GRENOBLE
ECOLE DE
MANAGEMENT
TECHNOLOGY & INNOVATION

Analyse des Disparités Régionales et Prévision des Prix du Gasoil Non Routier

Projet de thèse professionnelle en vue de l'obtention du titre

Mastère Spécialisé

en

Big Data

Réalisé par
Yassine NADRANI

Supérvisé par:
Dr. Alain RIVET

Grenoble Ecole de Management
ENS d'Informatique et des Mathématiques Appliquées
Grenoble, France

27 juin 2024

Table des Matières

Résumé

Cette thèse professionnelle se concentre sur l'analyse des disparités régionales et la prévision des prix du gasoil non routier (GNR) en France, essentiel pour les secteurs agricole, forestier et des travaux publics. L'objectif est de combler les lacunes existantes en utilisant des techniques avancées de data science et d'intelligence artificielle (IA). Les données sont collectées via web scraping de plusieurs distributeurs, puis structurées et stockées dans des bases de données relationnelles. Les techniques utilisées incluent le clustering (K-means, DBSCAN), la régression géographiquement pondérée (GWR), les modèles de séries temporelles (ARIMA, GARCH), et les techniques de machine learning (forêts aléatoires, réseaux de neurones). Ces modèles prédictifs permettront d'anticiper les variations futures des prix du GNR, en prenant en compte les facteurs économiques, politiques et environnementaux influençant ces prix. Les résultats offriront des perspectives nouvelles pour les décideurs politiques, les entreprises et la recherche académique.
Mots-clés : Gasoil Non Routier (GNR), Disparités Régionales, Modélisation Prédictive, Machine Learning, Régression Géographiquement Pondérée, Séries Temporelles

Abstract

This professional thesis focuses on the analysis of regional disparities and the prediction of non-road diesel (NRD) prices in France, which is crucial for the agricultural, forestry, and public works sectors. The aim is to fill existing gaps using advanced data science and artificial intelligence (AI) techniques. Data is collected through web scraping from multiple distributors, then structured and stored in relational databases. Techniques used include clustering (K-means, DBSCAN), geographically weighted regression (GWR), time series models (ARIMA, GARCH), and machine learning techniques (random forests, neural networks). These predictive models will enable the anticipation of future NRD price variations, considering the economic, political, and environmental factors influencing these prices. The results will provide new perspectives for policymakers, businesses, and academic research.

Key words : Non-Road Diesel (NRD), Regional Disparities, Predictive Modeling, Machine Learning, Geographically Weighted Regression (GWR), Time Series

CHAPITRE 1

Introduction

1.1 Objet de l'étude

Le gasoil non routier (GNR) est crucial pour les secteurs agricole, forestier et des travaux publics, où les fluctuations de prix impactent directement les coûts opérationnels et la compétitivité des entreprises. Cette thèse explore l'utilisation de la data science et de l'intelligence artificielle (IA) pour analyser les disparités régionales des prix du GNR et développer des modèles prédictifs pour anticiper les variations futures.

1.2 Contexte et motivation

1.2.1 Contexte Réglementaire et Économique

Le GNR, crucial pour les secteurs industriels comme l'agriculture et la construction, est sujet à des fluctuations de prix affectant les coûts opérationnels et la compétitivité. Depuis son autorisation en 2011 en France, le GNR est devenu obligatoire pour réduire les émissions polluantes. Les politiques énergétiques influencent fortement la demande et les prix du GNR [Eneb; EUR16; Leg11],

En janvier 2024, le gouvernement français a annulé la hausse progressive de la taxe sur le GNR prévue jusqu'en 2030, visant à alléger la fiscalité pour les agriculteurs, ce qui pourrait augmenter la consommation et la pression sur les prix si l'offre ne suit pas [ind24]

1.2.2 Importance de l'Analyse des Disparités Régionales pour le GNR

Les disparités régionales relatives aux prix des produits pétroliers en France sont bien documentées, influencées par divers facteurs tels que les coûts logistiques, les taxes lo-

cales, la disponibilité des produits et les politiques environnementales spécifiques à chaque région. [Fin; Enea] Cependant, contrairement à d'autres produits pétroliers, il existe une lacune notable en ce qu'il n'existe pas d'étude spécifique relative à ces disparités.[Enea]

Cette thèse vise à combler cette lacune en fournissant des outils pour les entreprises et les décideurs.

