École nationale de la statistique
et de l'analyse de l'information

Projet de de fin d'études

Analyse de données d'essais moteurs

Groupe 15 :
Félicie Bizeul
Maël Cayrefourcq
Clotilde Patarin
Baptiste Wolf

Tutrice :
Catherine Duveau

2020-2021

Remerciements

Nous tenons à remercier notre encadrant Arthur KATOSSKY pour l'organisation et le bon déroulement des différentes phases du projet.

Nous remercions également notre tutrice Catherine DUVEAU ainsi qu'Emmanuel MARGERIE et Laurent DESCAMPS, pour l'aide et le soutien apportés tout au long de ce projet. Un merci tout particulier à Catherine DUVEAU pour sa présence et son accompagnement important et régulier.

Abstract

SAFRAN a pour ambition de développer de nouveaux moteurs d'hélicoptères. Pour cela, de nombreuses analyses sont effectuées sur chacun des composants du moteur. Dans le cadre de ce projet, nous analysons l'évolution de la température et de la pression de la turbine au cours d'un essai, composant indispensable pour le fonctionnement d'un turboréacteur. De nombreux capteurs de pression et de température ont été fixés sur le matériau par les ingénieurs de *SAFRAN*. Les mesures liées à ces capteurs forment alors de grosses bases de données. Il est fréquent que des capteurs se décrochent, ou ne soient plus en état de fonctionnement au cours d'un essai.

L'un des objectifs de ce projet est d'identifier ces capteurs défectueux que le régime du moteur soit stabilisé ou transitoire, à travers différentes méthodes de détection de valeurs aberrantes. Une fois les capteurs identifiés, il s'agit d'estimer les valeurs qu'ils auraient dû prendre selon certaines données d'entrées, telles que la pression extérieure, la température extérieure, le régime moteur, et la position du capteur. Pour ce faire, nous développons des modèles afin de prédire un comportement moteur.

Table des matières

Remerciements **I**

Abstract **II**

Introduction **1**

1 Contexte de l'étude **2**
1.1 L'étude . . . 2
1.2 Les données . . . 2

2 Analyse des données en régime stabilisé **4**
2.1 Des données fortement dépendantes les unes des autres . . . 4
2.2 Détection des anomalies . . . 5
2.2.1 Travail sur les variables à l'écart . . . 5
2.2.2 Travail sur les observations en écart . . . 7
2.2.3 Capteurs et observations en écart : conclusion . . . 8
2.3 Prédictions du comportement des moteurs . . . 8
2.3.1 Partial Least Squares . . . 8

3 Analyse des données en régime transitoire **12**
3.1 Détection des anomalies . . . 12
3.1.1 Travail sur les variables à l'écart . . . 12
3.1.2 Distinction de groupes de capteurs selon leur temps de réponse . . . 13
3.2 Détection d'anomalies parmi les groupes de capteurs . . . 15
3.2.1 Détection des anomalies par analyse de la variance intra-groupes . . . 15
3.2.2 RMSE . . . 16
3.2.3 Étude des écarts entre les instants de réponse . . . 16
3.2.4 Détections des capteurs atypiques : conclusion . . . 17
3.3 Appréciation de la stabilisation par groupes de capteurs . . . 18
3.3.1 L'algorithme de détection de changement de point . . . 18
3.3.2 L'algorithme PELT pour des mesures allant jusqu'à 8000 secondes de durée . . . 21
3.3.3 Détection d'instants atypiques sur des phases de température stables . . . 21
3.4 Prédictions du comportement des moteurs : une mesure transitoire en fonction des autres . . . 22
3.4.1 Modèle Lasso . . . 23
3.4.2 Résultats . . . 24

Conclusion **25**

4 Annexes **27**
4.1 Compléments à l'analyse descriptive des données . . . 27
4.2 Complément à la détection d'anomalies en régime stabilisé : capteurs et observations . . . 28
4.2.1 Quelques graphiques . . . 28
4.2.2 Compléments sur l'étude des observations à l'écart en régime stabilisé . . . 30

4.2.3 Détection des individus aberrants par Leave-One-Out selon les variances 31
4.3 Complément à la prédiction du comportement des moteurs en régime stabilisé 36
4.3.1 La régression PLS 36
4.4 Compléments à la détection d'anomalies en régime transitoire : capteurs et observations . . . 36
4.4.1 Quelques graphiques 36
4.4.2 Complément à l'étude des résultats de l'algorithme PELT 39
4.4.3 Complément à la détection d'instants atypiques parmi les groupes de capteurs 39
4.5 Une mesure transitoire estimée selon la moyenne des autres 40

Table des figures

1.1 Exemple d'un graphe de capteur normal . . . 3
1.2 Exemple d'un graphe de capteur défectueux . . . 3
1.3 Exemple d'un graphe d'un capteur normal et ses différentes phases . . . 3

2.1 Heatmap et clustering hiérarchique des capteurs YB_92 . . . 4
2.2 Étendue de chacune des variables . . . 6
2.3 RMSE de nos variables expliquées par les autres variables . . . 6
2.4 Valeur du Q^2 en fonction du nombre de composantes . . . 9
2.5 Critère VIP pour l'ensemble des variables de la table X . . . 10
2.6 Valeurs prédites en fonction des valeurs de l'échantillon de validation . . . 11
2.7 Valeurs prédites en fonction des valeurs de l'échantillon de validation . . . 11

3.1 Variances de chacun de capteurs . . . 12
3.2 Comportement des capteurs anormaux en comparaison avec des capteurs normaux . . . 13
3.3 A gauche : Dendrogramme post-clustering des variables à expliquer en régime transitoire A droite : Dendogramme regroupant nos capteurs de même temps de réponse . . . 14
3.4 Exemples de graphes de capteurs du groupe 1, 2 et 3 respectivement . . . 14
3.5 Variances intra-groupes selon les capteurs du groupe 1 (à gauche), du groupe 2 (au milieu) et du groupe 3 (à droite) . . . 15
3.6 RMSE selon les capteurs du groupe 1 (à gauche), du groupe 2 (au milieu) et du groupe 3 (à droite) . . . 16
3.7 Somme des écarts entre instants selon les capteurs du groupe 1 (à gauche), du groupe 2 (au milieu) et du groupe 3 (à droite) . . . 17
3.8 Exemple d'un signal pour lequel BS n'a pas détecté les bonnes zones . . . 18
3.9 Exemple d'un signal pour lequel BS a détecté les bonnes zones . . . 18
3.10 Analyse du bon fonctionnement de pelt . . . 19
3.11 Variance de l'indicateur selon les capteurs du groupe 2 . . . 20
3.12 Comparaison performance pelt . . . 21
3.13 Graphes des capteurs YA_92_30_1 et YB_92_17_1 avec un zoom sur les observations après 6000 secondes . . . 22
3.14 Évolution de la MSE en fonction de la valeur de la pénalité selon les capteurs du groupe 1 (à gauche), du groupe 2 (au milieu) et du groupe 3 (à droite) . . . 23
3.15 Boxplot des valeurs prises par la RMSE pour chacune des modélisations des capteurs. À gauche pour la régression linéaire, à droite pour le modèle Lasso . . . 24

4.1 Matrice de corrélations entre les variables YB_92 . . . 27
4.2 Pourcentage de variance expliquée par chacun des axes de l'ACP sur les capteurs YB_92 . . . 28
4.3 Graphe d'un des capteurs détectés en écart en régime stabilisé . . . 28
4.4 Corrélation entre les variables explicatives . . . 29
4.5 Graphes de l'ensemble des RMSE calculés par LOO sur chacun des 4 capteurs détectés en écart 30
4.6 Graphes de l'ensemble des variances calculées par LOO sur chacun des 4 capteurs détectés en écart . . . 31
4.7 ACP sur l'ensemble des capteurs YB_92 (graphe des individus, axes 2 et 3) . . . 32

4.8 Local Outlier Factor des observations sur les capteurs Yb_92 33
4.9 ACP sur l'ensemble des capteurs YB_92 (graphe des individus, axes 2 et 3) 33
4.10 Score des observations par la méthode de l'Isolation Forest sur les capteurs YB_92 34
4.11 ACP sur l'ensemble des capteurs YB_92 (graphe des individus, axes 2 et 3) 35
4.12 Cercle des corrélations et carte des observations respectivement 36
4.13 Graphes de deux capteurs du groupe 1 détectés en écart après analyse de leurs variances . . . 36
4.14 Graphes des deux autres capteurs du groupe 1 détectés en écart après analyse de leurs variances 37
4.15 Graphes de deux capteurs du groupe 2 détectés en écart après analyse de leurs variances . . . 37
4.16 Graphes de deux capteurs du groupe 2 détectés en écart après analyse de leurs variances . . . 37
4.17 Graphes des deux capteurs du groupe 3 détectés en écart après analyse de leurs variances . . 38
4.18 Graphes des trois capteurs détectés en écart par chacune des méthodes 38
4.19 Capteurs pouvant comporter des anomalies selon PELT 39
4.20 Scores d'anomalies des capteurs YA_92_30_1 et YB_92_17_1 39
4.21 Scores d'anomalies des capteurs YA_92_57_0 et YB_92_48_0 40
4.22 Graphes des capteurs YA_92_57_0 et YB_92_48_0 avec un zoom sur les observations après 6000 secondes 40
4.23 Figure montrant la performance de prédiction de 2 capteurs du groupe 1 41

Liste des tableaux

2.1 Individus notables pour chacune des variables retenues comme anormales 7

3.1 Tableau récapitulatif des variables en écart par groupes de capteurs et par méthodes 17
3.2 Comparaison des régressions Linéaire et Lasso à partir du critère MAPE (en %) 24

Introduction

Pour la conception d'un turboréacteur, plusieurs moteurs dits de développement sont nécessaires avant d'aboutir à la définition série. Ces moteurs sont largement instrumentés avec de nombreux capteurs de pression, de température d'air et de température métal sur différentes pièces afin de comprendre le comportement du moteur et de valider/recaler les prédictions/modélisations numériques. Ces capteurs réalisent des milliers de mesures chaque jour qu'il convient d'étudier. Faciliter l'analyse de ces mesures est l'objectif demandé par *SAFRAN*.

"Motoriste aéronautique depuis plus de 110 ans, Safran, via sa filiale Aircraft Engines, conçoit, développe, produit et commercialise, seul ou en coopération, des moteurs pour avions civils et militaires, ainsi que pour satellites. Face aux enjeux d'un marché aéronautique en plein développement, Safran Aircraft Engines prépare l'avenir et anticipe les besoins de ses clients, en investissant continuellement dans le développement de nouvelles technologies et en concevant des moteurs toujours plus respectueux de l'environnement."[1]

Pour réaliser cette étude, nous procéderons en plusieurs étapes. Nous commencerons par une analyse des données en régime stabilisé, puis en régime transitoire. Dans le premier cas, le moteur fonctionne sur un régime constant pendant une durée suffisamment longue pour que toutes les mesures soient stables tandis que dans le deuxième cas le régime moteur et les mesures évoluent au cours du temps avec une fréquence d'acquisition de 1 Hz. Chacune de ces deux études consistera en la détection d'observations et/ou de variables en écart et en la prédiction du comportement des moteurs pour par exemple imputer d'éventuelles valeurs manquantes.

Chapitre 1

Contexte de l'étude

1.1 L'étude

Comme présenté en introduction, le but de l'étude est de faciliter l'analyse des mesures récoltées par les différents capteurs sur les moteurs d'avions. Or, le nombre de ces mesures est relativement important, avec environ 500 types de mesures (capteurs) différents sur un moteur. L'acquisition de ces mesures est donc réalisée en "régime stabilisé", mais également en "transitoire" et sont réalisées durant de nombreuses séquences d'essais pendant plusieurs jours voire semaines.

Il arrive que des capteurs soient défectueux pendant une partie de l'essai ou définitivement perdus (données manquantes) au cours de la campagne. Actuellement ces données sont analysées manuellement, et uniquement de façon uni ou bivariée. Pour l'étude, nous distinguons les trois phases principales suivantes :

1. la détection automatique de valeurs en écart : détection des observations et variables en écart dans un cadre multivarié
   - En régime stabilisé : nous souhaitons ici détecter, soit des mesures ponctuellement en écart, soit toujours en écart, soit qui dérive à partir de certains points d'essais
   - En régime transitoire : nous souhaitons ici détecter soit les mesures très bruitées, soit celles incorrectes ponctuellement, soit celles qui dérivent dans le temps
2. la détection automatique de ruptures dans les signaux pour identifier les périodes de défaillance des capteurs (en régime transitoire notamment grâce à l'algorithme PELT)
3. la prédiction du comportement du moteur, à partir d'un modèle de prédiction de certaines variables en fonction des données d'entrée moteur (mise en place des modèles, sélection des variables, et validation des modèles)

1.2 Les données

Les données dont nous disposons ont été anonymisées. Nous avions initialement l'intégralité des variables correspondant à chacun des capteurs. Cependant, nos tuteurs ont par la suite décidé de restreindre l'étude aux seuls capteurs Y_92, correspondant à des températures rotor, que le régime soit stabilisé ou transitoire. Aucune des variables retenues ne présentaient de valeurs manquantes.

En régime stabilisé, nous comptons 107 observations nommées, correspondants à des points d'essais (associés à leur numéro) et 81 variables dont 45 sont les mesures récoltées par les capteurs YB_92 (variables à expliquer) et 36 sont les variables de données d'entrée moteur (variables explicatives X, A et V).
En régime transitoire, nous comptons 9597 observations qui sont des valeurs mesurées en fonction du temps par les capteurs, 83 variables à expliquer YA_92 et YB_92, ainsi que les données d'entrée moteur correspondants aux 18 variables explicatives X, A et V.

Afin de visualiser les comportements des capteurs, voici aux figures 1.1 et 1.2, en régime transitoire, les graphes d'un capteur normal et d'un capteur défectueux que nous nous devons de relever.

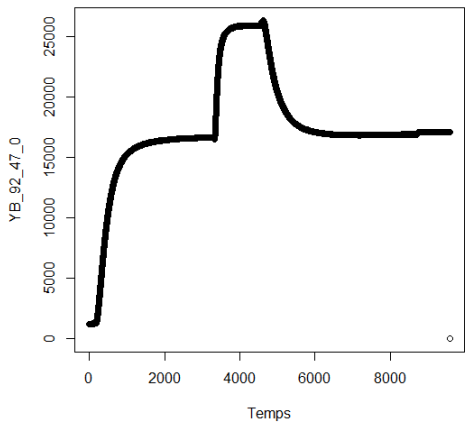

FIGURE 1.1 – Exemple d'un graphe de capteur normal
Source : SAFRAN

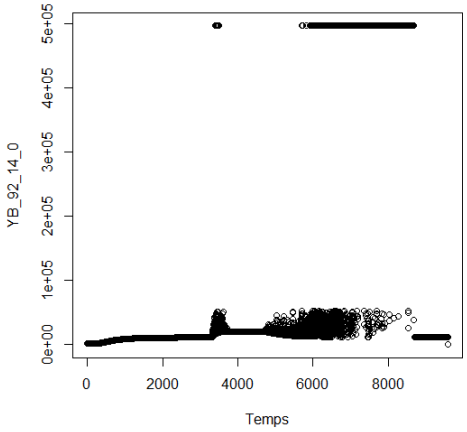

FIGURE 1.2 – Exemple d'un graphe de capteur défectueux
Source : SAFRAN

Pour détecter les périodes de rupture des signaux et identifier ainsi les périodes de défaillance des capteurs, nous étudierons par la suite les différentes phases des prises de mesure en transitoire, et nous nous pencherons sur celles où la température est stabilisée. Étudier ces zones censées être stables pour tous les capteurs devraient nous permettre de détecter facilement et automatiquement les capteurs ayant connu des problèmes.

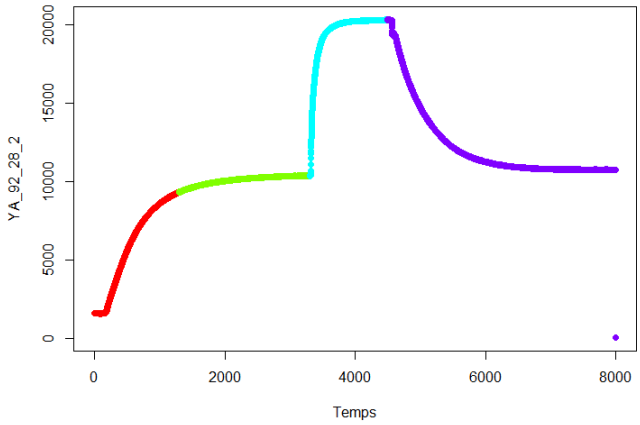

FIGURE 1.3 – Exemple d'un graphe d'un capteur normal et ses différentes phases
Source : SAFRAN

D'après la figure 1.3, nous identifions a priori quatre phases. La première correspond au lancement du moteur, le temps de chauffe, puis on observe une première stabilisation de la température. Le grand saut suivant correspondrait au décollage de l'avion pour lequel on observe également par la suite un temps de stabilisation de la température. La dernière chute pourrait correspondre à une diminution de la température du moteur puis à une stabilisation jusqu'à ce que l'avion atteigne sa vitesse de croisière.
Nous allons donc maintenant commencer la détection de valeurs en écart pour les deux régimes.

Chapitre 2

Analyse des données en régime stabilisé

Pour rappel, être en régime stabilisé signifie que le moteur fonctionne sur un régime constant pendant une durée suffisamment longue pour que toutes les mesures soient stables. Les observations sont ce qu'on appelle des points d'essais et les variables sont les différents capteurs mesurant par exemple des températures en ces différents points.

Dans la partie suivante, l'ensemble des analyses a été réalisé sur le sous-ensemble de 45 variables YB_92.

2.1 Des données fortement dépendantes les unes des autres

Nous commençons par un premier travail de description de nos données pour se les approprier. Tout d'abord, nous regardons la matrice des corrélations de nos capteurs YB_92 à travers la figure 4.1, page 27 en annexe. Nous notons immédiatement que les données sont très corrélées entre elles puisque les coefficients de corrélation approchent, pour la plupart, 1.

