Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 18 Septembre 2024

Par : Yapi Joseph Armel KOUASSI
Titre : Exploitation d'images de cartes dans la modélisation de la sinistralité en assurance habitation

Confidentialité : Non

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

Membre présent du jury de l'Institut des Actuaires :
Estelle De L'EPREVIER
Romain LAILY
Samuel STOCKSIECKER
Yufei LUO
Signature :

Entreprise :
Cardif IARD
Signature :

Membres présents du jury de l'EURIA :
Franck VERMET

Directeur de mémoire en entreprise :
Nicolas ZHANG
Signature :

Invité :

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 :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

L'évaluation du risque est un sujet central pour chaque compagnie d'assurance, notamment pour la tarification et la politique de sélection des risques. L'efficacité de la stratégie commerciale et la rentabilité d'un assureur dépendent fortement de sa capacité à évaluer précisément chaque risque assuré. Cette évaluation repose sur des modèles statistiques quantifiant pour chaque variable descriptive du risque son impact sur la sinistralité prévisionnelle.

Bien que les modèles linéaires généralisés restent, encore aujourd'hui, prépondérants en assurance non-vie pour réaliser cette modélisation, les dernières années ont vu l'émergence de nouveaux modèles basés sur l'apprentissage statistique pour améliorer les résultats. Cependant, la performance de ces algorithmes dépend de la disponibilité et de la qualité d'une quantité massive de données. Ainsi, la donnée apparaît comme un point stratégique central, aussi bien dans sa collecte que dans son stockage, son nettoyage et son utilisation. Toutefois, une importante part de ces données est disponible sous forme non structurée (image, audio ou vidéo), rendant leur exploitation par les algorithmes usuels de modélisation de la sinistralité impossible ou difficilement réalisable. Parmi elles, on trouve des cartes contenant des données géographiques (une carte des zones inondables en France, par exemple), mais dont la base de données d'origine n'est pas aisément accessible.

L'objectif des travaux de ce mémoire est de développer un outil algorithmique capable d'extraire des variables géographiques à partir de cartes légendées au format image. Ces variables pourront ensuite être exploitées pour enrichir la modélisation des risques. Après son développement, cet outil a notamment été mis à contribution dans le cadre de l'évaluation du risque de sécheresse et a permis la création d'un zonier simplifié.

Mots clefs: Traitement d'images, Apprentissage statistique, GLM, Assurance habitation, Sinistres, Zonier.

Abstract

Risk assessment is a central issue for every insurance company, particularly when it comes to pricing and risk selection policies. The effectiveness of an insurer's commercial strategy and profitability depend heavily on its ability to accurately assess each insured risk. This assessment is based on statistical models that quantify the impact of each risk descriptor on forecast claims experience.

Although generalised linear models still dominate non-life insurance modeling, recent years have seen the emergence of new models based on statistical learning to improve results. However, the performance of these algorithms depends on the availability and quality of a massive quantity of data. As a result, data is a key strategic issue, not only in terms of its collection, but also its storage, cleaning and use. However, a large proportion of this data is available in unstructured form (image, audio or video), making it impossible or difficult to use with the usual loss modelling algorithms. These include maps containing geographical data (a map of flood zones in France, for example), but for which the original database is not easily accessible.

The aim of this work is to develop an algorithmic tool capable of extracting geographical variables from maps with captions in image format. These variables can then be used to enhance risk modelling. Once it had been developed, the tool was used to assess the risk of drought, and was used to create a simplified zoning system.

Keywords: Image processing, Statistical learning, GLM, Home insurance, Claims, Zoning system.

Note de synthèse

Contexte et objectifs

Dans le secteur de l'assurance, l'accès à des données de qualité est devenu un enjeu stratégique majeur. Avec l'essor des technologies de l'information et l'augmentation des capacités de stockage et de calcul, les assureurs sont engagés dans une véritable course aux données. Ces données, qu'elles proviennent de sources internes ou externes, sont essentielles pour affiner les modèles de risque, améliorer la tarification, optimiser la gestion des sinistres, et anticiper les évolutions du marché.
L'acquisition de données géographiques, en particulier, est devenue cruciale pour la modélisation de nombreux types de risques, allant des catastrophes naturelles aux comportements des assurés dans des zones spécifiques, encore plus aujourd'hui dans le contexte de changement climatique. C'est dans ce contexte que s'inscrit ce mémoire, qui vise à développer un outil algorithmique capable d'extraire des variables géographiques à partir d'images de cartes de la France, afin de contribuer à l'enrichissement des bases de données actuarielles.
L'objectif principal de ce travail est donc de concevoir un outil algorithmique permettant d'exploiter de nouvelles sources de données géographiques pour améliorer la modélisation des risques en assurance. Il doit être capable de prendre en entrée une carte de la France légendée, au format image, et de sortir une base de données qui à la maille souhaitée (code postal ou code INSEE), associe la légende correspondante. En employant cet outil sur un cas pratique centré autour du risque sécheresse, ce mémoire se propose également d'évaluer les performances et l'utilité de l'outil dans différents contextes de modélisation.

Cadre juridique et légal

Avant de procéder à tout développement, la première étape a consisté à valider le cadre légal d'utilisation d'un tel outil auprès des services compétents. Il nous a été spécifiquement rappelé que :

— L'analyse doit être effectuée sur un volume de données suffisamment large pour garantir l'anonymat des personnes physiques, empêchant toute identification individuelle.
— L'utilisation des données doit se limiter à un usage interne, excluant toute exploitation à des fins de prospection commerciale.

— Les données reproduites ne doivent pas être altérées : lors de la conversion des échelles de couleurs en une échelle de 1 à 3, il est impératif de respecter les équivalences initiales et de préserver les niveaux originaux.
— Il est essentiel de respecter les conditions générales du site d'où proviennent les données et de s'assurer qu'aucune mention n'interdit leur reproduction ou réutilisation.

Tout au long de la réalisation des travaux de ce mémoire, une attention particulière a été portée à l'ensemble des images utilisées avec l'outil, afin de garantir le respect de la propriété intellectuelle et des droits d'auteur, tant en termes d'utilisation que de modification des contenus.

Conception de l'outil algorithmique

L'outil développé prend la forme d'une suite d'algorithmes visant à extraire méthodiquement l'information géographique contenue dans une carte de France légendée sous format d'image, en associant chaque commune à une couleur de légende. Cette méthode transforme des données visuelles en variables exploitables pour la modélisation des risques. La figure 3 résume les différentes étapes de fonctionnement de l'outil.

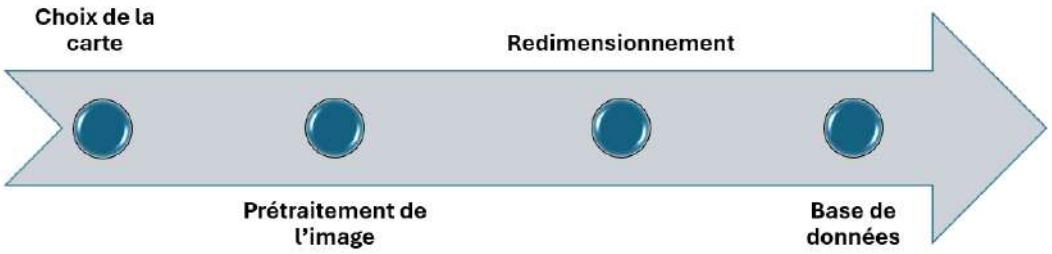


FIGURE 1 – Fonctionnement de l'outil algorithmique

Prétraitement des images

Le développement de l'outil commence par la sélection des cartes à traiter. Dans le cadre du mémoire, les cartes sélectionnées pour construire l'outil se devaient de fournir des informations potentiellement pertinentes pour la modélisation des sinistres en assurance habitation, avec une légende claire basée sur des couleurs permettant d'identifier les différentes zones.*(Mais rappelons que l'outil final a vocation à être universel et non pas relié au cas d'usage en assurance habitation.)* Elles doivent également couvrir toute la zone géographique d'intérêt *(dans ce mémoire, il s'agit de la France métropolitaine à l'exception de la Corse).*
Avant d'extraire les données, un prétraitement des images est réalisée pour analyser la composition colorimétrique de l'image. Cette étape cherche à classer les couleurs des pixels présents sur l'image selon les couleurs des pixels de la légende, à réduire le bruit

et à identifier les pixels dont la couleur ne figure pas dans la légende. Le bruit fait référence dans ce cas à tous les éléments pouvant altérer les couleurs de l'image ou rendre difficile la récupération des couleurs de certains pixels. Pour ce faire, on a recours à un clustering colorimétrique, qui consiste à segmenter les pixels de l'image en clusters de couleurs uniformes. Les couleurs des pixels sont analysées grâce au système RGB qui permet d'attribuer une valeur à une couleur. On utilise ensuite deux méthodes pour le clustering colorimétrique : les K-Means et l'Analyse en Composantes Principales (ACP). Les K-Means regroupent les pixels en clusters selon leurs couleurs, éliminant ainsi les dégradations et uniformisant les pixels similaires. L'ACP, quant à elle, projette les pixels dans un plan bidimensionnel, facilitant la détection de motifs et le regroupement de pixels similaires. Les pixels, dans ce contexte, ne sont définis que par leur couleur, et ceux présentant des couleurs similaires devraient se retrouver à proximité les uns des autres dans le plan principal de l'ACP. Ces étapes garantissent une meilleure qualité des données extraites.

Carte et base de référence

À présent, on construit une base de données de référence pour associer chaque commune française à une position en pixels sur une carte dite de référence. Cette étape s'inscrit dans l'idée générale de l'outil qui est d'avoir une carte sur laquelle les positions en termes de pixels sont connues pour tous les codes postaux français, puis de venir redimensionner la carte dont on veut extraire l'information afin qu'elle se superpose parfaitement à la carte de référence, les positions en pixels des communes seront alors connue.

Cette étape commence par la sélection de la carte de référence représentant la France métropolitaine (sans la Corse), qui doit respecter quelques critères. À partir de cette carte, les coordonnées en pixels de plusieurs villes sont recueillies manuellement. Ces données permettent ensuite de former un modèle de régression linéaire pour prédire les coordonnées des autres villes.

Deux approches de régression linéaire sont explorées pour modéliser ces coordonnées : l'une basée sur les latitudes et les longitudes, l'autre sur les coordonnées Lambert 93. Lambert 93 est le système de projections cartographiques officiel pour représenter la France. La comparaison des performances des deux approches montre que l'utilisation des coordonnées Lambert 93 offre une meilleure précision pour modéliser les coordonnées en pixels, comme l'indiquent les résultats des indicateurs RMSE et R^2. En plus des résultats quantitatifs, l'utilisation des coordonnées Lambert 93 permet d'accorder une certaine confiance au modèle en raison de la précision de la projection.

Redimensionnement

Cette étape consiste à ajuster les dimensions, l'inclinaison et la position des images pour qu'elles correspondent exactement à la carte de référence. Un masque de la carte de référence est construit et des transformations géométriques affines du plan (translation,

rotation, homothétie et réflexions) sont testées successivement jusqu'à aligner chaque image sur le masque. Ces transformations se font par calcul matriciel en multipliant les images sous Python et à l'aide de la théorie mathématique matricielle des transformations géométriques.
Pour converger vers une superposition parfaite, on a recours à des méthodes d'optimisation. Une fonction d'erreur est alors définie afin de mesurer la différence entre l'image traitée et le masque de référence. Cette fonction combine deux termes : l'un pénalise la présence de pixels non vides (la couleur du vide est définie pour regrouper toutes les couleurs présentes sur l'image, mais absente de la légende) à l'extérieur du masque, et l'autre pénalise la présence de pixels vides à l'intérieur du masque. Les méthodes d'optimisation de Powell et de Nelder-Mead, adaptées aux fonctions non différentiables, sont utilisées pour minimiser cette erreur. Bien que ces méthodes soient efficaces, elles ne garantissent pas toujours une superposition parfaite, laissant parfois des pixels vides à l'intérieur du masque. Pour corriger ces imperfections, l'algorithme des K plus proches voisins (K-NN) est appliqué, permettant de réattribuer la couleur des pixels vides en fonction des couleurs des pixels environnants.

Performance de l'outil

Quatre cartes ont été utilisées pour tester l'outil. Pour deux d'entre elles, les bases de données d'origine étaient disponibles, permettant d'évaluer les performances en termes de taux de classification des communes. Pour la carte A, l'ACP atteint un taux de classification de 89,63 %, tandis que K-Means atteint 81,47 %. Pour la carte B, l'ACP atteint 87,55 %, contre 85,71 % pour K-Means. Ces résultats montrent que l'ACP se révèle plus performante, surtout pour la carte A, où elle capture mieux les variations de couleur. Cependant, ces performances ne sont pas parfaites, et un biais subsiste dans les données récupérées, ce qui peut affecter la modélisation des sinistres. Concernant les deux autres cartes, en l'absence de bases de données d'origine, il n'est pas possible de calculer des taux de classification. Néanmoins, une comparaison visuelle entre les cartes originales et les cartes recréées avec l'outil montre que l'outil peut restituer assez fidèlement les informations des cartes, même si cette évaluation reste subjective.

FIGURE 2 – Exemple de carte[54] - originale vs recréée

Limites et améliorations

Bien que l'outil présente des performances encourageantes, plusieurs limites doivent être prises en compte. Premièrement, un indice de performance ne peut être calculé que pour les cartes dont les bases de données d'origine sont disponibles. Pour les autres cartes, l'évaluation reste subjective. Deuxièmement, la précision n'atteint pas 100 %, et des erreurs de classification subsistent. Enfin, la qualité des images joue un rôle crucial : des images de faible résolution ou fortement bruitées augmentent les erreurs de classification. Plusieurs pistes d'amélioration sont proposées, notamment l'intégration de techniques de machine learning avancées, telles que les réseaux de neurones convolutifs (CNN), ainsi que l'utilisation de filtres pour réduire le bruit. En abordant ces aspects, l'outil pourrait devenir encore plus performant et contribuer de manière significative à la modélisation des risques en assurance habitation. Il convient également de rappeler les questions éthiques et légales liées à l'utilisation des données géographiques, en respectant les droits d'auteur et les conditions d'utilisation des sources.

Cas d'usage : zonier simplifié pour la sécheresse

Dans le contexte de l'assurance habitation, Cardif IARD cherche à affiner son modèle de rentabilité technique, qui permet d'évaluer la rentabilité attendue des contrats d'assurance en utilisant diverses variables, y compris des données externes. Un modèle spécifique pour le péril sécheresse n'existait pas en raison de la réglementation fixant la prime Catastrophe Naturelle (CatNat) à l'avance. Toutefois, un tel modèle pourrait améliorer le suivi sur le risque sécheresse, en offrant une meilleure segmentation des zones à risque et une estimation plus précise des primes pures. Ce cas pratique vise à tester un modèle de prime pure espérée pour la sécheresse, en utilisant un outil d'extraction de données géographiques développé dans ce mémoire.
Le protocole suivi dans ce cas pratique comprend plusieurs étapes clés :

- Extraction des variables géographiques à partir de cartes pertinentes pour le risque sécheresse.
- Modélisation de la fréquence des sinistres à l'aide d'un GLM (modèle linéaire généralisé).
- Construction d'un zonier simplifié basé sur les résultats de la modélisation.
- Calcul et interprétation des pseudo-S/P pour les zones identifiées.

Extraction de variables géographiques

Cinq cartes ont été sélectionnées pour leur pertinence par rapport au risque de sécheresse : une carte des sols argileux (carte_rga), des cartes de chaleur et d'irradiation solaire, ainsi que des cartes de température moyenne et de précipitations. Ces cartes ont été traitées pour extraire les bases de données correspondantes pour être incluses dans le modèle. La méthode de traitement des images a inclus des opérations de clustering colorimétrique et de redimensionnement pour obtenir une base représentant fidèlement des cartes.

Modélisation de la fréquence

Le modèle 1C se concentre sur une seule variable explicative, la carte_rga, qui représente le risque de retrait-gonflement des argiles. Cette carte comporte quatre modalités : faible, sans info, moyenne, et forte. Les fréquences des sinistres par modalité montrent une tendance croissante avec le niveau de risque. L'analyse a montré que la fréquence des sinistres est nettement plus élevée dans les zones dans lesquelles le risque de retrait-gonflement est classé comme "forte", confirmant ainsi la pertinence de la carte_rga pour modéliser le risque de sécheresse. Le modèle a ensuite été évalué à l'aide de la courbe de Lorenz et du coefficient de Gini, avec un coefficient de 0,275, indiquant une segmentation modérée des zones de risque. Cette analyse a également permis de comparer la pseudo-prime pure liée à la sécheresse avec la prime moyenne CatNat, révélant des écarts significatifs, notamment dans les zones à haut risque.

Le passage à une modélisation plus complexe avec les modèles RI et RIT a permis d'intégrer plusieurs variables explicatives provenant de différentes cartes géographiques :

— Modèle RI : Ce modèle utilise deux variables : la carte_rga et la carte_irrad, qui mesure le niveau d'irradiation solaire.
— Modèle RIT : Ce modèle ajoute la carte_temp (température moyenne) aux variables du modèle RI.

Les *beta* des modèles, leur signification statistique et leur impact sur la fréquence des sinistres ont été analysés. Pour le modèle RIT, tous les coefficients significatifs montrent une logique cohérente avec la fréquence des sinistres. Par exemple, la carte_rga a des coefficients indiquant que les zones à "forte" exposition au risque ont une fréquence plus élevée de sinistres, tandis que la carte_temp montre que les zones avec des températures moyennes élevées sont plus susceptibles de subir des sinistres liés à la sécheresse.
Les métriques d'évaluations des modèles ont montré que le modèle RIT est légèrement supérieur au modèle RI, avec un coefficient de Gini de 0,327 contre 0,319 pour le modèle RI. Cette légère amélioration indique que l'ajout de la carte_temp a permis une meilleure segmentation des risques. Les résultats des modèles RI et RIT ont été utilisés pour constituer un zonier simplifié, divisant la France en cinq zones géographiques (A, B, C, D, E) selon la fréquence prédite des sinistres. Cette segmentation a ensuite été utilisée pour calculer les primes pures et les pseudo-S/P (charge sinistres sécheresse/primes CatNat) pour chaque zone : la zone A (risque le plus faible avec une pseudo-prime pure de 12,48 €) et la zone E (risque le plus élevé avec une pseudo-prime pure de 147,36 €.). Les graphiques de la courbe de Lorenz et les coefficients de Gini sur la base d'apprentissage et la base de test ont montré que le modèle RIT est plus robuste et offre une meilleure segmentation des risques que le modèle RI, confirmant sa pertinence pour modéliser le risque de sécheresse.

Discussions

Le chapitre se termine par une réflexion sur l'intérêt d'une segmentation interne basée sur le risque sécheresse et sur les biais potentiels liés à notre outil. Il est également souligné que l'outil développé pourrait avoir des applications plus larges, au-delà du cas d'usage, pour modéliser d'autres risques naturels ou pour être utilisé dans d'autres régions ou pays.

Conclusion

Dans le contexte actuel de changement climatique, l'accès à des données géographiques de qualité devient crucial pour les compagnies d'assurance. Ce mémoire présente un outil algorithmique capable d'extraire des variables géographiques à partir d'images de cartes pour enrichir les bases de données actuarielles. Bien que l'outil ait montré une précision supérieure à 85 %, des améliorations sont encore possibles pour affiner sa précision d'extraction et renforcer la fiabilité des données.

Executive summary

Context and objectives

In the insurance sector, access to high-quality data has become a major strategic issue. With the rise of information technologies and increased storage and computing capabilities, insurers are engaged in a true data race. These data, whether from internal or external sources, are essential for refining risk models, improving pricing, optimizing claims management, and anticipating market developments.
The acquisition of geographical data, in particular, has become crucial for modeling various types of risks, from natural disasters to the behavior of insured individuals in specific areas, especially in the context of climate change. This thesis aims to develop an algorithmic tool capable of extracting geographical variables from maps of France in image format to contribute to the enrichment of actuarial databases.
The primary objective of this work is to design an algorithmic tool to leverage new sources of geographical data to improve risk modeling in insurance. It must be able to take as input a labeled map of France, in image format, and produce a database that, at the desired level (postal code or INSEE code), associates the corresponding label. By applying this tool to a practical case focused on drought risk, this thesis also aims to evaluate the performance and utility of the tool in different modeling contexts.

Legal and regulatory framework

Before proceeding with any development, the first step was to validate the legal framework for using such a tool with the relevant authorities. It was specifically pointed out that :

— Analysis must be conducted on a sufficiently large volume of data to ensure the anonymity of individuals, preventing any individual identification.
— The use of data must be limited to internal purposes, excluding any exploitation for commercial prospecting.
— Reproduced data must not be altered : when converting color scales to a scale from 1 to 3, initial equivalences must be respected and original levels preserved.
— It is essential to adhere to the general conditions of the website from which the data originates and to ensure that no mention prohibits their reproduction or reuse.

Throughout the completion of this thesis, particular attention has been paid to all images used with the tool to ensure compliance with intellectual property and copyright laws, both in terms of use and modification of content.

Design of the algorithmic tool

The developed tool takes the form of a sequence of algorithms aimed at methodically extracting the geographical information contained in a labeled map of France in image format, associating each municipality with a legend color. This method transforms visual data into exploitable variables for risk modeling. Figure 3 summarizes the various steps of the tool's operation.

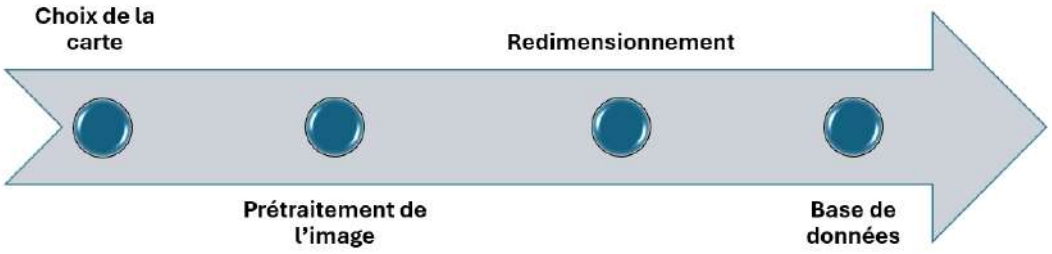


Figure 3 – Operation of the Algorithmic Tool

Image preprocessing

The development of the tool begins with the selection of maps to process. For this thesis, the selected maps needed to provide information potentially relevant to modeling home insurance claims, with a clear legend based on colors to identify different areas. *(However, it is important to note that the final tool is intended to be universal and not tied to home insurance use cases.)* They must also cover the entire geographical area of interest *(in this thesis, it is metropolitan France except Corsica).*
Before extracting the data, image preprocessing is performed to analyze the color composition of the image. This step aims to classify the colors of the pixels present in the image according to the colors of the legend, reduce noise, and identify pixels whose color is not in the legend. Noise refers to any elements that may alter the colors of the image or make it difficult to recover the colors of certain pixels. To achieve this, color clustering is used, which involves segmenting the image pixels into clusters of uniform colors. Pixel colors are analyzed using the RGB system, which assigns a value to a color. Two methods are then used for color clustering : K-Means and Principal Component Analysis (PCA). K-Means groups the pixels into clusters based on their colors, thus eliminating gradations and uniformizing similar pixels. PCA, on the other hand, projects the pixels into a two-dimensional plane, facilitating the detection of patterns and grouping similar pixels. In this context, pixels are defined only by their color, and those with similar colors

should be close to each other in the main PCA plane. These steps ensure better quality of the extracted data.

Reference map and database

Next, a reference database is constructed to associate each French municipality with a pixel position on a so-called reference map. This step fits into the general idea of the tool, which is to have a map where pixel positions are known for all French postal codes, and then to resize the map from which information is to be extracted so that it perfectly overlays the reference map, making the pixel positions of municipalities known.

This step begins with the selection of the reference map representing metropolitan France (excluding Corsica), which must meet certain criteria. From this map, pixel coordinates for several cities are manually collected. These data are then used to form a linear regression model to predict the coordinates of other cities.

Two linear regression approaches are explored for modeling these coordinates : one based on latitudes and longitudes, and the other on Lambert 93 coordinates. Lambert 93 is the official cartographic projection system for representing France. Comparing the performance of the two approaches shows that using Lambert 93 coordinates offers better accuracy for modeling pixel coordinates, as indicated by the RMSE and R^2 metrics. In addition to quantitative results, using Lambert 93 coordinates provides some confidence in the model due to the precision of the projection.

