Stage en data science
Rapport sur la détection des données manquantes

Ilyasse ALIOUI - 2SN

Département Sciences du Numérique - Deuxième année
2023-2024

Table des matières

Table des figures

1 Introduction

L'analyse des données est une étape incontournable en Data Science, car elle permet d'extraire des informations utiles, de détecter des tendances et de prendre des décisions éclairées. La qualité des données joue un rôle fondamental dans la fiabilité des résultats obtenus, car des données incorrectes ou incohérentes peuvent fausser l'ensemble du processus d'analyse et conduire à des conclusions erronées. Cependant, les données réelles collectées sur le terrain sont rarement parfaites et incluent souvent des valeurs manquantes. Ces valeurs manquantes peuvent résulter de diverses causes, telles que des erreurs humaines lors de la saisie, des défaillances dans les capteurs, des interruptions de la collecte ou simplement des informations indisponibles.

Si ces données manquantes ne sont pas correctement identifiées et traitées, elles peuvent avoir un impact négatif sur la qualité des modèles prédictifs et des analyses statistiques. C'est pourquoi la détection et la gestion des données manquantes sont des enjeux majeurs dans toute démarche de Data Science. Ce rapport présente en détail les travaux que j'ai réalisés dans le cadre de mon stage, durant lesquels j'ai exploré plusieurs techniques pour aborder ce problème complexe. J'ai étudié différentes approches de détection des données manquantes, en particulier en utilisant des modèles de machine learning, qui permettent d'automatiser ce processus de manière efficace. Par ailleurs, j'ai également expérimenté diverses méthodes pour gérer ces données manquantes, allant de l'imputation par la moyenne ou la médiane à des stratégies plus avancées comme l'imputation basée sur des modèles prédictifs.

2 Présentation du groupe OCP

Le Groupe OCP (Office Chérifien des Phosphates), fondé en 1920, est l'une des plus grandes entreprises marocaines et un leader mondial dans l'industrie des phosphates. Il joue un rôle clé dans l'extraction, la transformation et la commercialisation des phosphates, couvrant l'intégralité de la chaîne de valeur, de l'extraction minière à la production d'engrais complexes.

Ressources et Activités

Le Maroc possède environ 70% des réserves mondiales de phosphates, et l'OCP est responsable de l'exploitation de ces ressources naturelles. Avec plus de 100 ans d'expérience, le groupe s'est imposé comme un acteur stratégique dans le secteur agricole mondial en fournissant des solutions fertilisantes essentielles pour l'agriculture.

Les principales activités de l'OCP se décomposent en trois grands domaines :

— **Extraction des phosphates** : Le groupe gère plusieurs mines situées principalement à Khouribga, Benguerir, Youssoufia et Boucraâ. La mine de Khouribga est l'une des plus importantes au monde, produisant des millions de tonnes de phosphates chaque année.

— **Production d'acide phosphorique et d'engrais** : L'OCP transforme le phosphate brut en divers produits dérivés, notamment l'acide phosphorique, qui est un composant essentiel des engrais chimiques. Ses complexes industriels de Safi et Jorf Lasfar sont au cœur de cette activité, avec des capacités de production massives.

— **Commercialisation mondiale** : Le groupe exporte ses produits dans plus de 160 pays à travers le monde. Grâce à ses filiales et ses partenariats internationaux, l'OCP joue un rôle clé dans la sécurité alimentaire mondiale en fournissant des nutriments essentiels à l'agriculture.

Sites Industriels

Le groupe OCP dispose de plusieurs sites industriels et miniers stratégiques :

— **Khouribga** : Considéré comme le plus grand gisement de phosphates au monde, Khouribga contribue de manière significative à l'offre mondiale de phosphate.

— **Benguerir et Youssoufia** : Ces deux sites sont également essentiels à l'extraction de phosphate dans la région centrale du Maroc. Benguerir est un exemple de technologie minière moderne et de développement durable.

— **Safi et Jorf Lasfar** : Ces deux sites abritent des complexes industriels majeurs pour la production d'acide phosphorique et d'engrais. Jorf Lasfar est reconnu pour être l'un des plus grands hubs industriels mondiaux en termes de transformation de phosphates.

— **Boucraâ** : Situé au sud du Maroc, Boucraâ est une mine de phosphate à ciel ouvert, opérée dans des conditions géographiques difficiles.

Stratégie de Développement et Innovation

L'OCP investit constamment dans l'innovation et les technologies pour améliorer ses processus de production et réduire son empreinte écologique. Le groupe a adopté des stratégies de développement durable, comme l'optimisation de l'utilisation de l'eau et l'intégration de technologies vertes dans ses opérations.

Dans le cadre de sa stratégie de développement, l'OCP a également initié des partenariats internationaux et des joint-ventures, notamment en Afrique et en Asie, pour construire des usines de production d'engrais et étendre sa présence mondiale. Ces investissements permettent à l'OCP de répondre aux besoins croissants en fertilisants, particulièrement dans les régions en développement où l'agriculture est cruciale pour la sécurité alimentaire.