Afin de confirmer cette tendance, nous réalisons un partitionnement des données sur notre jeu YB_92. Cette méthode de clustering hiérarchique nous permet de mieux visualiser la corrélation entre nos variables. Nous obtenons dans un premier temps la heatmap suivante :

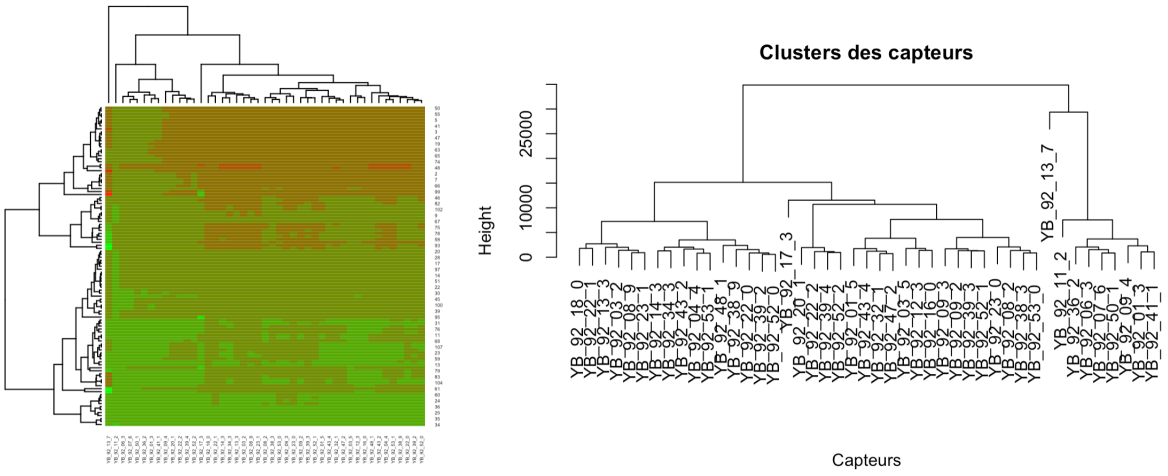

FIGURE 2.1 – Heatmap et clustering hiérarchique des capteurs YB_92
Source : SAFRAN / Observations : 107 observations nommées, 45 variables

Ce cluster permet de regrouper les capteurs. On voit ainsi apparaître des frontières horizontales qui confirment l'existence de trois catégories de capteur. Afin de mieux visualiser ce clustering, nous isolons le dendrogramme relatif aux capteurs du groupe YB_92. Cela va ainsi permettre d'illustrer l'arrangement des différents groupes générés par le regroupement hiérarchique. Celui-ci permet de porter plusieurs conclusions. Tout d'abord, nous pouvons isoler trois groupes distincts, à l'image de ce que nous avons pu observer sur la heat map. Ensuite, nous observons que les capteurs **YB_92_17_3**, **YB_92_13_7**, **YB_92_11_2** semblent isolés. Ces différentes tendances seront confirmées dans la suite de ce rapport.

Nous avons également lancé une ACP sur nos données et nous nous étions attardés à étudier le pourcentage de variance expliquée par chacun des axes, dont le graphe est disponible en annexe 4.2, page 28. On observe sur celui-ci que l'axe 1 explique plus de 96% de la variance totale. Cela montre une fois de plus la forte dépendance de nos données.

Maintenant qu'il est acquis que nos données sont fortement corrélées entre elles, nous allons nous servir de cette caractéristique pour détecter des variables ou des observations en écart. Au vu de la très forte corrélation, le moindre écart sera détectable car la variable ou l'observation en question se détachera grandement du lot.

2.2 Détection des anomalies

Comme présenté dans la partie sur le contexte de l'étude, nous souhaitons ici détecter, soit des mesures ponctuellement en écart, soit toujours en écart, soit qui dérivent à partir de certains points d'essais. Pour cela, nous allons travailler sur les variables en écart, mais aussi sur les individus en écart.

Pour répondre à ces objectifs, nous allons utiliser plusieurs méthodes que nous introduirons au fur et à mesure. Pour les variables en écart, nous étudions la variance des capteurs et la RMSE des capteurs expliqués par les autres. Pour les observations en écart, nous ré-étudions la variance et la RMSE des capteurs mais en Leave-One-Out afin de détecter la ou les observations qui rendraient la variance ou la RMSE d'un de nos capteurs différents des autres. Nous testerons également des méthodes plus exploratoires comme le Local Outlier Factor et l'Isolation Forest avec une représentation sur les axes de notre ACP.

Nous n'avons pas de références de normalités connues pour réaliser une telle analyse des variables et observations en écart, mais nous allons user de la forte dépendance entre toutes les mesures pour détecter les anomalies.

2.2.1 Travail sur les variables à l'écart

Nous regardons ici si l'on retrouve des différences en terme de variance entre nos capteurs mais aussi en terme de RMSE lorsqu'un capteur est expliqué par l'ensemble des autres capteurs.

1. Analyse de la variance

Dans un premier temps, nous nous intéressons aux variances des variables du groupe YB_92. D'après le graphe 2.2 à gauche ci-dessous, la plupart des variables a une étendue de valeurs similaire. Seules les valeurs de deux variables semblent réellement différentes des autres. Il s'agit des variables YB_92_13_7 et YB_92_17_3

Si l'on s'intéresse, cette fois, à la dispersion des variables en termes de variance, le graphique 2.2 à droite ci-dessous souligne que la variable YB_92_13_7 est réellement différente. En effet, elle a une variance beaucoup plus importante que chacune des autres variables.

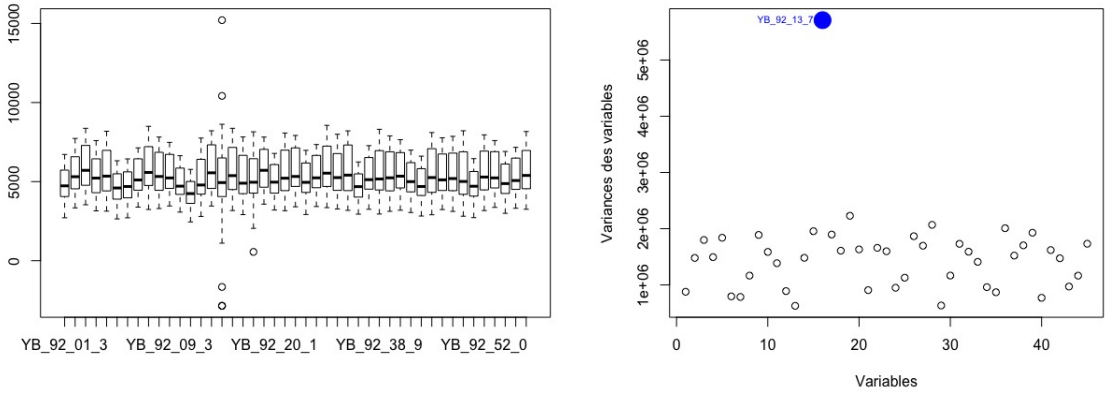

FIGURE 2.2 – Étendue de chacune des variables
Source : SAFRAN / Observations : 107 observations nommées, 45 variables

Il a été tout de suite confirmé par nos tuteurs que le capteur YB_92_13_7 était bien en écart des autres. Cette première méthode ets donc efficace pour détecter les capteurs en écart de façon assez extrême. Nous allons essayer une autre méthode pour peut-être détecter des capteurs qui seraient en écart moins important que le premier.

2. Détection de valeurs aberrantes en calculant les RMSE

Dans cette partie, nous approchons le problème par le calcul de la RMSE des variables. En effet, pour chaque variable, nous allons l'expliquer en fonction des autres. Nous implémentons une régression linéaire pour chaque variable en fonction des autres. Suite à cela, nous calculons les RMSE, et les comparons entre elles afin d'identifier des variables en écart. Nous espérons des valeurs de RMSE très faibles puisque les variables sont fortement corrélées entre elles, et de plus, nous notons certaines redondances de capteurs ainsi que des mesures très proches.

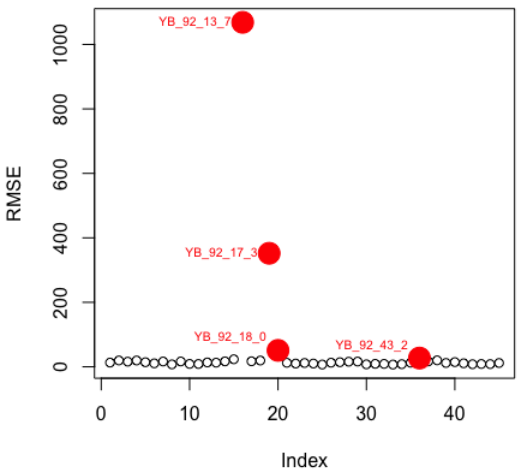

FIGURE 2.3 – RMSE de nos variables expliquées par les autres variables
Source : SAFRAN / Observations : 107 observations nommées, 45 variables

Nous constatons ici que deux variables sont particulièrement différentes des autres. Deux autres se détachent également de l'ensemble. En somme, nous notons les quatre variables suivantes comme ayant des comportement anormaux : YB_92_13_7, YB_92_13_7, YB_92_13_7 et YB_92_43_2. On relève bien à nouveau le capteur détecté par l'analyse des variances, ce qui confirme une fois de plus l'efficacité de nos méthodes.

Pour chacune des variables relevées par les deux méthodes précédentes, nous cherchons dès lors les observations en écart. Ces instants atypiques pourraient expliquer l'anormalité des capteurs.

2.2.2 Travail sur les observations en écart

Une observation en écart est une mesure qui ne suivrait pas la même tendance pour chacun des capteurs observés, que la plupart des autres observations.

Dans un second temps, nous allons regarder sur nos variables détectées comme en écart, les observations qui pourraient être la cause de cet écart. Pour cela, nous allons user de différentes méthodes expliquées dans ce qui suit, à savoir : détection par Leave-One-Out selon le RMSE et la variance des capteurs, ACP, Isolation Forest et Local Outlier factor.

1. Détection des individus aberrants par Leave-One-Out selon la RMSE

Comme expliqué ci-dessus, nous cherchons tout d'abord les observations qui pourraient expliquer l'anormalité des variables détectées précédemment.

Pour cela, nous appliquons la méthode du Leave-One-Out (LOO) sur les 4 variables. Nous calculons en fait la RMSE de chacune des variables 107 fois (correspondant donc aux nombres d'observations total) en enlevant à chaque fois une observation. Nous avons remarqué dans l'étude précédente que la RMSE de ces variables était particulièrement élevée. En calculant celle-ci avec toutes les observations sauf une à chaque fois, nous cherchons à identifier si l'on atteste d'une forte diminution de la RMSE. De la sorte, nous déterminons si une observation a un poids particulièrement important par rapport aux autres observations. Cette observation en question peut être donc être considérée en écart des autres.

Afin d'identifier facilement ces observations, nous représentons graphiquement les 107 RMSE calculées pour chacune des observations sur chacun des quatre capteurs. Nous repérons les RMSE les plus faibles associées à la suppression de telle ou telle observation, et nous retenons alors celles-ci comme en écart. Ces graphes sont disponibles en annexe, figure 4.5, page 30. Nous récapitulons dans le tableau 2.1 ci-dessous les individus aberrants pour chaque capteur détecté.

		YB_92_18_0		**YB_92_43_2**	
YB_92_13_7	**YB_92_17_3**	En rouge	En bleu	En rouge	En bleu
603	617	617	589	617	589
	601	601	503	603	350
	576	576	350	601	277
	603	603	277	576	
	286	286		286	

TABLE 2.1 – Individus notables pour chacune des variables retenues comme anormales
Source : SAFRAN / Observations : 107 observations nommées, 45 variables

Pour la plupart des variables, nous retrouvons les mêmes observations. En effet, pour les quatre variables, nous retrouvons l'observation 603. Pour trois variables, nous notons que les observations identifiées en rouge sur les graphiques précédents sont les mêmes. A nouveau, nous allons tester une autre méthode pour détecter des observations en écart.

2. Détection des individus aberrants par Leave-One-Out selon les variances

Nous appliquons de nouveau la méthode du Leave-One-Out expliquée précédemment, mais cette fois-ci à la variance de chacun de nos quatre capteurs. Pour chacune de nos variables, nous retirons tour à tour une observation et nous calculons de nouveau la variance de la variable sans cet individu, et ceci jusqu'à avoir retiré toutes les observations de notre jeu de données. Ceci nous permettra de comparer les variances entre elles et savoir quels sont les individus atypiques s'il y en a.

On se concentre sur les variables retenues précédemment, celles dont les individus sont les plus différents entre eux, et on relève les 5% d'observations les plus atypiques comme présentés sur les graphiques disponibles en annexe 4.6, page 31.
L'anormalité des variables YB_92_13_7 et YB_92_17_3 est confirmée par nos tuteurs : ces capteurs renvoient en effet ponctuellement des valeurs fortement en écart. Pour les variables YB_92_18_0 et YB_92_43_2, il faudrait faire à l'avenir des graphiques plus avancés pour vérifier et attester réellement de l'anormalité ou si ce n'est qu'une dispersion un peu plus élevée mais pas anormale.

Nous avons par la suite essayé de trouver d'autres observations à l'écart ou de confirmer les résultats précédents, à l'aide de méthodes de détections d'anomalies telles que le Local Outlier Factor ou encore l'Isolation Forest. Néanmoins, les résultats n'étaient pas concluants car l'écart d'un point d'essai est difficilement interprétable, les conditions d'essai étant chaque fois différentes. Les points d'essai détectés comme uniques ne sont pas nécessairement anormaux, c'est pourquoi il est plus intéressant de travailler sur les capteurs directement. L'ensemble des travaux menés est disponible en annexe de la page 32 à 35.

2.2.3 Capteurs et observations en écart : conclusion

Dans cette section, nous proposons une conclusion des analyses des capteurs et observations en écart. Nous notons que quatre capteurs sont particulièrement différents des autres. Il s'agit des capteurs YB_92_13_7, YB_92_17_3, YB_92_18_0, YB_92_43_2. De ces capteurs, nous notons que certaines observations sont atypiques. Des analyses présentées, nous concluons que les observations 603, 617, 576 et 286 sont les plus atypiques. En effet, que ce soit par la méthode de détection des individus aberrants par Leave-One-Out selon les variances ou selon la RMSE, ce sont les observations communes que nous retrouvons. Sur le graphe 4.3, nous mettons en avant ces observations sur le nuage de point d'un des quatre capteurs identifiés.

2.3 Prédictions du comportement des moteurs

Il s'agit de prédire la valeur d'un capteur selon des données d'entrée, qui sont le régime moteur, toutes les données relatives à la météo extérieure, la position du capteur, etc.

Dans un premier temps, nous analysons la matrice des covariables afin d'évaluer la corrélation entre celles-ci. A travers la figure 4.4 disponible en annexe, nous constatons que nos variables explicatives, X, A et V, sont très corrélées entre elles. Avec des données corrélées de la sorte, nous ne pouvons pas utiliser une simple régression linéaire. En effet, dans notre cas, la matrice des variables explicatives n'est pas de plein rang. Or, il s'agit d'une condition nécessaire pour effectuer une régression linéaire. De ce fait, nous allons utiliser une méthode prenant en compte les corrélations : la régression Partial Least Squares (PLS).

2.3.1 Partial Least Squares

La méthode des moindres carrés partiels permet de modéliser conjointement plusieurs variables à expliquer qui possèdent une structure de dépendance entre elles. Deux méthodes en découlent : la régression PLS 1 et la régression PLS 2. La méthode PLS 1 s'applique dans le cas où il n'y a qu'une seule variable dépendante, alors que la régression PLS 2 considère plusieurs variables dépendantes entre elles. Autrement dit, la méthode PLS 1 est un cas particulier de la régression PLS 2.

Dans un premier temps, nous allons nous attarder sur la régression PLS 2 sur l'ensemble des variables **YB_92**, avant d'utiliser la méthode PLS 1 dans un but prédictif vis à vis de la variable **YB_92_01_3**.

Dans cette partie, nous allons considérer la table des YB_92 (q = 45 variables) comme la table des variables à prédire notée $\mathbf{Y} \in \mathbb{R}^{nq}$, et la table regroupant les variables **X_**, **A_** et **V_** (p = 36 variables) notée $\mathbf{X} \in \mathbb{R}^{np}$, celle des variables prédictives.

L'enjeu de la régression PLS est de choisir des coefficients des combinaisons linéaires de sorte à construire des composantes qui maximisent la covariance entre les composantes de $\mathbf{X}$ et de $\mathbf{Y}$. Nous distinguons deux phases de l'algorithme : le choix de la première composante et le choix des composantes suivantes.

Pour la construction de la première composante, nous cherchons les vecteurs w_1 et c_1 tels que

$$\max_{w_1,c_1} \{\text{cov}\,(t_1 = Xw_1, u_1 = Yc_1)\} \text{ avec } \|w_1\|^2 = \|c_1\|^2 = 1 \tag{2.1}$$

On note que les vecteurs t_1 et u_1 sont obtenus via l'analyse Inter-Baterry de Tucker :

— $(Y^tX)^t (Y^tX) w_1 = \lambda_1 w_1$

— $(X^tY)^t (X^tY) c_1 = \lambda_1 c_1$

Nous obtenons finalement $\text{cov}(t_1, u_1) = \lambda_1$

Pour la construction des composantes suivantes (h = 2, ..., N), nous allons procéder de manière itérative. Dans un premier temps, nous allons déterminer les résidus, en considérant $P_i = c_i t_i$:

— $X^{(h)} = \left(Id - P_{t_{h-1}}\right) X^{(h-1)}$

— $Y^{(h)} = \left(Id - P_{t_{h-1}}\right) Y^{(h-1)}$

En reprenant la même démarche que pour la construction de la première composante, l'enjeu est de chercher les vecteurs w_h et c_h tel que :

$$\max_{w_h,c_h} \left\{\text{cov}\left(t_h = X^{(h)}w_h, u_h = Y^{(h)}c_h\right)\right\} \text{ avec } \|w_h\|^2 = \|c_h\|^2 = 1 \tag{2.2}$$

Ces vecteurs sont obtenus de la manière suivante :

— $\left(Y^{(h)t}X^{(h)}\right)^t \left(Y^{(h)t}X^{(h)}\right) w_h = \lambda_1^{(h)} w_h$

— $\left(X^{(h)t}Y^{(h)}\right)^t \left(X^{(h)t}Y^{(h)}\right) c_h = \lambda_1^{(h)} c_h$

La régression PLS cherche donc une suite de composantes PLS qui soient orthogonales entre elles et cela par construction. Ces composantes sont choisies comme maximisant la covariance (empirique) entre Y et une composante t.

