République Algérienne Démocratique et Populaire
Enseignement Supérieur et de la Recherche Scientifique
Université M'Hamed Bougara Boumerdès
Faculté des sciences
Département des mathématiques

Mémoire Présenté Pour
L'Obtention Du Diplôme De Master En Recherche Opérationnelle
Option : Recherche Opérationnelle Optimisation Et Management Stratégique (ROOMS)

Réalisé par :
Boubeguira Mohamed Ramzi
Douak Atif

Prédiction de la vitesse de corrosion sur base de paramètres chimiques dans les sites pétroliers

Devant le jury :

Présidente	Mme. K.BOURENNANI	M.A.A	U.M.B.B
Examinateur	M. A.LAOUZAI	M.C.B	U.M.B.B
Promotrice	Mme. W.DRICI	M.C.B	U.M.B.B
Co-Promoteur	M. S.TAHERBOUCHET	M.C.B	U.M.B.B

Année Universitaire 2022 - 2023

Remerciements

En préambule de ce mémoire, nous remercions avant tout ***ALLAH****, créateur de l'univers qui nous a maintenues en santé pour mener à bien cette année d'étude.*

Nous tenons d'abord à remercier nos parents, pour leurs conseils ainsi que pour leurs soutiens inconditionnels, à la fois moraux et économiques.

Un immense merci aussi pour l'équipe du département de corrosion pour leur chaleureux accueil, le partage de connaissances et leur disponibilité. Sans oublier la qualité de leur encadrement. Comme nous sommes très reconnaissantes envers Monsieur ***LAOUZAI*** *et Madame* ***ZERROUKI*** *de nous avoir acceptés et orientés au sein de l'entreprise et d'avoir mis à notre disposition tout ce dont nous avions besoin.*

Nous désirons aussi remercier notre encadreuse Madame ***DRICI*** *pour ses conseils avisés concernant la rédaction de ce mémoire et pour la qualité de l'enseignement offert pendant notre cursus universitaire. Nous tenons également à exprimer notre reconnaissance et notre profond respect envers les enseignants de notre formation pour avoir su nous transmettre leurs savoirs. Et bien sûr, Madame* ***BENMENSOUR****, qui a toujours été à notre écoute et a su nous apporter un soutien sans faille, notamment en ce qui concerne la programmation.*

Notre sincère remerciement s'adresse à tout le personnel du département "Mathématiques".

Un grand merci à nos familles surtout nos parents, qui nous ont aidées à suivre nos études dans les meilleures conditions et qui nous ont toujours soutenues et encouragées sans limite.

Nos remerciements s'adressent aussi aux membres du Jury Madame ***BOURENNANI*** *et Monsieur* ***LAOUZAI*** *qui nous ont fait l'honneur de juger ce modeste travail.*

Enfin, nous tenons à remercier toutes les personnes, de près ou de loin, qui nous ont conseillé lors de la réalisation de ce projet.

Boubeguira Mohamed Ramzi
Douak Atif

Dédicaces

À mes chers parents, frères, sœurs et amis,

C'est avec une immense reconnaissance que je dédie ce mémoire de fin d'étude à chacun d'entre vous. Votre amour inconditionnel, votre soutien constant et votre encouragement ont été des piliers essentiels dans ma vie et ont joué un rôle déterminant dans la réalisation de ce projet.

À mes parents, je vous suis reconnaissant pour vos sacrifices inlassables et votre dévouement sans faille. Votre soutien indéfectible, vos encouragements sans relâche et vos précieux conseils m'ont permis de croire en moi-même et de poursuivre mes objectifs avec détermination. Votre amour et votre confiance inébranlables ont été ma source d'inspiration tout au long de ce parcours académique.

À mon petit frère et ma petite sœur, vous êtes mes meilleurs amis et mes plus grands alliés. Votre présence constante, vos encouragements enthousiastes et votre soutien inconditionnel ont été d'une valeur inestimable. Vos mots d'encouragement et votre confiance en mes capacités ont été des moteurs de motivation qui m'ont aidé à persévérer dans les moments difficiles.

À mes amis, vous êtes ma deuxième famille. Vos encouragements, vos discussions stimulantes et votre présence joyeuse ont rendu ce voyage académique mémorable et agréable. Vos encouragements inépuisables et votre confiance en moi ont été une source d'inspiration et de soutien qui m'a donné la force de continuer à avancer.

*Je remercie tout particulièrement **Atif**, mon partenaire de mémoire, mon binôme, mon ami... sans qui rien n'aurait été pareil. Cette année fut riche en émotions et je tiens à te remercier pour ton soutien et ce lien tout particulier qui s'est créé entre nous.*

Ce mémoire de fin d'étude est dédié à chacun d'entre vous, car vous avez été mes piliers dans les moments de doute, mes moteurs de motivation lorsque la fatigue se faisait sentir, et mes sources de joie et de bonheur lorsque la réussite était au rendez-vous. Votre amour, votre soutien et votre amitié sont des trésors que je chérirai toujours.

Avec tout mon amour et ma gratitude,

Ramzi

Dédicaces

Je dédie ce modeste travail à mes chers parents, pour leur amour, leur tendresse, leurs prières, leur soutien et leurs sacrifices tout au long de mes études.

A mes chères frères et sœurs pour leur encouragement et leur soutien moral. Je dédie mon travail aussi à toute ma famille et mes amis pour le soutien tout au long de mon parcours universitaire.

Je remercie particulièrement mon partenaire de projet de fin d'étude, mon binôme et mon ami ***Ramzi****, qui a contribué à la réalisation de ce travail.*

À tous mes amis de l'association ***Aayla****, qui ont été ma deuxième famille et une source d'encouragement indéfectible. Votre amitié sincère, votre soutien mutuel et notre collaboration passionnée ont rendu ce projet possible.*

À tous les gens que j'aime et dont je n'ai pas cité les noms.

À tous mes camarades de la promotion sortante 2023 Recherche Opérationnelle.

Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fruit de votre soutien infaillible.

Merci d'être toujours là pour moi.

Atif

Résumé

La corrosion est un phénomène majeur dans l'industrie pétrolière et gazière, entraînant des dommages matériels, des problèmes de sécurité et des interruptions coûteuses de la production.

La relation entre les paramètres chimiques de l'eau et la vitesse de corrosion est d'une importance cruciale pour comprendre et prédire ce phénomène.

La régression linéaire multiple offre une méthode analytique permettant de modéliser cette relation complexe. En utilisant des techniques avancées de statistiques, il est possible d'identifier les paramètres chimiques de l'eau qui ont le plus d'impact sur la vitesse de corrosion.

Ces informations peuvent ensuite être utilisées pour prédire et contrôler la corrosion, contribuant ainsi à la durabilité des matériaux métalliques dans diverses applications industrielles et environnementales.

Notre objectif est de trouver un modèle mathématique en utilisant la régression linéaire multiple ce qui nous a donnée avec quelque tests statistiques tels que Fisher et Student le modèle recherché.

Nous avons aussi fait une comparaison entre 4 modèle avec le calcul des mesures statistiques tels que le AIC et le BIC.

Table des matières

Table des figures

Liste des tableaux

Introduction générale :

Dès 1830, le physicien Auguste De La Rive a proposait une théorie électrochimique en ce qui concerne le phénomène corrosion. Ces recherches n'ont véritablement pris leur essor qu'au XXe siècle. Leur but est double : déterminer le processus des phénomènes afin de leur trouver un remède, et définir les matériaux susceptibles d'être utilisés dans des conditions données pendant une durée qui est parfois de plusieurs décennies, comme c'est le cas pour certaines installations pétrolières tels que les sites pétroliers de SONATRACH.

La corrosion constitue un défi majeur pour l'industrie pétrolière et gazière, notamment pour les entreprises telles que SONATRACH qui exploite des pipelines pour le transport des hydrocarbures. Les canalisations sont exposées à des conditions environnementales extrêmes, y compris la présence d'eaux corrosives. Pour faire face à ce problème, SONATRACH doit trouver des solutions innovantes pour prédire et prévenir la corrosion dans ses canalisations.

Dans le cadre de ce stage en recherche opérationnelle chez SONATRACH, notre objectif est de développer un modèle mathématique ($Y = \beta_0 + \beta_1 * X_1 + \ldots + \beta_n * X_n$) qui permet de prédire la vitesse de corrosion dans les pipelines en fonction des paramètres chimiques des eaux utilisées, telles que les eaux d'injection ou celles présentes dans les puits de production. Compte tenu de la diversité des eaux rencontrées sur les différents sites pétroliers, il est essentiel de trouver une équation fiable qui estime la vitesse de corrosion, afin d'éviter la formation de piqûres dans les tuyaux et de permettre à SONATRACH de prévoir a efficacement quand et où ces phénomènes de corrosion se produiront. Ainsi, des mesures de maintenance adéquates pourraient être prises en temps voulu.

Ce travail d'étape est structuré en quatre chapitres.
Le premier chapitre présentera l'organisme d'accueil, SONATRACH, en mettant en évidence son rôle clé dans l'industrie pétrolière et gazière, ainsi que ses activités et ses défis spécifiques en matière de corrosion dans les pipelines.

Le deuxième chapitre sera consacré à une définition approfondie de la corrosion et à l'étude des eaux utilisées dans les sites pétroliers. Nous examinons la relation entre la corrosion et les paramètres chimiques de ces eaux, tels que le pH. Une compréhension approfondie de ces facteurs sera essentielle pour développer notre modèle de prédiction.

Le troisième chapitre fournit une introduction détaillée à la régression linéaire, une méthode statistique couramment utilisée pour estimer les relations entre les variables indépendantes et la variable dépendante. Nous discuterons des concepts clés de la régression linéaire multiple,

qui nous permettront d'intégrer plusieurs paramètres chimiques des eaux dans notre modèle de prédiction de la corrosion.

Enfin, le quatrième chapitre sera consacré à la partie pratique de notre travail de recherche. Nous présentons les données que nous avons utilisées, la méthodologie de collecte de ces données et les étapes de résolution du problème. Nous décrivons en détail la mise en œuvre de la régression linéaire multiple pour construire notre modèle de prédiction de la vitesse de corrosion. De plus, nous discuterons des résultats obtenus et de leur interprétation, ainsi que des implications pratiques de notre modèle pour SONATRACH.

En conclusion, ce travail met en évidence l'importance de prédire la vitesse de corrosion pour prévenir les dommages matériels, garantir la sécurité des sites et maintenir la production continue. Le modèle de régression de vitesse linéaire multiple a développé une solution prometteuse pour estimer cette corrosion en fonction des paramètres chimiques des eaux utilisées. Son utilisation pourrait permettre à SONATRACH de prendre des mesures de maintenance appropriées et de minimiser les risques associés à la corrosion des pipelines.

Chapitre 1

Présentation de l'organisme d'accueil

1.1 HISTORIQUE

SONATRACH est une entreprise nationale de dimension internationale et d'un poids économique considérable pour l'économie algérienne.
Au lendemain de l'indépendance, l'État Algérien a pris la décision de s'approprier ses richesses pétrolières et gazières, et de se doter d'un instrument de développement réunissant toutes les conditions de sa souveraineté, par la création de la SONATRACH (Société Nationale de Transport et Commercialisation des Hydrocarbures), le 31/12/1963 par le décret N° 63/491 paru dans le journal officiel le 10/01/1964.
En 1965, la SONATRACH a pu réaliser son premier défi qui était de concevoir et de poser le premier pipeline Algérien reliant le champ de HAOUD EL HAMRA-ARZEW d'un diamètre de 28 pouces et d'une longueur de 801 Km.
Le 22 Septembre 1966, le décret N° 66/296 a redéfinit la nouvelle mission de SONATRACH pour devenir "société nationale pour la recherche, la production, le transport, la transformation et la commercialisation des Hydrocarbures".
A l'orée des années 80, Sonatrach s'est engagée selon un plan quinquennal dans un nouveau processus de restructuration étendue, qui a abouti à la création de 17 entreprises telles que NAFTAL, GCB, ENTP, ENAC.
Aujourd'hui Sonatrach a consenti des efforts considérables en exploration, développement et exploitation de gisements, en infrastructures d'acheminement des hydrocarbures, en usines de liquéfaction de gaz naturel et en méthaniers. SONATRACH, qui est en tête des compagnies pétrolières en Afrique, pointe au 12ème rang mondial, et par ailleurs le deuxième exportateur de GNL (Gaz Naturel Liquéfié) et de GPL (Gaz de Pétrole Liquéfié) et le troisième exportateur de gaz naturel au monde.

FIGURE 1.1 – Logo SONATRACH

1.2 Les missions de SONATRACH

Les principales missions de SONATRACH sont :
— Prospection, exploration et exploitation.
— Développement, gestion et management de transport, stockage et moyens de chargement, Marketing, Transformation et raffinage.
— Liquéfaction du gaz naturel, traitement et valorisation des hydrocarbures gazeux.
— Mise en place de toutes formes d'activités en joint-venture à l'intérieur et à l'extérieur du territoire algérien, en collaboration avec des compagnies étrangères.
— Une fourniture constante d'hydrocarbures à usage domestique, la recherche pour la promotion et la valorisation de toute forme d'énergie et source.
— Le développement de toute activité en lien direct.

1.3 Les objectifs stratégiques de SONATRACH

Ses objectifs se basent sur :
— Un contrôle permanent des activités principales.
— Le renforcement de ses capacités technologiques et de gestion.
— Un partenariat authentique ainsi qu'une expansion internationale.
— La diversification de son portefeuille d'activités.

1.4 L'organisation de SONATRACH

Aujourd'hui, SONATRACH doit s'inscrire dans une nouvelle dynamique plus agile et efficace dans son organisation et son fonctionnement pour affronter les défis qu'il lui faut relever pour servir une Algérie plus prospère. La transformation de l'Entreprise donne le cap de la nouvelle politique mise en œuvre pour transformer l'entreprise en profondeur.

La macrostructure de Sonatrach s'articule autour des directions suivantes :
— La Direction Générale.
— Les Structures opérationnelles.
— Les Structures fonctionnelles.