Resizing

This step involves adjusting the dimensions, tilt, and position of the images to exactly match the reference map. A mask of the reference map is created, and affine geometric transformations of the plane (translation, rotation, scaling, and reflections) are tested successively until each image aligns with the mask. These transformations are carried out by matrix calculations using Python and matrix theory for geometric transformations.
To achieve a perfect overlay, optimization methods are employed. An error function is defined to measure the difference between the processed image and the reference mask. This function combines two terms : one penalizes the presence of non-empty pixels (the color of empty space is defined to group all colors present in the image but absent from the legend) outside the mask, and the other penalizes the presence of empty pixels inside the mask. Powell and Nelder-Mead optimization methods, suited for non-differentiable functions, are used to minimize this error. Although these methods are effective, they do not always guarantee a perfect overlay, sometimes leaving empty pixels inside the mask. To correct these imperfections, the K-Nearest Neighbors (K-NN) algorithm is applied to reassign the color of empty pixels based on the colors of neighboring pixels.

Tool performance

Four maps were used to test the tool. For two of them, the original databases were available, allowing for an assessment of classification performance for municipalities. For map A, PCA achieved a classification rate of 89.63

FIGURE 4 – Example of Map[54] - Original vs Recreated

Limitations and improvements

Although the tool shows promising performance, several limitations must be considered. First, a performance index can only be calculated for maps with available original databases. For other maps, the evaluation remains subjective. Second, accuracy does not reach 100

Use case : simplified drought zoning system

In the context of home insurance, Cardif IARD seeks to refine its technical profitability model, which assesses the expected profitability of insurance contracts using various variables, including external data. A specific model for drought risk did not exist due to regulation setting the Natural Catastrophe (CatNat) premium in advance. However, such a model could improve drought risk monitoring by offering better segmentation of risk areas and a more accurate estimation of pure premiums. This practical case aims to test a model of expected pure premium for drought using a geographical data extraction tool developed in this thesis.
The protocol followed in this practical case includes several key steps :

- Extraction of geographical variables from maps relevant to drought risk.
- Modeling of claim frequency using a GLM (Generalized Linear Model).
- Construction of a simplified zoning based on modeling results.
- Calculation and interpretation of pseudo-S/P for the identified zones.

Extraction of geographical variables

Five maps were selected for their relevance to drought risk : a map of clay soils (carte_rga), maps of heat and solar radiation, as well as maps of average temperature

and precipitation. These maps were processed to extract the corresponding databases to be included in the model. The image processing method included color clustering and resizing operations to obtain a base that accurately represents the maps.

Frequency modeling

Model 1C focuses on a single explanatory variable, the carte_rga, which represents the risk of clay shrinkage-swelling. This map has four categories : low, no info, medium, and high. The frequencies of claims by category show an increasing trend with the risk level. Analysis showed that the frequency of claims is significantly higher in areas where the risk of shrinkage-swelling is classified as "high," thus confirming the relevance of the carte_rga for modeling drought risk. The model was then evaluated using the Lorenz curve and the Gini coefficient, with a coefficient of 0.275, indicating moderate segmentation of risk areas. This analysis also compared the pseudo-pure premium related to drought with the average CatNat premium, revealing significant discrepancies, especially in high-risk areas.

The transition to a more complex modeling with RI and RIT models allowed for the inclusion of several explanatory variables from different geographical maps :

— RI Model : This model uses two variables : carte_rga and carte_irrad, which measures solar radiation level.
— RIT Model : This model adds carte_temp (average temperature) to the variables of the RI model.

The *beta* coefficients of the models, their statistical significance, and their impact on claim frequency were analyzed. For the RIT model, all significant coefficients show a consistent logic with claim frequency. For example, the carte_rga has coefficients indicating that areas with "high" exposure to risk have a higher frequency of claims, while the carte_temp shows that areas with high average temperatures are more likely to experience drought-related claims.
Model evaluation metrics showed that the RIT model is slightly superior to the RI model, with a Gini coefficient of 0.327 compared to 0.319 for the RI model. This slight improvement indicates that adding carte_temp allowed for better risk segmentation. The results of the RI and RIT models were used to create a simplified zoning, dividing France into five geographical zones (A, B, C, D, E) based on the predicted claim frequency. This segmentation was then used to calculate pure premiums and pseudo-S/P (drought claims/premiums CatNat) for each zone : zone A (lowest risk with a pseudo-pure premium of €12.48) and zone E (highest risk with a pseudo-pure premium of €147.36).
Lorenz curve graphs and Gini coefficients based on the training and test sets showed that the RIT model is more robust and offers better risk segmentation than the RI model, confirming its relevance for modeling drought risk.

Discussions

The chapter concludes with a reflection on the interest of internal segmentation based on drought risk and potential biases related to our tool. It is also highlighted that the developed tool could have broader applications beyond the use case, for modeling other natural risks or for use in other regions or countries.

Conclusion

In the current context of climate change, access to high-quality geographical data becomes crucial for insurance companies. This thesis presents an algorithmic tool capable of extracting geographical variables from map images to enrich actuarial databases. Although the tool has demonstrated an accuracy of over 85%, improvements are still possible to refine its extraction precision and enhance data reliability.

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude à Cardif IARD pour m'avoir offert l'opportunité de réaliser cette alternance. Un remerciement tout particulier à mon tuteur, Nicolas Zhang, pour son encadrement bienveillant et son implication constante tout au long de cette expérience. Je remercie également l'ensemble de l'équipe Tarification & Analytics, notamment Vincent Guien, Achraf Ennasser, et Kevin Farnault, pour leur disponibilité et leur aide précieuse qui ont grandement facilité mon intégration et mes travaux.

Je souhaite ensuite adresser mes remerciements à l'équipe pédagogique de l'EURIA pour la qualité de la formation dispensée, qui a été déterminante dans mon parcours académique et le sera sûrement dans mon parcours professionnel. Un merci tout particulier à Vincent Soulas et Franck Vermet, dont l'encadrement attentif a été essentiel à la réalisation de mon mémoire ainsi qu'au bon déroulement de mon alternance.

Je suis également reconnaissant envers mes amis du TTU pour leur soutien moral et leur aide précieuse lors de la relecture de ce mémoire.

Enfin, mes plus sincères remerciements vont à ma famille, dont l'amour et le soutien indéfectibles ont contribué à faire de moi la personne que je suis aujourd'hui.

« La passion et les rêves sont comme le temps, rien ne peut les arrêter. »

Table des matières

Introduction

L'assurance habitation joue un rôle important dans la société moderne en offrant une protection financière contre une série de risques, y compris les catastrophes naturelles, dont l'ampleur et la fréquence augmentent en raison du changement climatique. Cette protection préserve non seulement les biens des assurés, mais contribue également à stabiliser l'économie de la région touchée. Afin de fournir cette couverture de manière efficace et rentable, les assureurs doivent évaluer avec précision les risques associés à chaque bien assuré.

L'évaluation des risques d'assurance repose traditionnellement sur des modèles statistiques tels que les modèles linéaires généralisés. De plus, les progrès récents en matière d'apprentissage statistique ont introduit des algorithmes tout aussi puissants (voire plus) capables de traiter de grandes quantités de données et d'affiner les prédictions. Cependant, la performance de toutes ces méthodes dépend fortement de la disponibilité et de la qualité des données. Les données géographiques, telles que les zones inondables en France par exemple, sont particulièrement importantes pour l'évaluation des risques en assurance habitation. Pourtant, ces informations sont souvent disponibles sous une forme non structurée (images de cartes), ce qui complique leur intégration dans les modèles statistiques traditionnels.
Pour répondre à ce défi, une alternative est de développer un outil algorithmique capable d'extraire des variables géographiques à partir de cartes légendées au format image (sous réserve d'un cadre légal autorisant leur utilisation). Cet outil permettra de transformer des données non structurées en informations exploitables par les modèles de prévision, offrant ainsi la possibilité d'enrichir ces modèles et de les rendre plus efficaces.

Ce mémoire est structuré de manière à offrir une compréhension complète et détaillée de cette problématique. La première section présente une revue de la littérature qui explore les techniques de modélisation de la sinistralité en assurance habitation et l'importance des données géographiques. La deuxième partie détaille la méthodologie adoptée pour le développement de l'outil d'extraction de données, incluant les techniques de traitement d'image et la construction d'une base de données géographiques. Les résultats obtenus seront ensuite présentés, analysés et discutés pour évaluer la performance de l'outil. Enfin, un cas pratique sur le risque climatique illustrera l'application de cet outil, démontrant son utilité et son impact dans la modélisation et la gestion de la sinistralité.

Chapitre 1

Modélisation des risques en assurance habitation : revue de littérature

La modélisation des risques en assurance habitation est un processus essentiel pour les compagnies d'assurance. Ce processus permet non seulement de déterminer les primes d'assurance de manière juste et équitable, mais aussi de garantir la stabilité financière des assureurs en minimisant les pertes dues aux sinistres. Cette revue de la littérature explore les défis actuels de l'assurance habitation, le régime particulier des catastrophes naturelles, et les méthodes de sélection des risques employées par les assureurs, principalement dans le contexte actuel de changement climatique.

1.1 Assurance habitation et catastrophes naturelles

L'assurance habitation est un pilier essentiel du secteur des assurances, offrant une protection financière contre divers types de sinistres, tels que les catastrophes naturelles, les incendies, les cambriolages et autres risques. Les défis actuels dans ce domaine sont nombreux et variés, englobant non seulement la gestion efficace des sinistres et la satisfaction des assurés, mais aussi l'adaptation aux évolutions réglementaires et technologiques. Le changement climatique, avec l'augmentation de la fréquence et de la gravité des catastrophes naturelles, ajoute un défi supplémentaire et croissant pour l'ensemble de l'industrie.

1.1.1 Assurance habitation : Défis et enjeux

L'assurance habitation, communément appelée assurance multirisque habitation (MRH), est un contrat d'assurance visant à indemniser les assurés en cas de sinistres liés à leur domicile. Cette couverture s'applique aux propriétaires, locataires et occupants temporaires. La MRH constitue le type de contrat le plus répandu et le plus complet dans le

domaine de l'assurance IARD (Incendie, Accidents et Risques Divers). Elle comprend :

— **L'assurance responsabilité civile (RC) :**

1. RC vie privée : Elle couvre l'assuré contre les dommages non intentionnels qu'il pourrait causer à des tiers dans sa vie privée, que ce soit à l'intérieur ou à l'extérieur de son domicile, qu'il s'agisse de dommages matériels, corporels ou immatériels.
2. RC habitation : Elle prend en charge les dégâts causés par le logement de l'assuré aux tiers, dont l'assuré pourrait être tenu responsable. Par exemple, si un incendie se déclare dans l'appartement de l'assuré et endommage le logement du voisin, ou si, en tant que propriétaire, une tuile de sa maison tombe et abîme une voiture stationnée dans la rue.

— **L'assurance de l'habitation :** Celle-ci protège le logement de l'assuré (à la fois l'immobilier et le mobilier) contre les dommages qu'il pourrait subir. Par exemple, elle couvre les dégâts causés par une tempête qui arrache la toiture du logement, ou les dommages résultant d'un dégât des eaux.

Garanties. Les garanties d'un contrat d'assurance multirisque habitation (MRH) définissent précisément ce qui est couvert par l'assurance. Elles permettent à l'assuré de comprendre l'étendue de sa protection et servent de base pour le calcul de la prime d'assurance. Les contrats MRH incluent des garanties de base et peuvent être personnalisés avec des garanties optionnelles selon les besoins spécifiques de l'assuré.
Les principales garanties de base d'un contrat MRH comprennent :

— la responsabilité civile vie privée : Couvre les dommages causés à des tiers dans le cadre de la vie quotidienne.
— la responsabilité civile habitation : Couvre les dégâts causés par le logement de l'assuré aux tiers.
— les dégâts des eaux : Protège contre les dommages causés par les infiltrations d'eau, les fuites ou les ruptures de canalisation.
— les incendies et explosion : Couvre les dommages matériels résultant d'un incendie ou d'une explosion.
— le vol : Indemnise les pertes liées au vol ou à la tentative de vol dans le domicile assuré.
— les évènements climatiques : Couvre les dommages causés par des tempêtes, la grêle et la neige.
— les catastrophes naturelles : Indemnise les sinistres liés à des événements tels que la sécheresse, les inondations et les mouvements de terrain.

En plus des garanties de base, de nombreuses assurances habitation proposent des garanties supplémentaires qui permettent d'adapter la couverture aux besoins spécifiques de l'assuré. Parmi celles-ci, on trouve :

— la protection juridique : Couvre les frais de justice en cas de litige lié au logement.

— l'assistance : Offre des services d'urgence 24h/7j, tels que l'intervention d'un serrurier ou d'un plombier.
— les bris de glace : Couvre les dommages aux surfaces vitrées, comme les fenêtres et les miroirs.
— les dommages électriques : Protège les appareils électroménagers et électroniques contre les pannes ou les dommages.

Il est également essentiel de prendre en compte les plafonds et les franchises associés à chaque garantie. Les plafonds correspondent aux montants maximums d'indemnisation, généralement exprimés en euros ou en pourcentage de la valeur assurée. Les franchises, quant à elles, représentent le montant restant à la charge de l'assuré en cas de sinistre, et peuvent varier selon le type de garantie et le contrat.
Certaines particularités peuvent également être intégrées dans les contrats MRH. Par exemple, la valeur à neuf permet un remboursement à la valeur d'achat pour les biens récents, tandis que la renonciation à la règle proportionnelle engage l'assureur à ne pas appliquer de réduction d'indemnité en cas de sous-assurance jusqu'à un certain pourcentage. L'option "Tous risques sauf" peut également être proposée, offrant une couverture plus large en incluant tous les dommages sauf ceux expressément exclus.
Enfin, il est important de noter que chaque garantie comporte des exclusions spécifiques, détaillées dans les conditions générales du contrat. Certains événements, tels que la guerre ou la faute intentionnelle, sont généralement exclus de toutes les garanties. Il est donc crucial pour l'assuré de bien comprendre ces différents éléments afin de choisir un contrat adapté à ses besoins et d'éviter les mauvaises surprises en cas de sinistre. Par ailleurs, il est important de comprendre qu'en 2024, le secteur de l'assurance habitation fait face à de nombreuses problématiques, reflétant les changements dans l'environnement économique, climatique et sociétal.

Risque climatique et sinistralité accrue. Les événements climatiques extrêmes et les catastrophes naturelles sont en augmentation en raison des changements climatiques. Ces événements posent des défis majeurs pour les assureurs, notamment en termes de prévision et de gestion des sinistres.

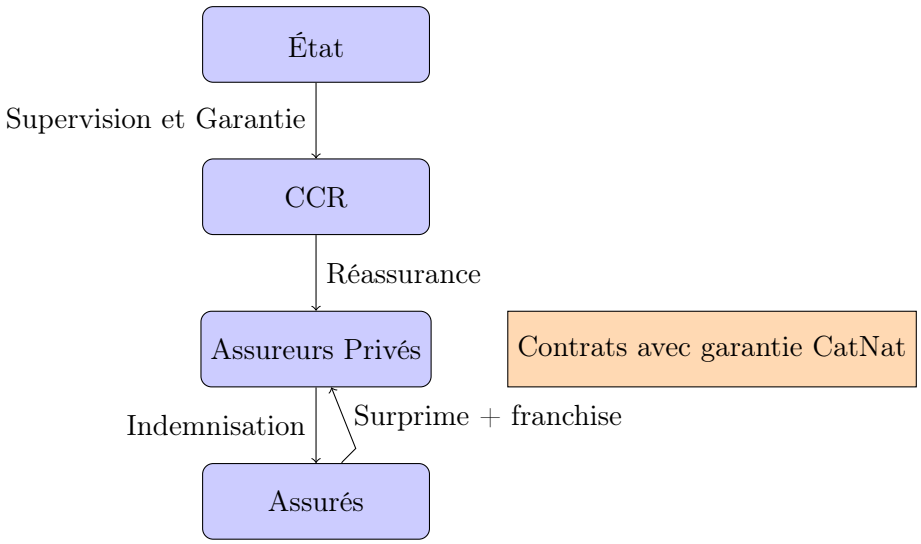
Impact de l'inflation. L'inflation économique, exacerbée par la récession mondiale, a un impact direct sur les coûts des sinistres. Les assureurs doivent gérer les hausses de coûts liées à l'inflation économique et sociale, incluant l'augmentation des frais de litige et des indemnités.

1.1.2 Le régime des catastrophes naturelles (CatNat)

Le régime d'indemnisation des catastrophes naturelles (CatNat) a été instauré en France par la loi du 13 juillet 1982, en réponse aux inondations dévastatrices qui ont frappé le Sud-Ouest du pays en 1981. Ce système unique vise à offrir une couverture d'assurance pour des risques naturels jugés inassurables par le marché privé, tout en ga-

rantissant une solidarité nationale face aux catastrophes naturelles. L'objectif principal est d'assurer une indemnisation rapide et efficace des sinistrés, indépendamment de leur situation géographique ou économique.

Le fonctionnement du régime CatNat repose sur un mécanisme spécifique. Il s'active par la publication d'un arrêté interministériel au Journal Officiel, déclarant l'état de catastrophe naturelle pour une zone et un événement spécifiques. Les risques couverts comprennent notamment les inondations, les sécheresses, les avalanches, les séismes et les mouvements de terrain. Il est important de noter que certains phénomènes naturels, tels que les tempêtes, la grêle, le poids de la neige ou les incendies de forêt, ne sont pas couverts par ce régime et relèvent d'autres garanties d'assurance. Cette exclusion s'explique par le fait que ces risques sont considérés comme assurables par le marché privé. En effet, les dommages causés par les tempêtes, la grêle et la neige sont généralement couverts par la garantie "Tempête, Grêle et Neige" (TGN) incluse dans les contrats multirisques habitation standards. Cette distinction permet de maintenir l'équilibre du régime CatNat en le réservant aux risques jugés "inassurables" par le seul secteur privé, tout en assurant une couverture adéquate pour les autres phénomènes naturels via les contrats d'assurance classiques.
La structure du régime CatNat s'appuie sur un partenariat public-privé innovant. Les assureurs privés proposent la garantie CatNat dans leurs contrats multirisques habitation et entreprise. La Caisse Centrale de Réassurance (CCR), détenue par l'État, offre une réassurance aux assureurs avec la garantie illimitée de l'État. Ce dernier joue un rôle crucial en fournissant cette garantie illimitée à la CCR et en supervisant l'ensemble du système. Le financement du régime est assuré par une surprime CatNat, actuellement fixée à 12% de la prime d'assurance dommages aux biens, qui passera à 20% en 2025. Une franchise légale est également appliquée, s'élevant à 380 € pour les biens à usage d'habitation (1520 € pour la sécheresse).



La procédure d'indemnisation dans le cadre du régime CatNat suit un processus défini. L'assuré doit déclarer le sinistre à son assureur dans les 10 jours suivant la publication de l'arrêté de catastrophe naturelle. Un expert mandaté par l'assureur procède ensuite à l'évaluation des dommages. L'indemnisation doit intervenir dans un délai de 3 mois à compter de la remise de l'état estimatif des pertes ou de la publication de l'arrêté. Ce système vise à garantir une prise en charge rapide des sinistrés, essentielle dans des situations souvent critiques. Aujourd'hui, le régime CatNat fait face à des enjeux et défis majeurs, principalement liés au changement climatique. L'augmentation prévue de la fréquence et de l'intensité des événements climatiques extrêmes pose un défi considérable. La CCR estime une hausse de 40% des coûts de sinistralité d'ici à 2050, ce qui soulève des questions sur l'équilibre financier à long terme du régime. L'épuisement progressif de la provision d'égalisation de la CCR est un signal d'alerte, indiquant la nécessité d'adapter le financement du régime face à l'augmentation des risques.

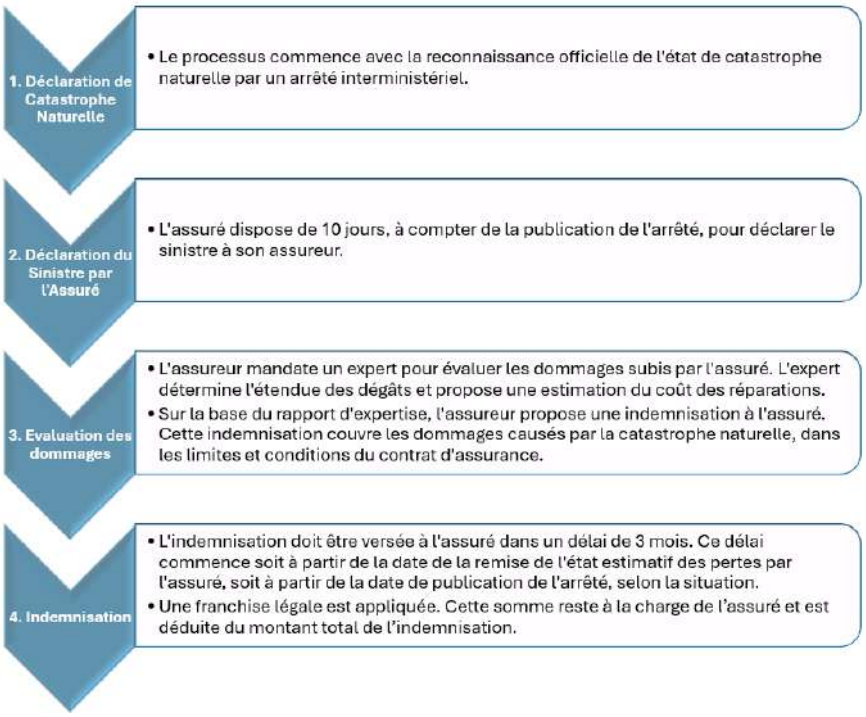


FIGURE 1.1 – Processus d'indemnisation CatNat

La problématique spécifique de la sécheresse mérite une attention particulière. Ce phénomène engendre des coûts importants et en augmentation, tout en présentant des difficultés dans la reconnaissance et l'évaluation des dommages. Cette situation souligne l'importance de renforcer les mesures de prévention et d'adaptation, un aspect crucial pour la pérennité du régime CatNat.

Face à ces défis, des réflexions sont en cours sur une possible réforme du financement du régime et sur le renforcement du lien entre indemnisation et prévention. L'adaptation des critères de reconnaissance des catastrophes naturelles, notamment pour la sécheresse, est également à l'étude. Ces évolutions potentielles visent à assurer la viabilité à long terme du régime CatNat, tout en maintenant son rôle crucial dans la protection financière des Français face aux catastrophes naturelles.

1.1.3 La sélection des risques : un atout dans la gestion de la sinistralité ?

La sélection des risques est le processus par lequel un assureur évalue le niveau de risque présenté par un potentiel assuré ou un bien à assurer, afin de décider s'il accepte de fournir une couverture et à quelles conditions. Selon Michel Fromenteau, Victor Ruol et Laurence Eslous : « *La sélection des risques est dans la nature même de l'activité d'assurance, c'est le métier de l'assureur et il conserve toujours la possibilité de refuser un risque* »[44]. En identifiant les risques élevés, les assureurs peuvent prendre des mesures pour limiter leur exposition, comme exiger des mesures de prévention supplémentaires.[21]
Pour effectuer une sélection des risques, les assureurs utilisent diverses méthodes. Le principe intrinsèque reste cependant le même pour chaque méthode, déterminer un seuil de risque à partir duquel l'assurance est refusée ou fortement conditionnée :

- Évaluation géographique : La zone géographique est analysée pour déterminer les risques naturels, les taux de criminalité, et d'autres facteurs locaux. Pour un logement se trouvant, par exemple, dans une zone proche d'un cours d'eau ou avec un indice de précipitation élevé, l'assureur peut majorer la prime d'assurance de 15% par rapport au tarif standard en raison du risque d'inondation.
- Historique de sinistres : L'historique des sinistres du demandeur est évalué pour identifier les tendances et les risques. Si cet historique montre des sinistres fréquents, l'assureur pourrait réviser les conditions du contrat, en ajustant les primes, pour mieux refléter le niveau de risque estimé.
- Algorithmes prédictifs : Les assureurs utilisent des données et des modèles statistiques pour affiner leur évaluation du risque. Ils déterminent un score, une probabilité d'occurrence d'un sinistre, qui peut être éliminatoire, à partir des informations du bien à assurer. Les données sur le bien à assurer sont généralement recueillies grâce à un questionnaire détaillé.