Afin d'optimiser le nombre de composante, on considère l'indicateur Q^2, que l'on va chercher à maximiser :

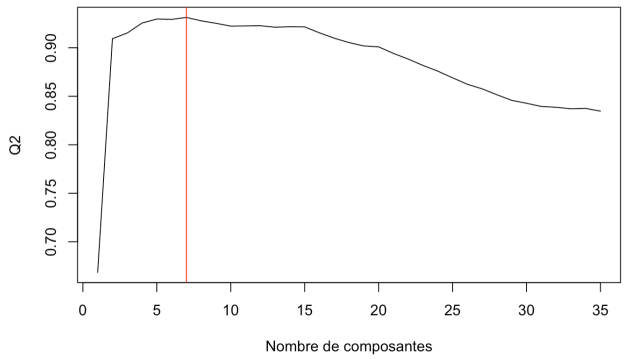

FIGURE 2.4 – Valeur du Q^2 en fonction du nombre de composantes
Source : SAFRAN / Observations : 107 observations nommées, 36 variables

On observe sur la figure 2.4 ci-dessus que le RMSE se stabilise à partir de 5 composantes. La régression PLS 2 permet la construction d'un cercle des corrélations, ainsi qu'une carte des observations, présentés sur la figure 4.12, page 36 en annexe. D'après la carte des observations, il semblerait que les observations 26, 33, 34, 35 et 44 semblent marginales. Nous décidons de les retirer du jeu de données pour la suite de ce chapitre.

L'un des principaux critères apportés par la régression PLS est le critère **VIP** (Variable Importance in the Prediction). Ce critère permet de mettre en évidence l'importance de la $j - ème$ variable dans la prédiction de Y dans un modèle à m composantes. La régression PLS 2 nous permet d'obtenir les indices VIP suivants :

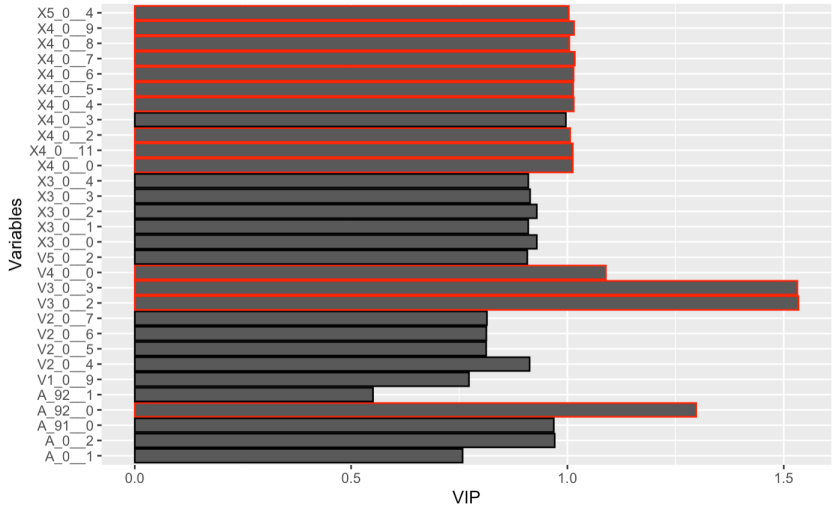

FIGURE 2.5 – Critère VIP pour l'ensemble des variables de la table X
Source : SAFRAN / Observations : 107 observations nommées, 36 variables

Les variables ayant un fort VIP (supérieur à 1 et coloré en rouge ci-dessous) sont les plus importantes dans la construction de Y.

Maintenant que le modèle est en place, nous allons essayer de prédire une variable pour à l'avenir pouvoir potentiellement imputer les valeurs manquantes d'une variable.

Prédiction de la variable YB_92_01_3

Nous allons maintenant utiliser la régression PLS 1 afin de mettre en place un modèle de prédiction pour la variable **YB_92_01_3**. Nous avons choisi cette variable arbitrairement, à titre d'exemple. Dans un premier temps, nous créons un ensemble d'entraînement et un ensemble de test, en retirant les observations 26, 33, 34, 35 et 44 explicitées ci-dessus. Nous allons ainsi construire un échantillon d'entraînement comportant 80% des 102 observations, et les 20% restants seront alloués à l'échantillon de validation.

Nous devons déterminer dans un premier temps le nombre k de composantes PLS à considérer. Pour ce faire, nous procédons par validation croisée. Cela permet d'annuler le biais d'échantillonnage responsable du sur-apprentissage. Le nombre de composantes optimales est donc déterminé par 10 CV. Nous prenons le MSEP comme critère pour savoir quel nombre de composantes permet d'optimiser la prédiction.

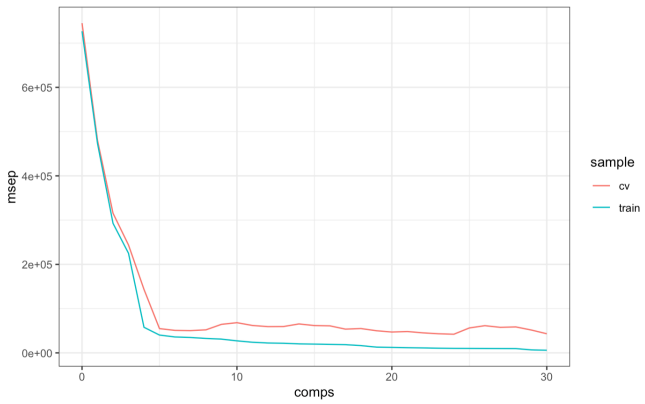

FIGURE 2.6 – Valeurs prédites en fonction des valeurs de l'échantillon de validation
Source : SAFRAN / Observations : 107 observations nommées, 36 variables

D'après la figure 2.6, le nombre de composantes optimales est 24. A travers cette figure, nous remarquons que l'écart entre 24 et 5 composantes est très faible. Nous choisissons pour la suite un modèle à 5 coefficients. Celui-ci sera plus robuste qu'un modèle à 24 coefficients. Nous pouvons à présent évaluer la capacité prédictive du modèle sur notre jeu de validation.

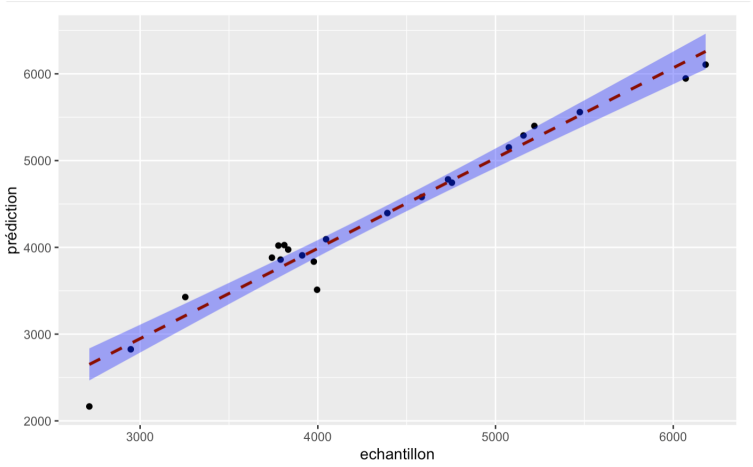

FIGURE 2.7 – Valeurs prédites en fonction des valeurs de l'échantillon de validation
Source : SAFRAN / Observations : 107 observations nommées, 36 variables

De ce modèle prédictif, nous obtenons différents indicateurs mettant en avant les performances de ce dernier. Ainsi, le **RMSE** est de 182,9, le R^2 de 0.961, alors que le **MAPE** (erreur absolue moyenne) est de 3,2%. Notre modèle est donc relativement performant.

Chapitre 3

Analyse des données en régime transitoire

En régime transitoire, le régime moteur et les mesures évoluent au cours du temps avec une fréquence d'acquisition de 1 Hz. Nos données peuvent donc être considérées comme des séries temporelles, car nos variables (nos capteurs) prennent des mesures (par exemple des températures) en un même point à une fréquence régulière.
Les capteurs étudiés sont ici les capteurs YA_92 et YB_92.

3.1 Détection des anomalies

3.1.1 Travail sur les variables à l'écart

Comme en régime stabilisé, nous étudions dans un premier temps la variance de chacun des capteurs pour repérer si certains se détachent du lot. A travers le graphe 3.1 ci-dessous, nous présentons les variances des capteurs YA et YB. Nous remarquons que quatre capteurs ont des variances très importantes comparées à celles des autres capteurs.

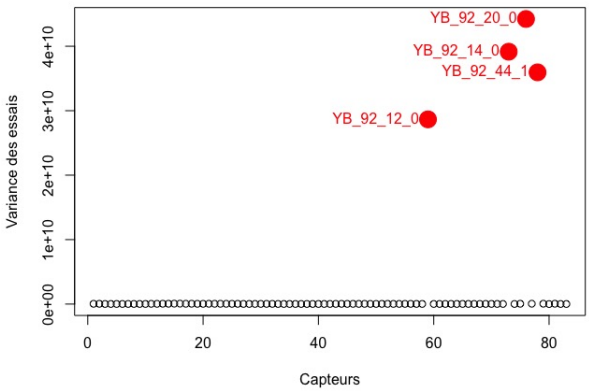

FIGURE 3.1 – Variances de chacun de capteurs
Source : SAFRAN / Observations : 9597 observations nommées, 83 variables

Nous étudions ci-dessous, figure 3.2, les évolutions de comportement de ces 4 capteurs repérés comme anormaux de part leurs variances très élevées.

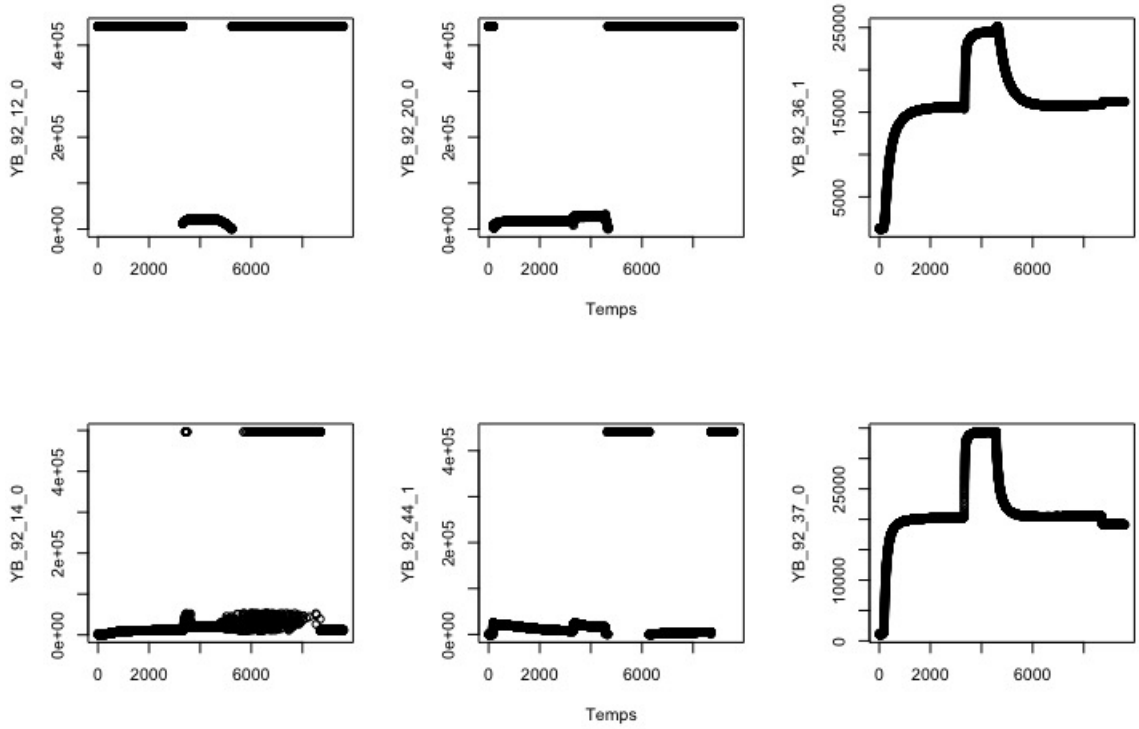

FIGURE 3.2 – Comportement des capteurs anormaux en comparaison avec des capteurs normaux
Source : SAFRAN / Observations : 9597 observations nommées, 1 variable par graphique

Pour faciliter la lecture, nous avons représenté à côté les graphes de capteurs normaux. On constate bien que leur allure est réellement différente des capteurs de la partie gauche. Pour faciliter la détection, nous allons maintenant séparer nos capteurs en groupes selon leur temps de réponse.

3.1.2 Distinction de groupes de capteurs selon leur temps de réponse

Pour pouvoir simplifier par la suite le travail d'appréciation des plages de stabilisation des capteurs, nous avons séparé nos capteurs en groupes selon leur temps de réponse (rapide, moyen, lent). Pour cela, nous avons réalisé un clustering de nos variables (des capteurs), puis divisé le dendrogramme en trois groupes.

Nous remarquons immédiatement sur la figure 3.3, à gauche ci-dessous, que cinq capteurs se détachent du lot (YB_92_12_0, YB_92_39_1,YB_92_14_0, YB_92_20_0, et YB_92_44_1). Quatre d'entre eux sont les mêmes que ceux détectés après l'étude de la variance.

Nous décidons donc d'ôter ces 5 capteurs de notre analyse et de relancer un clustering pour tenter d'obtenir des groupes plus homogènes et plus distincts. Le premier groupe était en effet ici constitué de 78 capteurs, et les deux autres se partageaient les 5 capteurs défectueux. De plus, la plupart des mesures avaient des variations brutes en fin de mesure, dues à l'extinction du moteur, au déplacement d'un capteur par action humaine etc. Pour être certains que ces variations ne biaisent pas le regroupement de nos capteurs, nous décidons, avec l'accord de nos tuteurs, de ne prendre les mesures que jusqu'à 8000 secondes. Dès lors, nous obtenons 3 groupes de capteurs plus équilibrés (figure 3.3 à droite) : le premier groupe constitué de 36 capteurs, le second de 30 capteurs et le dernier de 12 capteurs.

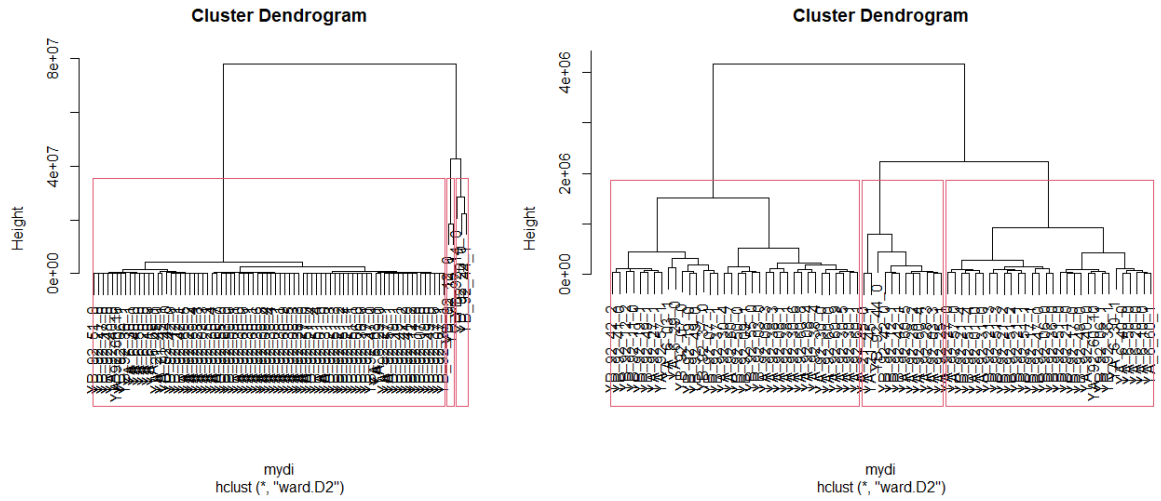

FIGURE 3.3 – A gauche : Dendrogramme post-clustering des variables à expliquer en régime transitoire
A droite : Dendogramme regroupant nos capteurs de même temps de réponse
Source : SAFRAN / Observations : 9597 observations

Nous proposons ci-dessous en figure 3.4, les graphes d'un capteur de chaque groupe à titre d'exemple pour visuellement déterminer les principales caractéristiques de ces groupes :

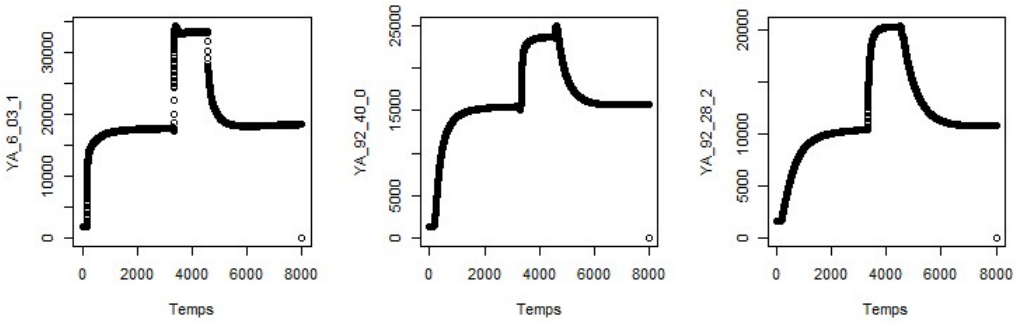

FIGURE 3.4 – Exemples de graphes de capteurs du groupe 1, 2 et 3 respectivement
Source : SAFRAN / Observations : 9597 observations

Le premier groupe semble être celui des temps de réponse rapides. Les passages d'une phase à l'autre sont très angulaires sur les graphes donc très brutaux et soudains. Les deuxième et troisième groupe affichent des formes de courbes assez similaires. Néanmoins, lorsque l'on se penche dessus, nous remarquons que la première phase du deuxième groupe (période de chauffe) se stabilise plus rapidement que celle du troisième groupe. De plus, si nous comparons les axes des ordonnées, les capteurs du deuxième groupe ont un saut (pour le décollage) autour de 15 000 et montent jusqu'à une valeur autour de 25 000. Les capteurs du troisième groupe ont, eux, leur saut autour de 10 000 et montent jusqu'à une valeur d'environ 20 000. Nous sommes donc amenés à penser que le groupe 1 est le groupe des réponses rapides, le groupe 2 des réponses moyennes, et le groupe 3 des réponses lentes.

Après observation de chacun des graphes des capteurs au sein de leur groupe respectif, il semblerait que le regroupement soit plutôt juste. Cette démarche permet d'établir, par la suite, de meilleures précisions quant

à la détection des capteurs anormaux.