1.4.1 La direction générale

Elle est assurée par le Président Directeur Général, assisté par :
— Un comité Exécutif.
— Un Secrétaire Général.
— Un cabinet.
La Direction Générale est dotée des autres comités suivants :
— Le Comité d'Examen des Projets (CEP).
— Le Comité de Coordination des Projets Internationaux (CPI).

— Le Comité d'Ethique.
— Le service Sureté Interne d'Ètablissement (SIE).

1.4.2 Les structures opérationnelles

Chaque activité exerce ses métiers, développe son portefeuille d'affaires et contribue dans son domaine de compétences, au développement des activités internationales de la Société. Chaque activité est placée sous l'autorité directe d'un vice président.

Les structures opérationnelles sont organisées comme suit :
— L'Activité Exploration-Production (EP) a pour mission la recherche, le développement, l'exploitation et la production des hydrocarbures.
— L'Activité Transport par Canalisation (TRC) a pour missions de développer le réseau d'infrastructures de Transport des hydrocarbures depuis les pôles de production au sud vers les pôles de demande et de transformation au nord (marché national et exportation).
— L'Activité Liquéfaction-Séparation (LQS) a pour mission la transformation des hydrocarbures par la liquéfaction du gaz naturel et la séparation des GPL.
— L'Activité Raffinage et Pétrochimie (RPC) a pour mission essentielle l'exploitation et la gestion de l'outil de production du Raffinage et d la Pétrochimie.
— L'Activité Commercialisation(COM) a pour mission de veiller aux approvisionnements énergétiques du marché national.

1.4.3 Les structures fonctionnelle

Elaborent et veillent à l'application des politiques et stratégies du groupe, elles sont organisées en plusieurs directions :
— La Direction Communication (CMN)est chargée de l'élaboration et de la mise en œuvre de la stratégie de communication de SONATRACH.
— La Direction Transformation (TRF) est chargée de la coordination et du suivi de la mise en œuvre du plan de transformation de la transformation de l'entreprise.
— La Direction Corporate Stratégie, Planification et Économie (SPE) est chargée de l'élaboration et le développement à moyen et long terme et d'évaluer leur mise en œuvre.
— La Direction Corporate Finances (FIN)est chargée d'élaborer les politiques et stratégies dans le domaine de la Finance. Elle évalue leur mise en œuvre et veille à la qualité de l'information financière.
— La Direction Corporate Business Développement et Marketing (BDM) est chargée de formuler la stratégie de croissance et de recherche des opportunités d'investissement pour la Société.
— La Direction Corporate Ressources humaines (RHU)est chargée de l'élaboration des politiques et stratégies en matière de ressources humaines et du contrôle de leur mise en œuvre.

— La Direction Centrale Procurement Logistique (P& L) a pour mission de piloter les processus d'achats et la Logistique pour le Groupe.
— La Direction Centrale Ressources Nouvelles (R& N)est chargée de piloter et d'exploiter, depuis le centre, les projets de Ressources Non Conventionnelles et l'Offshore.
— La Direction Centrale Engineering & Project Management (EPM), assure le pilotage et l'exécution des grands projets industriels du Groupe.
— La Direction Centrale juridique (JUR)est en charge de l'élaboration et de l'harmonisation des instruments juridiques et du contrôle de leurs applications.
— La Direction Centrale Digitalisation et Système d'information (DSI)est chargée de la définition et du contrôle de la politique informatique et de la digitalisation de la Société.
— La Direction Centrale Santé, Sécurité et Environnement (HSE)a en charge l'élaboration des politiques en matière d'environnement, de sécurité et de qualité de vie au travail. Elle assure le contrôle de leur application.
— La Direction Centrale de la Recherche et du Développement (R& D) est chargée de promouvoir et de mettre en œuvre la politique de la recherche appliquée et développement des technologies dans les métiers de base de la Société.

1.4.4 Organigramme de la marcostructure de SONATRACH

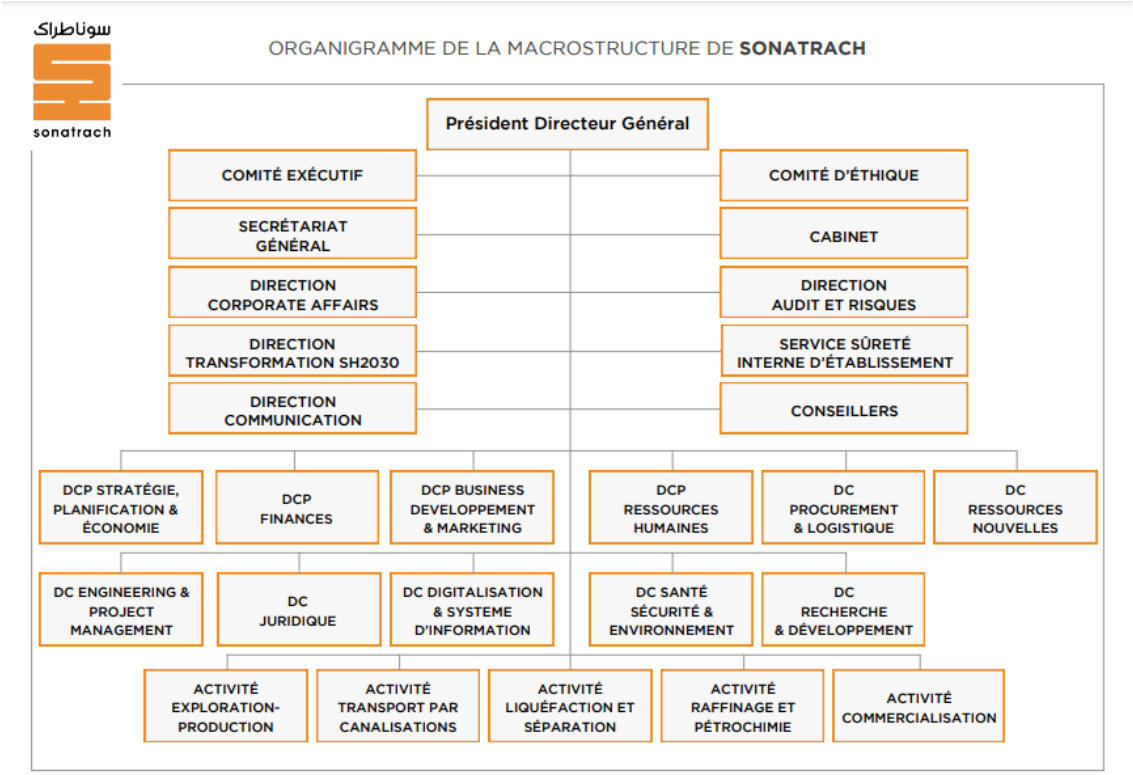

FIGURE 1.2 – Macrostructure De SONATRACH

1.5 Présentation de la division laboratoire de SONATRACH

L'activité Exploration-Production est basée sur les travaux d'explorations, le forage, les services au puits, le développement des gisements et l'exploitation des gisements, cette activité est divisée en plusieurs divisions y compris la Division Laboratoire.
La Division Laboratoire a été créée en 1973 à Dar El Beida, et a été installée à Boumerdès en 1975, cette structure est devenue un outil scientifique et technique indispensable pour les structures opérationnelles de la SONATRACH, aussi bien en amont qu'en aval du domaine pétrolier.
La division laboratoire est structurée d'une part, de cinq directions techniques :

1.5.1 Direction recherche

C'est la direction responsable de la recherche, de l'innovation et du développement technologique dans le domaine des hydrocarbures, avec pour objectif d'optimiser la découverte, l'exploitation et la récupération des ressources pétrolières et gazières en Algérie. Elle collabore souvent avec des instituts de recherche, des universités et d'autres partenaires de l'industrie pour promouvoir l'échange de connaissances, la recherche conjointe et le développement de solutions innovantes.

1.5.2 Direction gisement

Elle vise à optimiser la gestion des gisements pétroliers et gaziers de la société, en garantissant une exploitation efficace et durable des ressources hydrocarbures en Algérie.
La Direction Gisement se compose de trois département :

- Département Caractérisation des Réservoirs
- Département Études Thermodynamiques
- Département Caractérisation produits pétroliers stabilisés

1.5.3 Direction géologie

Est responsable de la gestion des activités géologiques liées à l'exploration et à la production d'hydrocarbures. Elle mène des études géologiques approfondies, interprète les données sismiques, crée des modèles de réservoirs et évalue les risques géologiques. Elle surveille également les opérations de forage et collabore avec d'autres disciplines telles que la géophysique et l'ingénierie des réservoirs. La Direction Géologie participe à la recherche et au développement pour améliorer les techniques d'exploration et d'exploitation des hydrocarbures. Son rôle est essentiel pour fournir des informations géologiques précieuses et contribuer à la prise de décisions stratégiques liées aux gisements d'hydrocarbures de SONATRACH.
Cette direction se compose de :

— Département sédimentologie
— Département géochimie
— Département stratigraphie

1.5.4 Direction assistance aux unités industries

— Département environnement
— Département traitement et contrôle des fluides
— Département corrosion

1.5.5 Direction laboratoires et cartothèque centrale

— Département analyses
— Département roches réservoirs
— Département cartothèque centrale
— Département administration générale

Et de trois directions de soutien :

Direction gestion personnel et moyens

— Département développement des ressources humaines
— Département gestion des ressources humaines
— Département moyens généraux
— Département approvisionnement

Direction finance et juridique

— Département budget et financement
— Département comptabilité et facturation
— Département juridique

Direction générale

— Département technique
— Département QHSE
— Département technologie de l'information
— Assistant sureté Interne
— Conseil scientifique
— Centre conservation d'exposition du patrimoine Géo & HC

1.5.6 Organigramme de la division laboratoire de SONATRACH

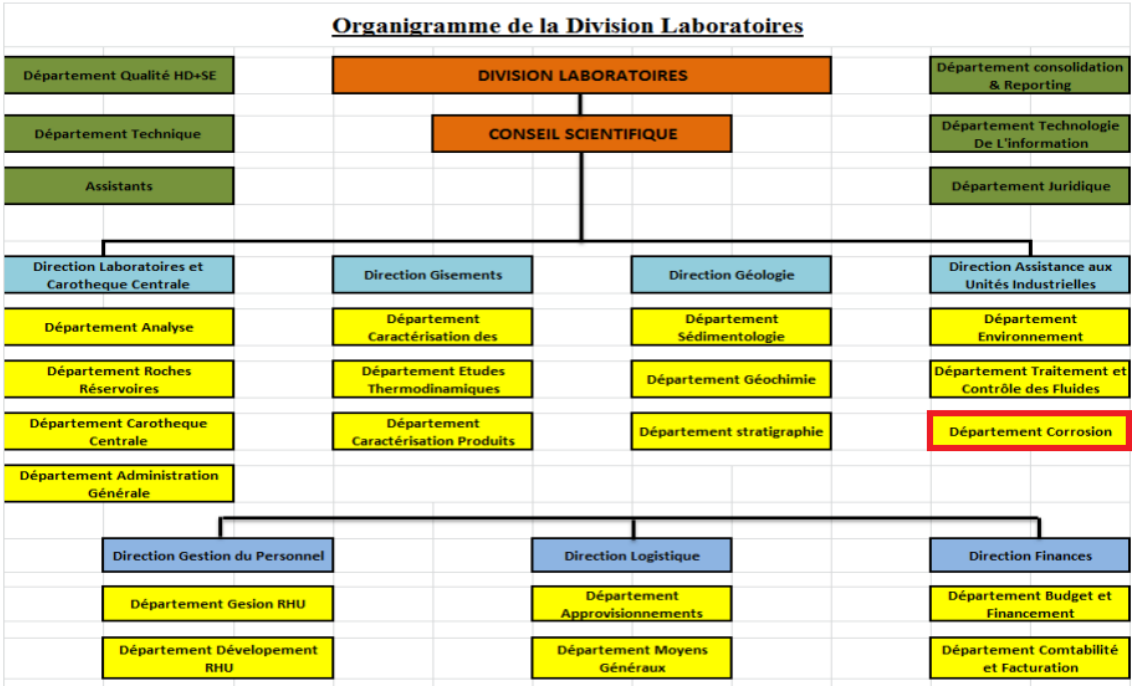

FIGURE 1.3 – Organigramme de la division laboratoire de SONATRACH

1.6 Département corrosion

1.6.1 Service corrosion électrochimique et métallurgique :

C'est le phénomène de corrosion le plus important et elle se manifeste lorsque le réactif est un liquide ou lorsqu'il existe une hétérogénéité soit dans le métal ou dans le réactif, présentant une dissymétrie de composition.
L'existence de ces hétérogénéités détermine la formation d une pile, alors un courant électrique circule entre anodes et cathodes dans le réactif et les zones qui constituent les anodes sont attaquées(corrodées).

1.6.2 Les méthodes d'évaluation de la résistance à la corrosion :

- Mesure de la vitesse de transfert de charge aux interfaces.
- La vitesse de transport de matière.
- Le degré d'adsorption.
- Les vitesses et les constantes d'équilibre des réaction chimique.

Pour ce faire, il faut le suivi des paramétrés potentiel(E), courant(l) et le temps (T) est réalise dans une cellule électrochimique a trois électrodes.

Chapitre 2

Corrosion et composition chimique des eaux sur les sites pétroliers

Introduction

Dans ce chapitre nous allons voir la relation entre la corrosion et la composition chimique des eaux, en effet cette dernière est un sujet d'intérêt crucial dans de nombreux domaines et principalement dans notre cas l'industrie pétrolière plus précisément dans les Piplines et les sites pétroliers. La corrosion, un processus destructeur qui altère les matériaux métalliques, peut être fortement influencée par la composition chimique des eaux en contact avec ces matériaux. Comprendre cette relation est essentiel pour prévenir et atténuer les problèmes de corrosion. Cet article explorera donc l'impact de la composition chimique des eaux sur la corrosion.

La composition chimique des eaux, y compris les niveaux de pH, la teneur en oxygène dissous, les concentrations d'ions et la présence de contaminants, joue un rôle significatif dans la corrosion des métaux.

De plus, la présence de contaminants dans l'eau, tels que les produits chimiques industriels, les métaux lourds ou les microorganismes, peut également avoir un impact significatif sur la corrosion. Certains contaminants peuvent agir comme des catalyseurs de corrosion, accélérant ainsi le processus.

