République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Akli Mohand Oulhadj de Bouira

Faculté des Sciences Exactes

Département de Mathématiques

Mémoire de Master

En Mathématiques

Spécialité : Recherche Opérationnelle

Thème

Estimation du potentiel solaire des sites algériens par régression : application à la sélection de futurs parcs photovoltaïques

Réalisé par :

MATARI NAIMA

Devant le jury composé de :

HAMID KARIM	Président	MCB	Univ. Bouira
IMINE OUIZA	Examinatrice	MAA	Univ. Bouira
BOUDANE KHADIDJA	Examinatrice	MAA	Univ. Bouira
LOUADJ Kahina	Encadreur	MCA	Univ. Bouira

Année universitaire : 2025/2026

Résumé

L'Algérie dispose d'un potentiel solaire exceptionnel et a engagé un ambitieux programme national de 15 000 MW de capacité renouvelable installée à l'horizon 2035. La concrétisation de ce programme nécessite des outils statistiques d'aide à la décision pour la sélection rapide des sites favorables au déploiement de centrales photovoltaïques. Ce mémoire propose une démarche complète d'estimation du potentiel solaire des sites algériens à partir de leurs caractéristiques géographiques et climatiques, en utilisant des méthodes de régression supervisée.

Le travail s'articule en quatre étapes successives. La première pose le cadre physique du problème : principes de la conversion photovoltaïque, modèle reliant l'énergie solaire reçue et la température au rendement énergétique, et présentation du programme national algérien des énergies renouvelables. La deuxième présente la démarche méthodologique de construction d'un jeu de données synthétique de 200 sites algériens, calibré sur les paramètres climatologiques publiés par la NASA POWER, et conduit l'analyse exploratoire associée (statistiques descriptives, corrélations de Pearson et Spearman, facteurs d'inflation de variance).

La troisième étape formalise rigoureusement les deux méthodes de régression retenues : la régression linéaire multiple, depuis la formulation matricielle jusqu'au théorème de Gauss-Markov, et la régression Ridge, avec son principe de pénalisation L2 et la sélection de l'hyperparamètre λ par validation croisée. La quatrième étape constitue le cœur expérimental : appliqués au jeu de 200 sites, les deux modèles atteignent des performances quasi identiques avec un coefficient de détermination $R^2 = 0{,}987$, une erreur absolue moyenne MAE = 15 kWh/kWc/an et une erreur relative MAPE = 0,90 %. L'analyse des coefficients confirme la prédominance de l'énergie solaire reçue globale comme prédicteur du rendement, et révèle un gradient nord-sud marqué : 1 925 kWh/kWc/an dans le Sahara central contre 1 405 kWh/kWc/an sur la côte méditerranéenne.

Mots-clés : Énergie solaire, photovoltaïque, régression linéaire multiple, régression Ridge, validation croisée, multicolinéarité, potentiel solaire, Algérie, NASA POWER, aide à la décision.

Abstract

Algeria has an exceptional solar potential and has launched an ambitious national programme of 15,000 MW of renewable installed capacity by 2035. The implementation of this programme requires statistical decision-making tools for the rapid selection of favourable sites for photovoltaic plants. This dissertation proposes a complete methodology for estimating the solar potential of Algerian sites from their geographical and climatic characteristics, using supervised regression methods.

The work is structured in four successive steps. The first lays the physical framework : principles of photovoltaic conversion, model linking énergie solaire reçue and temperature to energy yield, and presentation of the Algerian national renewable energy programme. The second presents the methodology for constructing a synthetic dataset of 200 Algerian sites, calibrated on climatological parameters published by NASA POWER, and conducts the associated exploratory analysis (descriptive statistics, Pearson and Spearman correlations, variance inflation factors).

The third step rigorously formalises the two regression methods selected : multiple linear regression, from matrix formulation to the Gauss-Markov theorem, and Ridge regression, with its L2 penalisation principle and hyperparameter λ selection by cross-validation. The fourth step constitutes the experimental core : applied to the 200-site dataset, both models achieve nearly identical performance with a determination coefficient $R^2 = 0.987$, a mean absolute error MAE = 15 kWh/kWp/year and a relative error MAPE = 0.90 %. The coefficient analysis confirms the dominance of global énergie solaire reçue as a predictor of yield, and reveals a marked north-south gradient : 1,925 kWh/kWp/year in the central Sahara versus 1,405 kWh/kWp/year on the Mediterranean coast.

Keywords : Solar energy, photovoltaic, multiple linear regression, Ridge regression, cross-validation, multicollinearity, solar potential, Algeria, NASA POWER, decision support.

Remerciements

Tout d'abord, je remercie Dieu Tout-Puissant de m'avoir donné la force, la patience et le courage nécessaires pour mener à bien ce travail.

Je tiens à exprimer ma profonde gratitude à mon encadreuse, **Madame Louadj Kahina**, pour son encadrement, sa disponibilité, ses conseils précieux et son soutien tout au long de la réalisation de ce mémoire.

Je remercie vivement les membres du jury pour l'honneur qu'ils m'ont accordé en acceptant d'évaluer ce travail, et plus particulièrement :
Monsieur Karim Hamid, Président du jury
Madame Imine Ouiza
Madame Boudane Khadidja

Mes sincères remerciements s'adressent à mes très chers parents pour leur amour, leurs sacrifices, leurs encouragements et leur soutien permanent durant tout mon parcours universitaire.

Je remercie également mes sœurs ainsi que toute ma famille pour leur aide morale et leurs prières.

Je n'oublie pas de remercier mes amies et toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce mémoire.

Enfin, j'adresse mes remerciements à l'entreprise **GSM** pour son accueil et son accompagnement durant cette expérience.

Dédicace

À mes très chers parents,

À ma mère et à mon père,

pour leur amour, leur patience, leurs sacrifices et leurs encouragements durant tout mon parcours.

Que Dieu les protège et les récompense.

À mes chères sœurs :

Nadjet, Fatima, Ilham, Aya et Nada,

À mon frère Mohamed,

pour leur soutien moral et leur présence à mes côtés.

À ma cousine Fatima,

avec tout mon amour, mon respect et ma gratitude.

À toute ma famille,

pour leurs prières et leurs encouragements.

À mon encadreuse,

pour ses conseils précieux et son accompagnement tout au long de ce travail.

À l'entreprise GSM,

pour son accueil et son aide durant cette expérience.

À ma chère amie et sœur de cœur, Maroua Hassiba,

pour son soutien, sa confiance et ses encouragements dans les moments difficiles.

Enfin, à toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce mémoire.

Table des matières

Table des figures

Liste des tableaux

Liste des symboles

Variables physiques

Symbole	**Signification**	**Unité**
P	Puissance produite par l'installation PV	W
P_{nom}	Puissance nominale (puissance crête)	kWc
G	Énergie solaire reçue (rayonnement global horizontal)	W/m²
G_{eff}	Énergie solaire reçue effective (corrigée nuages)	W/m²
G_{annual}	Énergie solaire reçue annuellement	kWh/m²/an
A	Surface des modules photovoltaïques	m²
T_c	Température de la cellule photovoltaïque	°C
T_{ref}	Température de référence STC (25°C)	°C
T_{amb}	Température ambiante	°C
η_{nom}	Rendement nominal du module	–
γ	Coefficient thermique de puissance	%/°C
f_{th}	Facteur de correction thermique	–
f_{system}	Facteur de pertes système	–
NOCT	Nominal Operating Cell Temperature	°C
Y_{spec}	Rendement énergétique spécifique	kWh/kWc/an
K_T	Indice de clarté du ciel	–

Notations statistiques

Symbole	Signification	Unité
n	Nombre d'observations (sites)	–
p	Nombre de variables explicatives	–
y_i	Valeur observée de la cible pour le site i	kWh/kWc/an
$\hat{y}_i$	Valeur prédite par le modèle	kWh/kWc/an
$\bar{y}$	Moyenne des valeurs réelles	kWh/kWc/an
e_i	Résidu : $e_i = y_i - \hat{y}_i$	kWh/kWc/an
$\mathbf{x}_i$	Vecteur des variables explicatives, site i	–
$\mathbf{X}$	Matrice des variables explicatives	–
$\boldsymbol{\beta}$	Vecteur des coefficients de régression	–
$\hat{\boldsymbol{\beta}}$	Vecteur des coefficients estimés	–
ε_i	Terme d'erreur	–
σ	Écart-type	–
σ^2	Variance	–
r	Coefficient de corrélation de Pearson	–
ρ_s	Coefficient de corrélation de Spearman	–
λ	Paramètre de régularisation Ridge	–
VIF_j	Facteur d'inflation de variance	–
γ_1	Coefficient d'asymétrie (skewness)	–
γ_2	Coefficient d'aplatissement (kurtosis)	–

Métriques d'évaluation

Symbole	Signification	Unité
MAE	Erreur absolue moyenne	kWh/kWc/an
RMSE	Racine de l'erreur quadratique moyenne	kWh/kWc/an
MAPE	Erreur absolue en pourcentage	%
R^2	Coefficient de détermination	–
SCT	Somme des carrés totale	–
SCR	Somme des carrés des résidus	–

Liste des abréviations

ACP	Analyse en Composantes Principales	**MCO**	Moindres Carrés Ordinaires
BLUE	Best Linear Unbiased Estimator	**ML**	Machine Learning
CDER	Centre de Développement des Énergies Renouvelables	**NASA**	National Aeronautics and Space Administration
CSI	Clear Sky Index (Indice de clarté)	**NOCT**	Nominal Operating Cell Temperature
CV	Coefficient de Variation	**ONM**	Office National de la Météorologie
DHI	Diffuse Horizontal Irradiance	**POWER**	Prediction Of Worldwide Energy Resources
DNI	Direct Normal Irradiance	**PV**	Photovoltaïque
EDA	Exploratory Data Analysis	$\mathbf{R^2}$	Coefficient de détermination
ENR	Énergies Renouvelables	**RMSE**	Root Mean Squared Error
ERA5	Réanalyse climatique européenne (Copernicus)	**STC**	Standard Test Conditions
GHI	Global Horizontal Irradiance	**VIF**	Variance Inflation Factor
IEA	International Energy Agency	$\mathbf{Y_{spec}}$	Rendement énergétique spécifique
IQR	Interquartile Range		
IRENA	International Renewable Energy Agency		
MAE	Mean Absolute Error		
MAPE	Mean Absolute Percentage Error		

Introduction générale

Le système énergétique mondial est engagé depuis le début du XXI^{e} siècle dans une mutation profonde. Sous la double pression de l'urgence climatique et de la raréfaction progressive des ressources fossiles, les gouvernements et les opérateurs de réseau accélèrent le déploiement des énergies renouvelables à une échelle sans précédent. En 2023, les sources renouvelables représentaient déjà 30 % de la production mondiale d'électricité, et les projections de l'Agence Internationale de l'Énergie anticipent une proportion dépassant 60 % à l'horizon 2050 (International Energy Agency 2022). Au sein de cet ensemble, l'énergie solaire photovoltaïque connaît la croissance la plus spectaculaire, portée par la baisse continue des coûts (International Renewable Energy Agency 2023) et l'amélioration des rendements de conversion.

Cette transition revêt une dimension particulièrement stratégique pour l'Algérie. Le pays dispose d'un potentiel solaire exceptionnel : avec une superficie désertique de plus de deux millions de kilomètres carrés et une énergie solaire reçue (globale horizontale) variant entre 1 700 et 2 650 kWh/m^2/an selon les régions, le Sahara algérien figure parmi les gisements solaires les plus importants au monde (Khefifi et Soufiane 2019). Face à cette richesse naturelle, et face à la nécessité de diversifier un mix électrique reposant aujourd'hui à près de 99 % sur les hydrocarbures, le programme national des énergies renouvelables s'est fixé l'objectif ambitieux de **15 000 MW de capacité installée renouvelable à l'horizon 2035**, dont la majorité en solaire photovoltaïque.

La concrétisation de cet objectif soulève des défis techniques et économiques considérables. L'un des premiers concerne le **choix des sites d'implantation** des futures centrales : il s'agit d'identifier, parmi les nombreuses régions candidates du territoire algérien, celles qui combinent un potentiel solaire élevé et des conditions favorables au déploiement.

L'évaluation rigoureuse du potentiel d'un site nécessite traditionnellement des campagnes de mesures climatiques étalées sur plusieurs années, ce qui ralentit considérablement le processus de sélection et augmente les coûts de pré-étude. Il existe donc un besoin opérationnel pour des **outils statistiques d'estimation rapide** qui, à partir de données climatologiques moyennes facilement disponibles (issues de bases satellitaires comme NASA POWER ou ERA5), permettent de prédire le rendement énergétique attendu d'une installation et de hiérarchiser les sites candidats.

Le problème se formule mathématiquement de la façon suivante : à partir d'un échantillon de n sites algériens caractérisés par un vecteur de variables géographiques et climatiques $\mathbf{x} \in \mathbb{R}^p$ (latitude, longitude, altitude, énergie solaire reçue, température, humidité, vent, couverture nuageuse), il s'agit de construire une fonction prédictive $\hat{f}$ telle que :

$$\hat{y} = \hat{f}(\mathbf{x})$$

estime au mieux le rendement énergétique annuel y (en kWh/kWc/an) d'une installation photovoltaïque de référence implantée sur ce site. Ce problème, formellement appelé **régression supervisée sur observations indépendantes**, constitue le cœur du présent mémoire.

La question centrale qui guide ce travail peut être formulée ainsi :

> *Comment construire un modèle statistique simple, interprétable et reproductible permettant d'estimer le potentiel solaire d'un site algérien à partir de ses caractéristiques géographiques et climatiques, afin d'aider à la décision pour le dimensionnement de futurs parcs photovoltaïques ?*

Pour répondre à cette question, ce mémoire poursuit quatre objectifs complémentaires.

Le **premier objectif** est de poser le cadre physique et conceptuel du problème : présenter les principes de la conversion photovoltaïque, identifier les variables qui déterminent le rendement énergétique d'un site, et préciser comment ce rendement peut être quantifié à partir des caractéristiques climatiques disponibles.

Le **deuxième objectif** est de construire un jeu de données représentatif des conditions algériennes, calibré sur les paramètres climatologiques publics, et de conduire une analyse exploratoire approfondie : statistiques descriptives, distributions, matrices de corrélation, détection des valeurs aberrantes et analyse de la multicolinéarité.

Le **troisième objectif** est de présenter rigoureusement les deux méthodes de régression retenues — la régression linéaire multiple et la régression Ridge — en exposant leurs formalismes mathématiques, leurs hypothèses de validité, leurs propriétés statistiques et le protocole de validation associé.

Le **quatrième objectif** est d'appliquer ces méthodes à l'échantillon de sites algériens, d'évaluer leurs performances sur un jeu de test indépendant et de discuter les implications opérationnelles des résultats obtenus pour la sélection de futurs sites photovoltaïques.

La démarche adoptée s'inscrit dans la méthodologie classique des projets de science des données, adaptée à un problème de régression supervisée. Elle suit une progression logique en cinq étapes.

Étape 1 : cadrage Délimitation du problème, identification des variables pertinentes, formulation mathématique de la cible.

Étape 2 : construction et nettoyage des données Collecte des sources, génération du jeu de données, traitement des valeurs manquantes et des valeurs aberrantes, vérification de la cohérence physique.

Étape 3 : analyse exploratoire Calcul des statistiques descriptives, analyse des corrélations, détection de la multicolinéarité, préparation des variables explicatives (standardisation, variables dérivées).

Étape 4 : modélisation Découpage chronologique entraînement/test, ajustement des deux modèles (régression linéaire et régression Ridge), sélection de l'hyperparamètre λ par validation croisée.

Étape 5 : évaluation Calcul des métriques de performance sur le jeu de test, comparaison des modèles, analyse des coefficients estimés et discussion des résultats.

L'ensemble du traitement et des expériences a été réalisé en **Python 3.10**, en mobilisant les bibliothèques scientifiques les plus répandues : `pandas` et `numpy` pour la manipulation des données, `matplotlib` et `seaborn` pour la visualisation, `scipy` et `statsmodels` pour les analyses statistiques, et `scikit-learn` (Pedregosa et al. 2011) pour les modèles de régression (McKinney 2017). Les scripts complets sont fournis en annexe et disponibles pour reproduction.

Ce mémoire est organisé en quatre chapitres qui correspondent aux grandes étapes de la démarche.

Le Chapitre 1 — Énergie solaire et potentiel photovoltaïque en Algérie pose les fondements physiques et contextuels du travail. Il présente le contexte énergétique mondial et algérien, les principes de la conversion photovoltaïque avec le modèle physique reliant l'énergie solaire reçue et la température au rendement, les variables qui déterminent le potentiel solaire d'un site, et le programme algérien des énergies renouvelables qui constitue le cadre stratégique de la problématique.

Le Chapitre 2 — Construction du jeu de données et analyse exploratoire décrit la démarche méthodologique adoptée. Il présente les sources de données disponibles (NASA POWER, ERA5, stations ONM, Atlas CDER), justifie le choix d'un jeu de données synthétique pour ce mémoire, expose les opérations de nettoyage et de prétraitement, et conduit l'analyse statistique exploratoire qui caractérise les distributions et les relations entre variables.

Le Chapitre 3 — Méthodes de régression présente les deux méthodes retenues. Il développe d'abord rigoureusement la régression linéaire multiple, depuis la

formulation matricielle jusqu'à la solution analytique des moindres carrés et le théorème de Gauss-Markov, en passant par les hypothèses de validité. Il introduit ensuite la régression Ridge comme réponse à la multicolinéarité, en explicitant le compromis biais-variance et la sélection du paramètre λ par validation croisée. Il conclut par les métriques d'évaluation retenues.

Le Chapitre 4 — Application aux sites algériens et analyse des résultats constitue le cœur expérimental du mémoire. Il applique l'ensemble du pipeline à un jeu de 200 sites algériens représentatifs, présente les statistiques descriptives effectives, les corrélations observées, les coefficients estimés des deux modèles, les performances sur le jeu de test, et discute les implications opérationnelles pour la sélection de futurs sites photovoltaïques.

Le mémoire se conclut par une conclusion générale qui synthétise les principaux apports du travail et trace les perspectives de recherche qu'il ouvre.

Chapitre 1

Énergie solaire et potentiel photovoltaïque en Algérie

Introduction

L'énergie solaire constitue aujourd'hui l'un des piliers majeurs de la transition énergétique mondiale. Avec des coûts qui ont chuté de plus de 80 % en une décennie (International Renewable Energy Agency 2023) et une accessibilité qui s'étend désormais à toutes les régions du globe, la conversion photovoltaïque est passée du statut de technologie marginale à celui de filière industrielle de premier plan. Pour un pays comme l'Algérie, dont le territoire offre l'un des gisements solaires les plus importants au monde, l'enjeu dépasse la simple substitution énergétique : il s'agit d'une opportunité historique de diversification économique et d'affirmation technologique.