3.2 Détection d'anomalies parmi les groupes de capteurs

Comme présenté dans la partie sur le contexte de l'étude, nous souhaitons ici détecter soit les mesures très bruitées, soit celles incorrectes ponctuellement, soit celles qui dérivent dans le temps. Pour cela, nous travaillons sur les variables en écart, mais aussi les instants en écart.

Pour cela, nous utilisons plusieurs méthodes que nous introduirons au fur et à mesure. Pour les variables en écart, nous étudions une fois de plus la variance des capteurs et le RMSE des capteurs expliqués par les autres, mais aussi l'écart entre les instants de réponse, tout cela par groupes de capteurs ainsi définis.

3.2.1 Détection des anomalies par analyse de la variance intra-groupes

Dans le but de détecter les capteurs aux comportements les plus anormaux, nous avons dans un premier temps étudié les variances des valeurs prises par les capteurs. Plus la variance de ces derniers est grande, plus la variabilité des mesures des instants est dispersée. À l'inverse, plus la variance est petite, moins les données sont dispersées.

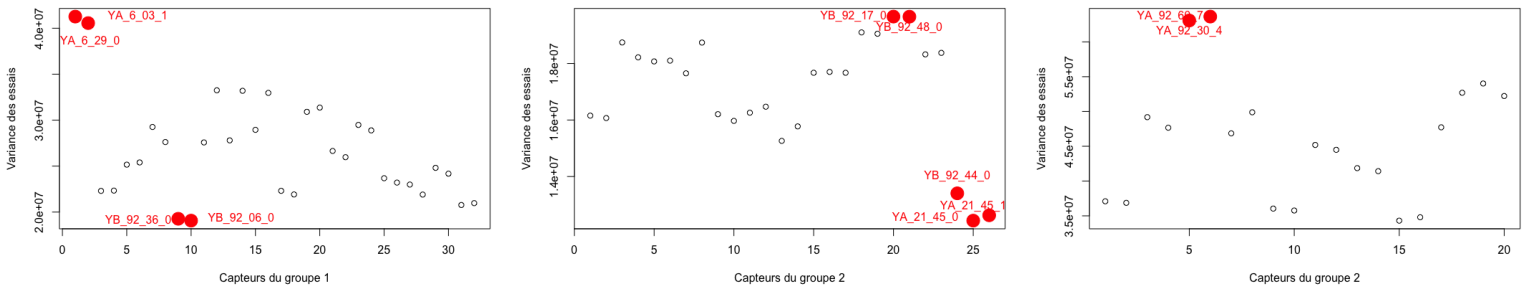

FIGURE 3.5 – Variances intra-groupes selon les capteurs du groupe 1 (à gauche), du groupe 2 (au milieu) et du groupe 3 (à droite)
Source : SAFRAN / Observations : 9597 observations

Parmi les capteurs du groupe 1, nous notons que deux d'entre eux ont des variances plus importantes que l'ensemble et deux moins importantes. Lorsque l'on s'intéresse à la distribution de ces capteurs respectivement, disponibles en annexes 4.13 page 36 et 4.14 page 37, nous remarquons que les instants des deux premiers sont plus étendus. En effet, les mesures atteignent l'échelle des 35 000 unités en ordonnée et réalisent des sauts importants, tandis que les deux derniers ont une distribution d'instants de sauts plus régulière, et s'apparente à des distributions normales. Le maximum des mesures est atteint aux alentours des 25 000 unités en ordonnée.

La méthode de l'analyse de la variance permet de relever cinq capteurs du groupe 2. En annexe 4.15 et 4.16 page 37 sont disponibles les graphiques des allures des capteurs. Les deux capteurs aux variances les plus importantes semblent avoir une allure plutôt classique par rapport aux capteurs du groupe. Des points atypiques sont notables, ceux-ci expliquant probablement l'importante variance du capteur. Enfin, au sein de ce groupe, les trois capteurs aux faibles variances ont des allures différentes de l'ensemble. En effet, les distributions sont beaucoup plus éparses, autrement dit les instants de sauts sont irréguliers. De plus, les maximums des mesures sont atteints aux alentours des 15 000 unités en ordonnée contre 25 000 pour les deux premiers capteurs identifiés.

Le groupe 3 met en avant deux capteurs dont la variance des mesures est importante. Lorsque l'on visualise leur distribution, disponibles en annexe 4.17 page 38, nous notons que les mesures atteignent les 40 000 unités en ordonnée et que les instants de sauts ne sont pas réguliers comme les capteurs normaux. Afin maintenant de détecter des capteurs différents de l'ensemble, nous essayons par la suite une seconde méthode.

3.2.2 RMSE

La méthode cette fois-ci consiste à expliquer un capteur selon le comportement des autres. Autrement dit, nous réalisons une régression linéaire d'un capteur en fonction des autres capteurs du même groupe. Une fois cela effectué, nous choisissons de calculer la RMSE pour chacune des régressions. De fait, nous obtenons une RMSE par capteur. Cette dernière correspond à la performance des capteurs utilisés dans la régression, à estimer le capteur en question. Une RMSE élevée comparée à l'ensemble des autres signifie que le capteur est mal estimé. Ceci informe que le capteur a un comportement différent des autres.

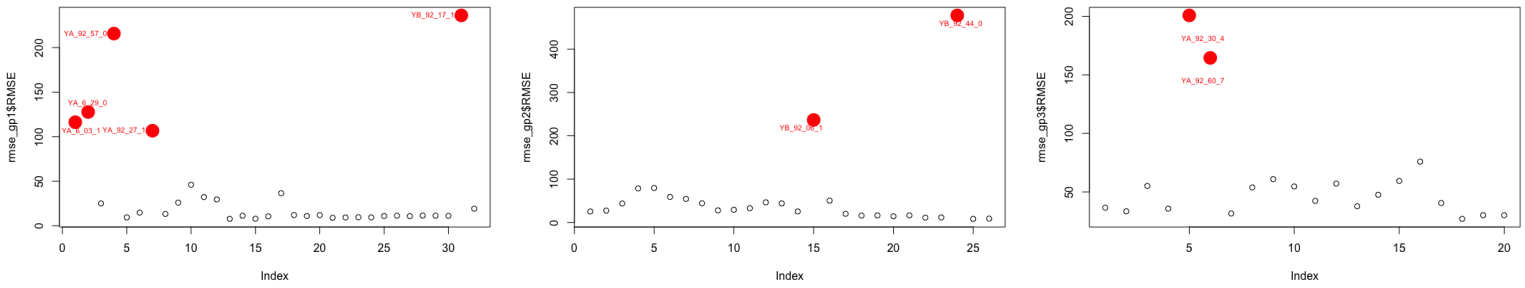

FIGURE 3.6 – RMSE selon les capteurs du groupe 1 (à gauche), du groupe 2 (au milieu) et du groupe 3 (à droite)
Source : SAFRAN / Observations : 9597 observations

Après avoir observé les distributions des comportements des capteurs identifiés, nous pouvons conclure que les instants de sauts sont très éparses et les maximums atteints très importants par rapport aux capteurs dits normaux. Nous passons maintenant à une dernière méthode pour tenter de détecter des capteurs en écart.

3.2.3 Étude des écarts entre les instants de réponse

Par les méthodes précédentes, nous avons retenus plusieurs capteurs atypiques. Par la méthode suivante, nous identifions des capteurs dont un ou plusieurs instants sont atypiques. Nous travaillons sur les écarts entre les instants. Nous nous intéressons aux différences entre les instants t et t-1. Une fois chaque différence calculée, nous procédons à la somme des écarts afin de pouvoir comparer les mesures des capteurs entre elles.

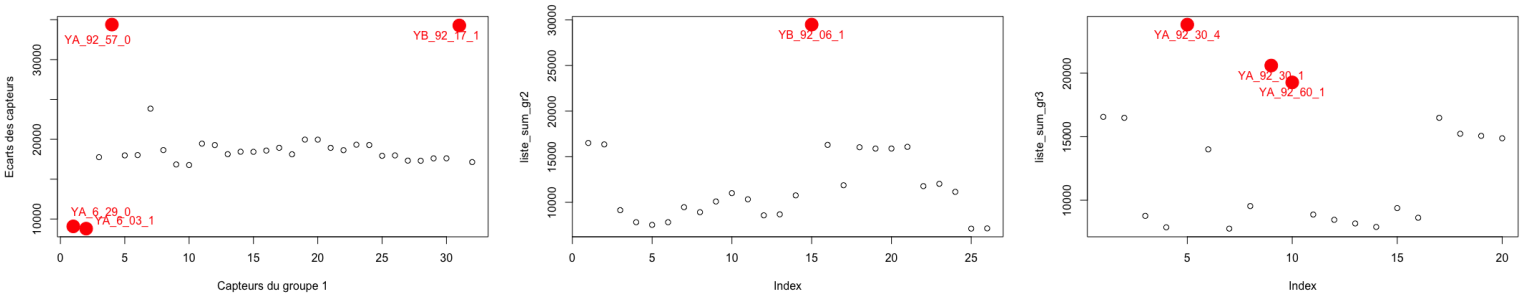

FIGURE 3.7 – Somme des écarts entre instants selon les capteurs du groupe 1 (à gauche), du groupe 2 (au milieu) et du groupe 3 (à droite)
Source : SAFRAN / Observations : 9597 observations

De même que pour les deux premières méthodes présentées ci-dessus, nous affirmons après observation sur les graphes que les capteurs identifiés sont correctement détectés. En effet, les instants de sauts sont très éparses et les maximums atteints très importants ou très faible par rapport aux capteurs de référence.

3.2.4 Détections des capteurs atypiques : conclusion

Suite aux trois méthodes de détections de capteurs anormaux présentées, nous proposons ci-dessous un tableau récapitulatif par groupes de capteurs et par méthodes. Ces capteurs s'ajoutent aux cinq retirés de l'étude précédemment.

	Variance	RMSE	Ecarts	Méthodes communes
Groupe1	YA_6_03_1 YA_6_29_0 YB_92_36_0 YB_92_06_0	YB_92_17_1 YA_92_57_0 YA_6_29_0 YA_92_27_1	YB_92_57_0 YA_6_29_0 YA_6_03_1 YB_92_17_1	YA_6_03_1 YA_6_29_0
Groupe 2	YB_92_A7_0 YB_92_48_0 YB_92_44_0 YB_21_45_1 YA_21_45_0	 YB_92_06_1 YB_92_44_0 	 YB_92_06_1 	 YB_92_44_0 YB_92_06_1
Groupe3	YA_92_60_7 YA_92_30_4	YA_92_30_4 YA_92_60_7	YA_92_30_4 YA_92_30_1 YA_92_60_1	YA_92_30_4 YA_92_60_7

TABLE 3.1 – Tableau récapitulatif des variables en écart par groupes de capteurs et par méthodes

En définitive, nous notons que trois capteurs (YA_6_03_1, YA_6_29_0, YA_92_30_4) ont été identifiés par les trois méthodes. Leurs distributions (disponibles en annexe 4.18, page 38) sont très similaires. Trois autres capteurs (YB_92_44_0, YB_92_06_1, YA_92_60_7) ont été identifiés par deux méthodes. Enfin, tout le reste des capteurs présentés dans le tableau 3.1 ont été identifié par l'une des trois méthodes.

Pour aller plus loin dans l'étude, nous avons pensé qu'il pourrait être intéressant d'essayer par la suite de détecter les instants atypiques des capteurs identifiés (sur leur zone de température stable). A l'instar du régime stabilisé, déterminer les observations (les moments i) en écart des autres. Nous notons cependant

que la méthode par Leave-One-Out ne pourrait pas être efficace puisque nous travaillons avec des séries temporelles. Les tests de Grubbs, Dixon, et Rosner ne peuvent pas non plus être appliqués car la distribution de nos données ne suit pas une loi normale. Nous allons donc essayer de déterminer ces phases de températures stabilisés puis pour certains capteurs essayer de détecter lors de ces phases si des observations sont en écart.

3.3 Appréciation de la stabilisation par groupes de capteurs

Nous cherchons à détecter des ruptures dans le signal. Pour cela, nous utilisons plusieurs algorithmes de détection de points de changement. D'abord, nous cherchons à détecter les quatre phases des capteurs par l'algorithme Binary Segmentation. Cela réalisé, l'algorithme PELT est utilisé sur chacune des phases. Ceci permet d'identifier les ruptures des signaux dans chacune de leurs phases. Si le signal est normal, il n'est pas censé avoir de points de changement.

3.3.1 L'algorithme de détection de changement de point

Binary Segmentation

Dans un premier temps, l'objectif est d'utiliser un algorithme de détection de points de changement afin d'identifier automatiquement les différents régime moteur. Pour cela nous allons utiliser un algorithme nommé "Binary Segmentation" (BS). Le principe est expliqué ci-après.
Dans un premier temps nous allons appliquer un test statistique de point de changement unique. Notons p la densité de probabilité de nos données et θ l'estimateur du maximum de vraisemblance de nos paramètres. Sous l'hypothèse alternative, nous considérons un modèle avec un point de changement à τ_1. Nous calculons par la suite deux nouveaux EMV sur chacun des intervalles. Selon l'article [8], le test statistique sont tels que :

$$ML(\tau_1) = \log p(y_{1:\tau_1|\theta_1}) - \log p(y_{(\tau_1+1):n|\theta_2})$$
$$\lambda = 2\left[\max_{\tau_1} ML(\tau_1) - \log p(y_{1:n}|\theta)\right]$$

Ce test nécessite la mise en place d'un seuil c tel que l'on rejette l'hypothèse nulle si $\lambda > c$. Cependant, la valeur de ce seuil reste toujours un sujet de recherche. Si un point de changement est détecté, le jeu de données est divisé en deux échantillons : un premier dont les bornes de l'intervalle sont 1 et τ_1 et un deuxième entre $\tau_1 + 1$ et n. Cette procédure est répétée sur chacun des deux échantillons, et ce programme est réitéré tant que le programme détecte des points de rupture sur les différents échantillons.

Ci-dessous, les figures 3.8 et 3.9 illustrent les performances de l'algorithme Binary Segmentation. Nous remarquons que les performances de détection de changements de phases varient selon les capteurs. De ce fait, nous analysons, à travers le graphe 3.10 le bon fonctionnement de l'algorithme selon les groupes.

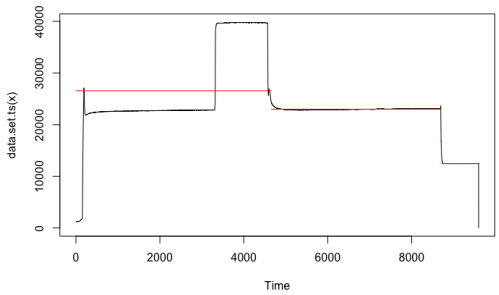

FIGURE 3.8 – Exemple d'un signal pour lequel BS n'a pas détecté les bonnes zones

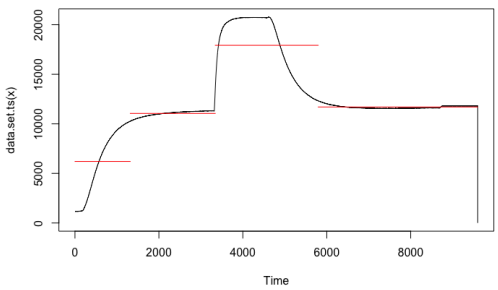

FIGURE 3.9 – Exemple d'un signal pour lequel BS a détecté les bonnes zones

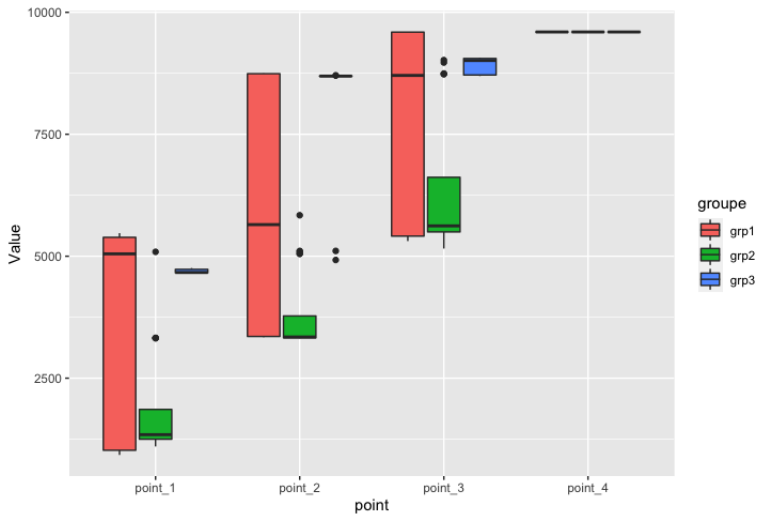

FIGURE 3.10 – Analyse du bon fonctionnement de pelt
Source : SAFRAN / Observations : 9597 observations

A travers la figure 3.10, nous constatons que les écarts types des points de ruptures identifiés (correspondant au début et à la fin de chaque phase) sont très différents selon les groupes. Pour le groupe 1, il est très grand, alors que pour les groupes 2 et 3 il est beaucoup plus petit. Cependant, pour le groupe 3 de capteurs, nous notons que les points 2 et 3 (correspondants au début et à la fin de l'accélération de l'avion) sont quasiment aux mêmes instants. Ainsi l'algorithme fonctionne bien uniquement sur le groupe de capteurs 2. De ce fait, nous choisissons de détecter les ruptures de signaux sur chacune des phases des capteurs du groupe 2. Ceci est proposé dans la partie suivante où nous analysons les différentes phases trouvées par l'algorithme sur le groupe 2.

Détection des zones de ruptures sur le groupe 2 : algorithme PELT

L'algorithme PELT (Pruned Exact Linear Time) fait partie des algorithmes permettant la détection de plusieurs points de changement. Tout comme l'algorithme des "Segment Neighbourhood", l'objectif est de minimiser la fonction :

$$\sum_{i=1}^{m+1} [C(y_{(\tau_{i-1}):\tau_i} + \beta] \tag{3.1}$$

avec m le nombre de segments après avoir identifié les points de changement, C la fonction de coût et le produit βm une pénalité afin d'éviter le sur-apprentissage.

Dans le cas des "Segment Neighbourhood", le nombre m du maximum de points de changement est pré-défini. Dans le cas de l'algorithme PELT, une recherche exhaustive est effectuée. Elle permet donc de préciser le nombre de segments finaux. Cependant, cette méthode peut être très coûteuse. Pour pallier ce problème, l'algorithme utilise une méthode de prunning. Ceci consiste à sélectionner uniquement des valeurs d'intérêt afin de réduire la complexité algorithmique. Dans notre cas, il s'agit d'enlever les valeurs de τ qui ne seront pas identifiées comme des changements dans le signal.