Impact Social et Environnemental

En plus de ses activités industrielles, l'OCP s'engage dans de nombreuses initiatives sociales et environnementales. Le groupe investit dans l'éducation, la formation, et le développement des communautés locales à proximité de ses sites. À travers la Fondation OCP, le groupe soutient des projets d'agriculture durable, de protection de l'environnement et de lutte contre le changement climatique.

L'OCP s'efforce également de minimiser son impact environnemental. Parmi ses initiatives figure l'utilisation d'énergie propre dans ses processus industriels, comme l'énergie solaire et éolienne, ainsi que la réduction de la consommation d'eau grâce à des technologies innovantes de recyclage et de gestion des ressources.

3 Objectifs de stage

Mon stage avait pour but principal de développer des modèles capables de détecter et de traiter les données manquantes dans des jeux de données complexes. Les objectifs spécifiques comprenaient :

- L'analyse des jeux de données pour identifier et comprendre les motifs sous-jacents des données manquantes.
- La mise en place de différentes stratégies pour traiter ces données manquantes, allant des méthodes simples comme l'imputation jusqu'à des approches plus avancées basées sur le machine learning.
- Le développement et la mise en œuvre de modèles de machine learning pour automatiser la détection des données manquantes.
- L'évaluation des performances des modèles à travers diverses métriques, afin de recommander les méthodes les plus efficaces.

4 Méthodologie

4.1 Analyse exploratoire des données

La première étape de mon travail a consisté à analyser le jeu de données fourni pour comprendre l'étendue et la nature des données manquantes. J'ai utilisé des outils de visualisation pour détecter des patterns qui pourraient indiquer des corrélations ou des tendances spécifiques dans les données manquantes.

4.1.1 Distribution des Mesures dans le Jeu de Données

	Measure ID	NGV	Date of Creation	Point Name
0	4211208	1.35	2022-01-22T06:11:43.000000Z	1RV-O
1	4061956	0.42	2022-01-16T05:25:21.000000Z	1RV-O
2	4055996	0.42	2022-01-15T21:08:32.000000Z	1RV-O
3	4130532	1.90	2022-01-19T21:08:55.000000Z	1RV-O
4	4100892	0.34	2022-01-18T07:49:33.000000Z	1RV-O

FIGURE 1 – Extrait des données des mesures avec les valeurs NGV.

La Figure 1 présente un extrait du jeu de données utilisé dans l'étude. Chaque observation est identifiée par un `Measure ID` et contient des informations telles que la valeur de `NGV` (Nombre de Gaz mesuré), la date de création de la mesure, et le nom du point de mesure. Cet extrait montre une distribution variée des valeurs de `NGV`, indiquant une diversité des mesures capturées par différents équipements.

4.1.2 Statistiques Descriptives des Données

	Measure ID	NGV
count	9.334000e+04	93340.000000
mean	1.341557e+07	1.571926
std	4.789173e+06	1.640130
min	4.055996e+06	0.024000
25%	8.069308e+06	0.263000
50%	1.436090e+07	1.297000
75%	1.579799e+07	2.210000
max	2.278229e+07	135.247000

	Measure ID	NGV
count	2.829500e+05	282950.000000
mean	1.665079e+07	1.415128
std	2.814695e+06	8.838158
min	1.257741e+07	0.039230
25%	1.436889e+07	0.737265
50%	1.589401e+07	1.229195
75%	1.867143e+07	1.741847
max	2.259814e+07	2307.053220

FIGURE 2 – Statistiques descriptives des mesures de NGV.

La Figure 2 présente les statistiques descriptives des données, réparties en deux ensembles distincts. On observe un premier ensemble de données comprenant 93 340 observations avec une moyenne de `NGV` de 1.57, et un second ensemble plus large avec 282 950 observations, ayant une moyenne légèrement inférieure de 1.41. Les deux ensembles montrent une dispersion significative, comme en témoigne l'écart-type, et la présence de valeurs maximales très élevées (jusqu'à 2 307 dans le second ensemble), suggérant la présence de valeurs aberrantes ou extrêmes dans le jeu de données.

