Projet de Mathématiques 3A

Druhet Joachim, Galice Mahaut

February 7, 2024

Contents

1 Remerciements

Nous souhaitons tout d'abord remercier les professeurs qui nous ont aider à mener à bien ce projet, Mme Vago et Mme. Bohnet, M. Jacquemard pour son aide et ses explications sur la transformée de Fourier ainsi que l'entreprise Bosch pour nous avoir fourni leurs données.

2 Introduction

La détection d'anomalies dans les données est une tâche cruciale dans de nombreux domaines, tels que la maintenance industrielle, la surveillance de la santé des systèmes, la détection de fraudes, etc. Dans le cadre de ce projet, nous nous sommes intéressés à la détection d'anomalies sur des radiateurs avec un set de données fournis par Bosh. L'objectif de cette démarche est de développer une méthode efficace pour identifier les comportements anormaux des radiateurs afin de faire de la maintenance prédictive par exemple. Nous allons maintenant nous intéresser au signal multidimensionnel, et pour ce faire, nous allons utiliser un modèle basé sur les vecteurs autorégressifs.

3 La Transformée de Fourrier

L'analyse de Fourier est une méthode mathématique fondamentale utilisée pour analyser la composition fréquentielle des signaux périodiques ou non périodiques. Elle est largement utilisée dans divers domaines, y compris les télécommunications, le traitement du signal, l'imagerie médicale, la physique et l'ingénierie.

3.1 Introduction à l'analyse de Fourier

3.1.1 Principes fondamentaux dans l'analyse de Fourier

Les principes utilisés pour effectuer une analyse de Fourier sont tout d'abord la transformation de Fourier qui est un outil mathématique qui décompose un signal temporel en ses composantes fréquentielles constitutives. Elle permet de passer du domaine temporel au domaine fréquentiel et est souvent utilisée pour analyser la structure harmonique des signaux. Il faut également prendre en compte le spectre de fréquencde du signal qui représente la distribution des fréquences qui composent ce signal. Il fournit des informations sur les différentes composantes fréquentielles présentes dans le signal, telles que les fréquences dominantes, les harmoniques et les bandes de fréquences.

3.1.2 Application de l'analyse de Fourier aux signaux temporels

Afin de réaliser l'analyse de Fourier, il faut tout d'abord décomposer notre signal temporel complexe en une somme de composantes sinusoïdales de fréquences et amplitudes variées.

Ensuite nous devons identifier les fréquences dominantes dans le spectre de fréquence du signal, l'analyse de Fourier nous permettant de détecter des motifs périodiques, pouvant être associés à des phénomènes eux aussi périodiques et donc des comportements réguliers dans les données.

L'analyse de Fourier nous offre donc une alternative facile pour comprendre la structure des signaux temporels et est, à ce titre, très largement utilisée dans la détection d'anomalies pour identifier des variations inhabituelles dans les données. Nous allons donc explorer dans la prochaine partie comment nous pouvons appliquer cette méthode à la détection d'anomalies dans les données de Bosch.

3.2 Méthodologie de détection d'anomalies basée sur l'analyse de Fourier

3.2.1 Prétraitement des données

Il est nécessaire d'avoir une plage de données adaptée pour effectuer l'analyse de fourrier. En effet une analyse sur un intervalle de temps trop petit peut être trop local et ne pas prendre en compte une certaine périodicité des données (Choisir que quelques heures ne prend pas en compte l'effet d'un cycle journalier). A l'inverse un intervalle de temps trop grand peut entraîner une perte de précision des données. Dans notre cas nous avons choisis une durée de 20 jours.

3.2.2 Calcul du spectre de fréquence et suppression du bruit

Pour ce faire on doit d'abord utiliser la transformation de Fourier discètre (TFD), qui est l'algorithme utilisé pour calculer le spectre de fréquence d'un signal discret. Elle transforme le signal du domaine temporel au domaine fréquentiel en calculant les amplitudes et les phases des différentes composantes fréquentielles présentes dans le signal.