Il est important de noter que la corrosion peut être un phénomène complexe et multifactoriel, et la composition chimique de l'eau est seulement l'un des nombreux facteurs qui peuvent influencer la corrosion. D'autres facteurs tels que la température, la vitesse de circulation de l'eau, la rugosité de surface et les propriétés spécifiques des matériaux métalliques jouent également un rôle.

En conclusion, la composition chimique des eaux a un impact significatif sur la corrosion des matériaux métalliques. ce qui nous mène dans ce chapitre a présenter toutes les notions et les définitions qui concerne la corrosion et la composition chimique présentes dans les sites pétroliers pour en fin comprendre la relation entre eux.

2.1 Définition :

Evans puis Wagner et Traud sont les premiers à avoir défini la corrosion en présence d'une phase liquide(corrosion dû à la présence de liquide), la corrosion est une réaction chimique ou électrochimique entre un matériau, généralement un métal, et son environnement qui entraine une dégradation du matériau et de ses propriétés,sous l'effet de l'environnement immédiat qui peut être le sol, l'atmosphère, l'eau ou d'autres fluides (gaz,pétrole). Compte tenu du nombre important de paramètres intervenant dans le processus électrochimique, la corrosion est un phénomène très complexe. Cette dernière peut être vue sous sa forme globale comme une réaction spontanée d'échange d'électrons à l'interface métal/environnement. C'est un phénomène naturel qui tend à faire retourner les métaux à leur état d'oxyde par une attaque plus ou moins rapide du milieu corrosif.
Dans ce chapitre nous nous intéresserons essentiellement à la corrosion aqueuse (corrosion électrochimique).[12]

2.2 Causes de la corrosion :

Dans la nature tous les métaux, à l'exception des métaux nobles tels que l'or (Au) et le platine (Pt), se présentent dans la nature sous forme d'oxydes et de sulfures métalliques. Cet état de point de vue thermodynamique est très stable. Cependant, l'énergie considérable fournit pour l'obtention des métaux de ces minerais fait que les métaux obtenus se trouvent dans un niveau énergétique élevé, ils sont thermodynamiquement instables. C'est pour cette raison que tous les métaux usuels ont tendance à retourner à leur état initial en énergie, cela se fait à l'aide du milieu environnant. [18]

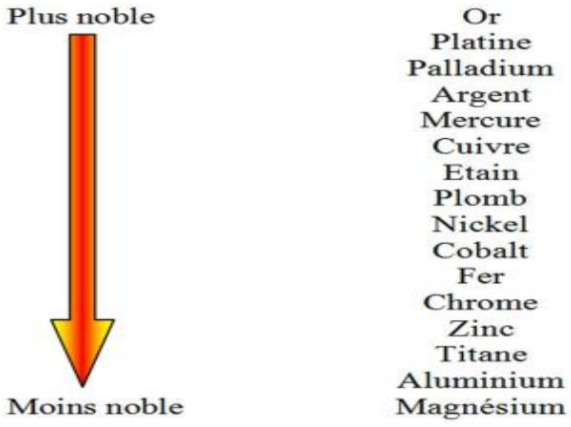

FIGURE 2.1 – Classification des métaux

2.3 Classification de la corrosion :

La corrosion se développe selon deux processus :
— La corrosion sèche.
— La corrosion humide.

2.3.1 La corrosion sèche :

En général, la corrosion des métaux est favorisée par la présence d'eau ou d'humidité dans l'environnement. Cependant, il existe également des gaz qui peuvent corroder les métaux en l'absence d'eau, comme les gaz sulfureux, les gaz acides, ou encore les gaz chlorés.
Ces gaz peuvent réagir avec la surface des métaux à des températures très élevés pour former des produits de corrosion qui peuvent affecter la résistance et la durabilité des matériaux.
Ce phénomène joue un rôle très important dans les appareils qui fonctionnent à haute température.

2.3.2 La corrosion humide :

C'est la plus répondue, elle se manifeste dans le couple métal / fluide, exemple la dégradation du matériau organique et du béton.[18]

2.4 Les différents modes de corrosion :

Il existe plusieurs modes d'agressivité parmi lesquels on distingue :
— La corrosion chimique.
— La corrosion électrochimique.
— La corrosion bactérienne.
— La corrosion en présence d'une sollicitation mécanique.
Que nous allons voir avec plus de détails.

2.4.1 La corrosion chimique :

Elle se manifeste par une attaque directe du métal lorsqu'il est en contacte avec des solutions non électrolytiques ou avec des gaz secs. Au cours de cette corrosion l'oxydation du métal et la réduction de l'oxydant se produisent en un seul acte. L'action de l'oxygène reste l'exemple typique de la corrosion chimique.

2.4.2 La corrosion électrochimique :

Elle se produit lorsqu'il existe une hétérogénéité, dans le métal, ou dans l'électrolyte, l'existence de ces hétérogénéités favorise la formation d'une pile.
Un courant électrique circule entre l'anode et la cathode. Ce sont en général les anodes qui sont attaquées. Le métal s'oxyde en réduisant le milieu corrosif.
La différence de potentiel due a sa présence d'un couple de matériaux, au contact avec des milieux différents et a la présence de zones a concentrations différentes, est un facteur électrochimique important.
Par ailleurs, la part de la corrosion par effet électrochimique dans la plupart des cas ne eut provenir et se déclencher que par la présence du dioxyde de carbone "CO2", des conditions d'exploitation propices et des paramètres opératoires favorables, engendrant la formation de l'acide carbonique "H2CO3" ; allant même dans certain cas de figures favoriser la naissance de l'acide chlorhydrique.[11]

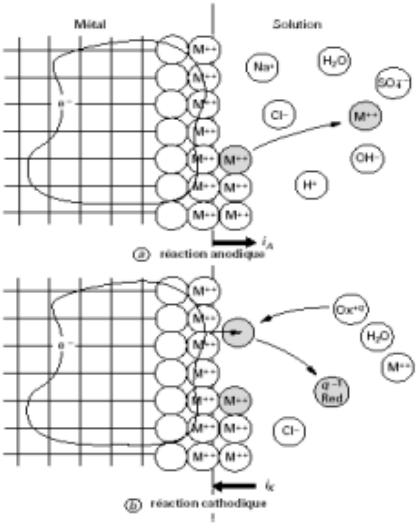

FIGURE 2.2 – Corrosion électrochimique

2.4.3 La corrosion bactérienne

La présence dans le milieu de certaines bactéries ; en liaison avec des substances organiques bien précises ; peut augmenter localement la vitesse de corrosion de l'acier en provoquant une dépolarisation accrue des sites cathodiques. Il s'agit principalement des bactéries réductrices de sulfates B.S.R.(Sulfaovibrio- desulfuricans) qui se développent quand les conditions physico-chimiques le permettent :

— Présence d'ions sulfates.
— Présence de matières organiques.
— Teneur en oxygène dissous négligeable.

La réaction de base du métabolisme de ces bactéries est la réduction de l'ion sulfate, catalysée par l'enzyme déshydrogénase SO.+8HS +4 H20
L'hydrogène peut être fourni par des matières organiques (alcools, protéines, amidon, hydrocarbures) ou bien, il peut s'agir de l'hydrogène produit par la corrosion de l'acier ; dans ce cas, la réaction globale est la suivante 4 Fe + 4 H.0 + SO. —Fes + 3 Fe (O) 2 +2OH On voit que la corrosion bactérienne de ce type conduit a la présence de sulfure de fer dans les produits de corrosion constitues en majeure partie de rouille formant des pustules.
Il est fréquent de rencontrer des souches de telles bactéries dans les eaux de gisement. Qui sont exemptes d'oxygène. Elles sont responsables de fréquentes piqures sur les tubes de productions dans des zones particulièrement favorables a leur prolifération.[18]

2.5 Les effets de la corrosion :

— Pollution de l'eau avec les produits de dégradation.
— Diminution de la résistance mécanique.
— Amincissement des parois.
— Perforations, fuites.
— Dégradation des équipements.

2.6 Morphologie de la corrosion :

Selon la nature de l'attaque, la corrosion peut présenter des aspects très divers regroupés en deux grandes familles :

— La corrosion uniforme (généralisée).
— La corrosion localisée.

2.6.1 La corrosion généralisée :

— Dissolution uniforme sur toute la surface du métal.
— Résultat d'une corrosion physique (érosion, cavitation..) ou chimique par réaction d'oxydoréduction.

— Se mesure en perte de masse / unité de surface / unité de temps ou par épaisseur / unité de temps.

FIGURE 2.3 – Corrosion Généralisée

2.6.2 Corrosion localisée :

La corrosion galvanique :

— Mise en contact de matériaux de potentiels électriques différents.
— Formation d'une pile de corrosion anode / cathode.
— Hétérogénéité de l'attaque sur le métal le moins noble.

La corrosion par piqures :

— Attaque très localisée : rupture du film passif en présence d'halogénures .
— Ou passivation incomplète.
— Perforation rapide après la phase d'amorçage et la propagation.

La corrosion caverneuse :

— Phénomène proche de la corrosion par piqures.
— Dans les zones confinées, faible volume, milieu stagnant, sous joints.
— Attaque liée à une modification locale de la composition du milieu.

La corrosion inter-granulaire :

— Localisée aux joints des grains - zone désordonnée / structure cristallo
— Liée aux opérations de soudage : précipitation de carbures de chrome qui réduit la résistance à la corrosion.

La corrosion erosion–corrosion cavitation :

- La corrosion érosion : écoulement turbulent, particules, jet qui érode le film protecteur ou le détruit, créant ainsi une pile de corrosion.
- La corrosion par cavitation : dégradation par implosion de bulles de cavitation (ondes de choc et fatigue)[18]

FIGURE 2.4 – Exemple 1 de l'eresion-corrosion cavitation

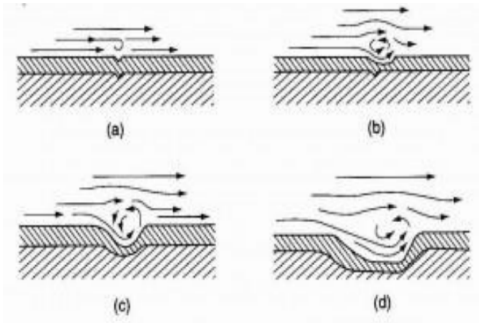

FIGURE 2.5 – Exemple 2 de l'eresion-corrosion cavitation

La corrosion sous contrainte :

- Contrainte mécanique + milieu agressif.
- Fissures inter ou transgranulaires perpendiculaires / contrainte principale.
- Contraintes résiduelles : écrouissage, laminage, cintrage.
- Contraintes thermiques : dilatation.
- Contraintes d'installation : suspension, soutènement, pression.[18]

2.7 Moyens de protection

Sachant que la corrosion avait pour origine l'eau produite rendue agressive par le gaz dissous, soit par les bactéries BSR. I existe deux moyens.

2.7.1 Inhibiteurs de corrosion

Se sont des substances a ajoutes a faible quantité dans le milieu, suppriment les inconvénients de ce milieu, sans le modifier.
D'une manière générale, Winhibiteur va former un film sur les surface métallique, suppriment ainsi le contact métal/eau. Toutefois, le filmage dépend du caractère ionique des produits utilises.

Nature et mode de fonctionnement des inhibiteurs du champ pétrolier

Les inhibiteurs utilises dans les puits de Hassi R'Mel sont des composes formant des films qui sont physiquement et chimiquement attires par la surface métallique pour former une barrière entre le métal et l'environnement. Les plus utilisées sont les AMINES, qui sont des chimiques contre la corrosion organique, ces derniers sont gras et insolubles dans l'eau, on peut les rendre solubles, en les simplifiant par HCL ou CH3COOH, ceci signifie en outre que la solubilité de ces produits peuvent varier suivant le choix de l'aminé ou de l'acide employé.[8]

2.7.2 Les biocides

Se sont des substances chimiques a ajouter a l'eau d'injection, ces dernières agissent directement sur les composantes de l'eau produite.[8]

2.8 Composition chimique des eaux :

La composition chimique des eaux peut varier en fonction de nombreux facteurs tels que l'origine géologique, la région géographique, les sources de pollution et les activités humaines. Cependant, certaines substances chimiques se retrouvent souvent dans les eaux naturelles. Voici une liste de quelques-unes des substances les plus courantes que l'on peut retrouver dans les eaux :

- Les ions minéraux tels que le calcium (Ca2+), le magnésium (Mg2+), le sodium (Na+), le potassium (K+), le chlorure (Cl-), le sulfate (SO42-) et le carbonate (CO32-).
- Les gaz dissous tels que l'oxygène (O2), le dioxyde de carbone (CO2) et le méthane (CH4).
- Les matières organiques telles que les acides humiques et fulviques, les sucres, les protéines et les lipides.
- Les éléments traces métalliques tels que le fer (Fe), le cuivre (Cu), le zinc (Zn), le plomb (Pb), le mercure (Hg), l'arsenic (As), le cadmium (Cd) et le nickel (Ni).
- Les composés chimiques d'origine anthropique tels que les pesticides, les herbicides, les hydrocarbures aromatiques polycycliques (HAP), les composés organochlorés (COCl) et les PCB (polychlorobiphényles).[15]

Il est important de noter que la présence de certaines substances chimiques dans les eaux peut avoir des effets néfastes sur la santé humaine et l'environnement, en particulier si elles dépassent certaines normes de qualité de l'eau. C'est pourquoi la surveillance de la qualité de l'eau est essentielle pour assurer la sécurité et la durabilité des ressources en eau.