Notons $F(s)$, le minimum de l'équation (3.1) sur l'intervalle $y_{1:s}$. L'algorithme utilise le résultat du théorème suivant afin de déterminer les changements de points optimaux :

Theorem 3.3.1. *Nous supposons qu'en introduisant un point de changement dans une séquence d'observations, le coût, C, de la séquence diminue. Plus formellement, nous supposons qu'il existe une constante K telle que pour tout* $t < s < T$,

$$C(y(t+1):s) + C(y(s+1):T) + K \leq C(y(t+1):T)$$

si,

$$F(t) + C(y(t+1):s) + K \geq F(s)$$

alors, à un moment futur $T > s$, t ne peut jamais être le dernier point de changement optimal avant T .

C'est pourquoi l'algorithme PELT utilisera ce théorème dans son implémentation afin de sélectionner les points de rupture.

Algorithm 1 Algorithme PELT

Input

— Un dataset Y

— Une mesure du coût qui dépend des données

— Une valeur de la pénalité qui ne dépend pas du nombre de points de changement

— La constante K qui satisfait l'équation ci dessus

Initialisation : n = length(Y), F(0)=-$\beta, cp(0) = NULL[], R_1 = 0$

for $\tau^* = 1,\ldots,n$ **do**

1. Calculer $F(\tau^*) = \min_{\tau \in R_\tau^*}[F(\tau) + C(y_{(\tau+1):\tau^*} + \beta]$
2. $\tau^1 = \arg\left\{\min_{\tau \in R_{\tau^*}} \left[F(\tau) + \mathcal{C}\left(y_{(\tau+1):\tau^*}\right) + \beta\right]\right\}$
3. On pose $cp\left(\tau^*\right) = \left[cp\left(\tau^1\right), \tau^1\right]$
4. On pose $R_{\tau^*+1} = \left\{\tau \in R_{\tau^*} \cup \{\tau^*\} : F(\tau) + \mathcal{C}\left(y_{\tau+1:\tau^*}\right) + K \leq F\left(\tau^*\right)\right\}$

end

return Liste cp(n)

Maintenant que nous avons expliqué la méthode, nous allons l'implémenter sur chacune de nos phases détectées par la méthode "Binary Segmentation". Pour cela nous travaillerons sur nos données différenciées, car nous recherchons des ruptures dans la variance. Une fois que l'algorithme PELT a identifier les différents points de ruptures, on va stocker la variance sur chacun des nouveaux segments. Nous sommes partis du principe que plus la variance explose, plus la rupture est forte, et plus la probabilité d'avoir un capteur présentant une anomalie est importante. Afin de comparer la valeur des ruptures intra-groupe (caractérisée par la variance des segments trouvés), nous avons choisi de représenter la variance des variances estimées par segment pour chacun des capteurs.

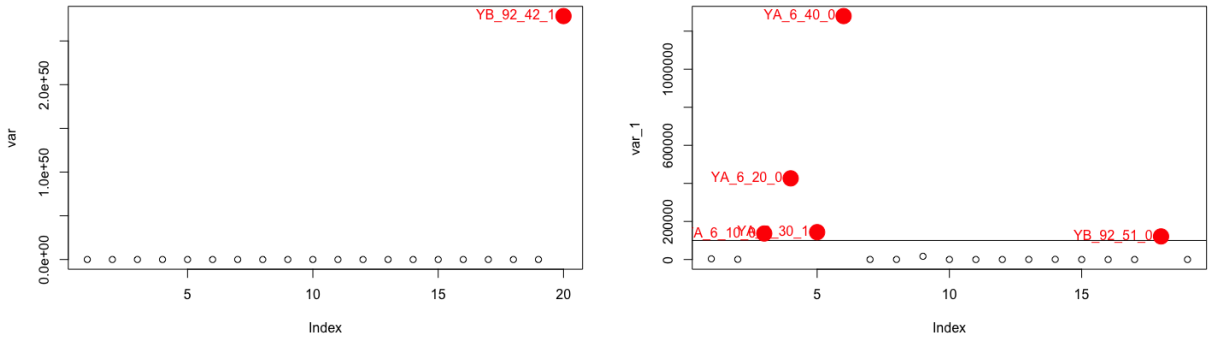

FIGURE 3.11 – Variance de l'indicateur selon les capteurs du groupe 2
Source : SAFRAN / Observations : 9597 observations

Nous constatons sur le graphe de gauche de la figure 3.11 que la variable *YB_92_42_1* a une valeur extrêmement grande comparée aux autres capteurs. Pour cela, nous reproduisons ce graphique sans la variable identifiée afin d'analyser les autres variances et s'intéresser aux potentielles anomalies d'autres capteurs. Sur le graphe 3.11, nous repérons cinq capteurs ayant une variance supérieure à 100 000 et supérieure à l'ensemble des autres capteurs du groupe 2. Nous nous intéressons aux anomalies de ces capteurs à travers les graphes de la figure 4.19, page 39 disponible en annexe. L'étude de ces graphes indique que les anomalies portent sur la fin du signal, qui correspond à l'arrêt du moteur. Nous pouvons donc approfondir l'analyse afin de voir si d'autres capteurs ont des anomalies antérieures à l'arrêt du moteur. L'arrêt moteur se trouvant autour de 8000 secondes, nous coupons le signal à cet instant.

3.3.2 L'algorithme PELT pour des mesures allant jusqu'à 8000 secondes de durée

Après avoir réitéré l'algorithme PELT sur les données coupées des instants supérieurs à 8000 secondes, nous constatons que beaucoup moins de signaux ont des comportements anormaux. Le graphique 3.12 illustre le fait que l'algorithme fonctionne beaucoup mieux. En effet, nous constatons sur la figure 3.12 que les boîtes à moustaches sont beaucoup moins étendues que sur la figure 3.10. L'algorithme semble très bien détecter les points de rupture pour les groupes 2 et 3 puisque les boîtes à moustaches sont excessivement rétrécies.

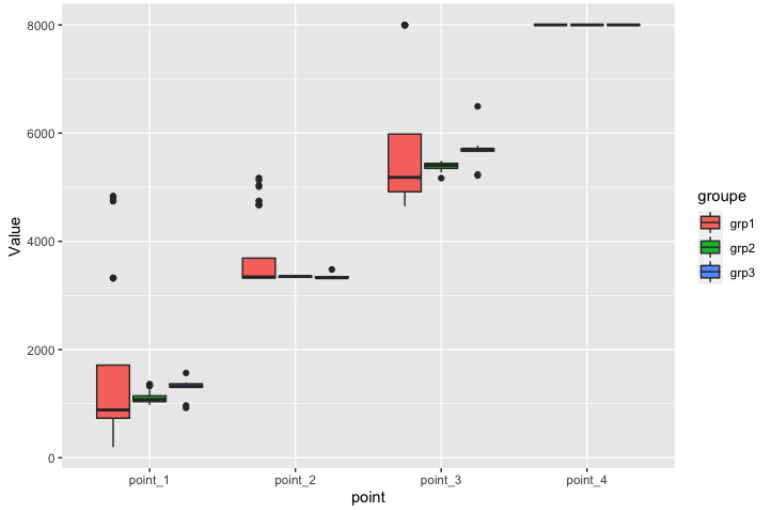

FIGURE 3.12 – Comparaison performance pelt
Source : SAFRAN / Observations : 9597 observations

3.3.3 Détection d'instants atypiques sur des phases de température stables

L'algorithme PELT a permis d'identifier les changements de régime. Suite à cela, nous avons donc une nouvelle fois essayé de mettre en place un modèle exploratoire : l'Isolation Forest. Celui-ci permettrait d'identifier sur les phases de températures stables seulement, si certains instants sont en écart des autres.

"L'isolation Forest calcule un score d'anomalie pour chaque observation du dataset. Ce score donne une mesure de la normalité de chaque observation en fonction de l'ensemble des données. Pour calculer ce score, l'algorithme isole la donnée en question de manière récursive : il choisit une variable au hasard et fixe un seuil de coupure au hasard, puis il évalue si cela permet d'isoler une observation en particulier."[2]

Nous avons sélectionné certains capteurs dont nous observions visuellement sur les graphes des observations en écart sur des phases de température stables. Le modèle une fois lancé, nous observons les scores d'anomalies. Nous représentons ensuite le graphique de chaque capteur en colorant les points détectés comme en écart. Nous présentons à travers la figure 3.13 deux exemples, deux autres sont disponibles en annexes 4.21 page

40. Nous définissons graphiquement un seuil d'anomalie en observant les graphes de scores disponibles en annexe 4.20, page 39. Si celui-ci n'est pas évident visuellement, il est usuel de prendre un seuil de 0,5.

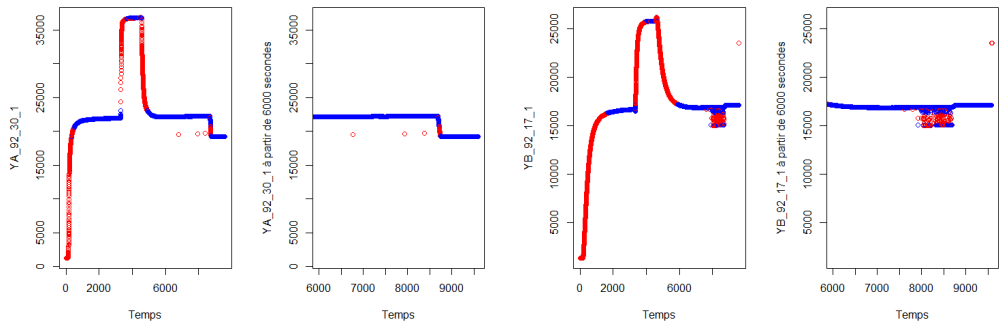

FIGURE 3.13 – Graphes des capteurs YA_92_30_1 et YB_92_17_1 avec un zoom sur les observations après 6000 secondes
Source : SAFRAN / Observations : 9597 observations / Note de lecture : Les points rouges sont les points détectés comme en écart par la méthode de l'Isolation Forest.

Nous ne prêtons ici attention qu'aux zones de températures stables (bien qu'on remarque d'ailleurs que l'Isolation Forest puisse potentiellement être une méthode de distinction des phases du régime moteur), d'où le zoom sur les observations supérieures à 6000 secondes sur les graphes de la figure 3.13. Les points en rouge sont les points détectés comme en écart par la méthode de l'Isolation Forest. Or visuellement, il semblerait que la méthode ait bien détecté de réels points en écart de la mesure de température stable. En plus de ces points singuliers, la méthode a également détecté la plupart des points de la fin de la prise de mesure de la température du capteur YB_92_17_1 où il semble y avoir eu un problème.

Nous avions également essayé une autre méthode exploratoire, à savoir celle du Local Outlier Factor que nous avions déjà essayé en régime stabilisé, mais les résultats n'étaient pas concluants.

Maintenant que nous avons détecté les observations et capteurs en écart ainsi que leurs différentes phases moteur, nous passons comme pour en régime stabilisé à la prédiction du comportement des moteurs.

3.4 Prédictions du comportement des moteurs : une mesure transitoire en fonction des autres

Rappelons que les capteurs mesurent une température ou une pression à une position donnée sur la turbine. Tout au long de cette étude, nous avons présenté des méthodes permettant de détecter si un capteur présente des anomalies. Ces anomalies peuvent être voulues, dûes à une accélération brutale du pilote ou alors à une position du capteur spéciale. Néanmoins, elles peuvent aussi être engendrées à cause d'un capteur endommagé ou perdu.
Le but ici est d'estimer le signal d'un capteur quand il est défectueux. Rappelons que nous travaillons uniquement sur une manoeuvre (un essai). Dans les faits, *SAFRAN* dispose de nombreux essais, avec les mêmes capteurs, aux mêmes positions, mais sous un régime moteur imposé par le pilote différent.
Dans ce contexte, nous allons donc simplement estimer un capteur à partir des autres selon plusieurs méthodes. Construire un modèle d'un capteur en fonction des autres avec l'essai que nous avons, permettra à *SAFRAN* de réutiliser ce modèle, avec les mêmes paramètres des modèles sur une autre manoeuvre si un capteur est défectueux.

Dans un premier temps nous avons utilisé une régression linéaire et dans un second nous avons construit un modèle pénalisé de type Lasso. Nous présentons uniquement la construction du modèle Lasso, mais nous

comparons ses performances avec celles de la régression linéaire.

3.4.1 Modèle Lasso

Même si le modèle de régression linéaire fonctionnait très bien, nous voulions paramétrer un modèle de régression pénalisée afin d'éviter le sur-apprentissage. En effet, les manoeuvres pratiquées par les pilotes peuvent être différentes. La pression et la température à l'intérieur de la turbine peuvent réagir de manière différente, ainsi les corrélations entre les capteurs peuvent être légèrement modifiées. Inclure à notre modèle une pénalité, réduira certainement la performance de notre modèle sur cet essai. Cependant, ceci permettra probablement de mieux estimer des capteurs défectueux sur des manoeuvres totalement différentes.

Comme dit précédemment, nous souhaitons effectuer un modèle Lasso et trouver la meilleure pénalité à imposer à notre modèle. Nous avons opté pour une régression Lasso et non Ridge afin d'utiliser les vertus parcimonieuses de cette régression. En effet, nos données demeurent très corrélées (même si au sein d'un même groupe on peut constater quelques variables moins corrélées, c'est ce qui nous a sans doute permis d'utiliser une régression linéaire). L'utilisation d'une régression pénalisée de type Lasso, nous permet de prendre en compte uniquement certaines variables : de sélectionner les variables les plus importantes.

Dans un premier temps, nous optimisons pour chaque régression pénalisée (une régression correspond à la modélisation d'un capteur en fonction des autres) la valeur du coefficient de pénalité. Rappelons que sous Rstudio, nous optimisons la fonction :

$$\min_{\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j z_{ij} \right)^2 + \lambda \left[\alpha \sum_{i=1}^{p} |\beta_p| + (1-\alpha) \sum_{i=1}^{p} \beta_i^2 \right]$$

Pour la régression Lasso, $\alpha = 1$. Ainsi, nous avons optimisé la valeur de λ par 10 fold cross validation. Ci dessous, nous visualisons l'évolution du **MSE** (Mean Square Error) en fonction de la valeur de λ.

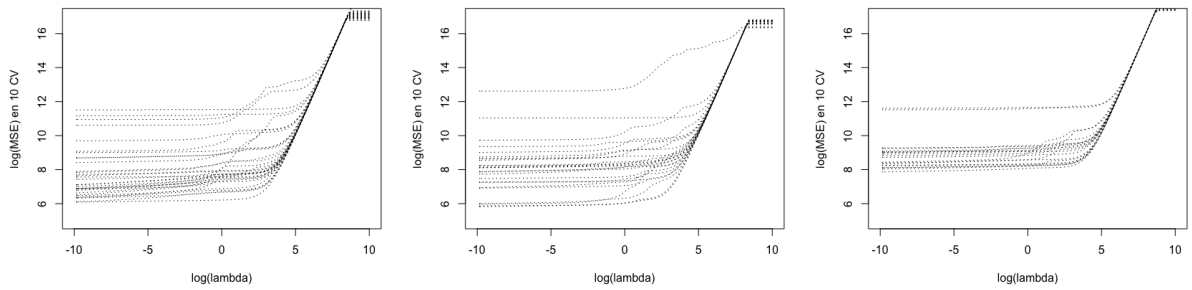

FIGURE 3.14 – Évolution de la MSE en fonction de la valeur de la pénalité selon les capteurs du groupe 1 (à gauche), du groupe 2 (au milieu) et du groupe 3 (à droite)
Source : SAFRAN / Observations : 9597 observations

Nous constatons que la valeur du **MSE** est stable pour des pénalités allant de 10^{-5} à 150. Lorsque nous analysons les coefficients pénalisés, nous constatons que pour chaque modélisation d'un capteur en fonction des autres, de nombreux capteurs sont pénalisés par le modèle.

3.4.2 Résultats

Sur la manoeuvre d'intérêt, nous comparons les performances de prédiction du modèle de régression linéaire et du modèle de régression pénalisée de type Lasso. Les calculs de **RMSE** ont été effectué sur l'échantillon de validation.

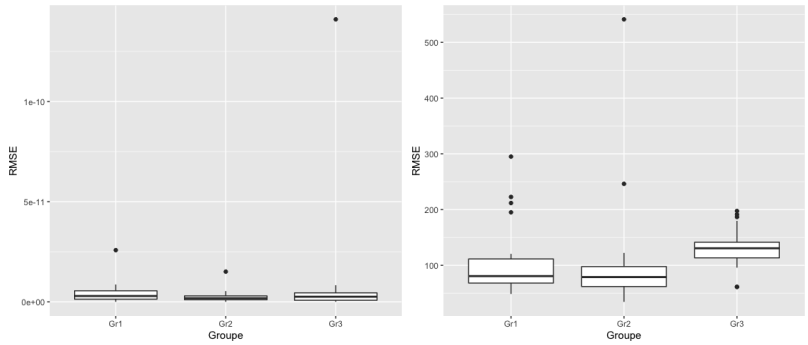

FIGURE 3.15 – Boxplot des valeurs prises par la RMSE pour chacune des modélisations des capteurs. À gauche pour la régression linéaire, à droite pour le modèle Lasso
Source : SAFRAN / Observations : 9597 observations

Suite à ces boîtes à moustaches, nous constatons que la valeur de la **RMSE** est extrèmement basse pour la régression linéaire (nous avons calculé un $R^2 = 1$). C'est ce résultat qui nous laisse penser que nous pourrions faire face à un problème de sur-apprentissage si l'on utilise d'autres manoeuvres avec la régression linéaire. Pour conclure nous avons calculé l'indicateur **MAPE** pour chacune de nos régression. Le tableau suivant synthétise les résultats du **MAPE** moyen par groupe et par modèle.

	Groupe 1	Groupe 2	Groupe 3
Régression linéaire	1.5 x 10^{-15}	9.1 x 10^{-16}	4.1 x 10^{-15}
Régression Lasso	5.2 x 10^{-3}	7.2 x 10^{-3}	5.8x 10^{-3}

TABLE 3.2 – Comparaison des régressions Linéaire et Lasso à partir du critère MAPE (en %)

Nous constatons que l'erreur effective de chacun des modèles est très faible. Ceci certifie que nos estimations sont justes. Cependant, notamment pour le modèle de régression linéaire, il ne faut pas omettre un potentiel phénomène de sur-apprentissage.