La sélection des risques a été décrite dans cette section comme un avantage pour les assureurs, leur permettant de refuser les risques trop élevés et ainsi de préserver leur rentabilité financière. Cependant, la situation est plus complexe, surtout dans le contexte actuel de hausse des risques climatiques. En effet, le changement climatique influence la sélection des risques de plusieurs façons :

Désengagement des assureurs. Le changement climatique pourrait engendrer un désengagement des assureurs dans les zones très exposées aux risques climatiques, qu'ils

jugeraient inassurables, ce qui constituerait un risque stratégique majeur. En réalité, le système français d'indemnisation des catastrophes naturelles repose sur la solidarité entre assurés. Le désengagement des assureurs dans les zones à haut risque peut entraîner une démutualisation du système, où les assureurs restants dans ces zones voient leurs pertes augmenter, ce qui les inciterait à augmenter les primes pour tous leurs clients, y compris ceux dans les zones à faible risque. Cela peut nuire à la compétitivité des assureurs et réduire leur capacité à mutualiser les risques.

Discrimination géographique. La discrimination géographique désigne une pratique selon laquelle les assureurs utilisent la localisation géographique comme critère principal pour déterminer l'éligibilité à l'assurance et le montant des primes. Cela signifie que les personnes vivant dans des zones à haut risque de catastrophe naturelle pourraient être obligées de payer des primes beaucoup plus élevées que celles vivant dans des zones à faible risque ou se voir refuser une couverture d'assurance dans le pire des cas. Cette pratique, souvent considérée comme une forme d'antisélection inversée, survient lorsque ce sont les assureurs, et non les assurés, qui cherchent à éviter les risques les plus élevés. Cela crée un déséquilibre, non plus dans la composition des portefeuilles d'assurance comme dans le cas de l'antisélection classique, mais dans l'accès même à l'assurance, limitant la couverture pour les populations les plus exposées. Ce phénomène est contraire au principe de solidarité entre assurés, qui est à la base du système français d'indemnisation des catastrophes naturelles.

En somme, la sélection des risques est un élément crucial de la gestion des assurances habitation. Elle permet aux assureurs de maintenir leur viabilité financière tout en offrant une couverture adaptée. Néanmoins, elle soulève également des questions d'équité et d'accessibilité, notamment face aux défis croissants posés par le changement climatique et l'évolution des risques, auxquelles il faut trouver des réponses. Dans leur rapport final sur le thème *Adapter le système assurantiel français face à l'évolution des risques climatiques*, Thierry Langreney, Gonéri Le Cozannet et Myriam Merad insistent sur la nécessité de créer une cartographie harmonisée et partagée des zones à plus forte exposition, pour identifier notamment les "zones rouges" (à aléa très élevé) et les "zones orange" (à aléa élevé)[55]. Le rapport propose également de mettre en place des mécanismes incitatifs pour maintenir l'engagement des assureurs dans ces zones à risque, notamment en envisageant des pénalités pour ceux qui se retireraient des zones les plus exposées. Cela vise à garantir que la solidarité entre assurés soit préservée, tout en répondant aux spécificités des risques climatiques locaux. L'objectif est de renforcer la résilience du système assurantiel en garantissant que les assureurs continuent à jouer un rôle actif dans les zones à risque, sans pour autant déséquilibrer la mutualisation des risques au niveau national.

1.2 Modélisation de la sinistralité en assurance habitation

La modélisation de la sinistralité en assurance habitation est un processus complexe qui vise à prédire la fréquence (le nombre de sinistres survenant au cours d'une période donnée) et la gravité des sinistres (le coût moyen des sinistres). Cela permet aux compagnies d'assurance de fixer des primes adéquates et de gérer efficacement les risques. Modéliser la sinistralité revient donc à estimer la prime pure, c'est-à-dire la somme des coûts attendus des sinistres. La sinistralité totale S peut ainsi être définie comme le produit entre la fréquence N et la gravité X :

$$S = N \times X = f(Y_1, ..., Y_n)$$

où $Y_1, ..., Y_n$ représentent des variables explicatives telles que les caractéristiques des contrats et des variables externes.
Ce processus repose sur l'utilisation de méthodes statistiques avancées et d'algorithmes de machine learning pour analyser des données historiques et déterminer cette fonction f.

1.2.1 Apprentissage statistique et modélisation des risques

L'apprentissage statistique est une branche de l'intelligence artificielle qui utilise des méthodes statistiques pour permettre aux systèmes informatiques d'apprendre et d'améliorer leurs performances à partir de données, sans être explicitement programmés. Dans le domaine de l'apprentissage statistique, il existe plusieurs types de méthodes utilisées pour modéliser et prédire des phénomènes à partir de données. Les principaux types d'apprentissage sont l'apprentissage supervisé, l'apprentissage non supervisé, et l'apprentissage par renforcement.

a - Apprentissage supervisé.

L'apprentissage supervisé est une méthode avec laquelle le modèle est entraîné sur un ensemble de données étiquetées, c'est-à-dire des données pour lesquelles la variable cible (ou la "vérité terrain") est connue. L'objectif est de créer un modèle capable de prédire correctement les résultats pour de nouvelles données non étiquetées. Il existe deux catégories d'apprentissage supervisé :

1. La régression : Dans ce cas, l'étiquette est une valeur numérique. La régression est utilisée pour prédire des valeurs continues, comme le montant d'une réclamation d'assurance ou le coût d'un sinistre.
2. La classification : Ici, l'étiquette est un groupe ou une catégorie auquel appartient la donnée. La classification est utilisée pour attribuer des observations à des catégories prédéfinies, comme déterminer si une réclamation est frauduleuse ou non.

En assurance habitation, l'apprentissage supervisé est le plus souvent utilisé pour prédire la fréquence et la gravité des sinistres en fonction de différentes variables explicatives, et déterminer la prime pure. Les modèles les plus courants incluent :

— Les modèles linéaires généralisés (GLM) : Ces modèles permettent de traiter des variables dépendantes suivant une distribution normale ou une autre, ce qui est souvent le cas en assurance. Ils sont utilisés pour modéliser la fréquence (loi de poisson) et le coût des sinistres (loi gamma), et même la prime pure dans certains cas, en fonction de variables explicatives.
— Les régressions pénalisées : La régression pénalisée est une technique de modélisation statistique utilisée pour améliorer la performance des modèles de régression en ajoutant une contrainte ou une pénalité aux coefficients du modèle. Cette méthode est particulièrement utile pour prévenir le surapprentissage (*overfitting*), surtout lorsqu'il y a un grand nombre de variables explicatives. Les deux types de régression pénalisée les plus couramment utilisés sont la régression Ridge et la régression Lasso.
— Les arbres de régression/classification et les forêts aléatoires : Les arbres de régression et de classification sont des modèles non paramétriques qui segmentent les données en sous-groupes en fonction des valeurs des variables explicatives. Ces modèles sont particulièrement populaires en assurance pour leur simplicité, leur capacité à capturer des interactions complexes et des relations non linéaires entre les variables, mais surtout l'interprétabilité. Les forêts aléatoires, quant à elles, sont une extension des arbres de décision qui améliorent la robustesse et la précision des prédictions en utilisant un ensemble de nombreux arbres de décision, chacun construit sur un échantillon aléatoire des données et des variables explicatives. Cette méthode réduit la variance des prédictions et améliore la performance globale du modèle.
— Le *gradient boosting* : C'est une technique qui améliore la précision des modèles en combinant plusieurs modèles simples (comme des arbres de décision) de manière itérative. Chaque nouveau modèle corrige les erreurs du précédent, ce qui permet d'obtenir un modèle final très précis et robuste. Utilisé en assurance habitation, il aide à prédire la fréquence et la gravité des sinistres, à déterminer des primes adéquates, et à détecter les fraudes.

Focus sur les GLM. Les modèles linéaires généralisés (GLM) permettent de modéliser les relations entre une variable dépendante et plusieurs variables explicatives. Formellement, l'espérance μ de la variable dépendante est liée linéairement aux prédicteurs X via une fonction de lien g :

$$\mu = g^{-1}(X\beta)$$

où β est le vecteur des coefficients des variables explicatives.
Les paramètres β sont estimés par la maximisation de la vraisemblance, ce qui implique de résoudre :

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta|\mathbf{x}) = \arg\max_{\theta} \prod_{i=1}^{n} f(x_i|\theta)$$

où $f(x_i|\theta)$ est la fonction de densité de probabilité pour l'observation x_i et le paramètre θ, et $\mathcal{L}(\theta|\mathbf{x})$ est la fonction de vraisemblance.

Pour un modèle avec une distribution normale, la fonction de lien est l'identité, ce qui signifie que l'espérance est directement une combinaison linéaire des prédicteurs :

$$E(Y|x_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

Dans ce cas, la variance est constante, ne dépend pas de la moyenne et on retrouve une régression linéaire.
Pour finir, il faut ajouter que lorsqu'on travaille avec des variables catégorielles dans un GLM, une catégorie spécifique est choisie comme base ou référence. Les coefficients des autres catégories sont alors interprétés par rapport à cette base. Par exemple, si l'on modélise l'effet de la localisation géographique sur la fréquence des sinistres avec des catégories comme "région A", "région B", et "région C", l'une de ces régions sera choisie comme référence, et les coefficients associés aux autres régions indiqueront leur effet relatif sur la fréquence des sinistres par rapport à la région de référence.

b - Apprentissage non supervisé.

L'apprentissage non supervisé est une méthode où le modèle est entraîné sur un ensemble de données non étiquetées. Le but est de trouver des structures cachées ou des motifs dans les données. Cette méthode est particulièrement utile pour des tâches exploratoires où l'objectif principal est de segmenter ou regrouper des données similaires. Les techniques courantes incluent :

1. Le clustering : Il s'agit d'une méthode utilisée pour diviser un ensemble de données en groupes (clusters) d'éléments similaires. Les méthodes de clustering les plus utilisées sont les K-Means et la classification ascendante hiérarchique (CAH).
2. L'analyse en composantes principales (ACP) : C'est une technique de réduction de la dimensionnalité qui transforme les données en un ensemble de variables non corrélées (composantes principales), tout en conservant autant que possible la variation totale des données d'origine. Un espace de N variables est projeté sur un plan (deux dimensions), ou un hyperplan de dimension H $<$ N, en gardant le maximum d'informations possible.

En assurance habitation, l'apprentissage non supervisé, notamment via des techniques de clustering, permet de regrouper les clients en segments basés sur des similarités dans leurs profils de risque. Par exemple, en analysant des données comme le type de logement, la localisation, ou l'historique des sinistres, le clustering identifie des groupes de clients présentant des risques similaires. Ces segments permettent ensuite de développer des stratégies de tarification plus précises et personnalisées. En effet, chaque segment peut se voir attribuer une prime qui reflète mieux le risque spécifique qu'il présente plutôt qu'une tarification uniforme.

c - Apprentissage par renforcement.

L'apprentissage par renforcement est une méthode avec laquelle un "agent" (comme un programme) apprend à prendre des décisions en interagissant avec un environnement.

À chaque action entreprise, l'agent reçoit une récompense ou une punition en fonction de l'effet de cette action, l'objectif étant de maximiser les récompenses cumulées sur le long terme. Par exemple, si une stratégie de tarification attire plus de clients sans augmenter le risque de sinistre, l'agent est récompensé. Une méthode courante dans ce domaine est le *Q-learning*, qui aide l'agent à évaluer la valeur des actions dans différents contextes. Bien que rarement utilisé en assurance habitation, l'apprentissage par renforcement pourrait optimiser les stratégies de tarification et de souscription en s'appuyant sur les données des sinistres et les comportements des assurés, permettant ainsi aux assureurs de s'adapter dynamiquement aux évolutions du marché.

L'apprentissage statistique, ou machine learning, a permis une importante amélioration de la modélisation des sinistres en introduisant des techniques avancées qui permettent d'analyser de vastes ensembles de données et de créer des modèles complexes. De plus, ces modèles peuvent être continuellement mis à jour avec de nouvelles données, permettant aux assureurs de s'adapter rapidement aux changements du marché et aux nouvelles tendances des sinistres. Les modèles résultants peuvent être plus précis, mais dépendants d'une quantité massive et fiable de données.

1.2.2 Modélisation du risque géographique

Le risque géographique en assurance habitation réfère à l'influence de la localisation géographique sur la probabilité et l'impact des sinistres[34]. En d'autres termes, il s'agit de l'impact de l'environnement géographique sur le risque assuré. Par exemple, des facteurs tels que la densité de population, la proximité de lieux à haut risque de criminalité, la présence de services de sécurité, ou le type de sols peuvent influencer le risque de vol ou de sécheresse lié à une habitation.

Modéliser le risque géographique consiste à utiliser des données géographiques pour prédire la fréquence et la gravité des sinistres. Cela implique de développer des modèles statistiques qui intègrent des variables géographiques pour segmenter les risques de manière précise et personnalisée. L'objectif est de mieux comprendre les corrélations entre les caractéristiques géographiques et les sinistres, afin de proposer des primes d'assurance qui reflètent mieux le risque réel, ce qui améliore la précision de la tarification et la compétitivité des offres. Pour modéliser le risque géographique, les actuaires ont recours à ce qu'on appelle un zonier.

Un zonier est un outil utilisé pour traiter le signal géographique dans les modèles de prédiction. Il permet de segmenter les risques géographiques de manière à créer des classes de risques homogènes en sinistralité. Le zonier est essentiel pour adapter les primes d'assurance au risque réel de chaque zone, ce qui est crucial pour la rentabilité et la compétitivité des assureurs.

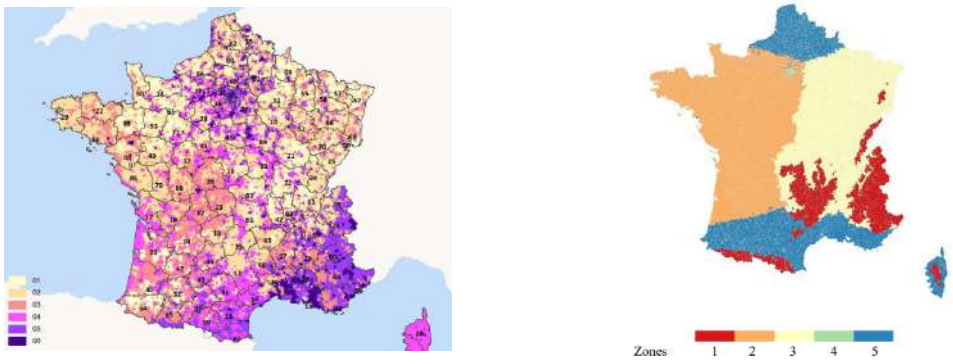


FIGURE 1.2 – Exemples n°1 de Zonier[15] (gauche) et n°2 de Zonier[43] (droite)

La méthode standard de construction d'un zonier, notamment dans le domaine de l'assurance habitation, repose sur l'identification des variations géographiques du risque à travers des techniques de modélisation statistique.
La première étape consiste à établir un modèle de risque à partir des données déclaratives, obtenues lors de la souscription de l'assurance, en l'absence totale de variables géographiques. On utilise généralement des modèles linéaires généralisés (GLM), présentés dans la section précédente, pour estimer le coût et la fréquence des sinistres.

— Modèle de fréquence des sinistres : Ce modèle prédit le nombre de sinistres pour une période donnée. La formule générale utilisée est basée sur un modèle de régression de Poisson, de quasi-poisson ou de loi binomiale négative en fonction de la dispersion des données :

$$E(N_i|X_1, ..., X_p) = \lambda_i \times Expo = exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}) \times Expo$$

où N_i est le nombre attendu de sinistres pour l'observation i, X_{ji} sont les variables explicatives, λ_i la fréquence des sinistres, $Expo$ l'exposition et les β_j sont les coefficients estimés du modèle.

— Modèle de coût des sinistres : Ce modèle évalue le coût moyen associé à chaque sinistre, en utilisant en général un modèle de régression Gamma (ci-dessous) ou log normale :

$$E(Y_i|X_1, ..., X_p) = \mu_i = exp(\alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + ... + \alpha_p X_{ip})$$

où Y_i est le coût du sinistre pour l'observation i, et α_j sont les coefficients estimés.

Les résidus R_i de ce modèle, c'est-à-dire la différence entre les valeurs observées et prédites (pour les résidus bruts), servent de cible pour le modèle géographique.
Avant de créer le zonier, il est important de confirmer que les résidus du modèle présentent une dépendance géographique significative. Une fois cette dépendance identifiée,

deux approches peuvent être envisagées. La première consiste à modéliser ces résidus en fonction des variables géographiques disponibles pour capturer la composante spatiale du risque. La seconde option est de réaliser directement une classification sur ces résidus afin de partitionner l'espace en zones homogènes de risque. Chaque zone ainsi définie est associée à un niveau de risque spécifique. Le choix du nombre de zones constitue un compromis entre la précision de la segmentation et les exigences commerciales, car un nombre trop élevé de zones pourrait entraîner des fluctuations tarifaires incomprises par les clients.

La construction d'un zonier implique l'intégration de données géographiques dans le modèle de tarification pour obtenir une meilleure discrimination du risque selon la localisation. Les modèles prédictifs utilisés doivent être robustes et adaptés aux particularités des données spatiales. Cette approche permet de capturer les variations locales du risque, ce qui est crucial pour l'amélioration de la précision tarifaire et la gestion efficace des portefeuilles d'assurance. Pour plus de détails, les mémoires de Catalina Sepulveda[9] et Guillaume Beraud-Sudreau[29] peuvent être consultés.

1.2.3 Collecte de données en actuariat

Il ressort des sections précédentes que les données jouent un rôle crucial dans la modélisation de la sinistralité, notamment en assurance habitation, que ce soit pour la modélisation des risques géographiques ou l'utilisation de technologies avancées d'apprentissage statistique. Plus la quantité de ces données est importante, plus les modèles gagnent en précision. Les travaux de ce mémoire visent à fournir une source supplémentaire de données pour alimenter ces modèles. De plus, les assureurs doivent collecter et rapporter des données pour se conformer aux exigences réglementaires. Il est donc essentiel de comprendre les types de données déjà utilisés et les méthodes employées par les actuaires pour les collecter jusqu'à ce jour.

Les méthodes traditionnelles de collecte de données utilisées par les actuaires ont considérablement évolué au fil du temps, s'adaptant aux avancées technologiques et aux besoins croissants en matière d'analyse de risques. Elles peuvent être regroupées en trois catégories.
Parmi les plus anciennes, il y a :

- Les registres manuels : Les actuaires compilaient manuellement les données des sinistres et des polices dans des registres physiques.
- Les enquêtes postales : Envoi de questionnaires par courrier aux assurés pour collecter des informations.

Ces méthodes permettaient une collecte de base des données essentielles, mais étaient lentes et sujettes aux erreurs. Dans les méthodes "intermédiaires" se trouvent :

- La saisie informatique : Utilisation de logiciels de base de données pour stocker et organiser les informations.
- Les enquêtes téléphoniques : Collecte d'informations plus rapide et interactive.

Elles ont permis l'amélioration de la précision et de la vitesse de traitement des données, permettant des analyses plus poussées. Enfin, les méthodes les plus récentes comprennent :

- La collecte automatisée : Utilisation de systèmes informatiques intégrés pour capturer automatiquement les données des polices et des sinistres.
- Les enquêtes en ligne : Collecte rapide et à grande échelle d'informations auprès des assurés.
- Le Big Data : Exploitation de vastes ensembles de données provenant de sources variées (IoT, réseaux sociaux, etc.).
- L'achat de données aux concurrents.
- L'exploitation de bases de données open source.

Sous réserve d'un cadre légal permettant leur utilisation, ces méthodes permettent une analyse plus fine et en temps réel des risques, une tarification plus précise, des études marketing et comportementales, une meilleure détection des fraudes et divers.

Par ailleurs, il est crucial de comprendre non seulement les types de données déjà utilisés, mais aussi les contraintes réglementaires qui influencent la collecte de ces données. En effet, le secteur de l'assurance en France est fortement réglementé, notamment sur la collecte et l'utilisation des données.

Réglementation stricte. La CNIL (Commission Nationale de l'Informatique et des Libertés) impose des règles strictes sur la collecte et l'utilisation des données personnelles, garantissant ainsi la protection de la vie privée des individus. Les actuaires doivent se conformer à ces règles lors de la collecte de données, ce qui peut restreindre l'accès à certaines informations sensibles. En plus des exigences de la CNIL, le RGPD (Règlement Général sur la Protection des Données) ajoute une couche supplémentaire de contraintes en matière de protection des données. Les entreprises doivent obtenir un consentement explicite des individus pour collecter et utiliser leurs données, ce qui peut compliquer le processus de collecte de données pour les actuaires. En outre, l'ACPR (Autorité de Contrôle Prudentiel et de Résolution) impose des exigences strictes en matière de reporting et de qualité des données. Les compagnies d'assurance doivent soumettre des rapports détaillés, ce qui nécessite une collecte rigoureuse des données. Cette régulation vise à garantir la stabilité financière et la solvabilité des assureurs, tout en protégeant les assurés.

Forte concurrence. Le marché français de l'assurance est très concurrentiel, ce qui rend l'accès à des données de qualité souvent coûteux. Pour gagner un avantage concurrentiel, certaines compagnies d'assurance investissent massivement dans l'achat de données externes. Cela inclut des données géospatiales, des données de comportement des individus, et d'autres types de données pertinentes pour affiner les modèles de risque.

Ces particularités obligent les actuaires français à être particulièrement vigilants dans leurs méthodes de collecte de données, en équilibrant la nécessité d'obtenir des informa-

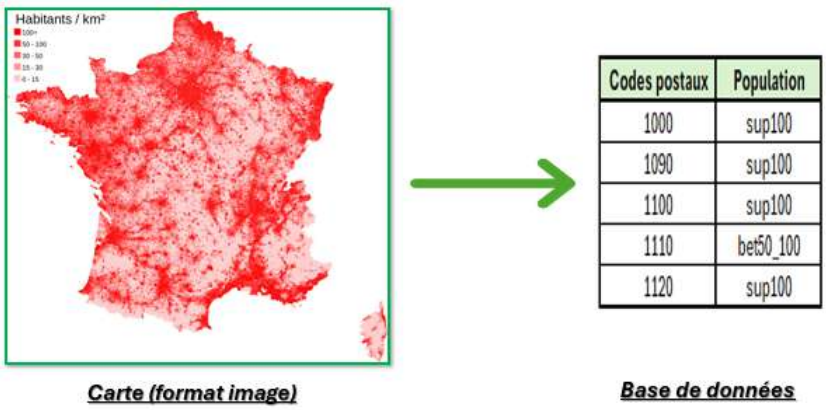
tions précises avec le respect de la réglementation et les contraintes budgétaires. L'innovation dans les méthodes de collecte et d'analyse des données devient donc un enjeu majeur pour rester compétitif tout en respectant le cadre légal.

Chapitre 2

Méthodologie et développement de l'outil

2.1 Objet et démarche générale

L'objectif principal de cette section est de détailler le développement d'un outil capable d'extraire des données géographiques à partir d'images de cartes. Toutefois, bien avant la conception effective de cet outil, la première étape a été de consulter le service juridique pour confirmer la légalité d'un tel outil et définir son cadre d'utilisation.
Par ailleurs, il convient de souligner la philosophie d'utilisation de cet outil. Il vise à extraire en masse des données à partir de cartes jugées pertinentes pour la modélisation de la sinistralité, tout en permettant de tester rapidement des données de carte dont l'intérêt reste incertain. Grâce à l'outil, il devient possible d'extraire rapidement des informations de ces cartes et de les tester dans les modèles de sinistralité, permettant ainsi d'évaluer leur impact potentiel sans limitation préalable.



Codes postaux	Population
1000	sup100
1090	sup100
1100	sup100
1110	bet50_100
1120	sup100

FIGURE 2.1 – Objet de l'outil algorithmique : passage d'une image de carte à une base de données

L'outil algorithmique doit donc répondre à plusieurs contraintes spécifiques pour garantir son efficacité et son adaptabilité.

Rapidité et simplicité d'utilisation. L'outil doit pouvoir, à partir d'une carte, sortir une base de données de façon relativement simple en un temps raisonnable. L'objectif étant de collecter de façon massive des données.

Adaptabilité. Il doit s'adapter à divers formats d'images (quelle que soit l'extension, la compression, la résolution de l'image) et à divers contenus (présence de labels, fleuves, ombres portées, position et format de la légende).

Performance et précision. Il doit restituer fidèlement les données et corriger ou lisser le bruit pour assurer la précision des informations extraites. Dans ce contexte, le bruit fait référence à une dégradation de l'image, une décoloration due aux différents traitements antérieurs de l'image.