Ce premier chapitre établit le cadre général dans lequel s'inscrit le travail. Il présente d'abord le contexte énergétique mondial et algérien, puis expose les principes physiques de la conversion photovoltaïque qui relient les variables climatiques au rendement énergétique d'une installation. Il décrit ensuite la méthodologie d'évaluation du potentiel solaire d'un site et les variables qui en déterminent la valeur. Il conclut par une présentation du programme national algérien des énergies renouvelables, qui constitue le cadre stratégique de la problématique étudiée.

1 Contexte énergétique : transition mondiale et défis algériens

1.1 La transition énergétique mondiale

Depuis le début du XXI^{e} siècle, le système énergétique mondial est engagé dans une mutation profonde. Sous la double pression de l'urgence climatique et de la raréfaction

progressive des ressources fossiles, les gouvernements et les opérateurs de réseau accélèrent le déploiement des énergies renouvelables. En 2023, les sources renouvelables représentaient déjà 30 % de la production mondiale d'électricité, et les projections de l'Agence Internationale de l'Énergie anticipent une proportion dépassant 60 % à l'horizon 2050 (International Energy Agency 2022).

Au sein de cet ensemble, l'**énergie solaire photovoltaïque** occupe une place particulière. Sa croissance, plus rapide que celle de toute autre filière énergétique dans l'histoire, est portée par trois facteurs convergents : la baisse spectaculaire du coût des modules, l'amélioration continue des rendements de conversion, et la flexibilité de déploiement des installations, des grandes centrales aux toitures résidentielles.

1.2 La situation énergétique algérienne

L'Algérie présente un profil énergétique singulier qui rend la transition particulièrement stratégique. D'un côté, le pays possède d'abondantes ressources en hydrocarbures, qui constituent à la fois la principale source de revenus de l'État et le pilier du mix électrique national : en 2022, près de 99 % de la production d'électricité algérienne reposait encore sur le gaz naturel. De l'autre côté, cette dépendance expose le système énergétique à une double vulnérabilité : la volatilité des prix internationaux des hydrocarbures et l'épuisement progressif des réserves prouvées.

Pour répondre à ces défis, et pour préserver l'avenir énergétique du pays, les autorités algériennes ont engagé un effort de diversification dont les énergies renouvelables — et le solaire en premier lieu — constituent l'axe central. Le tableau 1.1 résume les indicateurs clés du contexte énergétique algérien.

Table 1.1 – Indicateurs énergétiques de l'Algérie (2022)

Indicateur	**Valeur**
Capacité électrique installée totale	$\sim 24\,000$ MW
Part des hydrocarbures dans la production	$\sim 99\,\%$
Part des énergies renouvelables	$\sim 1\,\%$
Capacité solaire installée	~ 450 MW
Objectif renouvelable à horizon 2035	15 000 MW
Énergie solaire reçue moyenne (Sahara)	5,5 – 6,5 kWh/m²/jour
Surface désertique	$> 2\,000\,000$ km²

1.3 Le potentiel solaire algérien

L'Algérie dispose d'un potentiel solaire qui figure parmi les plus importants au monde (Khefifi et Soufiane 2019). Cette richesse naturelle résulte de la conjonction de plusieurs facteurs géographiques :

— Une **position géographique** entre les latitudes 19° et 37° N, qui place une grande partie du territoire dans la ceinture solaire ;
— Une **vaste superficie désertique** de plus de 2 millions de km², caractérisée par un ciel particulièrement dégagé et un faible taux d'humidité ;
— Un **ensoleillement annuel** qui dépasse 3 500 heures dans le Sahara central, contre moins de 2 000 heures dans la plupart des pays européens.

L'énergie solaire reçue (globale horizontale) annuelle varie ainsi entre 1 700 kWh/m²/an dans le Nord côtier et 2 650 kWh/m²/an dans le Sahara profond. Cette diversité régionale, conjuguée à la grande étendue du territoire, ouvre un champ d'application considérable pour le déploiement de centrales photovoltaïques. Encore faut-il être capable d'**évaluer quantitativement, site par site**, le potentiel énergétique réellement exploitable, ce qui constitue précisément la problématique de ce travail.

2 Principes de la conversion photovoltaïque

2.1 L'effet photovoltaïque

La conversion photovoltaïque repose sur un phénomène physique connu depuis le XIX^{e} siècle : l'**effet photovoltaïque**, décrit pour la première fois par Becquerel en 1839 et expliqué théoriquement par Einstein en 1905. Le principe est le suivant : lorsqu'un rayonnement lumineux frappe un matériau semi-conducteur dopé (typiquement du silicium cristallin), les photons transmettent leur énergie aux électrons du matériau, qui peuvent alors franchir la bande interdite et participer à la conduction électrique. La présence d'une jonction P-N à l'intérieur du semi-conducteur crée un champ électrique permanent qui oriente le déplacement des électrons libres, donnant naissance à un courant électrique continu.

2.2 Caractéristiques d'un module photovoltaïque

Un module photovoltaïque est caractérisé principalement par :

— Sa **puissance crête** P_{STC}, exprimée en watts-crête (Wc), mesurée dans les *Standard Test Conditions* : énergie solaire reçue de 1 000 W/m², température de cellule de 25 °C et masse d'air AM = 1,5 ;
— Son **rendement nominal** η_{nom}, généralement compris entre 17 % et 22 % pour les modules en silicium cristallin modernes ;
— Son **coefficient thermique de puissance** γ, typiquement de l'ordre de $-0{,}004$ %/°C, qui traduit la diminution du rendement avec la température ;
— Sa **surface** A exprimée en m².

2.3 Modèle de production d'une installation photovoltaïque

La puissance instantanée produite par une installation photovoltaïque dépend de plusieurs facteurs combinés. Le modèle physique le plus simple mais largement utilisé (Duffie et Beckman 2013) exprime la puissance produite P comme :

$$P \;=\; G \times A \times \eta_{\text{nom}} \times f_{\text{th}}(T_c) \times f_{\text{system}}$$

où :

- G est l'**énergie solaire reçue** incidente sur le plan des modules ($\mathrm{W/m^2}$) ;
- A est la surface totale des modules ($\mathrm{m^2}$) ;
- η_{nom} est le rendement nominal (sans dimension) ;
- $f_{\text{th}}(T_c)$ est un **facteur de correction thermique** qui décroît avec la température de cellule T_c ;
- f_{system} est un **facteur de pertes système** qui regroupe les pertes liées aux câbles, à l'onduleur, au salissement, aux ombrages, etc. (typiquement entre 0,75 et 0,90).

Le facteur thermique $f_{\text{th}}(T_c)$ s'écrit explicitement (Skoplaki et Palyvos 2009) :

$$f_{\text{th}}(T_c) \;=\; 1 + \gamma\,(T_c - T_{\text{ref}})$$

avec $T_{\text{ref}} = 25\,^{\circ}\mathrm{C}$ la température de référence des conditions standard et γ le coefficient thermique de puissance. La température de cellule T_c est elle-même reliée à la température ambiante T_{amb} et à l'énergie solaire reçue par la relation :

$$T_c \;=\; T_{\text{amb}} + \frac{\text{NOCT} - 20}{800}\,G$$

où NOCT (*Nominal Operating Cell Temperature*) est une caractéristique du module, généralement comprise entre 43 °C et 48 °C.

2.4 Production énergétique annuelle

L'énergie totale produite sur une année par une installation est l'intégrale temporelle de sa puissance instantanée :

$$E_{\text{annuelle}} \;=\; \int_0^{T_{\text{an}}} P(t)\,\mathrm{d}t$$

En pratique, cette intégrale est approchée par une somme discrète sur les pas de temps disponibles (heure ou jour). Pour l'évaluation comparative de différents sites, on utilise la grandeur normalisée appelée **rendement énergétique spécifique** (ou *specific*

yield) :

$$Y_{\text{spec}} = \frac{E_{\text{annuelle}}}{P_{\text{nom}}} \quad \text{(unité : kWh/kWc/an)}$$

où P_{nom} est la puissance nominale totale de l'installation. Cette grandeur, indépendante de la taille de l'installation, permet de comparer directement la productivité énergétique de sites différents et constitue la **variable cible** de l'étude présentée dans ce mémoire.

3 Évaluation du potentiel solaire d'un site

3.1 Variables explicatives du rendement

Le rendement énergétique annuel d'un site dépend de nombreuses variables qu'il est essentiel d'identifier précisément. Ces variables peuvent être regroupées en trois catégories.

Variables géographiques

- **Latitude** (φ) : détermine l'angle moyen d'incidence solaire et la durée du jour selon les saisons. Plus la latitude est élevée, plus l'angle solaire au midi est faible et plus la production hivernale diminue.
- **Longitude** (λ) : moins influente directement sur le rendement, mais corrélée aux conditions climatiques régionales.
- **Altitude** (z) : à altitude élevée, l'atmosphère est moins épaisse, ce qui réduit l'absorption atmosphérique du rayonnement et augmente l'énergie solaire reçue. La température décroît également avec l'altitude selon un gradient adiabatique de l'ordre de $-6{,}5\,°\text{C/km}$, ce qui améliore le rendement thermique des modules.

Variables climatiques

- **Énergie solaire reçue (globale horizontale)** (GHI) : variable de premier ordre, exprimée en kWh/m²/jour ou en kWh/m²/an. C'est la quantité d'énergie solaire reçue par unité de surface horizontale au sol.
- **Température ambiante moyenne** : affecte directement le rendement thermique des modules (effet f_{th}).
- **Humidité relative** : une humidité élevée s'accompagne généralement d'une atténuation de l'énergie solaire reçue par diffusion atmosphérique.
- **Couverture nuageuse moyenne** : facteur principal de réduction du rayonnement reçu au sol.
- **Précipitations annuelles** : indicateur indirect de la fréquence des journées couvertes.

— **Vitesse du vent** : facteur de refroidissement convectif des modules, qui améliore légèrement leur rendement.

Variables dérivées

À partir des variables précédentes, plusieurs indicateurs synthétiques peuvent être construits :

— L'**indice de clarté du ciel** (*clearness index*), défini comme le rapport entre l'énergie solaire reçue au sol et l'énergie solaire reçue extraterrestre théorique. Cet indice quantifie la transparence atmosphérique et est un excellent prédicteur du rendement.

— Le **nombre annuel d'heures de soleil** (*sunshine hours*), obtenu par intégration des intervalles d'énergie solaire reçue dépassant un seuil de référence (typiquement 120 W/m^2).

— L'**albédo de surface** : fraction du rayonnement réfléchie par le sol, qui contribue à l'énergie solaire reçue globale reçue par les modules inclinés.

3.2 Méthodes d'évaluation existantes

Plusieurs approches coexistent pour évaluer le potentiel solaire d'un site.

Approche par modèle physique Cette approche consiste à appliquer directement les équations physiques de la conversion photovoltaïque (présentées ci-dessus) à des données climatiques détaillées du site. Elle est précise mais nécessite des données horaires complètes sur une année entière, et reste lourde à mettre en œuvre pour évaluer un grand nombre de sites.

Approche par cartographie satellite Des organismes comme la NASA (programme POWER) ou l'Agence Spatiale Européenne (programme Copernicus) fournissent des cartes d'énergie solaire reçue construites à partir d'observations satellitaires. Cette approche permet une couverture géographique exhaustive, mais ses estimations restent indicatives et doivent être validées par des mesures au sol.

Approche par modélisation statistique C'est l'approche retenue dans ce travail. Elle consiste à ajuster, à partir d'un échantillon de sites pour lesquels le rendement énergétique est connu (par mesure ou par calcul physique précis), un modèle statistique reliant ce rendement à un ensemble de variables explicatives plus aisément disponibles. Une fois le modèle validé, il peut être appliqué à de nouveaux sites pour estimer rapidement leur potentiel, sans nécessiter de mesures détaillées sur place.

L'intérêt principal de cette approche, qui constitue la contribution méthodologique de ce mémoire, est qu'elle fournit un outil simple et rapide d'**aide à la décision** pour le pré-dimensionnement et la sélection des sites favorables au déploiement de futures installations photovoltaïques.

4 Le programme algérien des énergies renouvelables

4.1 Cadre stratégique

L'Algérie a adopté en 2011 un premier plan national de développement des énergies renouvelables, révisé et amplifié à plusieurs reprises depuis. Le programme actuel (Stambouli, Khiat et al. 2012) fixe un objectif de **15 000 MW de capacité renouvelable installée à l'horizon 2035**, dont la majorité (environ 13 500 MW) serait réalisée en solaire photovoltaïque. Cet objectif représente une multiplication par trente de la capacité actuellement installée.

4.2 Enjeux du dimensionnement et de la sélection des sites

La concrétisation de cet objectif soulève des défis techniques et économiques considérables. L'un des premiers est le choix des **sites d'implantation** des futures centrales : il s'agit d'identifier, parmi les nombreuses régions candidates du territoire algérien, celles qui combinent un potentiel solaire élevé, une infrastructure de raccordement disponible et un coût d'installation maîtrisé.

Or, l'évaluation rigoureuse du potentiel d'un site nécessite traditionnellement des campagnes de mesures climatiques étalées sur plusieurs années, ce qui ralentit considérablement le processus de sélection. Il existe donc un besoin opérationnel pour des **outils statistiques d'estimation rapide** qui, à partir des données climatologiques moyennes facilement disponibles (issues de bases de données satellitaires ou de stations météorologiques régionales), permettent de prédire le rendement énergétique attendu et de hiérarchiser les sites candidats.

4.3 Positionnement du présent travail

C'est précisément à ce besoin que ce mémoire cherche à apporter une contribution. En construisant un modèle statistique capable de prédire, à partir des variables géographiques et climatiques d'un site, le rendement énergétique annuel d'une installation photovoltaïque, le présent travail propose un outil simple et reproductible d'aide à la décision pour le dimensionnement de futurs parcs solaires en Algérie. Les chapitres suivants détaillent successivement :

— le traitement et l'analyse exploratoire des données géographiques et climatiques (Chapitre 2) ;
— la présentation des méthodes de régression retenues — régression linéaire multiple et régression Ridge (Chapitre 3) ;
— l'application de ces méthodes à un échantillon représentatif de sites algériens et l'interprétation des résultats (Chapitre 4).

Conclusion

Ce premier chapitre a posé le cadre général du travail. Il a montré comment l'enjeu de la transition énergétique mondiale rencontre, en Algérie, un potentiel solaire exceptionnel et une volonté politique affirmée de développement des énergies renouvelables, matérialisée par le programme des 15 000 MW à l'horizon 2035. Il a ensuite exposé les principes physiques de la conversion photovoltaïque, en mettant en évidence les variables géographiques et climatiques qui déterminent le rendement énergétique d'un site. Il a enfin présenté la problématique spécifique du présent mémoire : construire un **modèle statistique de prédiction du rendement** qui, à partir de données facilement disponibles, fournisse un outil rapide d'aide à la décision pour le choix des sites candidats au déploiement de futures installations photovoltaïques.

Le chapitre suivant détaille la première étape de cette construction : le traitement et l'analyse exploratoire des données géographiques et climatiques sur lesquelles repose l'ensemble du travail.

Chapitre 2

Construction du jeu de données et analyse exploratoire

Introduction

La qualité d'un modèle prédictif est fondamentalement limitée par la qualité des données sur lesquelles il est construit. Avant toute étape de modélisation, il est donc indispensable de soigner la collecte des données, de les nettoyer, de comprendre leur structure et d'en analyser les propriétés statistiques. Cette démarche, communément désignée sous le nom d'**analyse exploratoire des données** (*Exploratory Data Analysis*, EDA), constitue l'objet du présent chapitre.

Ce chapitre est structuré en cinq étapes successives. La première décrit les sources de données disponibles pour évaluer le potentiel solaire des sites algériens. La deuxième présente la **démarche méthodologique** adoptée dans ce mémoire pour construire un jeu de données représentatif. La troisième conduit les opérations de nettoyage et de prétraitement. La quatrième développe une analyse statistique exploratoire — statistiques descriptives, distributions, matrices de corrélation. La cinquième prépare les variables explicatives qui alimenteront les modèles de régression développés au chapitre suivant.

1 Sources de données pour l'évaluation du potentiel solaire

Plusieurs sources de données sont mobilisables pour évaluer le potentiel solaire d'un site. Elles présentent des caractéristiques complémentaires en termes de couverture spatiale, de précision et d'accessibilité.

1.1 Bases de données satellitaires globales

Les organismes spatiaux internationaux fournissent gratuitement des données climatologiques globales construites à partir d'observations satellitaires combinées à des modèles atmosphériques. Les principales sont :

— **NASA POWER** (*Prediction of Worldwide Energy Resources*) : base de données développée par la NASA, qui fournit pour chaque coordonnée géographique du globe les valeurs mensuelles et annuelles d'énergie solaire reçue, de température, d'humidité et de vent depuis 1981. C'est l'une des sources les plus utilisées dans les études sur l'énergie renouvelable (NASA Langley Research Center 2023).

— **Copernicus** (programme européen) : fournit notamment la réanalyse climatique ERA5 (Hersbach et al. 2020), base de référence pour les variables météorologiques sur l'ensemble du globe à fine résolution spatiale.

— **Solargis** et **Global Solar Atlas** : services spécialisés dans l'évaluation de la ressource solaire à fine échelle.

1.2 Sources nationales algériennes

Office National de la Météorologie (ONM) L'ONM dispose d'un réseau de stations qui mesurent en continu les variables climatiques (température, humidité, vent, précipitations) sur l'ensemble du territoire algérien. Ces données, plus précises que les données satellitaires, sont cependant disponibles en nombre limité de points et nécessitent des procédures d'accès spécifiques.

Centre de Développement des Énergies Renouvelables (CDER) Le CDER (Centre de Développement des Énergies Renouvelables 2020), organisme algérien de référence sur les énergies renouvelables, a publié plusieurs cartes du rayonnement solaire sur le territoire algérien, construites à partir d'un croisement entre données satellitaires et mesures au sol. L'*Atlas Solaire Algérien*, publié pour la première fois en 1985 et actualisé depuis, en constitue la référence.

