Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Maxime Segueni

Titre du mémoire :

Cartographie du risque feu de forêt au Chili

Confidentialité : ☒ NON ☐OUI (Durée : ☐ 1 an ☐2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus.

Membres présents du jury de la filière : Signature :

Entreprise : Willis Towers Watson

Nom :

Signature :

Directeur de mémoire en entreprise

Membres présents du jury de l'Institut des Actuaires : Signature :

Nom : Luc Schwartzbrod

Signature :

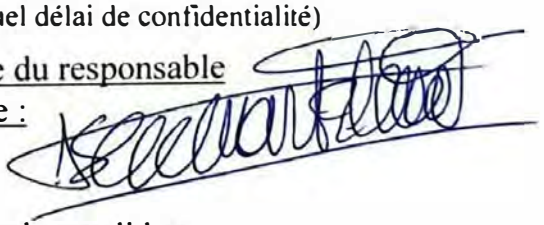
Invité :

Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise :



Signature du candidat :

Résumé

Les feux de forêt représentent un risque de plus en plus préoccupant pour les assureurs. Leur caractère imprévisible et potentiellement dévastateur pose des défis spécifiques pour le secteur de l'assurance, notamment en matière de prévision des sinistres et de gestion des coûts. Dans le contexte des couvertures d'assurance pour les biens naturels et agricoles, ce risque prend une place significative dans la prime pure observée, en raison des dommages souvent considérables qu'engendrent les incendies. Face à l'intensification des épisodes de sécheresse et l'allongement de la saison des feux de forêt observée dans le monde depuis les années 1980, il devient essentiel pour les assureurs de développer des modèles précis d'estimation du risque d'incendie afin d'optimiser la gestion de leur portefeuille et de proposer des solutions assurantielles adaptées aux risques pris.

Ce mémoire se propose de concevoir une approche de modélisation spécifique aux feux de forêt, appliquée au Chili central, une région particulièrement exposée. La méthodologie adoptée repose sur la collecte de données historiques des feux de forêt entre 1985 et 2018, ainsi que les caractéristiques géographiques et environnementales des zones étudiées. Une modélisation statistique est ensuite employée pour établir les liens existants entre les facteurs d'exposition et la probabilité d'occurrence d'incendies. Enfin, les résultats obtenus sont appliqués pour produire une carte de risque d'incendie.

Abstract

Wildfires represent an increasingly concerning risk for insurers. Their unpredictable and potentially devastating nature presents specific challenges for the insurance sector, particularly in the areas of claim forecasting and cost management. In the context of insurance coverage for natural and agricultural assets, this risk occupies a significant portion of the observed pure premium due to the often substantial damages caused by fires. In light of the intensification of drought episodes and the extended wildfire season observed in the world since the 1980s, it has become essential for insurers to develop accurate fire risk estimation models to optimize portfolio management and offer insurance solutions tailored to the level of risk involved.

This thesis aims to design a modeling approach specific to wildfire, applied to central Chile, a particularly vulnerable region. The adopted methodology is based on the collection of historical wildfire data from 1985 to 2018, as well as the geographical and environmental characteristics of the areas studied. Statistical modeling is then employed to establish the existing relationships between exposure factors and the probability of wildfire occurrence. Finally, the results are applied to produce a wildfire risk map.

Synthèse

Contexte de l'étude

Alors que l'incidence des catastrophes naturelles croît, les compagnies d'assurance développent des méthodes toujours plus sophistiquées pour estimer l'exposition aux sinistres. À la différence des risques traditionnels, les feux de forêt sont influencés par des facteurs environnementaux et climatiques complexes qui varient considérablement d'une région à l'autre. L'évolution rapide des technologies de collecte et d'analyse de données transforme cependant l'approche classique de mutualisation du risque en mettant à disposition des assureurs des outils de prédiction plus précis et plus accessibles qu'auparavant.

Le Chili est une région particulièrement exposée aux feux de forêt, où les défiés liés à ces feux et à la gestion forestière sont de dimension politique, sociale, environnementale et économique. Pour les assureurs et les exploitants agricoles en particulier, les risques d'incendie forestier représentent des enjeux majeurs : la vulnérabilité de leurs investissements ainsi que la durabilité de leurs activités dans un contexte de mégasécheresse sont directement affectées par ces événements dont l'ampleur record a marqué les esprits lors des saisons des feux 2017 et 2023.

Par ailleurs, le Chili étant un vaste pays avec une forte diversité de climats, de topographie et d'occupation des terrains, l'étude des feux de forêt y est multiforme, il est nécessaire de prendre en considération tous ces paramètres pour mieux gérer ce risque. C'est dans ce contexte qu'a été collecté un vaste ensemble de données à résolution fine afin d'entraîner un modèle prédictif basé sur la classification binaire du caractère brûlée ou non des parcelles de terrain chiliennes.

Démarche

Données collectées

Pour témoigner de la diversité des facteurs de risque d'incendie, les données climatiques, météorologiques, d'urbanisation et de classificaiton du terrain des régions centrales du Chili ont été collectées et traitées. Les cartes précises de zones brûlées après le passage de près de 8153 feux enregistrés entre 1985 et 2018 mesurés par satellites sont consolidées. Les données finales sont obtenues par le rassemblement de l'ensemble au même format et à la même résolution de 30m^2. Ces données ont aussi été moyennées et rassemblées à divers résolutions administratives avec des cartes par commune, par province et pour finir par région.

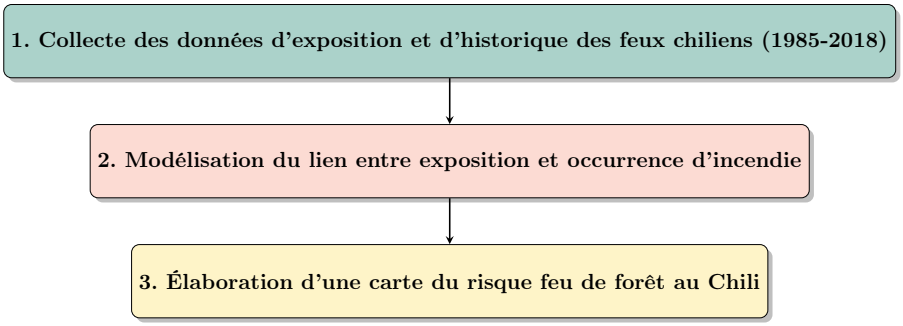


FIGURE 1 – Approche adoptée pour la modélisation du risque feu de forêt

Classification des communes

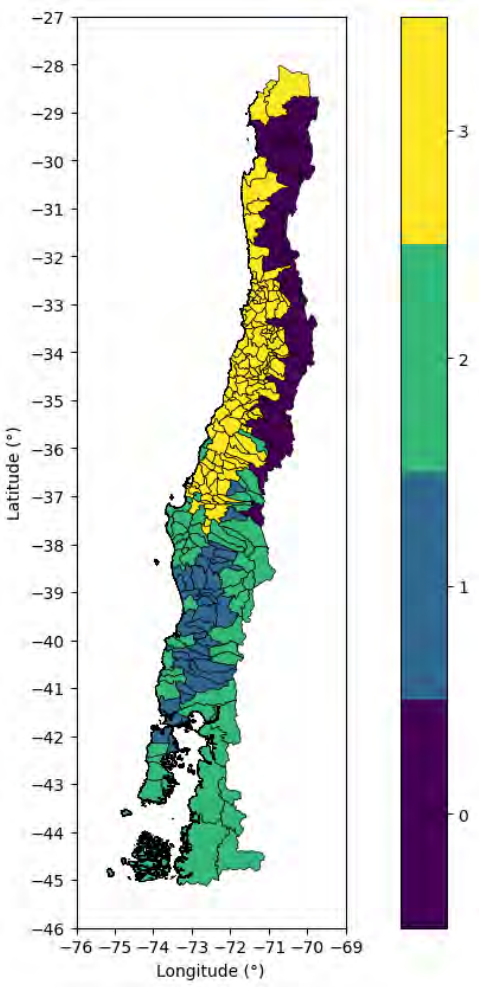


FIGURE 2 – Carte des communes colorées par cluster.

Afin d'améliorer la précision et la capacité des modèles à généraliser les observations, les communes ont été classées en groupes homogènes en fonction de leurs caractéristiques. Dans un premier temps, une analyse en composantes principales (PCA) a permis de réduire la dimensionnalité des données tout en conservant l'essentiel de l'information. Cette réduction à quatre

dimensions a simplifié l'analyse en isolant les principaux axes de variation des données. Ensuite, la méthode de classification des k-means a été appliquée sur les composantes principales obtenues, afin de regrouper les communes en quatre clusters distincts. Le résultat final est une démarcation nette entre les communes bien que le critère spatial n'ait pas été inclus dans les caractéristiques de clustering.

Modélisation

Deux modèles ont été choisis pour modéliser le risque de feu de forêt. Le GLM, bien adapté aux relations linéaires, repose sur une fonction de lien qui relie les variables explicatives (comme le climat et la végétation) aux probabilités d'incendie. Ce modèle est particulièrement efficace pour interpréter l'impact spécifique de chaque variable, ce qui en fait un outil transparent pour comprendre les facteurs de risque. À l'inverse, le random forest est un modèle d'ensemble non linéaire qui combine plusieurs arbres de décision pour capturer des interactions complexes dans les données. Cela lui permet de modéliser avec précision des relations plus subtiles et d'identifier des zones à haut risque, même lorsque les interactions entre variables sont non linéaires. Le random forest est également plus résistant au surapprentissage et gère mieux les données avec des distributions déséquilibrées, comme celles des feux de forêt.

L'approche suivie pour chaque modèle comprend plusieurs étapes. D'abord, les communes sont divisées en groupes d'entraînement et de test. Les variables explicatives sont normalisées pour le GLM, et le random forest est ajusté sans nécessiter de normalisation. La performance de chaque modèle est évaluée par des métriques comme le F1-score, la précision, le rappel, et l'AUC, permettant de mesurer la justesse des prédictions et d'équilibrer les erreurs entre les classes brûlées et non-brûlées.

Comparaison et résultat

Le random forest permet de produire une carte des probabilités d'incendie par cellule de 30 m^2 dans la région de la Araucania. La fréquence est ajustée pour témoigner de la moyenne de sinistralité mesurée par satellite dans la région sur cinq ans.

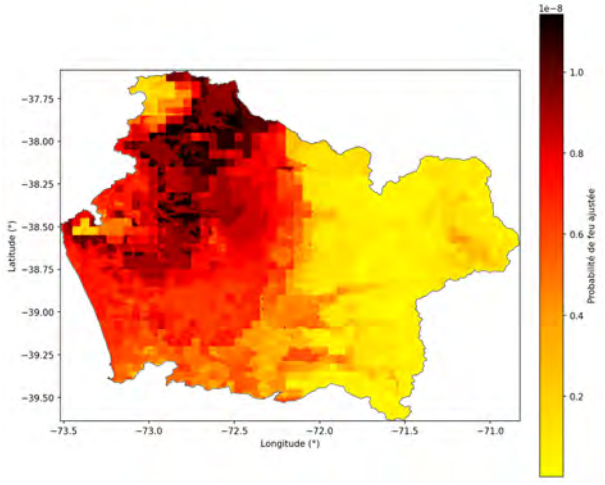


FIGURE 3 – Carte de fréquence de feu de la Araucania

Le GLM permet l'élaboration d'une carte où les quatre clusters sont modélisés séparément.

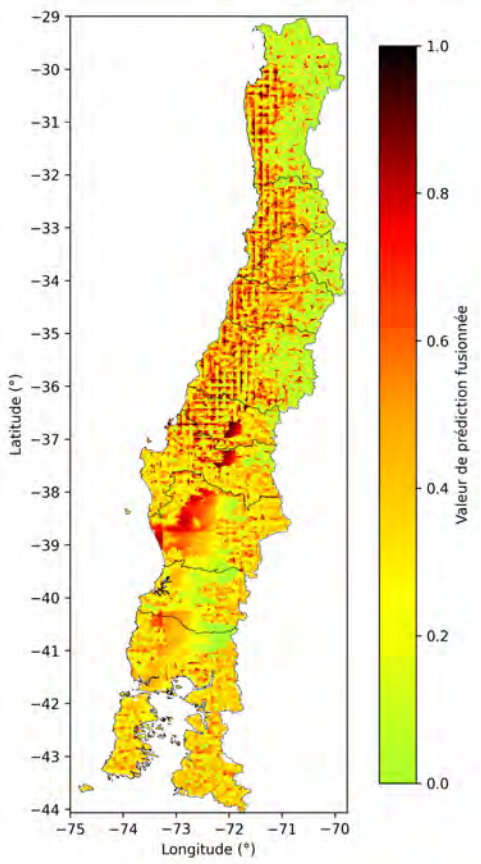


FIGURE 4 – Carte de fréquence consolidée du modèle GLM

Summary

Study background

With the rising incidence of natural disasters, insurance companies are developing increasingly sophisticated methods to estimate loss exposure. Unlike traditional risks, forest fires are influenced by complex environmental and climatic factors that vary significantly between regions. However, progress in data collection and analysis technology are transforming the classic approach to risk pooling, offering insurers more precise and accessible prediction tools than ever before.

Chile is particularly vulnerable to forest fires, where the challenges related to fires and forest management are political, social, environmental, and economic in scope. Forest fire risks represent major concerns for insurers and agricultural operators : the vulnerability of their investments and the sustainability of their activities in the context of a megadrought are directly affected by these events, whose record-breaking magnitude during the fire seasons of 2017 and 2023 left a lasting impact.

Moreover, given Chile's vast expanse and diversity in climate, topography, and land use, the study of forest fires here is multifaceted, requiring all these parameters to be considered to better manage risk. Within this context, a comprehensive dataset of fine-resolution data was collected to train a predictive model, based on a binary classification approach, to assess whether Chilean land parcels are likely to be burned or not.

Approach

Data collection

To capture the diversity of fire risk factors, climatic, meteorological, urbanization, and land classification data for central Chile were collected and processed. High-resolution burn area maps, covering approximately 8,153 fires recorded between 1985 and 2018 as measured by satellites, were consolidated. Final data were compiled uniformly at a 30 m^2 resolution and averaged to various administrative levels, producing maps for each commune, province, and region.

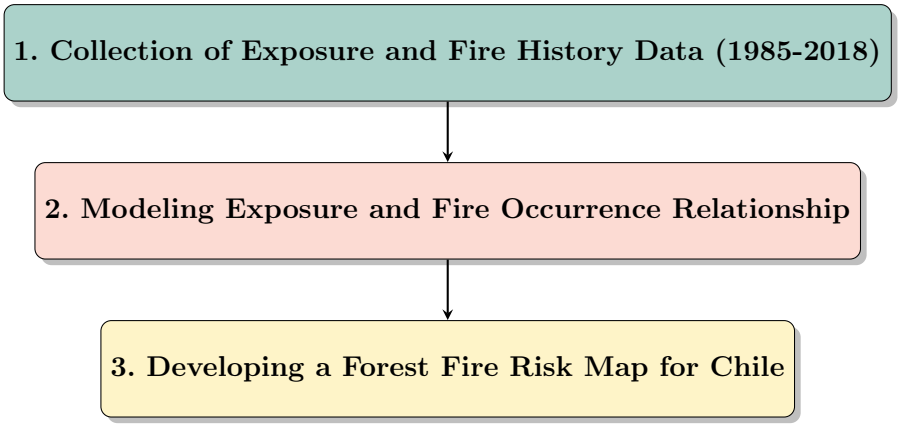


FIGURE 5 – Approach for Forest Fire Risk Modeling

Commune classification

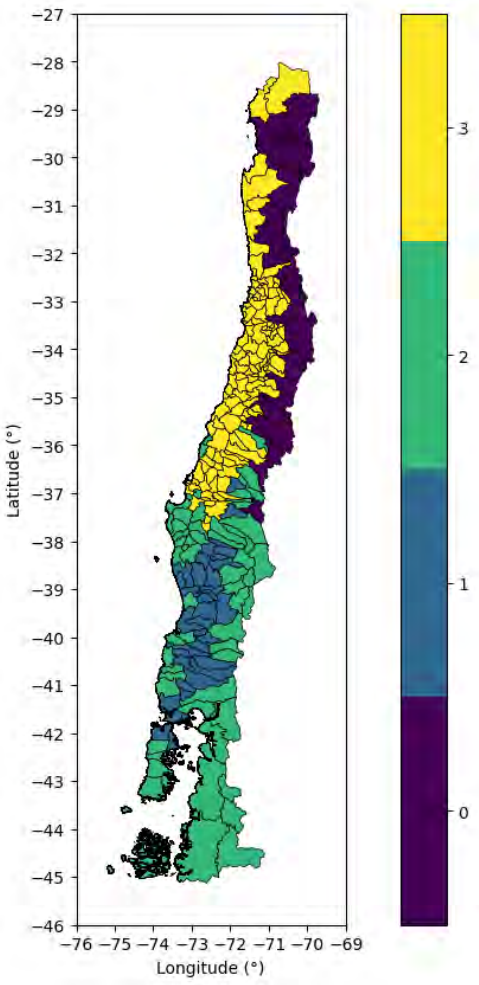


FIGURE 6 – Map of communes color-coded by cluster.

To enhance model accuracy and generalization capabilities, communes were classified into homogeneous groups based on their characteristics. Initially, a principal component analysis (PCA) was conducted to reduce data dimensionality while retaining essential information. This reduction to four dimensions simplified the analysis by isolating the primary axes of data variability. The k-means clustering method was then applied to the resulting principal components, grouping communes into four distinct clusters. The final outcome is a clear delineation among communes, even though spatial criteria were not included in the clustering characteristics.

Modeling

Two models were chosen to simulate forest fire risk. The generalized linear model (GLM), well-suited to linear relationships, relies on a link function connecting explanatory variables (such as climate and vegetation) to fire probabilities. This model is particularly effective for interpreting the specific impact of each variable, making it a transparent tool for understanding risk factors. In contrast, the random forest is a nonlinear ensemble model combining multiple decision trees to capture complex data interactions. It can accurately model more subtle relationships and identify high-risk zones, even when variable interactions are nonlinear. The random forest is also more resistant to overfitting and handles imbalanced data, such as forest fire events, more effectively.

The approach for each model involved several steps. First, the communes were divided into training and testing groups. Explanatory variables were normalized for the GLM, while the random forest was adjusted without requiring normalization. Model performance was evaluated using metrics such as F1-score, precision, recall, and AUC, which measure prediction accuracy and balance errors between burned and unburned classes.

Comparison and results

The random forest model enabled the creation of a fire probability map at a 30 m^2 resolution for the Araucania region, adjusted to reflect the average fire frequency measured by satellite over five years.

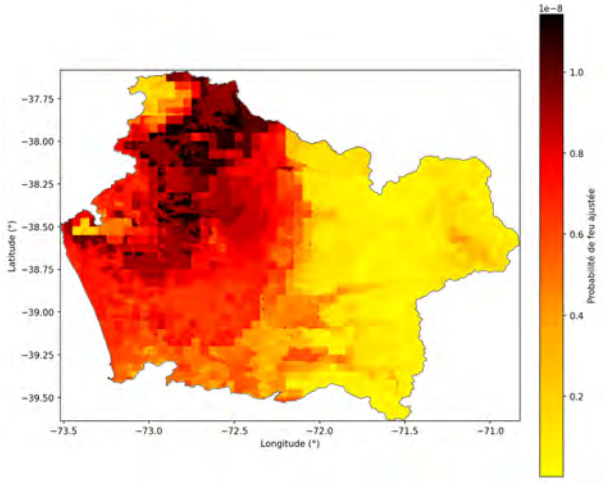


FIGURE 7 – Fire frequency map for the Araucania region

The GLM enabled the development of a map where the four clusters are modeled separately.

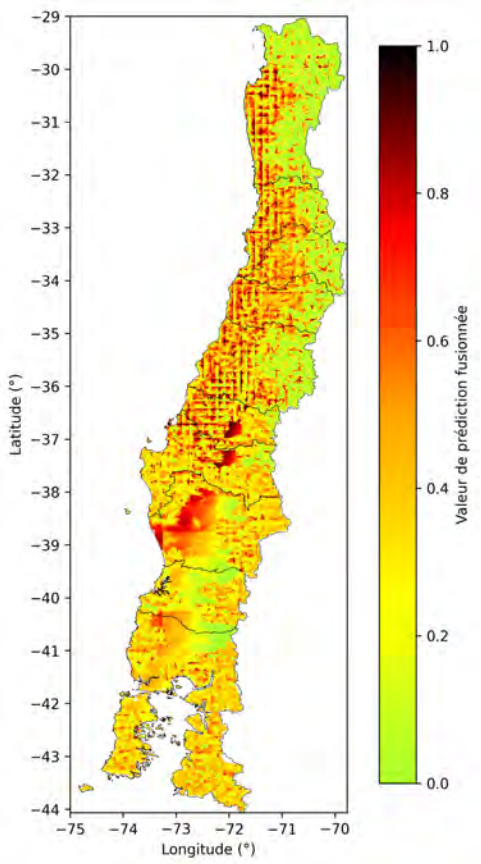


FIGURE 8 – Consolidated frequency map from the GLM model

Remerciements

Je souhaite tout d'abord remercier l'équipe *Alternative Risk Transfer* de *Willis Towers Watson* pour tous ces moments enrichissants et stimulants passés durant l'alternance. Je tiens en particulier à remercier Luc Schwartzbrod pour son aide, son accompagnement et surtout son soutien tout au long des travaux, Aurélien Schwachtgen pour son expertise avisée et Meriem Hafid pour ses conseils et sa bonne humeur.

Je remercie également l'équipe *ART* de *WTW London* qui m'a prodigué de nombreux conseils dans l'établissement de ce projet.

Enfin, merci à ceux qui ont contribué, de près ou de loin, à l'accomplissement de ce mémoire.

Table des matières

Introduction **14**

1 Contexte de l'étude et approche **15**
1.1 Généralités sur les feux de forêt . . . 15
1.1.1 Déclenchement et types des feux de forêt . . . 15
1.1.1.1 Conditions de déclenchement . . . 15
1.1.1.2 Les différents types de feux . . . 16
1.1.2 Origine des feux de forêt . . . 17
1.1.2.1 Causes naturelles . . . 17
1.1.2.2 Causes anthropiques . . . 17
1.1.3 Facteurs de risque et stratégies de prévention . . . 18
1.1.3.1 Facteurs environnementaux . . . 18
1.1.3.2 Facteurs humains . . . 19
1.2 Le cas du Chili . . . 20
1.2.1 Les forêts chiliennes, entre expansion et déforestation . . . 20
1.2.2 Les feux de forêt chiliens . . . 21
1.2.2.1 Origine des incendies . . . 21
1.2.2.2 L'influence du phénomène El Niño . . . 22
1.2.2.3 Enjeux . . . 23
1.2.2.4 Couverture assurantielle . . . 24
1.3 L'assurance des feux de forêt . . . 24
1.3.1 Mesure de sinistre : les aires brûlées . . . 24
1.3.1.1 Normalized Burn Ratio (NBR) . . . 25
1.3.1.2 Relative delta Normalized Burn Ratio (RdNBR) . . . 27
1.3.1.3 Normalized Difference Vegetation Index (NDVI) . . . 27
1.3.2 Les solutions d'assurance . . . 27
1.3.3 La modélisation des feux de forêt . . . 28
1.3.3.1 Le modèle catastrophe . . . 28
1.3.3.2 Démarche adoptée . . . 29

2 Constitution de la base de données **31**
2.1 Cartographie . . . 31
2.1.1 Sélection du périmètre . . . 31
2.1.2 Choix du système de projection . . . 31
2.1.3 Cartes des régions administratives du Chili . . . 32
2.1.4 Fichiers raster . . . 33
2.2 Données des feux historiques . . . 33
2.2.1 Répertoire de la CONAF . . . 34

2.2.2 La base de données Landscape Fire Scars . . . 34
2.2.2.1 Constitution de la base . . . 35
2.2.2.2 Attributs . . . 35
2.2.2.3 Traitement des données . . . 36
2.3 Données environnementales et urbaines . . . 36
2.3.1 Données climatiques et météorologiques . . . 36
2.3.1.1 Température et Précipitations . . . 36
2.3.1.2 Indices météo . . . 37
2.3.1.3 Irradiance . . . 39
2.3.1.4 Impacts de foudre . . . 40
2.3.2 Données topographiques . . . 40
2.3.2.1 Altitude et Pente . . . 40
2.3.3 Couverture du sol . . . 42
2.3.4 Données d'urbanisation . . . 44
2.3.4.1 Données démographiques . . . 44
2.3.4.2 LitPop . . . 44
2.4 Données finales . . . 45
2.4.1 Homogénéisation des formats . . . 45
2.4.2 Distance aux zones urbaines et aux points de feu . . . 46
2.4.3 Récapitulatif des données . . . 46
2.5 Analyse préliminaire . . . 47
2.5.1 Échelle parcellaire . . . 47
2.5.2 Échelle communale . . . 49
2.5.3 Échelle provinciale . . . 51
2.5.4 Échelle régionale . . . 53

3 Modélisation du risque de feu de forêt 55
3.1 Classification des communes . . . 55
3.1.1 Analyse en composantes principales . . . 55
3.1.2 La méthode des k-means . . . 57
3.1.3 Application aux communes . . . 58
3.2 Méthode de GLM . . . 62
3.2.1 Théorie . . . 62
3.2.2 Mesures de performance . . . 63
3.2.3 Mise en œuvre du modèle GLM . . . 66
3.3 Méthode de Random Forest . . . 71
3.3.1 Théorie . . . 71
3.3.2 Mise en œuvre du Random Forest . . . 73
3.4 Calcul de prime . . . 76

Annexe 81

Introduction

Avec de nouvelles recherches montrant que le monde pourrait atteindre un réchauffement de 2,7°C d'ici la fin de ce siècle, le changement climatique entraîne des feux de forêt plus grands et plus fréquents dans une grande partie du monde. Depuis les années 1980, la saison des feux de forêt s'est allongée de 27% à l'échelle mondiale, avec des augmentations particulièrement prononcées en Amazonie, en Méditerranée et dans les forêts de l'Ouest de l'Amérique du Nord. Des périodes prolongées de sécheresse sévère et d'humidité faible augmentent la durée de la saison des feux de forêt à un rythme nettement plus rapide que ce qui était initialement prévu par les modèles climatiques.

La nécessité de maîtriser le risque devient primordial, en particulier dans des zones fortement exposées comme le Chili central. Les feux de forêt représentent une menace pour les écosystèmes, les infrastructures et les communautés locales, nécessitant des approches prédictives et réactives pour anticiper et atténuer les pertes potentielles. Dans ce contexte, des modèles permettant de prédire le risque de feux de forêt à partir de données climatiques et géographiques peuvent s'avérer utiles pour améliorer la résilience des zones vulnérables.

Pour établir et construire un tel modèle, les données couvrant la période 1985-2018 sur les variables de risque ont été recueillies dans les régions centrales du Chili, englobant un large éventail de conditions environnementales et humaines. Ces données ont ensuite été adaptées à une résolution de $30m^2$ puis exploitées à l'aide des méthodes de GLM dans un premier temps puis de random forest, reconnues pour leur capacité à classifier efficacement les zones.
Après avoir décrit le contexte des feux de forêt au Chili et analysé les variables disponibles et leurs corrélations avec les incidents passés, la méthode de modélisation sera détaillée en explicitant les ajustements appliqués pour optimiser les méthodes. Enfin, les cartes seront dressées.

Chapitre 1

Contexte de l'étude et approche

Cette première section vise à introduire le cadre de l'étude sur le risque d'incendies de forêt au Chili, en présentant le contexte global, les facteurs de risque, les implications pour le secteur de l'assurance et les approches de modélisation. Cette analyse initiale permettra de justifier les choix méthodologiques et les approches spécifiques qui seront développés dans les sections suivantes. Pour cela, les différentes dimensions du risque de feu de forêt seront tout d'abord examinées, en identifiant les causes naturelles et anthropiques de l'origine des incendies, les types d'incendies et leurs processus de propagation. Ensuite, l'analyse se concentrera sur le cas spécifique du Chili, avec une étude des écosystèmes forestiers et des facteurs locaux influençant l'occurrence des incendies. Enfin, l'exploration des stratégies d'assurance associées aux feux de forêt, les méthodes d'évaluation des sinistres et les approches de modélisation appliquées seront abordées.

