beta\|_2^2$$

Remarque 1. *En fait le critère des moindres carrés est donné par :*

$$\frac{1}{2n} \sum_{i=1}^{n} \epsilon_i^2 = \frac{1}{2n} \|y - X\beta\|_2^2$$

le critère 3.2 est equivalent à la recherche de $\beta \in \mathbb{R}^{p+1}$ *tel que* $\forall \beta^{'} \in \mathbb{R}^{p+1}$ *on à :*

$$\frac{1}{2n} \|y - X\beta^{'}\|_2^2 \geq \frac{1}{2n} \|y - X\beta\|_2^2 \tag{3.3}$$

la solution de (3.3) *est equivalent à :*

$$min_{\beta \in \mathbb{R}^{p+1}} \|y - X\beta\|_2^2$$

On suppose que le rang de la matrice X de l'échantillon est p, et que le nombre d'observations n est supérieur au nombre de caractéristiques p,c'est à dire, $p \leq n$. Sous ces conditions, les colonnes $X^{(1)}, ..., X^{(p)}$ de la matrice X sont linéairement indépendantes. La fonction qui à β associe $\|y - X\beta\|_2^2$ est strictement convexe, et donc admet une unique solution notée $\tilde{\beta}$. Soit la résolution matricielle suivante :

$$\begin{aligned} SCR(\beta) = \|y - X\beta\|_2^2 &= (y - X\beta)^T (y - X\beta) \\ &= (y^T - (X\beta)^T)(y - X\beta) \\ &= y^T y - y^T (X\beta) - (X\beta)^T y + (X\beta)^T (X\beta) \\ &= y^T y - 2(X\beta)^T y + \beta^T X^T X\beta \end{aligned} \tag{3.4}$$

Pour trouver où la fonction ci-dessus a un minimum, on a la condition d'optimalité (annulation du gradient) fournit la solution :

$$\frac{\partial SCR(\beta)}{\partial \beta} = 2X^T X\beta - 2X^T y = 0$$

Ou :

$$X^T X\beta = X^T y$$

Maintenant, en supposant que la matrice $X^T X$ est inversible, nous pouvons multiplier les deux côtés par $(X^T X)^{-1}$ et obtenir

$$\beta = (X^T X)^{-1} X^T y \tag{3.5}$$

Quelle est l'équation normale.

3.2.2 Limitations de la régression linéaire

— La régression linéaire souffre de quelques inconvénients quand les variables sont corrélées : la solution n'est pas unique et les coefficients ont une grande variabilité, et l'interprétation est plus difficile.

— Une autre situation dans laquelle la solution n'est pas unique est celui où le nombre de variables est plus grand que celui d'observations $p > n$. Dans ce cas, la matrice $X^T X$ n'est pas inversible : X ne peut pas être de rang colonne plein en ayant plus de colonnes (variables) que de lignes (observations).

— Il y a donc un risque de sur-apprentissage.

3.3 Régression linéaire pénalisée

Le critère 3.2 n'est valable que lorsque $n \geq p$, Mais quand on a $p > n$, Le problème des moindres carrés n'a pas de solution unique du fait que $X^T X$ n'est pas inversible, d'où l'idée d'introduire une pénalisation en fonction de la norme de β dans la fonction à minimiser est :

$$\tilde{\beta} = argmin_{\beta \in \mathbb{R}^{p+1}} \left(\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j \right) \right)^2 + \lambda \Omega(\beta) \right) \tag{3.6}$$

En d'autres termes, On cherche l'hyperplan qui passe au mieux entre tous les points. Mais quand on a plus de variables explicatives (dimensions dans laquelles ces points sont exprimés) que de caractéristiques (points), il y a une infinité d'hyperplans qui minimise cette somme.

Deux choix usuels de $\Omega(\beta)$ sont :

— Régression Ridge :

$$\Omega(\beta) = \sum_{j=0}^{p} \beta_j^2 = \|\beta\|_2^2 \tag{3.7}$$

— Régression lasso :

$$\Omega(\beta) = \sum_{j=0}^{p} |\beta_j| = \|\beta\|_1 \tag{3.8}$$

$\lambda \geq 0$ est le paramètre de régularisation (à choisir par l'utilisateur) et $\Omega(\beta) \geq 0$ est le terme de régularisation. L'idée ici est de pousser certains paramètres β_j (ceux qui sont associés aux variables explicatives qui ne sont pas essentielles pour décrire la sortie y) en jouant sur la valeur de λ et la forme de $\Omega(\beta)$. À l'extrême, lorsque :

— $\lambda \to \infty$ on souhaite avoir $\Omega(\beta) \to 0$ c'est-à-dire $\beta = 0$ (aucune variable n'est sélectionnée)

— $\lambda = 0$, nous retrouvons la solution (3.5) de la régression linéaire (toutes les variables explicatives sont incluses dans le modèle)

. Les choix appropriés de λ et β permettront de déterminer un compromis entre ces deux situations extrêmes (3.7) et (3.8).

Il s'agit donc d'une méthode de sélection de variables et de réduction de dimension supervisée : les variables qui ne sont pas nécessaires à la prédiction de l'étiquette sont éliminées.

3.3.1 Interprétation géométrique

Il se trouve que, pour une valeur de $\lambda \in \mathbb{R}_+$ donnée, il existe un $t \in \mathbb{R}_+$ unique tel que :

$$\tilde{\beta}^{Lasso} = argmin_{\beta \in \mathbb{R}^{p+1}} \|y - X\beta\|_2^2 \text{ sous la contrainte } \|\beta\|_1 \leq t$$

$$\tilde{\beta}^{Ridge} = argmin_{\beta \in \mathbb{R}^{p+1}} \|y - X\beta\|_2^2 \text{ sous la contrainte } \|\beta\|_2^2 \leq t$$

Remarque 2. — *il y'a une équivalence direct entre t et* λ

— $t \to 0 \Leftrightarrow \lambda \to +\infty$: $\beta_j \to 0$ *(tous), variances des coefficients nulles.*

— $t \to +\infty \Leftrightarrow \lambda \to 0$: $\beta_{Lasso} = \beta_{Ridge} = \beta_{MCO}$

les deux figures ci-dessous donnent une comparaison de la régression Lasso et celle du Ridge avec les différentes contraintes exercées. :

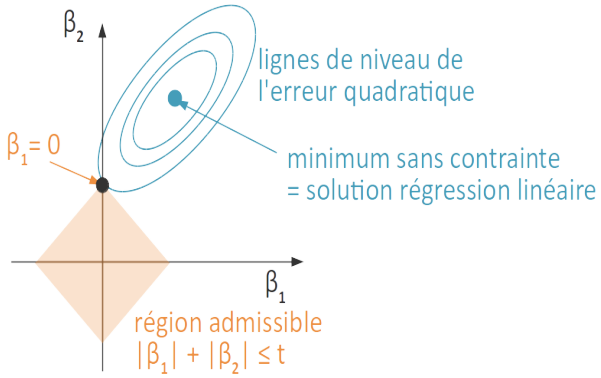


FIGURE 3.1 – Solution du Lasso

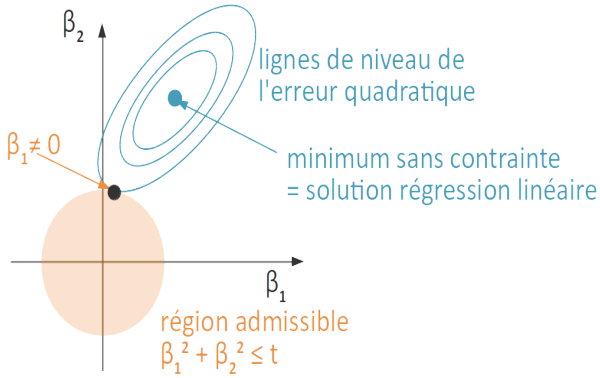


FIGURE 3.2 – Solution du ridge

Géométriquement, cela signifie que la solution du Lasso est un point situé à l'intersection d'une ligne de niveau du terme d'erreur $\|y - X\beta\|_2$ et de la région $\|\beta\|_1$ dite "admissible", c'est-à-dire, la région de $\mathbb{R}^p$ où la contrainte est verifiée. De plus, puisque le terme d'erreur sus-mentionné est quadratique en β alors la ligne de niveau est une ellipsoide. Par ailleurs, la région admissible $\|\beta\|_1$ est une boule de la norme l_1 de rayon t, autrement dit, un hypercube. Et enfin, comme cet hypercube a des sommets alors l'ellipsoide est susceptible de la rencontrer sur un de ces sommets, là où une ou plusieurs coordonnées sont nulles.
La solution de ce problème Ridge est l'intersection d'une courbe de niveau de l'erreur $\|y - X\beta\|_2$ avec la boule l_2 de rayon t c'est-à-dire, la boule Euclidienne "ronde" $\|\beta\|_2^2$. Cette fois-ci, il n'y a ici aucune raison que cette intersection se fasse à un endroit où une ou plusieurs coordonnées s'annulent.

3.3.2 Solution analytique de la régression linéaire pénalisée

Régression Ridge

Revenons à notre problème général résoudre l'equation (3.6) pour la régression Ridge est equivalent à résoudre :

$$min_{\beta\in\mathbb{R}^{p+1}}\{\|y-X\beta\|_2^2+\lambda\|\beta\|_2^2\} \tag{3.9}$$

on a la résolution matricielle suivant : $\|y-X\beta\|_2^2=(y-X\beta)^T(y-X\beta)$ et $\|\beta\|_2^2=\beta^T\beta$, et donc la solution de l'equation (3.9) est equivalent à minimiser la forme matricielle suivant :

$$SCR(\beta,\lambda)=(y-X\beta)^T(y-X\beta)+\lambda\beta^T\beta$$

ou encore

$$SCR(\beta,\lambda)=y^Ty-2\beta^TX^Ty+\beta^TX^TX\beta+\lambda\beta^T\beta$$

La condition d'optimalité (annulation du gradient) fournit la solution :

$$\frac{\partial SCR(\beta,\lambda)}{\partial\beta}=2(X^TX)\beta-2X^Ty+2\lambda\beta=0$$

Cette expression peut être encore simplifiée comme suit :

$$2(X^TX)\beta+2\lambda\beta=2X^Ty$$

$$(X^TX+\lambda I)\beta=X^Ty$$

La matrice symétrique X^TX étant définie semi-positive, toutes ses valeurs propres sont non négatives et donc $X^TX+\lambda I$ est symétrique et définie positive ($\lambda>0$), donc inversible. Alors, l'estimateur Ridge est donné par la formule explicite :

$$\tilde{\beta}^{Ridje}=(X^TX+\lambda I)^{-1}X^Ty$$

1. Avantages de la méthode Ridge :
   - — rétrécir les coefficients β
   - — très performante, en présence de corrélation entre les colonnes de X
   - — elle améliore l'erreur de prédiction, en réduisant la variance des estimate urs.
2. Inconvénients :
   - — c'est une méthode non appropriée pour la sélection des variables. En effet, si des prédicteurs sont fortement corrélés entre eux, leurs coefficients seront très proches les uns des autres
   - — elle ne produit pas de parcimonie dans le modèle. Autrement dit, la méthode ne pénalise pas les variables nuisibles par des coefficients exactement nuls.