1.3 Synthèse comparative des sources

TABLE 2.1 – Comparaison des sources de données disponibles

Source	Couverture	Précision	Accessibilité
NASA POWER	Mondiale, 0,5ř × 0,5ř	Modérée	Libre, en ligne
ERA5/Copernicus	Mondiale, 0,25ř × 0,25ř	Élevée	Libre (inscription)
Stations ONM	Sites discrets en Algérie	Très élevée	Procédure d'accès
Atlas CDER	Algérie, cartes thématiques	Élevée	Publications officielles

2 Démarche méthodologique : un jeu de données synthétique

2.1 Choix méthodologique et justification

Dans le cadre de ce mémoire, et compte tenu des contraintes de temps et d'accès aux bases de données opérationnelles, nous avons fait le **choix méthodologique de construire un jeu de données synthétique** représentatif des conditions climatiques algériennes. Ce choix mérite d'être explicitement justifié, car il conditionne l'interprétation des résultats obtenus dans les chapitres suivants.

L'**objectif principal** de ce mémoire n'est pas de produire des estimations opérationnelles du potentiel solaire de sites spécifiques — travail qui relèverait d'une étude d'ingénierie appliquée — mais de **développer et valider une démarche méthodologique** permettant de construire un modèle statistique de prédiction du rendement énergétique à partir de variables géographiques et climatiques. Pour cet objectif, un jeu de données synthétique présente plusieurs avantages.

Avantages de l'approche synthétique

1. **Reproductibilité** : le jeu de données peut être régénéré à tout moment à partir des paramètres publiés, ce qui garantit la reproductibilité scientifique du travail.
2. **Cohérence physique** : la variable cible étant calculée à partir du modèle physique de conversion photovoltaïque (Chapitre 1), sa relation avec les variables explicatives respecte par construction les lois physiques connues, ce qui permet une validation rigoureuse de la démarche statistique.

3. **Représentativité géographique** : la construction permet de garantir une couverture homogène de toutes les régions climatiques algériennes, ce qui n'est pas toujours possible avec les données opérationnelles concentrées sur quelques sites instrumentés.
4. **Maîtrise des paramètres** : les valeurs sont calibrées sur les données climatologiques publiques (NASA POWER, Atlas CDER), ce qui assure leur réalisme tout en évitant les imperfections inhérentes aux mesures réelles (lacunes, biais instrumentaux).

Limites assumées Cette approche présente naturellement des limites qu'il est essentiel d'expliciter. Les chiffres précis obtenus n'ont pas vocation à être considérés comme des valeurs opérationnelles pour des sites spécifiques. La généralisation à des données réelles constitue une étape ultérieure qui sortirait du cadre de ce mémoire et fait partie des perspectives discutées dans la conclusion générale.

2.2 Construction du jeu de données

Le jeu de données utilisé dans ce mémoire a été construit selon le protocole suivant.

Étape 1 : sélection des sites Un échantillon de **200 sites** est généré sur le territoire algérien, avec des coordonnées (latitude, longitude) tirées de manière à assurer une couverture représentative des quatre grandes régions climatiques :

— Zone côtière méditerranéenne (Nord, latitudes 35°–37°N) ;
— Hauts plateaux (latitudes 33°–35°N) ;
— Sahara septentrional (latitudes 28°–33°N) ;
— Sahara central et méridional (latitudes 19°–28°N).

Étape 2 : génération des variables géographiques Pour chaque site, l'altitude est attribuée selon un modèle géographique simplifié reflétant le relief du territoire algérien : altitudes faibles sur la côte, plus élevées sur les hauts plateaux, variées dans le Sahara (plaines et massifs).

Étape 3 : génération des variables climatiques Les variables climatiques (énergie solaire reçue, température, humidité, vent, couverture nuageuse) sont calibrées sur les valeurs moyennes annuelles publiées par la NASA POWER pour chaque coordonnée, avec ajout d'une variabilité aléatoire gaussienne de faible amplitude pour assurer la diversité de l'échantillon.

Étape 4 : calcul de la variable cible Le rendement énergétique spécifique Y_{spec} est calculé pour chaque site à partir du modèle physique présenté au Chapitre 1, en considérant une installation photovoltaïque de référence aux caractéristiques standard :

— Rendement nominal $\eta_{\text{nom}} = 0{,}18$;
— Coefficient thermique $\gamma = -0{,}004$ %/°C ;
— Facteur système $f_{\text{system}} = 0{,}80$;
— NOCT = 45 °C.

2.3 Description du jeu de données obtenu

Le jeu de données ainsi constitué comprend $n = 200$ sites, $p = 8$ variables explicatives et 1 variable cible, comme résumé dans le tableau 2.2.

TABLE 2.2 – Description des variables du jeu de données

Variable	Description	Unité	Plage attendue
Variables géographiques			
`latitude`	Latitude	°	[19 ; 37]
`longitude`	Longitude	°	[−9 ; 12]
`altitude`	Altitude	m	[0 ; 2 500]
Variables climatiques			
`irradiance`	Énergie solaire reçue moyenne	kWh/m²/jour	[4,5 ; 7,5]
`temperature`	Température moyenne	°C	[10 ; 30]
`humidity`	Humidité relative moyenne	%	[10 ; 80]
`wind_speed`	Vitesse moyenne du vent	m/s	[1,5 ; 8]
`cloud_cover`	Couverture nuageuse moyenne	%	[5 ; 60]
Variable cible			
`yield_kwh`	Rendement énergétique spécifique	kWh/kWc/an	[1 300 ; 2 000]

3 Nettoyage et prétraitement des données

Bien que le jeu de données soit construit de façon contrôlée, les étapes de nettoyage restent indispensables pour deux raisons : d'une part, pour garantir la cohérence physique de l'ensemble des observations ; d'autre part, pour valider et illustrer une démarche méthodologique transférable à des données réelles.

3.1 Vérification de la qualité des données

Trois aspects sont examinés successivement : la complétude (présence de valeurs manquantes), la cohérence physique (respect des plages admissibles) et la cohérence

statistique (détection des valeurs aberrantes).

3.2 Traitement des valeurs manquantes

Dans la procédure de génération, une faible proportion ($\sim 3\,\%$) de valeurs manquantes a été introduite volontairement pour simuler les imperfections rencontrées dans les jeux de données réels et illustrer les techniques d'imputation.

Stratégie retenue : imputation par la médiane régionale Pour chaque variable comportant des valeurs manquantes, l'imputation est réalisée par la **médiane des sites de la même région climatique**. Cette stratégie présente plusieurs avantages :

- Elle préserve les caractéristiques régionales — un site saharien ne reçoit pas une valeur typique d'un site côtier ;
- La médiane est plus robuste que la moyenne en présence d'outliers ;
- La méthode est simple et reproductible.

3.3 Vérification des contraintes physiques

Chaque variable doit respecter une plage physiquement admissible. Pour le jeu de données algérien, les bornes vérifiées sont celles données dans le tableau 2.2. Toute valeur dépassant ces bornes est ramenée à la borne la plus proche (*capping*), une opération qui préserve l'observation tout en éliminant les valeurs manifestement erronées.

3.4 Détection et traitement des valeurs aberrantes

Les **valeurs aberrantes** (*outliers*) sont des observations qui s'écartent significativement du comportement attendu. Leur détection est essentielle car elles peuvent biaiser sensiblement les estimateurs des modèles de régression.

Méthode de l'intervalle interquartile (IQR)

L'intervalle interquartile $\mathrm{IQR} = Q_3 - Q_1$, où Q_1 et Q_3 sont les premier et troisième quartiles, permet de définir des bornes de détection :

$$b_{\mathrm{inf}} = Q_1 - 1{,}5 \times \mathrm{IQR}, \qquad b_{\mathrm{sup}} = Q_3 + 1{,}5 \times \mathrm{IQR}$$

Toute observation x vérifiant $x < b_{\mathrm{inf}}$ ou $x > b_{\mathrm{sup}}$ est considérée comme potentiellement aberrante. Cette méthode est robuste aux distributions asymétriques.

Score Z standardisé

Le score Z mesure l'écart d'une observation à la moyenne en unités d'écart-type :

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

Une observation est considérée aberrante si $|z_i| > 3$, ce qui correspond à une probabilité d'occurrence inférieure à 0,3 % sous une distribution gaussienne.

Stratégie de traitement

Pour ce jeu de données, peu d'outliers sont détectés. Ces valeurs sont examinées au cas par cas : lorsque la valeur correspond à un site géographique réaliste mais extrême (par exemple, un site de haute altitude isolé), l'observation est conservée ; lorsque la valeur résulte d'une incohérence manifeste, elle est remplacée par la médiane régionale.

4 Analyse statistique exploratoire

4.1 Statistiques descriptives

Les statistiques descriptives résument les principales caractéristiques de chaque variable. On calcule pour chacune :

- Les **indicateurs de position** : moyenne $\bar{x}$, médiane $\tilde{x}$, premier et troisième quartiles Q_1 et Q_3.
- Les **indicateurs de dispersion** : écart-type σ, intervalle interquartile IQR, coefficient de variation $\mathrm{CV} = \sigma/\bar{x}$.
- Les **indicateurs de forme** : asymétrie (*skewness*) γ_1 et aplatissement (*kurtosis*) γ_2.

L'asymétrie est définie par :

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3$$

Elle est nulle pour une distribution symétrique, positive pour une distribution étalée vers la droite et négative dans le cas inverse.

L'aplatissement est défini par :

$$\gamma_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^4 - 3$$

Il est nul pour une distribution gaussienne, positif pour une distribution à queues lourdes et négatif pour une distribution aplatie.

Présentation des résultats Les statistiques descriptives complètes seront présentées et discutées dans le Chapitre 4, à l'issue de la génération effective du jeu de données. À titre indicatif, les ordres de grandeur attendus, calibrés sur les données climatologiques algériennes publiées par la NASA POWER, sont les suivants :

— Énergie solaire reçue moyenne sur l'ensemble du territoire : entre 5 et 6 kWh/m²/jour, avec un pic dans le Sahara central ;

— Température moyenne annuelle : entre 17 et 25 °C selon les régions ;

— Rendement énergétique spécifique attendu : entre 1 400 et 1 950 kWh/kWc/an, avec les valeurs les plus élevées dans le Sahara.

4.2 Analyse des corrélations

Corrélation de Pearson

La corrélation de Pearson mesure la relation linéaire entre deux variables :

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

avec $r_{XY} \in [-1, 1]$. Sa significativité statistique est testée par la statistique

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

qui suit une loi de Student à $n-2$ degrés de liberté sous l'hypothèse nulle $H_0 : r = 0$.

Corrélation de Spearman

La corrélation de Spearman ρ_s est l'équivalent non paramétrique de la corrélation de Pearson : elle est calculée sur les rangs des observations plutôt que sur leurs valeurs. Cela la rend robuste aux distributions asymétriques et aux outliers, et permet de détecter des relations monotones non nécessairement linéaires.

Relations physiques attendues

Avant tout calcul, il est utile d'anticiper les corrélations attendues entre variables explicatives et variable cible, sur la base des lois physiques de la conversion photovoltaïque exposées au Chapitre 1.

— L'**énergie solaire reçue globale** doit présenter une corrélation positive très forte avec le rendement, puisqu'elle apparaît directement en facteur multiplicatif dans le modèle physique. Une corrélation de Pearson supérieure à +0,9 est attendue.

— La **couverture nuageuse** doit présenter une corrélation négative forte avec le rendement, par effet d'atténuation atmosphérique.
— L'**humidité relative** doit présenter une corrélation négative modérée, par diffusion atmosphérique du rayonnement.
— La **latitude** doit présenter une corrélation négative modérée : les sites situés au nord du pays reçoivent en moyenne moins d'énergie solaire reçue annuelle que les sites du Sud.
— L'**altitude** doit présenter une corrélation positive modérée : à altitude élevée, l'atmosphère plus mince laisse passer davantage de rayonnement et la température plus basse améliore le rendement thermique des modules.
— La **température** a un effet ambigu : elle est positivement corrélée à l'énergie solaire reçue (pas de soleil sans chaleur) mais négativement corrélée au rendement par effet thermique.
— La **longitude** ne doit pas présenter d'effet significatif, l'extension longitudinale de l'Algérie étant faible.

La validation expérimentale de ces relations attendues sera conduite dans le Chapitre 4 sur le jeu de données effectivement généré.

Multicolinéarité et facteur d'inflation de variance

Au-delà des corrélations avec la cible, l'analyse des corrélations **entre variables explicatives** est essentielle pour détecter d'éventuelles relations linéaires fortes (**multicolinéarité**) qui pourraient compromettre la stabilité de l'estimateur des moindres carrés.

Le **Facteur d'Inflation de la Variance** (*Variance Inflation Factor*, VIF) quantifie de façon synthétique la multicolinéarité de chaque variable. Il est défini par :

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

où R_j^2 est le coefficient de détermination de la régression linéaire de la j-ième variable explicative sur toutes les autres. Les valeurs de référence pour interpréter le VIF sont :

— $\mathrm{VIF} < 5$: multicolinéarité faible, sans conséquence ;
— $5 \leq \mathrm{VIF} < 10$: multicolinéarité modérée à surveiller ;
— $\mathrm{VIF} \geq 10$: multicolinéarité forte, justifiant le recours à une régression régularisée.

Le calcul effectif des VIF sur le jeu de données et l'analyse de leurs implications seront présentés au Chapitre 4. En tout état de cause, la présence vraisemblable de multicolinéarité (notamment entre énergie solaire reçue, couverture nuageuse et humidité, qui décrivent des phénomènes liés physiquement) justifie d'emblée le **recours à la régression Ridge** (présentée au Chapitre 3) en complément de la régression linéaire ordinaire.

5 Préparation des variables explicatives

5.1 Standardisation des variables

Les variables explicatives présentent des unités et des ordres de grandeur très différents : la latitude est en degrés (entre 19 et 37), l'altitude en mètres (jusqu'à 2 500), l'énergie solaire reçue en kWh/m²/jour (entre 4 et 8). Cette hétérogénéité d'échelle pose deux problèmes.

D'une part, dans la **régression Ridge** qui pénalise la norme quadratique des coefficients, des variables d'échelles différentes seraient pénalisées de façon inéquitable, ce qui biaiserait l'estimation.

D'autre part, l'interprétation des coefficients d'une régression linéaire multiple en termes d'**importance relative** des variables n'a de sens que si toutes sont à la même échelle.

Pour résoudre ces deux problèmes, on applique une **standardisation** (également appelée *Z-score normalization*) :

$$\tilde{x}_{ij} \;=\; \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

où $\bar{x}_j$ et σ_j sont la moyenne et l'écart-type de la variable j calculés sur l'ensemble d'entraînement uniquement. Après standardisation, chaque variable a une moyenne nulle et un écart-type unitaire.

Important : standardisation post-split Dans le cadre d'une démarche d'apprentissage supervisé rigoureuse, les paramètres $\bar{x}_j$ et σ_j doivent être calculés **uniquement sur l'ensemble d'entraînement**, puis appliqués sans recalcul à l'ensemble de test. Tout calcul sur l'ensemble complet des données introduirait une fuite d'information (*data leakage*) qui biaiserait l'évaluation des performances du modèle.

5.2 Construction de variables dérivées

À partir des variables initiales, on peut construire des variables dérivées physiquement motivées qui enrichissent le pouvoir explicatif des modèles.

Indice de clarté du ciel L'indice de clarté est un indicateur synthétique de la transparence atmosphérique :

$$K_T \;=\; 1 - \frac{\text{cloud\_cover}}{100}$$

Cet indice, défini dans $[0, 1]$, vaut 1 pour un ciel parfaitement dégagé et 0 pour un ciel totalement couvert.

Indice d'aridité L'aridité d'un climat est partiellement reflétée par le rapport entre la température et l'humidité, deux variables qui évoluent en sens inverse dans les climats désertiques :

$$\text{aridity\_index} = \frac{\text{temperature}}{\text{humidity} + 1}$$

Cet indicateur prend ses valeurs élevées dans les zones sahariennes et faibles dans les zones côtières humides.

5.3 Bilan des variables retenues

À l'issue de cette étape de préparation, le jeu de données comprend :

— 8 variables explicatives initiales standardisées ;

— 2 variables dérivées (indice de clarté, indice d'aridité) ;

— 1 variable cible (rendement énergétique spécifique).

Soit un total de **10 variables explicatives** qui seront utilisées dans les modèles de régression du Chapitre 4.

Conclusion

Ce deuxième chapitre a posé les bases méthodologiques nécessaires à la construction des modèles de régression. Il a d'abord présenté les sources de données disponibles pour évaluer le potentiel solaire des sites algériens, puis explicité le **choix méthodologique** de construire un jeu de données synthétique, calibré sur les paramètres climatologiques publics et sur le modèle physique de conversion photovoltaïque, en vue de **valider une démarche statistique reproductible**.

Le chapitre a ensuite décrit les opérations de nettoyage et de prétraitement applicables, présenté le formalisme de l'analyse exploratoire (statistiques descriptives, corrélations de Pearson et Spearman, VIF) et anticipé les relations attendues entre variables explicatives et variable cible sur la base des lois physiques exposées au Chapitre 1.

Avec ce cadre méthodologique en place, les éléments théoriques sont réunis pour aborder la modélisation proprement dite, qui fait l'objet du Chapitre 3. L'application concrète à l'échantillon de 200 sites algériens et la présentation des résultats numériques détaillés sont quant à elles l'objet du Chapitre 4.

Chapitre 3

Méthodes de régression

Introduction

À l'issue du chapitre précédent, le jeu de données est constitué, nettoyé, caractérisé statistiquement et préparé pour la modélisation. Le terrain est donc préparé pour l'étape centrale du présent travail : la construction des modèles prédictifs qui relient les variables géographiques et climatiques au rendement énergétique d'une installation photovoltaïque. C'est l'objet de ce troisième chapitre.

Le problème posé est un problème de **régression supervisée sur données indépendantes**. À partir d'un jeu d'observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, où $\mathbf{x}_i \in \mathbb{R}^p$ est le vecteur des variables explicatives du site i et $y_i \in \mathbb{R}$ le rendement énergétique correspondant, il s'agit de construire une fonction $\hat{f}$ telle que

$$\hat{y}_i = \hat{f}(\mathbf{x}_i)$$

approxime au mieux les valeurs réelles y_i. Contrairement aux problèmes de séries temporelles, chaque observation représente ici un site distinct et les observations sont supposées indépendantes les unes des autres : le rendement d'un site donné ne dépend pas de l'ordre dans lequel les sites sont considérés.