1.1 Généralités sur les feux de forêt

1.1.1 Déclenchement et types des feux de forêt

1.1.1.1 Conditions de déclenchement

Pour qu'un feu de forêt puisse se déclencher, il est nécessaire que trois éléments soient réunis, souvent appelés le « triangle du feu » : un combustible, un comburant (naturellement présent dans l'atmosphère sous forme d'oxygène) et une source de chaleur ou d'ignition. Ces trois composantes interagissent pour déclencher et entretenir un feu, chacun jouant un rôle essentiel dans le processus de combustion.

— **Combustible** : Le combustible désigne la matière inflammable qui alimente le feu et lui permet de se propager. En milieu forestier, il s'agit principalement de végétation : feuilles mortes, branches, herbes sèches, arbres et broussailles. Cependant, tous les combustibles ne présentent pas le même risque d'inflammabilité. Leur propension à brûler dépend de leur type, de leur teneur en humidité et de leur disposition au sol. Une végétation dense et sèche constitue un combustible de haute qualité, car elle s'enflamme rapidement et contribue à intensifier le feu.

— **Comburant** : Le comburant, principalement constitué d'oxygène dans l'atmosphère, alimente la réaction chimique de combustion. L'air ambiant contient environ 21% d'oxygène, un taux suffisant pour soutenir la plupart des incendies naturels. La circulation de l'air, notamment par le vent, peut intensifier l'apport en comburant et ainsi favoriser la propagation du feu.

— **Source d'ignition** : La source d'ignition, ou de chaleur, est l'élément déclencheur initial, sans lequel le feu ne pourrait débuter. Elle peut provenir de phénomènes naturels (comme la foudre) ou d'activités humaines. Lorsqu'une source d'ignition entre en contact avec un combustible suffisamment sec, la chaleur dégagée initie la combustion, transformant ainsi le combustible en gaz et libérant de l'énergie sous forme de flammes.

FIGURE 1.1 – Le triangle du feu

1.1.1.2 Les différents types de feux

On distingue généralement trois types de feux de forêt en fonction de la zone qu'ils brûlent :

- **les feux de sol** : ce sont des incendies souterrains se propageant dans les couches de matière organique du sol, comme les racines, les tourbes et l'humus, riches en carbone et hautement inflammables une fois secs. Ils se caractérisent par une faible intensité mais une propagation lente et continue. Cette progression, couplée à leur invisibilité en surface, les rend difficiles à détecter et à éteindre. En l'absence de flammes apparentes, ils peuvent rester actifs pendant des semaines, des mois, voire des années, constituant un risque latent pour les écosystèmes souterrains. En raison de la difficulté d'accès aux combustibles souterrains, leur extinction est complexe et souvent peu efficace. Les régions boréales, comme le Canada, la Russie et l'Alaska, ainsi que les zones tropicales de tourbières en Indonésie, sont particulièrement vulnérables à ces feux de sol.
- **les feux de surface** : ces feux brûlent principalement la végétation au sol comme les herbes sèches, buissons, arbustes et débris organiques (feuilles mortes, branches) et représentent l'un des types d'incendies les plus courants dans les écosystèmes forestiers. Évoluant à faible hauteur, ils trouvent facilement prise dans des combustibles accessibles et s'étendent rapidement lorsque les conditions sont favorables, en particulier en présence de vent et de faibles niveaux d'humidité.
- **les feux de cimes** : ce type de feu se propage dans la canopée des arbres, affectant la strate supérieure de la forêt où se trouvent les branches et le feuillage des grands arbres. Les feux de cimes sont souvent déclenchés par des feux de surface qui atteignent les branches basses et évoluent de manière rapide et virulente surtout lorsque des vents forts facilitent leur progression à travers les cimes. En raison de leur intensité élevée et de leur propagation

rapide, ils sont particulièrement difficiles à maîtriser, nécessitant des interventions aériennes et des techniques de lutte spécialisées.

1.1.2 Origine des feux de forêt

1.1.2.1 Causes naturelles

Les feux de forêt peuvent être déclenchés par des causes naturelles, la foudre étant la plus fréquente et la plus significative. Environ 20% des incendies forestiers seraient déclenchés par des impacts de foudre. Cependant, toutes les activités orageuses ne se traduisent pas automatiquement par un risque accru de départ de feu, car la probabilité qu'un impact de foudre déclenche un incendie dépend de nombreux facteurs environnementaux. La foudre, en concentrant une quantité intense de chaleur en un point précis, peut enflammer des matériaux combustibles, en particulier lorsque les conditions de sécheresse augmentent la vulnérabilité de la végétation. Ce risque est amplifié pendant les périodes chaudes, lorsqu'une forêt déjà asséchée devient plus susceptible de s'embraser sous l'effet de l'énergie dégagée par un éclair.

Outre la foudre, les éruptions volcaniques constituent une autre cause naturelle, bien que moins fréquente, de feux de forêt, en particulier dans les régions situées à proximité de volcans actifs. Lors d'une éruption, les coulées de lave, les projections de cendres incandescentes et les débris en fusion peuvent facilement enflammer la végétation environnante. Les incendies causés par des éruptions volcaniques ont souvent une portée considérable, car les débris en fusion peuvent être projetés sur de grandes distances, créant des points de feu multiples et simultanés autour du site volcanique. Dans ces contextes, les incendies ainsi déclenchés sont difficilement contrôlables et peuvent entraîner des dégâts considérables sur les écosystèmes forestiers et les zones habitées adjacentes.

1.1.2.2 Causes anthropiques

Les activités humaines sont, de loin, les principales causes des feux de forêt. En effet, les études montrent que la majorité des incendies forestiers sont d'origine anthropique. En France, par exemple, l'Institut National de l'Information Géographique et Forestière (IGN) estime que 90 % des départs de feux de forêt ont pour origine les activités humaines [4]. Les causes sont variées :

1. **Malveillance** : Certains feux de forêt sont allumés intentionnellement par malveillance, motivés par des actes de vandalisme ou des conflits locaux. Ces incendies délibérés sont souvent déclenchés pour causer des dégâts ou manifester une opposition.
2. **Travaux et infrastructures** : Les lignes électriques défectueuses, les équipements industriels en forêt et les étincelles générées par des machines lors de travaux représentent des risques d'embrasement.
3. **Pratiques d'exploitation agricole** : Les incendies peuvent aussi provenir de pratiques agricoles. L'écobuage, utilisé pour brûler des résidus de culture ou défricher la végétation pour créer de nouvelles terres agricoles, peut facilement devenir incontrôlable.
4. **Négligence** : De nombreux incendies proviennent de négligences individuelles : mégots de cigarette, feux de camp non surveillés, feux d'artifice, ou barbecues près des zones boisées.

1.1.3 Facteurs de risque et stratégies de prévention

L'apparition et la propagation des feux de forêt sont fortement influencées par des facteurs environnementaux et humains, même s'il existe des facteurs de mitigation.

1.1.3.1 Facteurs environnementaux

Facteurs climatiques

Les incendies de forêt sont largement favorisés par des conditions climatiques spécifiques, souvent interconnectées. Les périodes de sécheresse prolongée assèchent la végétation et les sols, rendant les combustibles naturels plus inflammables. Les vagues de chaleur et températures extrêmes accélèrent ce processus en abaissant le seuil d'ignition des végétaux, surtout en canicule, où les températures nocturnes restent élevées, empêchant la réhydratation des plantes. L'humidité relative de l'air joue également un rôle : en été, de faibles taux d'humidité facilitent le dessèchement des matières organiques. Parallèlement, l'évapotranspiration, intensifiée par la chaleur, augmente le stress hydrique des plantes en accélérant la perte d'eau des sols, créant un environnement inflammable. La faible fréquence des précipitations accentue ces conditions, et plusieurs saisons sèches successives rendent les écosystèmes encore plus vulnérables.

Topographie

La morphologie d'une zone forestière modifie les microclimats et influence la répartition des combustibles, créant des zones à risque inégal. Les versants exposés au soleil reçoivent plus de rayonnement, asséchant la végétation plus rapidement. Les pentes inclinées facilitent la montée des flammes, car la chaleur préchauffe les végétaux en amont, tandis que les zones en altitude, bien que plus fraîches, sont exposées à des vents plus forts, dispersant les braises sur de longues distances. De plus, les couloirs naturels, comme les canyons, canalisent les vents et accélèrent la propagation dans des directions prédominantes, amplifiant ainsi les risques.

Structure et composition de la végétation

La nature et la densité de la végétation influencent fortement le risque d'embrasement dans les zones forestières, avec certaines espèces végétales comme les eucalyptus, les graminées inflammables, les résineux riches en composés volatils, ou les espèces à feuillage persistant telles que les pins qui détiennent une charge de combustible importante, augmentant le potentiel d'inflammation. Les végétaux naturellement secs dans les climats arides renforcent cette couverture inflammable au sol, rendant certaines zones plus propices au départ de feu. La densité du sous-bois et la présence de débris végétaux – feuilles mortes, branches et herbes sèches – alimentent encore davantage la charge de combustible, tandis qu'une canopée dense favorise la propagation verticale. De plus, la structure en strates de la végétation (herbes basses, arbustes, arbres) forme un réseau continu de combustibles qui facilite l'intensification du feu. Des couches épaisses d'humus sec contribuent également à la persistance des feux de sol, compliquant leur extinction, et la continuité horizontale des combustibles au sol, sans interruption naturelle, permet une propagation ininterrompue des flammes.

1.1.3.2 Facteurs humains

La multiplication des interfaces forêt-urbain

Une interface forêt-urbain (WUI, pour Wildland-Urban Interface) désigne une zone de transition où les espaces naturels non aménagés, comme les forêts, prairies et autres types de végétation, rencontrent les infrastructures humaines, notamment les zones résidentielles, commerciales et industrielles. Ces interfaces sont caractérisées par une forte proximité entre la végétation inflammable et les habitations ou bâtiments, ce qui les rend particulièrement vulnérables aux départs de feu et à la propagation rapide des incendies. La croissance démographique et l'urbanisation progressive des WUI augmentent l'interaction entre les espaces forestiers et les constructions humaines, accentuant ainsi les risques d'incendie et les défis associés à leur gestion. Dans ces interfaces, l'absence de zones tampons ou de bandes débroussaillées entre la forêt et les constructions résidentielles favorise une propagation directe des flammes.
Par ailleurs, les infrastructures humaines en zones boisées amplifient considérablement le risque d'incendie. Les lignes électriques peuvent déclencher des feux en cas de défaillance. Les chemins de fer, les routes forestières et décharges en périphérie facilitent l'accès aux zones naturelles et abritent des matières inflammables, augmentant ainsi le risque de départs de feu.

Les méthodes d'exploitation agricole

Certaines pratiques agricoles et forestières fréquentes dans certaines régions du globe économiquement fragiles augmentent les risques d'incendie en zones boisées. En Amazonie, le défrichement par brûlage, utilisé pour libérer des terres agricoles, devient incontrôlable en l'absence de mesures de gestion adéquates. En Asie du Sud-Est, comme en Indonésie et Malaisie, des méthodes similaires sont fréquentes pour l'expansion des plantations de palmiers à huile, provoquant souvent des feux non maîtrisés. L'écobuage, courant dans les pays méditerranéens comme l'Espagne et la Grèce, consiste à brûler les résidus de culture après récolte et est une source régulière d'incendies, particulièrement en période estivale. La déforestation par brûlage en Amazonie et dans le bassin du Congo réduit les barrières naturelles et rend les zones voisines plus vulnérables. Cette pratique, en transformant les paysages en surfaces inflammables, accroît le risque de propagation rapide des feux. Par ailleurs, les plantations inflammables, telles que les eucalyptus en Australie et au Portugal, libèrent des composés volatils sous l'effet de la chaleur, intensifiant les flammes. Enfin, les résidus de cultures saisonnières, comme le blé et le maïs, fréquemment présents en Amérique du Nord, constituent une charge combustible au sol après récolte, particulièrement dangereuse en période de sécheresse.

Les tensions locales

Les tensions géopolitiques, économiques et sociales augmentent les risques d'incendies de forêt, notamment dans les régions où l'utilisation des terres et l'accès aux ressources naturelles sont sources de conflits. Les divergences d'intérêts entre gouvernements, exploitants agricoles et communautés locales amènent parfois à recourir au feu pour contester les politiques ou marquer des revendications territoriales. Dans les zones où l'accès à la terre est limité, les tensions entre agriculteurs et éleveurs nomades, en particulier pour l'eau et les terres fertiles, entraînent aussi des incendies intentionnels pour éloigner d'autres utilisateurs. Les politiques de conservation, qui imposent des restrictions pour protéger les écosystèmes (comme l'interdiction de défricher ou de cultiver dans certaines zones), sont également perçues comme une menace par les populations locales. Pour des communautés dépendantes de l'agriculture ou du pâturage, ces politiques

imposent des contraintes à leur subsistance. En Amazonie, par exemple, les restrictions de défrichement sont parfois contournées par des brûlis illégaux pour protester contre des réglementations jugées restrictives. En allumant des feux, certains agriculteurs et éleveurs expriment ainsi leur désaccord, notamment dans les zones sous réglementation stricte. Par ailleurs, la pression urbaine fragilise les écosystèmes périurbains : le défrichement pour le développement immobilier, motivé par la demande de logements, accroît la vulnérabilité de ces zones aux incendies.

1.2 Le cas du Chili

Le Chili s'étend sur 4 620 kilomètres de long et 39° de latitude, en faisant le pays le plus long sur l'axe Nord-Sud. Son climat et sa géographie sont très variés, des déserts arides du nord vers le froid polaire de l'antarctique, en passant par les régions plus tempérées du Chili central. Il abrite aussi la Cordillère des Andes, une haute chaîne de montagne. Ce pays de 19,5 millions d'habitants et 756 106 km^2 de superficie possède une couverture forestière importante, en particulier dans les régions du centre-sud, où se trouvent les forêts tempérées humides et les plantations d'espèces exotiques comme les pins et les eucalyptus, largement utilisées dans l'industrie forestière. La forte concentration urbaine est également un trait distinctif du Chili, avec près de 88% des habitants vivant dans des villes, principalement dans des zones métropolitaines comme Santiago, Valparaíso et Concepción. Cette urbanisation rapide a entraîné un étalement urbain, où les villes s'étendent vers les périphéries, empiétant souvent sur les terres agricoles et forestières.

1.2.1 Les forêts chiliennes, entre expansion et déforestation

La surface de forêt a augmenté durant les deux dernières décennies au Chili. Cette expansion est le fruit de politiques gouvernementales encourageant la résolution des problèmes socio-économiques et environnementaux liée à une mauvaise gestion des forêts. De nombreuses tensions ont en effet émergé avec notamment la population indigène Mapuche qui revendique des droits sur certains territoires où le défrichage a eu lieu pour l'agriculture et l'exploitation industrielle. Des plantations ont ainsi été mises en place afin de pallier la dégradation des forêts primaires. Cependant, la perte de biodiversité liée à cette nouvelle végétation moins diversifiée a des conséquences négatives multiples telles que de l'érosion, un accès plus difficile à l'eau ainsi qu'une augmentation du risque incendie de forêt.

Le rôle de la CONAF

La Corporación Nacional Forestal, plus connue sous le nom de CONAF, est une institution chilienne de nature privée, mais qui opère sous l'égide du Ministère de l'Agriculture. Son rôle principal est de gérer et administrer les politiques forestières du Chili. Son objectif affiché est de soutenir le développement territorial, le bien-être des communautés autochtones et vulnérables, et à valoriser la biodiversité, en particulier face à la crise climatique. Pour cela, elle se fixe des objectifs tels que la conservation des écosystèmes, la promotion d'une gestion forestière durable, la création d'espaces verts, le dialogue politique sur le changement climatique, et l'instauration de collaborations dans le domaine forestier, tout en garantissant une certaine transparence.

La CONAF joue également un rôle dans la surveillance, la prévention et la lutte contre les incendies de forêt au Chili, en collectant des données exhaustives sur chaque incendie pour une gestion optimale des risques.

1.2.2 Les feux de forêt chiliens

Au Chili, la saison des feux s'étend principalement de décembre à mars, coïncidant avec l'été austral. Durant cette période, les régions centrales et méridionales, comme le Biobío, l'Araucanía et le Maule, présentent un risque accru en raison de leur végétation dense et des conditions météorologiques caractéristiques : sécheresse prolongée, vent fort et chaleur intense. Ces régions, où se concentrent des plantations de pins et d'eucalyptus, deviennent particulièrement inflammables.

1.2.2.1 Origine des incendies

L'activité orageuse étant limitée dans le pays, les causes humaines de déclenchement sont prévalentes au Chili. Ainsi d'après la CONAF, 88% des feux de forêt seraient d'origine accidentelle, 11% d'origine inconnue et 1% d'origine naturelle.

FIGURE 1.2 – Origine des incendies au Chili entre 1985 et 2018

Les causes accidentelles sont principalement liées au transit de personnes et de véhicules (30%), aux voies ferrées (27%) et aux activités récréatives (16%) tandis que celles intentionnelles sont liées au vandalisme, à la pyromanie et aux conflits indigène. Ces chiffres varient fortement entre les régions, en particulier entre les parts d'incendies volontaires et accidentels.

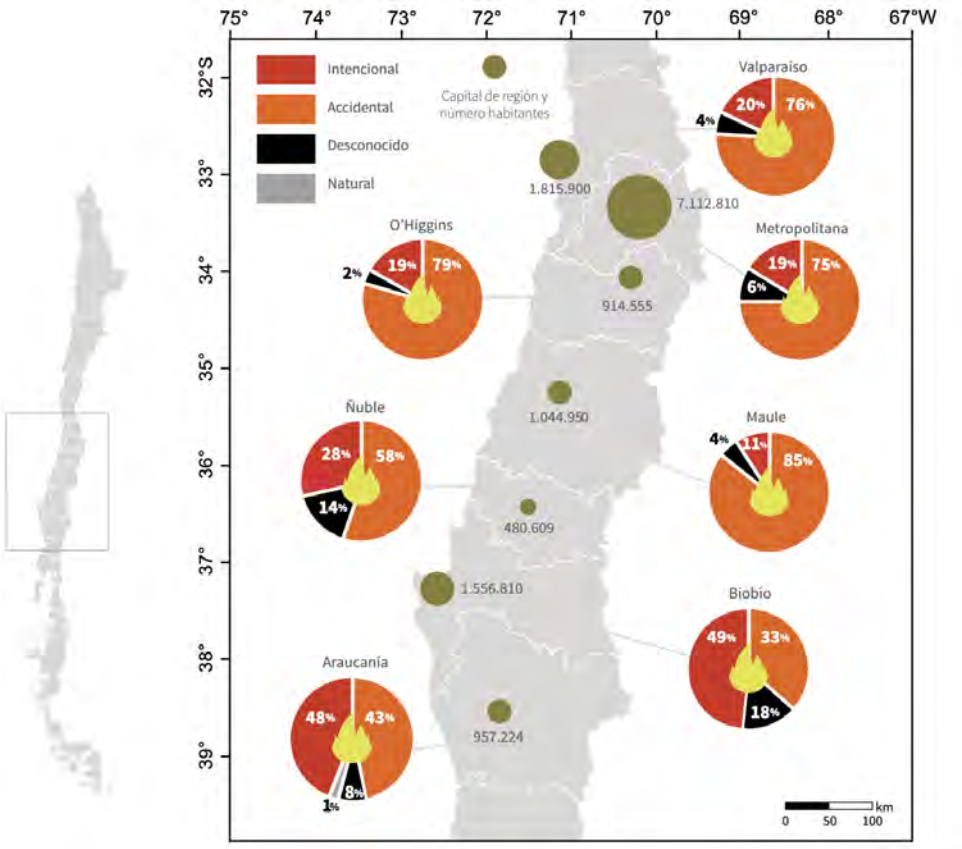


FIGURE 1.3 – Origine des incendies au Chili entre 1985 et 2018 entre les régions de Valparaiso et de la Araucanía

1.2.2.2 L'influence du phénomène El Niño

L'Oscillation Australe El Niño (ENSO) est un cycle climatique naturel qui comprend des phases de réchauffement (El Niño) et de refroidissement (La Niña) des eaux de surface dans l'océan Pacifique équatorial. Ce cycle affecte la circulation atmosphérique et influence le climat mondial, avec des effets qui peuvent se manifester jusqu'à plusieurs milliers de kilomètres de l'équateur. El Niño perturbe les schémas typiques des vents et des eaux, avec des répercussions mondiales sur le climat, les conditions météorologiques et les risques de catastrophes naturelles, y compris des impacts significatifs sur l'activité des incendies, les sécheresses et les anomalies de température.

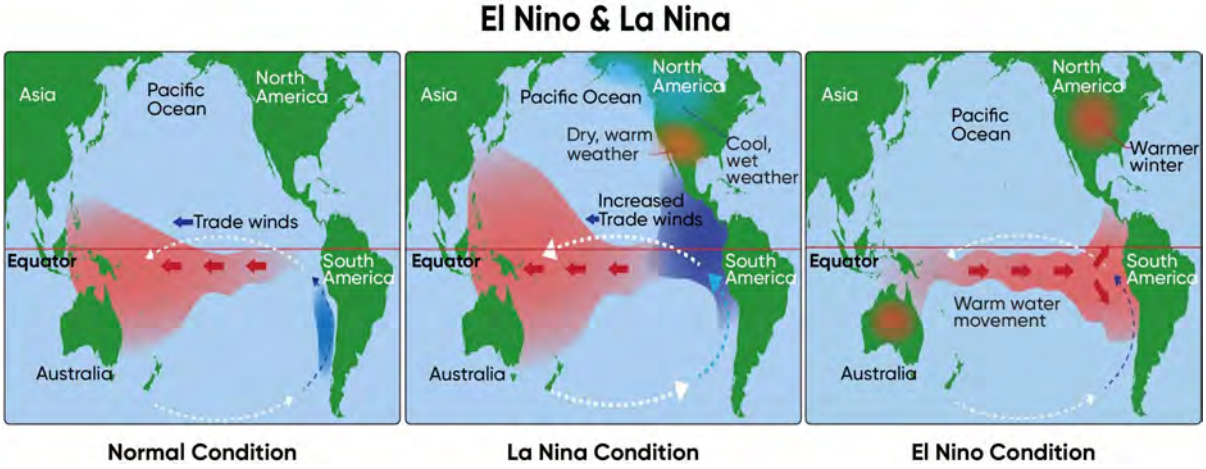


FIGURE 1.4 – Comparaison entre El Niño, La Niña et les conditions normales

En général, El Niño revient tous les deux à sept ans, marqué par un affaiblissement des vents qui ne parviennent plus à pousser les eaux de surface chaudes vers l'ouest. À la place, des eaux plus chaudes que la normale s'accumulent le long de la côte sud-américaine, modifiant les schémas météorologiques à travers les continents. Alors que El Niño apporte des conditions plus humides et plus chaudes à certaines parties de l'Amérique du Nord et du Sud, il tend à provoquer des conditions plus sèches dans d'autres régions, comme l'Australie et l'Asie du Sud-Est. Dans le centre du Chili, les eaux plus chaudes du Pacifique induisent des étés plus chauds, exacerbant les sécheresses, desséchant la végétation et favorisant des conditions météorologiques extrêmes propices aux incendies.

Ainsi, les deux dernières décennies ont connu une augmentation marquée de l'activité des feux de forêt chiliens, les conditions propices aux incendies devenant de plus en plus fréquentes en raison des événements récurrents d'El Niño et du changement climatique progressif. Depuis 2014, six des sept saisons d'incendies les plus destructrices du Chili ont été enregistrées, avec près de 1,7 million d'hectares brûlés en seulement une décennie, soit une augmentation de trois fois par rapport à la décennie précédente. Cette hausse de la fréquence et de l'intensité des incendies est largement attribuée à l'occurrence accrue de vagues de chaleur, de sécheresses persistantes et de températures exceptionnellement élevées qu'El Niño exacerbe.

1.2.2.3 Enjeux

La saison des feux au Chili nécessite une surveillance importante et des mesures de prévention spécifiques pour limiter les impacts des incendies sur les écosystèmes et les zones habitées. La méga-sécheresse qui dure depuis plus d'une décennie a réduit les réservoirs d'eau, induit du stress agricole et alimente les tensions sociales en raison de la rareté de l'eau. Les incendies ont infligé de graves pertes au secteur forestier chilien, compromis la qualité de l'air et causé des dommages substantiels à la productivité agricole, impactant à la fois les récoltes et le bétail. Cette situation est aggravée par la présence de vastes forêts de plantations, telles que le pin et l'eucalyptus, qui sont hautement inflammables. Les plantations forestières, couvrant environ 3 millions d'hectares au Chili, contribuent de manière disproportionnée à la surface totale brûlée, cette part ayant presque doublé lors des saisons record.

De plus, les pertes assurantielles causées par les feux de forêt au Chili se sont amplifiées avec la

répétition de saisons de feux particulièrement destructrices. L'instauration de primes d'assurance en adéquation avec le niveau de risque constitue donc un enjeu majeur. Cela passe par l'intégration d'une granularité plus grande et une modélisation fine et spécifique des risques pour chaque région, en prenant en compte les disparités géographiques et climatiques. L'approche actuarielle traditionnelle de mutualisation uniforme par secteur géographique régional ou provincial évolue vers une tarification plus différenciée, ajustée selon les facteurs de risque propres à chaque zone. Par exemple, les exploitations forestières situées dans des régions particulièrement exposées aux incendies se verraient appliquer des primes plus élevées que celles établies dans des zones à faible risque. Un tel ajustement tarifaire encouragerait les assurés à adopter des pratiques de gestion et de prévention renforcées, en offrant des incitations sous forme de réductions de prime pour des mesures telles que la création de pare-feu ou la diversification des plantations forestières, renforçant ainsi la résilience face aux incendies.

1.2.2.4 Couverture assurantielle

Les couvertures assurantielles pour les feux de forêt au Chili englobent souvent la plupart des dommages causés par les incendies forestiers, incluant la destruction partielle ou totale des biens immobiliers et mobiliers. Cette couverture peut également inclure les frais de déblaiement, les pertes de récoltes pour les agriculteurs, et même les pertes de revenus pour les entreprises directement impactées.

Pour souscrire à l'assurance, le propriétaire forestier doit généralement fournir des informations détaillées sur chaque plantation, incluant :

- le type d'espèce, ainsi que la superficie plantée et l'âge des plantations.
- la valeur estimée par hectare ainsi qu'une estimation totale pour chaque plantation.
- la présence de pare-feu et la distance moyenne entre ces dispositifs de sécurité.
- les pratiques de gestion appliquées, comme le débroussaillage, l'élagage et la gestion des résidus.
- les caractéristiques des parcelles avoisinantes (cultures, autres forêts, etc.).

Il peut également être amené à donner des informations sur les dispositifs de sécurité (tours de surveillance, équipements anti-incendie...), les risques aggravants (proximité de scieries, lignes électriques...) et les sinistres antérieurs. Ces éléments permettent à l'assureur d'ajuster la couverture en fonction du risque spécifique de chaque terrain. Le coût annuel de l'assurance incendie pour les forêts et plantations chiliennes est généralement estimé entre 1 et 1,5% de la valeur totale assurée. De plus, les producteurs forestiers peuvent bénéficier d'une subvention de l'État, réduisant le coût net de la prime en fonction de critères d'éligibilité.[2]

1.3 L'assurance des feux de forêt

1.3.1 Mesure de sinistre : les aires brûlées

Différentes méthodes existent pour quantifier précisément les étendues brûlées, reposant majoritairement sur des technologies de télédétection, de cartographie aérienne et d'analyse géospatiale. La méthode répandue aujourd'hui est le calcul d'indices comme le Normalized Difference Vegetation Index (NDVI) ou le Normalized Burn Ratio (NBR) à partir d'images satellites. Ces indices sont très sensibles aux changements dans la végétation et peuvent être utilisés pour évaluer l'étendue et la gravité des zones brûlées.