4.1.3 Évolution des Valeurs NGV en Fonction de la Date de Création

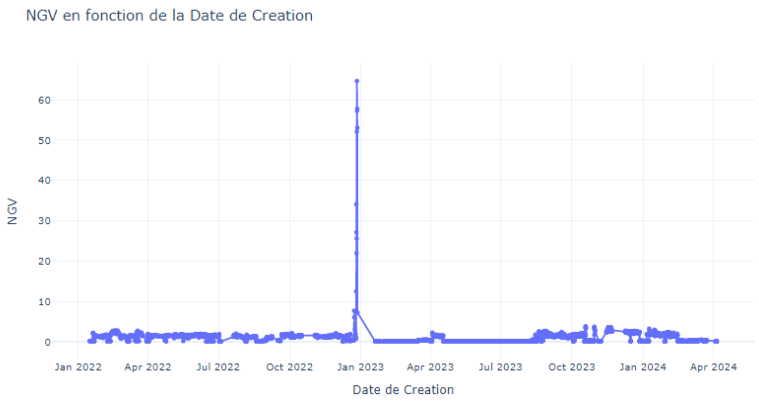

FIGURE 3 – Évolution des valeurs NGV en fonction de la Date de Création

La figure 3 illustre l'évolution des valeurs NGV en fonction de la date de création des mesures. En traçant les courbes des valeurs NGV en fonction du temps, nous avons pu observer les tendances

globales et identifier visuellement des interruptions soudaines ou des chutes drastiques des valeurs, indiquant potentiellement des périodes où des données n'ont pas été enregistrées.

4.1.4 Distribution des Intervalles de Temps Entre deux mesures consécutifs

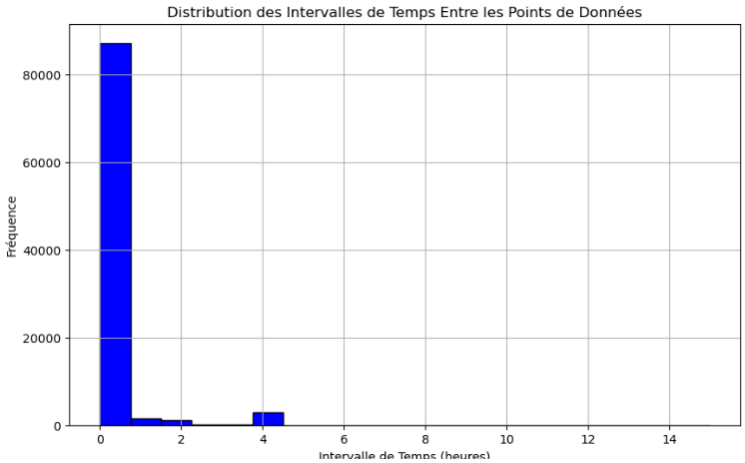

FIGURE 4 – Distribution des Intervalles de Temps

La figure 4 montre la distribution des intervalles de temps entre les mesures. Nous avons calculé les intervalles de temps entre les enregistrements consécutifs en analysant les colonnes de dates et heures de création des données. Cette approche permet de visualiser la distribution de ces intervalles de temps pour détecter des périodes anormalement longues sans enregistrement. Ces périodes prolongées sont un indicateur clé des données manquantes, surtout lorsque la fréquence normale de collecte des données est connue.

4.2 Traitement des données manquantes

Pour gérer les données manquantes, j'ai exploré plusieurs approches.

4.2.1 Approches statistiques

Dans le cadre de ce projet, nous avons mis en œuvre une méthode basée sur l'analyse des quantiles pour détecter efficacement les anomalies dans des séries temporelles. Cette méthode est particulièrement adaptée à la détection des anomalies causées par des données manquantes ou erronées.

Détection des Anomalies par les Pentes

Tout d'abord, nous avons calculé la pente des données NGV en fonction du temps, en déterminant la variation de NGV par rapport à la différence de temps entre chaque mesure. Cette pente permet d'identifier les variations soudaines ou anormales dans les données qui pourraient indiquer une anomalie, telle qu'une coupure de données ou une perte de signal.

Utilisation des Quantiles pour Définir les Seuils

Plutôt que d'utiliser des seuils basés sur la moyenne et l'écart type, ce qui pourrait ne pas être suffisamment robuste face à des données avec de fortes variations, nous avons opté pour une méthode basée sur les quantiles. En particulier, nous avons utilisé le 0.99ème quantile pour définir un seuil supérieur, ce qui nous permet de détecter les pentes extrêmement élevées, et le 0.01ème quantile pour fixer un seuil inférieur, permettant de repérer les pentes exceptionnellement faibles. Ces seuils sont particulièrement efficaces pour identifier les anomalies qui se traduisent par des pentes abruptes ou des stagnations dans les données.

Interpolation et Visualisation des Données

Après avoir identifié les anomalies, nous avons procédé à une interpolation des données pour combler les lacunes ou corriger les valeurs erronées. Cette étape permet de maintenir la continuité et la qualité des données pour des analyses ultérieures. Enfin, nous avons visualisé les résultats en traçant l'évolution des valeurs NGV au fil du temps, en mettant en évidence les anomalies détectées et les périodes interpolées.