Régression Lasso

Le problème Lasso s'écrit :

$$\tilde{\beta}=argmin_{\beta\in\mathbb{R}^{p+1}}\left(\sum_{i=1}^{n}\left(y_i-\left(\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j\right)\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right) \tag{3.10}$$

La difficulté du lasso est la non-dérivabilité de la fonction $|\beta|$ en 0, En revanche, $\tilde{\beta}$ n'admet pas de solution analytique et se détermine via des algorithmes d'optimisation itératifs. Un exemple d'algorithme de résolution du problème lasso est le suivant :

— Initialisation :
  — centrer les sorties $y_i \leftarrow y_i - \bar{y}$ et les entrées $x_i - \bar{x}$, $\forall i = 1, ..., n$
  — fixer $\beta^0 = 0$ et $t = 0$
— Répétez :
  — $\tilde{\beta}_1^t \leftarrow argmin_{\beta_1} \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - \sum_{j \neq 1} x_{ij}\beta_j^{t-1})^2 + \lambda|\beta_1|$
  — $\tilde{\beta}_2^t \leftarrow argmin_{\beta_2} \sum_{i=1}^{n} (y_i - x_{i1}\beta_1^t - x_{i2}\beta_2 - \sum_{j=3}^{p} x_{ij}\beta_j^{t-1})^2 + \lambda|\beta_2|$
  — ...
  — $\tilde{\beta}_p^t \leftarrow argmin_{\beta_p} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p-1} x_{ij}\beta_j^t)^2 + \lambda|\beta_p|$
  — $t \leftarrow t + 1$
— Jusqu'à convergence c.à.d $\beta_j^t \simeq \beta_j^{t-1}$

1. Avantages de la méthode Lasso :
   — elle crée une parcimonie. Cela veut dire qu'elle élimine les variables nuisibles dans le modèle en estimant leur coefficients dans le modèle par des zéros.
   — c'est une bonne méthode pour choisir les variables qui contribuent le plus dans le modèle
   — elle rétrécit les coefficients β vers zéro.
2. Inconvénients :
   — c'est une méthode non appropriée pour la sélection des groupes des prédicteurs. En effet, si des prédicteurs sont fortement corrélés entre eux, la méthode Lasso choisit un prédicteur et pénalise les autres avec des coefficients nuls.
   — dans le cas où $p > n$, l'approche Lasso choisie au maximum n variables.

3.3.3 Choix du paramètre de régularisation λ

La sélection des paramètres de régularisation joue un rôle très important dans la performance des méthodes de régularisation. Ainsi, nous avons besoin d'une technique efficace pour choisir ces paramètres. Dans la littérature, la méthode la plus utilisée dans ce domaine est la validation croisée. C'est une méthode simple pour estimer les paramètres de régularisation λ.

La validation croisée4.1.2

La validation croisée est un outil important dans l'application pratique d'un grand nombre d'approches de l'apprentissage statistique. Par exemple, elle peut être utilisée pour estimer l'erreur associée à une méthode d'apprentissage statistique donnée, pour évaluer la performance d'un modèle à l'échelle de la population des données. Ainsi, la méthode consiste à estimer les paramètres du modèle sur un jeu de données appelé jeu de données d 'apprentissage et valider la performance du modèle en calculant une statistique qui mesure les écarts entre les données observées et ce qui est prédit dans un jeu de données qui n'a pas été utilisé pour estimer les paramètres du modèle.
Supposons que l'on a un échantillon d 'apprentissage, et que l'on désire estimer l'erreur de test d 'une méthode d 'apprentissage statistique de façon appropriée. L'approche de l'ensemble de validation est une technique simple que l'on utilise pour atteindre un tel but. L'idée de cette méthode consiste à diviser l'ensemble sous étude arbitrairement en

deux ensembles, un ensemble pour l'apprentissage, et un autre pour la validation. On utilise l'échantillon d'apprentissage pour ajuster le modèle et estimer les paramètres, et l'échantillon de validation pour choisir le bon modèle, c-à.d le modèle qui donne la plus petite valeur de l'erreur de validation MSE définie par :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y_i})^2$$

3.4 Algorithme descente par coordonnée

Parfois la résolution analytique n'est pas possible, parce que le nombre de paramètre est élevé par exemple, ou parce que le calcul serait trop coûteux donc implique une approximation avec une approche itérative. l'algorithme de descente par coordonnée est le plus utilisé pour l'éstimation des paramètres d'un modèle de régression linéaire, cela ce fait par minimisation d'une fonction objectif celle de Lasso, Ridge, moindres carrés, ..., il s'ajout un autre algorithme LARS (least-angle regression) [7] dans ce mémoire en s'intérèsse que à la descente par coordonnée.

La descente par coordonnée [10], ou coordinate descente, est souvent utilisé car très général en plus d'être performant. Il a été utilisé dans le cadre de la régression parcimonieuse initialement par [W. J. Fu, 1998] [11]. Le principe est de résoudre le problème en minimisant la fonction objective f par rapport à une unique coordonnée, en effectuant une boucle sur toutes les coordonnées jusqu'à convergence. L'intérêt de cette méthode est que généralement on peut obtenir à bas coût la solution de la minimisation sur une coordonnée.Autrement dit, en commençant par les valeurs de variable initiales :

$$x^0 = (x_1^0, x_2^0, ..., x_p^0)$$

définir x^{k+1} à partir de x^k par résolution itérative des problèmes d'optimisation à variable unique :

$$x_j^{k+1} \in argmin_{x_j \in \mathbb{R}} f(x_1^{k+1}, ..., x_{j-1}^{k+1}, x_j, x_{j+1}^k, ..., x_p^k) \quad j = 1, 2, ..., p$$

Ainsi, on commence par une première estimation x^0 pour un minimum local de f et obtient une séquence $x^0, x^1, x^2, ...$ itérativement. En effectuant une recherche par ligne à chaque itération, on a automatiquement :

$$f(x^0) \geq f(x^1) \geq f(x^2) \geq ...$$

L'algorithme de descente par coordonnées n'est pas nécessairement convergent, mais la séparabilité des fonctions de perte et de régularisation utilisées assure la convergence dans le cas de la régression parcimonieuse.

Le lemme suivant donne des conditions à satisfaire par une fonction à plusieurs variables pour qu'elle admette un minimum global en x.

***Lemme* : 1.** *Soit $f : \mathbb{R}^p \mapsto \mathbb{R}$, une fonction convexe et dérivable . S'il existe un point $x \in \mathbb{R}^p$, tel que f est minimisée en chaque coordonnée de x, alors f admet un minimum global en x.*

***Preuve* : 1.** *Nous savons d'après l'énoncé que*

$$\forall i = 1, ..., p : \frac{\partial f(x)}{\partial x_i} = 0$$

Alors, nous avons

$$\nabla(f) = \left(\frac{\partial f(x)}{\partial x_1}, ..., \frac{\partial f(x)}{\partial x_p} \right) = 0$$

Puisque f est convexe, x est un minimum global. D'où le résultat.

Remarque 3. *Dans le cas où f est une fonction convexe non dérivable, f n'admet pas de minimum global. Voici un contre exemple, la fonction $f(x,y) = x^2 + y^2 + 2\lambda|x-y|, (x,y) \in [-\lambda, \lambda]^2$ admet 0 comme minimum pour l'axe des abscisses et pour l'axe des ordonnées. Cependant, le point $(0,0)$ n'est pas un minimum global pour la fonction f.*

***Lemme* : 2.** *Soit $f : \mathbb{R}^p \mapsto \mathbb{R}$, qui peut s'écrire de la façon suivante*

$$f(x) = g(x) + \sum_{i=1}^{p} h_i(x_i)$$

où g est une fonction convexe et dérivable, et $h_i, i = 1, ..., p$ sont des fonctions convexes.

S 'il existe un point $x \in \mathbb{R}^p$ tel que f est minimisée en chaque coordonnée de x, alors f admet x comme minimum global.

***Preuve* : 2.** *On cherche à montrer que $x \in \mathbb{R}^p$ est un minimum global pour f sachant que les entrées de x sont les minimums de f pour chacune des coordonnées. Il faut montrer que*

$$\forall z \in \mathbb{R}^p : f(z) \geq f(x)$$

En effet,

$$f(z) - f(x) = g(z) - g(x) + \sum_{i=1}^{p} [h_i(z_i) - h_i(x_i)]$$

D'abord, nous allons montrer que

$$g(z) - g(x) \geq \nabla g(x)^T (z - x)$$

En effet, dans le cas $p = 1$, puisque g est convexe, alors pour tout $0 \leq t \leq 1$, nous avons

$$g(tz + (1-t)x) \leq tg(z) + (1-t)g(x)$$

Ce qui implique

$$\frac{g(tz + (1-t)x) - g(x)}{t} \leq g(z) - g(x)$$

Puisque g est dérivable, on fait tendre t vers zéro pour avoir

$$g(z) - g(x) \geq \nabla g(x)^T (z - x) \tag{3.11}$$

Pour le cas général, nous allons poser

$$G(t) = g(tz + (1-t)x)$$

Puisque g est convexe et dérivable, alors G est convexe et dérivable. Ensuite, nous pouvons écrire

$$G(1) \geq G(0) + G^{'}(0)$$

En effet, nous avons appliqué l'équation (3.11) pour la fonction G entre les points 0 et 1. Ce qui équivaut à

$$g(z) - g(x) \geq \nabla g(x)^T (z - x)$$

Alors, nous pouvons écrire

$$\begin{aligned} f(z) - f(x) &\geq \nabla g(x)^T (z-x) + \sum_{i=1}^{p} [h_i(z_i) - h_i(x_i)] \\ &= \sum_{i=1}^{p} [\nabla_i g(x)(z_i - x_i) + h_i(z_i) - h_i(x_i)] \\ &\geq \sum_{i=1}^{p} [\nabla_i g(x)(z_i - x_i) + h^{'}(x_i, d)] \quad ,où \quad d = z_i - x_i \\ &\geq 0 \end{aligned} \tag{3.12}$$

L'algorithme de la descente par coordonnée [27] nous montre comment calculer le minimum global d'une fonction dans le contexte du lemme 2

Algorithme 1 : Descente par coordonnée

Entrées : f, K
Initialisation : $k = 0$ et $\beta^{(0)} = 0 \in \mathbb{R}^{p+1}$
pour $k = 0, ..., K$ faire :
$\beta_0^{(k+1)} \leftarrow argmin_{\beta_0 \in \mathbb{R}} f(\beta_0, \beta_1^{(k)}, \beta_2^{(k)}, ..., \beta_{p-1}^{(k)}, \beta_p^{(k)})$
$\beta_1^{(k+1)} \leftarrow argmin_{\beta_1 \in \mathbb{R}} f(\beta_0^{(k+1)}, \beta_1, \beta_2^{(k)}, ..., \beta_{p-1}^{(k)}, \beta_p^{(k)})$
$\beta_2^{(k+1)} \leftarrow argmin_{\beta_2 \in \mathbb{R}} f(\beta_0^{(k+1)}, \beta_1^{k+1}, \beta_2, ..., \beta_{p-1}^{(k)}, \beta_p^{(k)})$
.
.
.
$\beta_p^{(k+1)} \leftarrow argmin_{\beta_1 \in \mathbb{R}} f(\beta_0^{(k+1)}, \beta_1^{(k+1)}, \beta_2^{(k+1)}, ..., \beta_{p-1}^{(k+1)}, \beta_p)$
Sorties : $\beta^{(K)}$