1.2.3 Pertinence Économique et Stratégique

Comprendre les disparités régionales et prévoir les variations futures des prix du GNR est crucial pour les décideurs politiques et les gestionnaires d'entreprises. Les résultats permettront de formuler des politiques économiques plus équitables et efficaces, d'optimiser les ressources et d'améliorer la compétitivité des entreprises.

1.2.4 Innovations Technologiques et Méthodologiques

La data science et l'IA offrent des opportunités sans précédent pour analyser les données complexes et prévoir les tendances économiques. Les techniques avancées de machine learning permettent d'analyser de grandes quantités de données avec précision et efficacité. L'application de ces technologies à l'analyse des prix du GNR peut révéler des insights précieux et des modèles difficiles à identifier autrement [Hon21; Csp]

1.3 Impacts et objectifs du projet

Cette étude comblera une lacune dans la littérature en fournissant une analyse régionale détaillée des prix du GNR en France. Les résultats offriront des perspectives nouvelles pour les décideurs politiques, les entreprises et la recherche académique, permettant de mieux anticiper et gérer les variations de coûts du GNR.

Cette thèse professionnelle se propose donc de répondre à plusieurs questions essentielles : quelles sont les disparités régionales des prix du GNR en France, quels sont les facteurs économiques, politiques et environnementaux qui influencent les prix du GNR, comment la data science et l'IA peuvent-elles être utilisées pour prévoir les prix du GNR avec précision, et quels modèles prédictifs offrent les meilleures performances pour la prévision des prix du GNR.

Ces objectifs visent à fournir des outils pratiques et théoriques pour mieux comprendre et gérer les fluctuations des prix du GNR, contribuant ainsi à une gestion plus efficace des ressources et à une prise de décision informée.

CHAPITRE 2

État de l'Art et Approches Techniques

2.1 Introduction

Dans ce chapitre, nous présentons une revue des méthodes et approches couramment utilisées pour analyser les disparités régionales des prix des produits pétroliers, ainsi que pour développer des modèles prédictifs anticipant leurs variations. Bien que notre focus principal soit le gasoil non routier (GNR), nous tirerons également des enseignements des études existantes sur d'autres produits pétroliers pour enrichir notre méthodologie. Ce chapitre couvre les sources de données disponibles, les techniques de collecte de données, les facteurs influençant les prix, et les approches de modélisation prédictive pertinentes.

2.2 Sources de Données et Techniques de Collecte

2.2.1 Sources de Données

Pour analyser les disparités régionales et prévoir les prix du GNR, des données précises provenant de diverses sources sont essentielles.

- Données Gouvernementales et Institutionnelles : Les agences gouvernementales telles que l'Agence Internationale de l'Énergie (AIE), le Ministère de la Transition Écologique en France, et les bases de données européennes fournissent des informations détaillées sur les prix des carburants et les politiques énergétiques.
- Données des Distributeurs de Carburants : Les sites web des principaux distributeurs de carburant offrent des informations sur les prix du GNR dans différentes

régions.

- Bases de Données Ouvertes et API : Des plateformes comme Open Data France et des API spécifiques permettent d'accéder à des données économiques, météorologiques et environnementales qui peuvent influencer les prix des carburants.