Pour atteindre ces objectifs, la démarche générale suit plusieurs étapes clés :

1. Choix des images : La sélection des images est cruciale pour garantir la qualité des données extraites. Les images doivent être claires, de bonne résolution, et bien légendées. Les cartes doivent couvrir les zones géographiques d'intérêt et inclure des informations utiles pour la modélisation des risques.
2. Analyse et correction colorimétrique : L'analyse colorimétrique implique l'utilisation de techniques de clustering comme les K-Means pour regrouper les pixels par couleur. Cette étape permet de séparer les zones géographiques des éléments de la légende. Des corrections colorimétriques sont appliquées pour standardiser les couleurs et améliorer la distinction des zones.
3. Sélection d'une carte de référence : Une carte de référence est choisie pour servir de base à la normalisation des coordonnées géographiques. Une base de données des codes postaux est créée, associant chaque commune à ses coordonnées (x, y) sur la carte de référence. Cette base de données sert à traiter les autres cartes utilisées.
4. Algorithme de redimensionnement : L'algorithme de redimensionnement utilise des transformations géométriques pour aligner les coordonnées des cartes étudiées sur celles de la carte de référence. Des techniques comme les transformations affines sont employées pour ajuster les proportions et les échelles.
5. Scan de l'image et constitution de la base de données : Le scan des images redimensionnées permet d'extraire les couleurs associées à chaque coordonnée de commune. Les couleurs sont comparées à celles de la légende pour identifier les zones géographiques et la légende correspondante à chaque zone. Ces informations sont stockées dans une base de données structurée, prête à être utilisée pour la modélisation des risques.
6. Analyse des performances : L'analyse des performances consiste à comparer les données extraites avec des données de référence pour évaluer la précision de l'outil. Des métriques comme le taux de classifications correctes sont utilisées pour mesurer l'efficacité

de l'extraction. Les résultats sont analysés pour identifier les améliorations potentielles et optimiser l'algorithme.
Ces étapes expliquent la conception de l'outil, son fonctionnement est résumé par la figure 2.2.

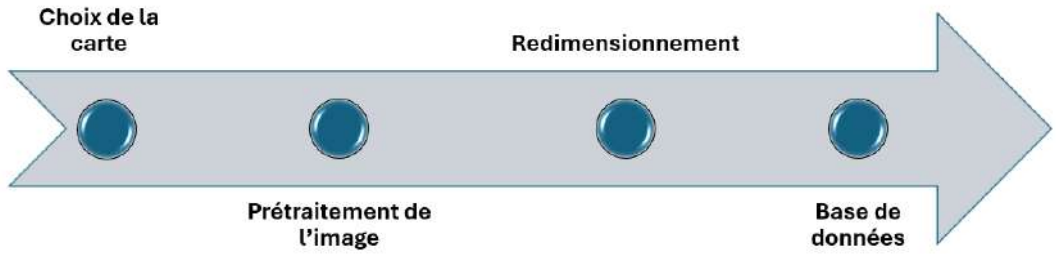


FIGURE 2.2 – Fonctionnement de l'outil algorithmique

En répondant aux contraintes de rapidité, simplicité, adaptabilité, performance et précision, cet outil permettra de collecter massivement des données de qualité, essentielles pour affiner les modèles de sinistralité et améliorer la précision des prévisions.

2.2 Initialisation

Les travaux de ce mémoire visent à traiter des images dans le but d'obtenir des données géographiques. Il est donc crucial de comprendre les spécificités des cartes utilisées, comment elles sont sélectionnées, et le choix des outils informatiques utilisés pour effectuer ces travaux de manière efficace et précise.
Les cartes sélectionnées pour ce projet doivent répondre à plusieurs critères spécifiques pour garantir la qualité et l'exactitude des données extraites, dont le premier est pour rappel, l'accord du service juridique qui en assure la légalité.

Données utiles pour la tarification : Les cartes doivent contenir des informations pertinentes pour la tarification en assurance habitation, telles que les zones de risque d'inondation, les zones à risque de sécheresse, etc. Ces informations sont essentielles pour affiner les modèles de sinistralité. Ce critère doit permettre une première sélection des cartes, mais il ne faut pas excessivement limiter le choix uniquement à celles que l'on juge utiles à priori. Les algorithmes peuvent ensuite évaluer la pertinence des données contenues dans chaque carte et déterminer si elles sont significatives.

Cartes open source et base de données indisponible : Il est important de mentionner que les cartes proviennent d'open data (voir les références de bibliographie dans les légendes des cartes) telles que Data.gouv.fr, IGN, Météo France et d'autres. Cependant, les jeux de données associés sont indisponibles, limitant les ressources utilisables qu'aux cartes. De plus, l'outil développé permet d'utiliser une maille géographique très fine au choix (code postal ou INSEE) qui n'est pas nécessairement disponible dans les

formats standard des *open data*, offrant ainsi une meilleure granularité pour la modélisation des risques en assurance.

Légende avec échelle colorimétrique : Les cartes doivent inclure une légende détaillée avec une échelle colorimétrique claire. Cette échelle est essentielle pour différencier les zones géographiques et attribuer correctement la légende correspondante. Il est possible de développer un outil algorithmique capable de traiter d'autres types de légende, mais ces travaux se limiteront aux cartes avec une légende à échelle colorimétrique.

Carte de la France : Les cartes doivent couvrir le territoire français de manière exhaustive, incluant toutes les communes. Cela permet d'assurer que toutes les zones d'intérêt pour la modélisation des risques soient incluses. Ces travaux se limitent à la France métropolitaine (à l'exception de la Corse), mais il est également possible de développer un outil spécifique à la Corse et aux autres territoires.
Les deux cartes ci-dessous sont données à titre d'exemple de cartes respectant les critères susmentionnés, respectivement pour les garanties inondation et dégâts des eaux.

FIGURE 2.3 – Nombre de cas d'inondations[12] (gauche) et Cumul des précipitations[18] (droite)

La sélection des cartes et le choix du langage de programmation sont des étapes cruciales dans le développement de l'outil d'extraction de données géographiques. Les cartes doivent répondre à des critères précis pour garantir la qualité des données extraites, et Python est choisi pour sa robustesse et sa flexibilité dans le traitement d'images. Ces choix sont fondamentaux pour atteindre l'objectif du mémoire.

Pour la suite des travaux de ce chapitre, et dans un souci d'illustration, nous considérerons les quatre cartes A, B, C et D ci-dessous. Ces cartes représentent respectivement : une carte des gonflements de sols argileux, une carte de densité de la population, une carte de production annuelle d'électricité et une carte du risque de sécheresse. En ce qui concerne les deux premières cartes (A et B), les bases de données d'origine sont connues, ce qui simplifiera l'analyse des résultats, contrairement aux deux dernières cartes (C et D).

Il est important de rappeler que ces cartes ont été choisies principalement pour dé-

montrer les performances et les limites intrinsèques de l'outil. D'autres cartes seront utilisées dans le cas pratique pour établir le lien avec la modélisation de la sinistralité. Cette sélection vise à fournir une évaluation complète de l'outil, en mettant en lumière ses capacités à traiter différents types de cartes et à en extraire fidèlement des informations.

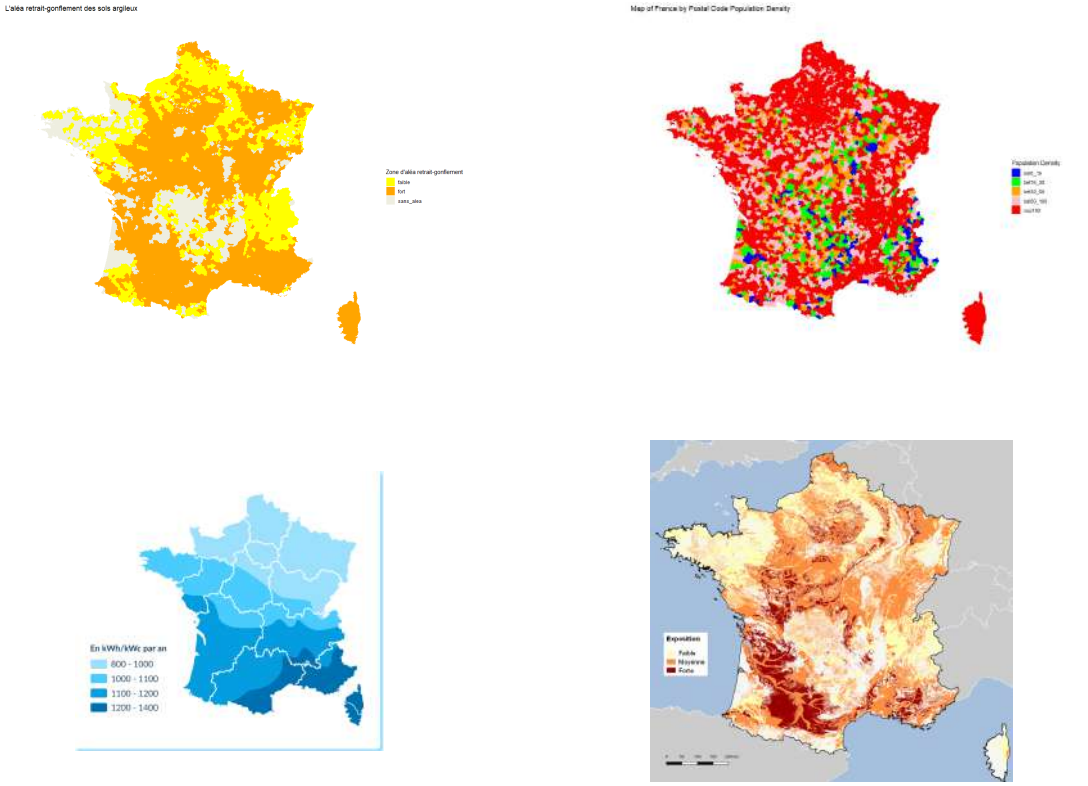


FIGURE 2.4 – Cartes A[25], B[6], C[30] et D[54] (de gauche à droite)

2.3 Prétraitement des images

Le prétraitement des images est une étape essentielle dans le processus d'extraction de données géographiques à partir de cartes. Cette étape vise à améliorer la qualité des images et à préparer les données pour une analyse plus précise et fiable. En assurant une préparation adéquate des images, on peut garantir que les données extraites sont à la fois fidèles et exploitables pour les analyses ultérieures.

Chaque carte est unique et comporte souvent des éléments qui ne doivent pas être pris en compte, tels que les fleuves, les frontières, ou les noms de régions et de villes. De plus, les images ne se limitent pas toujours aux contours de la France ; elles incluent parfois la mer, des pays voisins, ou encore des échelles et légendes qui peuvent interférer avec l'analyse, comme on peut le voir sur la carte D de la figure 2.4. Il est également rare que les couleurs utilisées dans la légende d'une carte correspondent exactement à celles visibles sur la carte elle-même. Il peut y avoir des variations subtiles, si bien que deux pixels semblant identiques à l'œil nu peuvent en réalité différer.

À la lumière de ces observations, le prétraitement des images devient indispensable. Il comprend plusieurs opérations clés, comme la séparation des couleurs de la légende de celles présentes sur le reste de la carte (par exemple, le bleu de la mer sur la carte D de la figure 2.4) et la réduction du bruit visuel. Ces étapes sont cruciales pour garantir une interprétation correcte des informations géographiques et éliminer les artefacts visuels susceptibles d'interférer avec l'extraction des données. En isolant les couleurs de la légende des autres éléments de la carte, on peut identifier avec précision les pixels représentant la France et ses environs, ce qui est nécessaire pour géolocaliser les informations pertinentes sur la carte.

Dissocier les couleurs de la légende de celles du reste de la carte. Il faut s'assurer que chaque couleur sur la carte est correctement associée à la catégorie de données qu'elle représente dans la légende. En effet, pour fluidifier l'algorithme de redimensionnement évoqué lors de la présentation de la démarche générale (et qui sera développé par la suite), il faudra uniformiser le hors légende avec une seule couleur. De plus, il est inévitable de connaître la couleur de chaque pixel pour pouvoir la rattacher à une couleur de la légende. Ainsi, pour une commune donnée, à partir du pixel qui la représente sur l'image, il sera possible de savoir à quelle catégorie de la légende, elle est associée.

Réduire le bruit des images. Le bruit sur les images fait référence aux variations de couleurs des pixels ou toutes autres déformations non désirées qui peuvent dégrader la qualité de l'image. Cela inclut des artefacts visuels comme les taches, les distorsions et les variations de couleur non intentionnelles. Réduire le bruit d'une image revient alors à améliorer la qualité de l'image pour faciliter une extraction de données plus précise. Cela permet d'obtenir des données plus fidèles et d'augmenter la fiabilité des analyses ultérieures.

2.3.1 Clustering colorimétrique

Le clustering colorimétrique est une technique utilisée pour segmenter les pixels d'une image en groupes (*clusters*) basés sur leurs valeurs de couleur. Cette technique est particulièrement utile pour :

— distinguer les couleurs de la légende de celles du reste de la carte ;
— distinguer les couleurs au sein de la légende ;
— éliminer le bruit ;
—

Avant d'aborder les méthodes pour réaliser un clustering colorimétrique, il est important de comprendre ce qu'est la valeur d'une couleur. La valeur d'une couleur permet de quantifier une couleur à partir d'une échelle de couleur. Il s'agit d'un système de représentation des couleurs qui permet de décrire les différentes teintes, saturations et luminosités de manière standardisée. Les échelles de couleur sont utilisées dans divers domaines, y compris le traitement d'images, la cartographie et les graphiques de données, pour représenter visuellement des informations quantitatives.

Système RGB. Le système RGB (Red, Green, Blue) est un modèle de couleur additif dans lequel les couleurs sont créées par la combinaison de différentes intensités de lumière rouge, verte et bleue. C'est le modèle de couleur le plus couramment utilisé dans les dispositifs électroniques, comme les écrans d'ordinateur, de télévision et les appareils photo numériques. Chaque couleur est définie par une combinaison de trois valeurs (R, G, B), chacune variant de 0 à 255, nommées coordonnées RGB ou tout simplement RGB. À titre d'exemple, la couleur rouge a pour RGB (255,0,0), la couleur verte (0,255,0) et la couleur blanche (255,255,255). C'est le système qui sera utilisé pour les travaux de ce mémoire.

Système HSV. Le système HSV (Hue, Saturation, Value) est un modèle de couleur qui représente les couleurs de manière plus intuitive pour les humains en séparant la teinte (Hue), la saturation (Saturation), et la valeur (Value ou luminosité). La teinte correspond à la couleur proprement dite et est mesurée en degrés sur une roue de couleur (de 0 à 360 degrés), la saturation indique l'intensité ou la pureté de la couleur (de 0 % à 100 %) et la valeur représente la luminosité de la couleur (de 0 % à 100 %). Cependant, les distances entre les couleurs dans l'espace HSV ne reflètent pas toujours les différences perçues de manière cohérente, ce qui peut fausser les résultats lors du clustering. Dans le cadre de ce mémoire, où la représentation des couleurs doit être simple et efficace pour traiter des images de cartes, le système RGB s'avère mieux adapté.

Système CIELAB. Le système CIELAB, ou Lab*, est un modèle de couleur conçu pour refléter la vision humaine. Il utilise trois axes : L* pour la luminosité (de 0 à 100), a* pour la teinte entre rouge et vert, et b* pour la teinte entre jaune et bleu. Contrairement à d'autres modèles, CIELAB est indépendant des dispositifs, garantissant des couleurs uniformes quel que soit l'appareil utilisé. Cependant, bien que ce système soit très précis, il est plus complexe à utiliser et nécessite des calculs plus lourds que le système RGB.

Cette complexité peut ralentir le traitement d'images, surtout lorsque l'on doit analyser rapidement de grandes quantités de données, comme c'est le cas dans ce mémoire. De plus, pour des images géographiques où l'objectif est d'extraire des informations simples à partir de cartes, le niveau de précision offert par CIELAB n'apporte pas de bénéfices significatifs par rapport à la simplicité du système RGB.

Nuances de gris. Les nuances de gris représentent les couleurs sans aucune information de teinte, se basant uniquement sur les variations de luminosité. Chaque pixel est défini par une valeur unique qui représente l'intensité lumineuse, allant du noir (0) au blanc (255). Cette échelle de couleurs est souvent utilisée pour les images en niveaux de gris, permettant de se concentrer uniquement sur les détails de luminosité. Elle n'est donc pas totalement adaptée à l'objectif de l'étude.

Le but principal du clustering colorimétrique est de regrouper les pixels ayant des couleurs similaires, facilitant ainsi l'analyse et le traitement ultérieur de l'image. Deux méthodes principales sont utilisées dans ce mémoire :

1. Méthode K-Means : Cette méthode se base sur l'algorithme des K-Means pour la distinction des couleurs.
2. Méthode ACP : Cette méthode se base, quant à elle, sur l'analyse en composantes principales pour la distinction des couleurs.

Le choix d'utiliser la méthode du K-means ou l'ACP pour réaliser le clustering colorimétrique dépend des caractéristiques de la carte étudiée. Tout au long de notre étude, pour chaque carte en entrée, les 2 méthodes ont systématiquement été lancées et la meilleure des 2 cartes *nettoyées* étaient gardées pour l'étape suivante.
Lors de nos utilisations successives de l'outils, l'ACP a été plus souvent retenue que le K-means du fait notamment de la flexibilité qu'il laisse à l'utilisateur d'ajuster manuellement ses clusters.

Toutefois, il apparaît donc nécessaire de présenter ces deux outils statistiques (K-means et ACP) afin de comprendre comment les utiliser pour réaliser un clustering colorimétrique.

a - Algorithme K-Means

L'algorithme des K-Means est une méthode populaire de clustering utilisée pour partitionner un ensemble de données en K groupes (clusters) distincts, où chaque point appartient au cluster avec le centroïde (le barycentre des points du cluster, appelé également "moyenne") le plus proche. La distance utilisée pour déterminer cette proximité est souvent la distance euclidienne, bien que d'autres métriques telles que la distance de Manhattan puissent être utilisées en fonction des spécificités des données. L'objectif de K-Means est de minimiser la variance intra-cluster, c'est-à-dire de rendre chaque cluster aussi homogène que possible, tout en maximisant la variance inter-clusters, ce qui signifie que les clusters sont bien séparés. L'algorithme est itératif et suit les étapes suivantes :

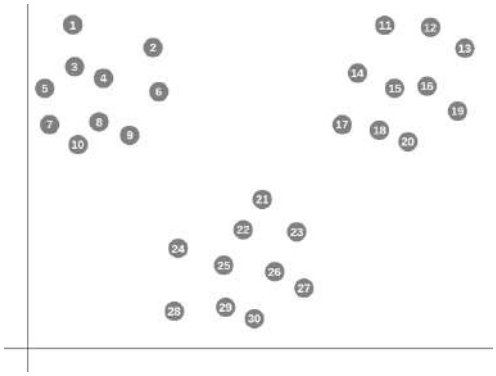


FIGURE 2.5 – K-Means - Echantillon[23]

On considère cet échantillon de points ci-dessus pour illustrer les étapes de l'algorithme.

1. Initialisation : Choisir K points initiaux comme centroïdes. Ces points peuvent être sélectionnés aléatoirement ou selon des méthodes plus sophistiquées comme K-Means++ pour améliorer la convergence. K-Means++ commence par sélectionner un premier centroïde de manière aléatoire, puis choisit les suivants en maximisant la distance par rapport aux centroïdes déjà sélectionnés. Pour notre exemple, K=3 avec A, B et C pour centroïdes.

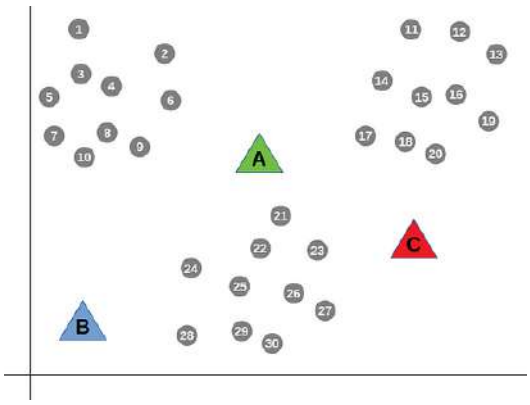


FIGURE 2.6 – K-Means - Initiation[23]

2. Attribution / affectation : Déterminer la distance entre chaque point et chaque centroïde et assigner chacun de ces points de données au centroïde le plus proche, formant K clusters. Ci-dessous, le point 21 est associé au cluster de centroïde A.
3. Mise à jour : Recalculer les centroïdes de chaque cluster en prenant la moyenne des points affectés à ce cluster :

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \tag{2.1}$$

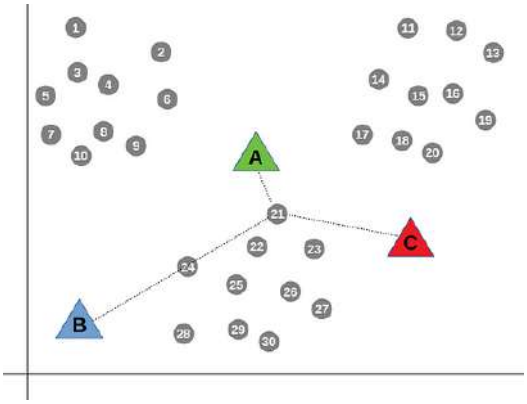


FIGURE 2.7 – K-Means - Attribution[23]

où :
— μ_j est le nouveau centroïde du cluster C_j,
— $|C_j|$ est le nombre de points dans le cluster C_j,
— x_i sont les points appartenant au cluster C_j.

4. Itération : Répéter les étapes d'affectation et de mise à jour jusqu'à ce que les centroïdes ne changent plus (ou changent très peu) ou jusqu'à ce qu'un nombre maximum d'itérations soit atteint.

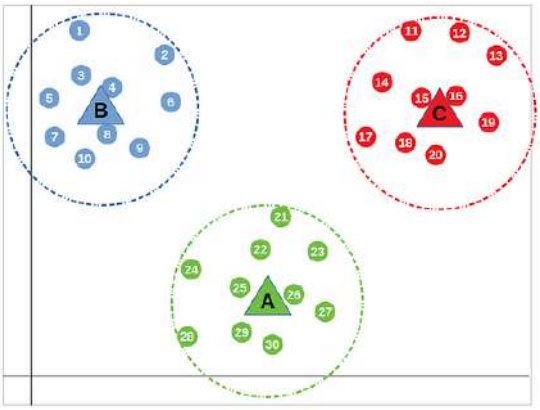


FIGURE 2.8 – K-Means - Résultats[23]

L'algorithme des K-Means présente plusieurs avantages et inconvénients. Parmi les avantages, on note sa simplicité et sa facilité d'implémentation, ce qui en fait un outil accessible même pour les débutants. De plus, il est rapide et efficace pour les ensembles de données de taille modérée, et convient bien aux données sphériques avec des clusters de forme circulaire. Cependant, K-Means a aussi ses limites. Il est sensible aux choix des centroïdes initiaux, pouvant conduire à des solutions locales plutôt qu'optimales. Il ne

fonctionne pas bien avec des clusters de tailles et de densités très différentes, et les clusters doivent idéalement être convexes pour des résultats optimaux. Enfin, il est également possible d'améliorer la qualité du clustering en utilisant des critères tels que le critère de Ward, qui minimise la variance intra-cluster en utilisant une approche hiérarchique basée sur la minimisation de la somme des carrés des distances entre les points d'un même cluster. Ce critère peut être utilisé en combinaison avec K-Means pour affiner la segmentation des données.
En dépit de ces inconvénients, l'algorithme reste largement utilisé en raison de sa rapidité et de son efficacité dans de nombreux cas pratiques.

b - Analyse en composantes principales (ACP)

L'Analyse en Composantes Principales (ACP) est une méthode statistique de réduction de la dimensionnalité utilisée pour transformer un ensemble de variables corrélées en un ensemble de variables non corrélées appelées composantes principales. Cette technique est largement utilisée pour simplifier les ensembles de données tout en conservant le maximum de variance possible. L'objectif principal de l'ACP est de réduire la complexité des données tout en minimisant la perte d'information. Cela se fait en transformant les variables d'origine en un nouvel ensemble de variables orthogonales (c'est-à-dire non corrélées) ordonnées de manière à capturer la plus grande partie possible de la variance totale des données.
Pour mieux comprendre l'intuition de l'ACP, il faut se dire qu'un ensemble de données à plus de deux variables est impossible à représenter dans le plan. L'ACP permet de réduire le nombre de variables (à deux en général) en perdant le minimum d'informations possible par rapport à l'ensemble complet des variables, afin de pouvoir représenter les données de plan. L'analyse en composantes principales se fait également en plusieurs étapes :

1. Standardisation des données : Les données sont standardisées pour avoir une moyenne nulle et une variance unitaire. Cela est crucial lorsque les variables d'origine ont des unités de mesure différentes.
2. Calcul de la matrice de covariance : Une matrice de covariance est calculée pour comprendre comment les variables varient ensemble.
3. Calcul des valeurs propres et des vecteurs propres : Les valeurs propres (qui représentent la quantité de variance capturée par chaque composante) et les vecteurs propres (qui définissent la direction de chaque composante principale) sont calculés à partir de la matrice de covariance.
4. Sélection des composantes principales : Les composantes principales sont triées par ordre décroissant de leurs valeurs propres. Les premières composantes principales, qui capturent le plus de variance, sont sélectionnées pour réduire la dimensionnalité des données.
5. Projection des données : Les données d'origine sont projetées sur les composantes principales sélectionnées pour obtenir un ensemble de données réduites en dimension.