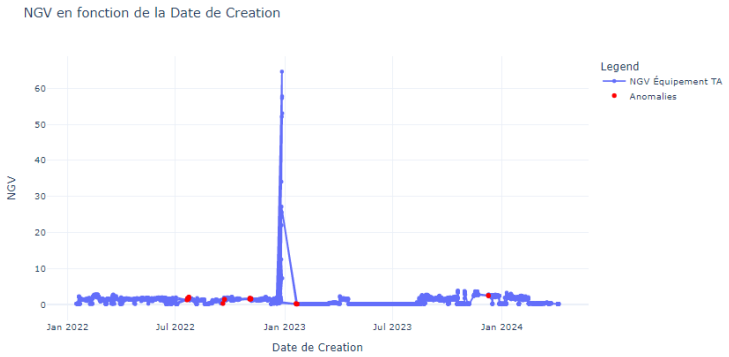

FIGURE 5 – Distribution des Intervalles de Temps

Limites de cette méthode

Une des limites de la méthode utilisée réside dans le fait qu'elle ne prend pas en compte la fréquence de détection des anomalies. En effet, la méthode se concentre principalement sur l'identification des pentes élevées dans les données de NGV, ce qui est souvent interprété comme un signe de perte de données ou d'anomalie. Cependant, une pente élevée n'indique pas nécessairement une anomalie dans tous les cas.

Il est possible que la fréquence d'émission des données ait changé, ce qui peut entraîner des variations abruptes dans les valeurs de NGV sans pour autant signifier une perte de données. Par exemple, si les données sont émises à une fréquence plus élevée ou plus faible qu'auparavant, les pentes calculées pourraient être faussement interprétées comme des anomalies. Cela souligne la nécessité d'intégrer des méthodes complémentaires ou des ajustements pour tenir compte des variations de fréquence dans l'analyse des anomalies.

4.2.2 Développement des modèles de machine learning

J'ai ensuite conçu et mis en œuvre plusieurs modèles de machine learning destinés à automatiser la détection des données manquantes.

Isolation forest

J'ai utilisé en premier l'Isolation Forest. Cet algorithme est bien adapté à la détection d'anomalies, car il isole les anomalies plutôt que de modéliser les instances normales. Voyons les résultats qu'il a donné.

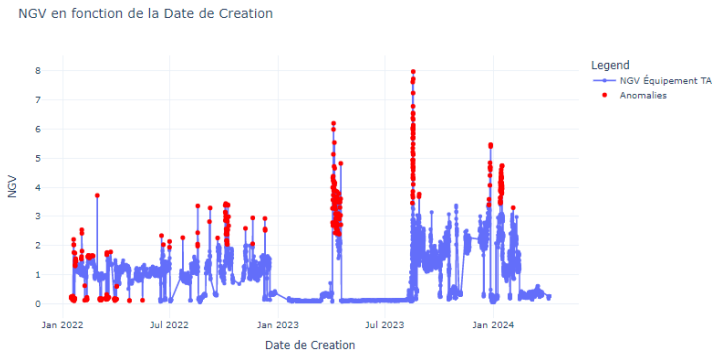

FIGURE 6 – Distribution des Intervalles de Temps

Evaluation du modele : Comme nous pouvons le voir dans la figure, aprés avoir entrainé le modèle il a considéré que les anomalies sont les valeurs extremistes de NGV, cependant c'est pas le cas, donc c'est clairement pas le modèle à prendre en compte.

Réseaux de neurones

Nous avons formulé le problème comme une tâche de prédiction de séquence où le modèle apprend à prédire l'observation actuelle en se basant sur les valeurs passées. Les écarts significatifs entre les valeurs prédites et réelles sont signalés comme des anomalies. Le modèle de réseau de neurones tente d'apprendre le modèle sous-jacent, et les anomalies peuvent indiquer des données manquantes, des changements soudains ou un comportement anormal.

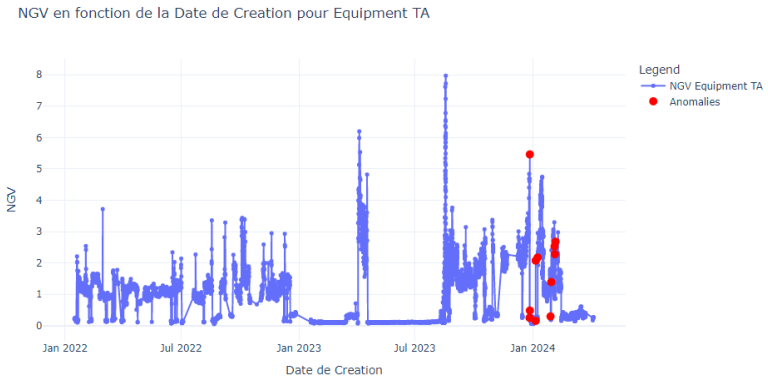

FIGURE 7 – Distribution des Intervalles de Temps

Evaluation de modèle :Il semble que la méthode utilisée pour détecter les anomalies avec le régressseur MLP n'ait pas fonctionné comme prévu pour identifier les données manquantes. Les données manquantes pourraient ne pas présenter les mêmes schémas que les anomalies liées aux valeurs aberrantes ou aux points de données inhabituels.