2.8.1 Les eaux des sites pétroliers

A travers toute les ères, l'eau a été indiscutablement un élément indispensable, voire une source de vie pour les etre humains, d'où la nécessite de bien l'extraire, la transporter et l'exploiter.
Son importance dans l'économie ne cesse de croitre, son approvisionnement devient ainsi de plus en plus difficile.
Dans l'industrie pétrolière, l'eau est vitale pour la production et le traitement du pétrole, d'où l'utilisation de divers source d'eau tels que les eaux souterraines qu'on divisera en quatre types :

Eau de gisement

L'eau de gisement est une eau souterraine qui se trouve dans des formations géologiques profondes et qui est souvent associée à des gisements de pétrole ou de gaz naturel.
Cette eau peut être piégée dans les pores de la roche réservoir où se trouve le pétrole ou le gaz, ou elle peut être présente dans des couches géologiques adjacentes.
L'eau de gisement peut varier considérablement en termes de qualité et de composition chimique en fonction de la géologie locale. Elle peut être salée, chargée de minéraux tels que le calcium, le magnésium ou le fer, ou même contenir des traces de gaz naturel.
L'eau de gisement peut être extraite pendant la production de pétrole ou de gaz naturel, et peut être utilisée pour diverses applications, telles que l'injection d'eau pour augmenter la récupération de pétrole ou de gaz, ou pour d'autres usages industriels. Cependant, il est important de surveiller attentivement la qualité de l'eau de gisement, car elle peut contenir des contaminants tels que des métaux lourds, des produits chimiques ou des micro-organismes qui peuvent nuire à la santé humaine ou à l'environnement.
Dans certaines régions, l'eau de gisement peut être la seule source d'eau douce disponible, mais elle doit être traitée avant d'être utilisée pour la consommation humaine ou l'irrigation, car elle peut contenir des contaminants potentiellement dangereux.
Il existe différents types d'eaux de gisements tels que :

Eaux de condensation
Elles correspondent la fraction d'eau en phase vapeur accompagnant les fluides de gisement.
Elles sont théoriquement moins chargées en éléments chimiques.
Elles sont produites a la tête des puits par condensation en quantité relativement faible.
Eaux de formation
Elles sont communément attribuées à l'aquifère du gisement et accompagnent la mise en place des hydrocarbures.
Elles sont variées et classées suivant les éléments chimiques dominants quelles renferment.
Eaux interstitielles
Ce sont des eaux que l'on retrouve dans les petits espaces entre les minuscules grains d'une roche.

Elles sont de deux types :

— Cynégétique : formée au même moment que la roche mère.
— EPI génétique : générée par des infiltrations dans la roche.

Eaux connées

Le mot "connée" veut dire née, produite ou générée ensemble.
Une eau cannée peut être considérée comme une eau interstitielle d'origine cynégétique.
Elle est donc une eau fossile qui est restée sans contact avec l'atmosphère durant une grande partie d'une période géologique. Le fait de dire qu'une eau est née avec la formation de la roche mère.[15]

Eau Albien

L'eau Albien est une eau fossile qui s'est formée pendant la période géologique de l'Albien, il y a environ 100 millions d'années. Elle a été emprisonnée dans des formations géologiques profondes depuis cette époque, sans contact avec l'atmosphère actuelle.
L'eau Albien peut être située dans des réservoirs souterrains profonds et isolés qui peuvent contenir des quantités importantes d'eau douce. Cependant, comme pour l'eau Barrémien, elle peut également contenir des minéraux ou des composés chimiques potentiellement dangereux qui peuvent affecter la qualité de l'eau.
L'exploitation de l'eau Albien peut être difficile car elle peut être piégée sous des couches de roches dures et difficiles à pénétrer.Cependant, avec les avancées technologiques, il est devenu possible d'extraire l'eau Albien à des profondeurs de plus en plus grandes.
Il est important de prendre des mesures pour protéger la qualité de l'eau Albien si elle est exploitée, afin de minimiser l'impact sur les écosystèmes locaux et de préserver cette ressource pour les générations futures.

Eau miopliocene

L'eau miopliocène est une eau fossile qui s'est formée au cours de la période géologique du Miocène, il y a environ 5 à 23 millions d'années.
Cette eau a été emprisonnée dans des formations géologiques et a été préservée depuis lors, sans contact avec l'atmosphère moderne.
L'eau miopliocène peut être riche en minéraux et peut avoir des caractéristiques chimiques uniques en raison des conditions géologiques dans lesquelles elle s'est formée et a été stockée.
Cette eau peut être utilisée pour des applications industrielles, telles que l'irrigation, ou pour la consommation humaine, mais elle nécessite généralement un traitement pour enlever les contaminants potentiels et pour atteindre les normes de qualité de l'eau potable.
L'eau miopliocène est souvent considérée comme une ressource précieuse en raison de son ca-

ractère rare et de sa qualité supposée, mais son exploitation doit être effectuée de manière responsable pour éviter toute perturbation des écosystèmes locaux ou des formations géologiques environnantes.

Eau d'injection

C'est le procédé le plus ancien (fin XIX siècle), et encore le plus employé. Son but est d'augmenter la récupération, mais aussi d'accélérer la production, ou plus précisément de diminuer son déclin. Le moyen utilisée est souvent un maintien de pression.
L'injection peut être soit du type reparti dans la zone a l'huile, soit du type périphérique dans un aquifere existant.
Avec une injection d'eau, le rapport de mobilité M est souvent favorable pour une huile légère (viscosité de l'huile faible) et pas trop défavorable pour une huile plus lourde.
L'efficacité c'est-a-dire la récupération, sera donc élevée ou moyenne. Quant aux sources en eau, i s'agit le plus souvent de couches aquiferes situées a faible profondeur, de l'eau de nier ou de l'eau en surface a terre (lacs, rivières).
L'injection d'eau est favorable pour le gisement hétérogène dont la roche est mouillable a l'eau, ce qui est souvent le cas, sauf pour certains réservoirs carbonate.
Par ailleurs, il faut que l'eau soit injectable, perméabilité, suffisante et compatibilité avec l'eau du gisement. En effet, le mélange d'eau injectée avec l'eau en place peut provoquer des précipites inscrutables (BasO4) qui bouche les puits.[2]

2.9 La relation entre la composition chimique des eaux et la corrosion

La composition chimique des eaux peut également avoir une incidence sur la corrosion des métaux et des alliages.
Les eaux naturelles peuvent contenir des ions tels que le calcium, le magnésium, le fer, le cuivre, le zinc, le chlorure, le sulfate et le carbonate, qui peuvent affecter la corrosion des métaux. Par exemple, les ions chlorure et sulfate peuvent augmenter la corrosion des métaux ferreux, tandis que les ions carbonate peuvent réduire la corrosion.
En outre, la présence de polluants chimiques dans les eaux, tels que les acides, les produits chimiques industriels, les métaux lourds, les sels et les matières organiques, peut également affecter la corrosion des métaux. Certains polluants chimiques peuvent accélérer la corrosion, tandis que d'autres peuvent la ralentir.
Il est important de comprendre la composition chimique de l'eau dans les systèmes où la corrosion est un problème potentiel, afin de sélectionner les matériaux appropriés pour chaque application et de prendre des mesures préventives pour minimiser les effets de la corrosion,telles que l'utilisation de traitements inhibiteurs de corrosion et le contrôle des niveaux de pH et de la teneur en oxygène dans l'eau.

2.10 Problématique

Pour faire face à la corrosion dans les canalisations, SONATRACH est confrontée à une contrainte majeure : l'impossibilité d'installer des capteurs électroniques en raison de la présence de produits inflammables dans les canalisations. Cependant, afin d'éviter les dommages matériels et de garantir la sécurité des installations, il est impératif que la société ne puisse en aucun cas interrompre la production de ses sites pétroliers.

Afin de répondre à ces défis, nous avons développé un modèle mathématique capable de prédire la vitesse de corrosion en fonction des paramètres chimiques des eaux utilisées (eaux d'injection) ou présents dans les puits de production. Les caractéristiques des eaux peuvent varier d'une région à une autre, rendant difficile l'obtention d'une vitesse de corrosion précise. C'est pourquoi notre objectif est de trouver une équation permettant d'estimer cette vitesse de manière approximative. Ceci permettra à SONATRACH d'anticiper les problèmes de corrosion, de localiser les zones à risque et de prendre les mesures nécessaires pour assurer la maintenance de ses sites en conséquence.

En mettant en place ce modèle prédictif, nous visons à prévenir les problèmes de piquage dans les pipelines, offrant ainsi à SONATRACH une meilleure connaissance des moments et des lieux où ces incidents pourraient se produire. Cette approche permettra à l'entreprise de prendre des mesures préventives correctement et d'assurer la continuité de ses activités tout en garantissant la sécurité de ses installations.

2.11 Conclusion

Ce deuxième chapitre porte sur différents aspects de la corrosion, en mettant l'accent sur la corrosion généralisée, la corrosion galvanique, la corrosion par piqûres.....ect.
Le chapitre aborde également les moyens de protection contre la corrosion, tels que les inhibiteurs de corrosion et les biocides. Il souligne l'importance de la composition chimique des eaux, en particulier dans les sites pétroliers, et présente différents types d'eaux souterraines, tels que l'eau de gisement, l'eau de condensation, l'eau de formation, l'eau interstitielle, l'eau connée, l'eau Albien et l'eau Miopliocene.

Chapitre 3

Régression linéaire

Introduction

La régression linéaire est une technique statistique couramment utilisée pour modéliser la relation entre une variable dépendante continue et une ou plusieurs variables indépendantes. On distingue la régression simple, lorsqu'on s'intéresse à la relation entre deux variables, et la régression multiple, lorsque la relation porte entre une variable et plusieurs autres variables. Cette technique consiste à trouver la meilleure relation linéaire entre les variables indépendantes qui sont les paramètres chimiques dans notre cas et la vitesse de corrosion qui est la variable dépendante , en minimisant l'écart entre les valeurs réelles de la variable dépendante et les valeurs prédites par le modèle.

3.1 Moindres carrés ordinaires

La méthode des moindres carrés ordinaires (MCO) est une technique courante pour la régression linéaire, qui permet de trouver les coefficients du modèle qui minimisent la somme des carrés des résidus entre les valeurs prédites et les valeurs réelles. Cependant, cette méthode suppose que les erreurs de la régression ont une variance constante, également appelée homoscédasticité. Dans les cas où l'hétéroscédasticité est présente, c'est-à-dire lorsque la variance des erreurs n'est pas constante, la méthode des moindres carrés ordinaires peut être insuffisante pour modéliser les données avec précision.

3.1.1 Définition 1

Une fonction affine est une fonction qui s'écrit sous la forme $y = ax + b$, où a, b sont des constantes.

3.1.2 Définition 2

Les points (x_i, y_i) étant donnés, le but est maintenant de trouver une fonction affine f telle que la quantité $\sum_{i=1}^{n} L(y_i - f(x_i))$ soit minimale. Pour pouvoir déterminer f, encore faut-il préciser la fonction de coût L. Deux fonctions sont classiquement utilisées :

— u représente la différence entre la valeur réelle de la variable dépendante y_i et la valeur prédite par la fonction affine pour un point (x_i, y_i)
— Le coût absolu $L(u) = |u|$;
— Le coût quadratique $L(u) = u^2$.

Les deux ont leurs vertus, mais on privilégiera dans la suite la fonction de coût quadratique. On parle alors de méthode d'estimation par moindres carrés (terminologie due à Legendre dans un article de 1805 sur la détermination des orbites des comètes).

3.2 Modélisation

Dans de nombreuses situations, en première approche, une idée naturelle est de supposer que la variable à expliquer y est une fonction affine de la variable explicative x, c'est-à-dire de chercher f dans l'ensemble F des fonctions affines de $\mathbb{R}$ dans $\mathbb{R}$. C'est le principe de la régression linéaire simple.

Définition 1 (Modèle de régression linéaire simple)

Un modèle de régression linéaire simple est défini par une équation de la forme :

$$\forall i \in \{1, \ldots, n\}, \quad y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, ...(1)$$

Les quantités ε_i viennent du fait que les points ne sont jamais parfaitement alignés sur une droite. On les appelle les erreurs (ou bruits) et elles sont supposées aléatoires. Pour pouvoir dire des choses pertinentes sur ce modèle, il faut néanmoins imposer des hypothèses les concernant. Voici celles que nous ferons dans un premier temps :

— (H1) : $\mathbb{E}[\varepsilon_i] = 0$ pour tout indice i
— (H2) : $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$ pour tout couple (i, j)

Les erreurs sont donc supposées centrées, de même variance (homoscédasticité) et non corrélées entre elles (δ_{ij} est le symbole de Kronecker, c'est-à-dire $\delta_{ij} = 1$ si $i = j$, $\delta_{ij} = 0$ si $i \neq j$). Notons que le modèle de régression linéaire simple de la définition 1 peut encore s'écrire de façon vectorielle :

$$Y = \beta_0 + \beta_1 X + \varepsilon, ...(2)$$

où :

— le vecteur $Y = [y_1, \ldots, y_n]'$ est aléatoire de dimension n,
— le vecteur $X = [x_1, \ldots, x_n]$ est un vecteur de dimension n donné (non aléatoire),
— les coefficients β_1 et β_2 sont les paramètres du modèle,
— le vecteur $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]'$ est aléatoire de dimension n. [4]

3.3 Présentation du modèle de régression linéaire multiple

On peut présenter le modèle de la régression linéaire multiple sous la forme suivante :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i .....(3) \quad i = 1, 2, ..., n$$

[3]

ou sous la forme matricielle :

$$Y = X\beta + \varepsilon$$

$$\text{Avec } Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Y : désigne le vecteur à expliquer (exogène, dépendante) de taille n,

X : la matrice explicative (endogènes, indépendantes) de taille $n \times (p+1)$,

ε : le vecteur de variable que représente le terme d'erreur de taille n.

β : le vecteur des paramètres.

Les hypothèses du modèle :

1. H_1 : Y_i sont des variable aléatoire $\forall i = 1, \cdots, n$
2. H_2 : X_i sont des variable non aléatoire.
3. H_3 : $Var(\varepsilon_i) = \sigma^2$, $\forall i = 1, \cdots, n$
4. H_4 : $Cov(\varepsilon_i, \varepsilon_j) = 0. \Leftrightarrow E(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j, \quad \forall i, j = 1, \cdots, n$
5. H_5 : ε_i sont des termes d'erreur, non observés, indépendants et identiquement distribués de loi $\mathcal{N}(0, \sigma^2)$.
6. H_6 : $Cov(x_{ji}, \varepsilon_i) = 0 \quad \forall j = 1, \cdots, p \quad \forall i = 1, \cdots, n$
   signifie qu'il n'y a pas de corrélation linéaire entre la variable explicative x_{ji} et l'erreur ε_i
   $\Rightarrow (X^T X)$ régulière et $(X^T X)^{-1}$ existe.
7. H_7 : le nombre d'observation et supérieur aux nombre de variable explicatives $(n > p)$
   [3]

3.4 Les estimateurs des moindres carre ordinaire

3.4.1 Estimation des coefficients de régression

Soit $\hat{\varepsilon} = (\hat{\varepsilon_1}; \ldots; \hat{\varepsilon_n})^t$ le vecteur de dimension $(n \times 1)$ des résidus défini par

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta}$$

où $\hat{Y} = X\hat{\beta}$ représente les valeurs estimées par le modèle, on les appelle aussi valeurs ajustées.