Dans la figure 2.9 ci-dessous, on considère un jeu de données avec 12 variables que sont les températures moyennes par mois en France et par villes (qui sont les individus). La

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Angouleme	4.2	4.9	7.9	10.4	13.6	17.0	18.7	18.4	16.1	11.7	7.6	4.9
Angers	4.6	5.4	8.9	11.3	14.5	17.2	19.5	19.4	16.9	12.5	8.1	5.3
Besancon	1.1	2.2	6.4	9.7	13.6	16.9	18.7	18.3	15.5	10.4	5.7	2.0
Biarritz	7.6	8.0	10.8	12.0	14.7	17.8	19.7	19.9	18.5	14.8	10.9	8.2
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21.0	18.6	13.8	9.1	6.2

FIGURE 2.9 – ACP - Données illustratives[35]

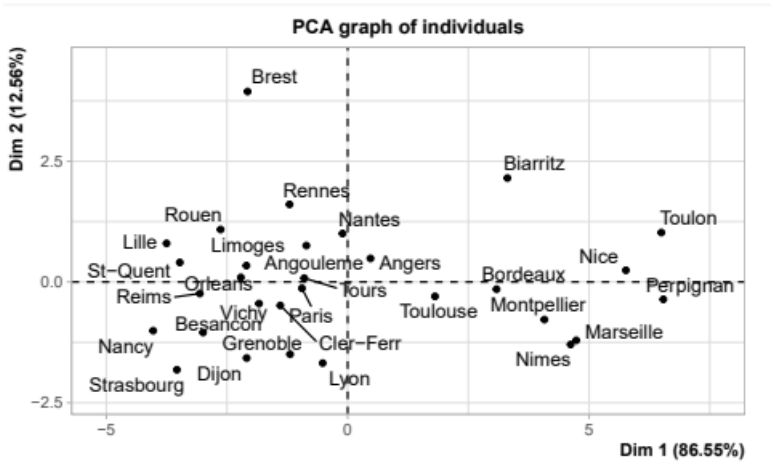


FIGURE 2.10 – ACP - Représentation dans le plan principal[35]

figure 2.10 montre la représentation des individus (villes) dans le premier plan principal. Il n'y a plus 12 variables, mais deux axes principaux qui contiennent respectivement 86,55% et 12,56% de l'information totale contenue dans l'ensemble des 12 variables de départ. L'enjeu sera ensuite d'interpréter ces deux axes principaux.

L'Analyse en Composantes Principales (ACP) présente plusieurs avantages et inconvénients. Parmi les avantages, l'ACP permet de réduire la dimensionnalité des ensembles de données complexes, simplifiant ainsi leur analyse tout en conservant, dans certains cas, une grande partie de la variance initiale (99,11% dans l'exemple précédent). Elle aide à identifier des motifs cachés et des structures dans les données, facilitant leur visualisation dans des espaces de moindre dimension comme la 2D ou la 3D. Cependant, l'ACP présente aussi des inconvénients, notamment la difficulté d'interpréter les composantes principales, car elles sont des combinaisons linéaires des variables d'origine, ce qui rend leur signification moins claire. De plus, une partie de la variance est perdue, surtout si seules quelques composantes principales sont retenues. Enfin, l'ACP ne capture que les relations linéaires entre les variables, ce qui peut limiter son efficacité pour des données présentant des relations non linéaires. Malgré ces limitations, l'ACP reste un outil puis-

sant pour l'analyse exploratoire et la visualisation des données.

Les deux principaux outils statistiques utilisés pour le clustering colorimétrique étant présentés, la prochaine étape est de mettre en exergue leur utilisation afin d'identifier les couleurs de la légende sur une image et de réduire le bruit. En outre, pour bien comprendre la nature des données sur lesquelles est appliqué l'algorithme, il faut visualiser une carte comme une base de données de pixels où les variables sont les coordonnées RGB :

N° Pixel	Red	Green	Blue
0	255	255	255
1	255	255	255
2	255	255	255
3	255	255	255
⋮	⋮	⋮	⋮
747995	255	255	255

TABLE 2.1: Carte C sous forme de base de données

c - Méthode K-Means

On considère les cartes A, B, C et D présentées dans la section 2.2 Initialisation. Il est relativement simple, en regardant les cartes, de connaître le nombre de couleurs appartenant à la légende ou pas. Les figures ci-dessous illustrent les centroïdes obtenus en appliquant l'algorithme de K-Means sur les quatre cartes.

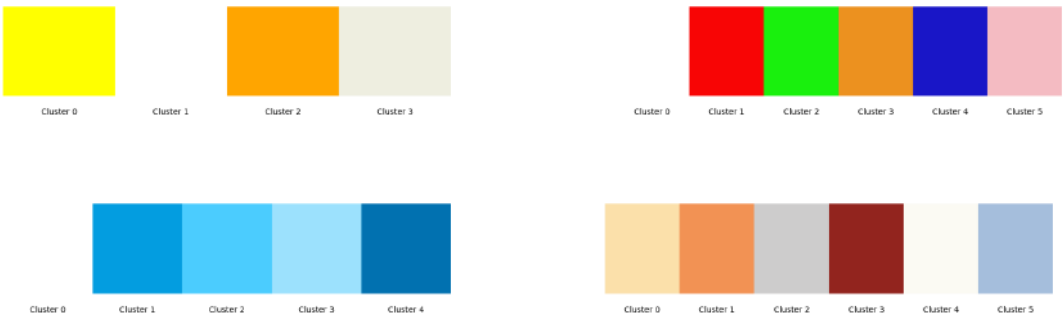


FIGURE 2.11 – Centroïdes — cartes A, B, C et D

Les observations montrent que le hors-légende est de couleur uniforme (très proche du blanc) sur les trois premières cartes (il suffit de rogner la carte C pour faire disparaître le contour bleu) et que les couleurs de la légende ont été correctement détectées par l'algorithme des K-Means.
En ce qui concerne la carte D, le hors-légende est principalement composé de trois couleurs représentant la mer, les autres pays, et une zone blanche présente en France, mais absente de la légende. Il est supposé que cette zone blanche correspond à des zones sans risques ou sans information. Pour simplifier le traitement, une couleur unique (le blanc) est attribuée

au hors-légende de la carte D. Sur Python, cela revient à blanchir tous les pixels de l'image appartenant aux clusters dont les centres ont été identifiés comme ayant une couleur hors légende. De plus, de manière pratique, une des couleurs du hors-légende est utilisée pour éliminer les informations superflues (le cadre, la légende, la Corse, etc.) avant d'appliquer l'algorithme K-Means.
Il est important de comprendre que cette approche est limitée, l'algorithme autorise le choix de la valeur de K, le nombre de clusters, mais ne permet pas de choisir les centroïdes. Des couleurs précises (celles de la légende et du hors légende) peuvent être choisies comme centroïdes initiaux, mais si une de ces couleurs est très peu présente sur l'image, l'algorithme peut ne pas la garder comme centroïde finale. C'est le cas de la carte E dans la figure 2.3.1 ci-dessous. Sur la carte originale, la légende indique une couleur blanche correspondant aux zones présentant une absence de sols (qu'on peut retrouver au niveau de l'Île-de-France). Pourtant, cette couleur étant très minoritaire, l'algorithme des K-Means ne la choisit pas pour être un centroïde final, mais elle est plutôt absorbée par le cluster 0 (qui représente les autres pays) dont le centroïde a une couleur proche du blanc, très présente sur la carte. Étant donné l'absence d'une couleur de la légende dans la liste des couleurs des centroïdes finaux et le fait que le nombre de clusters a été défini à l'avance, il existe une couleur hybride suffisamment présente sur la carte pour être choisie par l'algorithme (le cluster 6) à la place de la couleur de la légende manquante. L'existence de cette couleur sur la carte illustre également la présence de bruits sur l'image, une dégradation de certaines couleurs à plusieurs endroits, causée probablement par le style d'ombre autour de la France sur la carte originale. La méthode K-Means n'est donc pas adaptée à cette carte.

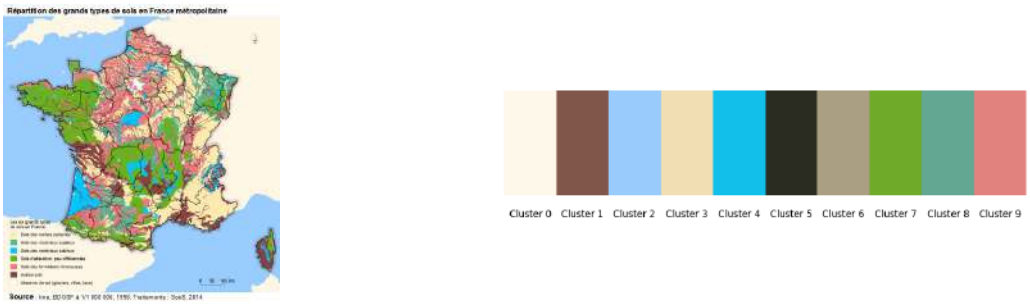


FIGURE 2.12 – Carte E[40] et ses centroïdes

L'algorithme K-Means a déterminé les couleurs de la légende et celles du hors-légende, ainsi que des clusters qui regroupent tous les pixels en fonction de leurs couleurs. Les cartes peuvent donc être reconstituées en indiquant à l'algorithme que chaque pixel doit prendre la couleur du centroïde du cluster auquel il a été assigné. En effet, dans un cluster, tous les pixels n'ont pas obligatoirement la même couleur, mais si un pixel est associé au cluster 0, par exemple, cela signifie que la couleur de ce pixel est plus proche de celle du centroïde du cluster 0 que des couleurs des centroïdes des autres clusters. Normalement, pour une carte sans bruit, tous les pixels d'un cluster doivent avoir la même couleur, mais ce n'est pas toujours le cas.

FIGURE 2.13 – K-Means — cartes A et B reconstituées

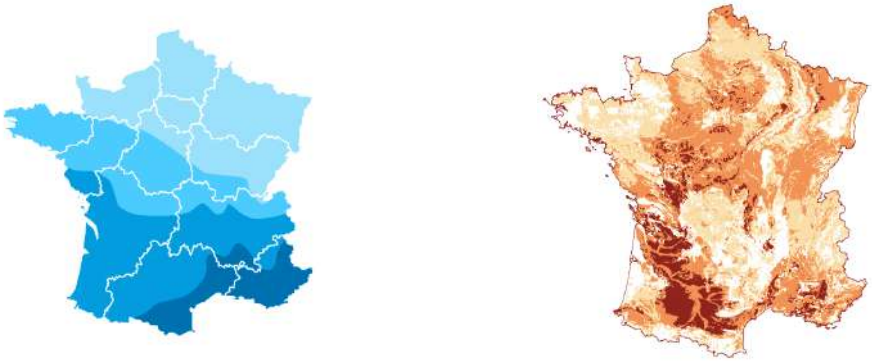


FIGURE 2.14 – K-Means — cartes C et D reconstituées

d - Méthode ACP

Les cartes A, B, C et D de la section 2.2 Initialisation, ainsi que leur format de base de données, sont toujours considérées. Par définition, l'Analyse en Composantes Principales (ACP) permet de passer d'un espace à trois dimensions (RGB) à un plan dans lequel il

est possible d'afficher tous les pixels de l'image. Cet affichage en deux dimensions permet de détecter des motifs et des groupes de pixels homogènes. En effet, dans la base de données, un pixel n'est défini que par sa couleur. Par conséquent, les pixels de couleurs similaires doivent se regrouper dans le plan principal de l'ACP.

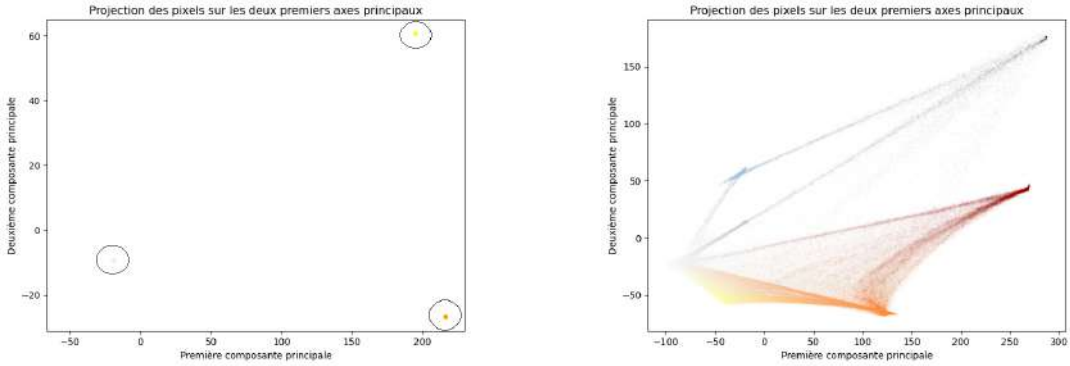


FIGURE 2.15 – Cartes A et D - Représentation des pixels dans le plan principal

Pour la carte A, il n'y a que quatre points (trois visibles sur la figure 2.15 de gauche et un quatrième de couleur blanche), ce qui signifie que les points représentant les pixels de même couleur sont confondus. Il en découle qu'il y a exactement quatre couleurs sur l'image, indiquant une carte parfaitement sans bruit.

Le cas de la carte D est différent. Un grand nombre de points sont présents et les pixels qui semblent avoir des couleurs identiques ne sont pas représentés au même endroit. Cela s'explique par le fait que ces pixels n'ont pas exactement la même couleur (au moins une coordonnée RGB est différente d'un pixel à l'autre), rendant la carte très bruitée. Néanmoins, il est possible de repérer et de former des groupes de pixels de couleurs similaires.
Pour ce faire, des points fictifs, représentant les couleurs originales de la légende, sont placés sur le plan. Si l'image était totalement dépourvue de bruit, les pixels correspondraient exactement (seraient confondus) à ces points fictifs, comme illustré sur la figure 2.15. Ensuite, des cercles sont tracés autour de ces points pour regrouper les pixels selon leur couleur. Par analogie avec la méthode K-Means, ces points fictifs jouent le rôle de centroïdes, et tous les pixels situés dans un cercle autour de chaque centroïde forment un cluster correspondant à cette couleur. Le rayon des cercles est modifié et ajusté manuellement en fonction du rendu visuel (voir figure 2.16).

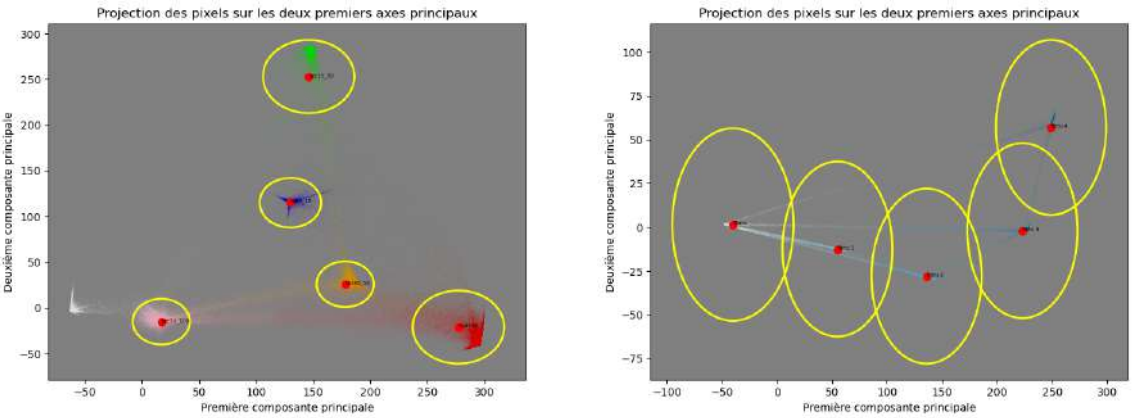


FIGURE 2.16 – Carte B et C - Représentation des clusters dans le plan principal

Sur la figure 2.16, les points fictifs sont représentés en rouge. Il est important de noter que lorsque deux cercles se croisent, un point appartenant à la fois aux deux cercles est associé au cercle dont le centre est le plus proche dans la représentation plane. De plus, lorsque plusieurs groupes de couleurs sont indissociables, ce qui peut arriver lorsque les centres de deux cercles sont très proches, on considère l'union des deux cercles, et chaque point est associé au cercle dont la couleur d'origine du centre est la plus proche de celle du point. En réalité, associer un point au cercle le plus proche dans le plan revient à l'associer au cercle dont la couleur du centre est la plus proche de la sienne. En effet, dans la base de données, un pixel est défini par ses coordonnées RGB et donc par sa couleur. L'ACP ne fait que projeter cet espace de couleurs en trois dimensions sur un plan. Par conséquent, la proximité entre les couleurs de deux pixels doit se refléter par une proximité équivalente dans le plan, c'est-à-dire que la distance entre les points représentant ces pixels dans le plan doit correspondre à la différence entre leurs couleurs respectives. Néanmoins, cette équivalence n'est pas totale puisque l'ACP ne projette pas toujours 100% de l'information contenue dans les variables d'origine. Cependant, on peut admettre cette équivalence lorsque le pourcentage de variance expliquée est proche de 100%. Il est par ailleurs tentant de simplement récupérer la base de données des RGB, de calculer la distance entre la couleur de chaque pixel et celles de la légende afin de former les clusters. Bien que cette tentative reste cohérente dans sa logique et engendrerait des clusters satisfaisants, elle serait malgré tout coûteuse en espace et en temps de calcul. En fait, il faudra calculer pour chaque pixel de l'image, la distance entre sa couleur et toutes les couleurs de la légende ; d'où la complexité algorithmique et le rejet de cette méthode au profit des cercles.

Après avoir formé les clusters, tout comme dans la méthode K-Means, il faut reconstituer les cartes en uniformisant les couleurs à l'intérieur des clusters. Toutefois, dans la méthode ACP, il peut y avoir des points qui ne sont associés à aucun cluster. En général, il s'agit de points considérés comme étant du bruit, car la couleur des pixels associés à ces points est incertaine. Pour ces pixels non classés, on leur attribue une couleur neutre, la

couleur du hors légende uniformisé. Dans la suite des travaux, on attribuera une couleur à ces pixels par un algorithme des K plus proches voisins, après l'étape de redimensionnement qui sera explicité dans la suite.

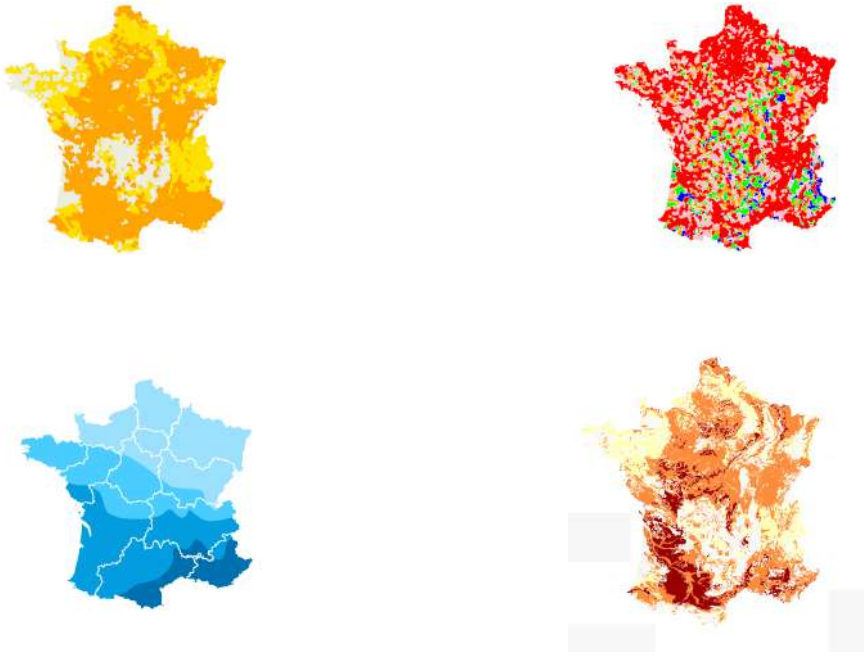


FIGURE 2.17 – ACP — cartes A, B, C et D reconstituées

Les cartes A et C ne semblent pas avoir subi de modifications. Sur les cartes B et D par contre, des trous sont apparus à l'intérieur de la France. Ces trous sont dus à des pixels bruités qui ont été retirés dans un souci de nettoyage de la carte, comme expliqué précédemment.

Interprétation de l'ACP. L'interprétation d'une ACP consiste à comprendre les relations entre les variables initiales et les nouvelles variables (composantes principales) créées par l'ACP, ainsi qu'à analyser la distribution des observations dans l'espace des composantes principales. Pour éviter de répéter les mêmes informations et fournir une analyse concise, l'interprétation se limitera aux résultats de l'ACP de la carte D.

Dimension	Variance expliquée	% variance expliquée	% variance expliquée cumulée
Dim1	1,95	65,0	65,0
Dim2	1,01	34,0	99,0
Dim3	0,03	1,0	100,0

TABLE 2.2 – Variance expliquée par chaque dimension — carte D

La table de variance expliquée montre que les deux premières composantes principales (Dim1 et Dim2) expliquent 99% de la variance totale des données. Cela signifie que les deux premières composantes capturent presque toute l'information présente dans les données d'origine, ce qui justifie l'utilisation d'un plan bidimensionnel.

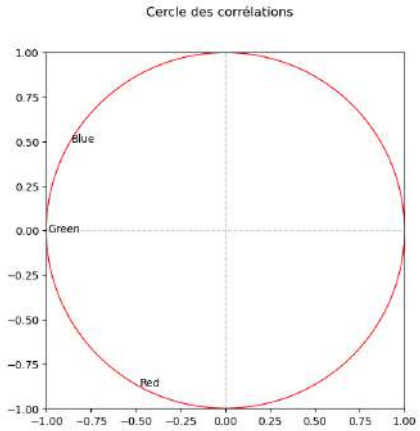


FIGURE 2.18 – Cercle de corrélation — carte D

Le cercle de corrélation montre la relation entre les variables originales (Red, Green, Blue) et les nouvelles composantes principales. Les canaux Green et Blue sont fortement anticorrélés avec Dim1, ce qui signifie que Dim1 capte principalement les variations dans ces deux canaux de couleur. Red est positionné de manière opposée à Dim2, indiquant une forte corrélation négative avec cette composante, ce qui signifie que les variations dans le canal rouge sont bien capturées par Dim2.
Ces résultats indiquent que les composantes Dim1 et Dim2 sont principalement influencées par les canaux de couleur rouge et bleu, respectivement.

Étant donné l'espace en 3 dimensions de couleurs, l'ACP projette cet espace sur un plan dans lequel chaque axe principal représente une échelle de couleurs :

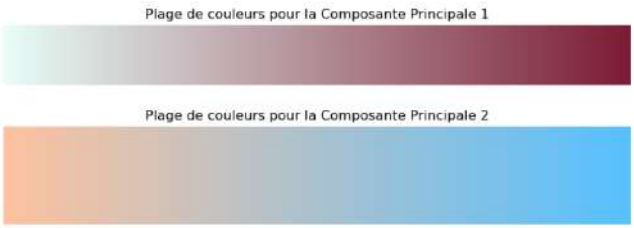


FIGURE 2.19 – Échelle de couleurs des axes principaux — carte D

Les plages de couleurs pour les composantes principales montrent comment les différentes nuances de couleurs se distribuent le long des axes. Dim1 présente une variation allant des couleurs avec des composantes Green (G) et Blue (B) élevées jusqu'aux couleurs avec des composantes Green et Blue faibles (donc principalement rouge). En revanche, Dim2 montre une évolution des couleurs où la composante Red (R) est dominante jusqu'aux couleurs où la composante rouge est absente. Cette distribution indique que Dim1 capture principalement les variations dans les canaux Green et Blue, tandis que Dim2 est fortement influencée par les variations dans le canal Red. Cela correspond à la corrélation observée grâce au cercle, où Green et Blue sont anticorrélés avec Dim1, et Red est anticorrélé avec Dim2.