2.2.2 Techniques de Web Scraping

Le web scraping est crucial pour collecter des données en ligne de manière automatisée. Les outils et bibliothèques comme Beautiful Soup, Scrapy et Selenium sont couramment utilisés pour extraire des données de sites web : [Myr]

- Beautiful Soup : Une bibliothèque Python utilisée pour extraire des données de fichiers HTML et XML. [Cru]
- Scrapy : Un framework Python permettant de construire et exécuter des spiders pour le scraping de données web.[Scr]
- Selenium : Un outil d'automatisation de navigateur web qui peut être utilisé pour interagir avec des pages web dynamiques. [Sel]

2.3 Collecte des Données via Web Scraping

2.3.1 Justification de l'Approche

La collecte de données via web scraping est une méthode efficace pour obtenir des informations en temps réel sur les prix du GNR de différents fournisseurs. En effet, les sites des distributeurs apportent des informations détaillées sur les prix, reflétant les conditions locales du marché. Le scraping permet donc d'accéder à des données précises et actualisées, cruciales pour analyser les disparités régionales. Cette approche permet d'obtenir des données récentes, car nous n'avons pas accès à un historique actualisé, et certaines données ne sont disponibles qu'à l'échelle nationale et non régionale. En résumé, le web scraping fournit des données essentielles pour une modélisation prédictive fiable des prix du GNR.

2.3.2 Méthodologie Technique

La mise en place d'un pipeline de web scraping comprend plusieurs étapes essentielles pour assurer une collecte de données efficace et structurée.

Tout d'abord, la conception des scripts de scraping se fait à l'aide de bibliothèques évoquées dans la partie 2.2.2. Beautiful Soup est utilisé pour extraire des données de fichiers HTML et XML, offrant une manipulation intuitive de la structure des pages web. Scrapy permet de construire et d'exécuter des spiders pour parcourir le web de manière autonome et extraire des informations. Selenium est utilisé pour interagir avec des pages web dynamiques, exécutant des scripts JavaScript et simulant des actions utilisateur.

Ensuite, l'automatisation des tâches est cruciale pour garantir une mise à jour régulière des données. Cela implique la planification de l'exécution des scripts de scraping à intervalles réguliers. Dans notre projet, nous utiliserons Google Cloud Platform (GCP) pour orchestrer ces tâches avec des services comme Cloud Scheduler et Cloud Functions, permettant de planifier et d'exécuter des tâches en réponse à des événements .[Goob; Gooc]

Le stockage des données collectées se fera dans des bases de données relationnelles comme Google Cloud SQL. Ces services offrent une structuration rigoureuse, une flexibilité et une capacité de mise à l'échelle adaptées aux grandes quantités de données . Pour l'analyse des données, nous utiliserons BigQuery, un entrepôt de données sans serveur, hautement évolutif et économique qui permet d'exécuter des requêtes SQL sur de vastes ensembles de données de manière rapide et efficace. [Gooa]

Enfin, le traitement des données, comprenant le nettoyage et la structuration, sera effectué avec des outils comme Pandas en Python. [Che]

2.3.3 Considérations Éthiques

Le web scraping soulève des questions éthiques et légales importantes. Il est crucial de respecter les conditions d'utilisation des sites web pour éviter toute violation des termes de service. La protection des données personnelles est impérative, et il convient de se conformer aux régulations telles que le RGPD en Europe.[CNI16]

La collecte de données publiquement accessibles, telles que les prix affichés, réduit les préoccupations liées à la protection des données personnelles. Toutefois, il convient de

rester prudent et de s'assurer que les informations collectées ne sont pas soumises à des restrictions spécifiques par les distributeurs.

L'impact sur les sites web doit être minimisé en concevant des scripts de scraping efficaces qui évitent les requêtes excessives et les perturbations du service. Il est recommandé de mettre en place des délais entre les requêtes et de respecter les règles de courtoisie pour éviter de surcharger les serveurs des sites.[VLL]

2.4 Analyse des Disparités Régionales

2.4.1 Visualisation des Données

La visualisation des données est essentielle pour identifier et comprendre les disparités régionales dans les prix du GNR. Les outils de visualisation, tels que Matplotlib, Seaborn et Plotly, permettent de créer des graphiques clairs et informatifs.

Les cartes chloroformes sont utilisées pour représenter les variations régionales des prix sur une carte géographique. Les graphiques en barres et en lignes permettent de comparer les prix moyens dans différentes régions et leur évolution dans le temps.