La somme des carrés des résidus est donnée par :

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \left(Y - X\hat{\beta}\right)^T (Y - X\hat{\beta})$$

Pour estimer les paramètres $\beta_0, \ldots, \beta_p$ du modèle linéaire, on utilise la méthode des moindres carrés ordinaires (MCO), la méthode des moindres carrés est une notion mathématique permettant d'apporter à un nombre d'éléments susceptibles de comporter des erreurs un ajustement afin d'obtenir des données proches de la réalité. cette méthode consiste Donc à trouver les valeurs de $\beta_0, \ldots, \beta_p$ qui minimisent la somme des carrés des résidus :

$$Argmin \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

On cherche a trouver le vecteur $\hat{\beta}$ qui minimise la quantité suivante :

$$\begin{aligned} |\hat{\varepsilon}|^2 &= (Y - X\beta)^T (Y - X\beta) \\ &= Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta \\ &= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta \end{aligned}$$

Car $(\beta^T X^T Y)$ et $(Y^T X\beta)$ est un scalaire. Donc il est égal à sa transposé.

La dérivée de $|\hat{\varepsilon}|^2$ par rapport à β est alors :

$$-2X^T Y + 2X^T X\beta$$

Nous cherchons $\hat{\beta}$ qui annule cette dérivée. Donc nous devons résoudre l'équation suivante :

$$X^T X\hat{\beta} = X^T Y$$

Nous trouvons après avoir inversé la matrice $X^T X$ (il faut naturellement vérifier que $X^T X$ est carrée et inversible)

$$\hat{\beta} = (X^T X)^{-1} X^T Y \ldots\ldots (4)$$

[3]

3.4.2 Propriétés des estimateurs MCO

Théorème L'estimateur $\hat{\beta}$ des moindres carrés ordinaire est sans biais, c.à.d $E(\hat{\beta}) = \beta$, et sa matrice de variance covariance, notée par varcov($\hat{\beta}$) ou par $S^2_(\hat{\beta})$, est :

$$\text{Varcov}(\hat{\beta}) = \sigma^2(X^TX)^{-1} \text{.....(5)}$$

Remarque La matrice de variance-covariance (varcov($\hat{\beta}$)) des coefficients de dimension $(p+1; p+1)$ est donnée par :

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \text{Var}\left(\hat{\beta}_0\right) & \text{Cov}\left(\hat{\beta}_0,\, \hat{\beta}_1\right) & \cdots & \text{Cov}\left(\hat{\beta}_0,\, \hat{\beta}_p\right) \\ \cdots & \text{Var}\left(\hat{\beta}_1\right) & \cdots & \text{Cov}\left(\hat{\beta}_1,\, \hat{\beta}_p\right) \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \text{Var}\left(\hat{\beta}_p\right) \end{bmatrix}$$

Cette matrice est symétrique, sur sa diagonale principale on observe les variances des coefficients estimés (var($\hat{\beta}_0$), ..., var($\hat{\beta}_p$)). [3]

Preuve :

Montrer que $\mathbb{E}(\hat{\beta}) = \beta$

$$\begin{aligned} \hat{\beta} &= (X^TX)^{-1}X^TY \\ &= (X^TX)^{-1}X^T[X\beta + \varepsilon] \\ &= (X^TX)^{-1}(X^TX)\beta + (X^TX)^{-1}X^T\varepsilon \\ &= \beta + (X^TX)^{-1}X^T\varepsilon \end{aligned}$$

Alors, l'espérance mathématique de $\hat{\beta}$ est :

$$\begin{aligned} E(\hat{\beta}) &= E[\beta + (X^TX)^{-1}X^T\varepsilon] \\ &= \beta + E[(X^TX)^{-1}X^T\varepsilon] \\ &= \beta + (X^TX)^{-1}X^TE[\varepsilon] \quad ;\text{car } X \text{ est non aléatoire.} \end{aligned}$$

Et sous l'hypothèse que $E(\varepsilon) = 0$ il vient que :

$$E(\hat{\beta}) = \beta$$

Montrer que $\text{Varcov}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

On procède de même, on a $\hat{\beta} = \beta + (X^T X)^{-1} X^T \varepsilon$ donc :

$$\begin{aligned}
\text{Varcov}(\hat{\beta}) &= \text{Var}\left(\beta + (X^T X)^{-1} X^T \varepsilon\right) \\
&= (X^T X)^{-1} X^T \,\text{Var}(\varepsilon) X (X^T X)^{-1} \\
&= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$

On trouve :

$$\text{Varcov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Pour estimer cette matrice de variance-covariance de $\hat{\beta}$,Il suffit de remplacer la variance théorique des résidus σ^2 par son estimateur.

3.4.3 Estimation de la variance du résidu σ^2

Lemme 1.2.3 Soit un vecteur Z composé de n variables aléatoires d'espérances nulles, et tel que $\text{Var}(Z) = \sigma_z^2 I_n$, et A une matrice symétrique non aléatoire, alors

$$\mathbb{E}[Z^T A Z] = \sigma_z^2 tr(A)$$

où $tr(A)$: est la trace de la matrice A.

Théorème 1.2.3 soit $\hat{\varepsilon} = Y - X\hat{\beta}$, alors

$$\mathbb{E}[\hat{\varepsilon}^t \varepsilon] = (n - p - 1)\sigma^2$$

Preuve : Soit

$$\begin{aligned}
\hat{Y} &= X\beta \\
&= \left(X(X^T X)^{-1} X^T\right) Y \\
&= HY; \text{ avec } H = X(X^T X)^{-1} X^T
\end{aligned}$$

Le vecteurs des résidus, noté par $\hat{\varepsilon}$, est donné par la relation :

$$\hat{\varepsilon} = Y - \hat{Y} = Y - HY = (\mathbb{I}_n - H)Y$$

où $\mathbb{I}_n$ est la matrice unité de dimension (n, n).

La matrice H est appelée la matrice chapeau de taille(n, n). Elle vérifie les deux propriétés suivantes :

$$\text{Symétrique :} \quad H^t = H \tag{3.1}$$
$$\text{Idémpotente :} \quad H^2 = H \tag{3.2}$$

La matrice$(\mathbb{I}_n - H)$ a les mêmes propriétés :

1. $(\mathbb{I}_n - H)^t = (\mathbb{I}_n - H)$
2. $(\mathbb{I}_n - H)^2 = (\mathbb{I}_n - H)$

Donc

$$\begin{aligned}\hat{\varepsilon} = Y - \hat{Y} = Y - HY &= (\mathbb{I}_n - H)Y \\ &= (\mathbb{I}_n - H)(X\beta + \varepsilon) \\ &= X\beta - HX\beta + \varepsilon - H\varepsilon\end{aligned}$$

où $HX = (X(X^TX)^{-1}X^T)X = X$

Ce qui donne :

$$\hat{\varepsilon} = (\mathbb{I}_n - H)\varepsilon$$

On obtient :

$$\begin{aligned}[\hat{\varepsilon}^t\varepsilon] &= ((\mathbb{I}_n - H)\varepsilon)^t(\mathbb{I}_n - H)\varepsilon \\ &= \varepsilon^t(\mathbb{I}_n - H)^t(\mathbb{I}_n - H)\varepsilon \\ &= \varepsilon^t(\mathbb{I}_n - H)\varepsilon \qquad \text{(d'après la propriété (3.2)} \\ &= \varepsilon^t\mathbb{I}_n\varepsilon - \varepsilon^tH\varepsilon\end{aligned}$$

Donc, d'après le **lemme 1.2.3**, on obtient :

$$\begin{aligned}\mathbb{E}[\varepsilon\varepsilon^t] &= \mathbb{E}[\varepsilon^t\mathbb{I}_n\varepsilon - \varepsilon^tH\varepsilon] \\ &= \mathbb{E}[\varepsilon^t\mathbb{I}_n\varepsilon] - \mathbb{E}[\varepsilon^tH\varepsilon] \\ &= \sigma^2\mathrm{tr}(\mathbb{I}_n) - \sigma^2\mathrm{tr}(H) \\ &= \sigma^2[\mathrm{tr}(\mathbb{I}_n), \mathrm{tr}(H)]\end{aligned}$$

Où $tr(\mathbb{I}_n) = n$ et :

$$\begin{aligned} tr(H) &= tr(X(X^TX)^{-1}X^T) \\ &= tr(X^TX(X^TX)^{-1}), \text{ puisque } tr(AB) = tr(BA) \\ &= tr(\mathbb{I}_{p+1}), \text{car } X^TX \text{ est une matrice carrée de taille } (p+1) \\ &= p+1 \end{aligned}$$

Donc

$$\mathbb{E}[\hat{\varepsilon}^t\varepsilon] = (n-p-1)\sigma^2$$

D'après **le théorème 1.2.3**, on peut construire l'estimateur sans biais pour σ^2 qui est :

$$\begin{aligned} S^2 &= \frac{\varepsilon^t\varepsilon}{n-p-1} \\ &= \frac{\sum_{i=1}^n \hat{\varepsilon}i^2}{n-p-1} \end{aligned}$$

[3]

3.5 Lois des estimateurs et intervalles de confiance

Après avoir obtenu l'estimateur, son espérance, et une estimation de sa variance, il ne reste plus qu'à calculer sa loi de distribution pour construire des intervalles de confiance ou des tests d'hypothèses sur β.

Proposition 1.3.1 (Lois des estimateurs)

$\hat{\beta}$ est le vecteur de estimateur par la méthode de MCO.

1. $\hat{\beta}$ est un vecteur gaussien de moyenne β et de variance $\sigma^2(X^TX)^{-1}$

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$$

.

2. $\hat{\beta}$ et S^2 sont indépendants.
3. $$(n-p-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

Proposition 1.3.2 (Intervalle de confiance)

1. Pour tout $j = 0, \cdots, p$ un intervalle de confiance de niveau $(1-\alpha)$ pour β_j est :

$$\left[\hat{\beta}_j - t_{1-\frac{\alpha}{2},(n-p-1)} s_{\hat{\beta}_j}; \hat{\beta}_j + t_{1-\frac{\alpha}{2},(n-p-1)} s_{\hat{\beta}_j}\right]$$

où
— $t_{\frac{\alpha}{2},(n-p+1)}$ est le quantile de niveau $(1-\frac{\alpha}{2})$ d'une loi de Student à $n-p-1$ degrés de liberté.
— $\hat{\beta}_j$ est l'estimation de paramètre β.
— $S_{\hat{\beta}_j}$ est l'écart type estimé pour l'estimateur de β_j.

2. Un intervalle de confiance de niveau $(1-\alpha)$ pour σ^2 est :

$$\left[\frac{(n-p-1)s^2_{\hat{\varepsilon}}}{c_{(1-\frac{\alpha}{2}),n-p-1}}; \frac{(n-p-1)s^2_{\hat{\varepsilon}}}{c_{(\frac{\alpha}{2}),n-p-1}}\right]$$

où
— $c_(1-\frac{\alpha}{2}), n-p-1$: est le quantile d'ordre $(1-\frac{\alpha}{2})$ d'une loi du (khi deux) à $n-p-1$ degrés de liberté.
— $c_(\frac{\alpha}{2}), n-p-1$: est le quantile d'ordre $(\frac{\alpha}{2})$ d'une loi du (khi deux) à $n-p-1$ degrés de liberté. [3]

3.6 Tests statistiques

3.6.1 P-valeur

On considère des hypothèses de la forme :

H_0 : A contre H_1 : contraire de A.

La p-valeur est le plus petit réel $\alpha \in]0,1[$ calculé à partir des données tel que l'on puisse se permettre de rejeter H_0 au risque $100\alpha\%$. Autrement écrit, la p-valeur est une estimation ponctuelle de la probabilité critique de se tromper en rejetant H_0 ou en affirmant H_1 alors que H_0 est vraie.
Les degrés de significativité sont les suivants :
Le rejet de H_0 sera :

— "significatif" si la p-valeur $\in]0.01, 0.05]$, symbolisé par $*$,
— "très significatif" si la p-valeur $\in]0.001, 0.01]$, symbolisé par $**$,
— "hautement significatif" si la p-valeur < 0.001, symbolisé par $***$,
— "presque significatif" si la p-valeur $\in]0.05, 0.1]$, symbolisé par . (un point).[6]

3.6.2 Test de Student

Soit $j \in \{0, \ldots, p\}$. Le test de Student permet d'évaluer l'influence de X_j sur Y.

On considère les hypothèses :

$H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.

On calcule la réalisation t_{obs} de
$T^* = \frac{\beta_{bj}}{\sigma_b(\beta_{bj})}$

On considère une variable $T \sim T(\nu)$.
Alors la p-valeur associée est

$$\text{p-valeur} = P(|T| \geq |t_{\text{obs}}|).$$

Si :
— $*$, l'influence de X_j sur Y est "significative",
— $**$, l'influence de X_j sur Y est "très significative",
— $***$, l'influence de X_j sur Y est "hautement significative". [6]

3.6.3 Test global de Fisher

L'objectif du test global de Fisher est d'étudier la pertinence du lien linéaire entre Y et $X_1, \ldots, X_p$.