Objectif : trouver une solution approchée de $\text{argmin}_{\beta \in \mathbb{R}^{p+1}} f(\beta)$

3.4.1 Optimiser le problème des moindres carrés par une coordonnée à la fois

Soit $f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$ où $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, avec colonnes $X_1, ..., X_p$, on a :

$$0 = \nabla_j f(\beta) = X_j^T (X\beta - y) = X_j^T (X_{-j}\beta_{-j} + X_j\beta_j - y)$$

Nous prenons :

$$\beta_j = \frac{X_j^T (y - X_{-j}\beta_{-j})}{X_j^T X_j} = \frac{X_j^T (y - X_{-j}\beta_{-j})}{\|X_j\|_2^2}$$

Algorithme 2 : Descente par coordonnée pour la régression des moindres carrés

Entrer : le nombre d'iteration K, et la paramètre de regularisation λ
Initialiser $\beta^{(0)} = 0$
pour k=0 :K faire :
pour j=0 :p faire :
$\beta_j^{(k+1)} = \frac{X_j^T (y - X_{-j}\beta_{-j}^{(k)})}{\|X_j\|_2^2}$
Sorties : $\beta^{(K)}$

3.4.2 Optimiser le problème de ridge par une coordonnée à la fois

Soit $f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2 + \frac{1}{2}\lambda\|\beta\|_2^2$ où $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\lambda \geq 0$, on a :

$$0 = \nabla_j f(\beta) = X_j^T(X\beta - y) + \lambda\beta_j = X_j^T(X_{-j}\beta_{-j} + X_j\beta_j - y) + \lambda\beta_j$$

Nous prenons :

$$\beta_j = \frac{X_j^T(y - X_{-j}\beta_{-j})}{X_j^T X_j + \lambda} = \frac{X_j^T(y - X_{-j}\beta_{-j})}{\|X_j\|_2^2 + \lambda}$$

Algorithme 3 : Descente par coordonnée pour la régression ridge

Entrer : le nombre d'iteration K, et la paramètre de regularisation λ
Initialiser $\beta^{(0)} = 0$
  pour k=0 :K faire :
    pour j=0 :p faire :
      $\beta_j^{(k+1)} = \frac{X_j^T(y - X_{-j}\beta_{-j}^{(k)})}{\|X_j\|_2^2 + \lambda}$
Sorties : $\beta^{(K)}$

3.4.3 Optimiser le problème de lasso par une coordonnée à la fois

Soit $f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2 + \frac{1}{2}\lambda\|\beta\|_1$ où $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\lambda \geq 0$, on a :

$$\begin{aligned} 0 = \nabla_j f(\beta) &= -X_j^T(y - X\beta) + \frac{1}{2}\lambda\partial_{\beta_j}|\beta_j| \\ &= -X_j^T(y - X_{-j}\beta_{-j} - X_j\beta_j) + \frac{1}{2}\lambda\partial_{\beta_j}|\beta_j| \\ &= -\underbrace{X_j^T(y - X_{-j}\beta_{-j})}_{=P_j} + \underbrace{\|X_j\|_2^2}_{=Z_j}\beta_j + \frac{1}{2}\lambda\partial_{\beta_j}|\beta_j| \\ &= Z_j\beta_j - P_j + \begin{cases} -\frac{\lambda}{2} & si\beta_j < 0 \\ [-\frac{\lambda}{2}, -\frac{\lambda}{2}] & si\beta_j = 0 \\ \frac{\lambda}{2} & si\beta_j > 0 \end{cases} \end{aligned} \tag{3.13}$$

— Cas 1 : $(\beta_j < 0) : Z_j\beta_j - P_j - \frac{\lambda}{2} = 0$, $\beta_j = \frac{P_j + \frac{\lambda}{2}}{Z_j}$
  — Pour $\beta_j < 0 : P_j < -\frac{\lambda}{2}$
— Cas 2 : $(\beta_j = 0)$
  — Pour $\beta_j = 0$ on a : $-P_j + \frac{\lambda}{2} \geq 0 \rightarrow P_j \leq \frac{\lambda}{2}$ $\quad -P_j - \frac{\lambda}{2} \leq 0 \rightarrow P_j \geq -\frac{\lambda}{2}$ $\quad -\frac{\lambda}{2} \leq P_j \leq \frac{\lambda}{2}$
— Cas 3 : $(\beta_j > 0) : Z_j\beta_j - P_j + \frac{\lambda}{2} = 0$ $\beta_j = \frac{P_j - \frac{\lambda}{2}}{Z_j}$
  — Pour $\beta_j > 0 : P_j > \frac{\lambda}{2}$

Donc :

$$\beta_j = \begin{cases} \frac{P_j + \frac{\lambda}{2}}{Z_j} & siP_j < -\frac{\lambda}{2} \\ 0 & si -\frac{\lambda}{2} \leq P_j \leq \frac{\lambda}{2} \\ \frac{P_j - \frac{\lambda}{2}}{Z_j} & siP_j > \frac{\lambda}{2} \end{cases} \tag{3.14}$$

Algorithme 4 : Descente par coordonnée pour la régression lasso

Entrer : le nombre d'iteration K, et la paramètre de regularisation λ

Initialiser $\beta^{(0)} = 0$

pour k=0 :K faire :

pour j=0 :p faire :

$$\beta_j^{(k+1)} = \begin{cases} \frac{X_j^T(y - X_{-j}\beta_{-j}^{(k)}) + \frac{\lambda}{2}}{\|X_j\|_2^2} & si P_j < -\frac{\lambda}{2} \\ 0 & si -\frac{\lambda}{2} \leq P_j \leq \frac{\lambda}{2} \\ \frac{X_j^T(y - X_{-j}\beta_{-j}^{(k)}) - \frac{\lambda}{2}}{\|X_j\|_2^2} & si P_j > \frac{\lambda}{2} \end{cases}$$

Sorties : $\beta^{(K)}$

3.5 Application d'algorithme et résultats de l'analyse

L'une des tâches principales de la détection des anomalies est la détection des cause racines de ces anomalies, une série y (dans ce cas y=E-RAB :DR) d'un comportement anomalie peut être la cause d'une anomalie d'autres séries appelées séries condidats et déterminées par des experts. Ceci justifie le choix du modèle Lasso, les caractéristiques du modèle Lasso (voire : Avantages de la méthode Lasso1)telles que l'élimination des variables non pertinents, la convergence rapide vers une solution optimale le rend préférable pour la résolution de ce problème.

Nous voulions comprendre l'effet causal de la variable du traitement x_j, $j = 1,...,9$ sur le résultat y=E-RAB :DR. Si nous trouvons que le β_j est positif, cela signifie qu'une augmentation de x_j entraîneune augmentation de y. De même, un β_j négatif indique qu'une augmentation de x_j entraînera une diminution de y.

Nom KPI	série
RACH : Cont based Stp SR	1
RRC connections : Conn Stp SR	2
E-RAB : Stp SR	3
E-RAB : Initial Accessibility	4
E-RAB : Active time, all	5
HARQ Retrans ratio : DL	6
Quality : Avg cqi	7
Quality : Avg SINR PUCCH	8
Quality : Avg SINR PUSCH	9

Table 3.1 – Nom des kpis lié aux séries da la table 3.2 de 1 à 9

E-RAB :DR	1	2	3	4	5	6	7	8	9
21.43	100.0	100.00	100.00	100.00	33.97	4.96	10.39	6.71	13.79
25.00	100.0	100.00	100.00	100.00	65.83	2.45	9.77	7.65	14.65
37.50	100.0	100.00	90.00	100.00	88.35	2.24	9.49	7.79	15.83
100.00	100.0	100.00	100.00	100.00	52.75	3.55	9.58	4.48	12.41
12.50	100.0	100.00	100.00	100.00	69.93	4.35	10.81	5.83	13.55
11.11	100.0	100.00	100.00	81.82	76.88	3.60	11.32	6.45	12.99
23.53	100.0	100.00	96.15	95.12	63.90	4.24	9.37	7.00	13.93
64.29	100.0	100.00	92.86	87.50	104.42	4.82	9.66	6.49	13.19
22.22	100.0	100.00	100.00	100.00	63.33	5.31	9.91	7.13	13.35
50.00	100.0	100.00	100.00	100.00	39.82	5.66	10.53	7.37	10.83
16.67	100.0	100.00	100.00	100.00	88.90	3.54	10.50	7.73	13.56
33.33	100.0	100.00	100.00	100.00	99.87	2.75	10.17	7.43	14.56
12.50	100.0	100.00	100.00	100.00	93.37	4.20	10.67	7.91	14.47
15.38	96.3	100.00	100.00	100.00	77.38	3.26	10.47	8.04	13.39
27.27	100.0	100.00	100.00	100.00	105.35	4.63	11.65	8.41	16.15
100.00	100.0	100.00	100.00	100.00	159.15	4.62	11.68	8.16	15.18
11.11	100.0	100.00	100.00	100.00	85.45	4.84	10.95	7.82	14.96
25.00	100.0	100.00	100.00	77.78	102.53	5.83	11.38	7.60	14.94
11.11	100.0	100.00	100.00	100.00	93.62	6.47	10.53	8.38	15.22
16.67	100.0	88.89	100.00	88.89	82.55	7.63	12.86	9.04	16.52
16.67	100.0	96.77	100.00	96.77	38.53	6.65	10.58	7.78	11.96
25.00	100.0	100.00	25.00	33.33	100.53	7.00	11.48	7.89	14.07
12.73	100.0	100.00	98.21	100.00	4.58	4.23	12.74	10.82	11.34
50.00	100.0	100.00	100.00	100.00	92.63	7.27	11.56	7.51	15.28
14.29	100.0	100.00	100.00	100.00	54.43	7.06	11.87	7.15	13.88
37.04	100.0	100.00	96.30	100.00	45.12	7.24	11.38	8.44	15.87

TABLE 3.2 – Nous essayons de découvrir lesquel des enregistrements 1 : 9 a un effet sur la série d'anomalies E-RAB : DR

série	$\tilde{\beta}^{Lasso}$	$\tilde{\beta}^{Ridge}$
1	0.000000e+00	0.262634
2	0.000000e+00	0.262931
3	0.000000e+00	-0.357176
4	0.000000e+00	0.591031
5	7.647342e-17	0.303969
6	-0.000000e+00	0.060851
7	-0.000000e+00	-0.258102
8	-0.000000e+00	-0.893550
9	-0.000000e+00	-0.779853

TABLE 3.3 – Ces résultats sont obtenus en appliquant Lasso et Ridge sur les données de table3.2, Lasso détecte une influence entre E-RAB :DR et la série 5, $\beta_5 = 7.647342e-17$, $\lambda = 232.573375$, par conter Ridge n'a éliminé aucune des séries avec meilleur $\lambda = 232.573375$