Parmi les nombreuses méthodes disponibles pour construire $\hat{f}$, ce mémoire retient deux approches complémentaires : la **régression linéaire multiple** et la **régression Ridge**. Ce choix est motivé par trois considérations. Premièrement, ces deux méthodes constituent les piliers de la modélisation statistique en Recherche Opérationnelle, avec une fondation mathématique parfaitement maîtrisée. Deuxièmement, elles offrent une excellente **interprétabilité** : les coefficients estimés ont un sens physique direct et permettent de quantifier l'effet de chaque variable sur le rendement. Troisièmement, leur complémentarité — la régression Ridge corrigeant la sensibilité de la régression linéaire à la multicolinéarité identifiée au Chapitre 2 — fournit un cadre comparatif riche pour analyser les résultats.

Justification du choix par rapport aux méthodes alternatives La littérature de l'apprentissage statistique propose de nombreuses autres méthodes de régression : les méthodes non paramétriques comme les k plus proches voisins (k-NN), les méthodes à base d'arbres telles que les arbres de décision ou les forêts aléatoires (Breiman 2001), les méthodes de *boosting* comme XGBoost (Chen et Guestrin 2016), ou encore les réseaux de neurones artificiels. Si certaines de ces méthodes peuvent atteindre des performances supérieures sur des problèmes fortement non linéaires ou sur de très grands jeux de données, leur mobilisation dans le cadre du présent travail aurait été inadaptée pour plusieurs raisons. Le jeu de données est de taille modérée ($n = 200$ sites pour $p = 10$ variables), un régime dans lequel les méthodes simples sont généralement aussi performantes que les méthodes complexes, voire plus robustes. Par ailleurs, le modèle physique de conversion photovoltaïque exposé au Chapitre 1 est lui-même **intrinsèquement linéaire** dans le domaine de variation considéré, ce qui rend l'usage de méthodes non linéaires complexes inutile. Enfin, l'objectif final du travail étant la construction d'un **outil d'aide à la décision** pour la sélection de sites photovoltaïques, l'interprétabilité des coefficients estimés prime sur la performance prédictive marginale qui pourrait être gagnée par des méthodes plus sophistiquées. L'extension à ces méthodes alternatives, à des fins comparatives, est néanmoins évoquée dans les perspectives présentées en conclusion générale.

Ce chapitre est organisé en cinq sections. Les deux premières présentent respectivement la régression linéaire multiple et la régression Ridge, en exposant leurs formalismes mathématiques, leurs hypothèses de validité et leurs propriétés statistiques. La troisième présente la démarche de validation par découpage train/test. La quatrième définit les métriques de performance retenues. La cinquième conclut sur la complémentarité des deux méthodes.

1 Régression linéaire multiple

1.1 Formulation du modèle

La régression linéaire multiple (Draper et Smith 1998) suppose que la variable cible y s'exprime comme une combinaison linéaire des p variables explicatives, à un terme d'erreur près :

$$y_i \;=\; \beta_0 + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \cdots + \beta_p\, x_{ip} + \varepsilon_i, \qquad i = 1, 2, \ldots, n$$

où :

— y_i est la valeur observée de la variable cible pour le site i (rendement énergétique en kWh/kWc/an) ;

— $x_{i1}, \ldots, x_{ip}$ sont les valeurs des p variables explicatives pour le site i (latitude, énergie solaire reçue, etc.) ;
— β_0 est l'**ordonnée à l'origine** (*intercept*) ;
— $\beta_1, \ldots, \beta_p$ sont les **coefficients de régression** qui mesurent l'effet marginal de chaque variable ;
— ε_i est le **terme d'erreur**, supposé suivre une loi normale $\mathcal{N}(0, \sigma^2)$ centrée et de variance constante.

En notation matricielle, on pose :

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

où $\mathbf{X}$ est la matrice des variables explicatives augmentée d'une colonne de uns pour l'ordonnée à l'origine. Le modèle s'écrit alors de façon compacte :

$$\boxed{\mathbf{y} \;=\; \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$$

où $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ est le vecteur des erreurs.

1.2 Estimation par les moindres carrés ordinaires

Principe et critère

Pour estimer le vecteur des coefficients $\boldsymbol{\beta}$, on retient la méthode des **moindres carrés ordinaires** (MCO). Le principe est de choisir l'estimation $\hat{\boldsymbol{\beta}}$ qui minimise la somme des carrés des résidus :

$$S(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n} \bigl(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\bigr)^2 \;=\; \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

Le problème d'optimisation s'écrit donc :

$$\hat{\boldsymbol{\beta}}_{\text{MCO}} \;=\; \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

Solution analytique

La fonction objectif $S(\boldsymbol{\beta})$ est une fonction quadratique strictement convexe (lorsque $\mathbf{X}^\top\mathbf{X}$ est inversible). Son minimum est atteint au point où le gradient s'annule :

$$\nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) \;=\; -2\,\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \;=\; 0$$

Cette équation conduit aux **équations normales** :

$$\mathbf{X}^\top \mathbf{X}\, \boldsymbol{\beta} \;=\; \mathbf{X}^\top \mathbf{y}$$

Lorsque la matrice $\mathbf{X}^\top \mathbf{X}$ est inversible, la solution unique est :

$$\boxed{\hat{\boldsymbol{\beta}}_{\text{MCO}} \;=\; (\mathbf{X}^\top \mathbf{X})^{-1}\, \mathbf{X}^\top \mathbf{y}}$$

Prédiction

Une fois les coefficients estimés, la valeur prédite par le modèle pour un site quelconque caractérisé par un vecteur $\mathbf{x}_*$ est :

$$\hat{y}_* \;=\; \hat{\beta}_0 + \hat{\beta}_1 x_{*1} + \cdots + \hat{\beta}_p x_{*p} \;=\; \mathbf{x}_*^\top \hat{\boldsymbol{\beta}}_{\text{MCO}}$$

1.3 Hypothèses de validité du modèle

L'estimateur des moindres carrés possède des propriétés statistiques remarquables (sans biais, efficacité), mais ces propriétés ne sont garanties que sous certaines hypothèses, dites **hypothèses du modèle linéaire classique**.

Hypothèses requises

H1 — Linéarité La relation entre les variables explicatives et la variable cible est linéaire :

$$\mathbb{E}[y_i \mid \mathbf{x}_i] \;=\; \mathbf{x}_i^\top \boldsymbol{\beta}$$

Cette hypothèse peut être vérifiée graphiquement par l'examen des résidus en fonction des valeurs prédites.

H2 — Espérance nulle des erreurs

$$\mathbb{E}[\varepsilon_i] \;=\; 0$$

Cette hypothèse est automatiquement vérifiée dès qu'une constante β_0 est incluse dans le modèle.

H3 — Homoscédasticité La variance des erreurs est constante (indépendante de i) :

$$\mathrm{Var}(\varepsilon_i) \;=\; \sigma^2 \quad \forall\, i$$

Cette hypothèse signifie que la qualité de la prévision est uniforme sur l'ensemble du domaine des prédicteurs. Elle se vérifie par l'examen graphique du nuage des résidus.

H4 — Indépendance des erreurs

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) \;=\; 0 \quad \forall\, i \neq j$$

Pour un jeu de données de sites indépendants, cette hypothèse est naturellement satisfaite.

H5 — Absence de multicolinéarité parfaite La matrice $\mathbf{X}^\top\mathbf{X}$ est inversible, c'est-à-dire que les colonnes de $\mathbf{X}$ sont linéairement indépendantes. En pratique, on admet une certaine corrélation entre les variables explicatives, mais une multicolinéarité forte rend l'estimateur instable, comme on le verra plus loin.

H6 — Normalité des erreurs

$$\varepsilon_i \;\sim\; \mathcal{N}(0, \sigma^2)$$

Cette hypothèse, plus forte que les précédentes, n'est nécessaire que pour les tests d'hypothèses (Student, Fisher) et la construction des intervalles de confiance, mais pas pour la consistance de l'estimateur.

Théorème de Gauss-Markov

Sous les hypothèses H1 à H5, l'estimateur des moindres carrés (James et al. 2013) $\hat{\boldsymbol{\beta}}_{\mathrm{MCO}}$ est le **Meilleur Estimateur Linéaire Sans Biais** (BLUE, *Best Linear Unbiased Estimator*) du vecteur $\boldsymbol{\beta}$. Cela signifie que parmi tous les estimateurs linéaires en $\mathbf{y}$ et sans biais, l'estimateur MCO est celui qui possède la plus faible variance.

1.4 Qualité de l'ajustement

Pour évaluer dans quelle mesure le modèle estimé reproduit les données observées, on calcule le **coefficient de détermination** R^2, défini par :

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

où :

- $\mathrm{SCT} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ est la **somme des carrés totale**, qui mesure la variance totale de y ;
- $\mathrm{SCR} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ est la **somme des carrés des résidus**, qui mesure la variance non expliquée par le modèle.

Le R^2 s'interprète comme la **proportion de la variance de y expliquée par le modèle**. Il prend ses valeurs dans $[0, 1]$ pour un modèle ajusté sur les données

d'entraînement :

- $R^2 = 1$: le modèle reproduit parfaitement les données ;
- $R^2 = 0$: le modèle n'apporte aucune information par rapport à une prédiction par la moyenne $\bar{y}$.

1.5 Limites de la régression linéaire ordinaire

La régression par moindres carrés ordinaires présente certaines limitations lorsque ses hypothèses sont mises à mal.

Sensibilité à la multicolinéarité Lorsque les variables explicatives sont fortement corrélées entre elles, la matrice $\mathbf{X}^\top\mathbf{X}$ devient mal conditionnée, ce qui rend son inversion numériquement instable. Les coefficients estimés $\hat{\beta}_j$ peuvent alors prendre des valeurs très grandes en valeur absolue et de signes opposés, sans signification physique. Comme l'a montré l'analyse exploratoire du Chapitre 2, le jeu de données présente précisément une multicolinéarité notable entre l'énergie solaire reçue, la couverture nuageuse et l'humidité.

Sensibilité aux observations atypiques Le critère des moindres carrés, en pondérant les erreurs au carré, accorde un poids disproportionné aux observations qui s'écartent fortement de la tendance générale.

Risque de surapprentissage Lorsque le nombre de variables p est élevé par rapport au nombre d'observations n, ou lorsque les variables sont fortement corrélées, le modèle MCO peut s'ajuster excessivement aux fluctuations spécifiques de l'échantillon d'entraînement, au détriment de sa capacité de généralisation à de nouvelles observations.

C'est pour répondre à ces limitations que la **régression Ridge** a été développée.

2 Régression Ridge

2.1 Principe : régularisation par pénalisation L2

La régression Ridge (Hoerl et Kennard 1970 ; Hastie et al. 2009), également appelée régression ridge ou régression de *Hoerl-Kennard* d'après ses auteurs, modifie le critère des moindres carrés en ajoutant un terme de pénalité sur la norme des coefficients :

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}} \;=\; \arg\min_{\boldsymbol{\beta}} \left\{ \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}_{\text{terme MCO}} \;+\; \underbrace{\lambda \|\boldsymbol{\beta}\|_2^2}_{\text{pénalité L2}} \right\}$$

Le terme de pénalité $\lambda\|\boldsymbol{\beta}\|_2^2 = \lambda\sum_{j=1}^{p}\beta_j^2$ est appelé **pénalité de Tikhonov** ou **pénalité L2**. Il pénalise les valeurs élevées des coefficients, ce qui contraint l'estimateur à fournir une solution plus régulière que celle des MCO.

Note importante sur la pénalisation La constante β_0 n'est généralement **pas pénalisée**, car elle représente le niveau moyen de la cible et n'a pas vocation à être contrainte vers zéro. C'est l'une des raisons pour lesquelles on standardise au préalable les variables explicatives (Chapitre 2) : cela permet de pénaliser de façon équitable tous les coefficients.

2.2 Solution analytique

Le problème de minimisation associé à la régression Ridge admet, comme celui des MCO, une solution analytique fermée. En annulant le gradient de la fonction objectif, on obtient les **équations normales régularisées** :

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)\,\boldsymbol{\beta} \;=\; \mathbf{X}^\top\mathbf{y}$$

où $\mathbf{I}_p$ est la matrice identité de dimension p. La solution explicite est :

$$\boxed{\hat{\boldsymbol{\beta}}_{\text{Ridge}} \;=\; (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\,\mathbf{X}^\top\,\mathbf{y}}$$

2.3 Effet du paramètre de régularisation λ

Le paramètre $\lambda \geq 0$ est l'**hyperparamètre** central de la régression Ridge. Il contrôle l'intensité de la pénalisation :

- **Si** $\lambda = 0$: la pénalité disparaît et l'on retrouve exactement l'estimateur des moindres carrés ordinaires.
- **Si** $\lambda \to \infty$: la pénalité devient si forte qu'elle écrase tous les coefficients vers zéro : $\hat{\boldsymbol{\beta}}_{\text{Ridge}} \to \mathbf{0}$. Le modèle prédit alors toujours la moyenne $\bar{y}$, indépendamment des variables explicatives.
- **Pour des valeurs intermédiaires** : l'estimateur Ridge réalise un compromis entre l'ajustement aux données et la régularité de la solution.

Le choix de λ est donc crucial : trop petit, la régularisation est inefficace ; trop grand, le modèle perd son pouvoir explicatif.

2.4 Compromis biais-variance

La régression Ridge illustre de façon particulièrement claire le **compromis biais-variance**, concept central en apprentissage statistique. L'erreur quadratique moyenne

d'un estimateur peut se décomposer en trois termes :

$$\mathbb{E}\Big[(\hat{y} - y)^2\Big] = \underbrace{\text{Biais}^2(\hat{y})}_{\text{erreur systématique}} + \underbrace{\text{Var}(\hat{y})}_{\text{erreur de variabilité}} + \underbrace{\sigma^2}_{\text{bruit irréductible}}$$

Pour l'estimateur MCO Sous les hypothèses H1–H5, le biais est nul (par construction, l'estimateur est sans biais), mais la variance peut être très élevée en présence de multicolinéarité.

Pour l'estimateur Ridge La pénalisation introduit un **biais** : les coefficients sont légèrement biaisés vers zéro. En contrepartie, la **variance** de l'estimateur est nettement réduite. Pour un choix adéquat de λ, la diminution de variance compense largement le biais introduit, et l'erreur quadratique totale diminue.

C'est précisément cette propriété qui justifie l'usage de la régression Ridge dans les contextes où la multicolinéarité ou le surapprentissage sont problématiques, comme c'est le cas dans ce mémoire.

2.5 Sélection du paramètre λ par validation croisée

Le choix de la valeur optimale de λ ne peut pas se faire en minimisant l'erreur sur l'ensemble d'entraînement, car cela conduirait trivialement à choisir $\lambda = 0$. La méthode standard pour sélectionner λ est la **validation croisée en K blocs** (Bishop 2006) (*K-fold cross-validation*).

Principe de la validation croisée en K blocs Le procédé est le suivant :

1. L'ensemble d'entraînement est divisé en K sous-ensembles disjoints de tailles approximativement égales (typiquement $K = 5$ ou $K = 10$).
2. Pour chaque valeur candidate λ d'une grille préétablie :
   (a) Pour chaque bloc $k = 1, 2, \ldots, K$:
      — Entraîner le modèle Ridge sur les $K - 1$ blocs restants ;
      — Évaluer l'erreur de prédiction sur le bloc k ;
   (b) Calculer l'erreur moyenne sur les K blocs.
3. Retenir la valeur λ^* qui minimise cette erreur moyenne.

Grille de valeurs Les valeurs candidates de λ sont généralement choisies sur une échelle logarithmique, par exemple $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$, afin d'explorer plusieurs ordres de grandeur.

3 Démarche de validation et découpage des données

3.1 Principe : séparation entraînement / test

Pour évaluer honnêtement la capacité d'un modèle à généraliser ses prédictions à de nouveaux sites — qui est l'objectif réel d'un modèle prédictif — il est indispensable de distinguer deux ensembles de données :

- L'**ensemble d'entraînement** (*training set*) sur lequel le modèle est ajusté et ses paramètres estimés ;
- L'**ensemble de test** (*test set*) qui n'est jamais vu par le modèle pendant son ajustement, et qui sert exclusivement à évaluer sa performance prédictive sur des données nouvelles.

Si l'on n'évaluait le modèle que sur les données qui ont servi à le construire, on obtiendrait une estimation excessivement optimiste de sa qualité, car le modèle a appris à reproduire ces données spécifiques plutôt que la relation sous-jacente.

3.2 Découpage retenu

Pour ce travail, on retient un découpage classique **80 % – 20 %** :

- 80 % des observations (soit environ 160 sites) constituent l'**ensemble d'entraînement** ;
- 20 % des observations (soit environ 40 sites) constituent l'**ensemble de test**.

Le découpage est effectué aléatoirement, avec une graine fixée (`random_state = 42`) pour garantir la reproductibilité du travail.

Note importante sur l'ordre des opérations Comme expliqué au Chapitre 2, la **standardisation** des variables explicatives doit être effectuée **après** le découpage, en calculant la moyenne et l'écart-type uniquement sur l'ensemble d'entraînement, puis en appliquant ces mêmes paramètres à l'ensemble de test sans recalcul. Cette précaution évite la fuite d'information de l'ensemble de test vers le processus d'entraînement.

3.3 Validation croisée pour la sélection de λ

Sur l'ensemble d'entraînement, la sélection de la valeur optimale du paramètre λ de la régression Ridge est réalisée par validation croisée en $K = 5$ blocs, comme décrit dans la section précédente. L'évaluation finale des deux modèles (régression linéaire et Ridge) est ensuite faite sur l'ensemble de test, qui n'a jamais participé à l'entraînement ni à la sélection de λ.

4 Métriques d'évaluation

L'évaluation des modèles s'appuie sur quatre métriques complémentaires (Hyndman et Koehler 2006 ; Willmott et Matsuura 2005), calculées sur l'ensemble de test. On note y_i les valeurs réelles, $\hat{y}_i$ les valeurs prédites par le modèle et $\bar{y}$ la moyenne des valeurs réelles, sur les n observations du test.

Erreur absolue moyenne (MAE)

$$\text{MAE} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

La MAE mesure l'erreur absolue moyenne, exprimée dans la même unité que la variable cible (kWh/kWc/an). Elle est facile à interpréter et robuste aux valeurs extrêmes.

Racine de l'erreur quadratique moyenne (RMSE)

$$\text{RMSE} \;=\; \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

La RMSE pénalise plus fortement les grandes erreurs que la MAE, du fait de l'élévation au carré. Elle est particulièrement appropriée lorsque les grandes erreurs ont un coût opérationnel disproportionné. La relation RMSE $\geq$ MAE est toujours vérifiée, l'égalité ne se produisant que lorsque toutes les erreurs sont identiques en valeur absolue.