On doit ensuite supprimer le bruit en enlevant les composantes de trop hautes fréquences, ainsi la transformée de Fourier inverse nous a permi de repasser dans le domaine temporel avec un signal lissé.

3.2.3 Identification des anomalies

Pour détecter les anomalies, nous avons calculé l'erreur quadratique entre le signal lissé après la transformée de Fourier et le signal réel. Si l'erreur quadratique en un point était trop importante alors ce point est considéré comme une anomalie.

3.3 Implémentation de l'analyse de Fourier pour la détection temporelle

L'implémentation de l'analyse de Fourier pour la détection d'anomalies implique plusieurs étapes pratiques, allant de la sélection des paramètres à l'intégration dans le flux de travail existant. Voici une décomposition détaillée de cette implémentation.

3.3.1 Sélection des paramètres

La taille des données utilisée pour calculer la transformée de Fourier discrète (TFD) peut avoir un impact sur la résolution fréquentielle et la sensibilité de la détection d'anomalies. Comme expliqué plus haut, nous avons choisis une durée de 20 jours pour faire nos tests. Cette durée à été trouvé de manière empirique, il s'agit de la durée nous permettant d'avoir une signal lissé nous rapprochant le plus possible du signal réel, tout en n'étant pas trop affecté par le bruit/anomalie.

Le choix d'un seuil de détection d'anomalies est crucial pour distinguer les composantes fréquentielles normales des anomalies. Ce seuil peut être déterminé empiriquement ou en utilisant des techniques statistiques pour définir des critères de détection d'anomalies robustes. Dans notre cas, nous l'avons fait de manière empirique en s'inspirant de certains informations trouvées sur internet pour nous donner un point de départ.

3.3.2 Développement de l'algoruthme de détection d'anomalies

Nous avons programmé la même méthode qu'expliquée précédemment en ajoutant la condition suivante pour la détection d'anomalies. En notant s le signal réel s_i la valeur d'un signal au point i et s' le signal lissé obtenu après la transformation de Fourier Inverse, on considère una anomalie au point i si : $(s_i - s'_i)^2 < seuil$

3.3.3 Affiner les résultats

Ayant observé des effets de bords dans notre analyse, afin de réduire le risque de faux positifs (détection d'anomalie à un point donnée alors qu'il y en a pas), nous avons effectué plusieurs fois l'analyse de Fourier et gardé que les anomalies qui sont détectés plusieurs fois. Par exemple si nous voulons détecter les anomalies entre le 10 et le 15 janvier, nous allons effectuer une analyse entre le 1er et le 20 janvier, et une entre le 5 et le 25 janvier. Si une anomalie est détecté au point i (i étant un moment entre le 10 et le 15 janvier) à la fois dans la première et dans la deuxième analyse, il est probable qu'il s'agisse bien d'une anomalie.

3.4 Données traitées

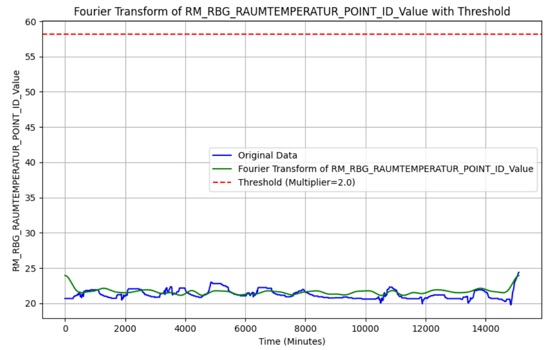

Figure 1: Résultats de notre transformée de Fourier

Le graphique obtenu nous montre une erreur dans la réalisation de notre transformée de Fourier et rend donc son interprétation impossible dans notre cas.

3.5 Conclusions avec Fourier

L'approche basée sur l'analyse de Fourier pour la détection d'anomalies présente certaines limitations et considérations qu'il est important de prendre en compte lors de son utilisation :

1. Sensibilité aux variations de fréquence : L'analyse de Fourier est sensible aux variations de fréquence dans les signaux. Les anomalies qui entraînent des variations significatives de fréquence peuvent ne pas être détectées efficacement par cette approche, car elles peuvent perturber le spectre de fréquence et masquer les anomalies.