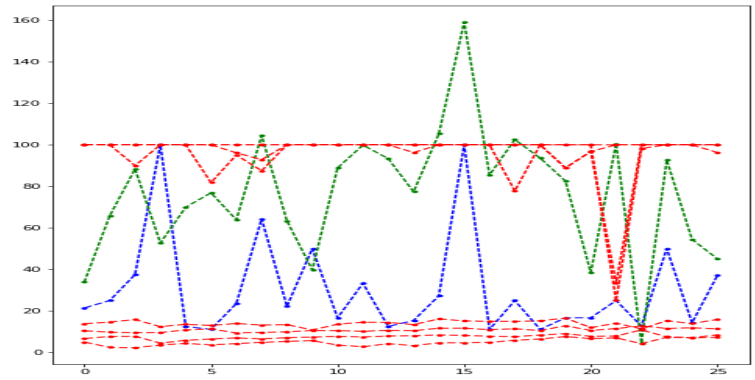


FIGURE 3.3 – cette figure est une présentation graphique des 10 séries d'études, la série verte représente "E-RAB : Active time, all", la série bleue représente E-RAB :DR, les séries rouge sont les 8 autres séries donc le lasso est une méthode de détection efficace

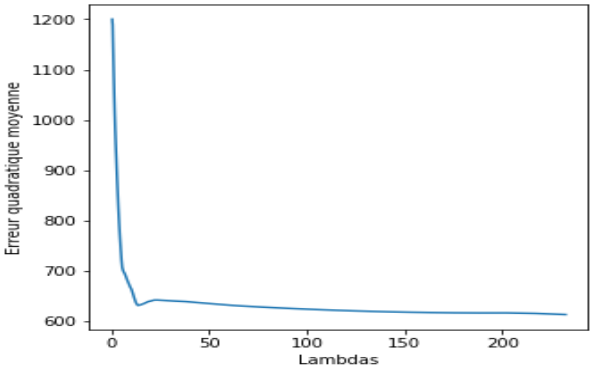


FIGURE 3.4 – Erreur quadratique moyenne du Lasso

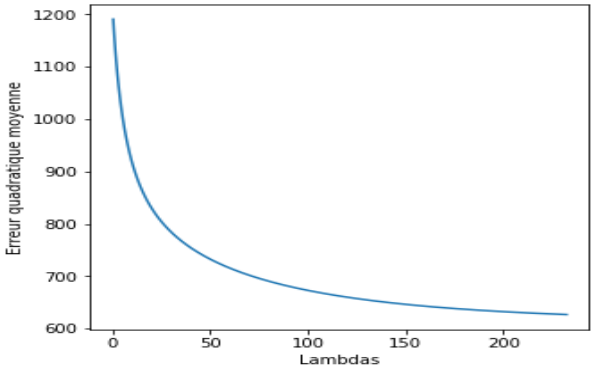


FIGURE 3.5 – Erreur quadratique moyenne du ridge

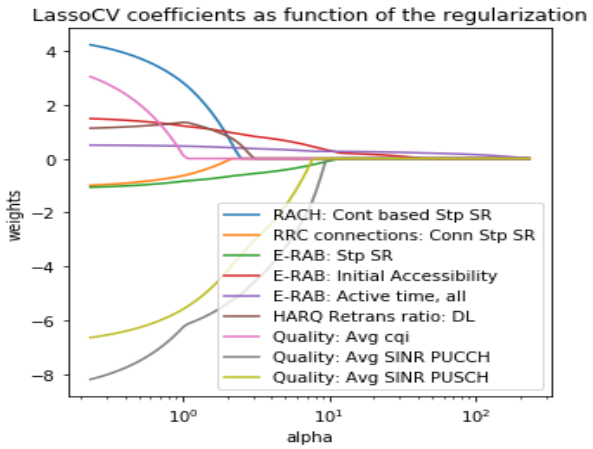


FIGURE 3.6 – Chemin de régularisation pour la régression Lasso

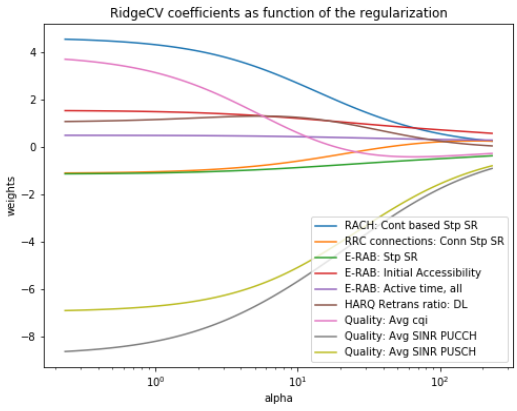


FIGURE 3.7 – Chemin de régularisation pour la régression ridge.

On constate que le lasso est plus efficace pour sélectionner les variables pertinentes et donne un modèle parcimonieux. Le ridge ne met explicitement un paramètre à 0 que pour de très grandes valeurs de λ.

3.6 conclusion

— L'idée des méthodes de sélection, abordées dans ce chapitre est de choisir le sous-modèle dont l'estimation du risque de prévision est minimale.

— Du point de vue optimisation on contraint la norme du vecteur des coefficients doit respecter une borne supérieure (on parle de rétrécissement, "shrinkage"). Si on utilise la norme l_2 on obtient le modèle ridge et si on utilise la norme 'l_1 on obtient le modèle lasso.

— L'utilisation des normes l_1 ou l_2 donne des solutions bien différentes malgré le même but recherché : la solution lasso est parcimonieuse contrairement à la solution ridge.

— En théorie, la régression lasso est intéressante puisqu'elle permet à la fois une erreur en généralisation plus faible et une solution parcimonieuse.

— Quand $p > n$ (données de grande dimension), la méthode lasso ne sélectionne que n variables.

— Si plusieurs variables sont corrélées entre elles, la méthode lasso ne sélectionnera qu'une seule d'entre elles et ignorera les autres.

— Dans de nombreux cas classiques avec $n > p$, s'il y a de fortes corrélations entre les variables explicatives, on trouve empiriquement que la méthode ridge donne de meilleures performances que la méthode lasso.

— Utilité de la descente par coordonnée est quand p est très grand.Il est souvent, difficile de trouver le minimum pour toutes les coordonnées, mais facile pour chaque coordonnée.

Chapitre 4

Implémentation et résultats

L'objectif de ce chapitre est de présenter les étapes de l'implémentation du modèle LassoCV (Least Absolute Shrinkage and Selection Operator Cross Validation) proposée dans le cadre de la détection d'anomalies. Nous commençons tout d'abord par la présentation des ressources et du langage que nous avons utilisé. Puis les étapes de la réalisation du modèle et on termine par les tests effectués.

4.1 Outils et Environnement

Pour le développement du modèle, la plate-forme Anaconda a été utilisée, qui est un distribution haute performance de packages de science des données pour Python et R. La plate-forme donne accès à des centaines de packages, mais pour ce mémoire, le Python la bibliothèque scikit-learn a été principalement utilisée. La bibliothèque scikit-learn est une bibliothèque open source simple et efficace qui contient un large éventail de puissants outils des algorithmes d'apprentissage automatique [21]. Son API simple mais puissante en fait un excellent choix pour ce type de problème d'apprentissage automatique.

Figure 4.1 – Logo anaconda

4.1.1 Python

Python est un langage de programmation de haut niveau utilisé pour la programmation générale. Créé par Guido van Rossum et sorti en 1991, Python a une philosophie de conception qui met l'accent sur la lisibilité du code, notamment en utilisant des espaces importants. Il fournit des constructions qui permettent une programmation claire à petite et à grande échelle. Python dispose d'un système de type dynamique et d'une gestion automatique de la mémoire. Il prend en charge de multiples paradigmes de programmation, y compris orientés objet, impératifs, fonctionnels et procéduraux,

et dispose d'une bibliothèque standard vaste et complète. Les interpréteurs de Python sont disponibles pour de nombreux systèmes d'exploitation.

Figure 4.2 – Logo python

4.1.2 Bibliothèques utilisées

Numpy

Est une bibliothèque permettant d'effectuer des calculs numériques avec Python. Il implémente des calculs sur des tableaux multidimensionnels et matrices. `https://numpy.org/`

Pandas

Pandas est une bibliothèque d'analyse de données de haut niveau efficace. Il offre des structures de données rapides, flexibles et faciles à manipuler, à savoir "DataFrame" et "Séries". `https://pandas.pydata.org/docs/`

Scikit-learn

Scikit-learn est une bibliothèque ML. Il met en œuvre la régression, la classification, et les algorithmes de clustering ainsi que certains prétraitements opérations telles que le nettoyage et l'interpolation des données. `https://scikit-learn.org/stable/`

Matplotlib

Matplotlib est une bibliothèque de visualisation de données. Il met en œuvre différents types de graphiques (nuages de points, histogrammes, camemberts, etc.). `https://matplotlib.org/`

Tkinter

Tkinter est généralement livré avec Python, et est le framework GUI standard de Python. Il est célèbre pour sa simplicité et utile pour crée des interfaces graphique. Il est open-source et disponible sous la licence Python. `https://wiki.python.org/moin/TkInter`

4.2 Les données

Tout élément de réseau contient des compteurs de mesure statistique de la performance que l'élément de réseau télécharge périodiquement vers une base de données

OSS [19]dans le centre O&M(Operation and Managing Center). Les données du compteur peuvent être enregistrées et collectées à des fréquences différentes (15 minutes, 60 minutes ou moins souvent), il sera important d'effectuer une agrégation sur les données afin que toutes les séries temporelles aient la même fréquence. La fréquence d'une heure est sélectionnée pour l'agrégation car la plupart des indicateurs clés de performance semblent avoir au moins une valeur dans une heure donnée et avec une fréquence moindre, nous pourrions ne pas capturer d'informations importantes. Lorsqu'un KPIs a plus d'une valeur en une heure, la moyenne des valeurs est utilisée pour cette heure. Exemple de données de séries temporelles avant et après l'agrégation temporelle :

Temps	Valeur
9 :00	100
9 :15	95
9 :30	98
9 :45	93
10 :00	99
10 :15	95
10 :30	90
10 :45	85

Table 4.1 – Séries temporelles avant agrégation

Temps	Valeur
9 :00	$(100 + 95 + 98 + 93)/4 = 96.5$
10 :00	$(99 + 95 + 90 + 85)/4 = 92.25$

Table 4.2 – Séries temporelles après agrégation

Les données collectées de l'OSS sont agréger et compresser dans un fichier CSV, dont chaque colonne représente une série de valeurs mesure une fonctionnalité dans le réseau LTE, tel que le nom de l'équipement utilisé, la zone gérer ou bien la cellule, des KPIs spécifique,. . . .etc, et chaque ligne représente la description de ces fonctionnalités enregistrée sur la même heure et la même date.