Deux principaux satellites sont utilisés pour ces analyses : MODIS et Sentinel-2. Le premier, MODIS, lancé par la NASA, prend des images de la Terre quotidiennement avec une résolution de 500 m x 500 m, ce qui limite sa capacité à détecter des zones brûlées de moins de 25 hectares. Bien que ce satellite permette de construire des historiques de sinistres sur plus de deux décennies grâce à ses données depuis l'an 2000, les images du produit de surface brûlée comportent un délai de traitement de trois mois. En raison de sa résolution relativement faible, MODIS est particulièrement adapté pour identifier de grandes étendues brûlées et fournir une estimation rapide. Le satellite Sentinel-2 quant à lui, lancé par l'ESA en 2015, possède une résolution de 30 m x 30 m, ce qui en fait un outil plus précis pour la détection des zones brûlées de petite échelle bien que son historique soit limité à environ 7 ans.

Concept de cicatrice de feu

Une cicatrice de feu est la zone visible sur le sol ou la végétation qui a été affectée par un incendie. Elle représente l'étendue de la surface brûlée, où la végétation a été détruite ou fortement endommagée par les flammes. Cette cicatrice se manifeste par des changements visibles dans la composition du sol et de la couverture végétale, souvent identifiables grâce à des techniques de télédétection.

1.3.1.1 Normalized Burn Ratio (NBR)

Le *Normalized Burn Ratio* (NBR) est un indice spectral conçu pour identifier et quantifier les zones brûlées dans les images satellitaires. Les satellites permettent de détecter la radiation électromagnétique sur l'ensemble du spectre visible et infrarouge, ce qui est particulièrement utile dans l'analyse de la végétation et des sols affectés par un incendie. Le NBR utilise les longueurs d'onde du proche infrarouge (NIR) et de l'infrarouge à courte longueur d'onde (SWIR) pour identifier les contrastes de réflectance entre les zones de végétation saine et celles affectées par le feu. La végétation en bonne santé présente une réflectance élevée en NIR et une réflectance faible en SWIR, alors que les surfaces brûlées montrent un comportement inverse, soit une réflectance faible en NIR et élevée en SWIR.

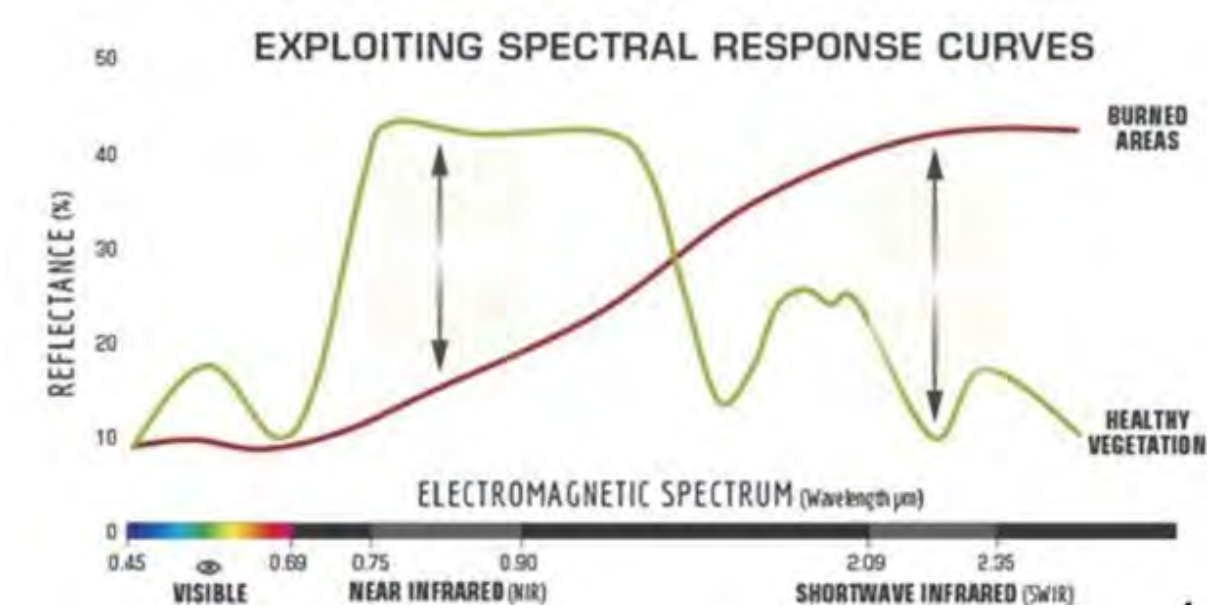


FIGURE 1.5 – Réponse spectrale de la végétation : comparaison entre zones brûlées et non brûlées

L'indice NBR est calculé selon la formule suivante :

$$\mathrm{NBR} = \frac{\mathrm{NIR} - \mathrm{SWIR}}{\mathrm{NIR} + \mathrm{SWIR}}. \tag{1.1}$$

Avant un incendie, la réflectance dans la bande infrarouge proche (NIR) est généralement élevée dans les zones de végétation dense, tandis que la réflectance dans la bande SWIR est faible, produisant des valeurs de NBR négatives. Cependant, suite à un incendie, les valeurs de NBR deviennent positives, en raison des dommages causés à la végétation, rendant les zones brûlées facilement identifiables par une diminution du NIR et une augmentation de SWIR.

Pour analyser la sévérité de l'incendie, deux images sont nécessaires : une image avant l'incendie (pré-feu) et une image après l'incendie (post-feu). La différence entre le NBR avant et après le feu est utilisée pour calculer le *dNBR* (delta NBR), une mesure qui permet de quantifier la sévérité des dommages et la zone brûlée. Le *dNBR* est défini comme suit :

$$\mathrm{dNBR} = \mathrm{NBR}_{\text{pré-feu}} - \mathrm{NBR}_{\text{post-feu}}. \tag{1.2}$$

Des valeurs de dNBR positives indiquent généralement des zones de végétation brûlée, alors que des valeurs négatives peuvent parfois signaler une régénération ou une repousse de la végétation après l'incendie. Dans les images de dNBR, les pixels ayant une valeur positive représentent les zones brûlées, tandis que ceux ayant une valeur négative sont souvent associés à des zones en récupération.

L'absence de produit pré-traité pour les surfaces brûlées nécessite un calcul interne du dNBR à partir d'images individuelles, un processus qui peut varier d'un assureur à l'autre en fonction des seuils appliqués. Des valeurs seuils ont été déterminées sur la base du programme Fire Effects Monitoring and Inventory Protocol (FireMON) de l'USGS [5].

Burn Severity Class	ΔNBR
Enhanced Regrowth, High	< -0.25
Enhanced Regrowth, Low	- 0.25 to -0.10
Unburned	- 0.10 to 0.10
Low Severity	0.10 to 0.27
Moderate-low Severity	0.27 to 0.44
Moderate-high Severity	0.44 to 0.66
High Severity	> 0.66

FIGURE 1.6 – Table des seuils de dNBR[5]

Les valeurs de dNBR plus élevées indiquent une sévérité accrue des dommages, et des seuils de référence, comme celui de 0,27 peuvent être utilisés pour évaluer les pertes.

1.3.1.2 Relative delta Normalized Burn Ratio (RdNBR)

Contrairement à l'indice *dNBR*, qui mesure directement le changement de l'indice NBR avant et après l'incendie, le *Relative delta Normalized Burn Ratio* (RdNBR) normalise cette différence en prenant en compte la densité initiale de la végétation. La formule du RdNBR est la suivante :

$$\text{RdNBR} = \frac{\text{NBR}_{\text{pré-feu}} - \text{NBR}_{\text{post-feu}}}{\sqrt{|\text{NBR}_{\text{pré-feu}}|}}. \tag{1.3}$$

L'ajustement effectué sur la valeur initiale de NBR (pré-feu) rend le RdNBR particulièrement sensible aux changements dans les zones où la végétation est dense. De plus, ce changement permet de mieux évaluer l'impact du feu dans des zones présentant des densités végétales variées. Les valeurs de RdNBR s'interprètent à la même échelle que la table pour le dNBR.

1.3.1.3 Normalized Difference Vegetation Index (NDVI)

Le *Normalized Difference Vegetation Index* (NDVI) est un autre indice spectral permettant de surveiller la santé et la densité de la végétation à partir de données satellitaires. Il repose sur la réflectance différentielle entre le proche infrarouge (NIR) et le rouge visible (RED), avec la formule suivante :

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}}. \tag{1.4}$$

Les valeurs de NDVI, variant entre -1 et +1, permettent d'identifier les zones de végétation saine (valeurs proches de +1) et les surfaces non végétalisées ou affectées (valeurs proches de 0 ou négatives). Une comparaison des valeurs de NDVI avant et après un incendie permet d'observer une diminution de l'indice dans les zones brûlées, indiquant une perte de végétation. Bien que le NDVI ne soit pas spécifiquement conçu pour évaluer la sévérité des feux de forêt, il offre une première estimation de l'impact du feu en termes de couverture végétale.

1.3.2 Les solutions d'assurance

Assurance classique

Dans le cadre de l'assurance classique appliquée aux feux de forêt, les indemnisations sont calculées sur la base des pertes réelles subies par les assurés, en tenant compte de la valeur des biens détruits et des dépenses nécessaires à leur remise en état. Cette méthode repose sur une combinaison de données issues à la fois des relevés sur le terrain et de l'imagerie satellite pour

estimer précisément la superficie affectée par l'incendie. Cependant, pour déterminer le montant final de l'indemnisation, une expertise terrain est souvent requise pour évaluer plus en détail les pertes économiques, notamment en prenant en compte la valeur spécifique des cultures ou des plantations affectées, l'âge des arbres, leur densité, ainsi que les coûts supplémentaires de reboisement, d'enlèvement des débris et de restauration des sols. L'indemnisation reflète alors de manière plus fine la valeur assurée des biens détruits.

Assurance paramétrique sur la mesure de la superficie brûlée

L'assurance paramétrique, s'éloignant du modèle de l'assurance indemnitaire traditionnelle, se base sur des critères ou "paramètres" spécifiques, préalablement établis, pour déclencher une indemnisation. Ces critères sont souvent des indices ou des événements mesurables et vérifiables. dans le contexte des incendies de forêt, une assurance paramétrique est souvent structurée pour indemniser l'assuré dès qu'un certain nombre d'hectares brûlés est atteint. Cette approche élimine le besoin pour l'assuré de justifier des dommages matériels ou financiers subis, une exigence courante dans l'assurance indemnitaire traditionnelle. Par sa simplicité et son caractère automatique, elle propose une réponse immédiate pour faire face aux conséquences économiques des incendies. Cette assurance alternative est donc intéressante lorsque la mesure des pertes réelles est ou pourrait s'avérer complexe comme pour les grandes exploitations forestières ou agricoles, où l'étendue des surfaces rend le coût d'expertise difficilement supportable.

1.3.3 La modélisation des feux de forêt

1.3.3.1 Le modèle catastrophe

La modélisation des catastrophes naturelles ou modélisation Cat Nat est une approche répandue. Cette méthode permet de représenter et d'anticiper les impacts de divers événements extrêmes, tels que les tremblements de terre, les inondations et les ouragans.

La modélisation Cat Nat repose généralement sur quatre modules distincts, chacun traitant un aspect spécifique du risque. Ensemble, ces modules offrent une approche structurée pour évaluer l'exposition aux catastrophes naturelles et estimer les pertes potentielles :

— **Module exposition** : Ce module identifie et quantifie les actifs ou populations exposés au risque, selon leur localisation et leur valeur. Il fournit une estimation des éléments susceptibles d'être endommagés lors d'une catastrophe, qu'il s'agisse de bâtiments, d'infrastructures, de terres agricoles ou de végétation.

— **Module aléa** : Ce module simule la fréquence et la distribution des événements catastrophiques. Pour chaque type de catastrophe, un catalogue de scénarios fictifs est créé, chacun avec une probabilité et une intensité spécifiques. Cela permet aux assureurs de visualiser et d'évaluer l'impact potentiel de divers événements extrêmes sur le territoire.

— **Module vulnérabilité** : Ce module évalue comment les actifs exposés réagissent face aux événements du module aléa, en quantifiant leur résistance aux catastrophes simulées. Il traduit la sévérité des événements en niveaux de dégâts probables pour chaque scénario, permettant d'estimer la proportion de végétation détruite ou la surface de terres impactée.

— **Module financier** : Ce module convertit les dommages physiques en pertes économiques, prenant en compte les conditions d'assurance, telles que franchises, limites de couverture, et clauses de réassurance. Il fournit ainsi une estimation des coûts potentiels pour chaque sinistre, aidant les assureurs à gérer le capital et les réserves nécessaires.

Ces modules donnent un point de départ intéressant pour l'étude des feux de forêt. Cependant, plutôt que de générer des scénarios fictifs comme dans le module aléa pour ce risque, il est courant de s'appuyer sur des données historiques pour établir des cartes de probabilité ou d'exposition.

1.3.3.2 Démarche adoptée

Le choix porte sur une modélisation directe de la probabilité d'incendie pour chaque parcelle du périmètre d'étude, défini au chapitre 2. L'approche ainsi retenue est synthétisée figure 1.7.

Tout d'abord, les données d'exposition et historiques des feux de forêt sont collectées, couvrant la période 1985-2018 et incluant les caractéristiques géographiques, environnementales, et les occurrences de feux passés. Ensuite, une modélisation statistique est réalisée pour établir le lien entre les facteurs d'exposition et la probabilité d'incendie. Enfin, ces résultats sont appliqués pour produire une carte de risque d'incendie de forêt.

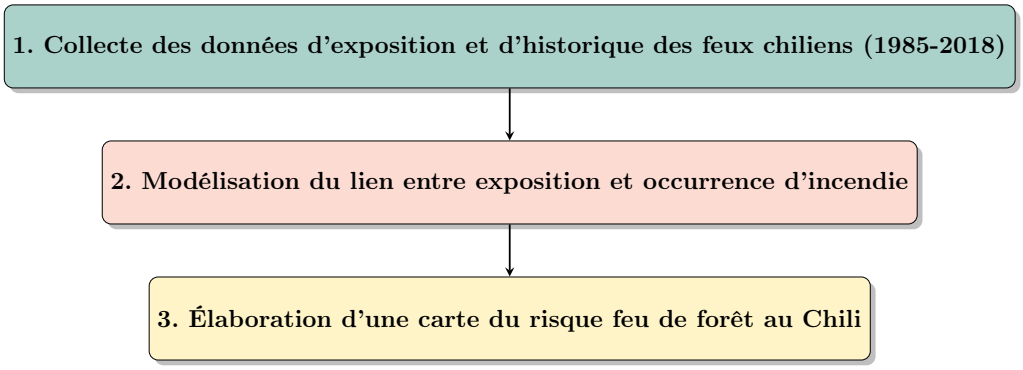


FIGURE 1.7 – Approche adoptée pour la modélisation du risque feu de forêt

Pour structurer l'approche de modélisation des feux de forêt au Chili et la rendre fonctionnelle, les hypothèses suivantes sont mises en place :

— **Parcelles de 30 m²** : Les zones assurées sont représentées sous forme de mosaïques constituées de parcelles de 30 m², une résolution largement répandue dans les données satellite. Cette échelle permet un compromis entre précision et faisabilité, capturant les dynamiques locales tout en restant gérable pour les calculs.

— **Parcelles constituées de végétation** : L'étude se concentre sur des biens assurés composés uniquement de végétation, tels que les forêts et les plantations, excluant ainsi les bâtiments, habitations et infrastructures. Ce choix évite la complexité liée aux biens hétérogènes, chacun ayant des vulnérabilités spécifiques face aux feux de forêt. En se limitant aux portefeuilles de végétation, la modélisation peut se concentrer sur les pertes propres aux zones naturelles et agricoles, en cohérence avec les polices d'assurance de végétation, sans empiéter sur les couvertures incendie des biens construits.

— **Valeur assurée fixée** : La modélisation ne prend pas en compte les valeurs assurées individuelles, car celles-ci sont couramment déterminées par déclaration de l'assuré dans ce type de couverture.

— **Taux de destruction binaire** : Chaque parcelle est considérée comme totalement brûlée ou intacte après le passage d'un feu, évitant la complexité d'une modélisation des dégâts partiels.

— **Absence d'interdépendance spatiale entre les parcelles** : Chaque parcelle est évaluée indépendamment, sans modélisation de la propagation du feu. Cela simplifie l'analyse en fournissant une estimation statique du risque pour chaque parcelle.

— **Variations annuelles** : L'étude est réalisée sur une moyenne annuelle, sans prise en compte des fluctuations saisonnières, en cohérence avec les polices d'assurances feu de forêt qui sont en général annuelles.

Chapitre 2

Constitution de la base de données

2.1 Cartographie

2.1.1 Sélection du périmètre

La partie centrale du Chili est la plus vulnérable et pertinente pour l'étude des feux de forêts chiliens. En effet, ce périmètre bordé au nord par le désert d'Atacama dans la région éponyme et par les montagnes et fjords de la région d'Aysén del General Carlos Ibáñez del Campo englobe plus de 94% de l'aire brûlée au Chili durant la période 1985-2018 d'après les données de la CONAF. Le périmètre d'étude comprend ainsi les régions (du nord au sud) de : Coquimbo, Valparaiso, Region Metropolitana de Santiago, Libertador General Bernardo O'Higgins, Maule, Nuble, Bio Bio, la Araucania, Los Rios et Los Lago pour une surface cumulée d'environ 268 000 km^2.

2.1.2 Choix du système de projection

L'outil QGIS sera principalement utilisé pour la représentation et le traitement des données. Ce logiciel de système d'information géographique permet d'analyser, d'éditer et de visualiser des données spatiales sur divers formats cartographiques tout en offrant une multitude d'outils pour la cartographie, l'analyse spatiale et la gestion des bases de données géospatiales. Il convient donc d'établir le système de coordonnées approprié pour les représentations. L'utilisation d'une projection de référence pour toutes les données permet de faciliter par la suite leur rassemblement et leur manipulation. Dans cette optique,la projection EPSG :4326 - WGS 84 est choisie, un système de coordonnées géographiques qui repose sur des degrés de latitude et de longitude. Ce système est couramment utilisé dans les applications géospatiales en raison de sa compatibilité avec une large variété de données. Il présente plusieurs avantages : il est standardisé au niveau mondial, largement accepté par les bases de données géospatiales, et permet de maintenir une cohérence entre les sources de données qui seront utilisées, comme les données satellites, météorologiques et les bases de données topographiques.
Cependant, il est important de noter que cette projection, bien que globalement adaptée pour l'étude, présente certaines limites, notamment lorsqu'elle est appliquée à des zones géographiquement étendues, comme le Chili. Chaque type de projection repose sur des hypothèses simplificatrices qui peuvent distordre les formes, distances ou surfaces. Dans le cas de l'EPSG :4326 - WGS 84, l'étendue latitudinale du périmètre d'étude fait apparaître

des variations notables dans les distances représentées par les longitudes à différentes latitudes. En effet, un degré de longitude représente une distance qui diminue à mesure que la latitude augmente vers le sud en raison de la convergence des longitudes aux pôles. Alors qu'à l'équateur, un degré de longitude équivaut à environ 111 kilomètres, cette distance se réduit à environ 96 kilomètres au niveau du 30e parallèle sud, qui est la latitude du centre de la région la plus au nord du périmètre. En considérant le centre de la région la plus au sud et situé proche du 41e parallèle, Los Lagos, un degré de longitude est proche de 84 kilomètres. La relation mathématique traduisant cette variation est donnée par :

$$l(\text{latitude}) = 111.3 \times \cos(\text{latitude})$$

où l(latitude) représente la longueur d'un degré de longitude à une latitude spécifique.

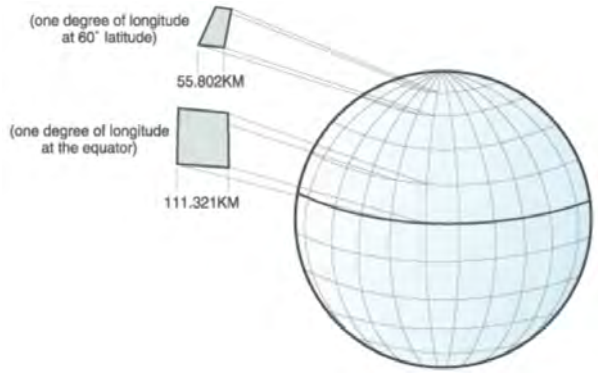


FIGURE 2.1 – Illustration de la variation de longueur d'un degré de longitude à deux latitudes différentes [12]

2.1.3 Cartes des régions administratives du Chili

Les shapefiles constituent un format couramment utilisé pour stocker et représenter des entités géographiques en trois composantes principales : localisation, forme (géométrie), et attributs (informations associées aux entités). Les attributs correspondent à des données descriptives, telles que le nom, la population, la superficie ou tout autre élément lié à l'entité géographique. Ce format permet de manipuler des données vectorielles et assure une représentation précise des limites, frontières et caractéristiques géographiques des zones étudiées.

Les cartes sélectionnées sont celles des régions, provinces et communes administratives du Chili, mises à disposition par le Bureau de la coordination des affaires humanitaires des Nations Unies (UNOCHA[10]). Ces shapefiles fournissent une description géospatiale officielle et actuelle des régions chiliennes, incluant les délimitations des frontières administratives et les noms. La flexibilité de ce format permet de filtrer, sélectionner ou fusionner des zones géographiques, et ainsi pour l'étude d'isoler les régions du Chili sélectionnées précédemment. Cette carte permettra de contraindre les différentes données amassées (climatiques, topographiques, points d'incendie) à la forme des régions.

FIGURE 2.2 – Chili et périmètre sélectionné

2.1.4 Fichiers raster

Les fichiers de type raster sont largement utilisés pour représenter des données géospatiales sous forme d'une grille régulière de cellules (ou pixels), où chaque cellule est associée à une valeur numérique. Ces valeurs correspondent généralement à une mesure ou une caractéristique d'une zone géographique donnée, comme l'altitude, la température ou encore l'intensité de la végétation. Contrairement aux données vectorielles qui décrivent des entités par des formes géométriques (points, lignes, polygones), les données raster capturent des phénomènes continus ou variables sur une surface. Ce format est particulièrement adapté à l'analyse et à la visualisation de données environnementales ou satellitaires, car il permet une représentation fine des variations spatiales sur une grande étendue. Dans la suite, c'est principalement sous ce format que les données de l'étude seront collectées.

2.2 Données des feux historiques

Pour l'établissement d'un modèle statistique prédisant les zones les plus à risque, il est essentiel de connaître les données historiques d'évènements ainsi que de récolter et déterminer les variables les plus importantes et qui expliquent le mieux ce risque. Divers types de données qui pourraient expliquer l'apparition et la propagation des feux de forêt ont donc été rassemblées ainsi que les données d'incendies enregistrés.

2.2.1 Répertoire de la CONAF

La CONAF recueille et met à disposition plusieurs données qu'elle utilise pour ses opérations. Elle collecte des informations détaillées sur chaque incendie, telles que la localisation, la date d'ignition et de contrôle, la cause, et la superficie brûlée. Jusqu'en 2003, la géolocalisation se faisait via une grille alphanumérique de 1 km^2 sur des cartes de l'Institut Géographique Militaire (IGM), mais après cette date, le GPS a permis une localisation plus précise. La superficie affectée est mesurée en hectares, en distinguant les types de végétation touchés, comme les forêts natives, les plantations forestières ou les prairies. Une autre mission de la CONAF consiste à identifier la cause des incendies, qu'ils soient d'origine naturelle (foudre) ou anthropique (négligence humaine, activités agricoles, etc.), afin de mieux comprendre les facteurs de risque et d'adapter les stratégies de prévention. De plus, chaque incendie est associé à une région administrative et une commune, facilitant ainsi l'analyse locale et la gestion des ressources. La CONAF publie régulièrement des rapports régionaux pour suivre l'évolution des incendies et des risques dans différentes zones du pays. Ces données permettent une gestion efficace des incendies de forêt et une meilleure compréhension de leur dynamique sur le territoire chilien.

FIGURE 2.3 – Feux enregistrés par la CONAF entre les saisons de feux 1985 et 2018

2.2.2 La base de données Landscape Fire Scars

La Landscape Fire Scars Database a été élaborée sous la direction d'Alejandro Miranda et une équipe de chercheurs chiliens afin de cartographier les incendies de forêt au Chili pour la période allant de la saison de feux 1985 à 2018[6][3]. Elle a pour ambition de compléter les

bases de données à haute résolution relatives aux feux de forêt dans les pays en développement, où la précision et la granularité des données satellitaires, en particulier dans les régions isolées et sous surveillance limitée, sont souvent insuffisantes. Cette base de données repose sur une exploitation systématique des images satellitaires des missions Landsat à travers la plateforme Google Earth Engine (GEE), permettant d'obtenir des informations précises sur les zones brûlées et la gravité des incendies de forêt enregistrés. Ce projet s'appuie sur les données enregistrées par la CONAF, et se focalise sur une analyse géospatiale des incendies ayant eu lieu dans les régions administratives allant de Coquimbo à Los Lagos, couvrant un total de 12 250 incendies enregistrés par la CONAF. En prenant en compte la disponibilité, la qualité et la résolution spatiale des images Landsat 5 (1984-2013), Landsat 7 (1999-2021), et Landsat 8 (2013 à 2018), les chercheurs ont sélectionné uniquement les incendies de plus de 10 hectares. Ce seuil a été choisi car il représente plus de 93% des surfaces brûlées selon les rapports de la CONAF pour cette période et pour éviter que les petits incendies, souvent liés à des brûlis agricoles, ne soient systématiquement inclus, car ils peuvent être confondus avec des pratiques agricoles traditionnelles.

2.2.2.1 Constitution de la base

Pour chaque incendie analysé, des images satellites pré- et post-incendie ont été évaluées dans Google Earth Engine 100 jours avant et après l'événement, afin d'identifier les zones brûlées et de définir la cicatrice de feu à l'aide de l'indice spectral RdNBR (Relative Delta Normalized Burn Ratio). Cette approche a permis de reconstruire un total de 8153 cicatrices d'incendies, soit environ 66,6% des incendies répertoriés officiellement sur la période et les régions considérées. Chaque incendie est décrit à travers une série de fichiers raster et vectoriels. Cela inclut les images avant et après l'incendie et les cartes de la gravité de l'incendie. Tous les fichiers sont projetés dans le système de coordonnées EPSG :4326 et les couches raster possèdent une résolution d'environ 30 mètres, enregistrées au format GeoTIFF. Les fichiers sont systématiquement nommés en fonction du type de produit représenté (image, indice, cicatrice de feu), du code ISO de la région administrative concernée et du numéro d'identification de l'incendie. La couverture des incendies reconstitués est plus importante pour les événements plus récents grâce à une meilleure disponibilité des images satellites : 35% des incendies entre 1985 et 1994, 63% entre 1995 et 2004, 82% entre 2005 et 2014, et 93% entre 2015 et 2018 sont ainsi reconstitués. Cette amélioration est particulièrement notable dans les régions méridionales du Chili, où la couverture nuageuse constituait un obstacle majeur à l'obtention de données satellitaires de haute qualité. L'évaluation des cicatrices de feu reconstruites a révélé une correspondance globale de 79 % avec les données officielles de la CONAF. Des divergences ont été constatées, notamment en ce qui concerne les zones non brûlées incluses par erreur dans certains périmètres digitalisés par la CONAF. Toutefois, les erreurs de commission (zones incorrectement identifiées comme brûlées) étaient limitées à 7%, tandis que les erreurs d'omission (zones brûlées non détectées) atteignaient 28%.

2.2.2.2 Attributs

Chaque enregistrement de feu dans la *Landscape Fire Scars Database* est doté d'attributs détaillés, regroupés dans une table d'attributs. Ces informations permettent de suivre les caractéristiques et l'impact de chaque feu. Voici les principaux attributs inclus dans la base :

— **FireID** : Identifiant unique attribué à chaque incendie.
— **FireSeason** : Saison des incendies correspondant à l'année de l'incident (définie de juillet de l'année précédente à juin de l'année en cours).

— **RegionCode** et **Region_CONAF** : Code de la région et nom de la région administrative où s'est produit l'incendie, selon les délimitations de la CONAF.

— **FireName_CONAF** : Nom de l'incendie assigné par la CONAF.

— **Area_CONAF [ha]** : Superficie totale brûlée, exprimée en hectares, enregistrée par la CONAF.

— **IgnitionDate_CONAF** et **ControlDate_CONAF** : Dates de début et de maîtrise de l'incendie selon les registres de la CONAF.

— **Latitude [°]** et **Longitude [°]** : Coordonnées géographiques du point d'ignition de l'incendie enregistrés par la CONAF.

— **FireScar** : Indication binaire de la présence d'une reconstitution de la cicatrice du feu.