4.2.3 Solutions envisageables

En examinant de plus près nos données et leurs visualisations, nous constatons une fréquence récurrente dans la réception des valeurs NGV. Cette fréquence reste stable sur des périodes significatives, mais elle présente également des variations notables. Ces variations constituent un élément central de l'approche que nous envisageons d'adopter.

Dans un premier temps, nous analyserons un échantillon de nos données pour mieux comprendre ces variations. À cette fin, nous sélectionnons aléatoirement un fichier Excel de notre jeu de données, traçons les valeurs NGV en fonction des dates, et visualisons les différences temporelles entre les échantillons reçus.

FIGURE 8 – Zoom des valeurs de NGV en fonction de la date

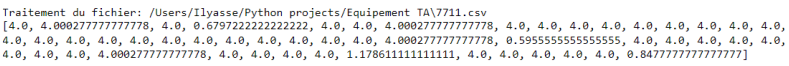

FIGURE 9 – Différence temporelles des valeurs de NGV reçu

Solution 1 :Proposition d'un algorithme

Étape 1 : Initialisation

— **Variables :**
  — `prev_freq` : fréquence précédente (initialisée à la première fréquence observée).
  — `current_state` : état actuel (0, 1 ou 2, selon les conditions décrites).
  — `start_time` : stocke le temps de début lorsque des "1" ou des "2" sont détectés.
  — `is_missing_data` : un indicateur pour savoir si des données manquantes ont été détectées.

Étape 2 : Traitement des Fréquences

— **Boucle sur les Fréquences :** Pour chaque fréquence `current_freq` dans la série de données :
  — **Cas 1 :** `current_freq == prev_freq`
    — Assignez `current_state = 0`.
    — Si le `current_state` précédent était 1, réinitialisez `is_missing_data` à `False` (pas de données manquantes).

— Continuez à la prochaine fréquence.

— **Cas 2 :** `current_freq > prev_freq`
  — Si c'est le premier "1" après un 0, définissez `start_time` à la date actuelle.
  — Si le `current_state` précédent était 1 et que plusieurs "1" se suivent, mettez à jour `is_missing_data = True`.
  — Si un "0" suit plusieurs "1", signalez les données manquantes de `start_time` à la date actuelle - 1 seconde.
  — Assignez `current_state = 1` et mettez à jour `prev_freq` avec `current_freq`.

— **Cas 3 :** `current_freq < prev_freq`
  — Si c'est le premier "2" après un 0, définissez `start_time` à la date actuelle.
  — Si un "0" suit un "2", réinitialisez `is_missing_data` à `False` (pas de données manquantes).
  — Si un "1" suit un "2", comparez la `current_freq` avec la fréquence du dernier 0 :
    — Si elles sont égales, assignez `current_state = 0`.
    — Si la `current_freq` dépasse celle du dernier 0, signalez les données manquantes de `start_time` à la date actuelle - 1 seconde.
    — Sinon, ne faites rien, continuez à la prochaine mesure.
  — Assignez `current_state = 2` et mettez à jour `prev_freq` avec `current_freq`.

Étape 3 : Finalisation

— À la fin du traitement, si des "1" ou des "2" sont restés sans résolution, signalez les données manquantes pour ces périodes.