Conclusion

Ce projet avait pour ambition l'analyse de données d'essais moteurs dans deux régimes distincts : un premier régime dit "stationnaire" et un régime "transitoire" durant lequel notre moteur subit différentes phases, à la fois d'accélération et de décélération. L'enjeu était alors de mettre en place différents modèles de détection d'anomalies, puis d'utiliser ces données afin d'effectuer des prédictions sur le comportement du moteur.

La première difficulté résidait dans la forte dépendance entre nos variables. Afin de contourner cette dernière, nous avons été amené à utiliser de nouveaux outils statistiques, tel que la régression PLS (Partial Least Squares) qui représente une alternative efficace à la régression linéaire, qui ne peut être utilisée dans le cas de forte colinéarité comme c'est le cas dans ce projet. Nous avons basé notre analyse sur trois mesures distinctes nous ayant permis la détection d'un nombre important de capteurs supposés être "atypiques" : la variance, la RMSE ainsi que les écarts entre les instants.

Concernant l'analyse des données transitoires, un travail a été effectué sur la recherche de point de rupture, notamment grâce à l'utilisation de l'algorithme PELT, dont l'intérêt était de pouvoir isoler des phases de "stabilité" sur lesquelles un travail de détection de valeurs aberrantes a été effectué. Les différents résultats détaillés au cours de ce rapport ont été validés par nos tuteurs, confirmant la fiabilité des méthodes utilisées.

Un travail complémentaire peut être effectué afin de préciser nos analyses et poursuivre les recherches. Comme explicité précédemment, nous considérons les données transitoires comme des séries temporelles. Ainsi, des méthodes statistiques comme la décomposition du signal par ondelettes, l'utilisation de méthodes de clustering ou de modélisation spéciales séries temporelles, telles que dynamic time warping et shapelet, seraient pertinentes pour compléter ce projet.

Bibliographie

[1] Site internet officielle de l'entreprise SAFRAN. Page de présentation de la société. https://www.safran-aircraft-engines.com/fr/societe, visité le 19 novembre 2020.

[2] Algorithme N°3 - Comprendre l'isolation forest en 5 min 03/07/20 17 :52 par Équipe Data Science. https://blog.ysance.com/algorithme-n3-comprendre-lisolation-forest-en-5-min, visité le 19 novembre 2020.

[3] https://maths.cnam.fr/IMG/pdf/REGRESSION_PLS-1-2-Disc-Logist.pdf

[4] Diplôme Postgrade en Statistique - La régression PLS https://core.ac.uk/download/pdf/20641325.pdf

[5] http://irma.math.unistra.fr/~fbertran/PLS_17122010_Sabatier.pdf

[6] http://members.cbio.mines-paristech.fr/~thocking/change-tutorial/RK-CptWorkshop.html

[7] https://channel9.msdn.com/events/useR-international-R-User-conferences/useR-International-R-User-2017-Conference/Introduction-to-optimal-changepoint-detection-algorithms?term%3Dhocking

[8] Rebecca Killick, Idris A. Eckley, 2014, *Journal of Statistical Software : changepoint : An R Package for Changepoint Analysis*

[9] Killick, R., Fearnhead, P. and Eckley, I.A., 2012, *Optimal detection of changepoints with a linear computational cost.* https://arxiv.org/pdf/1101.1438.pdf

Chapitre 4

Annexes

4.1 Compléments à l'analyse descriptive des données

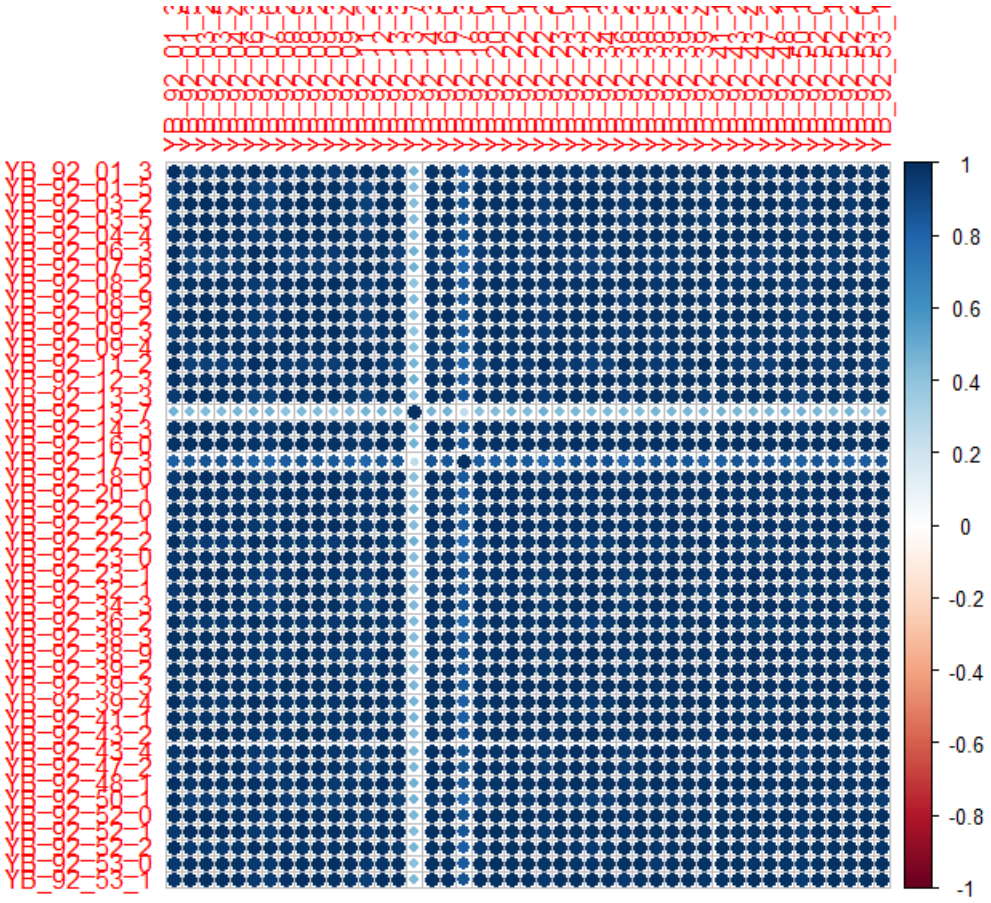

FIGURE 4.1 – Matrice de corrélations entre les variables YB_92
Source : SAFRAN / Observations : 107 observations nommées, 45 variables

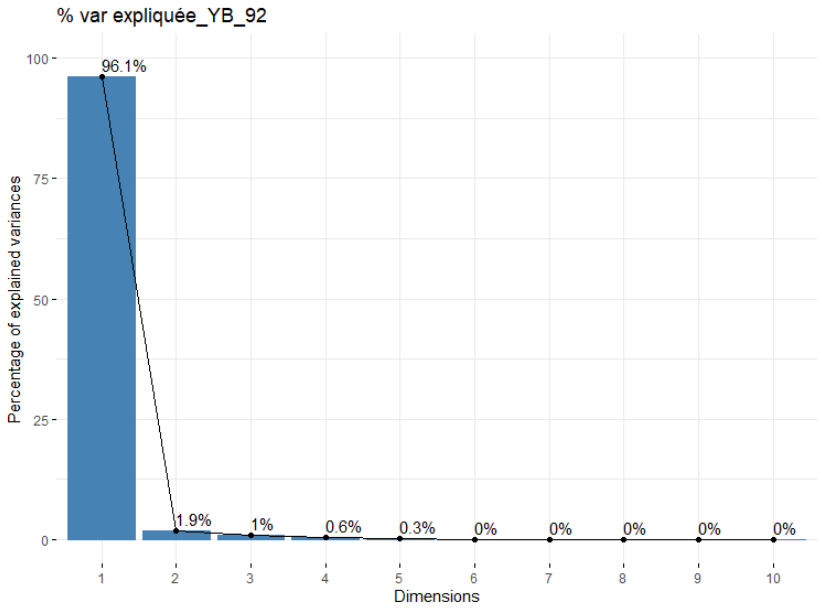

FIGURE 4.2 – Pourcentage de variance expliquée par chacun des axes de l'ACP sur les capteurs YB_92
Source : SAFRAN / Observations : 107 observations nommées, 45 variables

4.2 Complément à la détection d'anomalies en régime stabilisé : capteurs et observations

4.2.1 Quelques graphiques

Graphique d'un capteur défectueux et ses observations atypiques

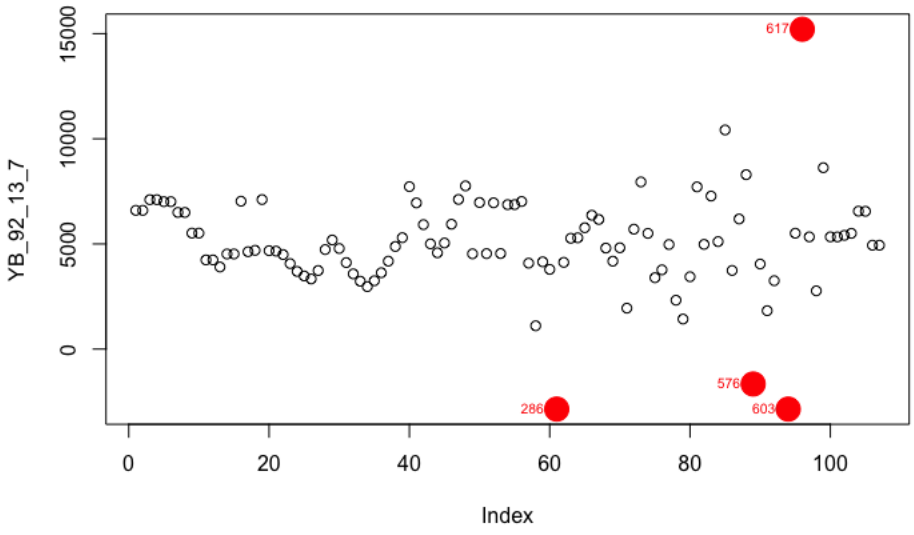

FIGURE 4.3 – Graphe d'un des capteurs détectés en écart en régime stabilisé
Source : SAFRAN / Observations : 107 observations, points rouges : instants atypiques

Corrélation des variables en régime stabilisé

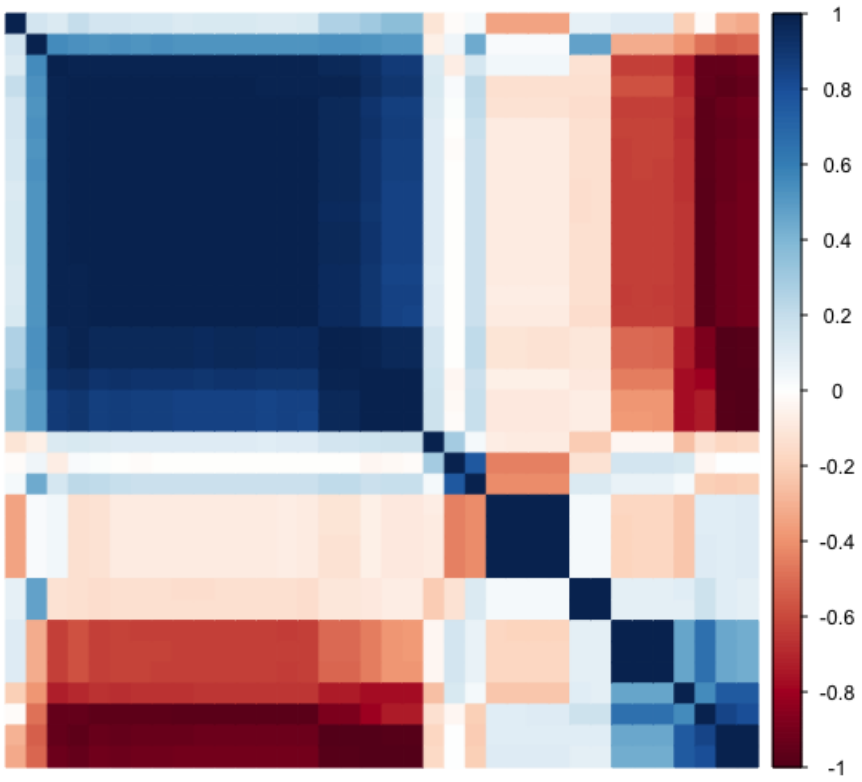

FIGURE 4.4 – Corrélation entre les variables explicatives
Source : SAFRAN / Observations : 107 observations nommées, 36 variables

4.2.2 Compléments sur l'étude des observations à l'écart en régime stabilisé

Détection des individus aberrants par Leave-One-Out selon le RMSE

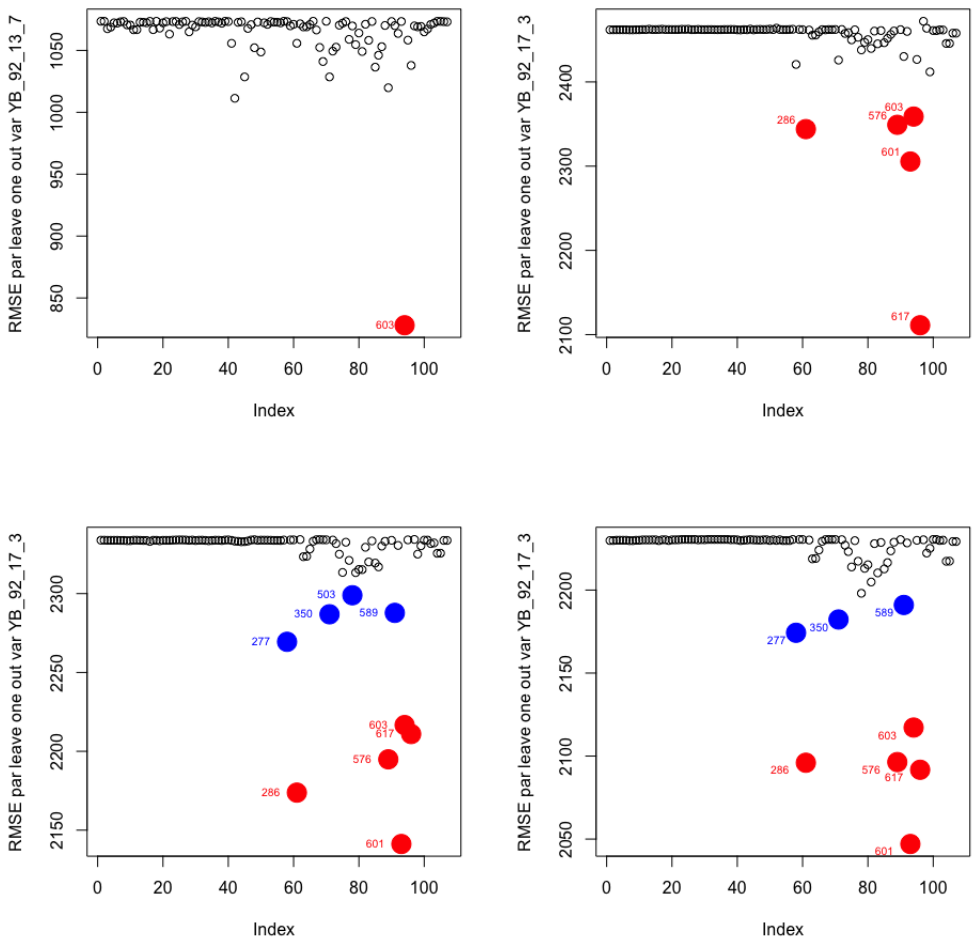

FIGURE 4.5 – Graphes de l'ensemble des RMSE calculés par LOO sur chacun des 4 capteurs détectés en écart

Source : SAFRAN / Note de lecture : en rouge on représente les 5 individus ayant la RMSE la plus élevée, en bleu les autres se détachant des valeurs standards.

4.2.3 Détection des individus aberrants par Leave-One-Out selon les variances

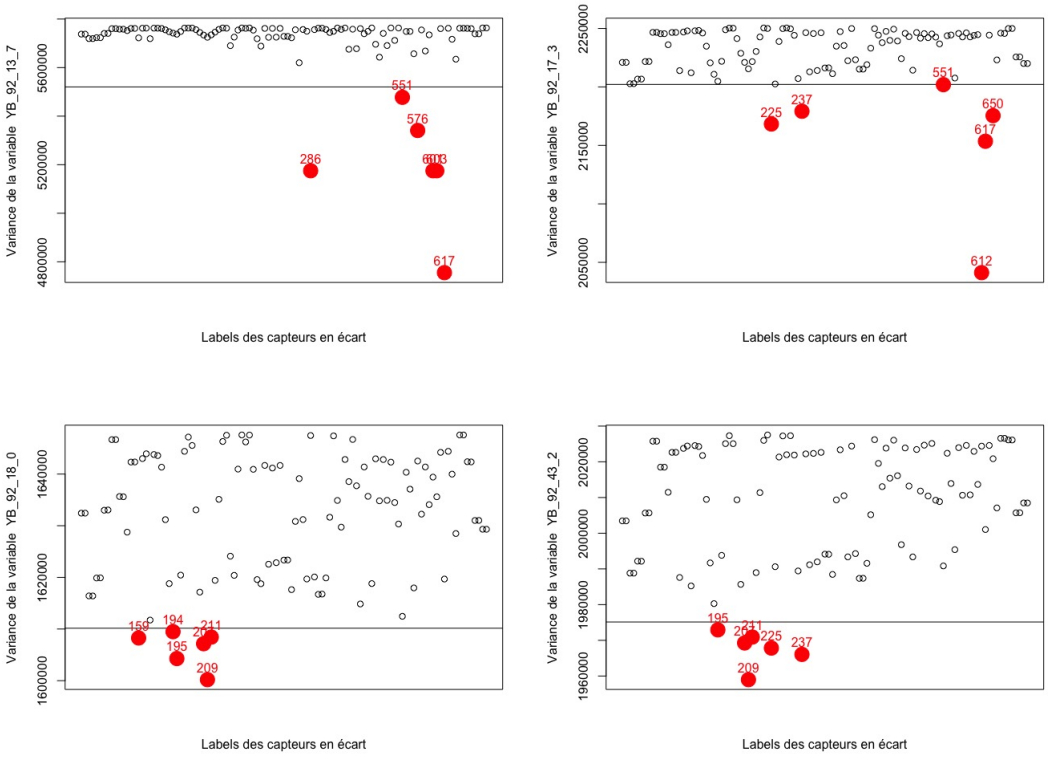

FIGURE 4.6 – Graphes de l'ensemble des variances calculées par LOO sur chacun des 4 capteurs détectés en écart
Source : SAFRAN

L'ACP

Après étude du pourcentage de variance expliquée par chacun des axes et discussion avec nos référents, nous décidons donc vis-à-vis de l'étude des observations en écart du reste de restreindre celle-ci à l'étude des axes 2 et 3 de l'ACP. En effet, nous pourrions ainsi détecter avec sans doute plus de précision les mesures en marge des autres, qu'en considérant l'axe 1 qui englobe toute l'information ou presque.