— **NorthBoundLatitude [°]**, **SouthBoundLatitude [°]**, **WestBoundLongitude [°]**, **EastBoundLongitude [°]** : Limites géographiques de l'incendie.

— **TotalArea [m²]** : Surface totale de la cicatrice en mètres carrés.

— **AreaUnchS [m²]**, **AreaLowS [m²]**, **AreaModS [m²]**, **AreaHighS [m²]** : Surfaces des cicatrices en fonction des catégories de gravité (inchangé, faible, modéré, élevé).

— **OverlapIDs** : Identifiants des cicatrices d'incendies qui se superposent avec la cicatrice actuelle.

— **Observations** : Informations supplémentaires liées à l'incendie.

2.2.2.3 Traitement des données

Chaque cicatrice d'incendie reconstituée a été géoréférencée avec précision et intégrée dans un système de coordonnées commun (EPSG :4326 - WGS 84). Dans le but de cartographier les zones qui ont historiquement brûlé dans le périmètre étudié, il convient de consolider les données des 8153 feux issus de la Landscape Fire Scars Database en créant une unique mosaïque géospatiale, permettant ainsi d'obtenir une vue d'ensemble des points brûlés sur la période de 1985 à 2018 couvrant les régions administratives de Coquimbo à Los Lagos en préservant la résolution spatiale d'environ 30 mètres.

2.3 Données environnementales et urbaines

Les données climatiques, météorologiques, topographiques, de couverture du sol et urbaines du périmètre ont été collectées par divers organismes publiques chiliens ou mondiaux. Lorsque cela a été possible d'inclure une étendue temporelle des mesures, les données ont été recueillies entre 1985 et 2018 puis moyennées.

2.3.1 Données climatiques et météorologiques

2.3.1.1 Température et Précipitations

La végétation sèche est un combustible de choix pour les feux, facilitant l'ignition et la propagation des incendies. Or, les hautes températures assèchent la végétation par le phénomène de l'évapotranspiration, et réduisent aussi le taux d'humidité de l'air. Par ailleurs, les précipitations jouent un rôle important dans l'humidification de la végétation. Dans des périodes de faibles précipitations ou de sécheresse prolongée, la probabilité et l'intensité des feux de forêt peuvent augmenter considérablement. Il est également à noter que les hausses de température peuvent

augmenter la fréquence des événements météorologiques extrêmes, comme les vagues de chaleur, ce qui renforce encore la probabilité des incendies. Dans cette étude, les données climatiques issues du jeu de données CR2MET du CR2 (Centro de Ciencia del Clima y la Resilienca)[11] sont utilisées. Elles fournissent des informations sur les précipitations et les températures (proches de la surface) maximales, minimales et moyennes à une résolution spatiale de 0,05° pour le Chili continental sur la période 1960-2021. Ces données sont dérivées de modèles statistiques calibrés à partir de relevés d'observation et enrichis par des données de réanalyse, des paramètres topographiques et des estimations de température de surface provenant du capteur satellite MODIS (Moderate Resolution Imaging Spectroradiometer). Les températures sont agrégées sous forme de moyennes mensuelles, tandis que les précipitations sont exprimées en millimètres mensuels, cumulant les précipitations journalières.

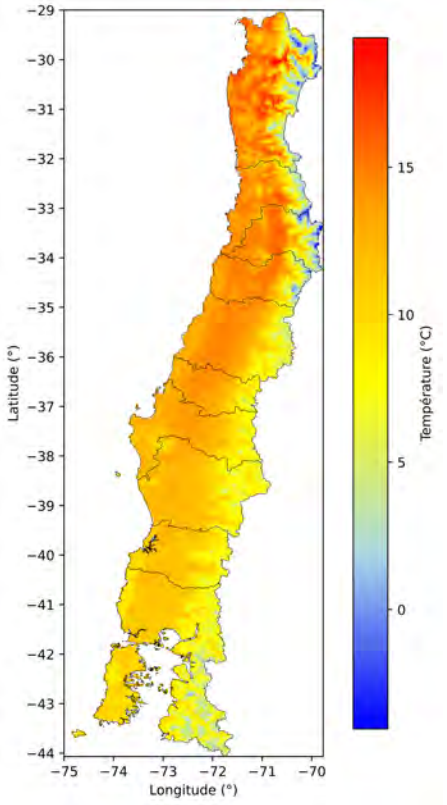


FIGURE 2.4 – Carte de température

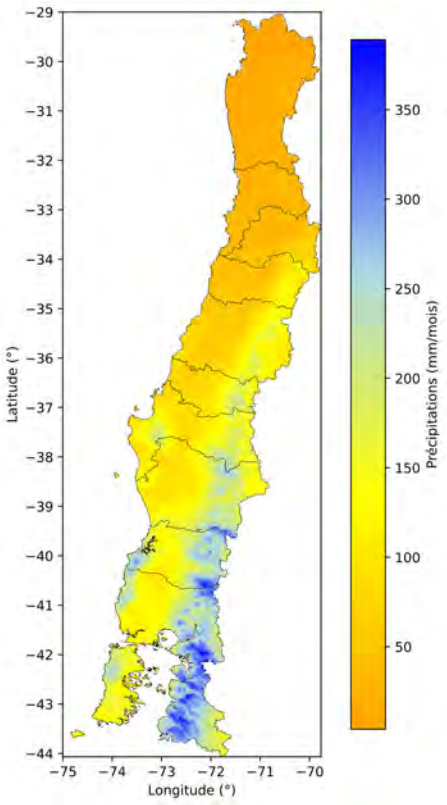


FIGURE 2.5 – Carte des précipitations

Les cartes des précipitations et des températures du Chili révèlent une transition climatique du nord au sud. Le nord, aride et chaud, reçoit peu de précipitations, tandis que la région centrale, au climat méditerranéen, bénéficie de précipitations modérées et de températures douces. Au sud, le climat devient tempéré et humide, avec des précipitations abondantes, surtout dans les Andes, et des températures plus fraîches influencées par l'altitude et les masses d'air polaires.

2.3.1.2 Indices météo

L'Indice météorologique de feu (FWI, pour Fire Weather Index) est un indice basé sur la météorologie et utilisé dans le monde entier pour estimer le danger d'incendie. Il est calculé à

partir d'autres indices mesurés ou calculés et qui prennent en compte les effets de l'humidité des sols et des combustibles, de la sécheresse des sols en profondeur et du vent sur le comportement et la propagation du feu. Plus l'indice FWI est élevé, plus les conditions météorologiques sont favorables au déclenchement d'un incendie de forêt. Il a été développé au Canada et son utilisation s'est étendue à de nombreux autres pays en raison de son efficacité.

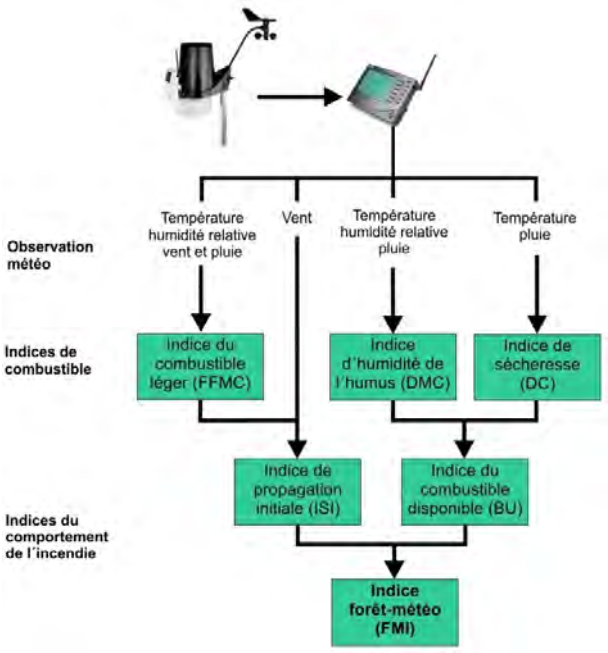


FIGURE 2.6 – Construction de l'indice forêt météo

Le FWI est constitué de différents indices météorologiques :

- **Fine Fuel Moisture Code (FFMC) :** lié à l'humidité du combustible tels que les feuilles et les herbes. Ce code est sensible aux variations de la température, de l'humidité relative et des précipitations. Un FFMC élevé indique que ces matériaux sont secs et donc plus susceptibles de s'enflammer.
- **Duff Moisture Code (DMC) :** est aussi lié à l'humidité. Il estime la teneur en humidité du sol et des matériaux organiques en décomposition, souvent appelés "duff". Tout comme le FFMC, le DMC est influencé par la température, l'humidité et les précipitations, mais à une échelle temporelle plus longue en raison de la nature plus dense des matériaux concernés.
- **Drought Code (DC) :** mesure la sécheresse du sol en profondeur. Ce code est particulièrement utile pour évaluer le risque d'incendies souterrains qui peuvent brûler pendant des périodes prolongées. Le DC est principalement influencé par la température et les précipitations.
- **Initial Spread Index (ISI) :** est lié au comportement du feu et estime la vitesse de propagation initiale d'un feu. Il est calculé à partir du FFMC et de la vitesse du vent. Un ISI élevé signifie que les conditions sont favorables à une propagation rapide du feu.

– **Build-Up Index (BUI) :** estime la quantité totale de combustible disponible pour la combustion. Il est calculé en utilisant le DMC et le DC, ce qui en fait une mesure composite de la disponibilité totale du combustible.

Les données moyennes mensuelles du Copernicus Emergency Management Service pour les cinq indices présentés ainsi que l'indice forêt météo sont ainsi inclus dans l'étude. Ces données sont disponibles à une résolution de 0.25° depuis 1940 dans le monde et au Chili.

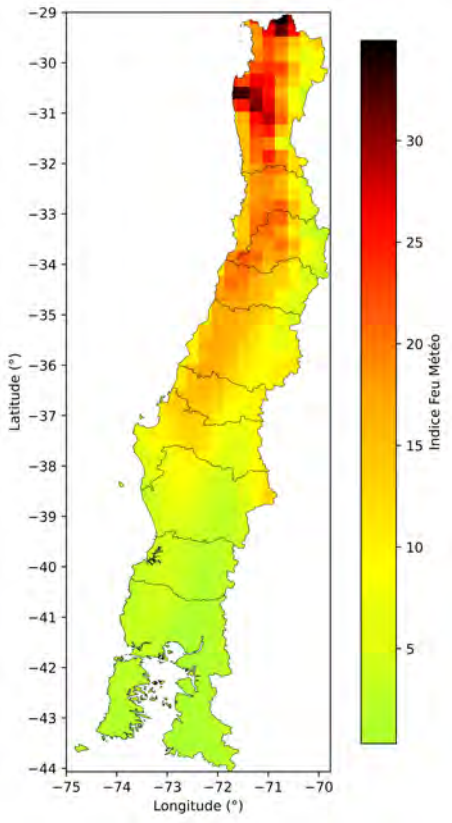


FIGURE 2.7 – Carte d'indice feu météo

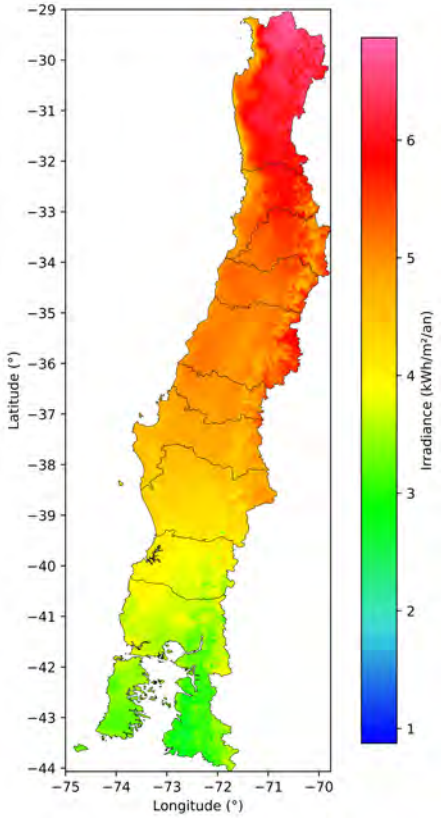


FIGURE 2.8 – Carte d'irradiance

Les cartes montrent qu'au nord du Chili, l'indice feu météo et l'irradiance solaire sont élevés. En allant vers le centre, ces valeurs diminuent progressivement, tandis que dans le sud, l'indice feu est faible et l'irradiance solaire est réduite.

2.3.1.3 Irradiance

L'irradiance solaire désigne la quantité d'énergie solaire reçue par unité de surface et permet d'évaluer l'énergie solaire disponible à un endroit donné. Différents types d'irradiances sont couramment utilisés pour caractériser l'énergie solaire :

– l'Irradiance Directe Normale (DNI) : Elle mesure le rayonnement solaire direct qui est reçu par une surface perpendiculaire aux rayons du soleil. Cette mesure peut servir pour les systèmes de concentration solaire qui nécessitent une mesure précise du rayonnement solaire direct.

– l'Irradiance Diffuse Horizontale (DHI) : Elle quantifie le rayonnement solaire diffus reçu sur une surface horizontale. Cette irradiance provient de la diffusion des rayons du soleil dans l'atmosphère, sans tenir compte du rayonnement solaire direct.
– l'Irradiance Horizontale Globale (GHI) : C'est la quantité totale de rayonnement solaire direct et diffus reçue sur une surface horizontale. Elle est pertinente pour les applications comme les panneaux solaires installés à plat.

Ces différentes mesures d'irradiance permettent d'évaluer l'efficacité et la faisabilité des installations solaires en fonction de leur orientation et de leur type. Ensemble, elles fournissent une image complète du rayonnement solaire disponible à un endroit précis.
Néanmoins dans le contexte de la prévision des feux de forêt, la mesure sur laquelle l'étude se focalisera est l'Irradiance Horizontale Globale (GHI) :

$$\text{GHI} = \text{DNI} \times \cos(0^\circ) + \text{Irradiance Diffuse} \tag{2.1}$$

En effet, la GHI prend en compte toute la lumière du soleil qui frappe une surface horizontale, y compris la lumière directe et diffusée. Or, les arbres, comme le sol, reçoivent la lumière du soleil sous divers angles en raison des effets de diffusion de l'atmosphère, des nuages et des surfaces environnantes. L'étude de la quantité d'énergie solaire reçue par la végétation peut expliquer l'apparition et la propagation des feux de forêt puisqu'elle augmente la qualité des combustibles tout en exposant directement la végétation à une source d'énergie, le soleil.
Les données de Global Solar Atlas [1] qui fournissent la quantité d'irradiance moyenne journalière de type GHI reçue entre 1999 et 2018 au Chili à une résolution de 9 arcsec (environ 250 m) sont intégrées aux données.

2.3.1.4 Impacts de foudre

La foudre est un phénomène météorologique extrêmement puissant, capable d'enflammer des zones de végétation sèche en une fraction de seconde. Au Chili, comme dans de nombreux endroits du monde, les incendies peuvent être déclenchés par la foudre pendant la saison sèche.

Pour étudier l'impact de la foudre sur les incendies de forêt au Chili, les données de la NASA[7] qui fournissent une moyenne mensuelle des impacts de foudre sont utilisées. Ces données couvrent la période de 1995 à 2014 et à une résolution de 0,5°.

Les régions centrales et certaines zones du sud présentent des niveaux plus élevés d'activité d'éclairs tandis que la majeure partie du nord et les zones plus éloignées à l'extrême sud sont caractérisées par une faible fréquence.

2.3.2 Données topographiques

La connaissance de la topographie du terrain peut s'avérer utile comme facteur de sélection afin d'expliquer le risque de départ et de propagation de feux. Les hautes altitudes sont en effet naturellement moins propices au péril feu de forêt au vu des diverses conditions telles que les températures plus basses, les taux d'humidité relatifs naturellement plus élevés. Il est donc approprié de prendre en compte ces données. La pente joue un rôle très important puisque la gravité augmente la vitesse de propagation des feux notamment par la chute des combustibles.

2.3.2.1 Altitude et Pente

La Shuttle Radar Topography Mission (SRTM)[9] est une collaboration internationale entre la NASA et l'Agence spatiale italienne réalisée en 2000 qui avait pour but de créer une carte nu-

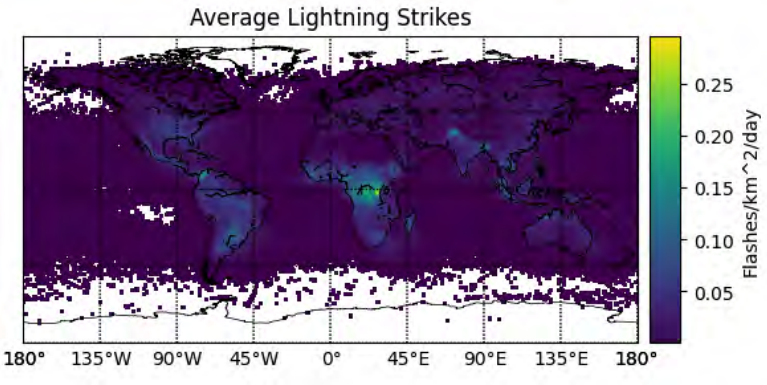


FIGURE 2.9 – Carte mondiale d'impact de foudre

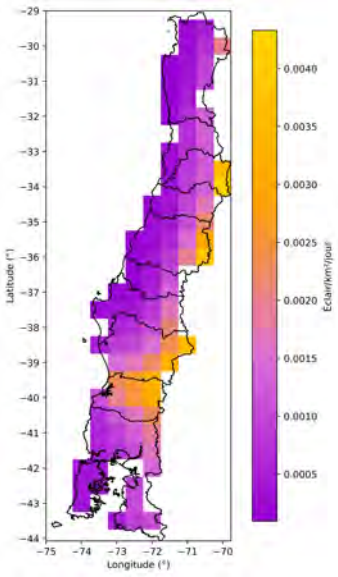


FIGURE 2.10 – Carte d'impact de foudre

mérique d'élévation à haute résolution de la Terre. La SRTM a utilisé un radar interférométrique monté à bord de la navette spatiale Endeavour pour cartographier la topographie de la majeure partie du globe entre les latitudes 56°S et 60°N, fournissant ainsi des données précieuses pour la recherche en géographie, géologie, hydrologie, ainsi que pour diverses applications pratiques comme la cartographie ou l'aménagement du territoire.
Les données d'altitude obtenues par la STRM à une résolution d'une arcseconde (environ 30m²) sont disponibles directement sur QGIS.

La pente se déduit par la méthode de Horn qui prend en compte les voisins immédiats d'un pixel dans un raster de données d'élévation (comme un DEM, ou Digital Elevation Model). La méthode de Horn prend en compte les valeurs des huit cellules entourant la cellule centrale, ce qui permet une meilleure approximation de la pente.

La formule générale pour le calcul de la pente S en degrés en utilisant la méthode de Horn est :

$$S = \arctan\left(\sqrt{\left(\frac{\Delta Z}{\Delta X}\right)^2 + \left(\frac{\Delta Z}{\Delta Y}\right)^2}\right)$$

où $\Delta Z/\Delta X$ et $\Delta Z/\Delta Y$ sont les variations de l'altitude par rapport aux distances dans les directions X et Y, respectivement.

Pour calculer ces variations ($\Delta Z/\Delta X$ et $\Delta Z/\Delta Y$), la méthode de Horn utilise une pondération des différences de hauteur entre la cellule centrale et les huit cellules voisines.

Mathématiquement, ces variations sont calculées comme suit pour la méthode de Horn :

$$\frac{\Delta Z}{\Delta X} = \frac{(Z1 + 2Z2 + Z3) - (Z7 + 2Z8 + Z9)}{8\Delta x}$$
$$\frac{\Delta Z}{\Delta Y} = \frac{(Z1 + 2Z4 + Z7) - (Z3 + 2Z6 + Z9)}{8\Delta y}$$

où $Z1, Z2, Z3, \ldots, Z9$ sont les valeurs d'altitude des cellules, comme illustré ci-dessous :

$$\begin{matrix} Z1 & Z2 & Z3 \\ Z4 & Z5 & Z6 \\ Z7 & Z8 & Z9 \end{matrix}$$

et Δx et Δy sont les dimensions de la cellule en unités de carte (généralement des mètres).

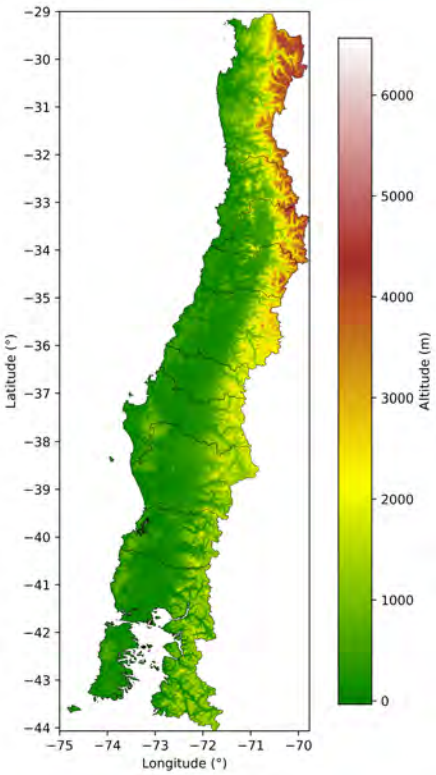


FIGURE 2.11 – Carte d'altitude

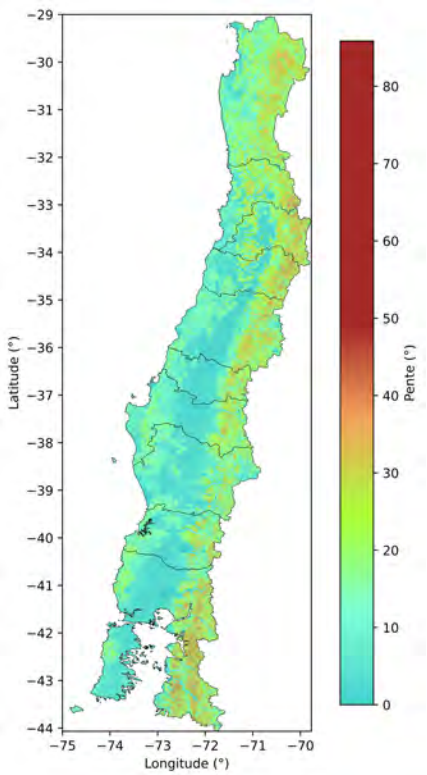


FIGURE 2.12 – Carte de pente

L'altitude est la plus élevée dans les Andes, avec des sommets dépassant les 6000 mètres dans le nord-est tandis que les régions côtières et du sud ont des altitudes plus basses. La carte des pentes montre des zones de forte inclinaison, particulièrement dans la chaîne andine tandis que les zones de faible pente se trouvent principalement dans les vallées et les plaines côtières.

2.3.3 Couverture du sol

L'analyse de la couverture du sol permet de mieux comprendre les interactions entre les incendies de forêt et l'environnement, notamment en ce qui concerne les types de végétation et les zones urbanisées. Différents types de couverture terrestre possèdent des caractéristiques de combustion variées : par exemple, les forêts denses présentent un potentiel de feux plus longs et plus intenses que les prairies ou les zones de végétation clairsemée.

Afin d'intégrer ces informations dans l'étude, la couverture des sols issue de la base de données *Global Land Cover Mapping and Estimation* (GLanCE30 v001 [8]) a été utilisée pour l'année 2019. Ce produit de données globales, fourni dans le cadre de l'initiative *NASA's Making Earth System Data Records for Use in Research Environments (MEaSUREs)*, propose des cartes de couverture terrestre à une résolution spatiale de 30 mètres. GLanCE30 dérive des images des

QA Value	QA Name	Description
1	Water	Zones couvertes d'eau toute l'année : cours d'eau, canaux, lacs, réservoirs et océans.
2	Ice/Snow	Zones où la glace et la neige couvrent plus de 50% de la surface tout au long de l'année.
3	Developed	Zones d'utilisation intensive : terres couvertes de structures, y compris les terres liées à des activités urbaines ou construites.
4	Barren/Sparsely Vegetated	Terres composées de sols naturels, de sable ou de roches, où moins de 10% de la surface est végétalisée.
5	Tree Cover	Terres où la couverture arborée dépasse 30%. Les zones déboisées (coupes à blanc) sont classées selon leur couverture actuelle (par ex. terrains stériles ou couverts de broussailles ou d'herbes).
6	Shrublands	Terres avec moins de 30% de couverture arborée, mais où la végétation totale dépasse 10% et la couverture arbustive dépasse 10%.
7	Herbaceous	Terres couvertes d'herbacées. La couverture végétale totale dépasse 10%, la couverture arborée est inférieure à 30%, et les arbustes couvrent moins de 10% de la zone.

TABLE 2.1 – Classification des types de couverture terrestre dans GLanCE

satellites Landsat 5, 7 et 8, et fournit des informations détaillées sur les types de couverture terrestre ainsi que les changements annuels observés sur la période 2001-2019. La carte est organisée en plusieurs couches, parmi lesquelles figurent la classification des types de couvertures terrestres, l'amplitude de l'indice de végétation (EVI2), ainsi que des métriques de changement interannuel.

Les classes de couverture terrestre incluent sept types principaux : les zones aquatiques, les surfaces glacées/neigeuses, les zones développées (urbanisées), les surfaces désertiques ou peu végétalisées, les forêts, les formations arbustives et les zones herbacées. La carte montre que dans les régions centrales du Chili, les zones forestières (Tree Cover) se trouvent principalement dans les basses et moyennes altitudes le long des Andes et de la vallée centrale, où les précipitations sont suffisantes pour soutenir cette végétation. Les régions herbacées (Herbaceous) et les broussailles (Shrublands) se concentrent dans le nord et certaines zones de basse altitude, adaptées aux conditions plus arides. Les zones de végétation clairsemée (Barren/Sparsely Vegetated) se situent principalement dans les zones montagneuses élevées et arides, où les conditions de sol et de climat limitent la végétation. Les zones de neige/glace (Ice/Snow) sont restreintes aux sommets andins les plus élevés et aux latitudes plus australes. Les développements humains (Developed) sont peu nombreux et localisés, et les étendues d'eau (Water) apparaissent dans certaines vallées et bassins, représentant les principaux réservoirs et lacs naturels.

Quatre variables sont ainsi créées pour chaque pixel, une variable TreeCover qui indique la présence de zone forestière, une variable Shrublands celle d'arbustes, une variable Herbaceous celle des prairies et une dernière variable NoVegetation pour les pixels d'eau, de neige ou de glace, de zone développée et de zone stérile ou avec peu de végétation.

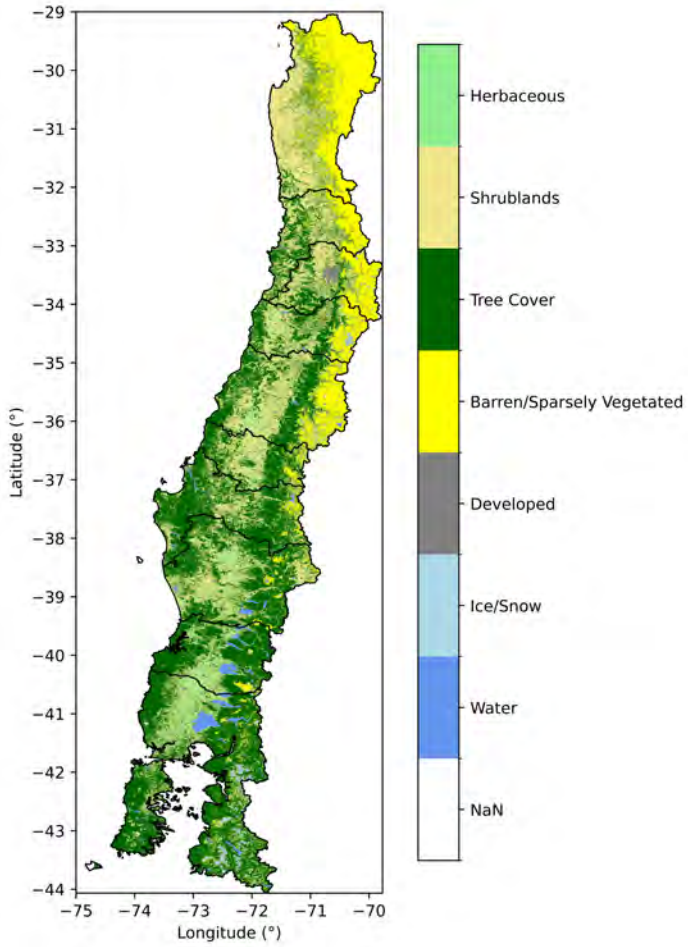


FIGURE 2.13 – Classification du terrain

2.3.4 Données d'urbanisation

La première cause de feu de forêt étant anthropique, il convient de prendre en compte les données d'urbanisation. Pour cela, deux mesures de la présence et de l'activité humaine ont été intégrées, la densité de population et l'activité économique lumineuse dite "Lit Pop".