On considère les hypothèses :

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{contre} \quad H_1 : \text{il y a au moins un coefficient non nul.}$$

On calcule la réalisation f_{obs} de

$$F^* = \frac{R_b^2}{1 - R_b^2} \frac{n - (p+1)}{p}.$$

On considère une variable $F \sim F(p, \nu)$. Alors la p-valeur associée est

$$\text{p-valeur} = P(F \geq f_{\text{obs}}).$$

Notons que ce test est moins précis que le test de Student car il ne précise pas quels sont les coefficients non nuls. Il est toutefois un indicateur utile pour déceler d'éventuelles problèmes (comme des colinéarités entre $X_1, \ldots, X_p$).[6]

3.7 Analyse de la variance et coefficient de détermination

3.7.1 Décomposition de la variance et tableau d'ANOVA

On peut aisément vérifier que, l'écart entre y_i et la moyenne des y_i en ajoutant puis retranchant $\hat{y}$ la valeur estimée de y par la droite de régression. Cette procédure fait apparaître une somme de deux écarts :

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

On peut obtenir la décomposition :

$$\begin{aligned}\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SCT} &= \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SCE} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SCR} \\ &= \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SCE} + \underbrace{\sum_{i=1}^{n}(\hat{\varepsilon}_i)^2}_{SCR}\end{aligned}$$

Où

— SCT : désigne la somme des carrés totaux (centrés).
— SCE : la somme des carrés expliqués (centrés).
— SCR : la somme des carrés des résidus.

Le tableau "d'analyse de la la variance" se présenté sous la forme suivante :

Source de variation	ddl	Somme des carrés	Carrés moyens
Expliquée	p	SCE	$CME = \frac{SCE}{P}$
Résiduelle	$n-p-1$	SCR	$CMR = \frac{SCR}{n-p-1}$
Totale	$n-1$	SCT	

TABLE 3.1 – analyse de la variance de la régression linéaire multiple

Remarque On peut réécrire l'équation d'ANOVA matriciellement comme suit :

$$\underbrace{(y - \bar{y})^t(\hat{y} - \bar{y})}_{SCT} = \underbrace{(\hat{y} - \bar{y})^t(\hat{y} - \bar{y})}_{SCE} + \underbrace{\hat{\varepsilon}^t\hat{\varepsilon}}_{SCR}$$

où $\bar{y}$ est le vecteur de $\mathbb{R}^n$ contenant n fois la moyenne de la variable y, c'est à dire

$$y = (\overline{y}; \cdots; \overline{y})^t$$

3.7.2 Coeffcient de détermination R^2

Le rapport entre SCE et SCT représente la proportion de variance expliquée et porte le nom de coeffcient de détermination, noté par R^2 :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$
$$1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SCR}{SCT}$$

Ce coefficient est compris entre 0 et 1 : plus il est proche de 1 est plus grande la part expliquée, autrement dit meilleure est la régression.

Inversement, un coefficient R^2 proche de 0 indique que la quantité SCR est élevée. [3]

3.8 La méthode des moindres carrés généralisés (MCG)

Dans ces situations, la méthode des moindres carrés généralisée (MCG) peut être utilisée pour améliorer la qualité de la modélisation.

La méthode des moindres carrés généralisée consiste à ajouter des termes de pénalité supplémentaires à la fonction de coût des moindres carrés ordinaires. Elle permet de prendre en compte la variance variable des erreurs en utilisant une fonction de variance qui lie la variance des erreurs à la valeur prédite.

Ainsi, la méthode des moindres carrés généralisée permet de mieux modéliser les données dans les situations où l'hétéroscédasticité est présente, contrairement à la méthode des moindres carrés ordinaires qui ne prend pas en compte la variance variable des erreurs.

La méthode des moindres carrés généralisés (MCG) permet de prendre en compte la variance variable des résidus contrairement aux moindres carrés ordinaires.[7]

3.9 Conclusion

La conclusion du ce chapitre porte sur les principaux points abordés dans le contexte du modèle de régression linéaire multiple. Ces points comprennent l'estimation des paramètres, la matrice de variance-covariance, l'estimation de la variance du résidu, la loi des estimateurs, les intervalles de confiance, les tests statistiques et la décomposition de la variance. En résumé, ce chapitre met en évidence les méthodes et les outils permettant d'estimer les coefficients de régression, d'évaluer leur significativité, d'estimer l'incertitude associée aux estimations et d'évaluer la proportion de variabilité expliquée par le modèle de régression linéaire multiple.

Chapitre 4

Résolution du problème

Introduction

Dans ce chapitre nous commençons par présenter le langage de programmation R que nous utiliserons pour faire une régression linéaire multiple à fin de résoudre le problème, à partir des données provenant de l'organisme d'accueil sous forme d'un fichier Excel, ainsi que tout les packages nécessaire pour mettre à bien notre travail.
Cette régression permet d'obtenir les coefficients de l'équation et aussi le nuage de points associé, représenter par un graphe et une corrélation.

4.1 Outils de Programmation

4.1.1 Le langage R

Reste un langage de programmation et un environnement logiciel utilisé principalement pour l'analyse statistique, la visualisation de données et le développement d'algorithmes. Il est largement utilisé dans les domaines de la science des données, de la recherche académique et de l'industrie.

Voici quelques caractéristiques principales du langage R :

- **Grande communauté et bibliothèques riches :** R bénéficie d'une vaste communauté d'utilisateurs qui contribue au développement de nombreuses bibliothèques et packages. Ces derniers offrent des fonctionnalités supplémentaires pour des tâches spécifiques, ce qui permet aux utilisateurs de bénéficier d'une large gamme d'outils statistiques et graphiques.
- **Analyse statistique avancée :** R est particulièrement bien adapté à l'analyse statistique. Il propose une grande variété de techniques statistiques, telles que les tests d'hypothèses, les régressions, les analyses de variance, les modèles linéaires généralisés, les méthodes non paramétriques, etc. Ces fonctionnalités font de R un choix privilégié pour les statisticiens et les scientifiques des données.
- **Visualisation de données :** R offre de puissantes capacités de visualisation de données. Il dispose de packages tels que ggplot2, lattice et plotly, qui permettent de créer des graphiques hautement personnalisables et esthétiques. Ces outils permettent de représenter graphiquement les données, de créer des graphiques interactifs et de produire des visualisations de qualité professionnelle.
- **Programmation fonctionnelle :** repose sur le paradigme de programmation fonctionnelle, ce qui signifie qu'il offre des fonctionnalités pour gérer les fonctions comme des objets de première classe. Cela permet une programmation plus expressive et flexible, notamment pour la création de fonctions personnalisées et la manipulation de données.
- **Intégration avec d'autres langages :** R dispose de fonctionnalités d'intégration avec d'autres langages de programmation tels que C, C++, Python et SQL. Cela permet aux utilisateurs de combiner les avantages de différentes langues pour des tâches

spécifiques, d'accéder à des bases de données externes et d'utiliser des bibliothèques externes.
— **Environnement interactif :** R est souvent utilisé avec RStudio, un environnement de développement intégré (IDE) populaire spécialement conçu pour R. RStudio offre une interface conviviale pour écrire, exécuter et déboguer du code R, ainsi que pour gérer les projets et les packages.
— **Gratuit et open source :** R est un logiciel libre et gratuit, ce qui signifie que vous pouvez l'utiliser, le distribuer et le modifier librement. De plus, la nature open source de R encourage la collaboration et le partage de connaissances au sein de la communauté.

4.2 Programmation et application de la régression linéaire multiple

En utilisant le langage de programmation R nous obtenons le code suivant :

```
#Commande pour lire les données depuis un fichier excel.csv
dds = read.csv("C:/Users/User/Downloads/ddsls.csv")

#commande pour faire la régression linéaire multiple
model1 = lm(Y~X1+ X2+ X3+ X4+ X5+ X6+ X7+ X8+ X9+ X10+ X11+ X12+ X13,dds)
summary(model1)
```

FIGURE 4.1 – Code de la régression linéaire sous langage R

Le code représente une régression linéaire multiple à l'aide de la fonction **"lm"** dans R. Nous avons spécifié un modèle linéaire avec la variable dépendante **Y** et treize variables indépendantes **X1** à **X13**. La régression linéaire multiple vise à déterminer la relation entre la variable dépendante Yet un ensemble de variables indépendantes, tel que :

— **Y :** Représente l'ensemble des vitesses de corrosion.
— **X1 :** Représente la valeur du **PH a 20°c**.
— **X2 :** Représente la quantité de **Ca++** présente dans l'échantillon.
— **X3 :** Représente la quantité de **Mg++** présente dans l'échantillon.
— **X4 :** Représente la quantité de **Na+** présente dans l'échantillon.
— **X5 :** Représente la quantité de **K+** présente dans l'échantillon.
— **X6 :** Représente la quantité de **Ba++** présente dans l'échantillon.
— **X7 :** Représente la quantité de **Fe++** présente dans l'échantillon.
— **X8 :** Représente la quantité de **Sr++** présente dans l'échantillon.
— **X9 :** Représente la quantité de **Cl-** présente dans l'échantillon.
— **X10 :** Représente la quantité de **Co3**– présente dans l'échantillon.
— **X11 :** Représente la quantité de **Hco3-** présente dans l'échantillon.
— **X12 :** Représente la quantité de **So4**– présente dans l'échantillon.

— **X13 :** Représente la quantité de **Extrait sec mg/l** présente dans l'échantillon.

Une fois que nous avons ajusté le modèle à nos données avec **lm**, on utilise **summary(model1)**pour obtenir un résumé des résultats de la régression linéaire.
Se qui nous permet d'obtenir le résultat dans la figure suivante :

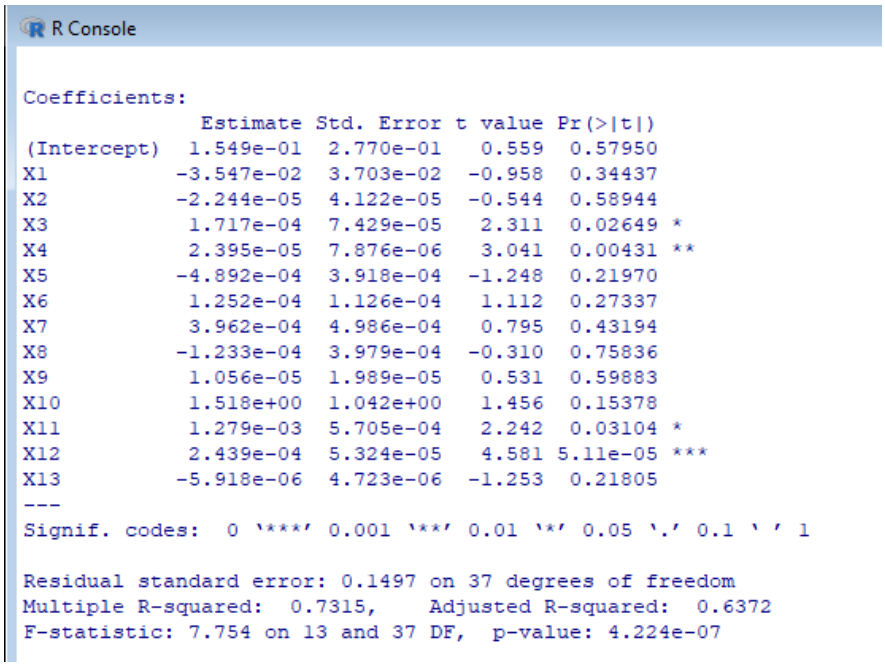

FIGURE 4.2 – Résultat du premier modèle

— **Coefficients :** Il affiche les estimations des coefficients de régression pour chaque variable indépendante. Ces coefficients quantifient l'impact de chaque variable indépendante sur la variable dépendante. Nous pouvons utiliser ces coefficients pour interpréter l'importance et la direction de l'effet de chaque variable.
Que nous avons trouver comme suite :

β_0	β_1	β_2	β_3	β_4	β_5	β_6
0.1549	-0,03547	-0,00002244	0,0001717	0,00002395	-0,0004892	0,0001252
β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}
0,0003962	-0,0001233	0,00001056	1,518	0,001279	0,0002439	-0,000005918

TABLE 4.1 – Les coefficients β_0,....,β13 obtenus après la régression

— **P-value :** La colonne "Pr($|t|$)" dans le résumé représente les valeurs p associées à chaque coefficient. Les valeurs p mesurent la significativité statistique de chaque variable indépendante. Plus la valeur p est faible (typiquement inférieure à 0,05), plus la variable est considérée comme **significative**.
— **Résidus :** Le résumé fournit des statistiques sur les résidus, qui représentent les écarts entre les valeurs prédites par le modèle et les valeurs réelles. Les résidus peuvent être utilisés pour évaluer la qualité de l'ajustement du modèle aux données.
— **Statistiques de l'ajustement :** Le résumé inclut également des statistiques de l'ajustement du modèle, telles que le R-carré (coefficient de détermination) et l'ajustement R-carré. Ces statistiques mesurent la qualité globale de l'ajustement du modèle aux données.
— **Multiple R-squared :** Indique qu'environ 73,15 % de la variance de la variable dépendante s'explique par les variables indépendantes de notre modèle de régression. Cela suggère que le modèle a un ajustement raisonnablement bon et que les variables indépendantes sont efficaces pour expliquer la variabilité de la variable dépendante.
— **Adjusted R-squared :** tient compte du nombre de prédicteurs dans le modèle et fournit une mesure de la proportion de variance expliquée par les variables indépendantes, ajustée pour les degrés de liberté. Cela pénalise l'ajout de prédicteurs inutiles au modèle.