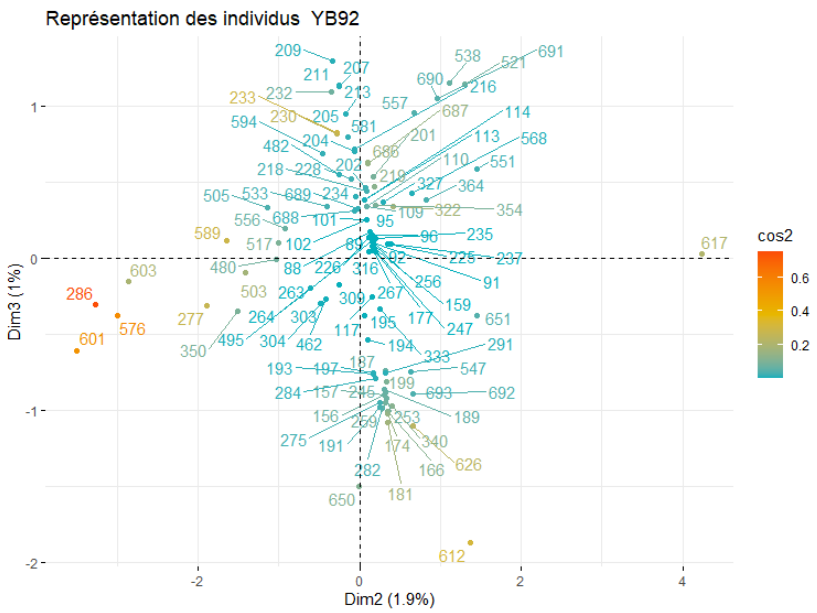

FIGURE 4.7 – ACP sur l'ensemble des capteurs YB_92 (graphe des individus, axes 2 et 3)
Source : SAFRAN / Observations : 107 observations nommées, 45 variables

On note tout de suite les quelques individus qui se détachent du lot, c'est-à-dire qui s'éloignent des clusters d'individus principaux visibles sur le graphe 4.7 de l'ACP sur l'ensemble des capteurs YB_92 entre les axes 2 et 3, à savoir les numéros 286, 576, 601, 603, 612 et 617. A priori, nos référents n'avaient pas retenu ces mesures comme anormales.
Nous avons ensuite décidé de mener des études supplémentaires sur la détection de mesures aberrantes. Pour cela, nous avons cherché à mettre en place deux méthodes exploratoires : celle du Local Outlier Factor et celle de l'Isolation Forest, expliquée ci-dessous.

Local Outlier Factor

L'algorithme LOF (Local Outlier Factor) est une méthode non supervisée de détection d'anomalies qui calcule l'écart de densité locale d'un point par rapport à ses voisins. Il considère comme aberrants les échantillons qui ont une densité sensiblement inférieure à celle de leurs voisins.

On regarde donc dans un premier temps la valeur pour chacune des observations de ce facteur :

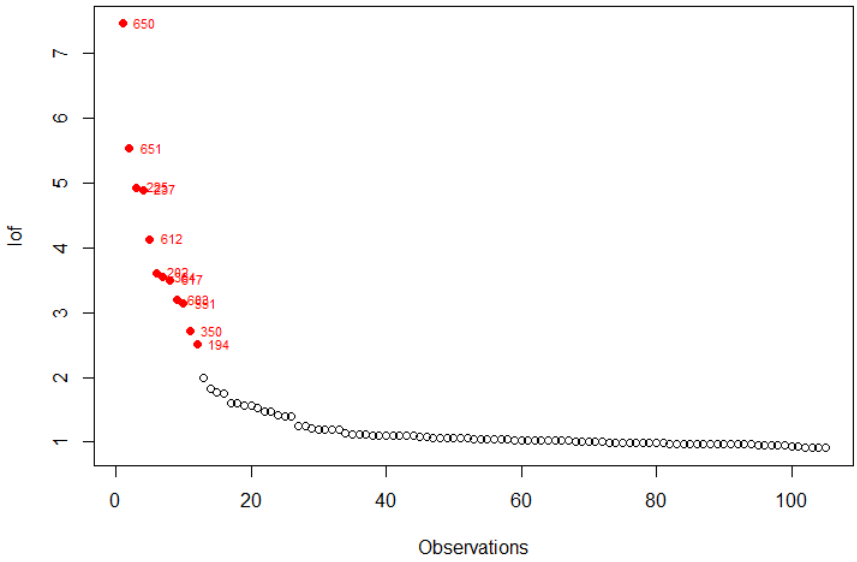

FIGURE 4.8 – Local Outlier Factor des observations sur les capteurs Yb_92
Source : SAFRAN / Observations : 107 observations nommées, 45 variables

On choisit de prendre pour seuil d'anormalité la valeur de 2 au regard du graphe 4.8. Les observations retenues sont donc les numéros 194, 202, 225, 237, 350, 364, 551, 603, 612, 617, 650 et 651.

On regarde ensuite sur l'ACP si ces observations semblent en effet se détacher des clusters principaux :

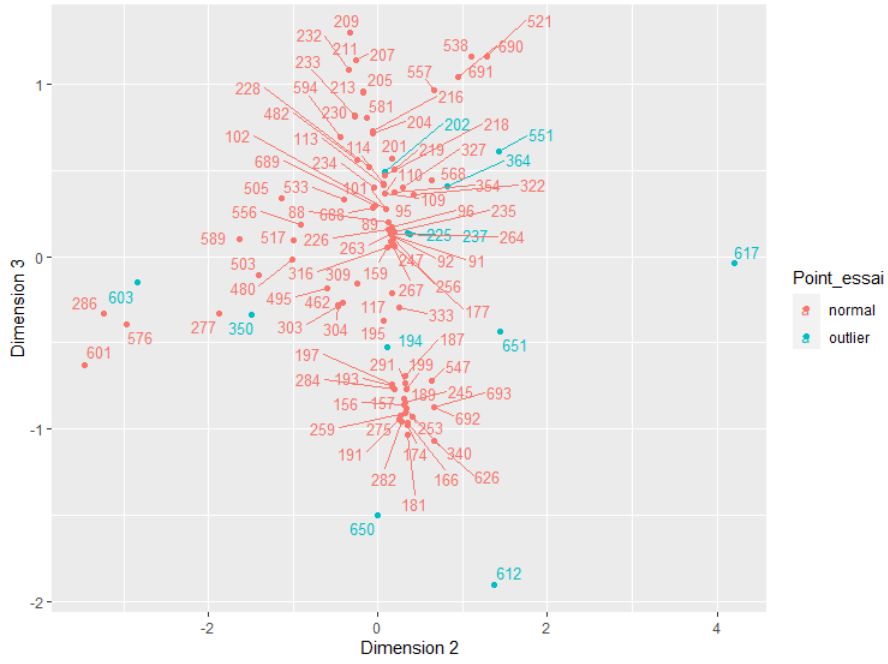

FIGURE 4.9 – ACP sur l'ensemble des capteurs YB_92 (graphe des individus, axes 2 et 3)
Source : SAFRAN / Observations : 107 observations nommées, 45 variables / Lecture : Les points en bleu sont les observations détectées comme en écart par la méthode LOF

La méthode LOF semble en effet repérer certaines observations en marge des autres. Après discussion avec nos référents, nous avons appris que les observations 650 et plus étaient naturellement différentes des autres car certains des capteurs au moment de la prise de mesure étaient hors-service. Le Local Outlier Factor a bien détecté des observations comme la 617, la 612 ou le 603 très écartées des clusters principaux, mais il ne semble pas avoir détecté les 286, 576 et 601 pourtant bien en marge du reste. Il a également détecté des mesures comme le 202, le 225 ou encore la 237 en plein coeur des clusters de notre représentation en deux dimensions ne contenant pas toute l'information.

Isolation Forest

"L'isolation Forest calcule un score d'anomalie pour chaque observation du dataset. Ce score donne une mesure de la normalité de chaque observation en fonction de l'ensemble des données. Pour calculer ce score, l'algorithme isole la donnée en question de manière récursive : il choisit une variable au hasard et fixe un seuil de coupure au hasard, puis il évalue si cela permet d'isoler une observation en particulier."[2]

On regarde donc dans un premier temps la valeur pour chacune des observations de ce score :

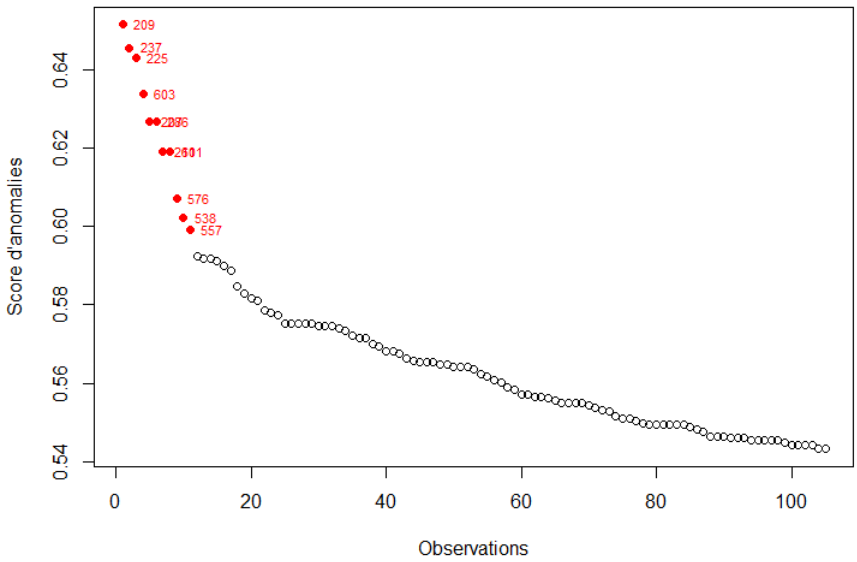

FIGURE 4.10 – Score des observations par la méthode de l'Isolation Forest sur les capteurs YB_92
Source : SAFRAN / Observations : 107 observations nommées, 45 variables

On choisit de prendre pour seuil d'anormalité la valeur de 0,599 au regard du graphe 4.10. Les observations retenues sont donc les numéros 207, 209, 211, 225, 237, 286, 538, 557, 576, 601, 603.

On regarde ensuite sur l'ACP si ces observations semblent en effet se détacher des clusters principaux :

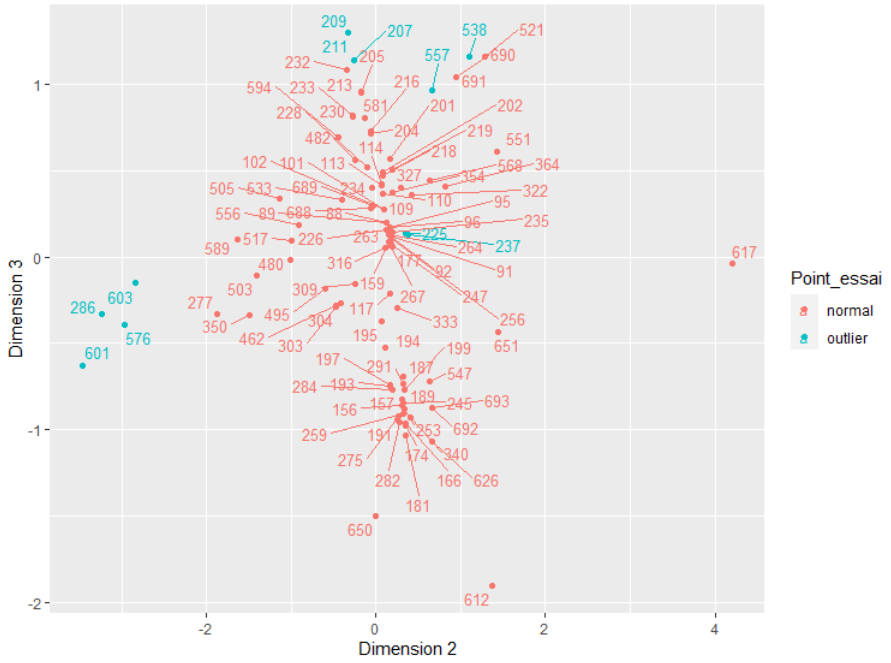

FIGURE 4.11 – ACP sur l'ensemble des capteurs YB_92 (graphe des individus, axes 2 et 3)
Source : SAFRAN / Observations : 107 observations nommées, 45 variables / Lecture : Les points en bleu sont les observations détectées comme en écart par la méthode IF

Contrairement à la méthode LOF, et hormis les observations 225 et 237, l'Isolation Forest a detecté des groupes de variables plutôt écartés des centres des clusters comme on peut le voir sur le graphe 4.11 (à gauche et en haut du graphe en bleu). Cette méthode n'a néanmoins pas détecté les mesures 612 et 617 très écartées des autres.

Comparaison des méthodes

Seules les observations 225, 237 et 603 sont détectées par chacune des deux méthodes LOF et Isolation Forest. Cependant, les mesures 225 et 237 sont plutôt au centre des clusters de notre représentation en deux dimensions ne contenant pas toute l'information. La mesure 603 est également détectée par la méthode Leave-One-Out. L'observation 617 est détectée par la méthode LOF et la méthode Leave-One-Out, et la 286 par la méthode de l'Isolation Forest et celle du Leave-One-Out.

On note donc bien qu'avec chacune des méthodes, certaines observations sont redondantes, ce qui est intéressant en terme d'efficacité des algorithmes. On retient cependant plutôt la méthode LOF car elle détecte plutôt des observations seules et en écart, ce que nous recherchons, alors que la méthode de l'Isolation Forest détecte plutôt des groupes d'observations en écart.

Néanmoins, après discussion avec nos tuteurs, nous décidons de laisser ces méthodes et leurs résultats de côté car l'anormalité d'un point d'essai est pour eux plus difficile à vérifier en interne qu'un capteur.

4.3 Complément à la prédiction du comportement des moteurs en régime stabilisé

4.3.1 La régression PLS

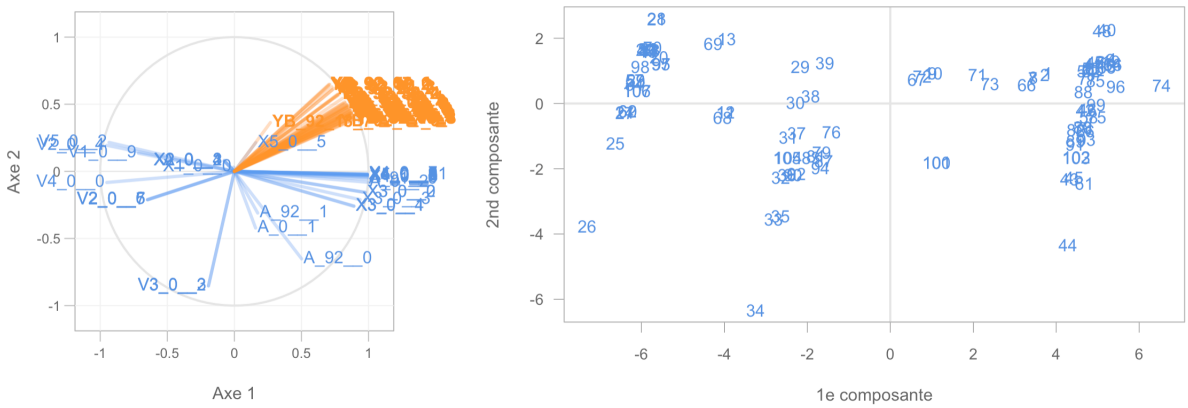

FIGURE 4.12 – Cercle des corrélations et carte des observations respectivement
Source : SAFRAN / Observations : 107 observations nommées, 36 variables

4.4 Compléments à la détection d'anomalies en régime transitoire : capteurs et observations

4.4.1 Quelques graphiques

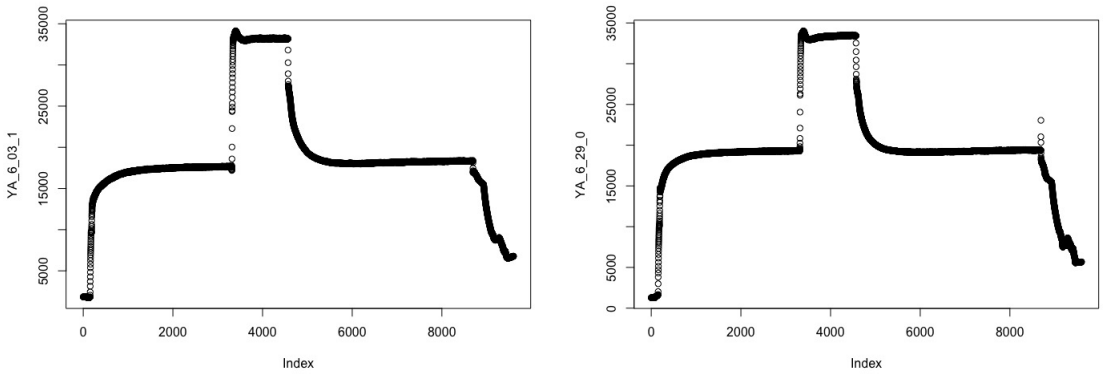

FIGURE 4.13 – Graphes de deux capteurs du groupe 1 détectés en écart après analyse de leurs variances
Source : SAFRAN / Observations : 9597 observations