Résultats de l'algorithme

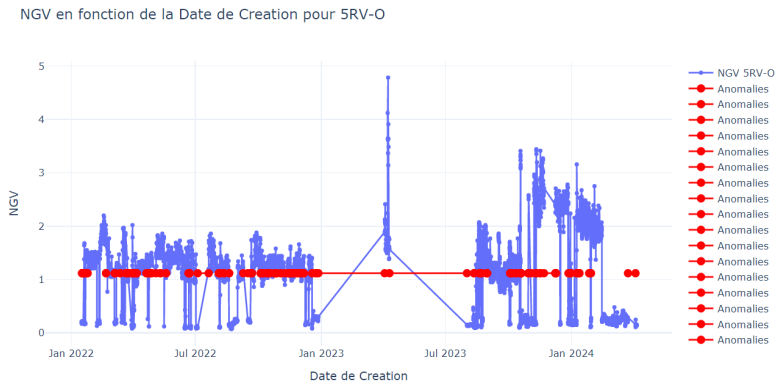

FIGURE 10 – NGV en fonction de la date

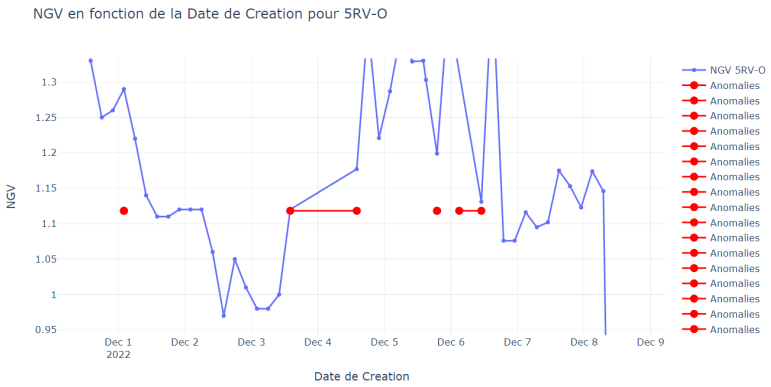

FIGURE 11 – Zoom des valeurs de NGV en fonction de la date

Nous pouvons remarquer que lorsqu'on a une différence de temps remarquable l'algorithme est capable de la détecter, de plus il prend en compte le cas de changement de fréquence, c'est à dire il ne déclare pas la manque d'une donnée dans le cas de changement de fréquence.

Solution 2 : DBSCAN

DBSCAN (density-based spatial clustering of applications with noise) est un algorithme de partitionnement de données proposé en 1996 par Martin Ester, Hans-Peter Kriegel, Jörg Sander et Xiaowei Xu1. Il s'agit d'un algorithme fondé sur la densité dans la mesure qui s'appuie sur la densité estimée des clusters pour effectuer le partitionnement.

Fonctionnement de DBSCAN

L'algorithme DBSCAN utilise 2 paramètres : la distance ϵ et le nombre minimum de points « MinPts » devant se trouver dans un rayon ϵ pour que ces points soient considérés comme un cluster.

DBSCAN fonction de la maniere suivante :

1 – DBSCAN commence par un point de données de départ arbitraire qui n'a pas été visité. Le voisinage de ce point est extrait en utilisant une distance epsilon ϵ.

2 – S'il y a un nombre suffisant de points (selon les minPoints) dans ce voisinage, le processus de mise en cluster démarre et le point de données actuel devient le premier point du nouveau cluster. Sinon, le point sera étiqueté comme bruit (plus tard, ce point bruyant pourrait devenir la partie du cluster). Dans les deux cas, ce point est marqué comme «visité».

3 – Pour ce premier point du nouveau cluster, les points situés dans son voisinage à distance se joignent également au même cluster. Cette procédure est ensuite répétée pour tous les nouveaux points qui viennent d'être ajoutés au groupe de cluster.

4 – Ce processus des étapes 2 et 3 est répété jusqu'à ce que tous les points du cluster soient déterminés, c'est-à-dire que tous les points à proximité du ϵ voisinage du cluster ont été visités et étiquetés.

5 – Une fois terminé avec le cluster actuel, un nouveau point non visité est récupéré et traité, ce qui permet de découvrir un nouveau cluster ou du bruit. Ce processus se répète jusqu'à ce que tous les points soient marqués comme étant visités. A la fin de tous les points visités, chaque points a été marqué comme appartenant à un cluster ou comme étant du bruit.

Motivation

Aprés calcul de fréquence, nous avons remarqué qu'il y a des valeurs de fréquence qui se répétent pendant une durée non négligeable, d'où l'idée de tracer un graphe qui présente le nombre d'occurence des différents fréquences et comme prévu nous avons obtenu une gaussienne, les deux figures ci-dessous répresentent ce graphe appliquée sur deux fichiers de données :

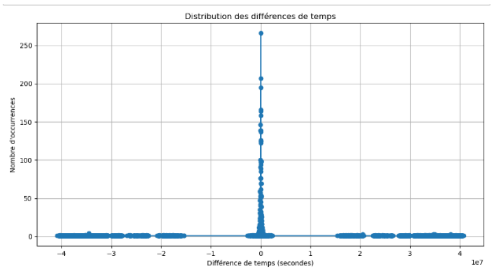

FIGURE 12

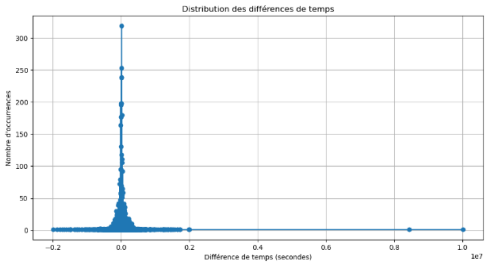

FIGURE 13

l'apparition de courbes en forme de gaussienne, comme illustré dans les Figures 12 et 13, est un indicateur fort de la récurrence des mêmes fréquences sur une période significative. Cette observation suggère une stabilité dans le comportement du système surveillé, avec des variations qui suivent une distribution normale autour de valeurs centrales.

La distribution gaussienne met en lumière des comportements répétitifs ou des patterns réguliers dans les données, ce qui peut être crucial pour détecter des anomalies. En effet, toute déviation marquée par rapport à cette distribution attendue pourrait indiquer un événement inhabituel ou une défaillance du système. C'est là que l'algorithme DBSCAN entre en jeu. En identifiant des clusters denses (où les points sont rapprochés et consistent en des fréquences récurrentes), l'algorithme permet de classifier les données en groupes homogènes. Les points ne correspondant pas à ces groupes sont alors étiquetés comme des anomalies ou du bruit.