2. Dépendance aux paramètres de prétraitement : Les résultats de l'analyse de Fourier peuvent être influencés par les choix de prétraitement des données, tels que la normalisation et le filtrage. Des paramètres de prétraitement inappropriés peuvent affecter la qualité des résultats et compromettre la capacité de l'approche à détecter efficacement les anomalies.

3. Sélection des seuils de détection : La sélection des seuils de détection d'anomalies peut être subjective et dépendante du contexte. Des seuils inappropriés peuvent entraîner des taux élevés de faux positifs ou de faux négatifs, ce qui compromet la fiabilité de la détection d'anomalies.

4. Analyse monodimensionnelle : Notre principal problème avec l'analyse de Fourier est le caractère monodimensionnelle, en effet nous avons seulement sélectionné le signal contenant la température de la pièce, nous n'avons pas pris en compte les autres signaux comme par exemple la température du sol, du plafond ou tout simplement la température demandée par le radiateur. Il est possible que nous détections une anomalie alors que la radiateur est simplement programmé pour s'allumer à une certaine heure. C'est pourquoi nous allons maintenant aborder la partie multidimensionnelle de notre projet.

4 Les Vecteurs auto régressifs

Pour atteindre notre objectif, nous avons choisi d'utiliser les vecteurs autorégressifs comme méthode de détection d'anomalies. Les vecteurs autorégressifs sont des outils puissants pour modéliser les dépendances temporelles dans les données multidimensionnelles. En utilisant des modèles probabilistes, ces vecteurs permettent de capturer les relations complexes entre les différentes dimensions des données et de prédire les valeurs futures en fonction des observations passées. Cette approche présente l'avantage de s'adapter aux variations temporelles et spatiales des données, ce qui en fait un choix prometteur pour la détection d'anomalies dans un contexte multidimensionnel comme celui des données de Bosch.

4.1 Explication générale des vecteurs autorégressifs

4.1.1 Définition et principes fondamentaux

Les vecteurs autorégressifs (VAR) constituent une classe de modèles statistiques largement utilisée pour modéliser les relations temporelles complexes entre les différentes dimensions des données. Ils sont particulièrement adaptés à l'analyse de séries temporelles multidimensionnelles, où plusieurs variables sont observées séquentiellement dans le temps. L'idée fondamentale des VAR est d'estimer un modèle qui représente les ≪lois≫ du système, le but final est de pouvoir prédire la valeur d'une variable en se basant sur ce modèle, nous verrons plus tard que l'application des VAR dans la détection d'anomalie se base sur la comparaison entre les valeurs réelles et les valeurs prédites par le modèle .
Formellement, un modèle VAR d'ordre p (VAR(p)) est défini par un ensemble de K équations de régression, une pour chaque variable dans le système. Mathématiquement, cela peut être représenté comme suit :
$Y_t = A_1Y_{t-1} + A_2Y_{t-2} + ... + A_pY_{t-p} + u_t$ Où :

- Y_t est un vecteur de taille K contenant les variables observées à l'instant t.
- $A_1, A_2, ..., A_p$ sont des matrices K*K de coefficients
- u_t est un terme d'erreur multivarié.

Afin d'estimer les coefficients A_i et u_t du modèle VAR, on utilise généralement la méthode des moindres carrés ordinaires (MCOà ou la méthode de maximisation de la vraisemblance.

L'utilisation de vecteurs autorégressifs pour modéliser les séries temporelles multidimensionnelles offre plusieurs avantages. Tout d'abord, ces modèles capturent efficacement les dépendances temporelles complexes entre les variables, ce qui leur permet de représenter fidèlement la dynamique du système. De plus, les VAR sont flexibles et peuvent être étendus pour inclure des variables exogènes, ce qui les rend adaptés à une grande variété de situations réelles.