2.3.4.1 Données démographiques

Pour évaluer l'impact potentiel des activités humaines sur les incendies de forêt, les données qui seront utilisées sont celles de densité de population fournies par WorldPop [13]. Ces données représentent la distribution spatiale de la population au Chili en 2020 à une résolution de 1 km.

2.3.4.2 LitPop

Afin de mieux quantifier l'exposition humaine et économique dans les zones à risque, les données du jeu de données LitPop[15] ont été intégrées. Développée par l'ETH Zurich, cette base de données fournit des cartes haute résolution des valeurs d'actifs exposés aux catastrophes naturelles. La méthode utilisée combine l'intensité lumineuse nocturne captée par satellite et les données démographiques pour produire une estimation précise de la répartition des actifs

nationaux. Cette approche permet de descendre à une résolution de 30 arcsec (1 km), offrant ainsi une granularité fine sur 224 pays, dont le Chili. Ces données sont essentielles pour modéliser le risque de catastrophes à grande échelle, en particulier en lien avec l'exposition aux incendies de forêt, et elles permettent de corréler les infrastructures éclairées à l'activité économique et à la population. Les données LitPop sont accompagnées d'une publication scientifique qui documente leur élaboration et leurs applications pratiques dans l'évaluation des risques physiques à l'échelle mondiale.

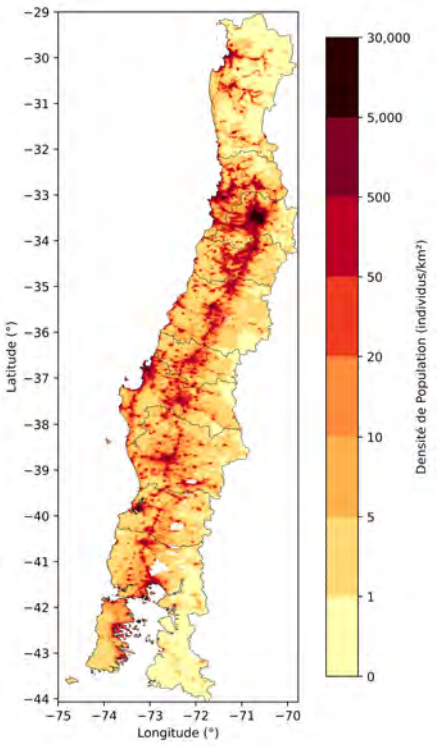


FIGURE 2.14 – Carte de densité de population

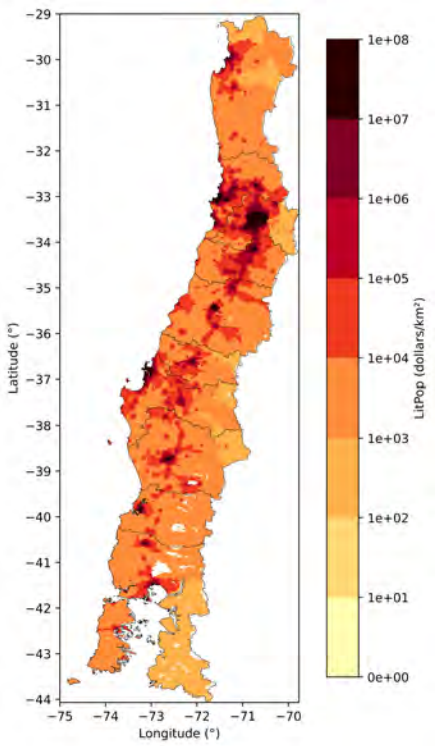


FIGURE 2.15 – Carte de LitPop

La densité de population est plus élevée dans les zones urbaines, notamment autour de Santiago et d'autres grandes villes. À mesure que l'on s'éloigne de ces centres, la densité de population et les valeurs de LitPop diminuent, laissant place à des régions moins peuplées et avec moins d'activité humaine, notamment dans les zones montagneuses et le sud rural.

2.4 Données finales

2.4.1 Homogénéisation des formats

A ce stade de l'étude, les divers types de données sont disjoints et présentent des résolutions spatiales variables. Afin d'homogénéiser l'ensemble des données, elles ont toutes été converties au format et à la résolution des feux historiques, soit dans le système de coordonnées EPSG :4326 - WGS 84 et à une résolution d'environ 30 mètres (correspondant à une arcseconde). Pour les données à résolution plus faible, une méthode de rééchantillonnage par plus proche voisin a été appliquée. Cette approche permet, pour chaque point de la nouvelle grille de 30 mètres, d'attribuer la valeur du point source le plus proche. Elle garantit ainsi la conservation des valeurs

discrètes des données, tout en évitant la création de nouvelles valeurs lors de l'augmentation de la résolution.

2.4.2 Distance aux zones urbaines et aux points de feu

Dans cette étude, un calcul a été réalisé pour déterminer la distance entre chaque point du périmètre et la zone urbanisée la plus proche, en se basant sur les informations de couverture des sols. Cette approche permet d'évaluer la proximité des zones habitées par rapport aux espaces naturels, ce qui peut s'avérer particulièrement pertinent pour l'analyse des risques liés aux incendies de forêt. En effet, la distance aux zones urbanisées constitue un indicateur clé pour identifier les zones à risque, où les interactions entre les feux de forêt et les infrastructures humaines peuvent être les plus critiques.

Par ailleurs, la variable "distance to fire" a été créée pour mesurer la distance entre chaque point du territoire et la cicatrice de feu la plus proche, quantifiant ainsi la proximité des différentes zones avec les incendies historiques. Elle permettra d'affiner la sélection des zones à inclure dans les modèles de risque afin de délimiter plus clairement les zones à risque des zones moins touchées.

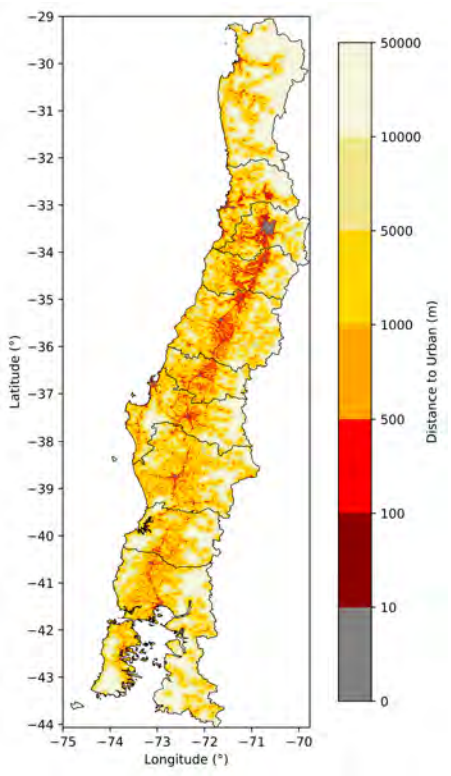


FIGURE 2.16 – Carte de distance à un point urbanisé

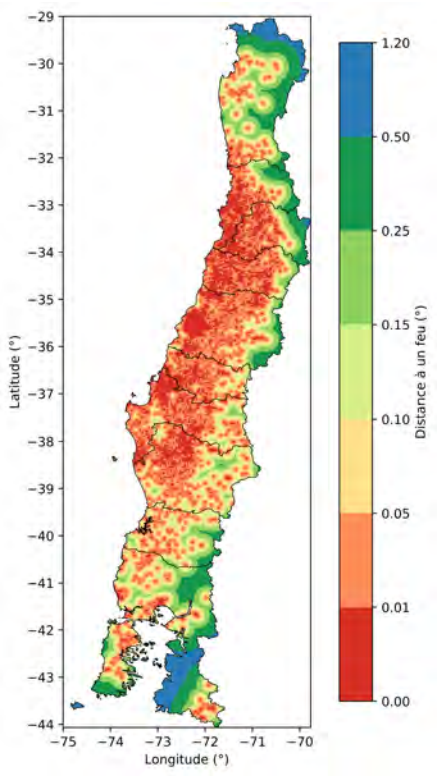


FIGURE 2.17 – Carte de distance à un point de feu passé

2.4.3 Récapitulatif des données

Ainsi, l'ensemble des données de l'étude est synthétisé dans la table 2.2.

Donnée	Source	Date	Résolution d'origine
Feux historiques	CONAF	1985-2018	30m
Distance to Fire	Calculée à partir des feux historiques	1985-2018	30m
Températures et Précipitations	CR2	1985-2018	0.05°
Indices feu météo	Copernicus	1985-2018	0.25°
Impacts de Foudre	NASA	1995-2014	1°
GHI	Global Solar Atlas	1999-2018	9″ (250m)
Altitude	Shuttle Radar Topography Mission (SRTM)	2015	1″ (30m)
Pente	Calculée à partir de l'altitude SRTM	2015	1″ (30m)
Land Cover	Déterminée à partir de Land Cover	2019	30m
TreeCover	Déterminée à partir de Land Cover	2019	30m
Shrublands	Déterminée à partir de Land Cover	2019	30m
Herbaceous	Déterminée à partir de Land Cover	2019	30m
No Vegetation	Déterminée à partir de Land Cover	2019	30m
Distance to Urban	Déterminée à partir de Land Cover	2019	30m
Densité Population	WorldPop	2020	1km
Lit Population	LitPop (ETH Zurich)	2020	30″ (1km)

TABLE 2.2 – Récapitulatif des données collectées

2.5 Analyse préliminaire

Il est possible de davantage traiter les données afin de mieux les appréhender et utiliser. Cette section fournit des statistiques préliminaires pour les données aux différentes échelles, allant de la parcelle aux hiérarchies administratives : la commune, la province et enfin la région.

Les données qui viennent d'être assemblées contiennent près de 750 millions de pixels, mais cela tient compte des valeurs manquantes en dehors du périmètre puisque le format utilisé est rectangulaire et englobe les données. Dans toute la suite de l'étude, les observations en dehors du périmètre ou contenant des valeurs manquantes sont filtrées, pour un nombre de pixels restants de 342 millions d'unité.

Les matrices de corrélation fournissent des informations sur les relations linéaires entre les variables d'un jeu de données et constituent ainsi une première étape importante pour son analyse. Chaque valeur dans la matrice représente la corrélation entre deux variables, mesurant la force et la direction de leur relation linéaire. La matrice est symétrique, c'est-à-dire que la corrélation entre les variables X et Y est identique à celle entre Y et X.

Les coefficients de corrélation varient de -1 à $+1$:

— Un coefficient de $+1$ indique une *corrélation positive parfaite* : les deux variables augmentent ou diminuent ensemble de manière proportionnelle.

— Un coefficient de -1 indique une *corrélation négative parfaite* : lorsque l'une des variables augmente, l'autre diminue proportionnellement.

— Un coefficient de 0 indique qu'il n'existe *aucune corrélation linéaire* entre les deux variables.

Les valeurs proches de -1 ou $+1$ révèlent une *relation linéaire forte* entre les variables, tandis que des valeurs proches de 0 indiquent une relation linéaire faible ou inexistante.

2.5.1 Échelle parcellaire

TABLE 2.3 – Statistiques globales descriptives

	mean	std	min	25%	50%	75%	max
Fire	0.050	0.210	0.00	0.00	0.00	0.00	1.00
Slope	12.94	10.70	0.00	3.54	11.00	21.47	85.85
Elevation	772.11	755.30	-36	183.00	482.00	1169.00	5550.00
Precipitation	121.73	42.38	5.52	36.51	107.28	173.69	389.15
Temperature	11.13	2.68	-3	8.84	11.61	13.79	19.39
LitPop	1.87e+06	2.40e+07	0.00	1813.61	3593.74	1.14e+04	9.20e+08
DistanceToUrban	5.56e+03	4.85e+03	0.00	1701.82	4107.81	8316.01	4.78e+04
PopDensity	77.89	618.09	2.16e-03	1.25	3.92	12.14	3.04e+04
GHI	4.64	0.24	1	3.82	4.85	5.27	6.80
Lightning	0.00	0.00	7.53e-05	4.13e-04	8.53e-04	1.34e-03	5.42e-03
FWI	10.06	3.56	0	3.34	9.04	16.32	33.70
TreeCover	0.400	0.430	0.00	0.00	0.00	1.00	1.00
Shrublands	0.220	0.380	0.00	0.00	0.00	0.00	1.00
Herbaceous	0.240	0.410	0.00	0.00	0.00	0.00	1.00
NoVegetation	0.130	0.320	0.00	0.00	0.00	0.00	1.00

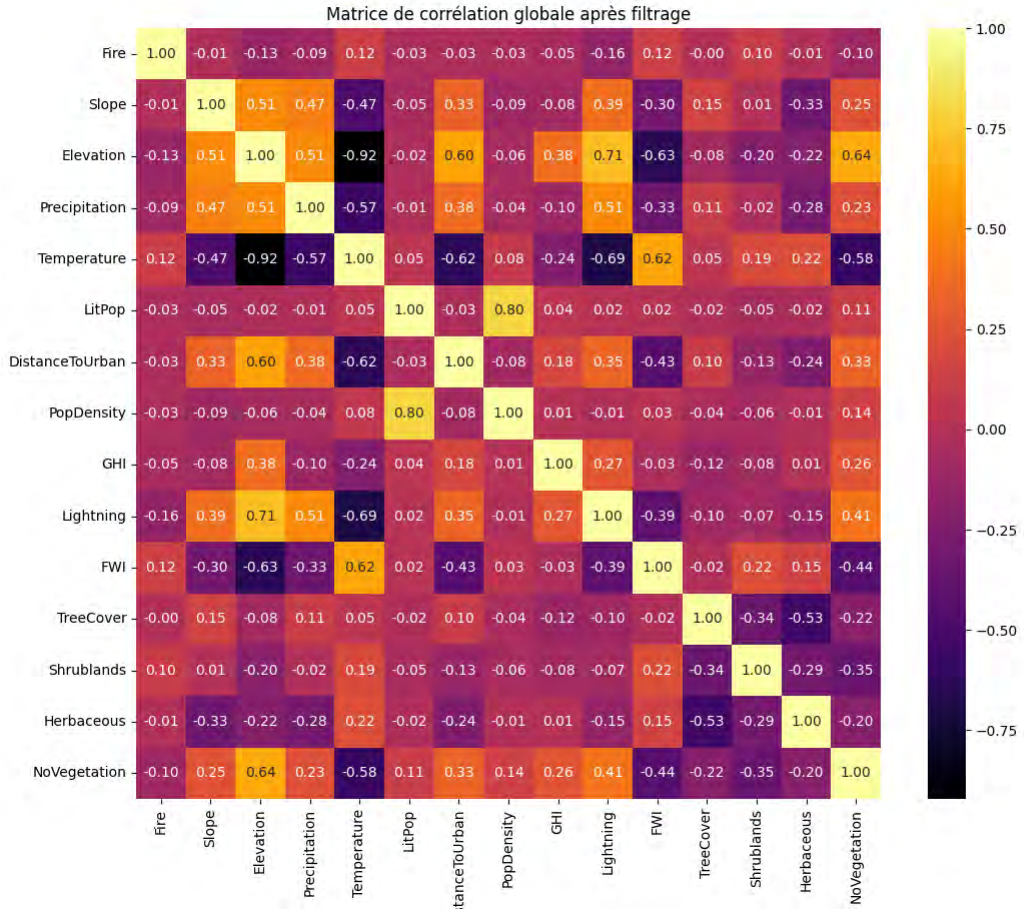
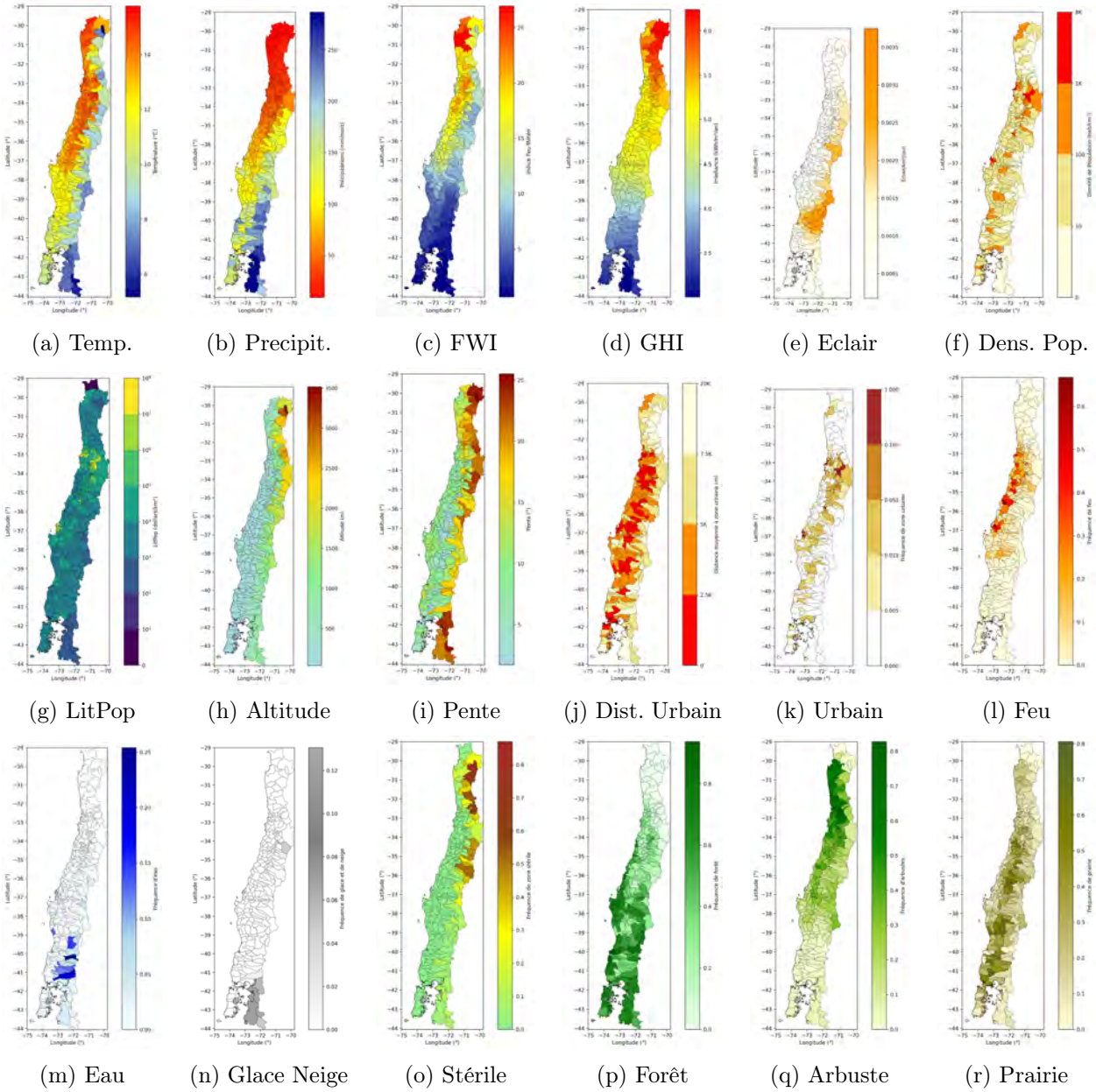


FIGURE 2.18 – Matrice de corrélation globale

2.5.2 Échelle communale



(a) Temp. (b) Precipit. (c) FWI (d) GHI (e) Eclair (f) Dens. Pop.

(g) LitPop (h) Altitude (i) Pente (j) Dist. Urbain (k) Urbain (l) Feu

(m) Eau (n) Glace Neige (o) Stérile (p) Forêt (q) Arbuste (r) Prairie

FIGURE 2.19 – Cartes des variables moyennes par commune dans les régions centrales du Chili

TABLE 2.4 – Statistiques descriptives pour chaque variable moyenne par commune

	mean	std	min	25%	50%	75%	max
Temperature	11.52	1.36	8.81	10.65	11.86	12.47	13.06
Precipitation	104	71	18	42	95	153	210
FWI	11.56	6.49	2.01	6.37	12.50	16.73	20.53
GHI	4.82	0.79	3.43	4.37	5.11	5.38	5.87
Lightning	1.15e-03	5.48e-04	5.54e-04	8.29e-04	9.60e-04	1.17e-03	2.24e-03
Population	122	214	22	29	40	68	684
LitPop	2.03e+04	3.30e+04	1.49e+03	4.03e+03	7.73e+03	1.97e+04	1.05e+05
Elevation	840	345	475	575	885	1009	1506
Slope	13.37	2.49	10.02	11.77	13.46	14.41	17.69
DistanceUrban	5497	1968	3652	4001	4749	6542	9637
Developed	0.01	0.01	0.00	0.00	0.01	0.01	0.05
Fire	0.06	0.05	0.00	0.00	0.07	0.08	0.13
Water	0.02	0.02	0.00	0.00	0.01	0.02	0.06
IceSnow	0.01	0.01	0.00	0.00	0.00	0.00	0.03
Barren	0.12	0.09	0.02	0.04	0.09	0.19	0.29
TreeCover	0.33	0.22	0.01	0.13	0.28	0.49	0.62
Shrublands	0.23	0.15	0.03	0.13	0.23	0.33	0.48
Herbaceous	0.21	0.07	0.11	0.15	0.23	0.26	0.33

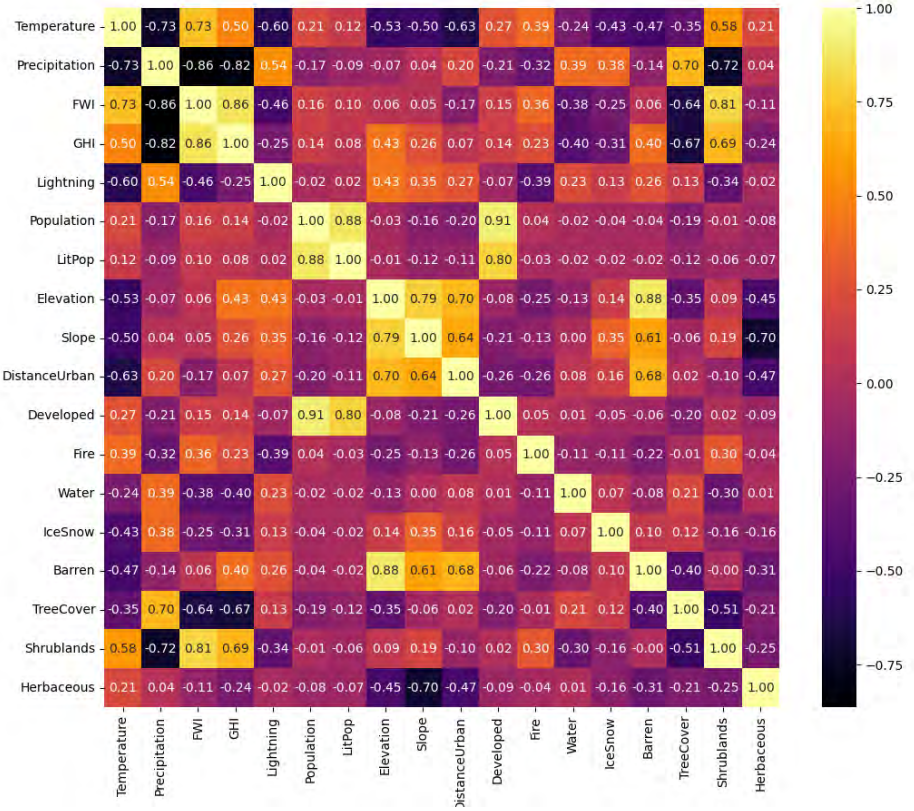
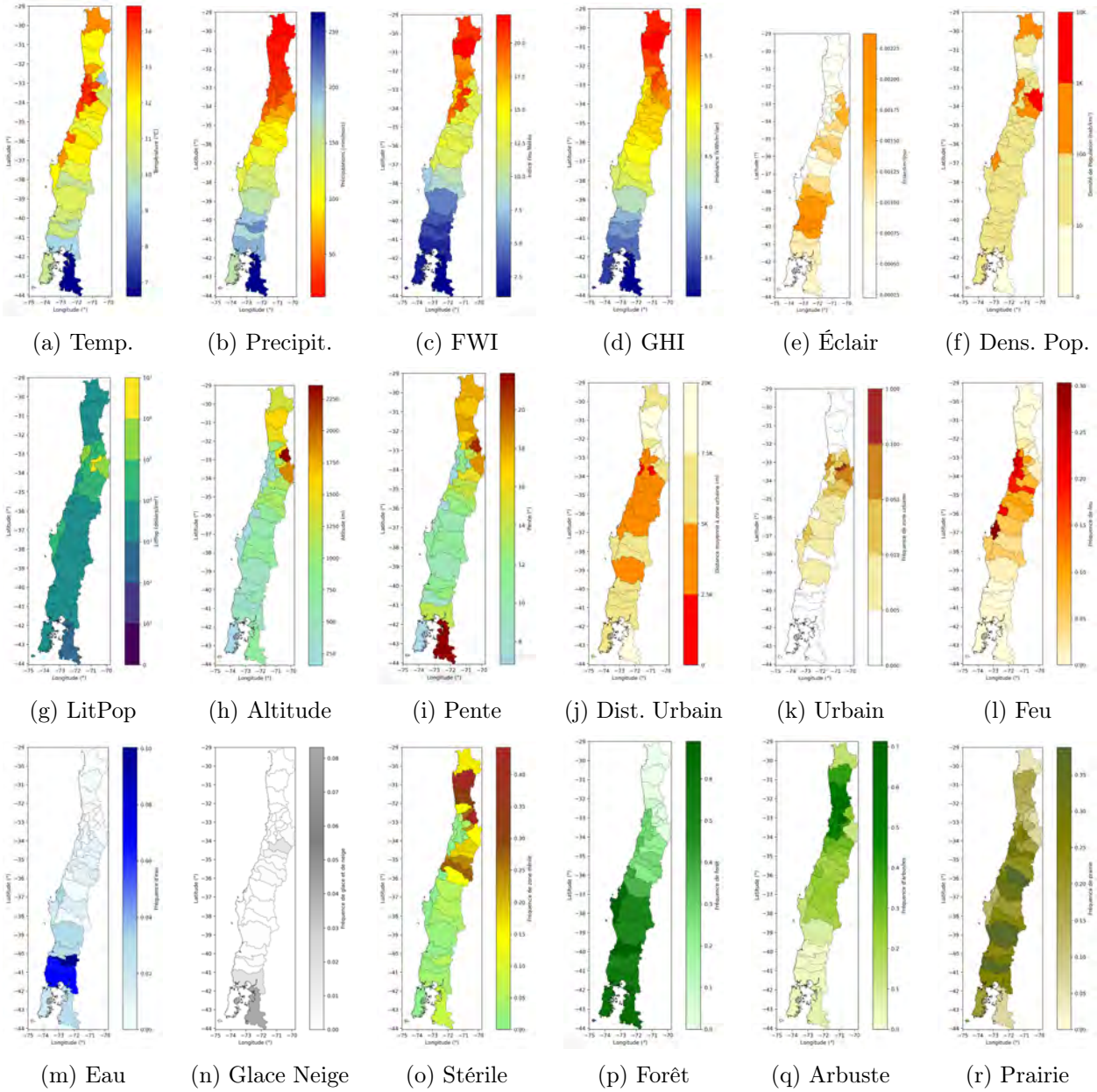


FIGURE 2.20 – Matrice de corrélation des variables moyennes par commune

2.5.3 Échelle provinciale



(a) Temp. (b) Precipit. (c) FWI (d) GHI (e) Éclair (f) Dens. Pop.