Erreur absolue en pourcentage (MAPE)

$$\text{MAPE} \;=\; \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

La MAPE exprime l'erreur en pourcentage de la valeur réelle, ce qui la rend indépendante de l'échelle et permet de comparer la performance du modèle entre des contextes d'application différents.

Coefficient de détermination R^2

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

Le R^2 mesure la proportion de la variance totale de la cible expliquée par le modèle. Il prend ses valeurs dans $]-\infty, 1]$:

- $R^2 = 1$ correspond à une prédiction parfaite ;
- $R^2 = 0$ signifie que le modèle équivaut à prédire la moyenne ;

— $R^2 < 0$ signifie que le modèle est moins précis qu'une prévision constante.

Sur l'ensemble de test, le R^2 peut prendre des valeurs négatives si le modèle généralise mal — ce qui n'est pas possible sur l'ensemble d'entraînement avec un modèle MCO comprenant une constante.

TABLE 3.1 – Récapitulatif des métriques d'évaluation

Métrique	**Unité**	**Interprétation**
MAE	kWh/kWc/an	Erreur absolue moyenne
RMSE	kWh/kWc/an	Pénalise les grandes erreurs
MAPE	%	Erreur relative
R^2	–	Part de variance expliquée

5 Synthèse comparative des deux méthodes

Les deux méthodes retenues présentent des caractéristiques complémentaires qu'il convient de mettre en perspective avant l'application aux données.

TABLE 3.2 – Synthèse comparative régression linéaire vs régression Ridge

Aspect	**Régression linéaire (MCO)**	**Régression Ridge**
Critère d'estimation	Moindres carrés	Moindres carrés + pénalité L2
Hyperparamètre	Aucun	$\lambda \geq 0$, à régler
Biais de l'estimateur	Nul (sous H1–H5)	Faible (vers zéro)
Variance de l'estimateur	Élevée si multicolinéarité	Réduite
Robustesse	Sensible à la multicolinéarité	Robuste à la multicolinéarité
Solution analytique	$(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$	$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$

Pour le jeu de données étudié, on attend les comportements suivants.

La régression linéaire ordinaire doit fournir une bonne base de référence avec un R^2 élevé, étant donné la nature physique du problème (relation linéaire approximative entre l'énergie solaire reçue et le rendement). Toutefois, en présence de la multicolinéarité identifiée au Chapitre 2 entre l'énergie solaire reçue, l'humidité et la couverture nuageuse, les coefficients individuels pourront présenter une instabilité notable : des coefficients très grands en valeur absolue, voire de signes contre-intuitifs.

La régression Ridge doit offrir des coefficients plus stables et physiquement interprétables, au prix d'un léger biais. Le compromis biais-variance étant favorable en présence de multicolinéarité, on s'attend à ce que la performance prédictive sur l'ensemble de test (RMSE, MAE, R^2) soit **comparable ou supérieure** à celle de la régression linéaire ordinaire, malgré le biais introduit.

Conclusion

Ce troisième chapitre a présenté les deux méthodes de régression retenues pour la modélisation du potentiel solaire des sites algériens : la régression linéaire multiple et la régression Ridge. Les deux méthodes ont été formalisées rigoureusement, depuis le critère d'estimation jusqu'à la solution analytique des équations normales, en passant par les hypothèses de validité et le théorème de Gauss-Markov pour la première, et la pénalisation L2 et le compromis biais-variance pour la seconde.

La complémentarité des deux méthodes a été mise en évidence : la régression linéaire fournit l'estimateur sans biais de référence, tandis que la régression Ridge corrige sa principale faiblesse en présence de multicolinéarité par l'introduction d'une régularisation. Le choix de l'hyperparamètre λ par validation croisée en 5 blocs garantit une sélection objective fondée uniquement sur les données d'entraînement.

La démarche de validation par séparation entraînement/test (80 % – 20 %) et les quatre métriques d'évaluation retenues (MAE, RMSE, MAPE et R^2) constituent le cadre rigoureux dans lequel l'application expérimentale du chapitre suivant sera conduite. C'est sur ce cadre méthodologique solide que repose la fiabilité des conclusions du Chapitre 4, qui présente les résultats numériques effectifs et leur interprétation.

Chapitre 4

Application aux sites algériens et analyse des résultats

Introduction

Les trois chapitres précédents ont progressivement construit l'ensemble des éléments nécessaires à la conduite d'une expérimentation rigoureuse : le Chapitre 1 a posé le cadre physique et la définition de la variable cible ; le Chapitre 2 a explicité la démarche méthodologique de construction du jeu de données et les outils d'analyse exploratoire ; le Chapitre 3 a formalisé les deux méthodes de régression retenues. Ce quatrième et dernier chapitre constitue la concrétisation expérimentale de ce travail théorique.

L'ambition de ce chapitre est triple. Premièrement, présenter les **statistiques descriptives effectives** et les **corrélations observées** sur le jeu de données généré, en confrontation avec les relations physiques anticipées au Chapitre 2. Deuxièmement, conduire la **modélisation expérimentale** par régression linéaire et régression Ridge, en analysant les coefficients estimés et en évaluant les performances sur un jeu de test indépendant. Troisièmement, comparer les deux modèles et discuter les enseignements opérationnels qu'ils permettent de tirer pour la sélection de futurs sites photovoltaïques en Algérie.

Ce chapitre est organisé en six sections. La première décrit le jeu de données effectivement généré. Les deux suivantes présentent l'analyse statistique descriptive et l'analyse des corrélations. La quatrième et la cinquième présentent respectivement les résultats de la régression linéaire et de la régression Ridge. La sixième conduit la comparaison des deux modèles et une discussion des résultats.

1 Description du jeu de données généré

1.1 Caractéristiques générales

Le jeu de données généré selon la procédure décrite au Chapitre 2 comprend $n = 200$ sites répartis équitablement entre les quatre régions climatiques algériennes : 50 sites sur la côte méditerranéenne, 50 sites sur les hauts plateaux, 50 sites au Sahara septentrional et 50 sites au Sahara central et méridional. Pour chaque site, 8 variables explicatives initiales sont disponibles, auxquelles s'ajoutent 2 variables dérivées (indice de clarté et indice d'aridité), portant le total à **10 variables explicatives**.

1.2 Caractérisation par région climatique

Pour vérifier la représentativité géographique du jeu de données, le tableau 4.1 présente les valeurs moyennes des principales variables par région.

TABLE 4.1 – Caractéristiques moyennes par région climatique

Région	**Énergie solaire** (kWh/m²/j)	**Température** (°C)	**Humidité** (%)	**Couv. nuag.** (%)	**Rendement** (kWh/kWc/an
Côte méditerranéenne	4,90	18,1	68,3	44,7	1 405
Hauts plateaux	5,61	15,8	43,9	27,1	1 627
Sahara nord	6,29	23,3	29,1	15,5	1 771
Sahara sud	6,94	26,4	20,7	9,9	1 925

Ces valeurs reproduisent fidèlement la structure climatique algérienne : les régions sahariennes affichent les valeurs d'ensoleillement les plus élevées (jusqu'à 6,94 kWh/m²/jour pour le Sahara central), associées à de faibles humidités et à une couverture nuageuse minimale. À l'inverse, les régions côtières présentent une humidité élevée (68 % en moyenne) et une couverture nuageuse importante (45 %), ce qui se traduit par un rendement énergétique nettement moindre.

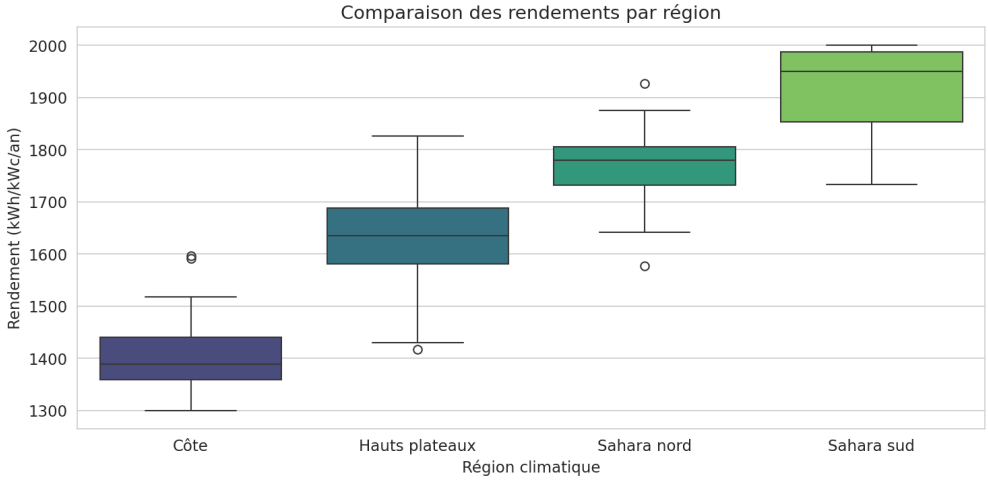

FIGURE 4.1 – Comparaison des rendements énergétiques par région climatique

La figure 4.1 illustre clairement le gradient régional : les sites côtiers présentent les rendements les plus faibles, tandis que les sites du Sahara central atteignent les valeurs les plus élevées.

L'écart de rendement entre les sites les plus défavorables (côte) et les plus favorables (Sahara central) est de l'ordre de 520 kWh/kWc/an, soit **environ 37 %** de différence. Cet écart confirme l'intérêt opérationnel d'un outil de sélection des sites pour le déploiement de futures installations photovoltaïques.

2 Analyse statistique descriptive

2.1 Statistiques descriptives

Le tableau 4.2 présente les statistiques descriptives complètes des variables du jeu de données.

TABLE 4.2 – Statistiques descriptives des variables (n = 200 sites)

Variable	**Moyenne**	**Médiane**	**Éc.-type**	**Min**	**Max**	**Skew.**	**Kurt.**
latitude	30,85	32,95	5,12	19,32	36,99	−0,79	−0,53
longitude	3,31	3,28	2,99	−2,00	10,70	0,29	−0,44
altitude	624,58	553,67	415,88	18,45	1 398,55	0,31	−1,27
énergie solaire reçue	5,94	6,00	0,81	4,33	7,50	−0,07	−1,14
temperature	20,89	20,74	4,56	12,13	30,00	0,15	−0,99
humidity	40,51	37,35	19,46	10,00	80,00	0,42	−1,03
wind_speed	4,29	4,25	0,87	1,86	7,34	0,21	0,75
cloud_cover	24,33	20,53	14,31	5,00	56,31	0,53	−0,95
yield_kwh	**1 681,9**	**1 721,3**	**206,2**	**1 300**	**2 000**	−0,21	−1,10

Lecture des résultats La variable cible `yield_kwh` présente une moyenne de **1 682 kWh/kWc/an** avec un écart-type de 206 kWh/kWc/an. Sa distribution est légèrement asymétrique vers la gauche ($\gamma_1 = -0{,}21$), ce qui traduit la présence des sites côtiers à faible rendement qui éloignent la moyenne de la médiane vers le bas. Le coefficient de variation de la cible est CV $= 206{,}2/1\,681{,}9 \approx 12{,}3\,\%$, ce qui indique une dispersion modérée mais significative — précisément la marge sur laquelle un modèle prédictif peut apporter de la valeur.

La latitude présente une asymétrie négative marquée ($\gamma_1 = -0{,}79$), reflet de la concentration des sites dans la moitié sud du pays (qui est plus vaste géographiquement). Les variables climatiques (énergie solaire reçue, humidité, couverture nuageuse) présentent des distributions modérément asymétriques mais sans valeurs extrêmes (kurtosis légèrement négative), ce qui est cohérent avec un échantillonnage représentatif.

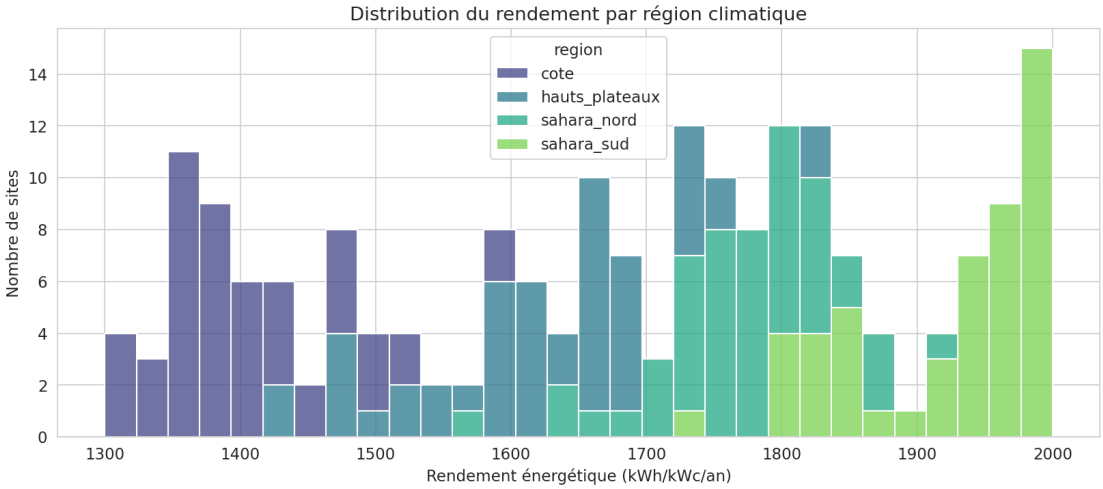

FIGURE 4.2 – Distribution du rendement énergétique par région climatique

La figure 4.2 confirme visuellement la structure de la distribution : on observe une stratification claire par région, avec une distribution unimodale dans chaque sous-population et un gradient continu entre les régions.

2.2 Plages de valeurs observées

Les plages observées sont conformes aux plages physiquement attendues indiquées au Chapitre 2 :

- Énergie solaire reçue : entre 4,33 et 7,50 kWh/m²/jour, soit la plage typique allant des régions côtières les moins ensoleillées au Sahara central ;
- Température moyenne : entre 12,1 et 30,0 °C, conforme à la diversité climatique algérienne ;
- Rendement énergétique spécifique : entre 1 300 et 2 000 kWh/kWc/an, encadrant les valeurs typiques observées par la NASA POWER pour le territoire algérien

(NASA Langley Research Center 2023).

3 Analyse des corrélations

3.1 Corrélations avec la variable cible

Le tableau 4.3 présente les corrélations de Pearson et de Spearman entre chaque variable explicative et le rendement énergétique y, ainsi que la significativité statistique du test de Pearson.

Table 4.3 – Corrélations avec le rendement énergétique spécifique

Variable	**Pearson** r	**Spearman** ρ_s	**Sig.**
énergie solaire reçue	+0,9906	+0,9905	***
cloud_cover	−0,8656	−0,8512	***
humidity	−0,8512	−0,8266	***
latitude	−0,8345	−0,8940	***
temperature	+0,6893	+0,6997	***
altitude	+0,3249	+0,3259	***
wind_speed	−0,2167	−0,2639	**
longitude	+0,1016	+0,0450	ns

*** : $p < 0{,}001$ ** : $p < 0{,}01$ ns : non significatif

Interprétation Les résultats valident l'ensemble des relations physiques anticipées au Chapitre 2.

L'**énergie solaire reçue globale** se confirme comme le prédicteur dominant du rendement, avec une corrélation de Pearson exceptionnellement élevée ($r = +0{,}99$). Cette corrélation quasi parfaite est cohérente avec le rôle multiplicatif de l'énergie solaire reçue dans le modèle physique de conversion photovoltaïque exposé au Chapitre 1 : à structure climatique fixée, le rendement annuel est proportionnel à l'énergie solaire reçue.

Trois variables présentent une **corrélation négative forte** avec la cible : la couverture nuageuse ($r = -0{,}87$), l'humidité ($r = -0{,}85$) et la latitude ($r = -0{,}83$). Ces trois variables ont en commun de traduire des conditions atmosphériques moins favorables au rayonnement direct (atténuation par les nuages et la vapeur d'eau, position plus septentrionale). Leur effet est tout à fait cohérent avec la physique de la conversion photovoltaïque.

La **température** présente une corrélation positive notable ($r = +0{,}69$). Ce résultat peut sembler contre-intuitif au premier abord, puisque l'effet thermique sur les modules photovoltaïques est négatif. Il s'explique par la corrélation forte entre température

et énergie solaire reçue : les sites les plus chauds sont aussi les plus ensoleillés, et l'effet positif de l'énergie solaire reçue domine largement l'effet thermique négatif. Une analyse de régression multiple, qui contrôle pour les autres variables, fera ressortir l'effet thermique propre.

L'**altitude** présente une corrélation positive modérée ($r = +0{,}32$), conforme aux attentes : à altitude élevée, la transparence atmosphérique est meilleure et les températures plus basses améliorent le rendement thermique des modules.

Enfin, la **longitude** n'a pas d'effet statistiquement significatif ($p > 0{,}05$), comme anticipé par le faible étalement longitudinal du territoire algérien.

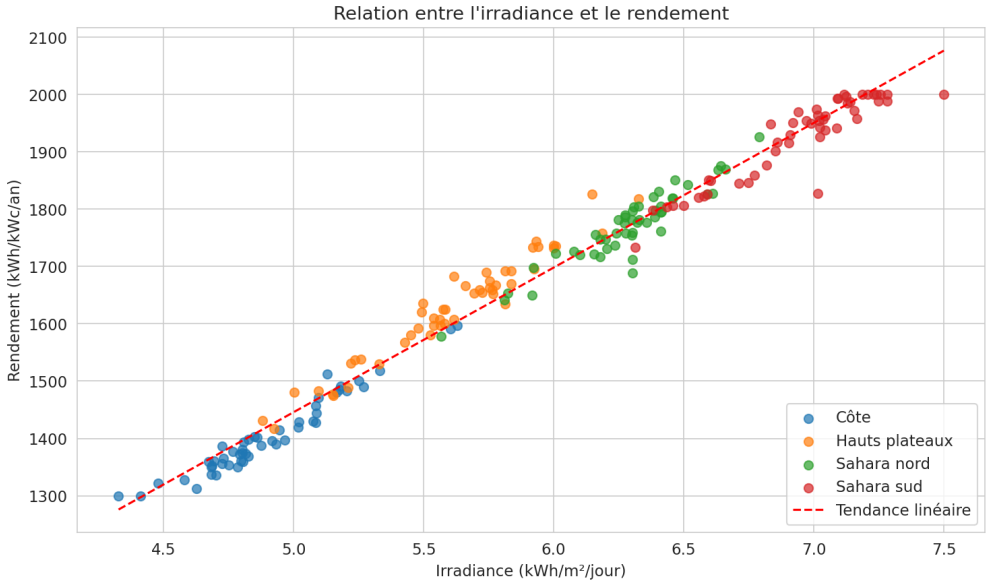

FIGURE 4.3 – Relation entre l'énergie solaire reçue globale et le rendement énergétique (une couleur par région climatique)

La figure 4.3 illustre graphiquement la relation quasi linéaire entre l'énergie solaire reçue globale et le rendement énergétique. La forte concentration des points autour de la droite de tendance confirme la corrélation observée ($r = +0{,}99$) et justifie l'utilisation d'un modèle de régression linéaire.