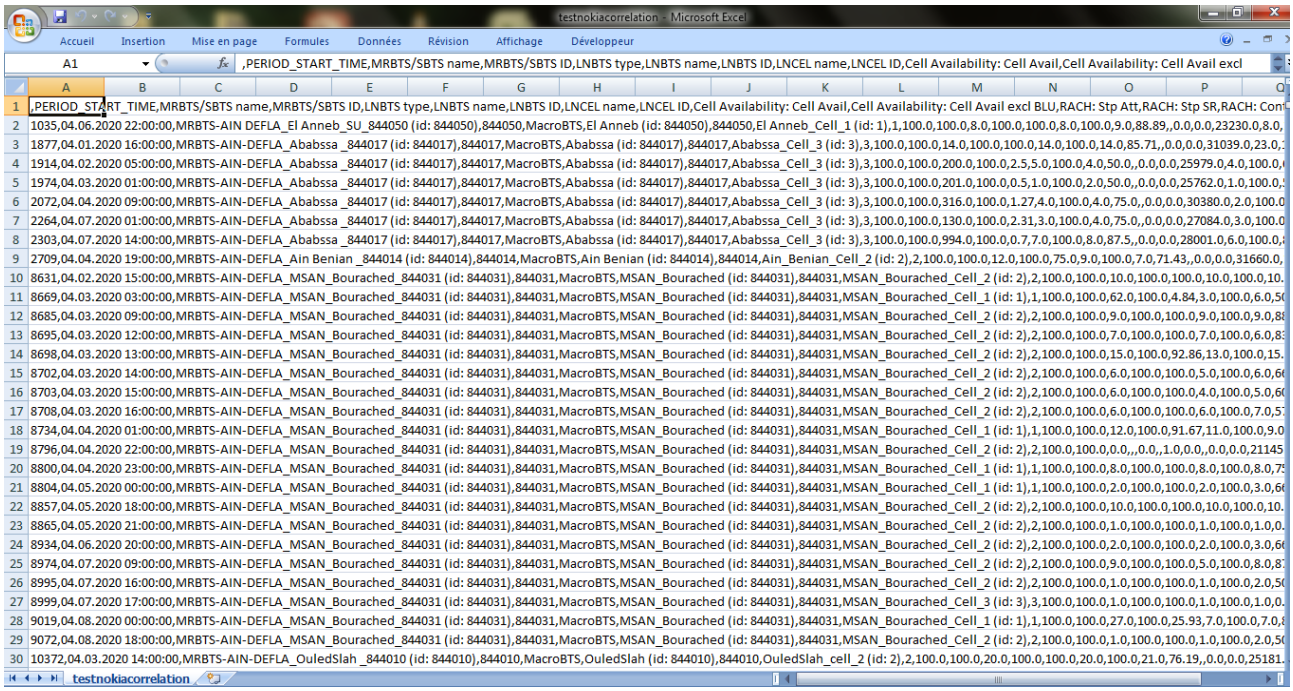


Figure 4.3 – Données csv

4.3 Implémentation et résultats

4.3.1 Visualisation de données

La visualisation des données est une compétence importante dans les statistiques appliquées et l'apprentissage automatique. La visualisation des données fournit une suite importante d'outils pour acquérir une compréhension qualitative.
Cela peut être utile lors de l'exploration et de la connaissance d'un ensemble de données et peut aider à identifier des modèles, des données corrompues, des valeurs aberrantes et bien plus encore. Avec un peu de connaissance du domaine, les visualisations de données peuvent être utilisées pour exprimer et démontrer des relations clés dans des graphiques et des graphiques qui sont plus viscéraux pour vous-même et les parties prenantes que des mesures d'association ou de signification.

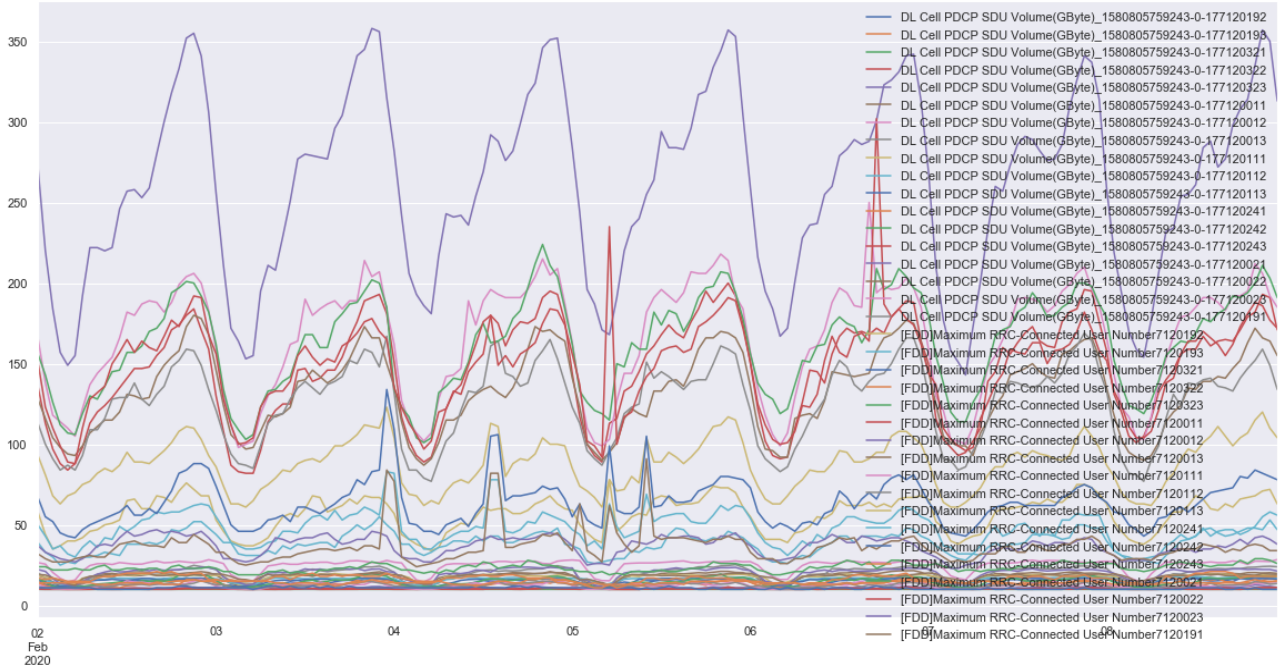


FIGURE 4.4 – visualiser les données à l'aide d'un Graphique linéaire nous a permet de remarquer une fort corrélation entre chaque 24 heurs donc une corrélation faible indiquera un comportement inormal

4.3.2 Data Preprocessing

Le travail de détection commence par la collecte des données de performance des cellules LTE. Une fois les données collectées, le preprocessing est effectué pour filtrer les valeurs incohérentes.Les algorithmes qui peuvent être appliqués n'acceptent que les valeurs numériques et les valeurs manquées doivent être filtrées. Étant donné que les performances diminuent en raison d'une panne de courant, les alarmes et la coupure de transmission ne doivent pas être considérées comme des anomalies, les valeurs de performance des cellules avec des alarmes et des cellules indisponibles sont également filtrées.

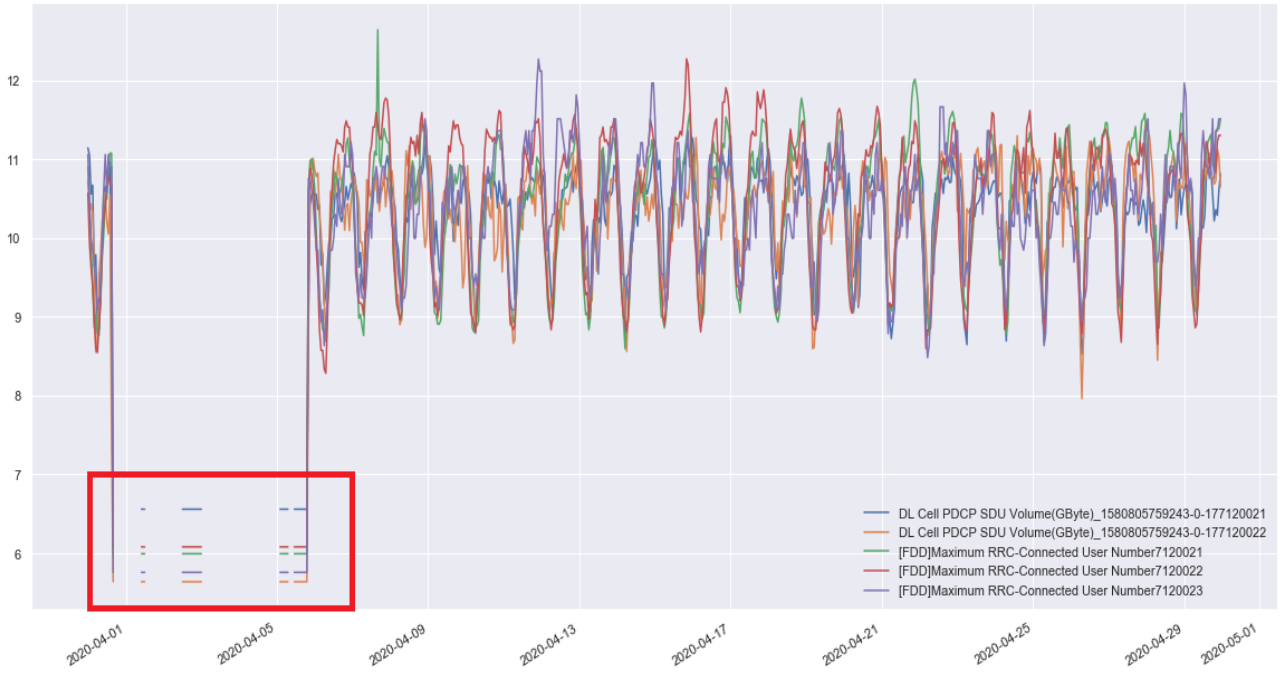


FIGURE 4.5 – dans notre analyse basé sur la corréllation les valeurs manquée sont éliminer des données