Cependant, l'interprétation des résultats des modèles VAR peut parfois être complexe en raison de la présence de rétroactions entre les variables et de la nécessité de spécifier l'ordre du modèle de manière appropriée, ce n'est pas parce qu'il y a 13 signal différent (par exemple) qu'il faut les utiliser les 13 pour détecter l'anomalie, le modèle risque d'être beaucoup plus efficace en utilisant un nombre réduit de signaux.

4.1.2 Fonctionnement des vecteurs autoregressifs dans la detection d'anomalies

Afin d'appliquer un modèle VAR à nos données, nous devons poser l'hypothèse que les données normales suivent un schéma régulier et prévisible, tandis que les anomalies se manifestent par des déviations significatives par rapport à ce schéma. Comme dans la transformée de fourrier, on peut utiliser le calcul de l'erreur quadratique et l'utilisation d'un seuil pour évaluer la déviation entre le modèle et les valeurs réelles. L'utilisation d'un seuil assure une certaine concordance entre les différentes prédiction d'anomalies. Cependant, ce seuil doit être adaptées à chaque données en effet le choix du seuil peut avoir un impact significatif sur la performance de la méthode de détection d'anomalie, le seuil que nous avons utilisé ne sera pas forcément le plus adéquat dans un autre jeu de données, de plus le seuil est soit déterminé «à la main»(de manière empirique), ou bien en utilisant certaines techniques statistiques.

Plus spécifiquement, dans le cadre de la détection d'anomalies, les vecteurs autorégressifs se décomposent en trois grandes parties :
1. Modélisation des données normales : Le modèle VAR est entraîné sur un ensemble de données historiques considérées comme normales. Le modèle capture les tendances et les motifs réguliers des données au fil du temps.
2. Prédiction des valeurs futures : Une fois le modèle VAR estimé, il est utilisé pour prédire les valeurs futures des variables observées. Ces prédictions servent de référence pour évaluer la normalité des observations futures.
3. Détection des anomalies : Lorsqu'une nouvelle observation est faite, elle est comparée à la prédiction du modèle. Si l'écart entre l'observation réelle et la prédiction dépasse un seuil prédéfini, cela peut être interprété comme une anomalie.

4.1.3 Avantages et Inconvénients

Les vecteurs autorégressifs (VAR) présentent divers avantages et inconvénients dans le contexte de la détection d'anomalies, qui doivent être pris en compte lors de leur utilisation dans des applications réelles.

Les avantages principaux sont :
La possibilité de modélisation des dépendances temporelles complexes, en effet les VAR sont capables de capturer efficacement les dépendances temporelles complexes entre les différentes dimensions des données. En modélisant les relations entre les variables au fil du temps, ils peuvent représenter fidèlement la dynamique du système.
La flexibilité dans la modélisation : Les VAR offrent une grande flexibilité dans la modélisation des séries temporelles multidimensionnelles. Ils peuvent être étendus pour inclure des variables exogènes et des retards temporels supplémentaires, ce qui les rend adaptés à une grande variété de situations réelles.
Et la capacité de prédiction car une fois estimés, les modèles VAR peuvent être utilisés pour prédire les valeurs futures des variables observées. Cette capacité de prédiction peut être utilisée pour évaluer la normalité des observations futures et détecter les comportements anormaux.

Pour autant il existe des inconvénients majeurs à prendre en compte, parmi lesquels nous pouvons noter
Une forte sensibilité aux variations et aux bruits : Les modèles VAR peuvent être sensibles aux variations soudaines ou aux bruits dans les données. Des changements brusques ou des valeurs aberrantes peuvent affecter négativement la précision des prédictions du modèle, ce qui peut entraîner une augmentation des faux positifs dans la détection d'anomalies.
La spécification des paramètres du modèle : La performance des modèles VAR dépend fortement de la spécification appropriée de l'ordre du modèle et des paramètres associés. Le choix inadéquat de ces paramètres peut entraîner des prédictions imprécises et une mauvaise détection d'anomalies.
Et surtout la difficulté d'interprétation des résultats : L'interprétation des résultats des modèles VAR peut parfois être complexe en raison de la présence de rétroactions entre les variables. Comprendre la signification des coefficients du modèle et interpréter les résultats de manière appropriée peut nécessiter une expertise approfondie dans le domaine d'application.