4.2.1 Matrice de variance-covariance

Pour calculer la matrice de variance-covariance des coefficients de régression, nous pouvons utiliser la fonction " **vcov()** "dans R. Cette fonction prend en entrée le modèle de régression comme suite :

```
#matrice VarCoVar
Mvcov = vcov(model1)
print(Mvcov)
```

FIGURE 4.3 – Code pour calculer la matrice de variance covariance du modèle 1 sous langage R

Ce qui nous renvoie la matrice de variance-covariance suivante :

	(Intercept)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
(Intercept)	7.675592e-02	-9.007522e-03	-1.804202e-06	4.179447e-06	9.307819e-07	-2.723094e-05	-1.047547e-06	1.512236e-05	6.759988e-06	8.084362e-07	9.051387e-02	-3.503521e-05	-3.659957e-06	-3.782489e-07
X1	-9.007522e-03	1.371083e-03	2.019829e-07	-4.702334e-07	-1.247894e-07	3.065325e-06	-5.008657e-07	-2.076015e-06	-1.227568e-07	-9.057103e-08	-1.665132e-02	-4.993934e-06	-3.674326e-07	4.301911e-08
X2	-1.804202e-06	2.019829e-07	1.698726e-09	-2.773938e-09	-1.451021e-12	1.549413e-08	-2.935756e-10	-2.032869e-08	1.150630e-09	-8.165404e-10	-7.160320e-06	2.368245e-09	-1.435261e-10	1.870901e-10
X3	4.179447e-06	-4.702334e-07	-2.773938e-09	5.519255e-09	1.317827e-10	-2.693942e-08	3.780730e-10	3.174918e-08	1.212665e-09	1.302011e-09	1.334819e-05	-4.552191e-09	1.282980e-10	-3.282152e-10
X4	9.307819e-07	-1.247894e-07	-1.451021e-12	1.317827e-10	6.202779e-11	-8.466696e-10	8.453012e-11	-1.973057e-10	1.290194e-09	2.485894e-12	1.796870e-06	-1.204697e-10	-2.113221e-11	-9.941220e-12
X5	-2.723094e-05	3.065325e-06	1.549413e-08	-2.693942e-08	-8.466696e-10	1.535154e-07	-3.700139e-09	-1.822620e-07	-1.375301e-08	-7.482786e-09	-8.378887e-05	3.021429e-08	-4.592627e-10	1.842253e-09
X6	-1.047547e-06	-5.008657e-07	-2.935756e-10	3.780730e-10	8.453012e-11	-3.700139e-09	1.267855e-08	4.694873e-09	-6.332373e-09	1.394199e-10	2.583607e-05	6.461650e-09	1.268331e-09	-3.090664e-11
X7	1.512236e-05	-2.076015e-06	-2.032869e-08	3.174918e-08	-1.973057e-10	-1.822620e-07	4.694873e-09	2.485927e-07	-3.127104e-08	9.861920e-09	8.569200e-05	-1.335444e-08	3.192308e-09	-2.205292e-09
X8	6.759988e-06	-1.227568e-07	1.150630e-09	1.212665e-09	1.290194e-09	-1.375301e-08	-6.332373e-09	-3.127104e-08	1.583046e-07	-7.709204e-10	1.058589e-06	-2.025026e-08	-2.593988e-09	-6.255998e-11
X9	8.084362e-07	-9.057103e-08	-8.165404e-10	1.302011e-09	2.485894e-12	-7.482786e-09	1.394199e-10	9.861920e-09	-7.709204e-10	3.957981e-10	3.410188e-06	-1.105579e-09	7.416484e-11	-9.033873e-11
X10	9.051387e-02	-1.665132e-02	-7.160320e-06	1.334819e-05	1.796870e-06	-8.378887e-05	2.583607e-05	8.569200e-05	1.058589e-06	3.410188e-06	1.086542e+00	6.845913e-05	1.106949e-05	-1.040845e-06
X11	-3.503521e-05	-4.993934e-06	2.368245e-09	-4.552191e-09	-1.204697e-10	3.021429e-08	6.461650e-09	-1.335444e-08	-2.025026e-08	-1.105579e-09	6.845913e-05	3.254142e-07	2.329499e-08	3.635267e-10
X12	-3.659957e-06	-3.674326e-07	-1.435261e-10	1.282980e-10	-2.113221e-11	-4.592627e-10	1.268331e-09	3.192308e-09	-2.593988e-09	7.416484e-11	1.106949e-05	2.329499e-08	2.834357e-09	-4.329061e-12
X13	-3.782489e-07	4.301911e-08	1.870901e-10	-3.282152e-10	-9.941220e-12	1.842253e-09	-3.090664e-11	-2.205292e-09	-6.255998e-11	-9.033873e-11	-1.040845e-06	3.635267e-10	-4.329061e-12	2.230427e-11

FIGURE 4.4 – Matrice de Var-Covariance

4.2.2 Calcul des écarts-types

Nous calculons les écarts-types à l'aide de la fonction **"sqrt(diag(vcov()))"**, cela nous donne :

```
#Calcul des Ecarts-types
Ecatype = sqrt(diag(vcov(model1)))
print(Ecats-types)
```

FIGURE 4.5 – Code pour calculer les écarts-types Sous R

Le code précédent nous a permis de obtenir les calculs présenter dans le tableau suivant :

(Intercept)	2.770486×10^{-1}
X1	3.702814×10^{-2}
X2	4.121560×10^{-5}
X3	7.429169×10^{-5}
X4	7.875772×10^{-6}
X5	3.918104×10^{-4}
X6	1.125991×10^{-4}
X7	4.985907×10^{-4}
X8	3.978751×10^{-4}
X9	1.989467×10^{-5}
X10	1.042373
X11	5.704508×10^{-4}
X12	5.323868×10^{-5}
X13	4.722739×10^{-6}

TABLE 4.2 – Résultat des écarts-types

4.2.3 L'intervalle de confiance

L'intervalle de confiance est une estimation statistique qui fournit une plage de valeurs plausibles pour un paramètre inconnu d'une population, basée sur un échantillon de cette population. Il est utilisé pour estimer l'incertitude associée à une estimation et pour déterminer la précision de cette estimation.

A fin de calculer l'intervalle de confiance, nous utilisons la commande **"confint()"** pour donner ce qui suit :

```
#Calcul de l'inervale de confiance
IntConf = confint(model1,conf.level = 0:95)
print(IntConf)
```

FIGURE 4.6 – Code pour calculer l'intervalle de confiance Sous R

Cela calculera les intervalles de confiance à 95 % pour chaque coefficient du **"modèle 1"** et les stockera dans la variable **"IntConf"**. Ensuite, la fonction **"print"** affichera les intervalles de confiance calculés.
Nous obtenons le résultat qui suit :

Variable	2.5%	97.5%
(Intercept)	-0.406	0.716
X1	-0.110	0.040
X2	-0.000106	0.000061
X3	0.000021	0.000322
X4	0.000008	0.000040
X5	-0.001283	0.000305
X6	-0.000103	0.000353
X7	-0.000614	0.001406
X8	-0.000929	0.000683
X9	-0.000030	0.000051
X10	-0.594	3.630
X11	0.000123	0.002435
X12	0.000136	0.000352
X13	-0.000015	0.000004

TABLE 4.3 – L'intervalle de confiance

4.2.4 ANOVA

On commence d'abord par calculer le SCR et le SCE ainsi que le SCT pour cela on a le programme suivant :

```
X <- dds[, -14]
col <- rep(1, nrow(X))
X1 <- cbind(col, X)
X2 = as.matrix(X1)

residus <- residuals(model1)
print(residus)
SCR <- sum(residus^2)
print(SCR)

Y <- dds[, 14]
SCT <- t(Y)%*%Y -51*(mean(Y)^2)
print(SCT)

coefficients <- coef(model1)
beta<- as.vector(coefficients)
SCE <- beta%*%t(X2)%*%Y -51*(mean(Y)^2)
print(SCE)
```

FIGURE 4.7 – Code pour calculer SCR, SCE et SCT Sous R

```
X <- dds[, -14]
```

Cette commande signifie que vous créez une nouvelle variable X en provoquant toutes les colonnes de la variable dds, sauf la colonne 14.

```
col <- rep(1, nrow(X))
```

Cette commande crée un vecteur col contenant des valeurs constantes égales à 1. La longueur de ce vecteur est déterminée par le nombre de lignes de la matrice ou du data frame X.

```
X1 <- cbind(col, X))
```

La commande précédente crée une nouvelle matrice X1 en concaténant le vecteur col(qui contient des valeurs constantes) avec la matrice X. La fonction cbind()est utilisée pour combiner des objets en colonnes.

```
X2 = as.matrix(X1)
```

Convertit la matrice X1 en une matrice X2 de type matrix. Cela peut être utile si la matrice X1 était initialement un data frame ou un autre type d'objet, et que vous souhaiteriez travailler avec une matrice.
On suite nous avons fait un calcul de résidus pour pouvoir obtenir la valeur de chaqu'une des sommes qui sont comme suite :

- **SCR = 0.8296328**
- **SCT = 3.089804**
- **SCE = 2.260171**

avec un

- **CME = 0.17385930**
- **CMR = 0.01626730**

et donc on obtien le tableau d'ANOVA suivant :

Source de variation	ddl	Somme des carrés	Carrés moyens
Expliquée	13	2.260171	0.17385930
Résiduelle	37	0.8296328	0.01626730
Totale	50	3.089804	0.0617960

TABLE 4.4 – analyse de la variance de la régression linéaire multiple

4.2.5 Coefficient de détermination

Le coefficient de détermination simple est calculer par :

R2 = SCE/SCT

On obtient :

$$R^2 = 0.7315 = 73\%$$

On remarque que $R^2 = 0.73$, ce qui implique que l'ajustement est presque bon et que ce modèle de régression linéaire multiple proposé explique 73% du la variation totale.

4.2.6 Test de Fisher

On calcule la statistique de test par :

$$F_{obs} = \frac{\frac{SCE}{p}}{\frac{SCR}{n-(p+1)}} = \frac{CME}{CMR} = 10.687041$$

D'après la table de la loi Fisher, la valeur critique $f^{1-\alpha}_{p,n-(p+1)}$ d'ordre $(1-\alpha)$ d'une loi de Fisher à (p,n-(p+1)) degré de liberté est :

```
qf(0.95, df1 = 13,df2= 37)
```

Comme la statistique F_{obs}égale à10.687041 est supérieure à la valeur critique égale à1.995221. On conclut que le test est significatif , et donc au moins une des variables explicatives contribue à expliquer le phénomène.

4.2.7 Test de Student

Test de Student permet de répendre à la question :

L'apport marginal d'une variable X_j est-il significative ?

On à tester les hypothèses :
— $H_0 : \beta_j = 0, \forall j = 1, \ldots, 13$
— $H_1 : \beta_j \neq 0, \forall j = 1, \ldots, 13$
On calcule la statistique de test T_{obs}par :

$$T_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)}$$

D'après la table de la loi Student, la valeur critique $t^{1-\frac{\alpha}{2}}_{n-(p+1)}$ d'ordre $(1-\frac{\alpha}{2})$ d'une loi Student à (p,n-(p+1)) degré de liberté est :

```
qt(0.975,df=37) = 2.026192
```

T_{β_0}	T_{β_1}	T_{β_2}	T_{β_3}	T_{β_4}	T_{β_5}	T_{β_6}
0.57950	0.34437	0.58944	0.02649	0.00431	0.21970	0.27337
T_{β_7}	T_{β_8}	T_{β_9}	$T_{\beta_{10}}$	$T_{\beta_{11}}$	$T_{\beta_{12}}$	$T_{\beta_{13}}$
0.43194	0.75836	0.59883	0.15378	0.03104	0.0000511	0.21805

TABLE 4.5 – Les résultats de la statistique de test Student

4.3 Observations

Nous trouvons donc 4 variables explicatives que nous garderons dans notre prochain modèle et qui sont X3, X4, X11 et X12 ce qui nous donne les commandes suivantes :

```
model2 = lm(Y~ X3+ X4+ X11+ X12,dds)
summary(model2)
```

Et nous obtenant donc le summary suivant :

```
Call:
lm(formula = Y ~ X3 + X4 + X11 + X12, data = dds)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31766 -0.16684 -0.05124  0.11910  0.43762

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.750e-01  1.421e-01   1.935   0.0592 .
X3           3.304e-05  1.309e-05   2.524   0.0151 *
X4          -7.370e-07  2.096e-06  -0.352   0.7268
X11         -1.243e-04  6.366e-04  -0.195   0.8461
X12          1.123e-04  6.113e-05   1.837   0.0726 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2154 on 46 degrees of freedom
Multiple R-squared:  0.309,     Adjusted R-squared:  0.2489
F-statistic: 5.143 on 4 and 46 DF,  p-value: 0.001647
```

FIGURE 4.8 – Code de la régression linéaire du deuxième modèle sous langage R

D'après la figure précédente nous remarquons que :

— L'intercept (constante) est estimé à 0.275 avec une erreur standard de 0.1421. Son p-value est proche de 0.05, ce qui suggère une tendance à être significatif, mais il ne l'est pas de manière statistiquement significative à ce niveau de signification (0.0592).
— Le coefficient de détermination multiple (R-carré multiple) est de 0.309, ce qui signifie que les variables explicatives dans le modèle expliquent environ 30.9% de la variance de la variable de réponse.
— L'ajustement du R-carré (Adjusted R-squared) est de 0.2489, ce qui tient compte du nombre de variables explicatives dans le modèle et ajuste le R-carré en conséquence.
— Le test F a une statistique de 5.143 et une p-value de 0.001647, indiquant une significativité globale du modèle.

Nous avons conclus que le deuxième modèle n'est pas approuvable ce pendant nous avons proposé d'essayer encore deux autres modèles le premier consiste à enlever les variables significatives précédentes (X3, X4, X11 ,X12) du modèle 1 et réessayer de trouver d'autres variables significatives donc on obtient :

```
model3 = lm(Y~X1+ X2+ X5+ X6+ X7+ X8+ X9+ X10+ X13,dds)
summary(model3)
```

```
Call:
lm(formula = Y ~ X1 + X2 + X5 + X6 + X7 + X8 + X9 + X10 + X13,
    data = dds)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33309 -0.12129  0.00477  0.10108  0.44218

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.962e-02  3.225e-01   0.092   0.9273
X1           4.268e-02  4.292e-02   0.994   0.3258
X2           4.925e-05  2.038e-05   2.417   0.0202 *
X5           4.462e-04  1.942e-04   2.298   0.0268 *
X6          -4.486e-05  1.462e-04  -0.307   0.7605
X7          -4.728e-04  2.967e-04  -1.594   0.1187
X8          -4.090e-04  4.832e-04  -0.846   0.4023
X9          -2.464e-05  1.153e-05  -2.137   0.0386 *
X10         -4.668e-01  1.323e+00  -0.353   0.7259
X13          5.147e-06  2.194e-06   2.346   0.0239 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2011 on 41 degrees of freedom
Multiple R-squared:  0.4632,    Adjusted R-squared:  0.3453
F-statistic: 3.931 on 9 and 41 DF,  p-value: 0.001144
```

FIGURE 4.9 – Code de la régression linéaire du troisième modèle sous langage R

— L'intercept (constante) est estimé à 0.02962 avec une erreur standard de 0.3225. Son p-value est élevé (0.9273), ce qui indique qu'il n'y a pas suffisamment de preuves pour conclure qu'il y a un effet significatif de l'intercept sur la variable de réponse Y.
— Le coefficient pour la variable X2 est estimé à 4.925e-05 avec une erreur standard de 2.038e-05. Son p-value est inférieur à 0.05 (0.0202), ce qui indique une significativité statistique, suggérant un effet significatif de la variable X2 sur la variable de réponse Y.
— Le coefficient pour la variable X5 est estimé à 4.462e-04 avec une erreur standard de 1.942e-04. Son p-value est inférieur à 0.05 (0.0268), indiquant une significativité statistique, ce qui suggère un effet significatif de la variable X5 sur la variable de réponse Y.
— Le coefficient pour la variable X9 est estimé à -2.464e-05 avec une erreur standard de 1.153e-05. Son p-value est inférieur à 0.05 (0.0386), indiquant une significativité statistique, ce qui suggère un effet significatif de la variable X9 sur la variable de réponse Y.
— Le coefficient pour la variable X13 est estimé à 5.147e-06 avec une erreur standard de 2.194e-06. Son p-value est inférieur à 0.05 (0.0239), indiquant une significativité statistique, ce qui suggère un effet significatif de la variable X13 sur la variable de réponse Y.
— Le coefficient de détermination multiple (R-carré multiple) est de 0.4632, ce qui signifie que les variables explicatives dans le modèle expliquent environ 46.32% de la variance

de la variable de réponse.
— Le test F a une statistique de 3.931 et une p-value de 0.001144, indiquant une significativité globale du modèle.