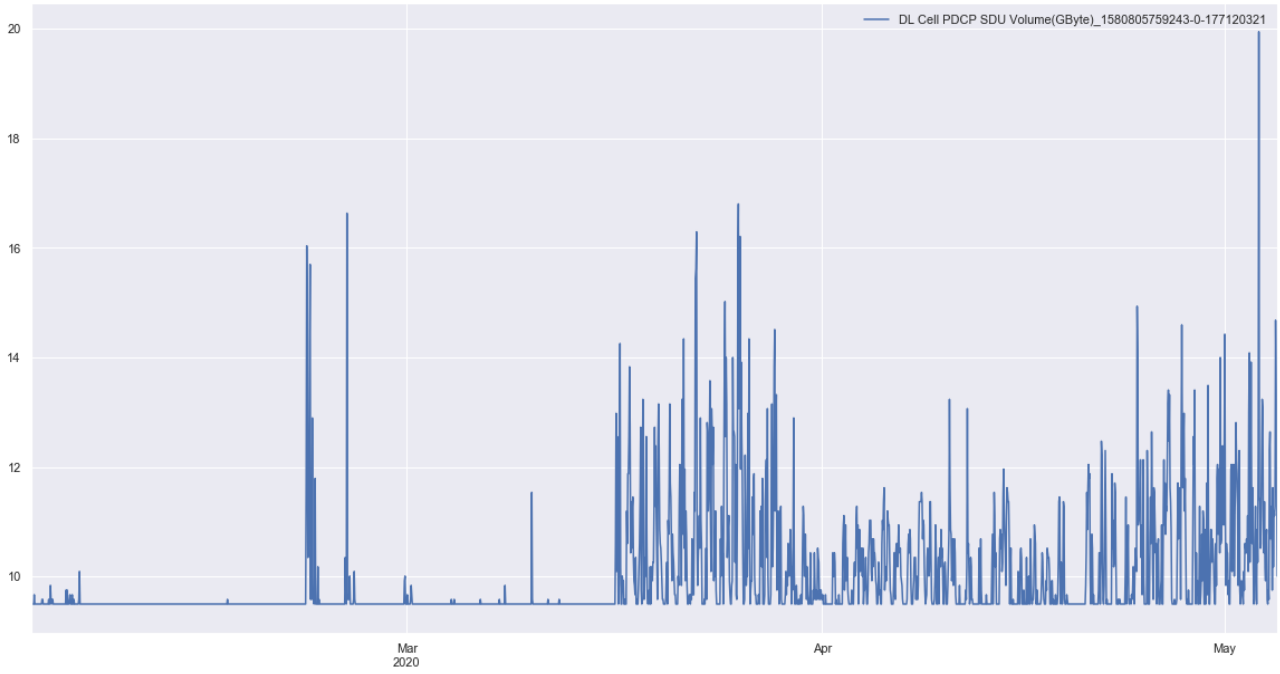


FIGURE 4.6 – Les séries incohérents sont éliminer car ne sont pas des indicateurs clés de performance

4.3.3 Feature Engineering

Feature Engineering est la science et l'art d'extraire de nouvelles fonctionnalités à partir des données brutes, en utilisant généralement des connaissances sur le terrain et une expérience ML accumulée, pour améliorer la précision du modèle ML. Une fonction raconte une histoire ou définit un aspect du problème qui aiderait le modèle à effectuer la tâche requise. Il peut être donné directement dans les données brutes ou peut être dérivé des caracteristiques des données.

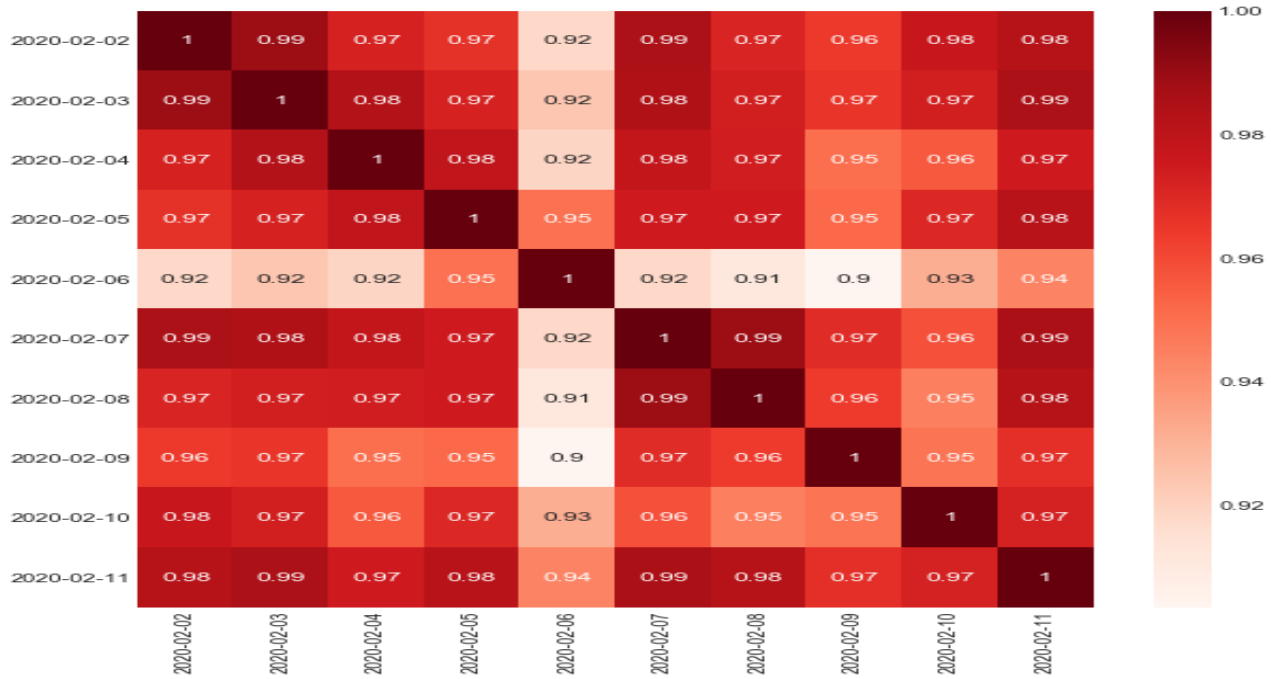


FIGURE 4.7 – illustration de la forte corrélattion entre fenêtre de 24 heurs d'une série temporelle

4.3.4 L'algorithmeLassoCV

LassoCV est parmi les algorithme de machine learning de package Scikit-learn4.1.2 sous python, et il est un modèle linéaire avec ajustement itératif le long d'un chemin de régularisation, implémenter pour résoudre le problème de **regression**, L'algorithme utilisé pour ajuster le modèle est la descente par coordonnées.Le meilleur modèle est sélectionné par validation croisée. L'objectif d'optimisation pour Lasso (où LassoCV) est :

$$\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Paramètres utilisé dans la résolution par LassoCV :

eps : Longueur du chemin, **n_alphas :** Nombre d'alphas le long du chemin de régularisation (alpha=λ dans la théorie), **alphas :** Liste des alphas pour calculer les modèles,**max_iter :** Le nombre maximum d'itérations, **cv :** Détermine la stratégie de fractionnement de validation croisée.

Les attributs de lassoCV sous Scikit-learn :

alpha_ : Le montant de pénalisation choisi par validation croisée, **coef_ :** vecteur de paramètre dans la formule de la fonction de coût, **mse_path_ :** erreur quadratique moyenne pour l'ensemble de test sur chaque groupe de test éffectué selon alpha, **alphas_ :** La grille d'alphas utilisée dans le modèle, **n_iter_ :** nombre d'itérations exécutées par le solveur de descente par coordonnées pour atteindre la tolérance spécifiée pour l'alpha optimal.

4.4 Présentation de l'interface graphique

Pour que le lecteur puisse utiliser notre application dans des bonnes conditions, nous allons consacrer cette partie à la présentation de l'application de notre méthode. Les

figures ci-dessous représentent la démarche à suivre pour utiliser notre application.

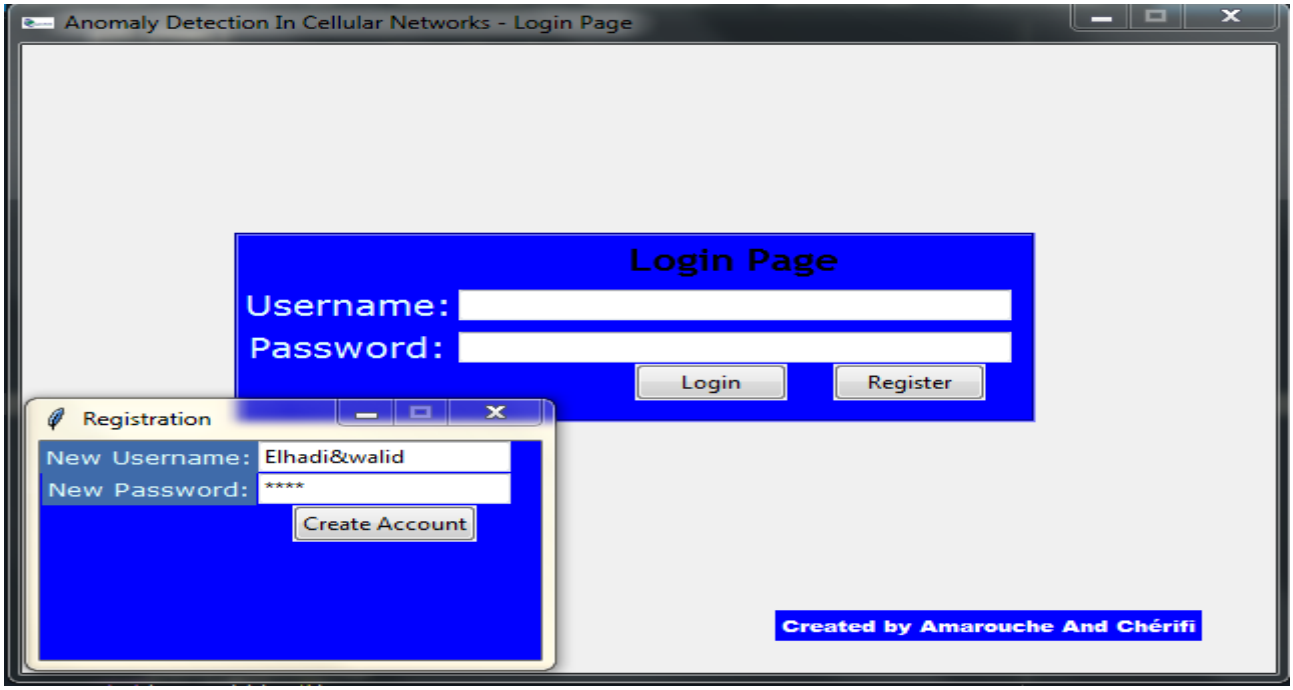


FIGURE 4.8 – La fenêtre de bienvenue.

Cette interface demande à l'utilisateur d'entrer sont pasword ou registration pour accéder à la prochaine interface.