2.3.2 Autres approches

Dans le cadre de ces travaux, plusieurs autres approches ont été envisagées et testées pour réaliser ce clustering colorimétrique et réduire les bruits afin d'améliorer la qualité des images. Bien que ces méthodes n'aient pas été retenues dans la version définitive de l'outil, elles présentent un potentiel intéressant pour des recherches futures et méritent d'être explorées davantage.

Réseaux de neurones convolutifs. Les réseaux de neurones convolutifs (CNN) sont une technique avancée de traitement d'image qui utilise des *layers* pour extraire les caractéristiques visuelles des images. Cette approche a été étudiée pour reconnaître directement la France sur les images de cartes. Les CNN offrent une haute précision dans la reconnaissance des formes et des objets et sont capables d'apprendre et de détecter des caractéristiques complexes et non linéaires. Cependant, cette méthode nécessite un grand nombre de données annotées pour l'entraînement, est complexe à mettre en œuvre, et demande des ressources computationnelles importantes. De plus, elle est sensible aux variations dans les données d'entrée, comme les différences de résolutions et de qualités d'image. Pour le clustering colorimétrique, les CNN n'ont pas été retenus en raison de leur complexité et surtout de la difficulté à obtenir suffisamment de données annotées pour un entraînement efficace, ainsi que des ressources computationnelles nécessaires, ce qui n'était pas compatible avec les contraintes des travaux. La figure 2.20 ci-dessous montre la capacité du Mask R CNN, qui fait partie de la famille des CNN, à reconnaître divers objets/formes sur des images, telles que des voitures ou des vélos.

FIGURE 2.20 – Mask R CNN - Exemples d'utilisation [52]

Utilisation de filtres gaussiens. Les filtres gaussiens sont couramment utilisés en traitement d'image pour lisser les images et réduire le bruit. Ils appliquent une convolution entre l'image et un noyau gaussien, ce qui a pour effet de flouter les détails fins et d'atténuer les variations de pixels non souhaitées. Bien que cette méthode améliore la qualité de l'image en réduisant le bruit et soit simple à mettre en œuvre, elle entraîne également une perte de détails fins, ce qui peut affecter la précision de l'extraction des données géographiques, et s'avère inefficace sur des images fortement bruitées ou avec des contours complexes. Ainsi, pour le clustering colorimétrique, les filtres gaussiens n'ont pas été retenus, car ils compromettent la précision nécessaire pour distinguer les couleurs proches de manière fiable.

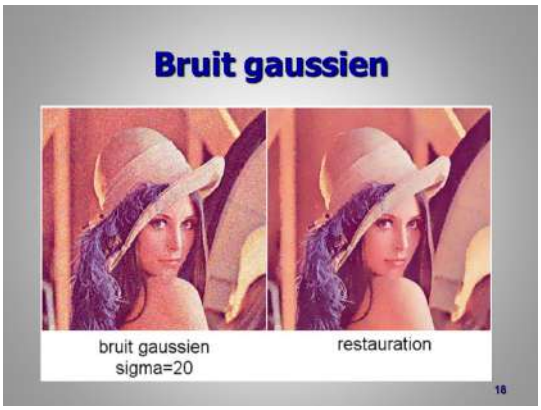


FIGURE 2.21 – Restauration par filtre gaussien[45]

2.4 Carte et base de données de référence

Pour pouvoir récupérer des informations sur plusieurs cartes, l'algorithme doit savoir où récupérer ces informations sur ces cartes. Dans le contexte de ce mémoire, l'algorithme doit déterminer où se trouve chaque commune (par son code postal ou son code INSEE) de la base de données, quelle que soit la carte. L'idée est donc d'aligner les positions des communes sur toutes les cartes avec celles d'une carte dite de référence, pour laquelle la position de chaque commune sur cette carte serait connue au pixel près. Pour ce faire, les autres cartes sont modifiées pour qu'elles aient la même forme que la carte de référence (voir section *2.5 Redimensionnement et scan des images*). Il faut également savoir qu'avec Python, une image est semblable à une matrice de pixels ; il est donc possible de connaître la position d'un pixel grâce à sa ligne et sa colonne dans la matrice, appelées dans la suite du mémoire "coordonnées en pixels".
Dans cette section, l'objectif est de créer une base de données de référence qui inclura la position de toutes les communes sur une carte sélectionnée comme référence. Le choix de cette carte de référence constitue donc la première étape du processus.

2.4.1 Principe

Premièrement, il est indispensable de choisir une carte de référence parmi toutes les cartes de la France disponible. Elle doit ainsi remplir certains critères : elle doit représenter l'intégralité de la France, inclure les positions et les noms des villes, et idéalement être conçue en utilisant la projection Lambert 93, qui est le système de projection officiel pour les représentations planes de la France (et qui sera détaillée ultérieurement). Ensuite, une base de données officielle des communes de France, disponible en open source sur `data.gouv.fr`, est utilisée. Elle contient les noms des communes, les codes postaux, les codes INSEE, ainsi que les latitudes et longitudes des villes. Enfin, les coordonnées en pixels de quelques villes sont récupérées manuellement pour entraîner un modèle de régression linéaire, permettant de déterminer la position précise de chaque commune sur la carte à partir de sa latitude et sa longitude.

2.4.2 Uniformisation des coordonnées

a - Choix de la carte de référence et base de données initiale

La carte de référence est importante, car elle sert de base pour toutes les autres cartes. Pour la suite des travaux, on considère comme carte de référence la figure 2.22 ci-dessous.

FIGURE 2.22 – Carte de référence[8]

Cette carte représente bien toute la France (les travaux de ce mémoire se limitant à la France métropolitaine sans la Corse) et les grandes villes. L'étape suivante est de former la base de données qui contiendra la position de toutes les communes. Pour cela, on récupère la base de données mentionnée au *2.4.1 Principe* comme base de données initiale :

Communes	Code INSEE	Code postal	Latitude	Longitude
ABERGEMENT CLEMENCIAT	01001	01400	46,151702	4,930600
ABERGEMENT DE VAREY	01002	01640	46,007131	5,424644
AMBERIEU EN BUGEY	01004	01500	45,957471	5,370568
AMBERIEUX EN DOMBES	01005	01330	45,999229	4,911872
AMBLEON	01006	01300	45,748314	5,592785
⋮	⋮	⋮	⋮	⋮
OUANGANI	97614	97670	-12,839254	45,138362
SADA	97616	97640	-12,864266	45,113440
TSINGONI	97617	97680	-12,782373	45,133400

TABLE 2.3 – Base de données initiale[19]

Les positions des communes sont recensées sur la carte de référence grâce aux coordonnées en pixels. Cependant, la base de données initiales localise chaque commune par ses coordonnées GPS (latitude et longitude). Il faut donc réaliser une correspondance entre les coordonnées GPS et les coordonnées en pixels sur la carte. Pour y parvenir, on récupère manuellement les coordonnées en pixels de quelques villes sur la carte afin d'entraîner un modèle de régression linéaire pour déterminer les coordonnées en pixels des autres villes. La régression linéaire est intuitivement choisie, car la représentation plane de la France est une projection à partir des latitudes et des longitudes. Il existe plusieurs types de projections, dont la projection Lambert 93, qui est la projection officielle pour la représentation plane de la France. Ainsi, deux régressions linéaires sont effectuées : la première tente de prédire les coordonnées en pixels directement à partir des latitudes et des longitudes, tandis que la deuxième approche convertit d'abord les latitudes et longitudes en coordonnées Lambert 93, puis prédit les coordonnées en pixels à partir de ces coordonnées Lambert 93.

b - Régression linéaire : première approche

Rappel sur la régression linéaire. La régression linéaire modélise la relation entre une variable dépendante Y et une ou plusieurs variables indépendantes $X_1, X_2, ..., X_p$. Le modèle est donné par :

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + W$$

où W représente l'erreur résiduelle. L'objectif est de minimiser la somme des carrés des écarts entre les valeurs observées et prédites. Pour que le modèle soit valide, certaines hypothèses doivent être respectées : linéarité, indépendance des résidus, homoscédasticité, normalité des résidus et absence de multicolinéarité.
Les coefficients de régression sont estimés en minimisant la somme des carrés des résidus :

$$\min_{\beta_0,\beta_1,...,\beta_p} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

où $\hat{Y}_i$ est la valeur prédite pour l'observation i. Les principaux indicateurs pour évaluer le modèle sont le coefficient de détermination R^2, la valeur p, et l'erreur standard des coefficients. La régression linéaire reste une méthode robuste pour modéliser des relations entre variables, à condition que les hypothèses soient respectées.

Modélisation et vérification des hypothèses.
La modélisation des coordonnées en pixels directement à partir de la latitude et de la longitude peut s'exprimer de la manière suivante :

$$Pixel_X = \beta_0 + \beta_1 Latitude + \beta_2 Longitude + W$$

$$Pixel_Y = \alpha_0 + \alpha_1 Latitude + \alpha_2 Longitude + Z$$

Les coordonnées en pixels de 15 villes sont récupérées manuellement, dont 10 sont utilisées pour l'entraînement du modèle et 5 pour le test. Bien que 10 codes postaux puissent sembler faible pour l'entraînement, ce choix s'explique par la complexité et le caractère chronophage de la collecte manuelle des coordonnées, toutefois les villes sélectionnées sont réparties de manière à "quadriller" toute la France. Cette répartition spatiale stratégique permet de capturer efficacement la diversité géographique du territoire, garantissant ainsi que le modèle représente correctement les variations locales. De plus, la forte corrélation entre les coordonnées géographiques et les positions en pixels renforce la capacité du modèle à généraliser à partir de cet échantillon ciblé, optimisant ainsi l'équilibre entre précision et faisabilité.

Communes	**INSEE**	**C. postal**	**Lat.**	**Long.**	**PixelX**	**PixelY**
LAVAULT STE ANNE	03140	03100	46,309297	2,6042	470	487
MONTLUÇON	03185	03100	46,342781	2,607941	470	487
TROYES	10387	10000	48,292395	4,076135	564	293
MARSEILLE 01	13201	13001	43,30018	5,382763	670	771
DIJON	21231	21000	47,33187	5,032219	629	387
QUIMPER	29232	29000	47,998163	-4,097206	31	311
TOULOUSE	31555	31000	43,600703	1,432843	392	749
BORDEAUX	33063	33000	44,862433	-0,584754	253	629
RENNES	35238	35000	48,115934	-1,688443	196	312
GRENOBLE	38185	38000	45,184234	5,71554	683	586
ST ETIENNE	42218	42000	45,424104	4,366467	595	567

TABLE 2.4 – Base de données d'entrainement

Communes	INSEE	C. postal	Lat.	Long.	PixelX	PixelY
DRUELLE BALSAC	12090	12000	44,379303	2,472372	470	675
LE MONASTÈRE	12146	12000	44,331477	2,589999	470	675
ONET LE CHÂTEAU	12176	12000	44,382374	2,561436	470	675
RODEZ	12202	12000	44,359078	2,569856	470	675
CAHORS	46042	46000	44,456511	1,439009	394	668
STRASBOURG	67482	67000	48,569074	7,762079	798	255
LYON 01	69381	69001	45,770104	4,826373	628	538
ANNECY	74010	74000	45,902355	6,126404	710	518

TABLE 2.5 – Base de données de test

Les figures ci-dessous permettent de vérifier l'hypothèse de linéarité entre la latitude et la variable pixelY, et entre la longitude et la variable pixelX.

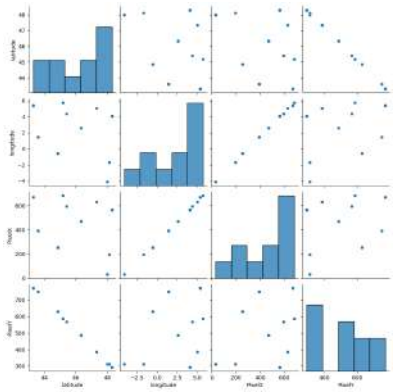


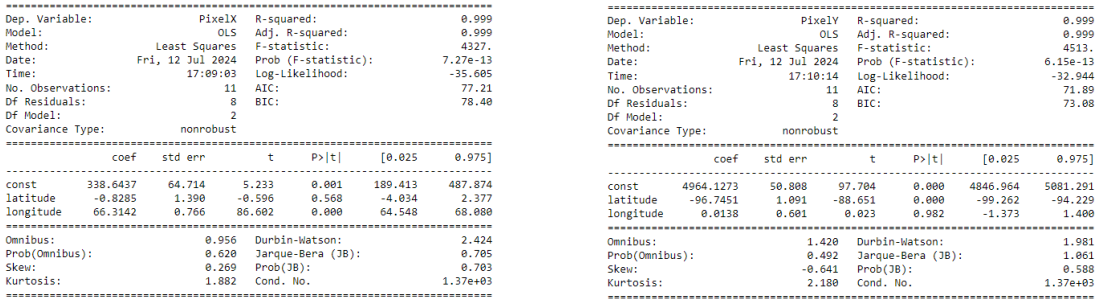
FIGURE 2.23 – Vérification de l'hypothèse de linéarité — approche n°1

Résultats et interprétations. L'objectif est de modéliser les coordonnées en pixels (PixelX et PixelY) en fonction de la latitude et de la longitude. L'expression finale du modèle est donc :

$$\text{PixelX} = 338.6437 + (-0.8285 \times \text{Latitude}) + (66.3142 \times \text{Longitude}) \tag{2.2}$$

$$\text{PixelY} = 4964.1273 + (-96.7451 \times \text{Latitude}) + (0.0138 \times \text{Longitude}) \tag{2.3}$$

Les résultats des deux régressions linéaires, pour PixelX et PixelY, sont présentés sur la figure 2.24 ci-après.



Dep. Variable:	PixelX	R-squared:	0.999
Model:	OLS	Adj. R-squared:	0.999
Method:	Least Squares	F-statistic:	4327.
Date:	Fri, 12 Jul 2024	Prob (F-statistic):	7.27e-13
Time:	17:09:03	Log-Likelihood:	-35.605
No. Observations:	11	AIC:	77.21
Df Residuals:	8	BIC:	78.40
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	338.6437	64.714	5.233	0.001	189.413	487.874
latitude	-0.8285	1.390	-0.596	0.568	-4.034	2.377
longitude	66.3142	0.766	86.602	0.000	64.548	68.080

Omnibus:	0.956	Durbin-Watson:	2.424
Prob(Omnibus):	0.620	Jarque-Bera (JB):	0.705
Skew:	0.269	Prob(JB):	0.703
Kurtosis:	1.882	Cond. No.	1.37e+03

Dep. Variable:	PixelY	R-squared:	0.999
Model:	OLS	Adj. R-squared:	0.999
Method:	Least Squares	F-statistic:	4513.
Date:	Fri, 12 Jul 2024	Prob (F-statistic):	6.15e-13
Time:	17:10:14	Log-Likelihood:	-32.944
No. Observations:	11	AIC:	71.89
Df Residuals:	8	BIC:	73.08
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4964.1273	50.808	97.704	0.000	4846.964	5081.291
latitude	-96.7451	1.091	-88.651	0.000	-99.262	-94.229
longitude	0.0138	0.601	0.023	0.982	-1.373	1.400

Omnibus:	1.420	Durbin-Watson:	1.981
Prob(Omnibus):	0.492	Jarque-Bera (JB):	1.061
Skew:	-0.641	Prob(JB):	0.588
Kurtosis:	2.180	Cond. No.	1.37e+03

FIGURE 2.24 – Résultats de l'approche n°1 - PixelX (gauche) et PixelY (droite)

Pour les modèles PixelX et PixelY, les R^2 et ${R_{adj}}^2$ très proches de 1 (0.999) montrent que les modèles expliquent presque toute la variance des coordonnées en pixels. Les interceptes sont significatifs ($p < 0.05$), indiquant une contribution importante aux modèles. Pour PixelX, la latitude n'est pas significative ($p = 0.568$), mais la longitude l'est fortement ($p = 0.000$). Pour PixelY par contre, la latitude a un impact significatif ($p = 0.000$), tandis que la longitude ne l'est pas ($p = 0.982$).
Les statistiques F élevées (4327.16 pour PixelX et 4513.49 pour PixelY) confirment la significativité globale des modèles. Les tests sur les résidus indiquent qu'ils suivent une distribution normale et sont indépendants, validant ainsi les modèles.
Ces résultats peuvent être utilisés pour comprendre comment les coordonnées géographiques se traduisent en coordonnées de pixels sur une carte.

c - Régression linéaire : seconde approche

Lambert 93.

Le système de coordonnées Lambert 93, ou RGF93 (Réseau Géodésique Français 1993), est une projection cartographique officielle utilisée en France. Il a été adopté pour remplacer le système Lambert II étendu et est basé sur le système géodésique européen ETRS89. Le système Lambert 93 est caractérisé par la projection conique conforme de Lambert, qui est adaptée pour les régions étendues en longitude, comme la France.
La projection Lambert 93 est particulièrement utile, car elle offre une représentation plane précise de la France, prenant en compte la courbure de la Terre pour minimiser les distorsions. Utiliser cette projection peut aider à obtenir des coordonnées en pixels plus précises pour les communes, améliorant ainsi l'uniformité et la fiabilité des cartes.
La méthodologie de projection, c'est-à-dire de conversion des latitudes et des longitudes en coordonnées Lambert 93, se base sur une théorie mathématique et sur des paramètres initiaux qui seront développés dans l'annexe A.

Modélisation et vérification des hypothèses.

Pour cette deuxième approche, on conserve les mêmes bases d'entraînement et de

test, mais on rajoute les coordonnées Lambert 93.

Communes	Lat.	Long.	XL93	YL93	PixelX	PixelY
LAVAULT STE ANNE	46,309297	2,6042	669 540,18	6 578 898	470	487
MONTLUÇON	46,342781	2,607941	669 846,51	6 582 615	470	487
TROYES	48,292395	4,076135	779 812,34	6 799 662	564	293
MARSEILLE 01	43,30018	5,382763	893 429,95	6 247 476	670	771
DIJON	47,33187	5,032219	853 455,54	6 694 368	629	387
QUIMPER	47,998163	-4,097206	171 387,38	6 790 192	31	311
TOULOUSE	43,600703	1,432843	573 431,92	6 279 217	392	749
BORDEAUX	44,862433	-0,584754	416 927,19	6 424 566	253	629
RENNES	48,115934	-1,688443	351 307,73	6 789 862	196	312
GRENOBLE	45,184234	5,71554	913 237,14	6 457 547	683	586
ST ETIENNE	45,424104	4,366467	806 856,34	6 481 444	595	567

TABLE 2.6 – Base de données d'entrainement — approche n°2

Résultats et interprétations.

L'objectif est de modéliser les coordonnées en pixels (PixelX et PixelY) en fonction des coordonnées en Lambert 93 nommées Xlambert et Ylambert. L'expression finale du modèle est donc :

$$\text{PixelX} = -24.8782 + (0.0009 \times \text{Xlambert}) + (-1.322e-05 \times \text{Ylambert}) \tag{2.4}$$

$$\text{PixelY} = 6207.7344 + (-1.917e-05 \times \text{Xlambert}) + (-0.0009 \times \text{Ylambert}) \tag{2.5}$$

La figure 2.25 ci-après permet de vérifier l'hypothèse de linéarité entre les variables Xlambert et pixelX, et entre les variables Ylambert et pixelY.

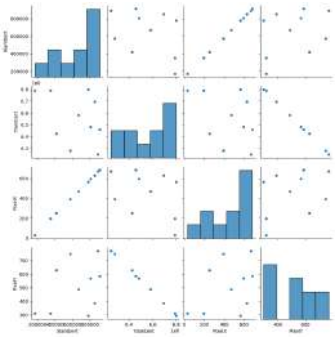
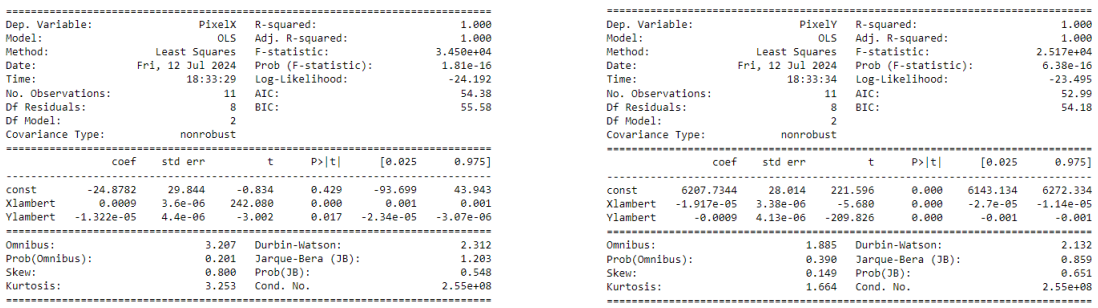


FIGURE 2.25 – Vérification de l'hypothèse de linéarité — approche n°2

Les résultats des deux régressions linéaires, pour PixelX et PixelY, sont présentés sur la figure 2.26 ci-dessous.

Pour les modèles PixelX et PixelY, les R^2 et ${R_{adj}}^2$ égaux à 1.000 montrent que les



Dep. Variable:	PixelX	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	3.450e+04
Date:	Fri, 12 Jul 2024	Prob (F-statistic):	1.81e-16
Time:	18:33:29	Log-Likelihood:	-24.192
No. Observations:	11	AIC:	54.38
Df Residuals:	8	BIC:	55.58
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-24.8782	29.844	-0.834	0.429	-93.699	43.943
Xlambert	0.0009	3.6e-06	242.080	0.000	0.001	0.001
Ylambert	-1.322e-05	4.4e-06	-3.002	0.017	-2.34e-05	-3.07e-06

Omnibus:	3.207	Durbin-Watson:	2.312
Prob(Omnibus):	0.201	Jarque-Bera (JB):	1.203
Skew:	0.800	Prob(JB):	0.548
Kurtosis:	3.253	Cond. No.	2.55e+08

Dep. Variable:	PixelY	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	2.517e+04
Date:	Fri, 12 Jul 2024	Prob (F-statistic):	6.38e-16
Time:	18:33:34	Log-Likelihood:	-23.495
No. Observations:	11	AIC:	52.99
Df Residuals:	8	BIC:	54.18
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6207.7344	28.014	221.596	0.000	6143.134	6272.334
Xlambert	-1.917e-05	3.38e-06	-5.680	0.000	-2.7e-05	-1.14e-05
Ylambert	-0.0009	4.13e-06	-209.826	0.000	-0.001	-0.001

Omnibus:	1.885	Durbin-Watson:	2.132
Prob(Omnibus):	0.390	Jarque-Bera (JB):	0.859
Skew:	0.149	Prob(JB):	0.651
Kurtosis:	1.664	Cond. No.	2.55e+08

FIGURE 2.26 – Résultats de l'approche n°2 - PixelX et PixelY

modèles expliquent parfaitement la variance des coordonnées en pixels. Pour PixelX, l'intercept n'est pas significatif (p = 0.429), tandis que Xlambert et Ylambert ont des impacts significatifs, positifs pour Xlambert (p = 0.000) et négatifs pour Ylambert (p = 0.017).

Pour PixelY, Xlambert et Ylambert ont tous deux un impact significatif et négatif (p = 0.000).

Les statistiques F très élevées (4327.16 pour PixelX et 4513.49 pour PixelY) confirment la significativité globale des modèles. Les tests sur les résidus montrent une distribution normale et une indépendance des résidus, validant ainsi les modèles.

Synthèse des deux régressions.

	PixelX		**PixelY**	
Indicateurs | Approches	**n°1**	**n°2**	**n°1**	**n°2**
RMSE (base d'apprentissage)	6,16	2,18	4,83	2,05
RMSE (base de test)	6,30	3,76	5,03	1,83
R^2	0,99	0,99	1	1

TABLE 2.7 – Comparaison des performances des approches pour PixelX et PixelY

Le tableau 2.7 présente les performances de deux approches pour modéliser les coordonnées en pixels (PixelX et PixelY) à partir des indicateurs $RMSE$ et R^2. L'approche n°2 montre des performances supérieures avec une RMSE plus faible pour les bases d'apprentissage et de test, indiquant une meilleure précision de cette méthode. Bien que les deux approches aient un coefficient de détermination R^2 élevé (proche de 1), signifiant que les modèles expliquent bien la variance des données, la précision accrue de l'approche n°2 en fait un meilleur choix.