2.4.2 Techniques de Clustering

Les techniques de clustering sont des outils puissants pour regrouper les régions en fonction des similitudes dans les prix du GNR. Elles permettent de segmenter les données en clusters homogènes, facilitant ainsi l'analyse des disparités régionales. Les méthodes de clustering les plus couramment utilisées dans ce contexte incluent l'algorithme K-means et l'algorithme DBSCAN.

L'algorithme K-means partitionne les données en K clusters, en minimisant la variance au sein de chaque cluster.[TRJ09] Cette méthode est particulièrement efficace pour des jeux de données volumineux et permet de déceler des tendances générales dans les prix du GNR. En utilisant K-means, les régions peuvent être regroupées en fonction de la similarité de leurs prix, révélant des modèles régionaux pertinents.[Mac67]

L'algorithme DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est une autre méthode de clustering qui identifie les clusters basés sur la densité des données.[M+96] Contrairement à K-means, DBSCAN peut identifier des clusters de forme

arbitraire et détecter les anomalies ou points aberrants. Cette méthode est utile pour traiter des données spatiales où les régions avec des variations de densité de prix du GNR peuvent exister.

2.4.3 Régression Géographiquement Pondérée (GWR)

La régression géographiquement pondérée (GWR) est une technique essentielle pour analyser les relations spatiales et comprendre comment les facteurs influencent les prix du GNR de manière variable selon les régions. Contrairement aux modèles de régression classiques qui supposent une relation uniforme sur l'ensemble du territoire, la GWR permet de capturer les variations locales en ajustant les coefficients de régression pour chaque point de données géographiques.[JG18]

L'avantage principal de la GWR réside dans sa capacité à fournir des prédictions précises au niveau régional, en tenant compte des particularités locales. En appliquant la GWR, il est possible de générer des cartes de coefficients qui montrent comment l'impact de chaque facteur varie spatialement.[SCM02]

2.5 Modélisation Prédictive des Prix

2.5.1 Modèles de Séries Temporelles

Les modèles de séries temporelles sont essentiels pour la prévision des prix du GNR, car ils permettent d'analyser les données chronologiques et d'identifier les tendances, les cycles et les schémas saisonniers. Parmi les modèles les plus utilisés figurent les modèles ARIMA et GARCH.

Le modèle ARIMA (AutoRegressive Integrated Moving Average) est une méthode statistique largement utilisée pour modéliser et prévoir des séries temporelles stationnaires.[JG18] Il combine trois composants principaux : l'auto-régression (AR), les moyennes mobiles (MA) et l'intégration (I) pour rendre la série temporelle stationnaire. L'ARIMA est efficace pour capturer les dépendances linéaires dans les données temporelles, ce qui en fait un choix populaire pour la prévision des prix du GNR.

En appliquant ces modèles de séries temporelles, il est possible de prévoir avec précision les futures variations des prix du GNR, fournissant ainsi des informations pré-

cieuses pour la planification et la prise de décision dans les secteurs concernés.

2.5.2 Techniques de Machine Learning

Les techniques de machine learning offrent des solutions puissantes pour la prévision des prix du GNR. Elles permettent de modéliser des relations complexes et de capturer des tendances non linéaires dans les données. Parmi les approches pertinentes figurent les forêts aléatoires et les réseaux de neurones.

Les forêts aléatoires sont des ensembles d'arbres décisionnels utilisés pour améliorer la précision des prédictions. Cette technique, introduite par Breiman (2001), combine les prédictions de plusieurs arbres construits sur des sous-échantillons aléatoires des données. Les forêts aléatoires sont particulièrement adaptées aux jeux de données volumineux et hétérogènes, permettant de capturer des interactions complexes entre les variables.[L01]

Les réseaux de neurones, en particulier les réseaux de neurones profonds (deep learning), sont capables de modéliser des relations très complexes et non linéaires dans les données. Ces modèles utilisent plusieurs couches de neurones pour extraire des caractéristiques hiérarchiques et sont particulièrement efficaces pour la prévision des séries temporelles et la reconnaissance de motifs. Les réseaux de neurones récurrents (RNN) et les réseaux de neurones convolutifs (CNN) sont des variantes spécialisées pour traiter des données séquentielles et spatiales respectivement.[GBC16]

En appliquant ces techniques de machine learning, il est possible de développer des modèles prédictifs robustes et précis pour anticiper les variations des prix du GNR, fournissant ainsi des outils précieux pour la prise de décision dans les secteurs affectés.