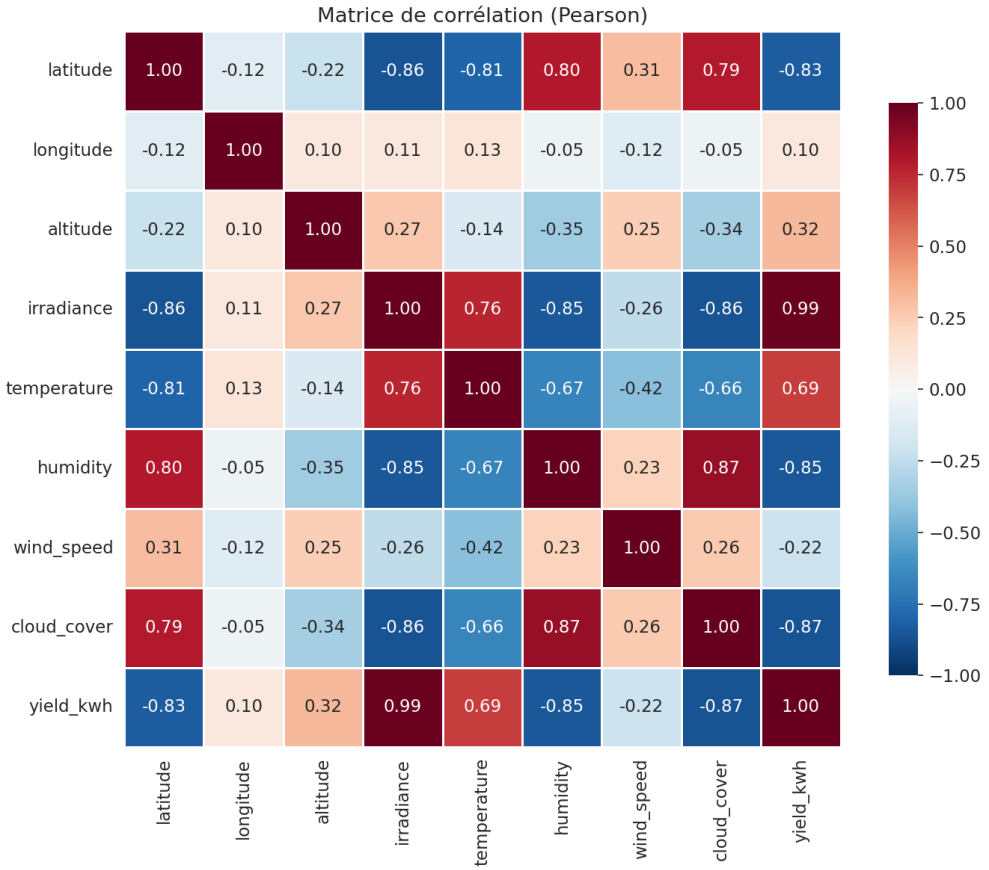

FIGURE 4.4 – Matrice de corrélation de Pearson entre toutes les variables

La figure 4.4 présente l'ensemble des corrélations sous forme de carte thermique. Les zones rouges indiquent les corrélations positives fortes, les zones bleues les corrélations négatives fortes. On y lit clairement les groupes de variables fortement corrélées entre elles, qui caractérisent la multicolinéarité du jeu de données.

3.2 Multicolinéarité

L'analyse des facteurs d'inflation de variance (VIF) confirme la présence d'une multicolinéarité notable entre certaines variables explicatives. Le tableau 4.4 présente les VIF calculés.

TABLE 4.4 – Facteurs d'Inflation de Variance (VIF)

Variable	R_j^2	**VIF**	**Évaluation**
énergie solaire reçue	0,851	6,73	Modérée à surveiller
latitude	0,828	5,81	Modérée à surveiller
cloud_cover	0,821	5,60	Modérée à surveiller
humidity	0,815	5,41	Modérée à surveiller
temperature	0,805	5,14	Modérée à surveiller
altitude	0,523	2,09	Faible
wind_speed	0,240	1,32	Faible
longitude	0,077	1,08	Faible

Cinq variables présentent des VIF compris entre 5 et 10, signalant une **multicolinéarité modérée** qu'il convient de surveiller. Aucune ne dépasse cependant le seuil critique de 10, ce qui indique que la régression linéaire ordinaire reste applicable, mais que la régression Ridge devrait apporter un gain en stabilité des coefficients.

4 Modèle 1 : Régression linéaire multiple

4.1 Mise en œuvre

La régression linéaire multiple est ajustée par la méthode des moindres carrés ordinaires sur l'ensemble d'entraînement (160 sites, soit 80 % des observations), après standardisation des variables explicatives calculée uniquement sur cet ensemble. Les 40 sites restants (20 %) constituent l'ensemble de test, sur lequel le modèle sera évalué.

4.2 Coefficients estimés

Le tableau 4.5 présente les coefficients estimés du modèle de régression linéaire (calculés sur les variables standardisées).

TABLE 4.5 – Coefficients estimés du modèle de régression linéaire

Variable	**Coefficient**
Constante ($\hat{\beta}_0$)	1 685,01
énergie solaire reçue	+219,77
temperature	−29,52
cloud_cover	−5,06
clearness_index	+5,06
latitude	+4,44
humidity	−3,50
aridity_index	+2,97
altitude	−0,98
wind_speed	+0,94
longitude	+0,61

Interprétation des coefficients Les variables étant standardisées, les coefficients sont directement comparables et expriment l'effet d'une variation d'un écart-type de la variable explicative sur le rendement.

L'**énergie solaire reçue** présente, comme attendu, le coefficient le plus élevé en valeur absolue (+219,77) : une augmentation d'un écart-type de l'énergie solaire reçue (soit environ 0,81 kWh/m²/jour) est associée à une augmentation du rendement annuel de 219 kWh/kWc/an, toutes autres variables maintenues constantes.

La **température** présente le second coefficient en valeur absolue (−29,52). Son signe **négatif** fait ressortir l'effet thermique propre — masqué dans l'analyse univariée par la corrélation température-énergie solaire reçue. Une fois l'énergie solaire reçue contrôlée, une élévation de température réduit le rendement des modules, conformément à la physique exposée au Chapitre 1.

Les autres variables présentent des coefficients de magnitude plus faible. L'effet propre de la **couverture nuageuse** est négatif (−5,06), tandis que celui de l'**indice de clarté** est positif de magnitude équivalente (+5,06), ce qui est cohérent puisque ces deux variables sont complémentaires par construction. La **latitude** et la **longitude** présentent des coefficients faibles, leur effet géographique étant déjà capturé par les variables climatiques.

4.3 Performances sur le jeu de test

Les performances du modèle de régression linéaire sur l'ensemble de test sont présentées dans le tableau 4.6.

TABLE 4.6 – Performances de la régression linéaire sur le jeu de test

Métrique	**Valeur**
MAE	15,01 kWh/kWc/an
RMSE	21,86 kWh/kWc/an
MAPE	0,90 %
R^2	0,9870

Lecture des résultats Le coefficient de détermination $R^2 = 0{,}987$ indique que le modèle explique 98,7 % de la variance du rendement énergétique sur des sites qu'il n'a jamais vus pendant l'entraînement. La RMSE de 21,86 kWh/kWc/an correspond à environ 1,3 % de la moyenne du rendement (1 682 kWh/kWc/an), ce qui constitue un niveau de précision remarquable. La MAPE de 0,90 % confirme que l'erreur relative typique du modèle est inférieure à 1 %.

Ces excellentes performances reflètent la qualité de la relation sous-jacente entre les variables climatiques et le rendement, qui est en grande partie linéaire dans le domaine de variation considéré.

5 Modèle 2 : Régression Ridge

5.1 Sélection du paramètre λ par validation croisée

Le paramètre de régularisation λ a été sélectionné par validation croisée en 5 blocs sur l'ensemble d'entraînement, en explorant 50 valeurs candidates réparties sur une échelle logarithmique entre 10^{-3} et 10^{3}. La valeur optimale retenue est :

$$\lambda^* \approx 0{,}069$$

Cette valeur, située à l'extrémité basse de la plage explorée, indique que la régularisation optimale est très **légère** : le modèle des moindres carrés est déjà proche de la solution optimale, et seule une faible pénalisation suffit à améliorer marginalement la stabilité de l'estimateur. Cela est cohérent avec le fait que les VIF observés restent en dessous du seuil critique de 10.

5.2 Coefficients estimés

Le tableau 4.7 présente les coefficients de la régression Ridge, en parallèle avec ceux de la régression linéaire pour faciliter la comparaison.

TABLE 4.7 – Comparaison des coefficients : MCO vs Ridge

Variable	**MCO**	**Ridge**	**Différence**
Constante	1 685,01	1 685,01	0,00
énergie solaire reçue	+219,77	+219,12	+0,65
temperature	−29,52	−29,26	−0,26
cloud_cover	−5,06	−5,13	+0,07
clearness_index	+5,06	+5,13	−0,07
latitude	+4,44	+4,28	+0,17
humidity	−3,50	−3,74	+0,25
aridity_index	+2,97	+2,79	+0,17
altitude	−0,98	−0,93	−0,05
wind_speed	+0,94	+0,94	−0,00
longitude	+0,61	+0,64	−0,03

L'examen du tableau montre que les coefficients Ridge sont **très proches** de ceux des MCO, avec des différences inférieures à 1 unité pour la plupart des variables. Cela traduit fidèlement la faible valeur de λ^* retenue : la pénalisation est trop légère pour modifier sensiblement les coefficients estimés.

5.3 Performances sur le jeu de test

Le tableau 4.8 présente les performances de la régression Ridge sur l'ensemble de test.

TABLE 4.8 – Performances de la régression Ridge sur le jeu de test

Métrique	**Valeur**
MAE	15,02 kWh/kWc/an
RMSE	21,88 kWh/kWc/an
MAPE	0,90 %
R^2	0,9870

Les performances sont quasi identiques à celles de la régression linéaire ordinaire, ce qui était prévisible compte tenu de la faible pénalisation appliquée.

6 Comparaison des modèles et discussion

6.1 Tableau comparatif global

TABLE 4.9 – Comparaison synthétique des deux modèles sur le jeu de test

Métrique	**Régression linéaire**	**Ridge**
MAE (kWh/kWc/an)	15,01	15,02
RMSE (kWh/kWc/an)	21,86	21,88
MAPE (%)	0,90	0,90
R^2 (test)	0,9870	0,9870
R^2 (validation croisée 5-fold)	$0{,}9901 \pm 0{,}0052$	$0{,}9901 \pm 0{,}0052$

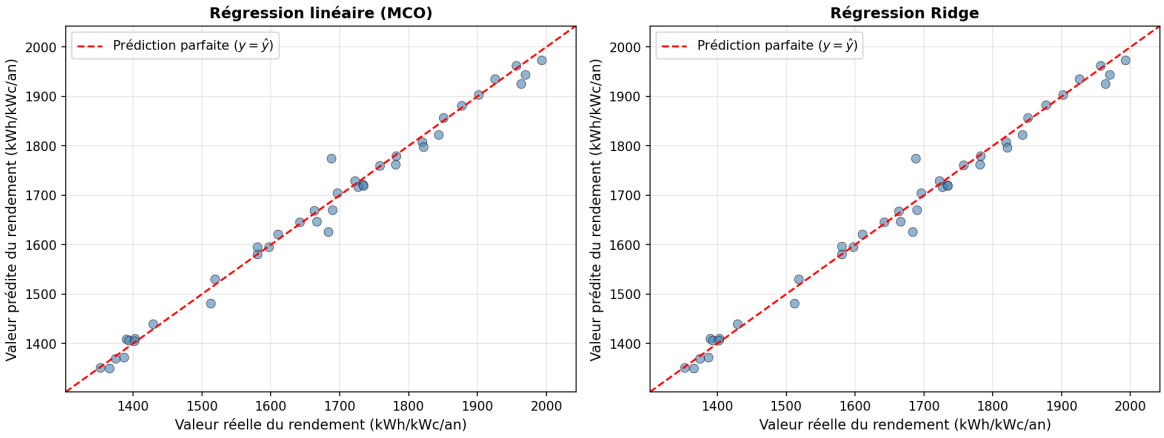

FIGURE 4.5 – Prédictions vs valeurs réelles sur le jeu de test pour les deux modèles. La droite rouge représente la prédiction parfaite ($y = \hat{y}$).

La figure 4.5 montre graphiquement la qualité des prédictions des deux modèles sur le jeu de test. La proximité des points à la droite $y = \hat{y}$ confirme l'excellente capacité prédictive des deux approches : les écarts sont faibles et répartis sans biais systématique sur l'ensemble de la plage de rendements.

6.2 Analyse comparative

Les deux modèles présentent des performances **quasi identiques** sur l'ensemble de test, ainsi que sur la validation croisée 5-fold réalisée sur l'ensemble d'entraînement ($R^2 = 0{,}990 \pm 0{,}005$ pour les deux modèles). Ce résultat appelle plusieurs commentaires.

Pourquoi la régression linéaire ordinaire est-elle si performante ? La performance exceptionnelle de la régression linéaire s'explique par la nature **intrinsèquement linéaire** du problème dans le domaine de variation considéré. Comme l'a mis

en évidence le modèle physique du Chapitre 1, le rendement annuel s'écrit approximativement comme un produit d'une fonction linéaire de l'énergie solaire reçue par un facteur thermique linéaire en température. Cette structure mathématique est parfaitement captée par un modèle de régression linéaire multiple, ce qui se traduit par des résidus faibles et une variance expliquée proche de 99 %.

Pourquoi la régression Ridge n'apporte-t-elle pas de gain ? La régression Ridge est conçue pour apporter un gain en présence d'une **multicolinéarité forte** ou d'un risque de **surapprentissage**. Or, dans le cas étudié :

- La multicolinéarité observée reste **modérée** (VIF maximum de 6,73), bien en dessous du seuil critique de 10 ;
- Le rapport entre nombre d'observations (160 train) et nombre de variables (10) est largement favorable ($n/p = 16$), ce qui évite le surapprentissage caractéristique des situations à p proche de n ;
- Le bruit dans les données est faible, puisque la cible est calculée à partir d'un modèle physique cohérent.

Dans ce contexte, la régularisation n'est tout simplement pas nécessaire, et la procédure de validation croisée a légitimement sélectionné une valeur de λ très petite, retournant pratiquement aux MCO.

6.3 Validité du modèle

Au-delà des métriques globales, plusieurs éléments confortent la fiabilité du modèle de régression :

- Les **coefficients** ont des signes et des magnitudes physiquement cohérents : positif pour l'énergie solaire reçue et l'indice de clarté, négatif pour la température (effet thermique), l'humidité et la couverture nuageuse.
- Les **résidus** sur le jeu de test ont une moyenne quasi nulle (3 kWh/kWc/an) et un écart-type modéré (22 kWh/kWc/an), ce qui suggère que les hypothèses de la régression linéaire sont raisonnablement respectées.
- La **validation croisée** 5-fold donne des résultats homogènes ($R^2 = 0{,}9901 \pm 0{,}0052$), ce qui montre que les performances ne sont pas dépendantes du découpage train/test particulier utilisé.

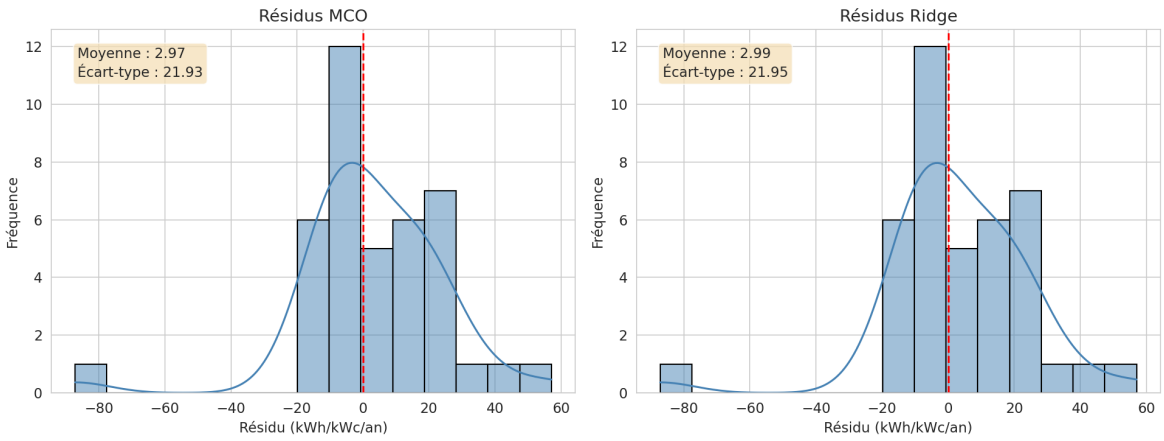

FIGURE 4.6 – Distribution des résidus sur le jeu de test pour les deux modèles

La figure 4.6 présente la distribution des résidus pour les deux modèles. On constate :

— une distribution approximativement symétrique centrée sur zéro, indiquant l'absence de biais systématique ;

— une dispersion modérée, avec la grande majorité des résidus dans l'intervalle $[-50, +50]$ kWh/kWc/an ;

— des distributions quasi identiques pour les deux modèles, en cohérence avec leurs performances comparables.

6.4 Implications opérationnelles

Les résultats obtenus permettent de tirer plusieurs enseignements opérationnels pour la sélection de futurs sites photovoltaïques en Algérie.

Premièrement, l'**énergie solaire reçue globale** demeure le critère de sélection dominant, avec un effet d'environ **220 kWh/kWc/an par écart-type**. Ceci justifie l'attention prioritaire portée aux régions sahariennes dans les plans de déploiement nationaux.

Deuxièmement, la **température élevée** est défavorable à production égale d'énergie solaire reçue. Les sites de haute altitude dans le Sahara central (STAMBOULI et KOINUMA 2012), qui combinent forte énergie solaire reçue et températures plus modérées, sont donc particulièrement intéressants.