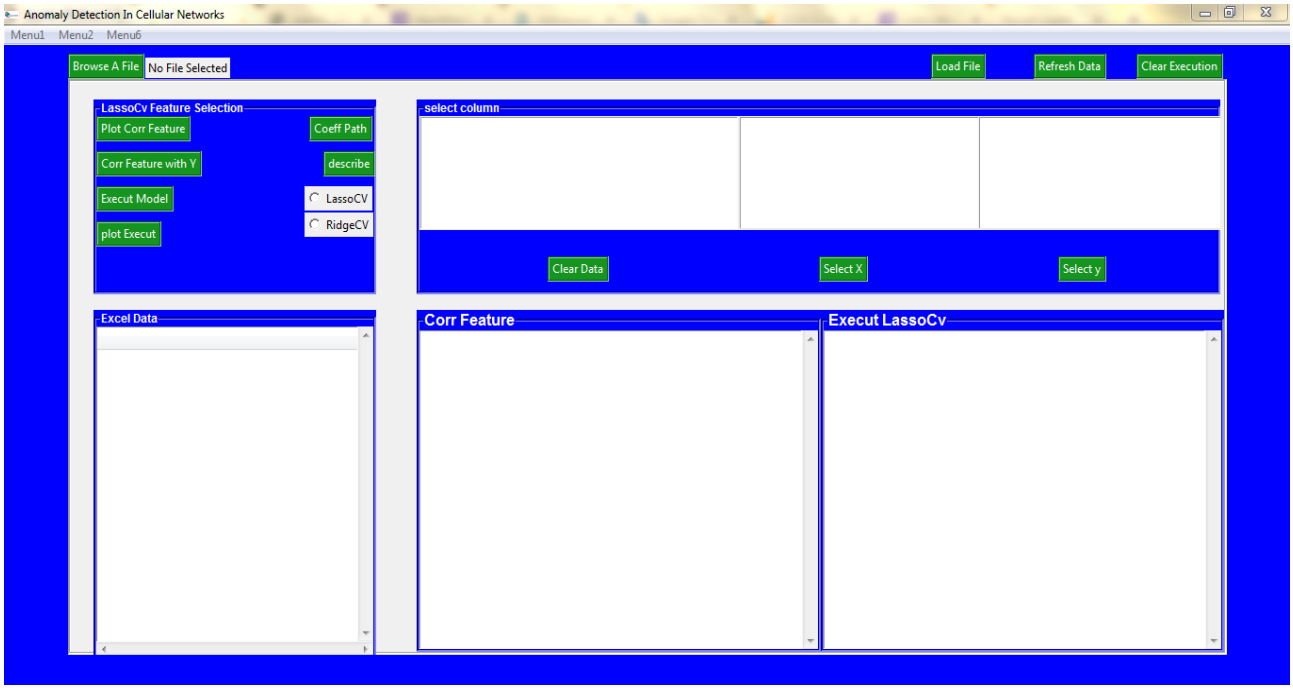


FIGURE 4.9 – Interface du résultat de problème.

Dans cette interface, nous avons plusieurs boutons pour diverses fonctionnalité.

Etapes d'exécution et résultats

1.Load File : chargement du fichier Excel,CSV. **2.Refrech Data :**actualiser les données.**3.Select X :** sélection des colonnes de la matrice X en cliquant sur les colonnes affichées dans le zoon "select column".**4.Select y :** sélection de la colonne y en cliquent

sur une colonne de la zoon "select column"..**5.describe :** décris est génére une description statistique des donnés. (écart-type, moyenne, variable maximale, variable minimale,...).**6. Plot Corr Feature :** tracé la corrélation des KPIs (des colonnes de X).**7. Corr Feature with y** trace la corrélation des colonnes X avec y..**8.LassoCV :** choisir le modèle lasso.**9.RidgeCV :** choisir le modèle Ridge.**10.Execut model :** exécuter le modèle.**11.Coeff Path :** affiche les coefficients du modèle.**12.plot Execut :** tracé les coefficients.**13.Clear Execution** effacer l'exécution.**14. Clear Data** effacer les données.

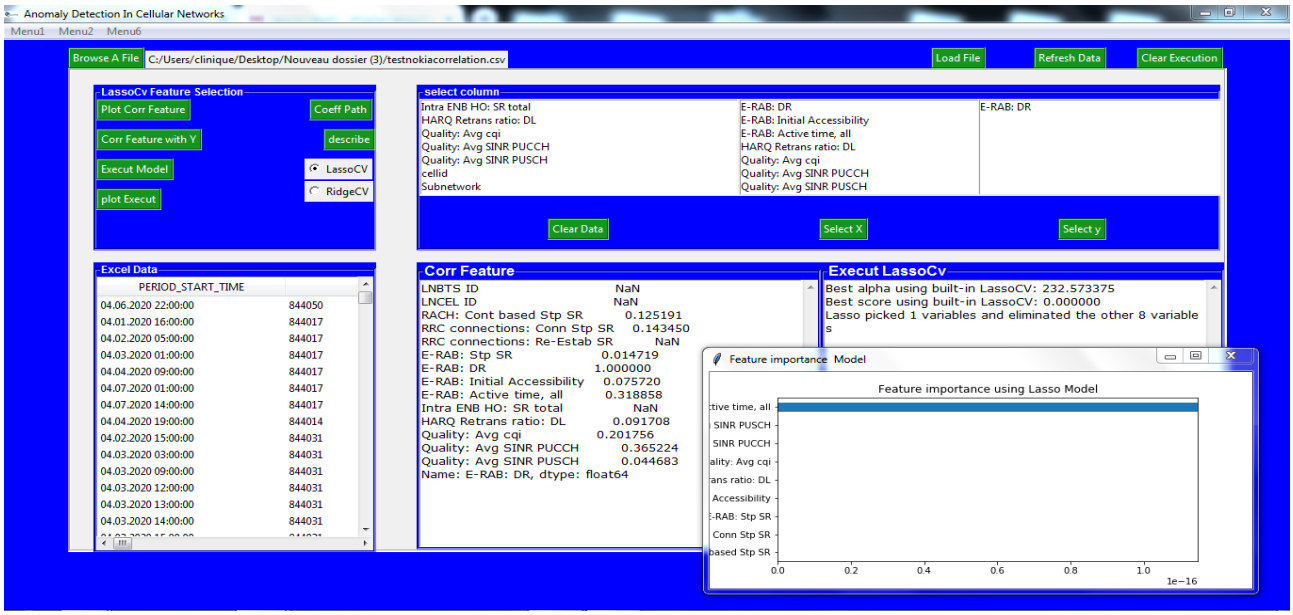


FIGURE 4.10 – Resultats d'exécution 1. Le Lasso détecte que "E-RAB : Active time, all" affecte E-RAB :DR

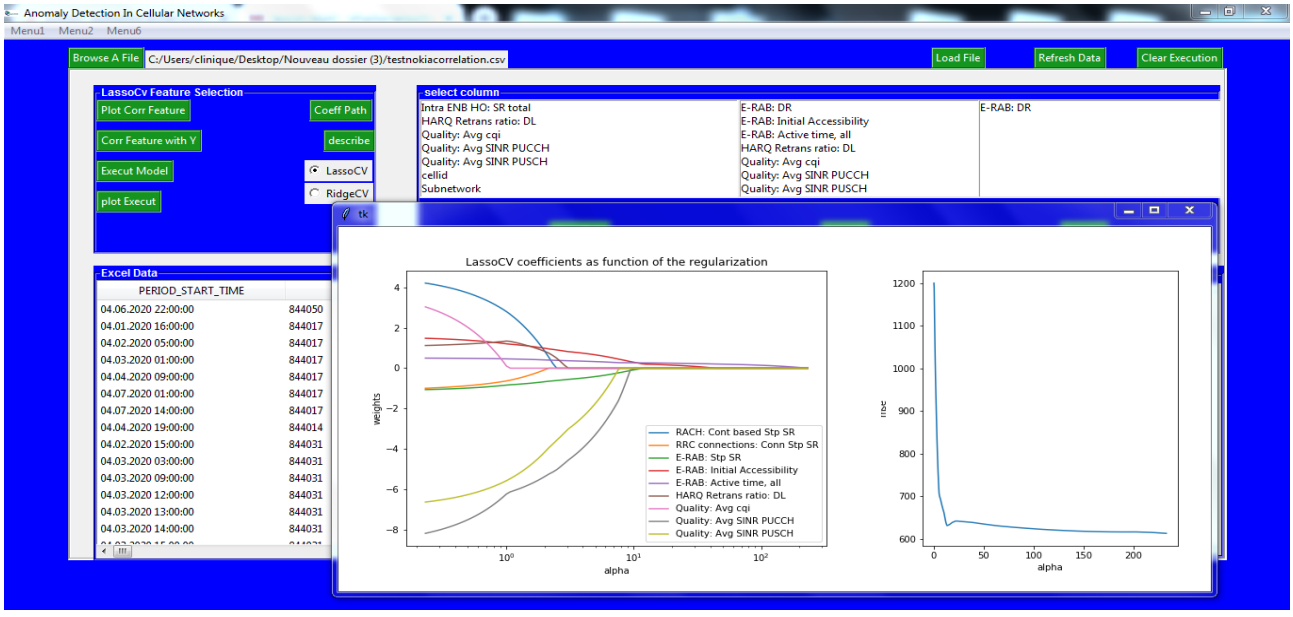


FIGURE 4.11 – Résultats d'exécution2, Les deux figures représentent le chemin de régularisation pour la régression Lasso (i.e comportement des paramètres estimé pour un échantillon des valeurs de lambdas) et le chemin de l'erreur relative pour les mêmes valeurs de lambdas ce qui preuve la convergence de l'algorithme descente par coordonnée et l'efficacité du lasso dans l'élimination des variables non pertinent

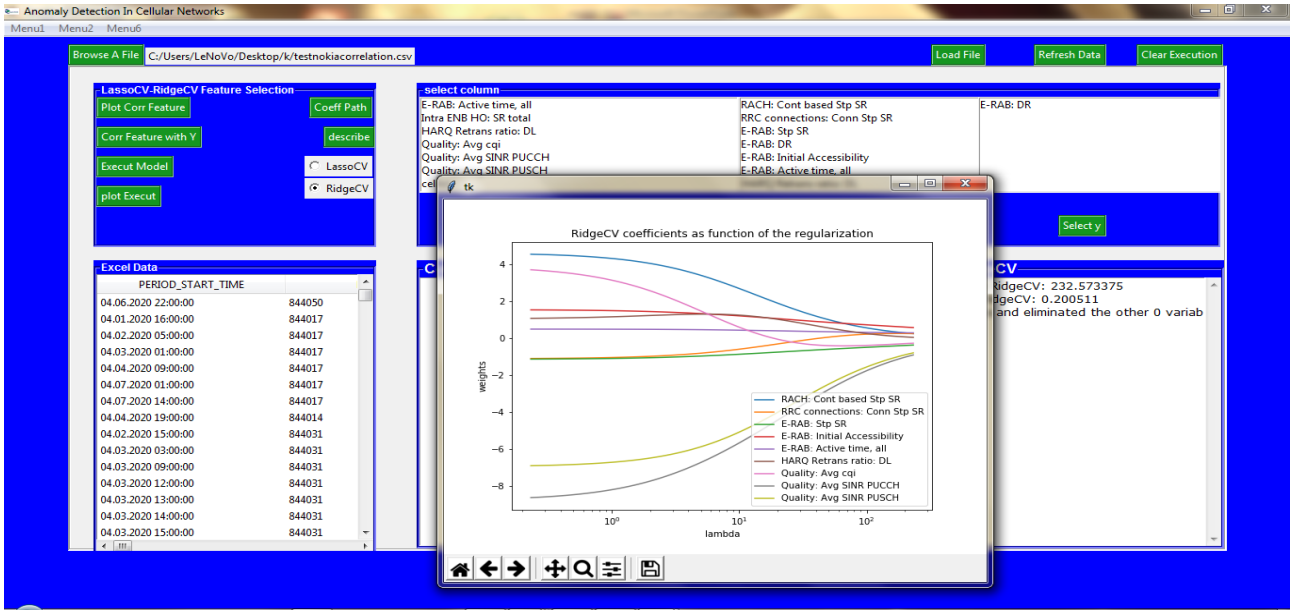


FIGURE 4.12 – Contrairement à la régression Lasso, la régression Ridge rétrécit les coefficients , mais ne les réduit pas à zéro

Conclusion

La statistique est la discipline qui étudie des phénomènes à travers la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre ces données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques. La RO est le consommateur final des solutions statistiques, l'analyse statistique est un étape de construction, d'identification et de test d'un modèle. Les outils mathématiques et numériques de la RO, théorie et algorithmes d'optimisation sont utilisées en statistiques quand les règles de decisions statistiques nécessitent une recherche de "meilleurs solutions". En termes de compétences les objectifs sont d'apprendre de modéliser un problème comme un problème d'optimisation (ou encore, de programmation mathématique), et d'utiliser les outils disponibles (Python, Excel, Machine Learning package, ...) pour le résoudre.