2.6 Facteurs Influents et Variables Utilisées

Pour assurer la pertinence des modèles prédictifs, il est crucial de sélectionner les variables influençant les prix du GNR. Ces variables peuvent être obtenues via des sources en ligne ou par scraping. Les principaux facteurs incluent les facteurs économiques, politiques, et environnementaux.

- Facteurs Économiques

  - Prix du baril Brent : Le prix du pétrole brut est un indicateur clé des prix des produits pétroliers. Les fluctuations des prix du pétrole brut sur les marchés mondiaux influencent directement les coûts des produits raffinés tels que le GNR. [Kos22]
  - Coûts de Transport et de Logistique : Les coûts liés au transport et à la logistique affectent également les prix régionaux du GNR. Dans cette étude, nous nous intéresserons particulièrement aux distances des dépôts de stockage et aux prix des carburants. Bien que les prix des carburants soient corrélés aux prix du baril, ils peuvent également varier en fonction des conditions locales de transport et de distribution. [Kos22]
  - Taux de Change : Les fluctuations des taux de change influencent les coûts d'importation des matières premières, ce qui peut affecter les prix du GNR [XEc]
  - Niveau de la Demande :La demande pour le GNR dans les secteurs agricoles, de la construction et des travaux publics est un facteur déterminant des prix. Les indicateurs de demande peuvent être obtenus via les bases de données gouvernementales, les rapports des associations industrielles et les publications des distributeurs de carburants. Une augmentation de l'activité dans ces secteurs, par exemple pendant les saisons de récolte agricole ou de construction, peut faire grimper les prix en raison de la demande accrue. [Kos22]

- Facteurs Politiques et Réglementaires

  Taxes et Subventions : Les politiques fiscales, y compris les taxes et subventions sur le GNR, influencent directement les prix. Ces informations sont disponibles via les sites gouvernementaux et les publications des ministères concernés. [Nat20]

- Facteurs Environnementaux

  Conditions Météorologiques : Les conditions météorologiques extrêmes, telles que les ouragans, peuvent perturber la production et la distribution de carburants, affectant ainsi les prix. Les données météorologiques sont disponibles via des API telles qu'OpenWeatherMap ou WeatherAPI. [Ope]

CHAPITRE 3

Faisabilité de l'Approche

3.1 Collecte des Données

Depuis le mois d'avril 2024, une collecte quotidienne des prix du GNR a été mise en place via web scraping pour plusieurs distributeurs/vendeurs dans plus de 1800 codes postaux différents. Cette collecte permet de capturer une grande variété de données spatiales et temporelles, offrant ainsi une base pour l'analyse des disparités régionales et la prévision des prix.Les données collectées sont automatiquement structurées et stockées dans des tables BigQuery grâce à l'automatisation mise en place sur Google Cloud Platform (GCP).

En plus des prix du GNR, d'autres types de données seront intégrés pour enrichir les analyses, notamment :

- Facteurs Économiques : prix du baril Brent, coûts de transport et de logistique, taux de change, niveau de la demande.
- Facteurs Politiques et Réglementaires : taxes et subventions sur le GNR.
- Facteurs Environnementaux : conditions météorologiques.

3.2 Accessibilité et Structuration des Données

La collecte de données via web scraping a permis de constituer une base de données structurée avec des prix quotidiens pour chaque distributeur et code postal. Les données additionnelles sur les facteurs économiques, politiques et environnementaux sont acces-

sibles via des API et des sources ouvertes. Ces données seront intégrées et structurées de manière à faciliter leur utilisation dans les différents modèles prédictifs et d'analyse.