(g) LitPop (h) Altitude (i) Pente (j) Dist. Urbain (k) Urbain (l) Feu

(m) Eau (n) Glace Neige (o) Stérile (p) Forêt (q) Arbuste (r) Prairie

FIGURE 2.21 – Cartes des variables moyennes par province dans les régions centrales du Chili

TABLE 2.5 – Statistiques descriptives pour chaque variable moyenne par province

	mean	std	min	25%	50%	75%	max
Temperature	11.96	1.82	6.60	10.75	12.06	13.52	14.69
Precipitation	89	67	12	34	71	135	268
FWI	12.83	6.31	0.99	8.83	12.82	18.77	22.12
GHI	4.92	0.73	3.11	4.66	5.14	5.39	5.97
Lightning	9.60e-04	5.42e-04	2.25e-04	5.40e-04	9.76e-04	1.22e-03	2.37e-03
Population	186	388	1	29	39	104	1672
LitPop	1.19e+05	3.38e+05	0.00	4.76e+03	7.81e+03	4.96e+04	1.53e+06
Elevation	805	547	158	375	678	1076	2356
Slope	13.09	4.01	6.82	10.03	12.34	16.11	21.89
DistanceUrban	5200	2502	1461	3704	4858	6109	10874
Developed	0.02	0.03	0.00	0.00	0.01	0.02	0.16
Fire	0.08	0.08	0.00	0.01	0.05	0.10	0.30
Water	0.01	0.02	0.00	0.00	0.00	0.01	0.10
IceSnow	0.00	0.01	0.00	0.00	0.00	0.00	0.08
Barren	0.10	0.12	0.00	0.01	0.05	0.17	0.44
TreeCover	0.30	0.22	0.00	0.09	0.28	0.49	0.69
Shrublands	0.28	0.19	0.00	0.13	0.24	0.44	0.71
Herbaceous	0.20	0.10	0.00	0.13	0.20	0.26	0.39

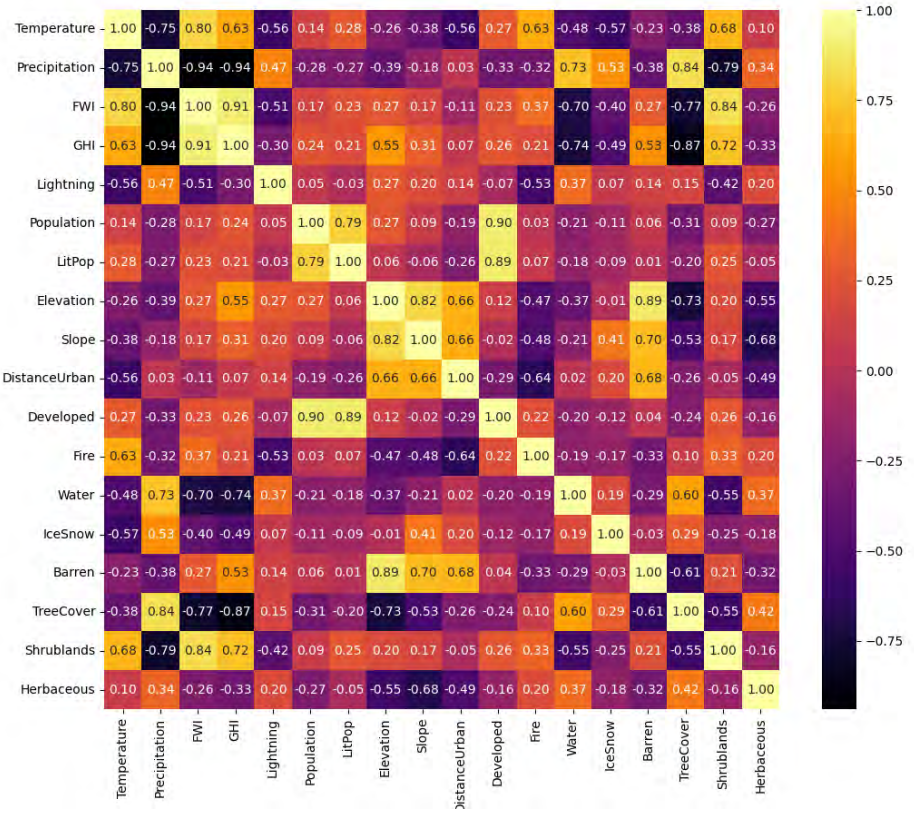
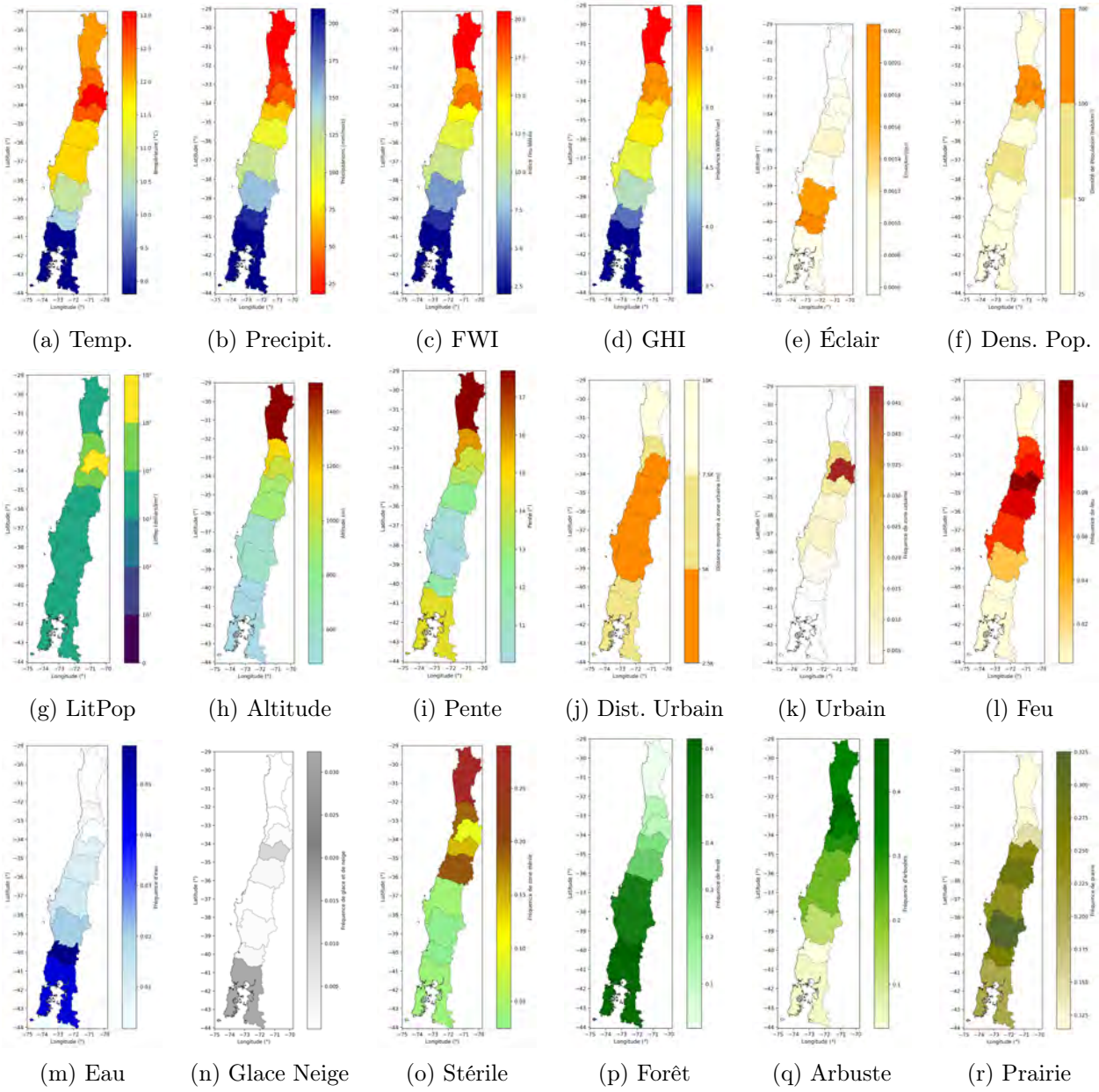


FIGURE 2.22 – Matrice de corrélation des variables moyennes par province

2.5.4 Échelle régionale



(a) Temp. (b) Precipit. (c) FWI (d) GHI (e) Éclair (f) Dens. Pop.

(g) LitPop (h) Altitude (i) Pente (j) Dist. Urbain (k) Urbain (l) Feu

(m) Eau (n) Glace Neige (o) Stérile (p) Forêt (q) Arbuste (r) Prairie

FIGURE 2.23 – Cartes des variables moyennes par région du Chili central

TABLE 2.6 – Statistiques descriptives pour chaque variable moyenne par région

	mean	std	min	25%	50%	75%	max
Temperature	11.52	1.36	8.81	10.65	11.86	12.47	13.06
Precipitation	104	71	18	42	95	153	210
FWI	11.56	6.49	2.01	6.37	12.50	16.73	20.53
GHI	4.82	0.79	3.43	4.37	5.11	5.38	5.87
Lightning	1.15e-03	5.48e-04	5.54e-04	8.29e-04	9.60e-04	1.17e-03	2.24e-03
Population	122	214	22	29	40	68	684
LitPop	2.03e+04	3.30e+04	1.49e+03	4.03e+03	7.73e+03	1.97e+04	1.05e+05
Elevation	840	345	475	575	885	1009	1506
Slope	13.37	2.49	10.02	11.77	13.46	14.41	17.69
DistanceUrban	5497	1968	3652	4001	4749	6542	9637
Developed	0.01	0.01	0.00	0.00	0.01	0.01	0.05
Fire	0.06	0.05	0.00	0.00	0.07	0.08	0.13
Water	0.02	0.02	0.00	0.00	0.01	0.02	0.06
IceSnow	0.01	0.01	0.00	0.00	0.00	0.00	0.03
Barren	0.12	0.09	0.02	0.04	0.09	0.19	0.29
TreeCover	0.33	0.22	0.01	0.13	0.28	0.49	0.62
Shrublands	0.23	0.15	0.03	0.13	0.23	0.33	0.48
Herbaceous	0.21	0.07	0.11	0.15	0.23	0.26	0.33

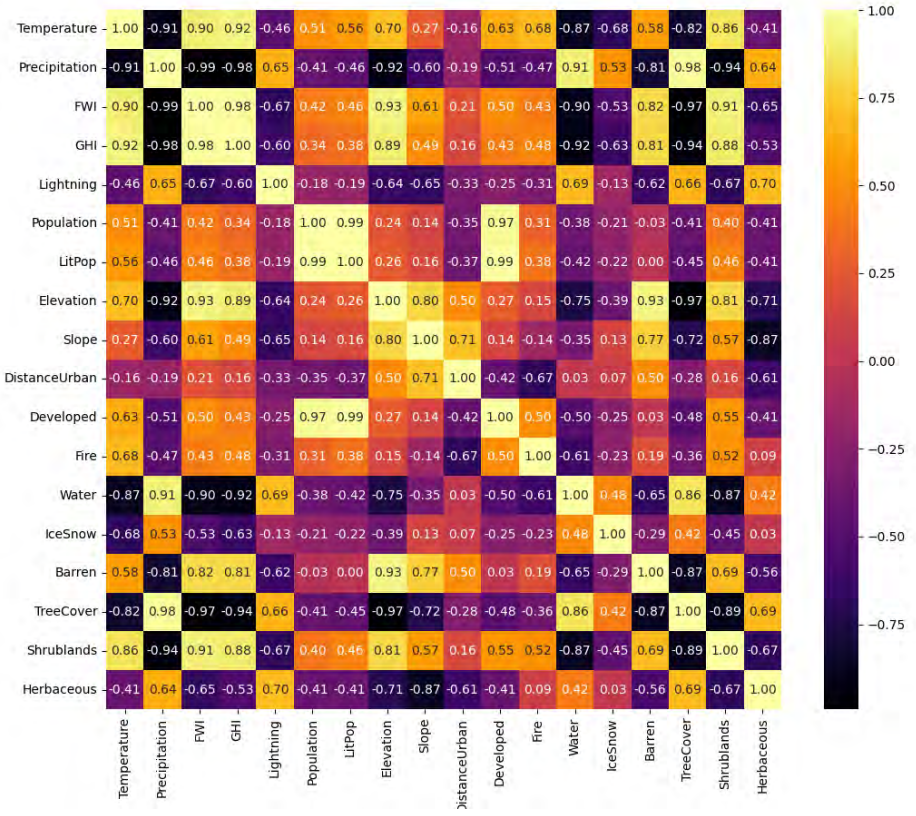


FIGURE 2.24 – Matrice de corrélation des variables moyennes par région

Chapitre 3

Modélisation du risque de feu de forêt

La démarche de modélisation du risque de feu de forêt choisie consiste à diviser les communes en groupes homogènes, en fonction de leurs caractéristiques, afin de permettre une calibration ciblée des modèles prédictifs. Chaque modèle tentera alors d'assigner une fréquence d'occurrence de feu à chaque parcelle dans ces groupes, dans le but d'obtenir des prédictions fines et adaptées.

3.1 Classification des communes

La première étape consiste à regrouper les communes présentant des caractéristiques similaires, afin de mieux calibrer les futurs modèles prédictifs pour chaque groupe spécifique. Cette classification permettra une segmentation plus fine et, par conséquent, une amélioration de la précision des modèles. Cette section décrit tout d'abord la méthode de classification utilisée c'est-à-dire une analyse en composantes principales (PCA) combinée à une classification par k-means. Les résultats de cette méthode appliquée à l'ensemble de communes étudié seront ensuite présentés.

3.1.1 Analyse en composantes principales

L'Analyse en Composantes Principales (PCA) est une technique de réduction de dimensionnalité qui simplifie des ensembles de données multivariées tout en conservant le maximum d'information possible. Son objectif principal est de réduire la dimension des données, rendant ainsi l'analyse plus rapide et moins complexe, tout en préservant une part significative de la variance. La PCA repose sur la recherche de combinaisons linéaires des variables initiales, appelées *composantes principales*, qui maximisent la variance et minimisent la redondance. Ces composantes sont orthogonales (donc non corrélées) et permettent de représenter les données dans un espace de faible dimension, facilitant ainsi l'interprétation et la visualisation tout en conservant la structure essentielle des données d'origine. Pour la suite seront présentés, la théorie, les calculs mathématiques et les décisions statistiques qui guident la PCA, ainsi que son interprétation.

Préparation des données

Pour que les variables contribuent de manière égale à l'analyse en composantes principales (PCA), indépendamment de leur échelle de mesure, il est d'usage de centrer et de normaliser les données. Soit X la matrice de données où chaque ligne représente une observation et chaque

colonne une variable, les données centrées et normalisées Z sont obtenues par :

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{\sigma_j}$$

où $\bar{X}_j$ est la moyenne de la j-ème variable et σ_j son écart-type.

Calcul de la covariance et des composantes principales

La PCA repose sur le calcul de la matrice de covariance des données. Pour les données standardisées X, la matrice de covariance Σ est définie par :

$$\Sigma = \frac{1}{n-1} X^T X$$

Cette matrice symétrique de taille $p \times p$ contient les covariances entre les paires de variables, capturant ainsi leur relation linéaire. Pour obtenir les composantes principales, on effectue une *décomposition en valeurs propres* de la matrice Σ :

$$\Sigma = P \Lambda P^T$$

où P est une matrice de vecteurs propres (ou directions principales) de taille $p \times p$, et Λ est une matrice diagonale contenant les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_p$ telles que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Chaque vecteur propre de P correspond à une direction dans l'espace des variables d'origine, et la variance expliquée par chaque composante est proportionnelle à sa valeur propre.

Sélection des composantes principales

La variance expliquée par chaque composante principale est donnée par la proportion de sa valeur propre sur la somme des valeurs propres :

$$\text{Variance expliquée par la composante } i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

On peut alors choisir un sous-ensemble de composantes principales qui cumule une part significative de la variance totale. Un critère courant consiste à retenir les premières composantes principales qui expliquent entre 90% et 95% de la variance totale. Une autre méthode est le critère du coude, qui consiste à tracer les valeurs propres en fonction du nombre de composantes et à identifier le point où l'ajout de nouvelles composantes n'apporte plus qu'un gain marginal de variance.

Projection

Après avoir sélectionné le nombre optimal de composantes principales k, on projette les données initiales dans cet espace réduit. Soit P_k la matrice des k premiers vecteurs propres, correspondant aux k plus grandes valeurs propres. La projection des données X dans cet espace est alors donnée par :

$$X_{\text{PCA}} = X P_k$$

Chaque ligne de X_{PCA} correspond aux coordonnées d'une observation dans l'espace des k composantes principales. Cela permet de réduire la dimension des données tout en conservant l'essentiel de leur structure.

Interprétation et visualisation des résultats

Pour interpréter les composantes principales, il convient d'examiner les coefficients (ou poids) des variables d'origine associés à chaque composante, donnés par les éléments de la matrice P_k où chaque colonne représente une composante. Les variables avec des poids élevés dans une même composante sont fortement corrélées et influencent significativement cette dimension, ce qui aide à identifier les caractéristiques dominantes des données.

La réduction de dimension par PCA permet également de visualiser les données dans un espace simplifié, notamment en 2D ou 3D lorsque $k = 2$ ou $k = 3$, facilitant ainsi l'identification de regroupements naturels. Le *biplot* est une option particulièrement informative, car il montre à la fois les observations et les variables d'origine dans l'espace des composantes, révélant la contribution de chaque variable et la structure des données.

Limites

Il est à noter certaines limitations de la PCA. D'abord, elle est une méthode linéaire, ce qui signifie qu'elle est moins adaptée pour des données ayant des relations non linéaires. De plus, la PCA est sensible aux valeurs aberrantes (outliers), qui peuvent affecter les vecteurs propres et déformer les résultats.

3.1.2 La méthode des k-means

Le clustering k-means est une méthode de partitionnement non supervisée qui permet de regrouper les observations en un nombre k de clusters spécifié, basé sur leur similarité. Dans ce contexte, chaque cluster est représenté par un centroïde, qui correspond à la moyenne des points appartenant à ce cluster. Le but de l'algorithme est de minimiser l'inertie intra-cluster, c'est-à-dire la somme des distances au carré entre chaque point et le centroïde de son cluster.

Cela se traduit mathématiquement par le fait de minimiser la fonction objective suivante :

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

où C_j est le j-ème cluster et μ_j est le centroïde de ce cluster, défini par :

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

Cette fonction J représente l'inertie intra-cluster, c'est-à-dire la somme des distances au carré entre chaque point x_i et le centroïde μ_j du cluster auquel il est assigné. En minimisant cette inertie, l'algorithme optimise la cohésion des clusters et maximise leur homogénéité interne.

Processus itératif du clustering k-means

L'algorithme k-means suit un processus itératif en quatre étapes principales :

— **Initialisation** : Initialiser aléatoirement k centroïdes qui représentent les centres initiaux des k clusters. Il est courant de relancer l'algorithme avec différentes initialisations pour éviter la convergence vers un minimum local.

— **Affectation des points aux clusters** : Pour chaque observation x_i, calculer la distance entre x_i et chacun des k centroïdes, puis assigner x_i au cluster du centroïde le plus proche selon la distance euclidienne :

$$c(i) = \arg\min_j \|x_i - \mu_j\|^2$$

où $c(i)$ est l'indice du cluster auquel l'observation x_i est assignée.

— **Mise à jour des centroïdes** : Recalculer chaque centroïde comme la moyenne des points assignés à son cluster :

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

Cela permet de recentrer chaque cluster autour de ses observations actuelles, améliorant la cohésion interne du cluster.

— **Convergence** : Répéter les étapes d'affectation et de mise à jour jusqu'à ce que les centroïdes ne changent plus ou que la variation soit minimale. Il est conseillé de relancer plusieurs fois l'algorithme pour maximiser les chances d'obtenir la solution optimale, car l'algorithme peut converger vers un minimum local.

Appliquer le clustering k-means dans un espace de dimension réduite, comme celui obtenu par l'analyse en composantes principales (PCA), présente plusieurs avantages. La réduction de dimension par PCA permet de supprimer le bruit et la redondance, rendant ainsi le clustering plus fiable et rapide.

Méthode du coude

Pour déterminer le nombre optimal de clusters k, on utilise souvent la *méthode du coude*, qui consiste à calculer l'inertie intra-cluster $J(k)$ pour différentes valeurs de k et à tracer $J(k)$ en fonction de k. L'inertie diminue généralement à mesure que k augmente, mais après un certain point, la réduction d'inertie devient marginale. Ce point, appelé le « coude » de la courbe, est interprété comme une estimation du nombre optimal de clusters. L'inertie intra-cluster pour chaque k est donnée par :

$$J(k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

où μ_j représente les centroïdes. En complément de cette méthode, on peut utiliser des indices de qualité de clustering comme le *coefficient de silhouette*, qui mesure la cohésion intra-cluster par rapport à la séparation inter-cluster.

Après avoir déterminé le nombre optimal de clusters, l'algorithme k-means est appliqué dans l'espace réduit en utilisant les k composantes principales sélectionnées. Cette application dans un espace simplifié rend le clustering plus efficace et facilite l'interprétation des résultats.

3.1.3 Application aux communes

Les régions du périmètre étudié ont été subdivisées en communes pour cette analyse. Au Chili, les régions couvrent des territoires étendus et variés, présentant des diversités géographiques et climatiques importantes. Cette subdivision en communes permet donc de classifier les données à une échelle plus homogène, réduisant ainsi l'impact de cette variabilité et aidant à rendre l'analyse plus pertinente pour des modèles de prédiction locaux.

Préparation des données

Le sous-ensemble de variables pertinentes a été sélectionné pour la PCA et le clustering : *FWI, Shrublands, LitPop, Elevation, DistanceToUrban, Temperature, Slope, Precipitation, Lightning, Herbaceous, TreeCover, GHI, NoVegetation* et *PopulationDensity*. Cependant, la variable *Fire* n'a pas été incluse dans l'analyse. Cette exclusion s'explique par le fait que *Fire* est une variable cible dans le modèle de risque de feu, et son inclusion pourrait biaiser la formation des clusters en introduisant des informations utilisées dans les prédictions ultérieures. Avant l'application de la PCA, les données ont été standardisées pour mettre toutes les variables sur une même échelle, empêchant ainsi les variables avec des valeurs plus élevées (comme l'altitude) de dominer l'analyse.

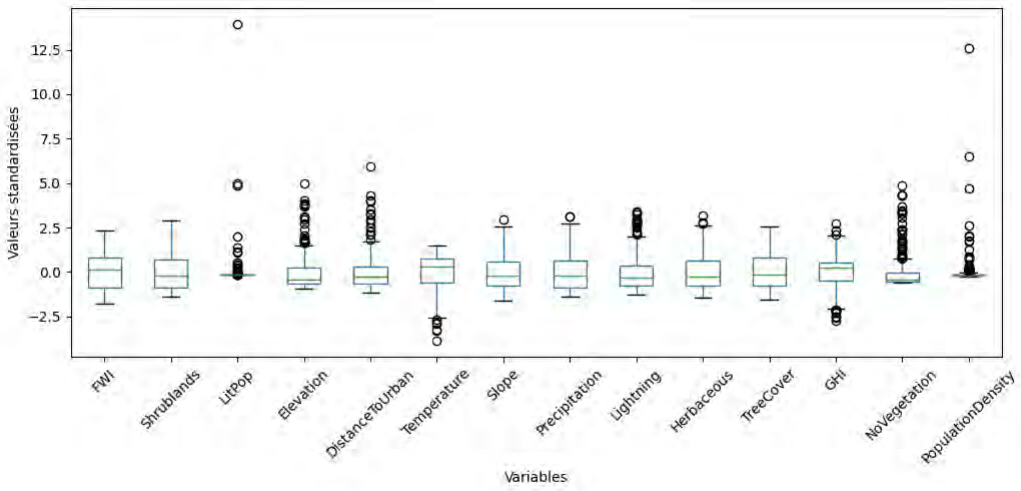


FIGURE 3.1 – Distribution des données standardisées par variable.

Choix du nombre de dimensions

La PCA a été appliquée aux variables sélectionnées pour calculer la variance expliquée par chaque composante principale, permettant d'identifier le nombre de composantes nécessaires pour conserver l'information significative des données. Quatre composantes principales ont été retenues, expliquant ainsi 85% de la variance totale. Un graphique de la variance expliquée cumulée a été dressé pour visualiser les contributions cumulatives de chaque composante principale. Cette représentation permet de déterminer le point au-delà duquel l'ajout de nouvelles composantes n'apporte plus d'amélioration substantielle à la variance expliquée.

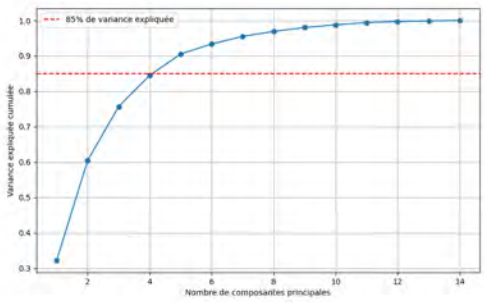


FIGURE 3.2 – Variance expliquée cumulée par nombre de composantes principales.

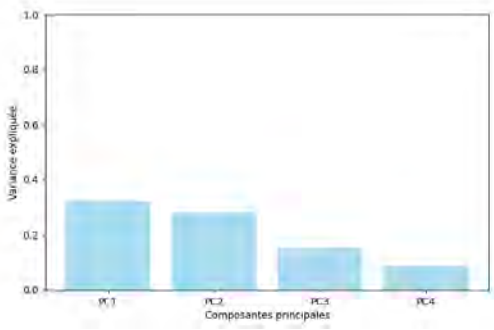


FIGURE 3.3 – Variance expliquée par chaque composante principale.

Choix du nombre de clusters

Après la réduction de dimensionnalité, la méthode du coude a été appliquée aux valeurs d'inertie pour déterminer le nombre optimal de clusters. Ici, le nombre de quatre clusters a été retenu.

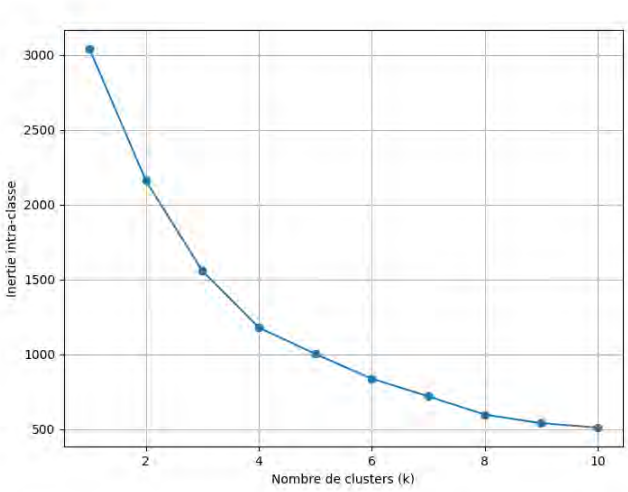


FIGURE 3.4 – Méthode du coude pour déterminer le nombre optimal de clusters.

Application et visualisation

Après avoir déterminé le nombre optimal de clusters, l'algorithme K-means a été appliqué aux données transformées par la PCA. La projection des données sur les deux premières composantes principales est représentée par un nuage de points colorés, permettant de visualiser les clusters identifiés par K-means. Chaque commune est ainsi assignée à un cluster en fonction de sa proximité avec le centre de celui-ci.

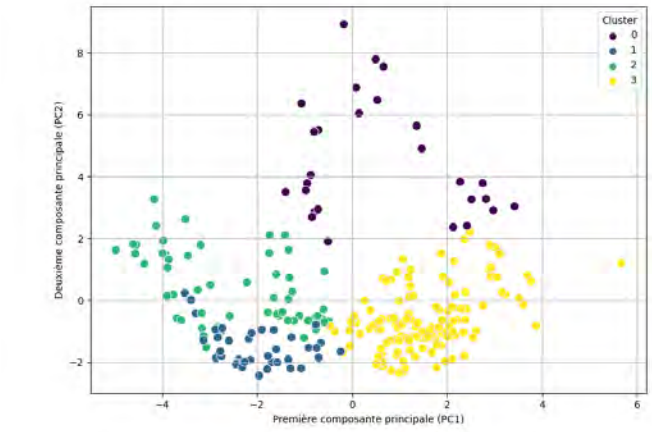


FIGURE 3.5 – Projection sur les deux premières composantes principales avec clusters.