En plus de ces résultats quantitatifs, l'utilisation des coordonnées Lambert 93, la projection cartographique officielle en France, a également influencé la décision. Ce système est

conçu pour offrir une meilleure précision géographique, particulièrement pour les grandes échelles, ce qui est crucial pour les applications cartographiques détaillées. Ainsi, même si les performances de l'approche n°2 avaient été légèrement inférieures à celles de l'approche n°1, elle aurait tout de même été retenue en raison des avantages structurels liés à l'utilisation des coordonnées Lambert 93. Cette considération permet de garantir une modélisation robuste et cohérente des coordonnées en pixels, adaptée aux besoins spécifiques de l'analyse géographique.

2.5 Redimensionnement et scan des images

Le redimensionnement des images est une étape cruciale dans le processus de fonctionnement de l'outil algorithmique, car, comme mentionné dans l'introduction de la section *2.4 carte et base de données de référence*, une carte de référence a été choisie pour uniformiser la position de toutes les communes, quelle que soit la carte dont on veut extraire les informations. Sachant les positions des communes sur la carte de référence, il est nécessaire de redimensionner ou de modifier la forme de l'image de carte étudiée, afin qu'elle ait les mêmes dimensions et le même positionnement de la France que la carte de référence. C'est l'essence du redimensionnement des images. De plus, le scan des images consiste à récupérer les informations de chaque commune sur la carte en associant la couleur du pixel représentant la commune à la légende de la carte.

Ce chapitre détaillera les étapes de restructuration des images et les processus de scan pour extraire les données géographiques nécessaires.

2.5.1 Restructuration des images par transformation affine

Le redimensionnement des images consiste à modifier ou ajuster les dimensions et le positionnement d'une image de carte pour qu'elle corresponde à une carte de référence choisie. Cette opération est essentielle pour garantir que chaque commune occupe la même position sur toutes les cartes utilisées, facilitant ainsi l'extraction et l'analyse des données géographiques.

Ce redimensionnement nécessite un masque de la carte de référence et se fait en plusieurs étapes. On applique des transformations affines (translation, rotation et homothétie) sur les images de carte pour qu'elles se superposent parfaitement ou presque avec le masque de la carte de référence. Pour appliquer ces transformations, on définit une fonction d'erreur (plus la carte est différente du masque de la carte de référence, plus l'erreur est grande) et on optimise le processus de sorte à minimiser l'erreur en faisant varier les transformations.

La conception du masque repose sur l'uniformité du fond de la carte de référence, qui est entièrement blanc en dehors de la France. Pour créer ce masque, une copie de la carte de référence est générée. À chaque pixel de cette copie est attribuée une valeur en fonction de sa couleur sur la carte de référence : s'il correspond à un pixel blanc, il

se voit attribuer la valeur 0 ; dans le cas contraire, il reçoit la valeur 1. Ainsi, tous les pixels représentant la France prennent la valeur 1, tandis que les autres (c'est-à-dire le hors légende) prennent la valeur 0. Pour visualiser le masque, les pixels ayant la valeur 1 sont affichés en gris (voir figure 2.27).

FIGURE 2.27 – Carte de référence[8] et masque

a - Transformations géométriques

Les transformations géométriques permettent de modifier l'image pour qu'elle se superpose à une autre image de référence. Les principales transformations utilisées sont les translations, les rotations, les mises à l'échelle et les réflexions. Ces transformations peuvent être combinées et représentées par des matrices homogènes 3×3.

La translation déplace chaque pixel de l'image d'un certain vecteur (t_x,t_y). La matrice de translation est :

$$\begin{pmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{pmatrix}$$

La rotation fait pivoter chaque pixel autour de l'origine d'un angle θ. La matrice de rotation est :

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

La mise à l'échelle (homothétie) modifie les dimensions de l'image par un facteur k. La matrice d'homothétie est :

$$\begin{pmatrix} k & 0 & 0 \\ 0 & k & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

La réflexion symétrise l'image par rapport à un axe. Par exemple, la réflexion par rapport à l'axe y est représentée par la matrice :

$$\begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Une image numérique est essentiellement une matrice dans laquelle chaque élément représente un pixel. Pour appliquer ces transformations à des images, chaque pixel de l'image est représenté par ses coordonnées homogènes (x,y,1) qui représente sa position dans la matrice de l'image, et la transformation géométrique est appliquée en multipliant la matrice de transformation par les coordonnées de chaque pixel. Pour chaque pixel (x,y) de l'image originale, les nouvelles coordonnées (x',y') sont calculées en utilisant la matrice de transformation.

Théorème (transformation affine) : *Une transformation affine dans le plan est une application $T : \mathbb{R}^2 \to \mathbb{R}^2$ de la forme :*

$$T(\mathbf{x}) = A\mathbf{x} + \mathbf{b},$$

où $\mathbf{x} \in \mathbb{R}^2$, A est une matrice 2×2, et $\mathbf{b} \in \mathbb{R}^2$. Cette transformation est composée d'une transformation linéaire représentée par la matrice A et d'une translation représentée par le vecteur $\mathbf{b}$. En coordonnées homogènes, la transformation affine peut être exprimée sous la forme matricielle suivante :

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} a & b & t_x \\ c & d & t_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = A \begin{pmatrix} x \\ y \\ 1 \end{pmatrix},$$

où a, b, c, d déterminent la transformation linéaire (rotation, homothétie, réflexion), et t_x, t_y déterminent la translation. A est la matrice de transformation et donc une composée de transformations affines.

Étant donné que les nouvelles coordonnées (x',y') peuvent ne pas correspondre exactement aux positions de la grille de pixels de l'image, une interpolation est nécessaire pour déterminer les valeurs des pixels transformés. L'interpolation bilinéaire est couramment utilisée pour ce processus, offrant un compromis entre précision et complexité de calcul.

b - Optimisation

Une fonction "transformation" est conçue pour aligner les images, avec précision, sur le masque de la carte de référence. Une fonction d'erreur est d'abord définie pour mesurer la différence entre une image et le masque de la carte de référence. La fonction "transformation" prend en entrée une image de carte, applique une transformation géométrique grâce à la matrice de transformation, puis calcule l'erreur entre l'image transformée et le masque de la carte de référence. L'objectif est de trouver, par optimisation, les coefficients a, b, c, d, t_x et t_y de la matrice de transformation qui permettent de minimiser l'erreur entre l'image de la carte et le masque de référence.

Fonction d'erreur. L'objectif de la fonction d'erreur est de mesurer numériquement le degré de superposition entre une image de carte et le masque de référence. Cette fonction d'erreur est composée de deux termes :

$$\text{Erreur} = \text{terme 1} + \text{terme 2}$$

Pour les différents termes de cette erreur, on considère que $\|.\|$ *est la distance euclidienne.*

1. Terme 1 :
$$\text{terme 1} = \sum_i \|\text{vide - pixel}_i\| \mathbf{1}_{\{\text{pixel}_i \notin \text{Masque}\}}$$
Ce terme pénalise la présence de pixels d'une couleur autre que celle du vide à l'extérieur du masque. Plus la distance entre la couleur du pixel et la couleur du vide est grande, plus l'erreur augmente. Cela signifie que l'algorithme favorise une correspondance entre les pixels extérieurs au masque et la couleur du vide.
2. Terme 2 :
$$\text{terme 2} = \sum_i \frac{1}{(\alpha + \|\text{vide - pixel}_i\|)} \mathbf{1}_{\{\text{pixel}_i \in \text{Masque}\}}$$
Ce terme pénalise la présence de pixels de la couleur du vide à l'intérieur du masque. α est une constante ajoutée pour éviter que le dénominateur ne soit nul. Une petite valeur de α garantit que la pénalisation reste forte pour les pixels identiques à la couleur du vide à l'intérieur du masque.

Minimiser cette fonction d'erreur revient donc à obtenir une superposition parfaite de l'image et du masque de référence, c'est-à-dire, ni pixels non vides à l'extérieur du masque, ni pixels vides à l'intérieur du masque. Cependant, un biais peut naître de la création de cette fonction. Dans certaines situations, le second terme peut dominer le premier, car les pénalités à l'intérieur du masque peuvent être amplifiées de manière disproportionnée par la petite valeur de α. Cela peut conduire à une optimisation qui privilégie excessivement l'élimination des pixels vides à l'intérieur du masque, au détriment de l'élimination des pixels non vides à l'extérieur du masque.
Pour pallier ce problème, une alternative serait d'ajouter une constante judicieusement choisie au terme 1, ce qui permettrait de rééquilibrer les deux termes. Ainsi, les deux critères de superposition seraient pris en compte de manière équilibrée lors du processus d'optimisation.

Méthodes d'optimisation. L'optimisation est le processus visant à trouver les valeurs des variables qui minimisent ou maximisent une fonction objectif. En d'autres termes, il s'agit de déterminer les conditions optimales sous lesquelles une fonction donnée atteint son minimum ou son maximum. Cela est crucial dans la transformation géométrique des images pour l'alignement avec le masque de référence. Il existe plusieurs méthodes d'optimisation, chacune ayant ses propres avantages et inconvénients en fonction de la nature du problème à résoudre.

— **La méthode de Powell** est une méthode d'optimisation sans dérivée qui ne nécessite pas le calcul des gradients. Elle est particulièrement utile pour les problèmes dans lesquels la fonction objectif est non différentiable, discontinue ou bruitée. La méthode fonctionne en optimisant successivement le long de différentes directions

de recherche, mises à jour au fur et à mesure (voir annexe B).

— **La méthode de Nelder-Mead**, également connue sous le nom de méthode du simplexe, est une autre méthode sans dérivée qui utilise un simplexe de points pour explorer l'espace de solution. Le simplexe est déformé à chaque étape par des opérations telles que la réflexion, l'expansion, la contraction et le rétrécissement pour approcher l'optimum.

— **La descente de gradient** est une méthode d'optimisation qui utilise les dérivées de la fonction objectif pour descendre progressivement vers un minimum local. À chaque itération, les paramètres sont mis à jour dans la direction opposée au gradient de la fonction objectif.[22]

Dans le contexte de l'optimisation des transformations géométriques pour aligner des images sur un masque de référence, les méthodes de Powell et de Nelder-Mead sont particulièrement adaptées, car elles ne nécessitent pas le calcul des dérivées. Cette caractéristique est essentielle, étant donné que la fonction d'erreur utilisée n'est pas différentiable en tout point.

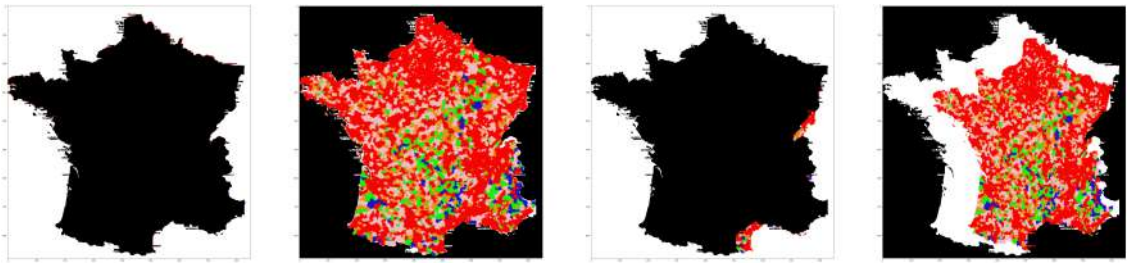


FIGURE 2.28 – Redimensionnement — carte B (méthode Powell à gauche et méthode Nelder-Mead à droite)

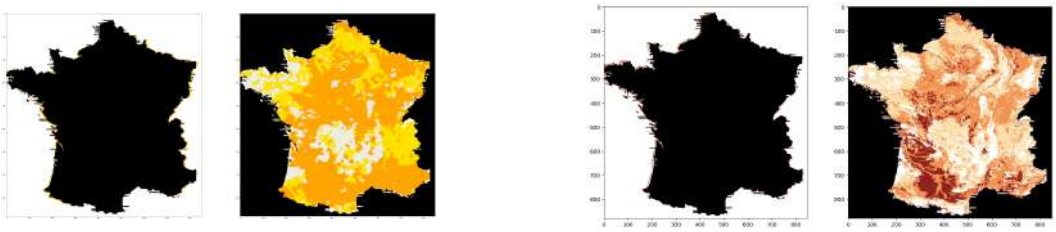


FIGURE 2.29 – Redimensionnement — cartes A et D (méthode de Powell)

Sur les figures ci-dessus, la superposition du masque de référence aux différentes cartes est observée. La première remarque que l'on peut faire grâce à la figure 2.28 est que la méthode de Powell permet généralement d'obtenir de meilleurs résultats que la méthode de Nelder-Mead. Ensuite, bien que les cartes redimensionnées soient obtenues, la superposition n'est pas parfaite et une faible proportion de pixels ayant la couleur du vide est encore présente à l'intérieur du masque.

A ce stade, l'intérieur du masque contient un certain nombre de pixels ayant la couleur du vide. Ces pixels ont pour origine :

1. le vide laissé par le léger décalage de superposition avec le masque, comme mentionné précédemment
2. les éléments assimilés à du hors légende après l'étape précédente du clustering colorimétrique (bruit, nom de ville, cours d'eau...)
3. le *vide* apparu suite aux transformations affines appliquées qui ont pu déformer le support d'origine.

Pour gérer ces pixels vides à l'intérieur de la France, un algorithme des K plus proches voisins est utilisé pour leur attribuer la couleur des pixels environnants.

Algorithme des K plus proches voisins (K-NN)

L'algorithme des K plus proches voisins (K-NN) est une méthode d'apprentissage supervisé utilisée principalement pour les tâches de classification et de régression. Il se base sur l'idée que des points de données similaires sont généralement proches les uns des autres dans l'espace des caractéristiques. Pour classifier un nouveau point, l'algorithme identifie ses K plus proches voisins et attribue la classe majoritaire parmi ces voisins. Pour la régression, il prédit la valeur en utilisant la moyenne des valeurs des K plus proches voisins. L'algorithme est itératif et suit les étapes suivantes :

1. Choix du paramètre K : Déterminer le nombre de voisins à considérer (K). Un petit K peut rendre l'algorithme sensible au bruit, tandis qu'un K trop grand peut inclure des points qui ne sont pas pertinents pour la classification ou la régression.
2. Calcul des distances : Pour chaque point de données de l'ensemble de test, calculer la distance entre ce point et tous les points de l'ensemble d'entraînement. La distance Euclidienne est couramment utilisée, mais d'autres mesures de distance peuvent également être employées.
3. Identification des K plus proches voisins : Trier toutes les distances et identifier les K plus proches voisins du point de données considéré.
4. Vote pour la classification ou moyenne pour la régression : Compter le nombre de voisins dans chaque catégorie et attribuer la catégorie ayant le plus grand nombre de votes (Classification), ou calculer la moyenne des valeurs des K plus proches voisins pour prédire la valeur de l'observation (Régression).

L'algorithme K-NN est simple à comprendre et à implémenter. Il est particulièrement efficace pour les ensembles de données de petite taille et les problèmes dans lesquels la relation entre les caractéristiques et la classe de sortie est complexe et non linéaire. Cependant, K-NN peut devenir très lent et gourmand en mémoire avec des ensembles de données de grande taille, car il nécessite le calcul de la distance entre chaque paire de points. De plus, la performance de l'algorithme dépend de la mesure de distance utilisée et de la valeur choisie pour K. Un autre inconvénient est qu'il est sensible aux caractéristiques non pertinentes et à l'échelle des données, nécessitant souvent une normalisation ou une mise à l'échelle préalable des caractéristiques.

FIGURE 2.30 – Cartes A et C (K-NN)

La figure 2.30 illustre l'efficacité de l'algorithme des K plus proches voisins pour remplir les "trous" laissés d'une part par la méthode ACP et d'autre part par le redimensionnement des images.

2.5.2 Constitution de la base de données de sortie

Une fois les images redimensionnées, la couleur (au format RGB) associée aux coordonnées en pixels de chaque commune de la base des codes postaux est récupérée. Les éléments de la légende sont attribués à chaque code postal en minimisant la distance quadratique entre la couleur de la commune et celles de la légende.

La distance quadratique, également connue sous le nom de distance euclidienne, est une mesure utilisée pour quantifier la différence entre deux points dans un espace tridimensionnel. Pour deux couleurs $\text{couleur}_1 = (R_1, G_1, B_1)$ et $\text{couleur}_2 = (R_2, G_2, B_2)$, la distance quadratique est calculée comme suit :

$$\text{distance} = \sqrt{(R_1 - R_2)^2 + (G_1 - G_2)^2 + (B_1 - B_2)^2}$$

Cette distance permet de déterminer la similitude entre deux couleurs. En minimisant cette distance, chaque commune peut être associée à l'élément de la légende dont la couleur est la plus proche de celle observée sur la carte redimensionnée.

Code INSEE	Code postal	pixelX	pixelY	exposition au risque sécheresse
01001	01400	627	497	moyenne
01002	01640	661	510	faible
01004	01500	657	515	faible
01005	01330	626	512	faible
01006	01300	673	534	sans
⋮	⋮	⋮	⋮	⋮
95680	95400	455	227	forte
95682	95720	454	221	moyenne
95690	95420	418	219	moyenne

TABLE 2.8 – Base de données complétée — carte D

2.6 Résultats et analyses

2.6.1 Présentation et interprétation des résultats

Parmi les cartes utilisées pour illustrer la méthodologie de conception de l'outil algorithmique d'extraction, les bases de données d'origine des cartes A et B sont disponibles. Il est donc possible d'obtenir un indice de performance numérique de l'outil en déterminant le pourcentage de communes correctement classées par l'outil, c'est-à-dire les communes pour lesquelles l'outil a attribué l'élément correct de la légende.

Les résultats présentés dans le tableau 2.9 montrent les taux de communes bien classées pour les cartes A et B, en utilisant deux méthodes différentes, à savoir l'ACP et la méthode K-Means. Pour la carte A, l'ACP atteint un taux de 89,63%, tandis que les K-Means atteignent 81,47%, indiquant une meilleure performance de l'ACP pour cette carte. Pour la carte B, l'ACP atteint 87,55%, et les K-Means atteignent 85,71%, montrant à nouveau une supériorité de l'ACP, bien que l'écart soit moindre. Ces résultats suggèrent que l'ACP est plus efficace pour capturer les variations de couleur et différencier les communes, surtout pour la carte A, où la différence de performance est plus marquée. La stabilité des résultats de l'ACP entre les deux cartes indique une robustesse accrue de cette méthode par rapport à la méthode K-Means, qui montrent plus de variabilité. Cependant, comme la performance n'est pas de 100%, une partie des données récupérées sera erronée. Ce biais peut avoir un impact sur la modélisation de la sinistralité, un aspect qui sera mis en exergue dans le *cas pratique*.

cartes	Taux de communes bien classées
carte A *(méthode ACP)*	89,63%
carte A *(méthode K-Means)*	81,47%
carte B *(méthode ACP)*	87,55%
carte B *(méthode K-Means)*	85,71%

TABLE 2.9 – Résultats (cartes A et B)

En conclusion, l'ACP se révèle être la méthode la plus performante et stable pour le classement des communes, ce qui en fait un outil précieux pour l'extraction des variables

géographiques à partir des images de cartes. Ces performances démontrent l'efficacité de l'outil algorithmique développé dans ce mémoire pour fournir des données géographiques assez précises et fiables, essentielles pour la modélisation des risques en assurance habitation.

Contrairement aux cartes A et B, les cartes C et D sont dépourvues de base de données d'origine, il n'est donc pas possible de calculer un taux de communes bien classées. On peut toutefois reconstituer une carte à partir des données fournies par l'outil et la comparer visuellement à l'image d'origine.

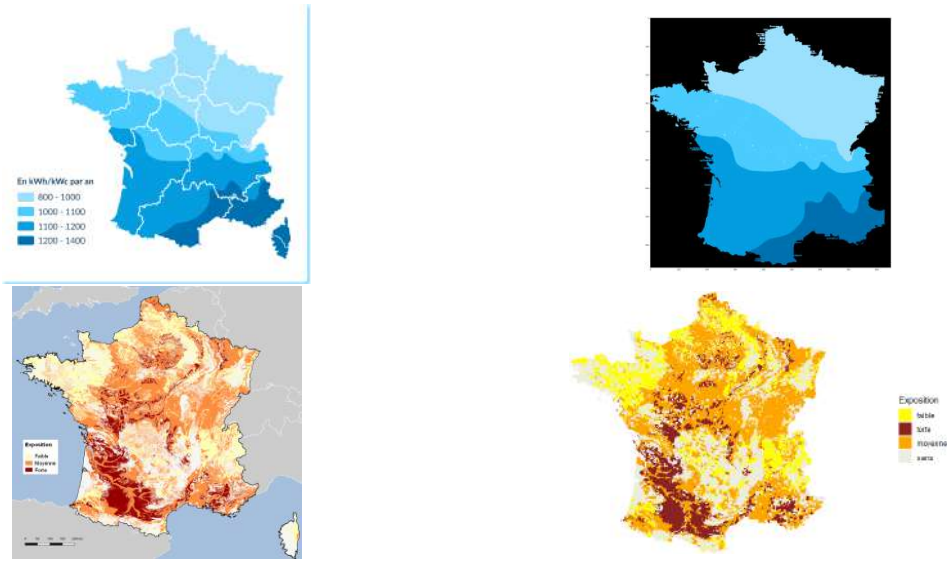


FIGURE 2.31 – Cartes C[30] et D[54] - originales vs recréées (méthode ACP)

Cette méthode ne permet pas d'obtenir une évaluation précise de la performance de l'outil, mais permet d'avoir un premier aperçu de la capacité de l'outil à restituer fidèlement les informations d'une carte.

2.6.2 Limites et perspectives d'amélioration de l'outil

L'outil algorithmique développé dans ce mémoire vise à extraire des variables géographiques à partir d'images de cartes. Bien que les résultats montrent une performance solide, l'outil présente certaines limites et offre diverses perspectives d'amélioration. Cette section explore ces aspects pour fournir une compréhension complète des défis actuels et des opportunités futures pour perfectionner cet outil algorithmique.

Limites de l'outil. Parmi les limites de l'outil, il y a : l'absence d'indice de performance, car il est seulement possible d'obtenir un indice de performance numérique pour les cartes dont la base de données d'origine est connue. Dans le cas contraire, l'évaluation

des performances de l'outil est limitée à une simple comparaison visuelle. La précision est incomplète, comme mentionné dans la section précédente, l'outil n'atteint pas une précision de 100%, ce qui signifie que certaines communes peuvent être mal classées. Ce biais dans les données récupérées peut affecter la modélisation de la sinistralité, rendant par exemple les prédictions moins précises. La sensibilité à la qualité des images est également une limitation, car la qualité des images utilisées influence fortement les résultats. Les images de faible résolution ou très bruitées peuvent entraîner des erreurs de classification plus fréquentes. La complexité des cartes est un autre défi : les cartes avec des détails complexes ou de multiples légendes posent des défis supplémentaires. L'algorithme peut avoir des difficultés à distinguer des éléments proches ou à gérer des couleurs similaires, notamment dans les cartes avec de nombreuses nuances de bleu. Enfin, la variabilité des formats de cartes (différentes projections, échelles, certaines cartes peuvent ne pas être des représentations fidèles de la France) peut compliquer l'alignement et la superposition des cartes avec le masque de référence. Des ajustements manuels peuvent être nécessaires, limitant ainsi l'automatisation complète du processus.

Perspectives d'amélioration. Pour aller plus loin et améliorer l'outil, plusieurs pistes peuvent être explorées. Il est possible d'améliorer la précision du redimensionnement en intégrant des techniques de machine learning plus avancées, comme les réseaux de neurones convolutifs (CNN), qui sont efficaces pour l'analyse d'images. Cela pourrait améliorer la capacité de l'outil à différencier les détails fins et à traiter les images de qualité variable pour un meilleur redimensionnement. De plus, les filtres gaussiens représentent une solution possible pour débruiter efficacement des images. (voir Autres approches)

Bien que l'outil algorithmique développé présente déjà des performances solides, plusieurs axes d'amélioration sont possibles pour augmenter sa précision, sa robustesse et son efficacité. En abordant ces limites et en explorant ces perspectives d'amélioration, l'outil pourrait devenir encore plus fiable et utile pour l'extraction des variables géographiques à partir des images de cartes. Ces améliorations contribueraient de manière significative à la modélisation des risques en assurance habitation.
Enfin, il est crucial de rappeler l'importance du respect des considérations éthiques et légales liées à l'utilisation des données de cartes pour garantir une application responsable et légitime de cet outil.