3.3 Quantité des Données

3.3.1 Modèles de Séries Temporelles (ARIMA, GARCH)

Avec des données quotidiennes depuis avril, nous disposons d'environ 210 points de données à ce jour. Bien que ce nombre soit suffisant pour des analyses initiales et pour détecter des tendances et des saisonnalités à court terme, un historique de données d'au moins un an (365 points) serait idéal pour améliorer la robustesse des modèles et permettre des prévisions à moyen et long terme. [JG18]

3.3.2 Techniques de Machine Learning (Forêts Aléatoires, Réseaux de Neurones)

La collecte quotidienne des prix dans plus de 1800 codes postaux permet de disposer de plusieurs milliers de points de données en peu de temps. Cette richesse en données est favorable pour les modèles de machine learning, qui nécessitent un grand volume de données pour capturer les relations complexes et les tendances non linéaires. L'intégration des facteurs économiques, politiques et environnementaux enrichit encore davantage les données, rendant les modèles plus robustes et précis. [L01; GBC16]

3.3.3 Techniques de Clustering (K-means, DBSCAN)

La diversité spatiale et la quantité de données collectées sont suffisantes pour appliquer des techniques de clustering. Ces méthodes permettront de détecter des patterns régionaux significatifs et de regrouper les régions selon les similitudes dans les prix du GNR, facilitant ainsi l'analyse des disparités régionales. [TRJ09; M+96]

3.3.4 Régression Géographiquement Pondérée (GWR)

La GWR bénéficiera grandement des données spatiales riches. Avec des données couvrant plus de 1800 codes postaux, il est possible de capturer les variations locales et de comprendre les influences spatiales sur les prix du GNR. Plus les données sont nom-

breuses, plus les modèles de GWR seront précis et pertinents pour les analyses régionales.-[SCM02]

3.4 Contraintes de Temps et Planification

3.4.1 Phase initiale (Avril à Mai)

Au début de la collecte, les scripts de web scraping ont été développés et mis en place pour collecter les données quotidiennes des prix du GNR de multiples distributeurs dans différents codes postaux. Les données collectées ont été automatiquement nettoyées et structurées, puis stockées dans des tables BigQuery grâce à des pipelines automatisés sur Google Cloud Platform. Cette phase a également inclus l'intégration de données historiques disponibles pour augmenter le volume de données et enrichir les analyses.

3.4.2 Phase de Modélisation (Juin à Août)

Pendant les mois de juin à août, le développement des modèles de séries temporelles sera effectué, avec l'implémentation des modèles ARIMA et GARCH pour chaque distributeur et code postal. Ces modèles seront initialement validés avec les données disponibles. L'objectif est de capturer les tendances et les saisonnalités à court terme. Les modèles seront ajustés et optimisés en fonction des résultats des validations initiales. Simultanément, les forêts aléatoires et les réseaux de neurones seront implémentés en utilisant les données enrichies. Ces modèles bénéficieront de la diversité des données collectées pour capturer des relations complexes. L'ingénierie des features sera réalisée pour inclure des variables additionnelles telles que les prix du baril Brent, les coûts logistiques, et les conditions météorologiques. Enfin, les méthodes de clustering, telles que K-means et DBSCAN, seront appliquées pour détecter des clusters régionaux en fonction des prix du GNR, permettant d'identifier des patterns régionaux et de segmenter les données en groupes homogènes.

3.4.3 Phase de Validation et d'Optimisation (Septembre à Octobre)

Au cours des mois de septembre et octobre, les modèles développés seront soumis à des méthodes de validation croisée pour ajuster et optimiser leurs performances. Cette phase inclura des tests avec des données récentes pour s'assurer de la robustesse et de la

précision des prévisions. Les modèles de machine learning et de séries temporelles seront ajustés en fonction des résultats des validations. Les tests finaux des modèles seront effectués avec des données récentes pour évaluer leur performance finale. Les ajustements nécessaires seront réalisés pour améliorer la précision des prédictions. Les résultats des tests seront analysés pour identifier les forces et les faiblesses des modèles.