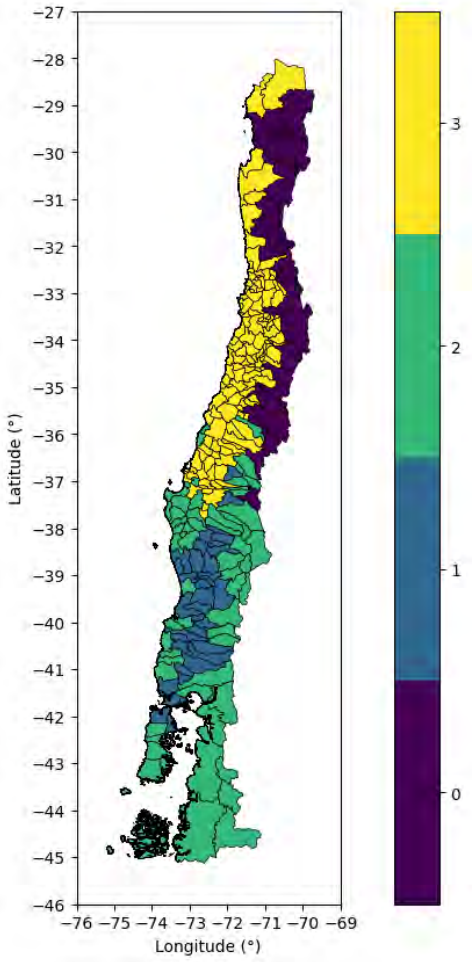


FIGURE 3.6 – Carte des communes colorées par cluster.

Interprétation

La lecture des résultats des clusters ainsi obtenus met en évidence des profils géographiques spécifiques parmi les communes. Il est intéressant de noter l'homogénéité spatiale des clusters : bien que les critères de regroupement n'incluent pas explicitement la localisation géographique, les clusters forment des ensembles presque totalement contigus. Cela suggère que les caractéristiques locales des communes, telles que les facteurs climatiques ou environnementaux, sont suffisamment similaires pour que les communes voisines partagent souvent un même cluster.

La carte révèle également des distinctions géographiques majeures, comme la différenciation entre les zones de la cordillère des Andes et les autres régions, reflétant des variations de facteurs environnementaux et potentiellement de risque de feu. Cependant, certaines zones de transition montrent une répartition un peu plus mélangée des clusters, indiquant des régions où les caractéristiques locales varient progressivement, créant des frontières moins marquées entre les clusters. Cela pourrait traduire des gradients environnementaux ou des zones de transition entre différents types de végétation ou de climat.

Cette segmentation géographique des clusters facilitera l'analyse et la modélisation prédictive dans la section suivante. En effet, les clusters ainsi identifiés pourront être utilisés pour entraîner des modèles prédictifs du risque de feu, ajustés aux caractéristiques propres de chaque groupe de communes, ce qui pourrait améliorer la précision des prédictions en fonction des spécificités de chaque zone.

TABLE 3.1 – Tableau récapitulatif des observations de feu par fichier et par cluster

Cluster	**Nombre d'observations**	**Nombre d'observations de feu**	**Pourcentage d'observations de feu (%)**
Cluster 0	78,711,141	866,068	1.10
Cluster 1	54,648,926	1,190,526	2.18
Cluster 2	112,672,930	2,032,435	1.80
Cluster 3	96,309,326	12,112,095	12.58

Le cluster 3 figure parmi les communes qui ont le plus brûlé, en nombre et en proportion avec plus de 12,5% des points touchés.

3.2 Méthode de GLM

3.2.1 Théorie

Les Modèles Linéaires Généralisés (GLM) sont construits pour modéliser des relations entre une variable réponse Y et des prédicteurs X dans des situations où les hypothèses classiques de normalité et de variance constante ne sont pas respectées. Les GLM reposent sur trois composants principaux : une distribution de la variable réponse dans la famille exponentielle, une fonction de lien reliant l'espérance conditionnelle de Y à une combinaison linéaire des prédicteurs, et un prédicteur linéaire η. Les observations de Y sont supposées indépendantes et la variance de Y est liée à sa moyenne par une fonction $V(\mu)$ spécifique à la distribution choisie. Ces hypothèses permettent de garantir l'inférence statistique et l'estimation correcte des coefficients β.

Distribution de la variable réponse

Dans un GLM, Y est supposée suivre une distribution de la famille exponentielle, englobant des distributions comme la normale, binomiale, Poisson, ou Gamma. La forme générale de densité de probabilité pour cette famille est donnée par :

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

où θ est le paramètre canonique, ϕ le paramètre de dispersion, et $a(\cdot), b(\cdot), c(\cdot)$ sont des fonctions spécifiques à chaque distribution. Un des avantages des GLM est leur capacité à modéliser des variances dépendantes de l'espérance $\mu = E(Y)$, permettant de traiter des données où la variance change avec la moyenne.

Fonction de lien

La fonction de lien $g(\cdot)$ relie l'espérance μ de Y au prédicteur linéaire η, de manière à linéariser la relation entre Y et X. La relation entre μ et $X\beta$ est ainsi :

$$g(\mu_i) = \eta_i = X_i\beta$$

où X_i est le vecteur de prédicteurs et β le vecteur des coefficients.

Cas de la classification binaire

Les GLM sont couramment appliqués à la classification binaire, un cas où la variable réponse Y prend seulement deux valeurs correspondant à deux classes distinctes. Dans ce contexte, la distribution de Y est modélisée par une loi binomiale, car chaque observation représente une occurrence binaire indépendante. Pour lier cette variable réponse à une combinaison linéaire des prédicteurs X, on utilise typiquement la fonction de lien logit, adaptée aux données binaires. La fonction de lien logit transforme l'espérance conditionnelle $\mu = E(Y)$ de Y en fonction de la probabilité de succès $P(Y = 1) = \mu$ en appliquant la transformation suivante :

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = X\beta$$

où $X\beta$ est le prédicteur linéaire, combinaison linéaire des variables explicatives X et des coefficients β. Cette transformation garantit que μ, qui représente la probabilité de succès, reste dans l'intervalle $(0, 1)$, et linéarise la relation entre les prédicteurs et le logarithme du rapport des cotes (ou odds ratio). Ainsi, chaque coefficient β_j indique l'effet d'une unité de variation de X_j sur le logarithme des cotes de la probabilité de succès.

L'estimation des coefficients β dans ce modèle est effectuée via la méthode du maximum de vraisemblance. Dans le cadre de la régression logistique, la fonction de vraisemblance pour n observations indépendantes est donnée par :

$$L(\beta) = \prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{1-y_i}$$

où $\mu_i = P(Y_i = 1|X_i) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$ est la probabilité de succès pour l'observation i.

3.2.2 Mesures de performance

Dans le cadre d'une classification binaire, il existe différentes mesures de performance pour évaluer la qualité du modèle. Ces métriques permettent d'évaluer non seulement la capacité du modèle à prédire correctement les classes positives, mais aussi son comportement face aux classes négatives. Les mesures classiques incluent la matrice de confusion, la précision, le rappel et le F1-score, la courbe ROC et l'AUC.

Matrice de confusion

La matrice de confusion est un outil pour évaluer la performance d'un modèle de classification. Elle présente une vue d'ensemble des prédictions correctes et incorrectes, séparées en quatre catégories :

$$\begin{bmatrix} \text{TP} & \text{FP} \\ \text{FN} & \text{TN} \end{bmatrix}$$

avec :

— **TP** (vrais positifs) : le nombre de cas où le modèle a correctement prédit un feu de forêt.

— **FP** (faux positifs) : le nombre de cas où le modèle a prédit un feu de forêt alors qu'il n'y en avait pas.

— **FN** (faux négatifs) : le nombre de cas où le modèle a prédit l'absence de feu alors qu'un feu était présent.

— **TN** (vrais négatifs) : le nombre de cas où le modèle a correctement prédit l'absence de feu.

La matrice de confusion est particulièrement utile pour visualiser les types d'erreurs commises par le modèle, et permet de dériver des métriques supplémentaires telles que la précision, le rappel et le F1-score.

Précision

La précision est définie comme la proportion de prédictions positives correctes parmi toutes les prédictions positives. En d'autres termes, elle mesure la fiabilité des prédictions positives du modèle :

$$\text{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Une précision élevée signifie que lorsque le modèle prédit un feu, il est souvent correct. Cependant, cette métrique ne tient pas compte des faux négatifs (FN), et peut donc être trompeuse si les faux négatifs sont nombreux.

Rappel

Le rappel (ou sensibilité) mesure la proportion de vrais positifs parmi tous les exemples réellement positifs, c'est-à-dire la capacité du modèle à identifier correctement les feux de forêt :

$$\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Le rappel est utile dans des contextes où il est important de minimiser les faux négatifs, par exemple lorsque un risque non détecté représente un danger important. Cependant, un rappel élevé peut parfois être obtenu au détriment de la précision, si le modèle produit trop de faux positifs.

F1-score

Le F1-score est la moyenne harmonique entre la précision et le rappel, formant ainsi un compromis entre les deux. Il est particulièrement utile dans des contextes où il est nécessaire d'équilibrer la précision et le rappel, surtout lorsque les classes sont déséquilibrées :

$$F1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Le F1-score prend des valeurs entre 0 et 1, où 1 représente une performance parfaite.

Courbe ROC

La courbe ROC (*Receiver Operating Characteristic*) est un outil graphique permettant d'évaluer la performance d'un modèle de classification binaire à différents seuils. Elle trace le **taux de vrais positifs** (TPR) en fonction du **taux de faux positifs** (FPR), illustrant ainsi la capacité du modèle à distinguer les classes positives des classes négatives. Chaque point de la courbe correspond à une combinaison de TPR et de FPR pour un seuil donné. En faisant varier ce seuil, on obtient l'ensemble de la courbe, montrant comment ces deux métriques évoluent en fonction de la sensibilité du modèle.

Un modèle idéal atteindrait un TPR de 1 et un FPR de 0, correspondant au point en haut à gauche du graphique (0,1). En revanche, un modèle qui fait des prédictions aléatoires se situerait le long de la diagonale de (0,0) à (1,1), où le TPR et le FPR augmentent proportionnellement.

Interprétation et compromis des seuils

La courbe ROC est particulièrement utile pour analyser le compromis entre sensibilité (capacité à détecter les feux) et spécificité (capacité à éviter les fausses alertes), à différents seuils de classification.

— Seuils bas : Un seuil bas rend le modèle plus sensible, augmentant ainsi le TPR, mais également le FPR. Ce compromis est parfois souhaitable lorsque les faux négatifs (feux non détectés) sont coûteux.

— Seuils élevés : À l'inverse, un seuil élevé réduit le nombre de faux positifs, mais diminue aussi le TPR, ce qui peut entraîner la non-détection de certains feux.

Plus la courbe ROC s'éloigne de la diagonale et se rapproche du coin supérieur gauche, meilleure est la performance du modèle.

Avantages et limites de la courbe ROC

La courbe ROC présente plusieurs avantages :

— Elle permet de visualiser la performance du modèle sur une large gamme de seuils, offrant une vue globale du compromis entre détection des feux et minimisation des fausses alertes.

— Elle est facile à interpréter et facilite la comparaison entre différents modèles.

Cependant, en présence de données fortement déséquilibrées (comme dans le cas des feux de forêt, où les événements positifs sont rares), la courbe ROC peut donner une impression exagérée de la performance du modèle. Dans ce cas, il est souvent pertinent de compléter l'analyse avec d'autres métriques, telles que le F1-score, qui sont plus sensibles aux classes minoritaires.

L'AUC

L'AUC (aire sous la courbe ROC) est une mesure quantitative de la performance globale d'un modèle de classification binaire, indépendante du seuil de classification. Elle résume en une seule valeur la capacité du modèle à discriminer correctement entre les classes positives et négatives sur tous les seuils possibles.

Cette aire est définie mathématiquement comme l'aire obtenue en intégrant le TPR en fonction du FPR pour tous les seuils possibles :

$$AUC = \int_0^1 \mathrm{TPR}(f)\, dFPR(f)$$

où f représente les différents seuils de classification.

Interprétation

L'AUC est comprise entre 0 et 1 :

— Une AUC de **1.0** indique un modèle parfait, qui sépare complètement les classes positives et négatives.

— Une AUC de **0.5** correspond à un modèle aléatoire, incapable de distinguer les classes.

— Une AUC inférieure à 0.5 suggère que le modèle fait pire qu'une prédiction aléatoire, potentiellement en inversant les classes.

Plus l'AUC est proche de 1, meilleure est la capacité du modèle à distinguer entre les classes. Elle mesure la probabilité qu'une observation positive soit classée avec une probabilité plus élevée qu'une observation négative.

Avantages et limites

L'AUC est utile pour évaluer les modèles de manière objective, car elle est indépendante du seuil de classification. Cependant, dans des jeux de données déséquilibrés, l'AUC peut masquer une faible performance sur la classe minoritaire. Par conséquent, bien que l'AUC offre une vue d'ensemble de la qualité du modèle, elle doit souvent être complétée par des métriques comme le F1-score pour mieux évaluer les performances sur les classes rares.

3.2.3 Mise en œuvre du modèle GLM

L'objectif de cette partie est de développer un GLM pour prédire le caractère brûlé ou non de cellules de 30 m^2 dans les régions centrales du Chili. Pour cela, la démarche suivante a été employée pour chaque cluster de communes établit précédemment :

1. **Répartition des communes en entraînement et test** : Les communes sont divisées en deux groupes équilibrés pour préserver la diversité spatiale et éviter la redondance dans les données.
2. **Échantillonnage de l'ensemble d'entraînement** : Un sous-échantillon de 2% des observations des communes d'entraînement est utilisé pour réduire les coûts computationnels.
3. **Construction d'un ensemble de test équilibré** : Un ensemble de test équilibré en termes de classes brûlées et non brûlées est construit à partir des observations des communes testées en prenant toutes les observations de classes brûlées et le même nombre aléatoire de non brûlées.
4. **Normalisation** : Les variables explicatives sont normalisées.
5. **Paramétrage** : Un seuil de décision de 0,5 est utilisé pour équilibrer la classification entre classes. Le nombre maximal d'itérations est fixé à 2000 pour assurer la convergence sans risque de boucle infinie.

6. **Sélection des variables** : Les variables sont réduites de manière itérative pour maximiser le F1-Score moyen des classes 0 et 1, optimisant ainsi la pertinence des variables restantes.
7. **Carte des probabilités** : Le modèle, une fois calibré, est appliqué aux régions centrales du Chili pour produire une carte prédictive des probabilités de brûlure par cellule de 30 m^2.

Sélection des communes pour l'entraînement et le test

La classification préalable des communes en clusters homogènes par K-means assure une certaine cohérence dans les caractéristiques de chaque groupe. Cela permet de diviser les communes de manière équilibrée en deux groupes (50% pour l'entraînement et 50% pour le test) tout en préservant la diversité spatiale. En effet, en échantillonnant simplement toutes les communes en un seul bloc, le risque de sélectionner des échantillons trop similaires apparaît, surtout à cette résolution où les données sont souvent redondantes spatialement.

Échantillonnage pour l'ensemble d'entraînement

Travaillant avec plusieurs dizaines de millions d'observations à haute résolution, il est possible de tirer parti de la redondance spatiale pour réduire les coûts computationnels : un sous-échantillon de 2% des données d'entraînement est donc utilisé. Il est à préciser qu'un déséquilibre des classes est présent dans ces données.

Création d'un ensemble de test équilibré

Pour évaluer le modèle sans introduire de biais lié à un déséquilibre des classes, l'ensemble de test est équilibré en sélectionnant un nombre égal de cellules brûlées et non brûlées. Cette approche permet d'obtenir une évaluation plus claire des performances du modèle, notamment en termes de précision et de rappel pour chaque classe, sans que l'une des classes, souvent majoritaire, ne domine les résultats des métriques globales.

Normalisation

La normalisation des variables explicatives assure une contribution équilibrée de chaque variable dans le modèle, ce qui est particulièrement important dans un modèle linéaire généralisé : les coefficients peuvent ainsi refléter plus fidèlement les contributions relatives de chaque variable. Ce prétraitement réduit le risque qu'une variable à grande échelle influence de façon disproportionnée les résultats, et rend l'interprétation des coefficients plus cohérente et comparable.

Paramétrage

Le seuil de décision de 0,5 est un choix standard en classification binaire, car il représente un équilibre où la probabilité d'appartenance à la classe positive (ici, brûlée) est aussi élevée que celle d'appartenance à la classe négative (non brûlée). Ce seuil peut être ajusté selon les besoins.

En régression logistique, l'objectif est de trouver les coefficients qui minimisent l'erreur entre les prédictions et les valeurs réelles. Cette minimisation passe par un processus itératif d'ajustement des coefficients pour réduire l'écart entre les probabilités prédites et les observations. Le modèle utilise une méthode d'optimisation, comme la descente de gradient, pour ajuster les coefficients progressivement. À chaque itération, le modèle affine ses coefficients pour se rapprocher d'un point de stabilité, où les ajustements n'ont plus d'effet significatif. Dans ce modèle, un

nombre maximal d'itérations est fixé à 2000 pour empêcher le modèle de tourner indéfiniment si la convergence est lente.

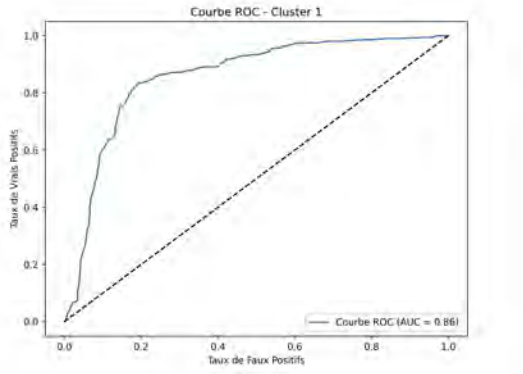
Sélection de variables

La réduction des caractéristiques est menée de manière itérative pour identifier et conserver la combinaison de variables qui maximise le F1-Score moyen des deux classes, 0 (non-brûlée) et 1 (brûlée). Le F1-Score moyen $F1_{\text{moyen}}$ se calcule en prenant la moyenne des F1-Scores des classes 0 et 1, où chaque F1-Score d'une classe c est défini comme suit :

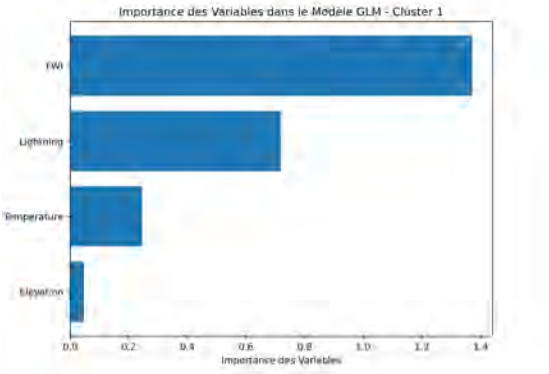
$$F1_{\text{moyen}} = \frac{1}{2}\left(2 \times \frac{\text{Précision}_0 \times \text{Rappel}_0}{\text{Précision}_0 + \text{Rappel}_0} + 2 \times \frac{\text{Précision}_1 \times \text{Rappel}_1}{\text{Précision}_1 + \text{Rappel}_1}\right)$$

où la précision et le rappel de chaque classe sont respectivement calculés comme le ratio des prédictions correctes parmi les prédictions positives et parmi les instances positives réelles.

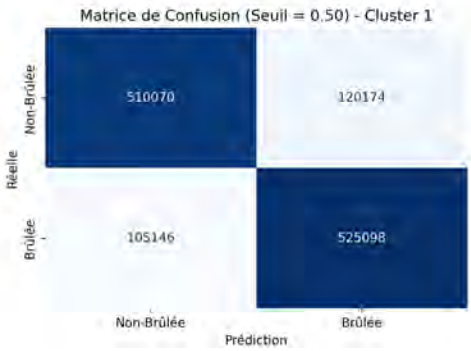
Lors de chaque itération de réduction, une des variables explicatives est temporairement retirée, puis le modèle est ajusté et son F1-Score moyen recalculé. Cette approche permet de comparer les impacts de chaque variable sur les performances globales du modèle et d'éliminer celles qui contribuent le moins à l'amélioration du F1-Score moyen, optimisant ainsi la robustesse et la pertinence des prédictions.



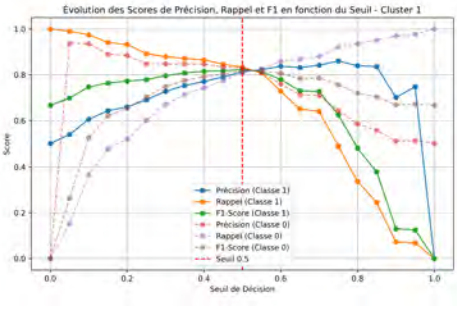
(a) Courbe ROC



(b) Importance des Variables



(c) Matrice de Confusion



(d) Évolution des Scores en fonction du Seuil

FIGURE 3.7 – Métriques du cluster 1

A l'issue de ce processus, les résultats obtenus sont intéressants pour le cluster 1 avec notamment un AUC de 0,86, ce qui indique une bonne classification. Cela se lit aussi sur la matrice de confusion où environ 82% des prédictions sont justes.

TABLE 3.2 – Résultats des modèles GLM pour chaque cluster

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**
ROC AUC	0.8977	0.8560	0.8468	0.7151
F1 Score (Classe 1)	0.8771	0.8233	0.8289	0.6873
F1 Score (Classe 0)	0.8535	0.8190	0.7753	0.6498
Précision (Classe 1)	0.8116	0.8136	0.7406	0.6524
Précision (Classe 0)	0.9445	0.8291	0.9191	0.6912
Rappel (Classe 1)	0.9542	0.8332	0.9410	0.7261
Rappel (Classe 0)	0.7785	0.8091	0.6705	0.6131
Accuracy	0.8664	0.8211	0.8057	0.6696

Les scores F1, précision et rappel sont élevés pour les clusters 0, 1 et 2, montrant une performance robuste dans la détection des feux dans ces clusters. Le cluster 3 présente des scores inférieurs, notamment pour le ROC AUC (0.7151), ce qui indique des difficultés pour le modèle à bien séparer les classes pour ce cluster.

TABLE 3.3 – Matrice de confusion pour chaque cluster

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**
Classe 0 (TN)	223235	509929	953801	3200778
Classe 0 (FP)	63519	120315	468744	2020021
Classe 1 (FN)	13128	105146	83968	1429772
Classe 1 (TP)	273626	525098	1338577	3791027

La proportion relativement élevée de faux positifs (FP) dans le cluster 3 pourrait être due à la variabilité des conditions dans ce cluster. Les communes de ce cluster ont en effet été sévèrement touchées par des incendies de grande envergure mais plus rares.

TABLE 3.4 – Importance des variables pour chaque cluster

Cluster	**Variables et Importances**
Cluster 0	**Variables :** TreeCover, DistanceToUrban, Elevation, FWI, NoVegetation **Importances :** 0.0883, 0.5835, 1.4337, 1.7826, 1.8211
Cluster 1	**Variables :** Elevation, Temperature, Lightning, FWI **Importances :** 0.0469, 0.2461, 0.7190, 1.3722
Cluster 2	**Variables :** LitPop, GHI, Lightning, Herbaceous, Shrublands, Temperature **Importances :** 0.0242, 0.1825, 0.2996, 0.3548, 0.5947, 0.7653
Cluster 3	**Variables :** LitPop, Lightning, Slope, GHI, TreeCover, Herbaceous, Shrublands **Importances :** 0.1152, 0.2924, 0.3646, 0.6497, 1.1252, 1.1354, 1.4850

Dans le cluster 0, NoVegetation et FWI sont les variables les plus importantes, ce qui s'explique par le fait qu'il y ait beaucoup de zones sans végétation dans les communes plus montagneuses et austères de ce groupe. La variable Lightning est influente dans les clusters 1, 2 et

3, indiquant son rôle potentiel dans la déclenchement de feux et ce en dépit de la qualité de résolution de cette donnée qui est bien plus faible que toutes celles des autres variables. Le rôle de Shrublands dans le cluster 3 avec une importance élevée pourrait indiquer des risques accrus d'incendie associés aux terres contenant des arbustes.

Carte des probabilités

Comme le modèle est entraîné et calibré, il convient de l'utiliser pour prédire les probabilités de feu pour chaque cellule de 30 m^2 de chacun des quatre clusters établis dans les régions centrales du Chili, permettant ainsi de construire une carte prédictive des zones à risque.

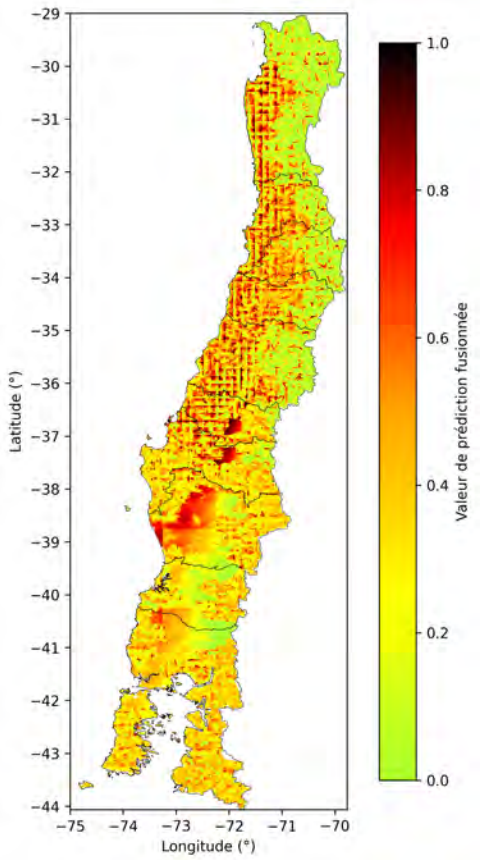


FIGURE 3.8 – Carte de fréquence consolidée du modèle GLM

Discussion

La carte générée à partir de l'algorithme de random forest présente une hétérogénéité de modélisation marquée entre les différents clusters. Cette variabilité découle des différences de résolution des données retenues par le Random Forest lors des ajustements de paramètres au cours du processus de sélection des caractéristiques les plus pertinentes pour chaque cluster de manière indépendante. En particulier, ce sont les variables de couverture végétale qui semblent ajouter cette granularité. Cela conduit à une carte finale où chaque cluster reflète des sensibilités et des critères distincts. Cette approche, bien qu'optimale localement, complexifie l'interprétation globale de la carte, car elle ne repose pas sur des bases homogènes pour l'ensemble des clusters. En conséquence, l'hétérogénéité des clusters rend cette carte difficile à justifier comme

une représentation cohérente et généralisable des risques, limitant ainsi son applicabilité pour des analyses ou des prédictions à plus grande échelle.

3.3 Méthode de Random Forest

3.3.1 Théorie

Les méthodes d'ensemble jouent un rôle central dans les techniques modernes de Machine Learning, grâce à leur capacité à améliorer les performances des modèles en combinant plusieurs apprenants faibles. Le modèle *Random Forest*, introduit par Breiman en 2001, est l'une des méthodes d'ensemble les plus utilisées, tant pour les tâches de classification que de régression.

Les méthodes d'ensemble

Le principe des méthodes d'ensemble repose sur l'idée qu'en combinant plusieurs modèles, appelés *apprenants*, on peut obtenir un modèle global plus performant que les modèles individuels. Mathématiquement, un ensemble peut être exprimé comme une somme pondérée des prédictions des apprenants individuels :

$$\hat{y} = \sum_{i=1}^{N} w_i h_i(x)$$

où $h_i(x)$ est la prédiction du i-ème modèle pour une observation x, et w_i est le poids associé à ce modèle. Dans le Random Forest, tous les poids w_i sont égaux, chaque arbre de décision contribuant de façon égale à la prédiction finale. Cela permet de capturer différentes perspectives sur les données, réduisant les erreurs induites par la variance des apprenants individuels.

Deux approches principales sont utilisées :

— **Bagging** (*Bootstrap Aggregating*) : Chaque modèle est entraîné indépendamment sur des échantillons bootstrap des données. Les prédictions sont ensuite agrégées par vote majoritaire (classification) ou par moyenne (régression). Le Random Forest est un exemple de bagging appliqué aux arbres de décision.

— **Boosting** : Contrairement au bagging, le boosting entraîne les modèles séquentiellement, en donnant plus de poids aux erreurs des modèles précédents, réduisant ainsi le biais et la variance. Toutefois, cette approche est plus sensible au bruit.