Résultats de DBSCAN

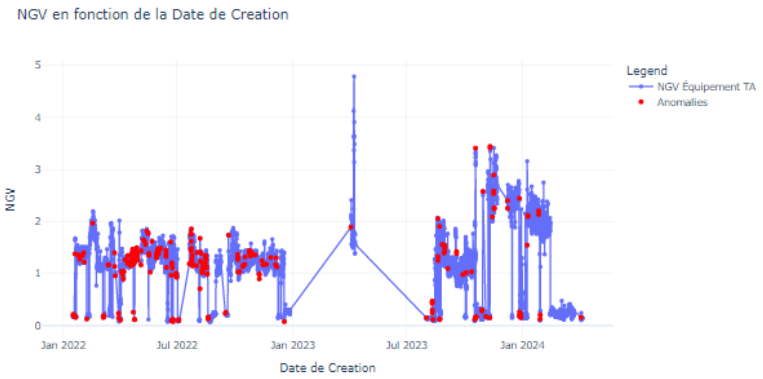

FIGURE 14 – NGV en fonction de la date

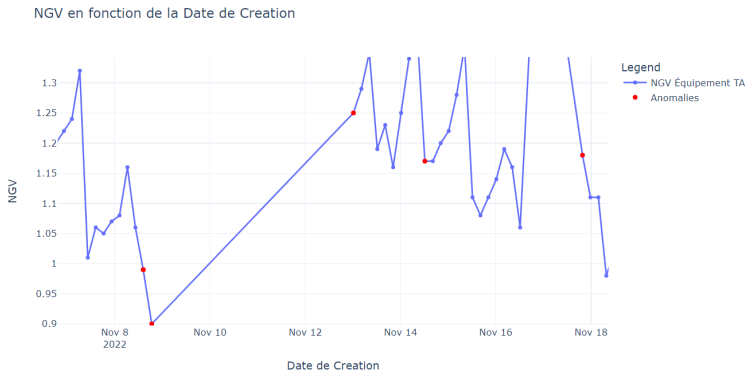

FIGURE 15 – Zoom des valeurs de NGV en fonction de la date

Les anomalies se produisent à des moments où la valeur de NGV soit chute considérablement, soit augmente brusquement par rapport aux points environnants. Par exemple, aux alentours du 8 novembre 2022, une chute de NGV en dessous de 1 a été identifiée comme une anomalie. De même, entre le 12 et le 16 novembre, il y a plusieurs fluctuations qui ont aussi été marquées comme des anomalies.

```
Nombre total de points dans la période : 186
Nombre d'anomalies détectées dans la période : 5

--- Anomalie détectée ---
Date de l'anomalie: 2022-11-13 00:15:44+00:00
Time diff de l'anomalie: 365954.0 secondes

Temps avant l'anomalie :
Date avant l'anomalie 2: 2022-11-08 14:22:31+00:00 avec time diff 207.0 secondes
Date avant l'anomalie 1: 2022-11-08 18:36:30+00:00 avec time diff 15239.0 secondes

Temps après l'anomalie :
Date après l'anomalie 1: 2022-11-13 04:15:44+00:00 avec time diff 14400.0 secondes
Date après l'anomalie 2: 2022-11-13 08:15:44+00:00 avec time diff 14400.0 secondes

--- Anomalie détectée ---
Date de l'anomalie: 2022-11-17 19:41:09+00:00
Time diff de l'anomalie: 69952.0 secondes

Temps avant l'anomalie :
Date avant l'anomalie 2: 2022-11-16 20:15:17+00:00 avec time diff 28800.0 secondes
Date avant l'anomalie 1: 2022-11-17 00:15:17+00:00 avec time diff 14400.0 secondes

Temps après l'anomalie :
Date après l'anomalie 1: 2022-11-17 23:41:09+00:00 avec time diff 14400.0 secondes
Date après l'anomalie 2: 2022-11-18 03:41:09+00:00 avec time diff 14400.0 secondes

--- Anomalie détectée ---
Date de l'anomalie: 2022-11-27 02:00:41+00:00
Time diff de l'anomalie: 37594.0 secondes

Temps avant l'anomalie :
Date avant l'anomalie 2: 2022-11-26 11:34:07+00:00 avec time diff 14400.0 secondes
Date avant l'anomalie 1: 2022-11-26 15:34:07+00:00 avec time diff 14400.0 secondes

Temps après l'anomalie :
Date après l'anomalie 1: 2022-11-27 06:00:41+00:00 avec time diff 14400.0 secondes
Date après l'anomalie 2: 2022-11-27 10:00:41+00:00 avec time diff 14400.0 secondes

--- Anomalie détectée ---
Date de l'anomalie: 2022-12-04 14:00:47+00:00
Time diff de l'anomalie: 86400.0 secondes

Temps avant l'anomalie :
Date avant l'anomalie 2: 2022-12-03 10:00:47+00:00 avec time diff 14400.0 secondes
Date avant l'anomalie 1: 2022-12-03 14:00:47+00:00 avec time diff 14400.0 secondes

Temps après l'anomalie :
Date après l'anomalie 1: 2022-12-04 18:00:47+00:00 avec time diff 14400.0 secondes
Date après l'anomalie 2: 2022-12-04 22:00:47+00:00 avec time diff 14400.0 secondes

--- Anomalie détectée ---
Date de l'anomalie: 2022-12-06 10:58:01+00:00
Time diff de l'anomalie: 43200.0 secondes

Temps avant l'anomalie :
Date avant l'anomalie 2: 2022-12-05 18:58:01+00:00 avec time diff 14402.0 secondes
Date avant l'anomalie 1: 2022-12-05 22:58:01+00:00 avec time diff 14400.0 secondes

Temps après l'anomalie :
Date après l'anomalie 1: 2022-12-06 14:58:01+00:00 avec time diff 14400.0 secondes
Date après l'anomalie 2: 2022-12-06 18:58:01+00:00 avec time diff 14400.0 secondes
```