En conclusion, bien que les vecteurs autorégressifs offrent de nombreux avantages pour la détection d'anomalies dans les séries temporelles multidimensionnelles, ils présentent également certains inconvénients qui doivent être pris en compte lors de leur utilisation, ainsi la plupart des résultats doivent être examinés (en tout cas au début) afin de s'assurer de la fiabilité du modèle.

4.2 Réalisation du projet

4.2.1 Implémentation des vecteurs autorégressifs pour la détection d'anomalies

Après avoir examiné les données de Bosch et mis en avant les différents signaux qui semblait les plus importants, nous avons intégré les modèles VAR à notre code afin de commencer la détection d'anomalies à plusieurs signaux. Cette phase du projet se décompose en plusieurs étapes :
1. Choix de l'ordre du modèle VAR : Avant de pouvoir estimer un modèle VAR, il est nécessaire de déterminer l'ordre du modèle, nous avons procédé de manière empirique pour cette partie.
2. Estimation des paramètres du modèle VAR : Une fois l'ordre du modèle déterminé, nous avons procédé à l'estimation des paramètres du modèle VAR à l'aide des données de Bosch. Ici aussi, nous avons procédé de manière empirique
3. Validation du modèle VAR : Après avoir estimé le modèle VAR, nous avons procédé à sa validation pour évaluer sa capacité à capturer correctement les dépendances temporelles dans les données. Afin de valider notre modèle nous avons notamment pris une plage de données de test où nous pensions qu'il n'y avait aucune anomalie, et vérifié que l'erreur du modèle était minime. Cette étape était importante pour garantir que notre modèle était robuste et fiable dans des situations réelles.
4. Détection d'anomalies : Maintenant que nous avons un modèle qui semble cohérent, il ne reste plus qu'à calculer l'erreur quadratique entre la prédiction du modèle et la valeur réelle pour chaque valeur dans un intervalle de temps précisé en entrée(nous avons fait nos tests avec une durée de 20 jours). Si cette erreur est supérieur au seuil fixé, alors nous considérons cette valeur comme une anomalie.

4.2.2 Analyse des résultats obtenus

Nous avons essayé d'examiner les exemples d'anomalies détectées par notre modèle pour comprendre leur nature et leur origine. Ainsi nous pouvons mieux identifier les signaux qui sont importants pour l'analyse et ceux qui le sont moins. Une manière de perfectionner le modèle est de modifier la base de notre modèle en choisissant seulement les signaux qui nous semblent le plus adapté après analyse des résultats, puis de répéter ce processus, jusqu'à résultat concluant.

Nous avons interprété les anomalies détectées en fonction du contexte des données. Nous nous sommes notamment interrogés sur les causes physiques de certaines anomalies et sur les faux positifs. En effet certaines anomalies ne le sont pas et cela peut être du à un changement brutal dans l'environnement du capteur ou encore au lieu où est situé le capteur. Ces faux positifs pourraient être traité avec un filtre post analyse par exemple.

4.3 Données traitées

Cette méthode permettant de prendre plusieurs variables en paramètres est donc plus intéressantes car nous permettant de comprendre si une anomalie en est vraiment une ou si elle est liée à d'autres paramètres. Cependant dans le but d'améliorer cette méthode nous pourrions adapter les seuils de détection plutôt que de les fixer arbitrairement. En effet, les adapter dynamiquement, à l'aide d'un apprentissage automatique par exemple, permettraient de diminuer le risque de détection de fausses anomalies.
Elle présente également une autre limite, car ne supposant que le fait qu'il s'agisse d'un signal stationnaire et donc suppose que le signal ne peut trop varier dans le temps