3.4.4 Phase Finale (Novembre)

En novembre, les résultats obtenus seront interprétés et visualisés pour illustrer les disparités régionales et les prévisions des prix du GNR. Les techniques de visualisation, telles que les cartes chloroformes et les graphiques en barres, seront utilisées pour présenter les résultats de manière claire et informative. Le rapport final de la thèse sera rédigé, incluant une discussion sur les implications des résultats et des recommandations pour les décideurs politiques et les entreprises. Cette phase inclura également une réflexion sur les limitations de l'étude et les perspectives de recherches futures.

Bibliographie

[Che] Daniel Y. Chen. *Pandas for Everyone: Python Data Analysis, 2nd edition* (citée à la page 6).

[CNI16] CNIL. *Le règlement général sur la protection des données - RGPD*. 2016 (citée à la page 6).

[Cru] Crummy. *Beautiful Soup Documentation* (citée à la page 5).

[Csp] Cspdailynews. *3 Examples of Artificial Intelligence in Fuel Pricing* (citée à la page 2).

[Enea] Connaissance des Energies. *Structuration des prix de l'essence et du gazole en France* (citée à la page 2).

[Eneb] Total Energies. *Tout savoir sur le GNR ou gazole non-routier.* (Citée à la page 1).

[EUR16] EUR-Lex. *Emission limits and type-approval rules for non-road mobile machinery.* 2016 (citée à la page 1).

[Fin] Ministère de l'Économie des Finances et de la Souveraineté industrielle et numérique. *Les prix des carburants en France* (citée à la page 2).

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. *MIT Press*. 2016 (citées aux pages 9, 12).

[Gooa] Google. *BigQuery Documentation* (citée à la page 6).

[Goob] Google. *Cloud Functions* (citée à la page 6).

[Gooc] Google. *Cloud Scheduler* (citée à la page 6).

[Hon21] Hongfang Lu et al. "Energy price prediction using data-driven models: A decade review." In: *Computer Science Review* (2021). DOI: https://doi.org/10.1016/j.cosrev.2020.100356 (citée à la page 2).

[ind24] Direction générale des douanes et droits indirects. *Dispositif de taxation du GNR agricole et ou forestier*. 2024 (citée à la page 1).

[JG18] Hyndman R. J. and Athanasopoulos G. *Forecasting: Principles and Practice*. 2018 (citées aux pages 8, 12).

[Kos22] Paul Kosakowski. *Understanding Oil Prices: Industry, Gas, and Oil Futures*. 2022 (citée à la page 10).

[L01] Breiman L. *Random Forests. Machine Learning, 45(1), 5-32*. 2001 (citées aux pages 9, 12).

[Leg11] Legifrance. *Article 2 - Arrêté du 10 novembre 2011 fixant pour le gazole, les gaz de pétrole liquéfiés et les émulsions d'eau dans du gazole des conditions d'emploi ouvrant droit à l'application du régime fiscal privilégié institué par l'article 265 du code des douanes en matière de taxe intérieure de consommation*. 2011 (citée à la page 1).

[M+96] Ester M. et al. *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. 1996 (citées aux pages 7, 12).

[Mac67] J. MacQueen. *Tome Methods for classification and Analysis of Multivariate Observations*. 1967 (citée à la page 7).

[Myr] Myriam Ertz et al. *Web Scraping Techniques and Applications: A Literature Review* (citée à la page 5).

[Nat20] Assemblée Nationale. *Les taxes sur les carburants en France*. 2020 (citée à la page 10).

[Ope] OpenWeatherAPI. *Weather API* (citée à la page 10).

[SCM02] Fotheringham A. S., Brunsdon C., and Charlton M. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley*. 2002 (citées aux pages 8, 13).

[Scr] Scrapy. *Scrapy Documentation* (citée à la page 5).

[Sel] Selenium. *The Selenium Browser Automation Project* (citée à la page 5).

[TRJ09] Hastie T., Tibshirani R., and Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2009 (citées aux pages 7, 12).

[VLL] Krotov Vlad, Johnson Leigh, and Silva Leiser. *Legality and Ethics of Web Scraping. Communications of the Association for Information Systems* (citée à la page 7).

[XEc] XE.com. *Currency Exchange Rates* (citée à la page 10).