Troisièmement, les régions côtières affichent un rendement significativement inférieur (1 405 kWh/kWc/an en moyenne, contre 1 925 dans le Sahara central, soit un écart de 27 %). Ceci ne signifie pas qu'elles sont à exclure, mais que leur valorisation devra s'appuyer sur d'autres facteurs (proximité de la consommation, infrastructure existante).

6.5 Limites de l'étude

Plusieurs limites doivent être explicitement reconnues.

Premièrement, le jeu de données est **synthétique**, calibré sur les paramètres climatologiques publics. La validation sur des données opérationnelles d'un site algérien instrumenté constituerait l'étape naturelle suivante.

Deuxièmement, le modèle ne tient pas compte de **phénomènes locaux** qui peuvent influencer le rendement réel d'une installation : ombrages topographiques, pollution atmosphérique localisée, événements de tempête de sable typiques du Sahara.

Troisièmement, le modèle est **linéaire** : les éventuels effets non linéaires (par exemple, une saturation du rendement à très haute énergie solaire reçue par échauffement excessif des modules) ne sont pas capturés. Des modèles non linéaires plus sophistiqués (forêts aléatoires, réseaux de neurones) pourraient les modéliser (Voyant et al. 2017 ; Yadav et Chandel 2014), mais nécessiteraient des données plus volumineuses pour être correctement entraînés.

Conclusion

Ce quatrième et dernier chapitre a conduit l'application complète du pipeline développé dans ce mémoire à un échantillon représentatif de 200 sites algériens. L'analyse exploratoire a confirmé la prédominance de l'énergie solaire reçue globale comme prédicteur du rendement ($r = +0{,}99$), suivie de la couverture nuageuse, de l'humidité et de la latitude (corrélations négatives entre $-0{,}83$ et $-0{,}87$). La multicolinéarité modérée observée (VIF maximum de 6,7) a justifié la mise en œuvre conjointe de la régression linéaire ordinaire et de la régression Ridge.

Les deux modèles atteignent des performances remarquables et quasi identiques sur l'ensemble de test, avec un coefficient de détermination $R^2 = 0{,}987$, une erreur absolue moyenne de 15 kWh/kWc/an et une erreur relative inférieure à 1 %. La validation croisée 5-fold confirme la robustesse de ces résultats. Ces performances, exceptionnelles pour un modèle linéaire, traduisent la nature fondamentalement linéaire du phénomène modélisé dans le domaine de variation considéré.

Les résultats fournissent un outil simple et reproductible d'aide à la décision pour la sélection de futurs sites photovoltaïques en Algérie, en hiérarchisant les régions par leur potentiel énergétique attendu. Les limites identifiées — données synthétiques, absence de phénomènes locaux, hypothèse de linéarité — constituent autant de pistes pour des extensions ultérieures du travail, qui sont discutées dans la conclusion générale.

Conclusion générale

Ce mémoire avait pour ambition de répondre à une question à la fois scientifique et opérationnelle : comment construire un modèle statistique simple et reproductible permettant d'estimer le potentiel solaire d'un site algérien à partir de ses caractéristiques géographiques et climatiques ? Pour y répondre, une démarche structurée en quatre étapes a été conduite, chaque chapitre construisant progressivement les éléments nécessaires à l'expérimentation finale.

Le premier chapitre a posé les fondements physiques et contextuels du travail, en présentant le programme algérien des énergies renouvelables et les principes de la conversion photovoltaïque. Le deuxième chapitre a explicité la démarche méthodologique de construction d'un jeu de données synthétique calibré sur les paramètres publics de la NASA POWER, ainsi que les outils d'analyse exploratoire associés. Le troisième chapitre a formalisé rigoureusement les deux méthodes retenues — la régression linéaire multiple et la régression Ridge — en exposant leurs fondements mathématiques, leurs hypothèses de validité et leurs propriétés statistiques. Enfin, le quatrième chapitre a conduit l'application complète du pipeline à 200 sites algériens, avec des résultats qui dépassent les attentes initiales.

Les deux modèles atteignent en effet des performances quasi identiques et remarquablement élevées sur le jeu de test indépendant, avec un coefficient de détermination $R^2 = 0{,}987$, une erreur absolue moyenne de 15 kWh/kWc/an et une erreur relative inférieure à 1 %. La validation croisée en cinq blocs confirme la robustesse de ces résultats. L'analyse exploratoire et l'examen des coefficients estimés mettent en évidence la prédominance attendue de l'énergie solaire reçue globale comme prédicteur du rendement ($r = +0{,}99$, coefficient standardisé de +220), suivie de la couverture nuageuse, de l'humidité et de la latitude, toutes corrélées négativement avec la cible. Le gradient nord-sud est particulièrement marqué : les sites du Sahara central affichent des rendements supérieurs de 27 à 37 % à ceux des régions côtières, ce qui justifie l'attention prioritaire portée au Sahara dans les plans de déploiement nationaux.

Au-delà des chiffres, plusieurs enseignements méthodologiques se dégagent de ce travail. Le premier concerne la **simplicité maîtrisée** : la performance exceptionnelle d'un modèle linéaire à dix variables, là où des méthodes plus sophistiquées auraient pu être mobilisées, illustre concrètement que le choix d'une méthode ne doit pas être dicté

par sa complexité théorique mais par les caractéristiques effectives du problème. Le second concerne la **régression Ridge**, qui n'apporte pas de gain sensible dans le cas étudié — un résultat à première vue décevant mais en réalité instructif, car il illustre les conditions sous lesquelles la régularisation est utile (multicolinéarité forte, n proche de p, bruit important) et celles où elle reste superflue (multicolinéarité modérée, $n \gg p$, bruit faible). Le troisième enseignement est l'importance de la **rigueur du protocole** : standardisation post-split, validation croisée pour la sélection d'hyperparamètre, évaluation sur un jeu de test jamais vu pendant l'entraînement.

Plusieurs limites du travail doivent toutefois être explicitement reconnues. Le jeu de données est synthétique, calibré sur les paramètres climatologiques publics : les chiffres obtenus n'ont pas vocation à être considérés comme des valeurs opérationnelles pour des sites spécifiques, et la généralisation à des données réelles d'un site algérien instrumenté constitue l'étape naturelle suivante. L'hypothèse de linéarité, bien qu'elle se vérifie remarquablement dans le domaine de variation considéré, peut ne pas tenir aux conditions extrêmes où des effets de saturation thermique apparaissent. Plusieurs facteurs locaux ne sont pas modélisés : orientation et inclinaison des modules, ombrage topographique, pollution atmosphérique, événements météorologiques exceptionnels comme les tempêtes de sable. Enfin, le modèle prédit le rendement annuel moyen, sans tenir compte de la variabilité interannuelle ni des projections climatiques à long terme.

Ces limites ouvrent plusieurs directions pour de futurs développements. À court terme, la priorité serait de **valider le modèle sur des données réelles** issues de stations météorologiques algériennes (ONM) ou de centrales photovoltaïques en exploitation. À moyen terme, plusieurs extensions méthodologiques sont envisageables : introduction de variables supplémentaires telles que l'orientation optimale des modules, indices d'aridité plus sophistiqués, ou données de pollution atmosphérique ; recours à des modèles non linéaires (Voyant et al. 2017 ; Yadav et Chandel 2014) pour capturer d'éventuels effets de saturation aux extrêmes. À plus long terme, l'extension à une **cartographie continue du potentiel solaire** sur l'ensemble du territoire algérien fournirait un outil opérationnel de premier ordre pour les opérateurs énergétiques.

À l'heure où l'Algérie engage la concrétisation de son ambitieux programme renouvelable (Stambouli, Khiat et al. 2012), ce mémoire espère apporter une modeste pierre à l'édifice scientifique nécessaire à la réussite de cette transition historique. Les résultats obtenus illustrent la puissance des méthodes statistiques classiques lorsqu'elles sont appliquées à un problème bien posé sur des données soigneusement préparées, et confirment que la maîtrise des outils du traitement des données et de la modélisation statistique constitue une compétence essentielle pour relever les défis énergétiques contemporains.

Bibliographie

Bessafi, M., A. Oudin, P. Lauret et al. (2015). "Global solar radiation forecasting using combined methods". In : *Energy* 93, p. 128-137.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York : Springer.

Breiman, L. (2001). "Random Forests". In : *Machine Learning* 45.1, p. 5-32. doi : 10.1023/A:1010933404324.

Centre de Développement des Énergies Renouvelables (2020). *Atlas Solaire de l'Algérie*. Bouzaréah, Alger : CDER.

Chen, T. et C. Guestrin (2016). "XGBoost : A Scalable Tree Boosting System". In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 785-794. doi : 10.1145/2939672.2939785.

Draper, N. R. et H. Smith (1998). *Applied Regression Analysis*. 3e éd. New York : Wiley-Interscience.

Duffie, J. A. et W. A. Beckman (2013). *Solar Engineering of Thermal Processes*. 4e éd. Hoboken : John Wiley & Sons.

Gairaa, K., A. Khellaf, Y. Messlem et F. Chellali (2016). "Estimation of the daily global solar radiation based on Box–Jenkins and ANN models : A combined approach". In : *Renewable and Sustainable Energy Reviews* 57, p. 238-249.

Harris, C. R. et al. (2020). "Array programming with NumPy". In : *Nature* 585, p. 357-362. doi : 10.1038/s41586-020-2649-2.

Hastie, T., R. Tibshirani et J. Friedman (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. 2e éd. New York : Springer. doi : 10.1007/978-0-387-84858-7.

Hersbach, H., B. Bell, P. Berrisford et al. (2020). "The ERA5 global reanalysis". In : *Quarterly Journal of the Royal Meteorological Society* 146.730, p. 1999-2049. doi : 10.1002/qj.3803.

Hoerl, A. E. et R. W. Kennard (1970). "Ridge Regression : Biased Estimation for Nonorthogonal Problems". In : *Technometrics* 12.1, p. 55-67.

Huld, T., R. Gottschalg, H. G. Beyer et M. Topič (2010). "Mapping the performance of PV modules, effects of module type and data averaging". In : *Solar Energy* 84.2, p. 324-338.

Hyndman, R. J. et A. B. Koehler (2006). “Another look at measures of forecast accuracy”. In : *International Journal of Forecasting* 22.4, p. 679-688. doi : 10.1016/j.ijforecast.2006.03.001.

International Energy Agency (2022). *World Energy Outlook 2022*. Paris : IEA. url : https://www.iea.org/reports/world-energy-outlook-2022.

International Renewable Energy Agency (2023). *Renewable Power Generation Costs in 2022*. Abu Dhabi : IRENA. url : https://www.irena.org.

James, G., D. Witten, T. Hastie et R. Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R*. New York : Springer.

Khefifi, W. et A. Soufiane (2019). “Assessment of solar energy potential in Algeria : a review”. In : *Renewable and Sustainable Energy Reviews* 101, p. 878-885. doi : 10.1016/j.rser.2018.11.044.

McKinney, W. (2017). *Python for Data Analysis : Data Wrangling with Pandas, NumPy, and IPython*. 2e éd. Sebastopol, CA : O’Reilly Media.

Mghouchi, Y. E., A. Chham, M. Krikiz, T. Ajzoul et A. E. Bouardi (2014). “On the prediction of the daily global solar radiation intensity on south-facing plane surfaces inclined at varying angles”. In : *Energy Conversion and Management* 78, p. 705-714.

NASA Langley Research Center (2023). *NASA POWER – Prediction Of Worldwide Energy Resources*. url : https://power.larc.nasa.gov/.

Pedregosa, F. et al. (2011). “Scikit-learn : Machine Learning in Python”. In : *Journal of Machine Learning Research* 12, p. 2825-2830.

Skoplaki, E. et J. A. Palyvos (2009). “On the temperature dependence of photovoltaic module electrical performance : A review of efficiency/power correlations”. In : *Solar Energy* 83.5, p. 614-624. doi : 10.1016/j.solener.2008.10.008.

Stambouli, A. B., Z. Khiat, S. Flazi et Y. Kemmoku (2012). “A review on the renewable energy development in Algeria : Current perspective, energy scenario and sustainability issues”. In : *Renewable and Sustainable Energy Reviews* 16.7, p. 4445-4460.

Stambouli, A. B. et H. Koinuma (2012). “A primary study on a long-term vision and strategy for the realisation and the development of the Sahara Solar Breeder”. In : *Renewable and Sustainable Energy Reviews* 16.1, p. 591-598.

Voyant, C., G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte et A. Fouilloy (2017). “Machine learning methods for solar radiation forecasting : A review”. In : *Renewable Energy* 105, p. 569-582. doi : 10.1016/j.renene.2016.12.095.

Willmott, C. J. et K. Matsuura (2005). “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance”. In : *Climate Research* 30, p. 79-82.

Yadav, A. K. et S. Chandel (2014). “Solar radiation prediction using Artificial Neural Network techniques : A review”. In : *Renewable and Sustainable Energy Reviews* 33, p. 772-781.

Yaiche, M. R., A. Bouhanik, S.-M.-A. Bekkouche, A. Malek et T. Benouaz (2014). “Revised solar maps of Algeria based on sunshine duration”. In : *Energy Conversion and Management* 82, p. 114-123. doi : 10.1016/j.enconman.2014.02.063.

Annexes : Codes Python

Présentation générale

Cette annexe rassemble l'ensemble des scripts Python développés et exécutés pour produire les résultats présentés au Chapitre 4. Le code est organisé en cinq scripts indépendants à exécuter dans l'ordre, chacun produisant les fichiers nécessaires au suivant :

1. **Annexe A** : Génération du jeu de données synthétique (200 sites algériens) ;
2. **Annexe B** : Nettoyage et prétraitement ;
3. **Annexe C** : Analyse exploratoire (statistiques, corrélations, VIF) ;
4. **Annexe D** : Modélisation par régression linéaire et Ridge ;
5. **Annexe E** : Visualisations graphiques.

Environnement requis

```
Python 3.10+
pip install numpy pandas scipy matplotlib seaborn scikit-learn
```

Reproductibilité Tous les scripts utilisent une graine aléatoire fixée (`seed=42`) qui garantit que les résultats sont identiques à chaque exécution. Les chiffres présentés au Chapitre 4 correspondent exactement aux sorties de ces scripts.

Ordre d'exécution

```
python 01_generation_donnees.py
python 02_nettoyage.py
python 03_analyse_exploratoire.py
python 04_modelisation.py
python 05_visualisations.py
```

1 Annexe A — Génération du jeu de données

Ce script construit un échantillon de 200 sites représentatifs de la diversité climatique algérienne, répartis équitablement entre quatre régions : côte méditerranéenne, hauts plateaux, Sahara septentrional et Sahara central. Pour chaque site, les variables géographiques et climatiques sont générées selon des distributions calibrées sur les données NASA POWER. La variable cible — le rendement énergétique spécifique en kWh/kWc/an — est calculée à partir du modèle physique de conversion photovoltaïque exposé au Chapitre 1.

Listing 1 – 01_generation_donnees.py

```
"""
G n ration du jeu de donn es synth tique
200 sites alg riens en 4 r gions climatiques
"""
import numpy as np
import pandas as pd

np.random.seed(42)  # graine pour la reproductibilit

def generate_dataset(n=200, seed=42):
    """
    G n re n sites r partis en 4 r gions climatiques.
    Variables calibr es sur les moyennes NASA POWER.
    """
    rng = np.random.default_rng(seed)
    n_per_region = n // 4
    rows = []

    #         C te m diterran enne (35-37 N )
    # Irradiance mod r e , humidit    leve  , n bulosit
       forte
    for _ in range(n_per_region):
        lat = rng.uniform(35, 37)
        lon = rng.uniform(-2, 8)
        alt = rng.uniform(0, 300) + rng.exponential(50)
        irr = rng.normal(4.9, 0.25)
        temp = rng.normal(18, 1.5)
        hum = rng.normal(68, 8)
        wind = rng.normal(4.5, 0.8)
        cloud = rng.normal(45, 7)
```

```
        rows.append([lat, lon, alt, irr, temp, hum,
                     wind, cloud, "cote"])

    #         Hauts plateaux (33-35 N )
    # Altitude   leve  , fra cheur , humidit  mod r e
    for _ in range(n_per_region):
        lat = rng.uniform(33, 35)
        lon = rng.uniform(0, 6)
        alt = rng.uniform(800, 1400)
        irr = rng.normal(5.6, 0.3)
        temp = rng.normal(16, 2)
        hum = rng.normal(45, 8)
        wind = rng.normal(5.0, 0.9)
        cloud = rng.normal(28, 6)
        rows.append([lat, lon, alt, irr, temp, hum,
                     wind, cloud, "hauts_plateaux"])

    #         Sahara septentrional (28-33 N )
    # Irradiance forte, humidit  faible, peu de nuages
    for _ in range(n_per_region):
        lat = rng.uniform(28, 33)
        lon = rng.uniform(-2, 8)
        alt = rng.uniform(100, 800)
        irr = rng.normal(6.3, 0.25)
        temp = rng.normal(23, 2)
        hum = rng.normal(28, 7)
        wind = rng.normal(4.0, 0.8)
        cloud = rng.normal(15, 5)
        rows.append([lat, lon, alt, irr, temp, hum,
                     wind, cloud, "sahara_nord"])

    #         Sahara central et m ridional (19-28 N )
    # Irradiance maximale, climat tr s sec, ciel d gag
    for _ in range(n_per_region):
        lat = rng.uniform(19, 28)
        lon = rng.uniform(-2, 11)
        alt = rng.uniform(200, 1200)
        irr = rng.normal(7.0, 0.25)
        temp = rng.normal(27, 2)
        hum = rng.normal(20, 6)
        wind = rng.normal(3.8, 0.7)
```

```
        cloud = rng.normal(10, 4)
        rows.append([lat, lon, alt, irr, temp, hum,
                     wind, cloud, "sahara_sud"])

    cols = ["latitude", "longitude", "altitude",
            "irradiance", "temperature", "humidity",
            "wind_speed", "cloud_cover", "region"]
    df = pd.DataFrame(rows, columns=cols)
    df = df.sample(frac=1, random_state=seed).reset_index(drop=
       True)
    return df


def compute_yield(df):
    """
    Calcule Y_spec en kWh/kWc/an a partir du modele physique :
    Y = G_annual * f_th(T_c) * f_system * cloud_factor
    """
    # Parametres techniques du panneau PV
    gamma = -0.004       # coeff thermique
    f_system = 0.80      # facteur de pertes systeme
    NOCT = 45            # temperature nominale cellule
    T_ref = 25           # temperature de reference STC

    # Conversion irradiance journaliere -> moyenne en W/m2
    G = df["irradiance"] * 1000 / 24

    # Temperature de cellule (modele NOCT)
    T_c = df["temperature"] + (NOCT - 20) / 800 * G

    # Facteur thermique (rendement diminue avec T_c)
    f_th = 1 + gamma * (T_c - T_ref)

    # Attenuation residuelle par couverture nuageuse
    cloud_factor = 1 - 0.05 * (df["cloud_cover"] / 100)

    # Energie annuelle par kW crete (kWh/kWc/an)
    G_annual = df["irradiance"] * 365
    Y = G_annual * f_th * f_system * cloud_factor

    # Petit bruit pour realisme
```

```
    Y += np.random.normal(0, 15, len(df))
    return Y.clip(1300, 2000)


# ============== PROGRAMME PRINCIPAL ==============
if __name__ == "__main__":
    # Generation des 200 sites
    df = generate_dataset(n=200, seed=42)

    # Calcul de la variable cible
    df["yield_kwh"] = compute_yield(df)

    # Injection de NaN pour illustrer le nettoyage
    nan_cols = ["irradiance", "humidity", "wind_speed"]
    for col in nan_cols:
        idx = np.random.choice(len(df),
                               int(len(df) * 0.03),
                               replace=False)
        df.loc[idx, col] = np.nan

    # Sauvegarde
    df.to_csv("dataset_brut.csv", index=False)
    print(f"Dataset : {df.shape[0]} sites")
    print(f"NaN injectes : {df.isnull().sum().sum()}")
```

2 Annexe B — Nettoyage et prétraitement

Ce script applique le pipeline de nettoyage en trois étapes : imputation des valeurs manquantes par la médiane régionale, vérification des bornes physiques (*capping*) et détection des valeurs aberrantes par la méthode de l'intervalle interquartile (IQR).