FIGURE 16

Pour démontrer l'efficacité de la méthode, j'ai choisi de non seulement indiquer la date à laquelle l'anomalie a été détectée, mais aussi de présenter les deux dates précédentes et les deux dates suivantes. Cela permet de constater que la fréquence reste stable avant et après l'anomalie,

alors qu'une différence significative de la fréquence est observée précisément au point d'anomalie.

4.3 Comparaison entre les solutions

Pour faire, j'ai choisi d'examiner les anomalies détectées par l'algorithme sur une période donnée et de les comparer à celles identifiées par DBSCAN.

```
Analyse des données manquantes pour le point : 5RV-O
Il y a des données manquantes entre 2022-11-08 18:22:31+00:00 et 2022-11-13 00:15:43+00:00
Il y a des données manquantes entre 2022-11-14 16:15:15+00:00 et 2022-11-14 16:15:16+00:00
Il y a des données manquantes entre 2022-11-16 16:15:17+00:00 et 2022-11-16 20:15:16+00:00
Il y a des données manquantes entre 2022-11-17 00:15:17+00:00 et 2022-11-17 19:41:08+00:00
Il y a des données manquantes entre 2022-11-20 07:34:05+00:00 et 2022-11-20 07:34:06+00:00
Il y a des données manquantes entre 2022-11-22 11:34:07+00:00 et 2022-11-22 15:34:07+00:00
Il y a des données manquantes entre 2022-11-26 19:34:07+00:00 et 2022-11-27 02:00:40+00:00
Il y a des données manquantes entre 2022-12-01 02:00:41+00:00 et 2022-12-01 02:00:46+00:00
Il y a des données manquantes entre 2022-12-03 14:00:47+00:00 et 2022-12-04 14:00:46+00:00
Il y a des données manquantes entre 2022-12-05 18:57:59+00:00 et 2022-12-05 18:58:00+00:00
Il y a des données manquantes entre 2022-12-06 02:58:01+00:00 et 2022-12-06 10:58:00+00:00
```

FIGURE 17

Tout d'abord, il est notable que le nombre d'anomalies détectées par DBSCAN est inférieur à celui détecté par l'autre algorithme. Examinons maintenant les anomalies supplémentaires trouvées par l'autre algorithme. En effet, toutes les anomalies détectées par DBSCAN sont incluses dans celles détectées par l'autre algorithme, mais ce n'est pas encore suffisant pour tirer une conclusion définitive. Il est important de noter que les données manquantes entre les périodes telles que 2022-11-14 16 :15 :15+00 :00 et 2022-11-14 16 :15 :16+00 :00, ainsi que celles entre 2022-11-16 16 :15 :17+00 :00 et 2022-11-16 20 :15 :16+00 :00... ne semblent pas indiquer des données manquantes significatives, étant donné la faible différence de temps.

Conclusion : La méthode de DBSCAN est plus efficace que l'algorithme proposés.

Explication théorique de la conclusion :L'algorithme proposé se base sur la comparaison des fréquences sans tenir compte de la valeur réelle de ces fréquences. Or, cette valeur est cruciale pour évaluer correctement si une donnée est réellement manquante ou non.

5 Conclusion

Dans ce rapport, nous avons pu explorer nos données, déterminer leurs caractéristiques afin de créer le meilleur modèle pour traiter le problème, aprés avoir tester différentes démarches et approches nous avons pu conclure l'efficacité de clustering avec l'algorithme DBSCAN pour résoudre le problème.