En conclusion, dans ce modèle de régression linéaire multiple, les variables X2, X5, X9 et X13 semblent avoir un effet significatif sur la variable de réponse Y, tandis que les autres variables ne montrent pas suffisamment de preuves pour être considérées comme ayant un effet significatif. Cependant, il est important de prendre en compte le contexte de l'étude et d'autres considérations statistiques avant de tirer des conclusions définitives.

A fin d'avoir le dernier modèle (le modèle 4) nous proposons de le composée de l'addition des variables significatives du modèle 1 (X3, X4, X11, X12)et celles du troisième modèle (X2, X5, X9, X13)ce qui nous permis d'obtenir le modele suivant :

```
model4 = lm(Y~X2+ X3+ X4+ X5+X9+ X11+ X12+X13 ,dds)
summary(model4)
```

Nous obtenons donc :

```
Call:
lm(formula = Y ~ X2 + X3 + X4 + X5 + X9 + X11 + X12 + X13, data = dds)

Residuals:
      Min        1Q    Median        3Q       Max
-0.212787 -0.064140 -0.005721  0.075242  0.277561

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.869e-02  1.226e-01  -0.234  0.81617
X2           9.348e-06  4.657e-06   2.007  0.05117 .
X3           1.199e-04  3.472e-05   3.454  0.00128 **
X4           2.096e-05  6.149e-06   3.410  0.00145 **
X5          -1.727e-04  9.537e-05  -1.811  0.07733 .
X9          -4.775e-06  1.698e-06  -2.813  0.00744 **
X11          1.124e-03  5.361e-04   2.097  0.04207 *
X12          2.164e-04  4.955e-05   4.368 8.04e-05 ***
X13         -2.067e-06  1.226e-06  -1.686  0.09920 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1464 on 42 degrees of freedom
Multiple R-squared:  0.7085,    Adjusted R-squared:  0.653
F-statistic: 12.76 on 8 and 42 DF,  p-value: 4.358e-09
```

FIGURE 4.10 – Code de la régression linéaire du dernier modèle sous langage R

Dans l'analyse de régression linéaire multiple effectuée, le modèle inclut les variables explicatives X2, X3, X4, X5, X9, X11, X12 et X13 pour prédire la variable dépendante Y. Voici un compte rendu détaillé des résultats obtenus :
— Coefficients : Les coefficients estimés représentent l'effet de chaque variable explicative sur la variable dépendante. Par exemple, pour une unité d'augmentation de X3, on s'attend à une augmentation de 0.0001199 dans la valeur attendue de Y, en tenant

compte des autres variables explicatives. Certains des coefficients ont des valeurs statistiquement significatives avec des codes d'indication de signification, tels que ** pour un niveau de confiance de 0.001 et * pour un niveau de confiance de 0.05.

— Intercept (Interception) : L'interception est le coefficient estimé pour la constante (Intercept) du modèle lorsque toutes les variables explicatives sont nulles. Dans ce cas, l'interception est estimée à -0.02869, mais il n'est pas statistiquement significatif (p-value = 0.81617).

— Residuals (Résidus) : Les résidus représentent les différences entre les valeurs observées de la variable dépendante et les valeurs prédites par le modèle. Les résidus ont une moyenne proche de zéro, ce qui indique que le modèle capture en grande partie les variations de la variable dépendante. Les résidus ont également une distribution approximativement symétrique, comme indiqué par les valeurs du premier quartile (1Q), de la médiane et du troisième quartile (3Q).

— Residual standard error (Erreur standard des résidus) : Cet indicateur est une estimation de l'écart-type des résidus du modèle. Dans ce cas, il est estimé à 0.1464, ce qui signifie que les valeurs observées de la variable dépendante varient en moyenne d'environ 0.1464 unité par rapport aux valeurs prédites par le modèle.

— R-squared (R-carré) et Adjusted R-squared (R-carré ajusté) : Le R-carré mesure la proportion de la variation totale de la variable dépendante expliquée par le modèle. Dans ce cas, le R-carré est de 0.7085, ce qui signifie que le modèle explique environ 70.85% de la variation de la variable dépendante. L'R-carré ajusté tient compte du nombre de variables explicatives et de la taille de l'échantillon, et il est donc légèrement inférieur à l'R-carré (0.653). Cela indique que les variables explicatives du modèle ont une bonne capacité à expliquer la variation de la variable dépendante.

— F-statistic (Statistique F) : Cette statistique évalue la significativité globale du modèle en comparant l'ajustement du modèle aux données par rapport à un modèle nul. Un F-statistic élevé (12.76) avec une p-value très faible (4.358e-09) indique que le modèle global est statistiquement significatif. En d'autres termes, il existe des preuves solides que les variables explicatives du modèle contribuent de manière significative à la prédiction de la variable dépendante.

En conclusion, le modèle de régression linéaire multiple montre une bonne adéquation aux données, avec des variables explicatives significatives pour certaines des variables. Cependant, en ce qui s'en suit nous allons faire une comparaison entre les 4 modèles obtenus et essayé de trouver le meilleur modèle.

4.4 Comparaison des modèles

4.4.1 L'AIC et le BIC

L'AIC et le BIC sont des mesures utilisées pour comparer la qualité des modèles statistiques, en prenant en compte à la fois l'ajustement du modèle et la complexité du modèle. Voici comment calculer ces deux critères :

AIC (Akaike's Information Criterion) :
L'AIC est calculé à l'aide de la formule suivante :

```
AIC = 2k - 2ln(L)
```

où k est le nombre de paramètres dans le modèle et L est la valeur maximisée de la fonction de vraisemblance du modèle.

BIC (Bayesian Information Criterion) :
Le BIC est calculé à l'aide de la formule suivante :

```
BIC = -2ln(L) + k * ln(n)
```

où k est le nombre de paramètres dans le modèle L est la valeur maximisée de la fonction de vraisemblance du modèle, et n est la taille de l'échantillon.

Dans les deux cas, un modèle avec un AIC ou un BIC plus bas est considéré comme préférable, car cela indique un meilleur équilibre entre l'ajustement du modèle et la complexité du modèle.

Nous allons choisir un modèle en tenant compte de trois paramètres qui sont comme suite :

— **AIC** qui est à minimiser.
— **BIC** qui est à minimiser.
— R^2 qui à maximiser.

Modèle	AIC	BIC	R^2
model1	-35.31675	-6.339369	0.7315
model2	-5.109436	6.481518	0.309
model3	-7.984657	13.26543	0.4632
model4	-41.12551	-21.80725	0.7085

TABLE 4.6 – Tableau de comparaison des modèles

4.5 Discussion de la comparaison

Après comparaison et en analysant ces résultats, on constate que le modèle 4 présente les valeurs les plus faibles pour l'AIC et le BIC, ce qui indique une meilleure adéquation du modèle aux données tout en évitant la surcomplexité. De plus, le modèle 4 obtient le R^2 le plus élevé après le premier modèle parmi tous les modèles, ce qui suggère qu'il explique de manière plus précise la variation des données.

Ainsi, en considérant les critères de minimisation de l'AIC et du BIC, ainsi que la maximisation du R^2, il est clair que le modèle 4 est le meilleur choix parmi les modèles étudiés.

4.6 Conclusion

Ce chapitre représente la partie pratique de notre travail, nous avons tout d'abord présenter le langage de programmation R qui dédié aux recherches statistiques.
Ensuite nous l'avons utilisé pour l'application de la régression linéaire multiple sur notre problème pour l'obtention et l'étude de plusieurs modèles.
Enfin nous avons calculé l'AIC et Le BIC qui sont des mesures statistiques pour la sélection d'un modèle et nous avons fait une comparaison les modèles obtenus en utilisant la régression, et en se fiant aux tests statistiques tel que Fisher et Student...ect nous avons finis par choisir un modèle.

Conclusion générale :

Ce travail nous a permis d'aborder un cas concret et de découvrir la méthodologie de résolution des problèmes, en passant par la modélisation de l'aspect pratique jusqu'à la résolution mathématique. Nous avons fait face à la complexité du problème et nous avons cherché à trouver une solution qui, bien que peut-être pas optimale, s'en approche autant que possible. La nature pratique du problème de la corrosion des pipelines présente des défis considérables, mais notre approche nous permet d'explorer différentes étapes pour aboutir à une solution.

Dans notre mémoire, nous nous sommes intéressés à l'application de la régression linéaire multiple sur un problème réels dans le but te trouver un modèle mathématique a fin de trouver une équation mathématique pour pouvoir prédire la vitesse de corrosion sur les sites pétroliers.

Le travail a permis à travers les résultats obtenus de trouver un modèle de prédiction. Cela nous permis de déduire en quelque sorte quels composantes chimiques pause problème et de proposer a l'entreprise de trouver un moyen pour éliminer quelques paramètres chimiques présent dans les eaux tel que Ba++, Fe++, Sr++ et Co3– pour réussir à augmenter l'efficacité de notre modèle, qui avec un plus grand nombre d'échantillons pourront trouver une solution plus exact a l'aide de l'intelligence artificielle (Maching Learning).

Bibliographie

[1] A.JENKINS, 2006, «Introduction to corrosion in oil and gas production,» Algiers, MI production.« MEMOIRE DE BOUTELDJA MALIKA, GHARBIE KHEIRA».

[2] BARCA Mokriv,1989 : Etude de synthese sur le suivi Hydrochimique des eaux de Hassi R'Mel (Rapports interne, sonatrach).

[3] BOURBONNAIS, Régis, 2019 *Économetrie : Cours et exercices corrigés*. 9e édition. Dunod.« MEMOIRE de AISSAOUI Fares, MEHBALI Ibrahim, OUDJEHANI Ammar ».

[4] C. DOUGLAS , Elizabeth A. Peck, and G. Geoffrey Vining, 2015, Montgomery, *Introduction to linear regression analysis*. Vol. 821. John Wiley & Sons. « MEMOIRE de AISSAOUI Fares, MEHBALI Ibrahim, OUDJEHANI Ammar ».

[5] Chaouche Mokriv A : 1989 contrôle de production des puits de Hassi Rmel par analyse Hydrochimique des eaux de condensation (seminaire : technique puits).

[6] Christophe Chesneau, 2017,« Modèle de régression », Université de Caen-Normandie.

[7] Dr HIDRA .Y.Cours Econométrie II L3 Economie Quantitative 2019-2020 : Vérifications des hypothèses du modèle de régression linéaire multiple.« MEMOIRE de AISSAOUI Fares, MEHBALI Ibrahim, OUDJEHANI Ammar ».

[8] F. Z. Mennad, 2015 «Etude de corrosion de l'acier API5CTGradN 80 dans des puits d'injection d'eau par l'inhibiteur N-(2-aminoéthyl),» chez mémoire de master, Ouargla Algerie, Université KasdiMerbeh.

[9] G. TRABANELLI, 1987, «Marcel Dekker. Y/ Mansfield, Corrosion Mechanism ».

[10] H .ARRIL Choeller, 1995 :Geochimie des eaux souterraine (Revue de l'institut Francais du petrole).

[11] H. T. C. W. H.B.WANG, , 2001 «Characterization of inhibitor and corrosion product film using electrochemical impedance spectroscopy (EIS),» chez Corrosion 2001, Nace International, Houston, TX, p. Paper n° 01023.

[12] J. Benard, A. Michel, J. Philibert, J. Talbot, 1969, Métallurgie Générale, Masson et C ie, Editeurs, Paris VI,49-52.

[13] J. P.BOUMERSBACH, 2005 «electrochemical characterization of a corrosion inhibitor : influence of temperature on the inhibition mechanism.,» chez 207th meeting of the electrochemical society, Quebec City (Canada, 15-20 mai 2005).

[14] M. AMRANI et N. Bondjadja 1998 : Etude et classification des eaux de condensation de Hassi Rmel w Corrosion Tubing x.

[15] M. Khalida, 2014 « Contribution à l'étude de l'incompatibilité entre un inhibiteur de corrosion et un inhibiteur de dépôt » mémoire de magister, p. 14.

[16] P. GARTLAND, 1998, «Choosing the right positions for corrosion monitoring on oil and gas pipelines,» Corr Ocean USA Nace, Houston, p. 83.

[17] Rodier : 1978 L'analyse de l'eau (eau naturelle, eau résiduaire, eau de mer).

Webliography

[18] Université de Relizane 2021, http://elearning-fr.univ-relizane.dz/moodle/pluginfile.php/30473/course/overviewfiles/CHAPITREI\%20g\%C3\%A9n\%C3\%A9ralit\%C3\%A9\%20sur\%20la\%20corrosion\%20\%28\%2021-22\%29-converti.pdf