La qualité des séparations est mesurée par des critères comme l'entropie et l'indice de Gini :

— **Entropie** : Mesure l'homogénéité des classes. Elle est minimale si toutes les instances appartiennent à la même classe et maximale si les classes sont uniformément réparties :

$$H(t) = -\sum_{i=1}^{K} p_i \log_2(p_i)$$

où p_i est la proportion des instances de la classe i.

— **Gain d'information** : Mesure la réduction de l'entropie après division d'un nœud :

$$\Delta H(X) = H(t) - \sum_{j} \frac{|t_j|}{|t|} H(t_j)$$

— **Indice de Gini** : Mesure la pureté d'un nœud en quantifiant la probabilité que deux instances choisies au hasard appartiennent à des classes différentes :

$$G(t) = 1 - \sum_{i=1}^{K} p_i^2$$

Avantages et limites

Le Random Forest a de nombreux atouts, notamment sa robustesse aux données bruitées et manquantes, sa capacité à évaluer l'importance des caractéristiques et sa résistance au sur-apprentissage. Grâce à l'agrégation et à l'échantillonnage aléatoire, il gère efficacement la variation des données et réduit les risques de sur-ajustement. Il peut traiter des données de grande dimension et diverses (catégorielles, continues ou mixtes), sans nécessiter de normalisation, tout en permettant une sélection des variables grâce à l'analyse de l'importance des caractéristiques. Ce modèle présente toutefois quelques inconvénients. D'une part, le coût computationnel est proportionnel au nombre d'arbres, chaque arbre nécessitant un temps de calcul significatif, ce qui peut rendre l'algorithme peu adapté aux applications en temps réel ou aux grands ensembles de données. Par ailleurs en termes de complexité, bien que le Random Forest permette d'obtenir des performances élevées, il reste difficile à interpréter du fait de La combinaison de nombreux arbres de décisions individuelles. Enfin, en présence de classes déséquilibrées, le modèle peut favoriser la classe majoritaire. Des ajustements, comme le rééchantillonnage ou la pondération des classes, sont parfois nécessaires pour améliorer les performances sur les classes minoritaires.

Les hyperparamètres

L'optimisation des hyperparamètres permet d'améliorer les performances du modèle en ajustant certains paramètres clés, par exemple :

— **n_estimators** : Ce paramètre représente le nombre d'arbres de décision construits dans la forêt. Un nombre plus élevé de n_estimators tend à réduire la variance du modèle et à améliorer sa stabilité, mais cela augmente également le temps de calcul.

— **max_depth** : Ce paramètre définit la profondeur maximale des arbres. Limiter la profondeur permet de contrôler la complexité du modèle et d'éviter le surapprentissage qui survient lorsque le modèle est trop adapté aux données d'entraînement, mais généralise mal aux nouvelles données.

— **min_samples_leaf** : Ce paramètre spécifie le nombre minimal d'échantillons requis pour créer une feuille dans un arbre. Une valeur plus élevée de min_samples_leaf réduit la variance des arbres individuels en forçant chaque feuille à contenir suffisamment de données, ce qui peut améliorer la robustesse du modèle.

L'optimisation de ces hyperparamètres se fait généralement en explorant différentes combinaisons afin de trouver celles qui maximisent une métrique de performance spécifique, comme le score F1 ou l'accuracy. Cette recherche d'optimisation permet d'équilibrer biais et variance dans le modèle, aboutissant ainsi à une forêt aléatoire qui généralise efficacement.

Comparaison avec les GLM

Aspect	Random Forest	GLM
Gestion du déséquilibre des classes	Gère bien le déséquilibre grâce à l'agrégation des arbres issus d'échantillons variés et à la méthode de bootstrap, ce qui réduit l'influence de la classe majoritaire	Souvent nécessaire d'ajuster manuellement pour éviter que la classe majoritaire domine, en pondérant davantage la classe minoritaire
Performance en prédiction	Haute performance, surtout pour les données complexes avec interactions non linéaires entre variables	Bonne performance pour des relations linéaires simples mais moins efficace pour les données avec interactions complexes
Interprétabilité	Moins interprétable en raison de la complexité des nombreux arbres ; cependant, l'importance des variables peut être visualisée	Très interprétable, chaque coefficient représente l'impact d'une variable sur la probabilité de la classe positive
Temps de calcul	Temps de calcul plus élevé dû à la création de multiples arbres, surtout pour de grands ensembles de données	Temps de calcul souvent plus faible, adapté aux ensembles de données de taille moyenne à grande
Résistance au surapprentissage	Résistant au surapprentissage grâce à l'agrégation des prédictions de chaque arbre et à la diversité des échantillons	Plus sensible au surapprentissage, particulièrement avec de nombreuses variables ou en présence de classes déséquilibrées
Précision des résultats	Très précis pour les données de grande dimension avec des interactions complexes	Moins précis lorsque les interactions entre variables sont complexes ou non linéaires

Table 3.5 – Comparaison entre Random Forest et GLM

3.3.2 Mise en œuvre du Random Forest

L'approche qui a été mise en œuvre dans un premier temps ici était similaire à celle utilisée pour la calibration du modèle GLM, à la différence qu'en plus de ne pas avoir besoin de centrer et réduire les variables, le modèle Random Forest ne compensait pas le déséquilibre des classes. En effet, les forêts aléatoires gèrent généralement mieux ce déséquilibre que les GLM. En contrepartie, les seuils de prédiction de zone brûlée nécessitaient un ajustement.

Toutefois, étant donné que les résultats étaient similaires à ceux du GLM en ce qui concerne la faible lisibilité des cartes avec cette manière de procéder, une autre voie a été empruntée. En se focalisant par région et en retirant dès le début les couvertures de sol qui ne sont pas de la végétation, la modélisation du random forest a été effectuée à l'échelle d'une région. Les zones d'incendie de la Araucania ont ainsi été prédites à partir de l'entraînement d'un modèle sur les neuf autres régions. Le choix de concentrer l'étude sur cette région provient du fait que cette région représente un juste milieu intéressant parmi les régions chiliennes pour le risque feu de forêt. En effet, elle est souvent touchée par des feux de forêt mais moins que les régions plus centrales. Il peut donc être intéressant pour un assureur d'y mesurer le risque.

Zone tampon

Une méthode parfois employée pour permettre à un modèle de classification de mieux discerner les classes est l'implémentation d'une zone tampon (ou buffer zone). Par exemple l'étude de Lingxiao Xie et al.[14] publiée en 2022 dans le journal Remote Sensing utilise un tel seuil dans la calibration d'un random forest destiné à la prédiction de feu de forêt court terme dans la préfecture autonome yi de Liangshan. Un seuil de 1 km est ainsi calibré dans les échantillons

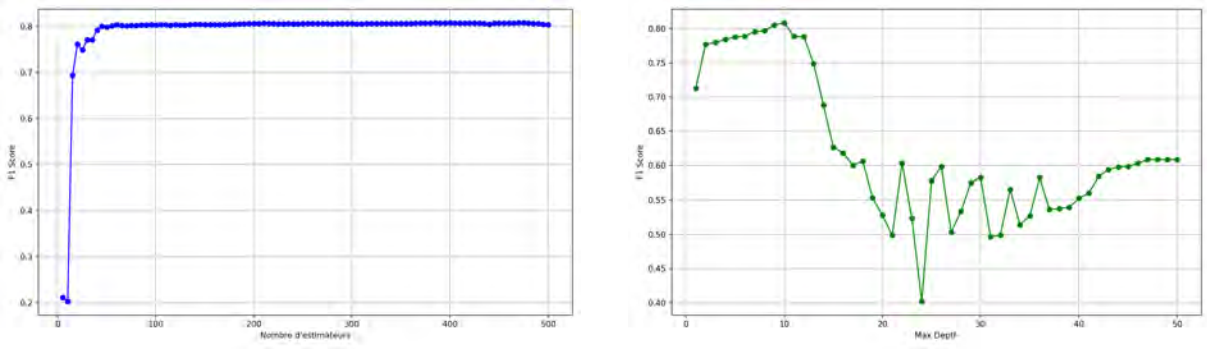
d'entraînement et de test à l'aide de la variable DistanceToFire.

Optimisation du modèle

Afin de ne garder que les variables qui améliorent le modèle, celles les moins influentes ont été successivement retirées en observant l'évolution du score F1 après chaque retrait. La meilleur combinaison de variables est composée de l'irradiance (GHI), des précipitations et de l'altitude. Il est à noter qu'au vu de la quantité importante de données, il devient vite long de chercher

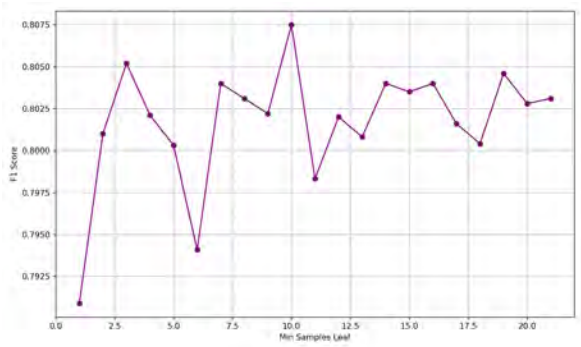
TABLE 3.6 – Résumé des variables retirées à chaque étape et du score F1 moyen obtenu

Étape	**Variable retirée**	**Score F1 après retrait**
1	FWI	0.6899
2	Population Density	0.7286
3	Lightning	0.7061
4	Temperature	0.7474
5	Slope	0.7651
6	Land Cover	0.7788
7	LitPop	0.7768
8	Distance to Urban	0.7859
9	GHI	0.7691
10	Precipitation	0.7400
-	**Elevation**	-



(a) F1 Score en fonction de n_estimators (b) F1 Score en fonction de max_depth

FIGURE 3.9 – Évolution du F1 Score en fonction des paramètres n_estimators et max_depth



(a) F1 Score en fonction de min_samples_leaf

FIGURE 3.10 – Évolution du F1 Score en fonction du paramètre min_samples_leaf

Les hyperparamètres retenus pour la forêt aléatoire sont les suivants :

— `n_estimators` = 100

— `max_depth` = 10

— `min_samples_leaf` = 10

Résultats

Le meilleur score obtenu présente une AUC de 0.8299 et un score F1 moyen de 0.8075, ce qui est plutôt correct. Après application du random forest calibré à la Araucanía, les résultats sans ajustement de la fréquence mettent en relief les zones les plus à risque.

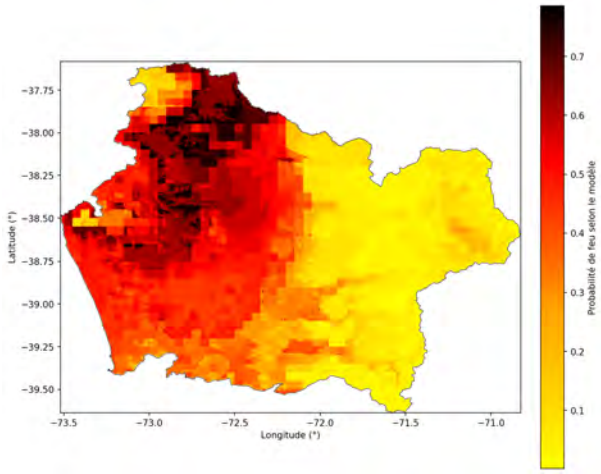


FIGURE 3.11 – Carte de probabilité de feu de la Araucanía

Calibrage de la fréquence

La méthode courante pour estimer les pertes en hectares d'une zone consiste à les comparer à une moyenne historique. En prenant les cinq dernières années de sinistralité pour la région de la Araucanía, mesurées à l'aide de la base de données Firescar, le calibrage de la fréquence est effectué en utilisant la moyenne annuelle des hectares brûlés au cours de cette période comme référence pour la carte des fréquences.

Ainsi, la fréquence est ajustée de manière à ce que la somme des superficies brûlées par parcelle corresponde à cette moyenne annuelle d'hectares brûlés. Étant donné que 1,0751% de la région a été affectée par des incendies au cours des cinq dernières années, on considère que 0,215% de la superficie de la région brûle chaque année.

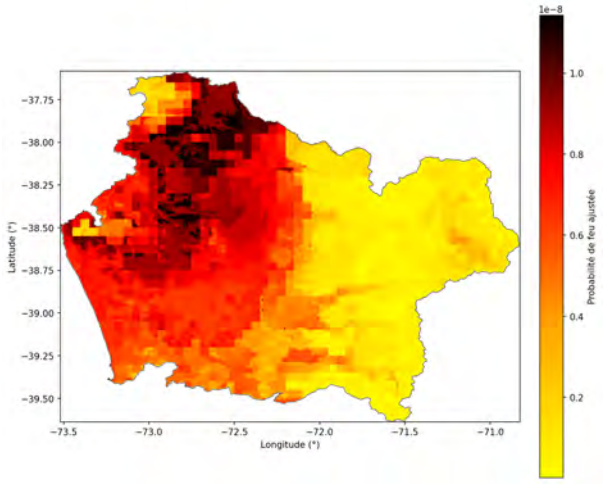


FIGURE 3.12 – Carte de probabilité de feu ajustée de la Araucanía

3.4 Calcul de prime

La suite logique de la modélisation entreprise serait une considération tarifaire. En assurance, la *prime pure* est définie comme le coût théorique attendu des sinistres pour un risque donné, sans inclure les frais administratifs, les taxes ou les marges de profit. Elle représente le montant moyen attendu des indemnisations que l'assureur devra payer pour couvrir le risque de sinistre. Mathématiquement, la prime pure pour un risque unique est donnée par l'espérance des sinistres S, soit :

— S est le coût total des sinistres pour une période donnée.

— N est la *fréquence des sinistres*, c'est-à-dire le nombre de sinistres sur cette période.

— X est la *gravité du sinistre*, c'est-à-dire le coût moyen par sinistre.

Cette formule repose sur la décomposition du coût des sinistres en fréquence et gravité, ce qui permet de modéliser séparément l'occurrence des sinistres et leur impact financier.

Prime pure par parcelle

Il est possible d'appliquer cela dans le cadre d'une assurance feu de forêt, en estimant d'une part la fréquence de brûlage d'une parcelle et, d'autre part, la valeur des biens susceptibles d'être détruits. Pour des parcelles forestières de 30 m^2 au Chili, chaque parcelle est considérée comme un risque individuel. L'assuré déclare la valeur de ses biens pour chaque parcelle, ce qui permet d'estimer le coût des dommages si la parcelle brûle complètement. Il est à disposition également une estimation de la fréquence annuelle à laquelle chaque parcelle brûle, obtenue par des analyses de données historiques d'incendies.

Pour une parcelle i, on définit :

— f_i est la *fréquence d'occurrence* annuelle d'un incendie pour cette parcelle, soit la probabilité que la parcelle brûle en une année.

— c_i est le *coût des dommages* si la parcelle brûle entièrement, basé sur la valeur des biens déclarée par l'assuré (plantations, infrastructure, etc.).

En supposant que la parcelle brûle dans sa totalité lorsqu'un incendie se produit, la prime pure pour cette parcelle i est donnée par :

$$\text{Prime Pure}_i = f_i \times c_i$$

Cette valeur représente le coût attendu des sinistres pour cette parcelle sur une année, en fonction de la probabilité d'occurrence de l'incendie et du coût total en cas de sinistre.

Prime totale

Pour calculer la prime pure totale pour une zone couvrant n parcelles de 30 m^2, il suffit de sommer les primes pures individuelles de chaque parcelle. Ainsi, la prime pure totale pour l'ensemble de la zone est donnée par :

$$\text{Prime Pure}_{\text{totale}} = \sum_{i=1}^{n} f_i \times c_i$$

Cette somme représente le coût attendu annuel des sinistres pour l'ensemble des parcelles assurées dans cette zone. En d'autres termes, elle fournit une estimation théorique des pertes dues aux incendies, sans ajustements pour les autres frais ou marges de l'assureur.

Conclusion

La très grande diversité des caractéristiques géographiques au sein des régions ainsi que l'étendue spatiale du pays soulèvent des problématiques qui poussent à incorporer plus de complexité dans la manière d'aborder et de traiter le sujet. Cette complexité se matérialise notamment par le besoin d'acquérir davantage de données de sources différentes. Ce mémoire s'est proposé d'explorer les facteurs et les dynamiques associés aux incendies de forêt dans les différentes zones du Chili à une maille fine. La première étape a consisté à constituer une base des événements de feux historiques puis un ensemble de donneés pouvant expliquer ce risque.

Une analyse par composantes principales suivie d'une classifications par la méthode des k-means a permi de regrouper les communes avec des propriétés similaires. Par suite, un modèle prédictif basé sur un GLM é été calibré. Une cartographie des risques de feu de forêt a pu ainsi être établie. L'ambition d'unifier les régions centrales du Chili sous une même modélisation homogène pour les feux de forêt s'est heurté à de nombreuses considérations géographiques et techniques. L'impact des divergences de résolutions des données sources sur l'interprètabilité des résultats s'est révélé important.

L'étape suivante d'implémentation d'un random forest spécifiquement dans la région de la Araucania a montré que finalement, un modèle avec trois variables robustes peut être plus efficace localement qu'un modèle tenant compte de l'ensemble des données à disposition. Ce compromis entre la généralité et le spécifique montre qu'une modélisation trop générale ne saurait capturer pleinement la complexité des risques.

Une autre manière à explorer pourrait être l'incorporation d'un modèle par propagation du feu, qui simule le comportement spatial et temporel des incendies à partir de paramètres comme le vent, la végétation, et la topographie. Ce type de modèle permettrait de mieux comprendre les zones à risque élevé, mais aussi la manière dont un feu pourrait se propager à travers différentes régions.

Table des figures

1 Approche adoptée pour la modélisation du risque feu de forêt 4
2 Carte des communes colorées par cluster. 4
3 Carte de fréquence de feu de la Araucania 5
4 Carte de fréquence consolidée du modèle GLM 6
5 Approach for Forest Fire Risk Modeling 8
6 Map of communes color-coded by cluster. 8
7 Fire frequency map for the Araucania region 9
8 Consolidated frequency map from the GLM model 10

1.1 Le triangle du feu 16
1.2 Origine des incendies au Chili entre 1985 et 2018 21
1.3 Origine des incendies au Chili entre 1985 et 2018 entre les régions de Valparaiso et de la Araucanía 22
1.4 Comparaison entre El Niño, La Niña et les conditions normales 23
1.5 Réponse spectrale de la végétation : comparaison entre zones brûlées et non brûlées 26
1.6 Table des seuils de dNBR[5] 27
1.7 Approche adoptée pour la modélisation du risque feu de forêt 29

2.1 Illustration de la variation de longueur d'un degré de longitude à deux latitudes différentes [12] 32
2.2 Chili et périmètre sélectionné 33
2.3 Feux enregistrés par la CONAF entre les saisons de feux 1985 et 2018 34
2.4 Carte de température 37
2.5 Carte des précipitations 37
2.6 Construction de l'indice forêt météo 38
2.7 Carte d'indice feu météo 39
2.8 Carte d'irradiance 39
2.9 Carte mondiale d'impact de foudre 41
2.10 Carte d'impact de foudre 41
2.11 Carte d'altitude 42
2.12 Carte de pente 42
2.13 Classification du terrain 44
2.14 Carte de densité de population 45
2.15 Carte de LitPop 45
2.16 Carte de distance à un point urbanisé 46
2.17 Carte de distance à un point de feu passé 46
2.18 Matrice de corrélation globale 48
2.19 Cartes des variables moyennes par commune dans les régions centrales du Chili . 49
2.20 Matrice de corrélation des variables moyennes par commune 50

2.21 Cartes des variables moyennes par province dans les régions centrales du Chili . . 51
2.22 Matrice de corrélation des variables moyennes par province 52
2.23 Cartes des variables moyennes par région du Chili central 53
2.24 Matrice de corrélation des variables moyennes par région 54

3.1 Distribution des données standardisées par variable. 59
3.2 Variance expliquée cumulée par nombre de composantes principales. 60
3.3 Variance expliquée par chaque composante principale. 60
3.4 Méthode du coude pour déterminer le nombre optimal de clusters. 60
3.5 Projection sur les deux premières composantes principales avec clusters. 61
3.6 Carte des communes colorées par cluster. 61
3.7 Métriques du cluster 1 68
3.8 Carte de fréquence consolidée du modèle GLM 70
3.9 Évolution du F1 Score en fonction des paramètres n_estimators et max_depth . 74
3.10 Évolution du F1 Score en fonction du paramètre min_samples_leaf 75
3.11 Carte de probabilité de feu de la Araucanía 75
3.12 Carte de probabilité de feu ajustée de la Araucanía 76
13 Matrice de corrélation pour la Región de La Araucanía 82

Annexe

TABLE 7 – Statistiques descriptives pour la Región de La Araucanía

	mean	std	min	25%	50%	75%	max
Fire	0.030	0.180	0.00	0.00	0.00	0.00	1.00
Slope	9.75	8.75	0.00	2.78	7.05	14.36	78.23
Elevation	596.45	523.43	-36	194.00	386.00	958.00	3087.00
Precipitation	148.77	49.19	70.48	105.05	140.04	185.38	333.35
Temperature	10.65	2.09	3.52	9.13	11.55	12.25	13.71
LitPop	1.60e+05	1.98e+06	0.00	2273.33	4921.81	1.06e+04	4.88e+07
DistanceToUrban	4.36e+03	3.65e+03	0.00	1597.06	3226.92	6211.81	2.14e+04
PopDensity	36.35	262.55	0	2.65	5.73	12.90	7310.42
GHI	4.39	0.23	2.65	4.24	4.36	4.53	5.06
Lightning	0.00	0.00	3.65e-04	9.06e-04	1.53e-03	2.40e-03	4.37e-03
FWI	6.53	2.43	2.06	4.55	6.47	8.30	16.11
TreeCover	0.490	0.500	0.00	0.00	0.00	1.00	1.00
Shrublands	0.140	0.340	0.00	0.00	0.00	0.00	1.00
Herbaceous	0.340	0.470	0.00	0.00	0.00	1.00	1.00
NoVegetation	0.030	0.170	0.00	0.00	0.00	0.00	1.00

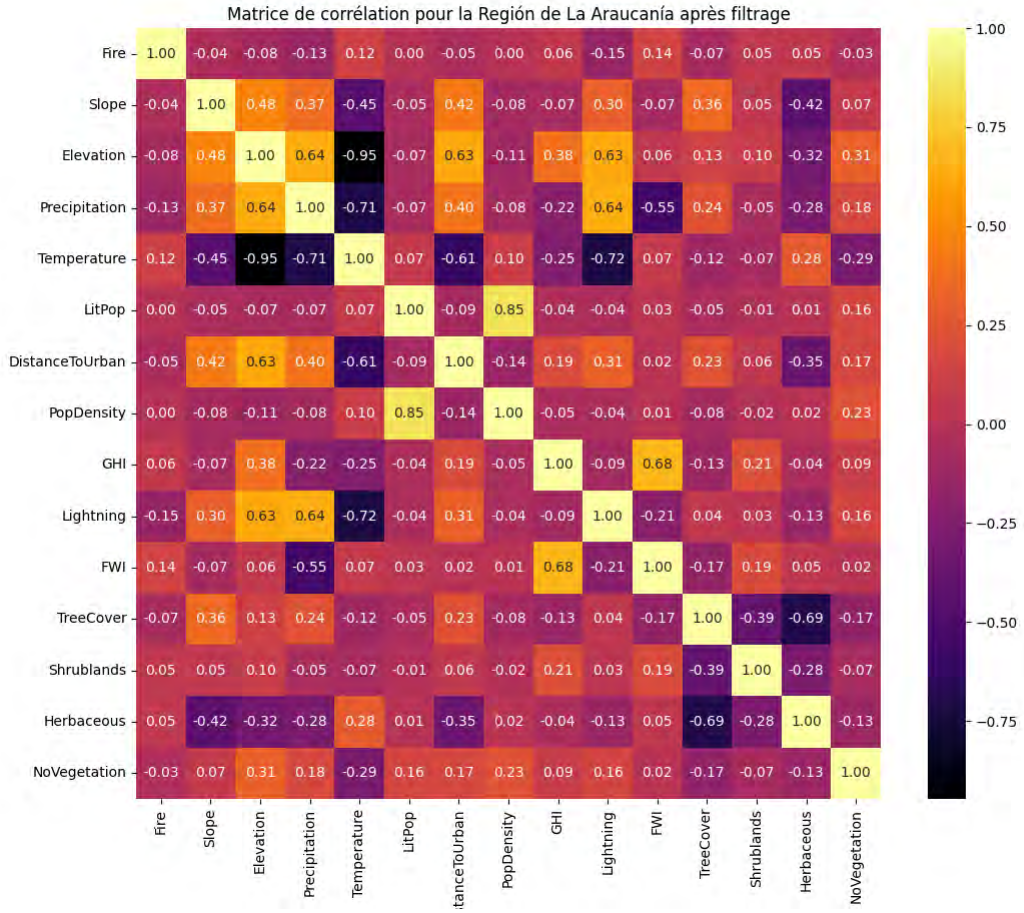


FIGURE 13 – Matrice de corrélation pour la Región de La Araucanía

Bibliographie

[1] Global Solar Atlas. *Global Solar Atlas - Irradiation solaire au Chili*. 2023. URL : https://globalsolaratlas.info/download/chile.

[2] ChileAtiende. *Subvention pour l'assurance forestière au Chili*. 2024. URL : https://www.chileatiende.gob.cl/fichas/61691-subsidio-al-seguro-forestal.

[3] Plataforma de Datos. *Données de cicatrices de feu - FireScar*. 2023. URL : https://www.plataformadedatos.cl/datasets/es/492F65B350AAF9D5.

[4] Institut National de l'Information Géographique et Forestière (IGN). *Les incendies de forêt et de végétation en France*. 2023. URL : https://foret.ign.fr/themes/les-incendies-de-foret-et-de-vegetation.

[5] Keane R.E. Caratti J.F. Key C.H. Benson N.C. Sutherland S. Gangi L.J. Lutes D.C. *FIREMON : Fire Effects Monitoring and Inventory System ; Tech. Rep. RMRS-GTR-164-CD*. 2006. URL : https://www.fs.usda.gov/rmrs/publications/firemon-fire-effects-monitoring-and-inventory-system.

[6] Mentler R. Moletto-Lobos Í. Alfaro G. Aliaga L. Balbontín D. Barraza M. Baumbach S. Calderón P. Cárdenas F. Castillo I. Contreras G. de la Barra F. Galleguillos M. González M. E. Hormazábal C. Lara A. Mancilla I. Muñoz F. Oyarce C. Pantoja F. Ramírez R. Miranda A. et V. Urrutia. *The Landscape Fire Scars Database : mapping historical burned area and fire severity in Chile, Earth Syst. Sci. Data, 14, 3599–3613*. 2022. URL : https://essd.copernicus.org/articles/14/3599/2022/.

[7] NASA. *Activité des éclairs*. 2023. URL : https://ghrc.nsstc.nasa.gov/lightning/.

[8] NASA. *Données GLanCE*. 2023. URL : https://lpdaac.usgs.gov/products/glance30v001/.

[9] NASA. *Données topographiques STRM*. 2023. URL : https://portal.opentopography.org/raster?opentopoID=OTSRTM.082015.4326.1.

[10] OCHA. *Régions administratives du Chili*. 2023. URL : https://data.humdata.org/dataset/cod-ab-chl.

[11] Centro de Ciencia del Clima y la Resiliencia (CR2). *Températures et précipitations au Chili*. 2023. URL : https://www.cr2.cl/datos-productos-grillados/.

[12] HCL Technologies. *Systèmes spatiaux - HCL Technologies*. 2023. URL : https://help.hcltechsw.com/onedb/1.0.1/spat/ids_spat_407.html.

[13] WorldPop. *Densité de population au Chili - WorldPop*. 2023. URL : https://data.humdata.org/dataset/worldpop-population-density-for-chile.

[14] Lingxiao Xie et al. *Wildfire Risk Assessment in Liangshan Prefecture, China Based on An Integration Machine Learning Algorithm*. Publié dans Remote Sensing, vol. 16, n° 1, pp. 45-60. 2024. URL : https://doi.org/10.3390/rs16010045.

[15] ETH Zurich. *Litpop - Données de population lumineuse*. 2023. URL : https://www.research-collection.ethz.ch/handle/20.500.11850/331316.