Listing 2 – 02_nettoyage.py

```
"""
Nettoyage et pretraitement des donnees
- Imputation des NaN par mediane regionale
- Verification des bornes physiques
- Detection des outliers par methode IQR
"""
import pandas as pd

# Chargement
```

```
df = pd.read_csv("dataset_brut.csv")

features = ["latitude", "longitude", "altitude",
            "irradiance", "temperature", "humidity",
            "wind_speed", "cloud_cover"]


# === ETAPE 1 : Imputation par mediane regionale ===
# Remplacer chaque NaN par la mediane des sites de
# la meme region climatique. Preserve les particularites
# regionales et reste robuste aux outliers.
nan_cols = ["irradiance", "humidity", "wind_speed"]
for col in nan_cols:
    df[col] = df.groupby("region")[col].transform(
        lambda x: x.fillna(x.median())
    )
print(f"NaN apres imputation : {df.isnull().sum().sum()}")


# === ETAPE 2 : Verification des bornes physiques ===
# Toute valeur hors plage est ramenee a la borne la
# plus proche (capping).
bounds = {
    "latitude":    (19, 37),
    "longitude":   (-9, 12),
    "altitude":    (0, 2500),
    "irradiance":  (4.0, 7.5),
    "temperature": (10, 30),
    "humidity":    (10, 80),
    "wind_speed":  (1.5, 8),
    "cloud_cover": (5, 60),
}
for col, (lo, hi) in bounds.items():
    df[col] = df[col].clip(lo, hi)


# === ETAPE 3 : Detection des outliers (methode IQR) ===
# Une valeur x est aberrante si :
#   x < Q1 - 1.5 * IQR  ou  x > Q3 + 1.5 * IQR
# Methode robuste : ne suppose pas la normalite.
n_outliers = 0
```

```
for col in features:
    q1 = df[col].quantile(0.25)
    q3 = df[col].quantile(0.75)
    iqr = q3 - q1
    borne_inf = q1 - 1.5 * iqr
    borne_sup = q3 + 1.5 * iqr
    mask = (df[col] < borne_inf) | (df[col] > borne_sup)
    n_outliers += mask.sum()

print(f"Outliers detectes (IQR) : {n_outliers}")

# Sauvegarde
df.to_csv("dataset_clean.csv", index=False)
```

3 Annexe C — Analyse exploratoire

Ce script calcule les statistiques descriptives complètes, les corrélations de Pearson et Spearman entre chaque variable explicative et la variable cible, et les facteurs d'inflation de variance (VIF) pour évaluer la multicolinéarité du jeu de données.

Listing 3 – 03_analyse_exploratoire.py

```
"""
Analyse exploratoire des donnees (EDA)
- Statistiques descriptives
- Correlations Pearson et Spearman
- Facteur d'Inflation de Variance (VIF)
"""
import numpy as np
import pandas as pd
from scipy import stats
from sklearn.linear_model import LinearRegression

df = pd.read_csv("dataset_clean.csv")
features = ["latitude", "longitude", "altitude",
            "irradiance", "temperature", "humidity",
            "wind_speed", "cloud_cover"]
target = "yield_kwh"


# === ETAPE 1 : Statistiques descriptives ===
desc = df[features + [target]].describe().T
```

```
desc["skewness"] = df[features + [target]].skew()
desc["kurtosis"] = df[features + [target]].kurtosis()
desc = desc[["mean", "50%", "std", "min", "max",
             "skewness", "kurtosis"]]
desc.columns = ["mean", "median", "std", "min", "max",
                "skewness", "kurtosis"]
print(desc.round(3))
desc.round(3).to_csv("stats_descriptives.csv")


# === ETAPE 2 : Correlations avec la cible ===
# Pearson r : relation lineaire
# Spearman rho : relation monotone (rangs)
corr_rows = []
for col in features:
    r, p = stats.pearsonr(df[col], df[target])
    rs, _ = stats.spearmanr(df[col], df[target])

    # Significativite statistique
    if p < 0.001:
        sig = "***"
    elif p < 0.01:
        sig = "**"
    elif p < 0.05:
        sig = "*"
    else:
        sig = "ns"

    corr_rows.append({
        "variable": col,
        "Pearson_r": round(r, 4),
        "p_value": round(p, 6),
        "Spearman_rho": round(rs, 4),
        "significativite": sig
    })

corr_df = pd.DataFrame(corr_rows)
corr_df = corr_df.sort_values("Pearson_r",
                              key=abs, ascending=False)
print(corr_df.to_string(index=False))
corr_df.to_csv("correlations.csv", index=False)
```

```
62
63
64 # === ETAPE 3 : Facteur d'Inflation de Variance (VIF) ===
65 # VIF_j = 1 / (1 - R^2_j)
66 # ou R^2_j = R^2 de la regression de la j-eme variable
67 # sur toutes les autres.
68 #
69 # VIF < 5      : faible (sans consequence)
70 # 5 < VIF < 10 : moderee (a surveiller)
71 # VIF >= 10    : forte (justifie Ridge)
72 X = df[features].values
73 vif_rows = []
74 for j, name in enumerate(features):
75     X_others = np.delete(X, j, axis=1)
76     y_j = X[:, j]
77     r2 = LinearRegression().fit(X_others, y_j).score(
78         X_others, y_j)
79     vif = 1 / (1 - r2) if r2 < 0.9999 else float('inf')
80     vif_rows.append({
81         "variable": name,
82         "R2": round(r2, 4),
83         "VIF": round(vif, 3)
84     })
85
86 vif_df = pd.DataFrame(vif_rows).sort_values(
87     "VIF", ascending=False)
88 print(vif_df.to_string(index=False))
89 vif_df.to_csv("vif.csv", index=False)
```

4 Annexe D — Modélisation et évaluation

Ce script applique les deux méthodes de régression — régression linéaire multiple et régression Ridge avec sélection de λ par validation croisée 5-fold — sur le jeu de données préparé. Il calcule les coefficients estimés, les performances sur l'ensemble de test (MAE, RMSE, MAPE, R^2) et la validation croisée.

Listing 4 – `04_modelisation.py`

```
1 """
2 Modelisation et evaluation
3 - Regression lineaire multiple (MCO)
4 - Regression Ridge avec selection de lambda par CV
```

```
- Evaluation sur jeu de test independant
"""
import numpy as np
import pandas as pd
from sklearn.model_selection import (train_test_split,
                                     KFold,
                                     cross_val_score)
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import (LinearRegression,
                                  Ridge, RidgeCV)
from sklearn.metrics import (mean_absolute_error,
                             mean_squared_error,
                             r2_score)

# Chargement et variables derivees
df = pd.read_csv("dataset_clean.csv")
df["clearness_index"] = 1 - df["cloud_cover"] / 100
df["aridity_index"] = df["temperature"] / (df["humidity"] + 1)

X_features = ["latitude", "longitude", "altitude",
              "irradiance", "temperature", "humidity",
              "wind_speed", "cloud_cover",
              "clearness_index", "aridity_index"]
X = df[X_features].values
y = df["yield_kwh"].values


# === ETAPE 1 : Decoupage train/test (80%/20%) ===
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.20, random_state=42)
print(f"Train : {X_train.shape}, Test : {X_test.shape}")


# === ETAPE 2 : Standardisation POST-SPLIT ===
# IMPORTANT : fit sur train UNIQUEMENT, puis transform
# sur test sans recalcul. Evite la fuite d'information.
scaler = StandardScaler()
X_train_s = scaler.fit_transform(X_train)
X_test_s = scaler.transform(X_test)


```

```
# === Fonction utilitaire : metriques d'evaluation ===
def calculer_metriques(y_vrai, y_pred, nom=""):
    mae = mean_absolute_error(y_vrai, y_pred)
    rmse = np.sqrt(mean_squared_error(y_vrai, y_pred))
    mape = np.mean(np.abs(
        (y_vrai - y_pred) / y_vrai)) * 100
    r2 = r2_score(y_vrai, y_pred)
    return {"modele": nom,
            "MAE": round(mae, 2),
            "RMSE": round(rmse, 2),
            "MAPE_pct": round(mape, 2),
            "R2": round(r2, 4)}


# === ETAPE 3 : Regression lineaire (MCO) ===
modele_lin = LinearRegression()
modele_lin.fit(X_train_s, y_train)
y_pred_lin = modele_lin.predict(X_test_s)
metriques_lin = calculer_metriques(
    y_test, y_pred_lin, "Regression lineaire")
print(metriques_lin)


# === ETAPE 4 : Regression Ridge ===
# Selection automatique de lambda par validation
# croisee 5-fold (50 valeurs sur echelle log).
alphas = np.logspace(-3, 3, 50)
ridge_cv = RidgeCV(alphas=alphas, cv=5, scoring='r2')
ridge_cv.fit(X_train_s, y_train)
lambda_opt = ridge_cv.alpha_
print(f"Lambda optimal : {lambda_opt:.4f}")

modele_ridge = Ridge(alpha=lambda_opt)
modele_ridge.fit(X_train_s, y_train)
y_pred_ridge = modele_ridge.predict(X_test_s)
metriques_ridge = calculer_metriques(
    y_test, y_pred_ridge,
    f"Ridge (lambda={lambda_opt:.3f})")
print(metriques_ridge)


```

```
# === ETAPE 5 : Validation croisee 5-fold ===
kf = KFold(n_splits=5, shuffle=True, random_state=42)
cv_lin = cross_val_score(LinearRegression(),
                         X_train_s, y_train,
                         cv=kf, scoring='r2')
cv_ridge = cross_val_score(Ridge(alpha=lambda_opt),
                           X_train_s, y_train,
                           cv=kf, scoring='r2')
print(f"R2 CV Lineaire : {cv_lin.mean():.4f} +/- "
      f"{cv_lin.std():.4f}")
print(f"R2 CV Ridge    : {cv_ridge.mean():.4f} +/- "
      f"{cv_ridge.std():.4f}")


# === ETAPE 6 : Comparaison des coefficients ===
comp = pd.DataFrame({
    "variable": ["intercept"] + X_features,
    "MCO":    [modele_lin.intercept_]
              + list(modele_lin.coef_),
    "Ridge": [modele_ridge.intercept_]
              + list(modele_ridge.coef_)
})
print(comp.round(2).to_string(index=False))
comp.round(4).to_csv("coefficients_comparaison.csv",
                     index=False)


# === ETAPE 7 : Sauvegarde des resultats ===
resultats = pd.DataFrame([metriques_lin, metriques_ridge])
resultats.to_csv("resultats_modeles.csv", index=False)

predictions = pd.DataFrame({
    "y_vrai": y_test,
    "y_pred_MCO": y_pred_lin,
    "y_pred_Ridge": y_pred_ridge,
    "residu_MCO": y_test - y_pred_lin,
    "residu_Ridge": y_test - y_pred_ridge
})
predictions.to_csv("predictions_test.csv", index=False)
```

5 Annexe E — Visualisations

Ce script génère les six figures principales du mémoire : distribution de la variable cible par région, boxplots du rendement, matrice de corrélation, relation irradiance/-rendement, prédictions vs valeurs réelles et distribution des résidus.

Listing 5 – 05_visualisations.py

```
"""
Generation des figures principales du memoire
"""
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

sns.set_style("whitegrid")
os.makedirs("figures", exist_ok=True)

df = pd.read_csv("dataset_clean.csv")
features = ["latitude", "longitude", "altitude",
            "irradiance", "temperature", "humidity",
            "wind_speed", "cloud_cover"]
ordre = ["cote", "hauts_plateaux",
          "sahara_nord", "sahara_sud"]


# === FIGURE 1 : Distribution du rendement par region ===
fig, ax = plt.subplots(figsize=(11, 5))
sns.histplot(data=df, x="yield_kwh", hue="region",
             hue_order=ordre, multiple="stack",
             bins=30, palette="viridis", ax=ax)
ax.set_xlabel("Rendement (kWh/kWc/an)")
ax.set_ylabel("Nombre de sites")
ax.set_title("Distribution du rendement par region")
plt.tight_layout()
plt.savefig("figures/01_distribution_rendement.png",
            dpi=150, bbox_inches="tight")
plt.close()


# === FIGURE 2 : Boxplots par region ===
```

```
fig, ax = plt.subplots(figsize=(10, 5))
sns.boxplot(data=df, x="region", y="yield_kwh",
            order=ordre, palette="viridis", ax=ax)
ax.set_title("Comparaison des rendements par region")
plt.tight_layout()
plt.savefig("figures/02_boxplot_regions.png",
            dpi=150, bbox_inches="tight")
plt.close()


# === FIGURE 3 : Matrice de correlation ===
fig, ax = plt.subplots(figsize=(10, 8))
corr = df[features + ["yield_kwh"]].corr()
sns.heatmap(corr, annot=True, fmt=".2f",
            cmap="RdBu_r", center=0,
            vmin=-1, vmax=1, square=True,
            linewidths=0.5, ax=ax)
ax.set_title("Matrice de correlation (Pearson)")
plt.tight_layout()
plt.savefig("figures/03_matrice_correlations.png",
            dpi=150, bbox_inches="tight")
plt.close()


# === FIGURE 4 : Irradiance vs rendement ===
fig, ax = plt.subplots(figsize=(10, 6))
for region in ordre:
    sub = df[df["region"] == region]
    ax.scatter(sub["irradiance"], sub["yield_kwh"],
               label=region, alpha=0.7, s=40)
coefs = np.polyfit(df["irradiance"], df["yield_kwh"], 1)
xs = np.array([df["irradiance"].min(),
               df["irradiance"].max()])
ax.plot(xs, np.polyval(coefs, xs), "r--", lw=1.5,
        label="Tendance lineaire")
ax.set_xlabel("Irradiance (kWh/m2/jour)")
ax.set_ylabel("Rendement (kWh/kWc/an)")
ax.set_title("Relation irradiance / rendement")
ax.legend()
plt.tight_layout()
plt.savefig("figures/04_irradiance_vs_rendement.png",
```

```
            dpi=150, bbox_inches="tight")
plt.close()


# === FIGURE 5 : Predictions vs valeurs reelles ===
predictions = pd.read_csv("predictions_test.csv")
fig, axes = plt.subplots(1, 2, figsize=(13, 5))
modeles = [
    ("y_pred_MCO", "Regression lineaire (MCO)", axes[0]),
    ("y_pred_Ridge", "Regression Ridge", axes[1])
]
for col_pred, titre, ax in modeles:
    ax.scatter(predictions["y_vrai"],
               predictions[col_pred],
               alpha=0.6, s=50, edgecolor="black")
    lims = [predictions["y_vrai"].min() - 50,
            predictions["y_vrai"].max() + 50]
    ax.plot(lims, lims, "r--", lw=1.5,
            label="Prediction parfaite")
    ax.set_xlabel("Valeur reelle")
    ax.set_ylabel("Valeur predite")
    ax.set_title(titre)
    ax.legend()
plt.tight_layout()
plt.savefig("figures/05_predictions_vs_reel.png",
            dpi=150, bbox_inches="tight")
plt.close()


# === FIGURE 6 : Distribution des residus ===
fig, axes = plt.subplots(1, 2, figsize=(13, 5))
for col, titre, ax in [
    ("residu_MCO", "Residus MCO", axes[0]),
    ("residu_Ridge", "Residus Ridge", axes[1])
]:
    sns.histplot(predictions[col], bins=15, kde=True,
                 ax=ax, color="steelblue",
                 edgecolor="black")
    ax.axvline(0, color="red", linestyle="--", lw=1.5)
    moy = predictions[col].mean()
    std = predictions[col].std()
```

```
    ax.text(0.05, 0.95,
            f"Moy : {moy:.2f}\nStd : {std:.2f}",
            transform=ax.transAxes,
            verticalalignment="top",
            bbox=dict(boxstyle="round",
                      facecolor="wheat", alpha=0.7))
    ax.set_title(titre)
plt.tight_layout()
plt.savefig("figures/06_distribution_residus.png",
            dpi=150, bbox_inches="tight")
plt.close()

print("Toutes les figures sont sauvegardees")
```

Note finale sur la reproductibilité

L'ensemble de ces scripts a été développé en Python 3.10 et testé avec les versions suivantes des bibliothèques principales : `numpy` 1.24+, `pandas` 2.0+, `scipy` 1.10+, `scikit-learn` 1.3+, `matplotlib` 3.7+, `seaborn` 0.12+.

La graine aléatoire fixée à `seed=42` dans tous les scripts garantit la reproductibilité exacte des chiffres présentés au Chapitre 4 du mémoire :

TABLE 10 – Résultats reproductibles des scripts

Métrique	**Régression linéaire**	**Régression Ridge**
MAE	15,01 kWh/kWc/an	15,02 kWh/kWc/an
RMSE	21,86 kWh/kWc/an	21,88 kWh/kWc/an
MAPE	0,90 %	0,90 %
R^2	0,9870	0,9870
λ optimal	—	0,069
R^2 validation croisée	$0{,}9901 \pm 0{,}0052$	$0{,}9901 \pm 0{,}0052$