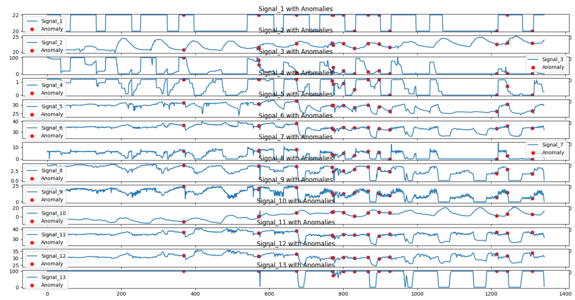

Figure 2: Tous les signaux traités avec mise en évidence des anomalies

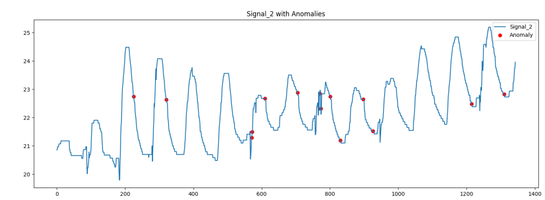

Figure 3: Température de la pièce avec les anomalies

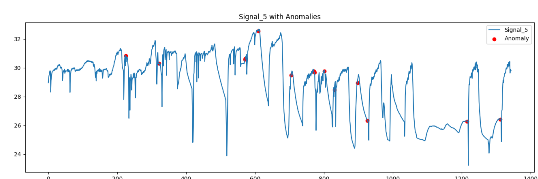

Figure 4: Température envoyée par le sol avec les anomalies

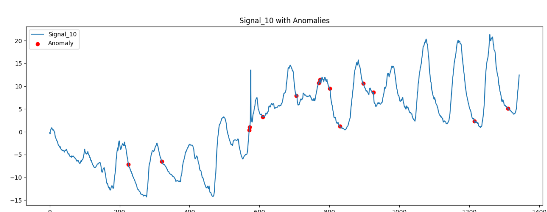

Figure 5: Température extérieure

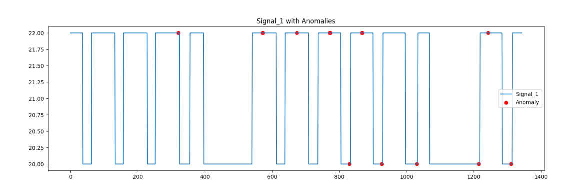

Figure 6: Température demandée dans la pièce

5 Conclusion

En conclusion, l'analyse de Fourier et les vecteurs auto régressifs sont deux méthodes efficaces pour la détection d'anomalies dans les signaux temporels. Chacune de ces approches présente des avantages et des limitations qui doivent être pris en compte lors du choix de la méthode utilisée pour détecter les anomalies.

L'analyse de Fourier est une méthode classique et largement utilisée pour analyser la composition fréquentielle des signaux. Elle est particulièrement utile lorsque les anomalies se manifestent sous forme de variations de fréquence dans les données. L'analyse de Fourier est également relativement simple à mettre en œuvre et rapide à calculer, ce qui en fait un choix attrayant pour les applications où la détection d'anomalies basée sur les caractéristiques fréquentielles est pertinente. Néanmoins elle n'est utile que pour les signaux unidimensionnel.

D'autre part, les vecteurs auto régressifs sont des modèles plus sophistiqués qui capturent les dépendances temporelles complexes entre les différents signaux. Ils sont capables de modéliser les relations dynamiques entre les différentes dimensions des données, ce qui les rend appropriés pour détecter des anomalies dans des situations multidimensionnelles. Les vecteurs auto régressifs peuvent également être adaptés pour prendre en compte la non-stationnarité des données et peuvent être plus flexibles dans la modélisation de structures de données complexes.

En général, l'analyse de Fourier peut être privilégiée lorsque les signaux sont dépendant d'une période (comme par exemple un cycle journalier ou hebdomadaire dans notre cas) et lorsque la modélisation des dépendances temporelles n'est pas nécessaire. D'autre part, les vecteurs auto régressifs sont plus appropriés lorsque les données présentent des tendances temporelles complexes ou des dépendances multidimensionnelles, et lorsque des modèles plus sophistiqués sont nécessaires pour capturer ces relations.