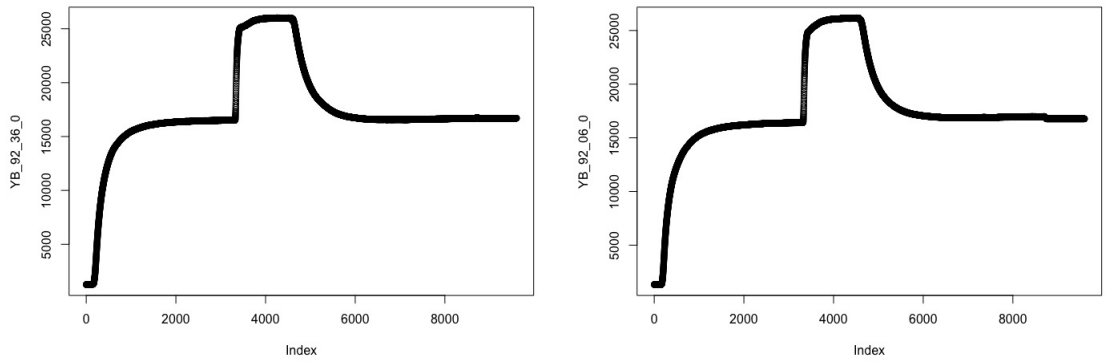

FIGURE 4.14 – Graphes des deux autres capteurs du groupe 1 détectés en écart après analyse de leurs variances
Source : SAFRAN / Observations : 9597 observations

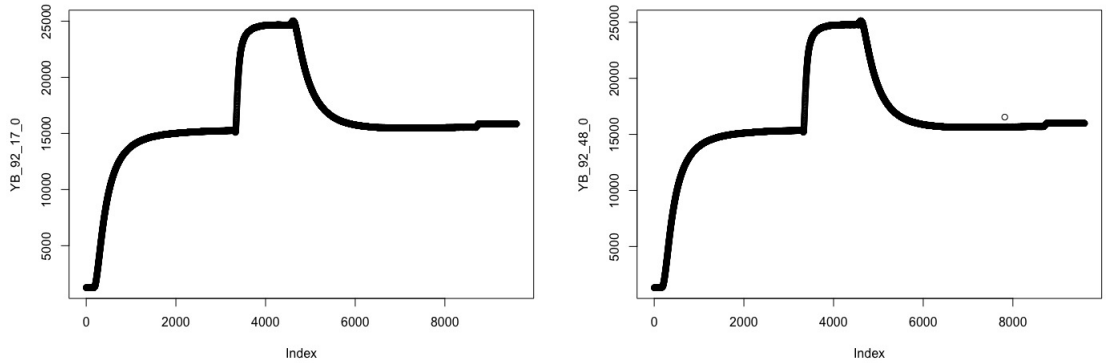

FIGURE 4.15 – Graphes de deux capteurs du groupe 2 détectés en écart après analyse de leurs variances
Source : SAFRAN / Observations : 9597 observations

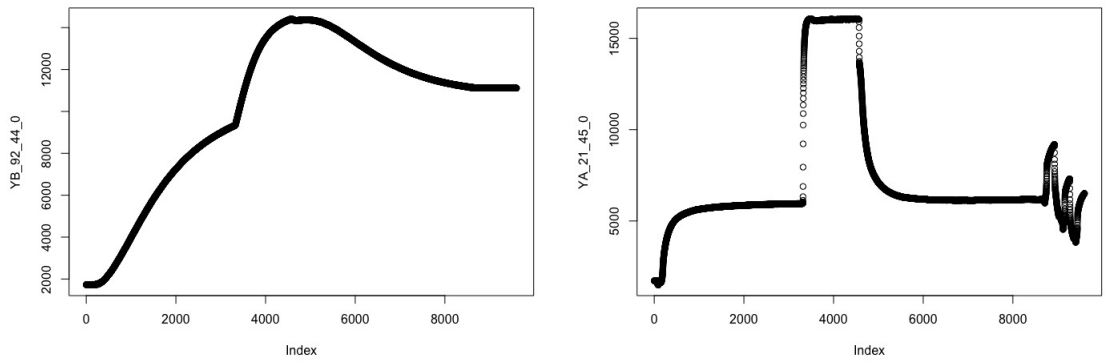

FIGURE 4.16 – Graphes de deux capteurs du groupe 2 détectés en écart après analyse de leurs variances
Source : SAFRAN / Observations : 9597 observations

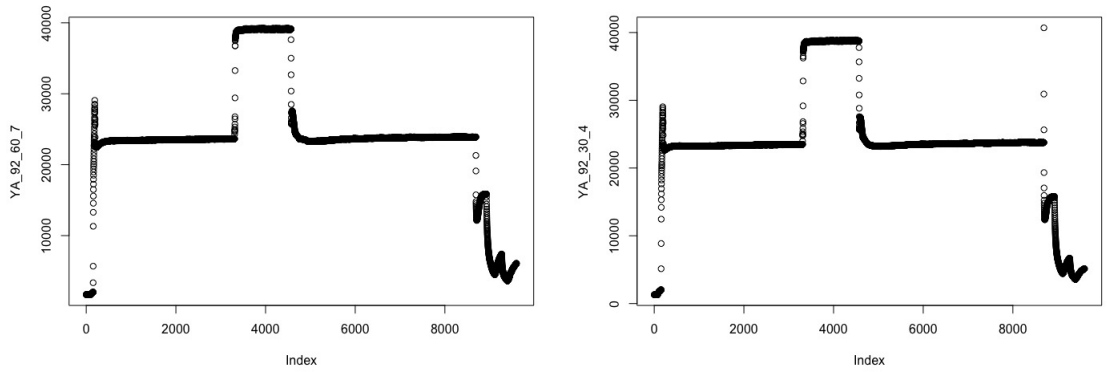

FIGURE 4.17 – Graphes des deux capteurs du groupe 3 détectés en écart après analyse de leurs variances
Source : SAFRAN / Observations : 9597 observations

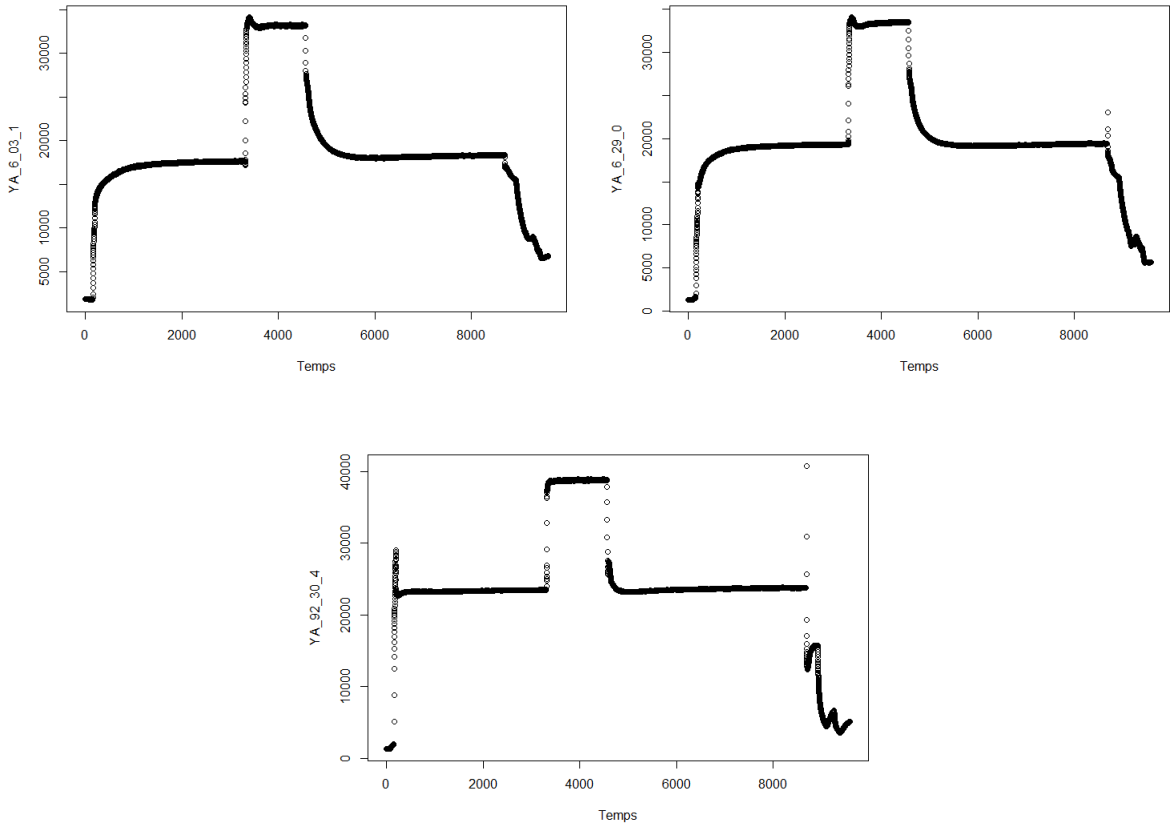

FIGURE 4.18 – Graphes des trois capteurs détectés en écart par chacune des méthodes
Source : SAFRAN / Observations : 9597 observations

4.4.2 Complément à l'étude des résultats de l'algorithme PELT

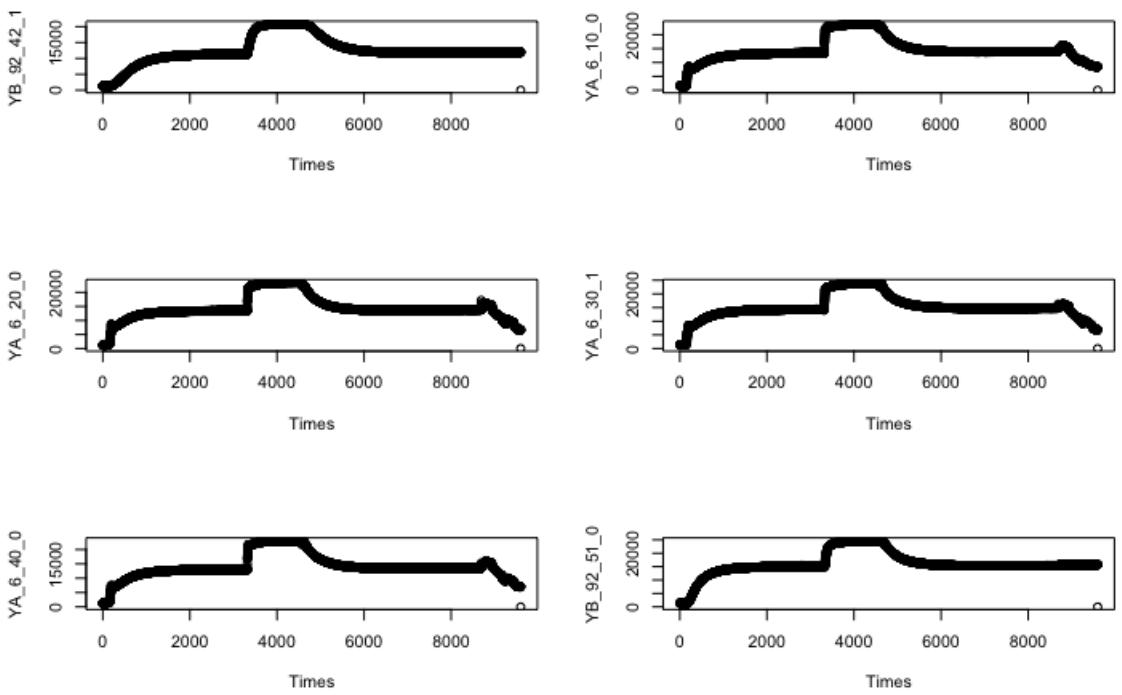

FIGURE 4.19 – Capteurs pouvant comporter des anomalies selon PELT
Source : SAFRAN / Observations : 9597 observations

4.4.3 Complément à la détection d'instants atypiques parmi les groupes de capteurs

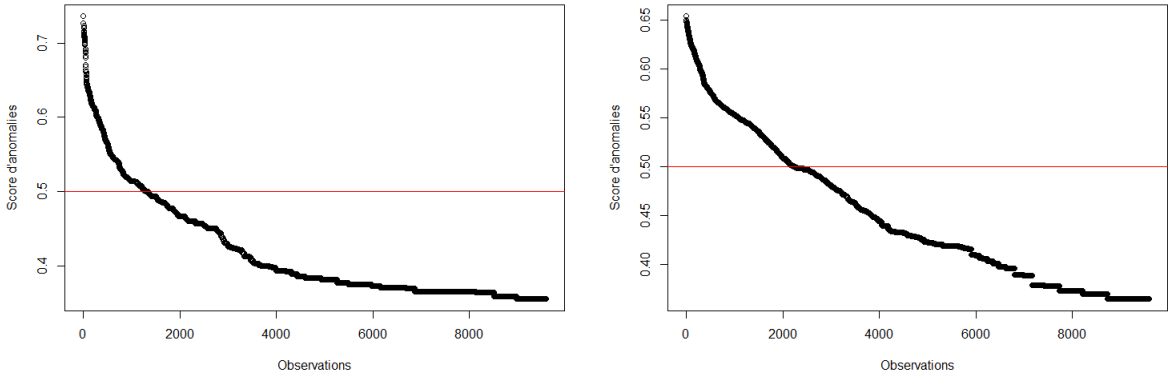

FIGURE 4.20 – Scores d'anomalies des capteurs YA_92_30_1 et YB_92_17_1
Source : SAFRAN / Observations : 9597 observations

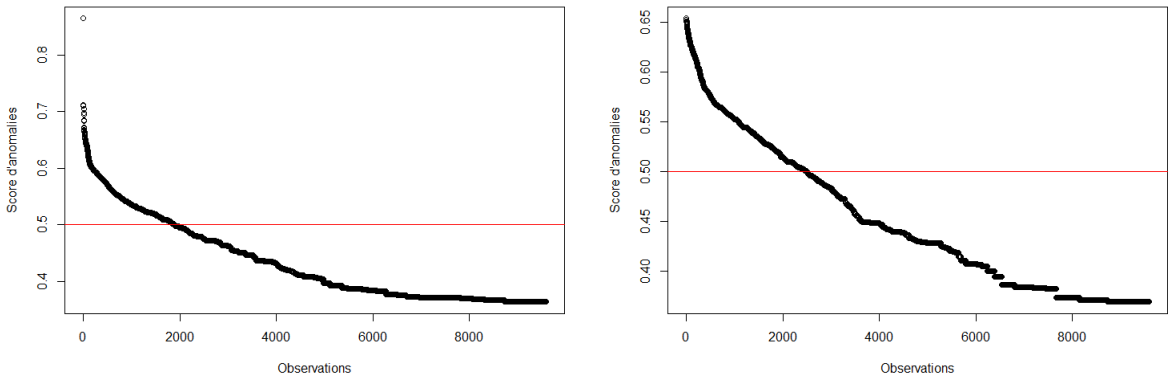

FIGURE 4.21 – Scores d'anomalies des capteurs YA_92_57_0 et YB_92_48_0
Source : SAFRAN / Observations : 9597 observations

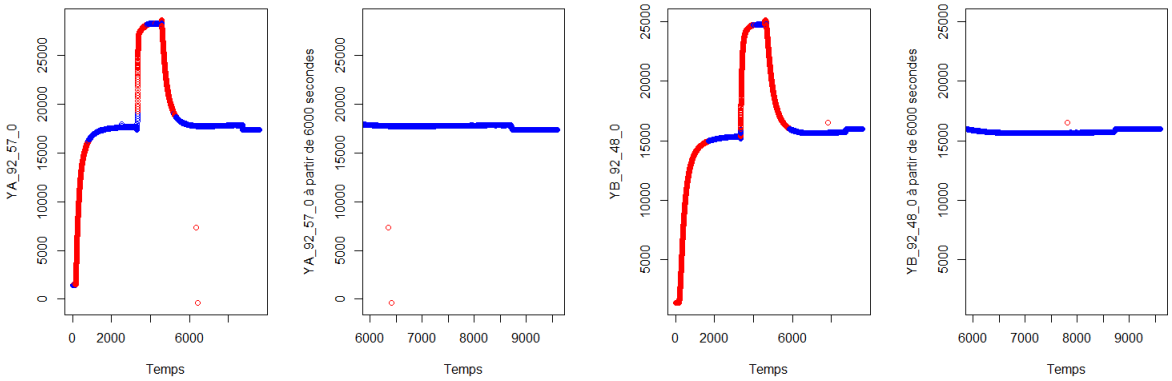

FIGURE 4.22 – Graphes des capteurs YA_92_57_0 et YB_92_48_0 avec un zoom sur les observations après 6000 secondes
Source : SAFRAN / Observations : 9597 observations / Les points rouges sont les points détectés comme en écart par la méthode de l'Isolation Forest.

4.5 Une mesure transitoire estimée selon la moyenne des autres

Nous considérons ici que nous avons aucune information sur les valeurs prises par le capteur défectueux. Nous connaissons uniquement les variables d'entrées, notamment sa position et son temps de réponse. Les experts de *SAFRAN* peuvent assimiler un capteur au bon groupe selon ces deux critères. Dans le cas de cette méthode, nous cherchons à estimer les instants d'un capteur non connu par les valeurs des instants des autres capteurs. Ces dernières mesures sont donc connues. Pour chaque instant du capteur à prédire, nous l'estimons par la moyenne de tous les autres capteurs à ce même instant.

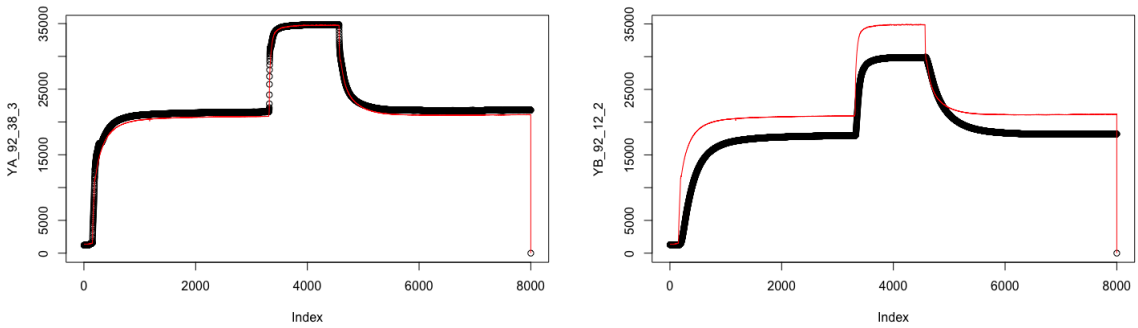

FIGURE 4.23 – Figure montrant la performance de prédiction de 2 capteurs du groupe 1
Source : SAFRAN / Observations : 9597 observations / En rouge : le signal prédit, en noir : le vrai signal

Cette méthode permet simplement d'avoir une approximation moyenne d'un signal d'un groupe. Les signaux qui sont proches du signal moyen de leur groupe sont bien estimés par cette méthode. Cependant, dès lors que les capteurs sont plus marginaux, la prédiction est moins juste. Ceci est illustré sur le graphique de droite de la figure 4.23 où le signal prédit est supérieur à la courbe du signal observé, en noir.