Annexe A

Quelques rappels de calcul différentiel, analyse et optimisation convexe et extremum

A.1 Problèmes d'optimisations :

L'optimisation[6] est plus généralement la recherche opérationnelle intervient lorsque l'outil mathématique est appliqué à une résolution, si le problème soit formalisable mathèmatiquement.

A.1.1 Extremum local et global :

soit f une fonction défini sur un intervalle D de $\mathbb{R}^n$ $f : D \subset \mathbb{R}^n \to \mathbb{R}$, $x_0 \in D$

Définition 2.

- *on dit que f admet un minimum local en x_0 sur D si :*
  $\forall x \in D$ *on a* $f(x_0) \leq f(x)$
- *on dit que f admet un minimum global en x_0 sur D si :*
  $\forall x \in D$ *on a* $f(x_0) < f(x)$
- *on dit que f admet un maximum local en x_0 sur D si :*
  $\forall x \in D$ *on a* $f(x_0) \geq f(x)$
- *on dit que f admet un maximum global en x_0 sur D si :*
  $\forall x \in D$ *on a* $f(x_0) > f(x)$

A.1.2 Matrice (semi) défini positive,(semi) défini négative :

Définition 3.
Soit Q une matrice symétrique $(n \times n)$

- *On dit que Q une matrice défini positive(DP) si :*
  $X^T QX > 0$, $\forall X \in \mathbb{R}^n$, $X \neq 0$
- *On dit que Q une matrice semi défini positive(SDP) si :*
  $X^T QX \geq 0$, $\forall X \in \mathbb{R}^n$
- *On dit que Q une matrice défini négative(DN) si :*
  $X^T QX < 0$, $\forall X \in \mathbb{R}^n$, $X \neq 0$
- *On dit que Q une matrice semi défini négative(SDN) si :*
  $X^T QX \leq 0$, $\forall X \in \mathbb{R}^n$

Caractérisation des optimums :

Définition 4. *Soit $x^* \subset \mathbb{R}^n$ et supposons $f : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable en x^*.on définit le gradient de f en x^* par :*

$$\nabla f(x^*) = [\frac{\partial f(x_1)}{\partial x_1}, ..., \frac{\partial f(x_n)}{\partial x_n}]^T$$

Si f possède un extremum local (ou global) en x^* alors.

$$\nabla f(x^*) = 0$$

les solution de $\nabla f(x^*) = 0$ sont appelées points stationnaires de f

Définition 5. *si $f : \mathbb{R}^n \to \mathbb{R}$ est de classe C^2 alors on définit la matrice Hessiene de f en x par :*

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}$$

Théorème 1.

- *Si f possède un minimum local (ou global) en x^* alors :*
  1. *$\nabla f(x^*) = 0$*
  2. *la matrice hessienne $\nabla^2 f(x^*)$ est semi-définie positive.*
- *Si f possède un maximum local (ou global) en x^* alors :*
  1. *$\nabla f(x^*) = 0$*
  2. *la matrice hessienne $\nabla^2 f(x^*)$ est semi-définie négative.*

A.1.3 Optimisation convexe

Définition 6.
soit $C \subset \mathbb{R}^n$ un ensemble convexe non vide et $f : C \to \mathbb{R}$

- *L'application f est convexe si :*
  $(x, y) \in C$ et $\forall \lambda \subset [0, 1] : f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y)$
- *L'application f est strictement convexe si :*
  $(x, y) \in C$ et $\forall \lambda \subset [0, 1] : f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda) f(y)$

Définition 7.
on dire qu'une fonction f, $f : C \to \mathbb{R}$ définie sur un ensemble convexe C est concave si :

- *L'application f est concave si :*
  $(x, y) \subset C$ et $\forall \lambda \subset [0, 1] : f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda) f(y)$
- *L'application f est strictement concave si :*
  $(x, y) \subset C$ et $\forall \lambda \subset [0, 1] : f(\lambda x + (1 - \lambda)y) > \lambda f(x) + (1 - \lambda) f(y)$

Propriété 1.
Soit f une fonction convexe defini sur un ensemble convexe $C \in \mathbb{R}^n \to \mathbb{R}$

- *L'ensemble M des points ou f atteint son minimum est convexe.*
- *Tout problème strictement convexe admet au plus une solution.*
- *Tout minimum local est un minimum global.*
- *Si f est strictement convexe, alors son minimum global est atteint en un seul point x_0.*

A.1.4 Quelques Notations

1. Pour tous $x, y \in \mathbb{R}^n$ on note par $< x, y > \in \mathbb{R}$ le produit scalaire de x et y, qui est donné par :

$$< x, y >= \sum_{i=1}^{n} x_i y_i$$

Deux vecteurs $x, y \in \mathbb{R}^n$ sont orthogonaux (on notera $x \perp y$) si $< x, y >= 0$.

2. Pour tout $x \in \mathbb{R}^n$ on note par $||x|| \geq 0$ la norme euclidienne de x, donnée par :

$$||x|| = \sqrt{< x, x >} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Désignent par $||x||_k = (\sum_{i=1}^{n} |x_i|^k)^{\frac{1}{k}}$, si $k = 2$, cette norme est la norme euclidienne. Rappellons les propriétés d'une norme (donc aussi de la norme euclidienne) :

(i) $||\lambda x|| = |\lambda| ||x|| \quad \forall \lambda \in \mathbb{R}, \forall x \in \mathbb{R}^n$

(ii) $||x + y|| \leq ||x|| + ||y|| \quad \forall x, y \in \mathbb{R}^n$

(iii) $||0|| = 0 \quad$ et $||x|| \geq 0 \quad$ si $x \in \mathbb{R}^n - \{0\}$

3. Pour tous $x \in \mathbb{R}^n$ et $r > 0$ on notera par $B(x, r)$ la boule ouverte du centre x et rayon r, donnée par :

$$B(x, r) = \{y \in \mathbb{R}^n, ||y - x|| < r\}$$

4. Rappellons aussi l'inégalité de Cauchy-Schwarz :

$$| < x, y > | \leq ||x||.||y|| \quad \forall x, y \in \mathbb{R}^n$$

Bibliographie

[1] Faraz Ahmed, Jeffrey Erman, Zihui Ge, Alex X Liu, Jia Wang, and He Yan. Detecting and localizing end-to-end performance degradation for cellular data services. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9. IEEE, 2016.

[2] Antoine Bonnefoy. *Elimination dynamique : accélération des algorithmes d'optimisation convexe pour les régressions parcimonieuses*. PhD thesis, Aix-Marseille, 2016.

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection : A survey. *ACM computing surveys (CSUR)*, 41(3) :1–58, 2009.

[4] Haiwen Chen, Guang Yu, Fang Liu, Zping Cai, Anfeng Liu, Shuhui Chen, Hongbin Huang, and Chak Fong Cheang. Unsupervised anomaly detection via dbscan for kpis jitters in network managements. *Computers, Materials & Continua*, 62(2) :917–927, 2020.

[5] Nina CİCHOCKİ. A thesıs submıtted to the faculty of the graduate school of the unıversıty of mınnesota by. *The Life Story Of The Çemberlitaş*, 2005.

[6] Ionel Sorin CIUPERCA. Cours optimisation. page 68.

[7] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2) :407–499, 2004.

[8] Sameh Faidi. *Finding Anomalous eNodeBs*. Metropolia Ammattikorkeakoulu, 2018.

[9] Valeria Fonti and Eduard Belitser. Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics*, 30 :1–25, 2017.

[10] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1) :1, 2010.

[11] Wenjiang J Fu. Penalized regressions : the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3) :397–416, 1998.

[12] Kevin Bellamy Guth. Anomaly detection using robust principal component analysis. 2018.

[13] Anders Høst-Madsen, Elyas Sabeti, and Chad Walton. Data discovery and anomaly detection using atypicality : Theory. *IEEE Transactions on Information Theory*, 65(9) :5302–5322, 2019.

[14] Shi Jin, Zhaobo Zhang, Krishnendu Chakrabarty, and Xinli Gu. *Anomaly-Detection and Health-Analysis Techniques for Core Router Systems*. Springer, 2020.

[15] Rachid Kharoubi. Une nouvelle approche pour la sélection des variables dans le cas de modèles de discrimination en grandes dimensions. 2016.

[16] Dapeng Liu, Youjian Zhao, Haowen Xu, Yongqian Sun, Dan Pei, Jiao Luo, Xiaowei Jing, and Mei Feng. Opprentice : Towards practical and automatic anomaly detection through machine learning. In *Proceedings of the 2015 Internet Measurement Conference*, pages 211–224, 2015.

[17] Junshui Ma and Simon Perkins. Time-series novelty detection using one-class support vector machines. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 3, pages 1741–1745. IEEE, 2003.

[18] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv :1607.00148*, 2016.

[19] Masataka Ohta, Fumitoshi Saito, Ken'ya Nishiki, Ken'ichi Yoshida, and D Eng. Operations support system solutions for ip networks. *Hitachi Review*, 49(4) :169, 2000.

[20] Kayur Patel, James Fogarty, James A Landay, and Beverly Harrison. Investigating statistical machine learning as a tool for software development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 667–676, 2008.

[21] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn : Machine learning in python. *the Journal of machine Learning research*, 12 :2825–2830, 2011.

[22] Damini Rai and Abhishek Dwivedi. Lte theory to practice-kpi optimization (a 4g wireless technology). *Int. J. Innov. Technol. Explor. Eng*, 8(2) :1–20, 2018.

[23] Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. Time-series anomaly detection service at microsoft. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3009–3017, 2019.

[24] Diptarka Saha, Debanjana Banerjee, and Bodhisattwa Prasad Majumder. Redclan-relative density based clustering and anomaly detection.

[25] Marina Thottan and Chuanyi Ji. Anomaly detection in ip networks. *IEEE Transactions on signal processing*, 51(8) :2191–2204, 2003.

[26] Marina Thottan, Guanglei Liu, and Chuanyi Ji. Anomaly detection approaches for communication networks. In *Algorithms for next generation networks*, pages 239–261. Springer, 2010.

[27] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3) :475–494, 2001.

[28] Jun Wu, Patrick PC Lee, Qi Li, Lujia Pan, and Jianfeng Zhang. Cellpad : Detecting performance anomalies in cellular networks via regression analysis. In *2018 IFIP Networking Conference (IFIP Networking) and Workshops*, pages 1–9. IEEE, 2018.

[29] Tong Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10(3), 2009.

[30] Changliang Zou and Peihua Qiu. Multivariate statistical process control using lasso. *Journal of the American Statistical Association*, 104(488) :1586–1596, 2009.