Chapitre 3

Cas pratique d'usage de l'outil : zonier simplifié en assurance habitation

L'évaluation précise des risques climatiques est devenue de plus en plus importante pour l'industrie de l'assurance habitation ; cela est dû à l'augmentation des impacts du changement climatique. Les modèles traditionnels, souvent basés sur des données structurées limitées, ne capturent pas pleinement les facteurs géographiques complexes qui influencent ces sinistres. L'outil algorithmique développé dans ce mémoire permet d'exploiter des cartes pour extraire des variables géographiques pertinentes, offrant ainsi une approche supplémentaire pour affiner les modèles actuariels.

3.1 Problématique à l'origine du cas d'usage

Pour améliorer le suivi du risque, Cardif IARD souhaite affiner son modèle de rentabilité technique, qui évalue la rentabilité technique espérée de chaque contrat en tenant compte des variables disponibles, telles que les informations de risque à la souscription, les données clients, et des données externes.

Ce modèle produit un indicateur clé : le rapport sinistres à primes (S/P), permettant de mesurer la rentabilité attendue d'un contrat dès sa souscription. Cet indicateur est essentiel pour le suivi du portefeuille, en identifiant les zones de rentabilité et en détectant les déformations structurelles potentielles.

Le modèle de rentabilité technique se ventile donc en sous-modèles par garantie du contrat habitation. À ce jour, aucun sous-modèle n'avait été développé pour estimer la prime pure attendue liée au péril sécheresse dans la garantie Catastrophe Naturelle, en raison de la réglementation qui fixe cette prime à l'avance. Cependant, un tel modèle pourrait offrir des avantages :

1. l'indicateur de S/P espéré produit par le modèle de rentabilité technique globale.

2. Comprendre l'exposition de Cardif IARD au risque sécheresse en la géolocalisant.
3. Comparer la sinistralité espérée attendue avec la prime réglementaire perçue.
4. Suivre les évolutions du portefeuille face à ce risque et mettre en évidence des tendances..

Cardif IARD souhaite donc tester la construction d'un modèle de prime pure attendue pour la sécheresse, envisageant une éventuelle intégration future. Le risque sécheresse étant influencé par des facteurs géographiques, l'outil développé dans ce mémoire pourrait jouer un rôle précieux dans ce développement. Ce cas pratique s'inscrit dans cette démarche.

Bien que ce modèle puisse être complexe, en raison des nombreux facteurs influençant la sinistralité sécheresse (géologiques, météorologiques, administratifs, etc.), l'objectif est de poser les bases d'un premier modèle avec un nombre restreint de facteurs explicatifs, pour une future amélioration. Cet exercice montrera comment des données externes non structurées, telles que des images de cartes, peuvent enrichir la compréhension et la prévision des sinistres liés à la sécheresse.

Selon France Assureurs, le coût de la sécheresse est celui qui connaît la plus forte croissance parmi les risques couverts par la garantie CatNat. Bien que la tarification de la sécheresse ne puisse être modifiée, la modélisation aboutira à un zonier simplifié du risque, utilisable de manière diversifiée.

3.2 Protocole

L'approche suit plusieurs étapes :

1. Extraction de variables géographiques : Dans un premier temps, l'outil d'extraction de données est utilisé pour obtenir des variables géographiques à partir de cartes jugées pertinentes par rapport à la sécheresse.
2. Modélisation de la fréquence des sinistres : La fréquence des sinistres est modélisée à l'aide d'un GLM. Ce modèle n'intègre que les variables géographiques extraites des cartes. Une distribution de Poisson est choisie pour modéliser le nombre de sinistres, avec une fonction de lien logarithmique pour refléter les relations multiplicatives.
3. Construction d'un zonier simplifié : À partir des résultats de la modélisation, une classification est réalisée pour créer des groupes de codes postaux homogènes en termes de fréquence pour le risque sécheresse.
4. pseudo-S/P et interprétations : Les primes pures, ainsi que les S/P suivant les zones créées sont calculés et interprétés.

3.3 Extraction de variables géographiques externes

Les variables géographiques recherchées doivent permettre d'expliquer le risque sécheresse (en particulier la fréquence). On s'intéresse donc aux cartes géographiques qui

contiennent des informations pertinentes sur les causes générales de sécheresse ou directement sur le risque de sécheresse. Il est donc nécessaire de comprendre la notion de sécheresse en assurance.

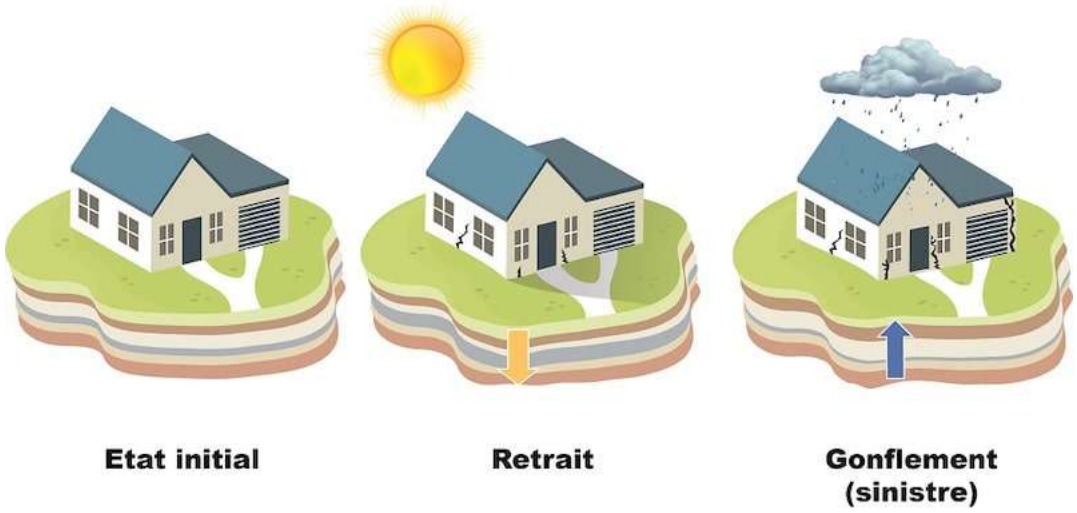


FIGURE 3.1 – Phénomène de retrait-gonflement d'argile [36]

Le phénomène de sécheresse en assurance habitation est principalement lié aux mouvements de terrain causés par l'alternance de périodes de sécheresse et de réhydratation des sols argileux. Ce processus est communément désigné sous le terme de "retrait-gonflement des argiles" ou principe de subsidence. Pendant les périodes de sécheresse prolongée, les sols argileux se contractent en raison de la perte d'humidité, entraînant un tassement du sol sous les fondations des bâtiments. À l'inverse, lors des périodes de réhydratation, comme après de fortes pluies, ces sols absorbent l'eau et gonflent, provoquant des soulèvements du terrain. Ces cycles de retrait et de gonflement exercent des pressions variables sur les fondations, créant des mouvements différentiels qui peuvent causer des dommages structurels aux bâtiments. Ces dommages se manifestent souvent sous forme de fissures diagonales, en particulier sur les murs extérieurs, suivant les joints de maçonnerie.

La recherche a abouti à la sélection de 5 cartes open source dont les sources sont mentionnées dans la bibliographie et référencées par des nombres au niveau des légendes des cartes :

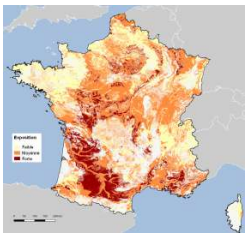


FIGURE 3.2 – Carte_rga [54]

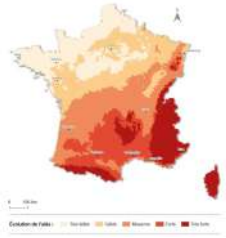


FIGURE 3.3 – Carte_chaleur [28]

FIGURE 3.4 – Carte_irrad [28]

FIGURE 3.5 – Carte_temp [18]

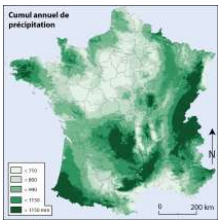


FIGURE 3.6 – Carte_pluie [18]

La carte_rga, également appelée carte A dans le chapitre 2, représente les zones argileuses en France, c'est-à-dire les régions où la composition du sol les expose plus ou moins au risque de retrait-gonflement. Cette carte est fondamentale pour modéliser le phénomène de sécheresse, car le retrait-gonflement des argiles est au cœur du phénomène. En complément, la carte_chaleur et la carte_irrad fournissent des informations sur les niveaux de chaleurs et d'irradiations solaires, deux facteurs liés à la contraction des sols argileux. Ces cartes sont donc particulièrement pertinentes pour évaluer les zones les plus exposées aux dommages liés à la sécheresse.
La carte_temp et la carte_pluie, quant à elles, représentent respectivement la température moyenne et le cumul annuel moyen des précipitations sur la période 1981 à 2010. Bien que la carte_temp reste pertinente pour évaluer l'impact des hautes températures sur le phénomène de sécheresse, la carte_pluie est plus discutable. En effet, la pluviométrie affecte le gonflement des sols argileux et donc le phénomène de sécheresse. Néanmoins, la période couverte par cette dernière est relativement ancienne (1981-2010) et ne reflète pas nécessairement les conditions climatiques actuelles, ce qui limite son utilité pour expliquer les tendances récentes liées au changement climatique. Toutefois, cette carte a été intégrée à titre de test, car l'outil algorithmique vise à faciliter la recherche de données, même lorsque leur impact potentiel est incertain. N'ayant pas trouvé de cartes de pluviométrie plus récentes et suffisamment détaillées, la carte_pluie est conservée, sachant que son utilisation pourrait n'apporter que peu de valeur ajoutée. L'objectif étant aussi de permettre aux modèles de tester ces données et de rejeter celles qui ne sont pas significatives.

Les images de cartes sont d'abord traitées pour garantir une qualité optimale pour l'extraction des données. Cela comprend la dissociation des couleurs, l'élimination des artefacts visuels, et la réduction du bruit pour assurer que les pixels représentent fidèlement les caractéristiques géographiques.

FIGURE 3.7 – carte_chaleur (ACP)

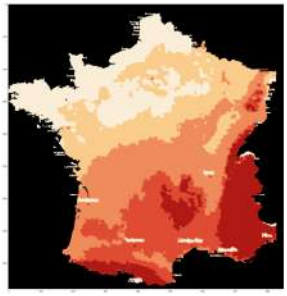


FIGURE 3.8 – carte_chaleur (K-Means)

FIGURE 3.9 – carte_irrad (ACP)

FIGURE 3.10 – carte_irrad (K-Means)

Les figures ci-dessus montrent les résultats du traitement des cartes 2 et 3 avec les méthodes ACP et K-Means. Il en ressort un redimensionnement et un clustering colorimétrique plutôt réussis qui permettent d'obtenir des représentations assez fidèles des cartes. Toutefois, quelques défauts sont remarquables, notamment au niveau des noms de quelques villes qui étaient présentes sur la carte originale, mais que l'algorithme a eu du mal à remplacer par la couleur adéquate. Le but de ces travaux étant d'incorporer les variables géographiques issues de ces cartes à un modèle, il est donc nécessaire de choisir entre les méthodes ACP et K-Means, qui permettent d'obtenir deux variables différentes pour la même carte. Étant donné qu'il s'agit de cartes dont les bases de données d'origine ne sont pas acquises, il n'est pas possible d'obtenir un réel indicateur de la performance globale des deux méthodes pour comparaison. Cependant, en plus d'une comparaison visuelle subjective qui laisse à penser que l'ACP arrive à mieux traiter les noms des quelques villes, et du fait que le chapitre précédent présentait l'ACP comme la meilleure méthode en général, l'étape de redimensionnement peut également servir d'indicateur pour choisir la meilleure méthode concernant une même carte. En effet, l'algorithme de redimensionnement se base sur une optimisation dans laquelle une fonction d'erreur est à minimiser, il est alors possible de choisir la méthode qui engendre la plus petite erreur. Prenons l'exemple de la carte_irrad, à travers la figure et le tableau des erreurs

ci-dessous :

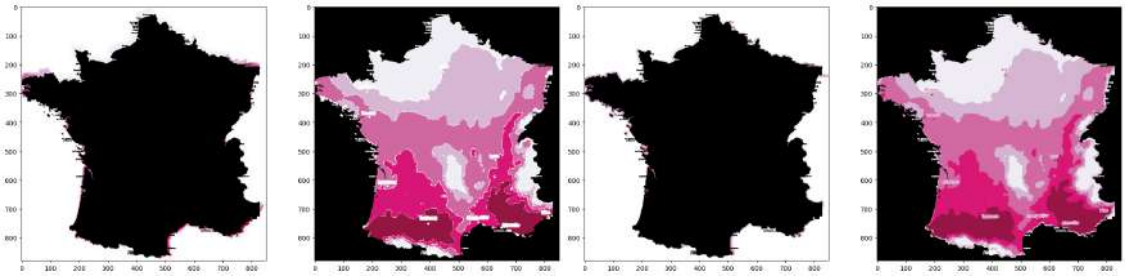


FIGURE 3.11 – Résultats du redimensionnement ACP (gauche) et K-Means (droite)

Méthodes	**ACP**	**K-Means**
carte_irrad	6 246 779,82	2 261 177,22

TABLE 3.1 – Comparaison des erreurs de redimensionnement des méthodes ACP et K-Means

Remarquable également sur la figure 3.11, le tableau 3.1 montre que la méthode K-Means est de loin la meilleure pour le traitement de cette carte avec une erreur presque trois fois moins élevée que celle de l'ACP. Cet écart pourrait paraître incohérent, mais en réalité, il est dû au fait que la fonction d'erreur pénalise fortement les pixels de la couleur du vide à l'intérieur du masque. De plus, cette erreur est obtenue d'une part par la distance entre chaque pixel et la couleur du vide, d'où l'ordre de grandeur de 10^6. Pour la suite du cas pratique, on conserve la méthode K-Means pour la carte_irrad et la méthode ACP pour les autres cartes.

Pour chaque carte, chacun des pixels est ensuite analysé pour déterminer son association avec une catégorie de données géographiques définie par la légende de la carte. La couleur d'un pixel est associée à la couleur de la légende la plus proche et la catégorie représentée par cette couleur est attribuée au pixel. Pour être plus précis et à titre de rappel, chaque pixel de l'image n'est pas analysé, mais cette classification se fait uniquement sur les pixels représentant des codes postaux de la base de données. Cela permet finalement de convertir les cartes en variables géographiques à plusieurs modalités.

En réalité, bien plus de cinq cartes ont été analysées et transformées en base de données. Toutefois, un premier filtrage a été effectué. Les cartes non présentées, bien que l'outil ait généré une base de données pour celles-ci, ont été éliminées en amont de la modélisation après une analyse univariée ou une analyse de corrélation avec la variable réponse.

INSEE	Codes P.	Carte_rga	Carte_chaleur	Carte_irrad	Carte_temp	Carte_pluie
01001	01400	moyenne	Moyenne	Moyenne	$< 11{,}4$	< 940
01002	01640	faible	Moyenne	Forte	$< 11{,}4$	> 1150
01004	01500	faible	Moyenne	Forte	$< 12{,}8$	< 1150
⋮	⋮	⋮	⋮	⋮	⋮	⋮
95607	95150	forte	Faible	Faible	$< 11{,}4$	< 710
95628	95760	moyenne	Faible	Très faible	$< 11{,}4$	< 710
95637	95490	moyenne	Faible	Très faible	$< 10{,}4$	< 710

TABLE 3.2 – Aperçu des variables géographiques obtenues

3.4 Modélisation de la fréquence des sinistres

Pour modéliser la fréquence, il est nécessaire de constituer une base de données dite de risque, contenant les variables utiles à la modélisation. Deux bases principales sont requises à cet effet : une base de contrats, qui comprend diverses informations sur les contrats d'assurance habitation (l'identifiant du contrat, les caractéristiques du bien assuré telles que déclarées au contrat et l'exposition), ainsi qu'une base de sinistres, qui recense tous les sinistres survenus depuis 2018. Par souci de simplification, la base de données est limitée aux contrats concernant les propriétaires de maisons, principaux concernés par les sinistres de sécheresse (les locataires et les appartements sont donc retirés de la base), et à la période 2020-2022. Cette période est privilégiée, car le délai de reconnaissance des sinistres de sécheresse (18 mois) ne permet pas d'affirmer que tous les sinistres de 2023 sont inclus dans la base. De plus, en raison du lancement des activités de Cardif IARD en 2018, les premières années (2018-2019) ne sont pas incluses (insuffisance de données). Un premier tri est effectué pour ne conserver que les variables nécessaires au calcul et à l'analyse de la fréquence, afin que le modèle explique uniquement l'effet des variables géographiques sur la fréquence des sinistres de sécheresse. Le tableau ci-dessous présente les variables finales de la base de risque utilisée pour la modélisation.

Variables	Descriptif
Code postal	Codes postaux des biens assurés
Nbsin sech	Nombre de sinistres sécheresse
Charge sech	Charges de sinistres sécheresse individuelles
Cot_cnat	Cotisation ou prime de la garantie CatNat
Expos	Expositions des contrats
Carte_rga	Risque de retrait-gonflement d'argile
Carte_chaleur	Niveau de chaleurs
Carte_irrad	Niveau d'irradiations solaires
Carte_temp	Température moyenne
Carte_pluie	Cumul annuel moyen de précipitations
Freq	fréquence individuelle calculée par : Nbsin sech/Expos

TABLE 3.3 – Tableau descriptif des variables de la base de risque

a - Modèle 1C (1 carte)

Dans un premier temps, l'analyse se concentre sur un modèle 1C, dans lequel la fréquence est expliquée par une seule carte. Ce modèle permet d'obtenir un premier aperçu de la capacité des variables géographiques à prédire la sécheresse et d'analyser l'évolution des résultats à mesure que le nombre de variables géographiques augmente. Pour cette approche, l'application d'un algorithme d'apprentissage statistique pour former des zones n'est pas nécessaire. La carte_rga, avec ses quatre modalités (faible, sans info, moyenne et forte), est utilisée. Ces modalités font référence au risque de retrait-gonflement des argiles. La catégorie "sans info" correspond à une couleur présente sur la carte, mais qui n'est rattachée à aucun élément de la légende.
Les fréquences par modalité de la carte_rga sont ensuite calculées, permettant d'observer la cohérence des résultats.
Les données de la base de risque utilisées dans tout le reste du mémoire sont fictives !

Carte_rga	**Expos(%)**	**Freq(%)**
faible	43,46	0,13
sans info	19,71	0,22
moyenne	27,82	0,34
forte	9,01	0,74

TABLE 3.4 – Fréquences par modalité (carte_rga)

L'analyse des données présentées montre une relation claire entre le risque de retrait-gonflement des argiles de la carte_rga et la fréquence des sinistres liés à la sécheresse.
La fréquence des sinistres suit une tendance croissante en fonction du niveau d'exposition au risque de retrait-gonflement des argiles. Les zones classées "forte" affichent une fréquence de sinistres nettement plus élevée (0,74 %) que celles classées "faible" (0,13 %). Cette différence notable entre les fréquences montre que le risque de sinistre est correctement capturé par la carte.
Le regroupement des catégories "sans info" et "moyenne" pourrait être envisagé pour stabiliser les estimations, pour pallier un risque d'instabilité de la catégorie "sans info".
Bien que les zones à fort risque représentent seulement 9,01 % des expositions, elles concentrent une part disproportionnée des sinistres. La majorité des expositions se trouvent dans les zones à faible et moyenne exposition (43,46 % et 27,82 % respectivement), mais ces zones sont associées à des fréquences de sinistres nettement inférieures. Cette situation indique une potentielle sous-estimation des zones à haut risque si la segmentation n'est pas suffisamment fine.
Ainsi, afin d'évaluer le niveau de segmentation de ce modèle 1C, la courbe de Lorenz et le coefficient de Gini sont utilisés (voir annexe C). Ces outils permettent de visualiser et de quantifier le pouvoir discriminant du modèle.

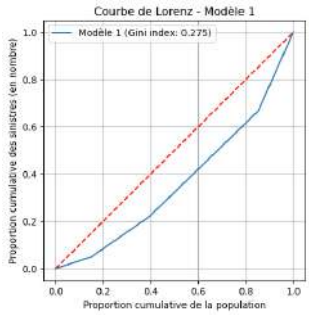


FIGURE 3.12 – Courbe de Lorenz et indicateur de Gini — modèle 1C

La courbe de Lorenz pour le modèle 1C, illustrée ci-dessus, montre la proportion cumulative des sinistres en fonction de la proportion cumulative des contrats (population). La courbe se situe en dessous de la ligne d'égalité parfaite, ce qui indique que les sinistres ne sont répartis de manière uniforme entre les différentes zones. Le coefficient de Gini, calculé à partir de la courbe de Lorenz, est de 0,275, ce qui traduit une segmentation modérée du portefeuille qui sera comparée à celle d'autres modèles décrits dans les sections suivantes. Cela signifie que l'exposition au risque, telle que mesurée par la carte_rga, permet une segmentation pertinente des zones en fonction de leur sinistralité.
Les primes pures par modalité sont également calculées en multipliant les fréquences de chaque modalité par un coût moyen unique, calculé sur l'ensemble de la base. L'utilisation d'un coût moyen global se justifie par le fait que les variables géographiques externes de cette étude ne sont pas adaptées à la modélisation du coût moyen. En effet, le coût d'un sinistre sécheresse dépend principalement des caractéristiques propres au logement endommagé. Les caractéristiques du bâtiment, telles que son type de construction, son âge et la profondeur de ses fondations, jouent un rôle crucial dans la détermination du montant des réparations. La nature des dommages, notamment l'étendue des fissures, la présence de dégâts structurels et les distorsions des portes et fenêtres, impacte directement l'ampleur des travaux nécessaires. Les facteurs économiques, comme les fluctuations du coût des matériaux et de la main-d'œuvre, peuvent faire varier significativement le montant final des réparations. Enfin, l'expertise et l'évaluation du sinistre, incluant la complexité des études géotechniques requises et le choix des méthodes de réparation, contribuent également à définir le coût global du sinistre.

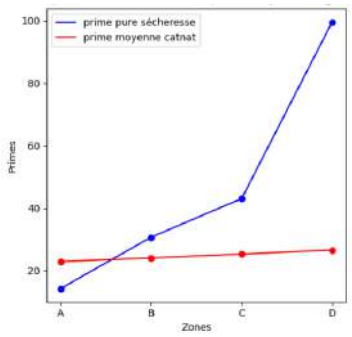


FIGURE 3.13 – Comparaison entre la prime pure sécheresse et la prime moyenne CatNat par zones

La figure 3.13 compare la pseudo-prime pure liée à la sécheresse (en bleu) avec la prime moyenne CatNat (en rouge) pour les différentes zones géographiques. Il s'agit d'une pseudo-prime pure en raison du coût moyen identique pour chaque zone, mais elle sera simplement appelée prime pure dans la suite. Les variations de cette prime dépendent donc uniquement de la fréquence des sinistres.
Cette comparaison met en évidence que la prime pure calculée pour le risque de sécheresse varie considérablement entre les zones. Une nette progression de la prime est observée en passant de la zone A à la zone D, reflétant l'augmentation du risque de sinistre lié à la sécheresse dans les zones géographiques les plus exposées. En revanche, la prime moyenne CatNat reste relativement constante à travers les zones, avec seulement une légère variation. Cela illustre de nouveau que, dans le cadre du régime CatNat, les primes ne tiennent pas compte des différences de risque entre les zones géographiques.

Le régime CatNat repose sur un principe de solidarité nationale, ce qui se traduit par une surprime uniforme appliquée à tous les contrats d'assurance dommages aux biens. Ce système garantit une redistribution des contributions entre les assurés, sans tenir compte du niveau de risque individuel. En effet, la prime CatNat n'est pas modulée en fonction de l'exposition géographique aux risques naturels. Dans ce contexte, le modèle 1C de segmentation du risque de sécheresse (une composante de la garantie CatNat) sur la base des modalités de la carte_rga, propose une approche plus fine en tenant compte des spécificités locales du risque.

L'intérêt peut alors se porter sur le ratio entre la prime pure calculée pour le risque de sécheresse et la prime moyenne CatNat pour chaque zone. L'analyse de ce ratio révèle une information intéressante :

$$\frac{\text{PP sech}}{\text{PrMoy CatNat}} = \frac{\text{PP sech} \times \text{Expos}}{\text{PrMoy} \times \text{Expos}} = \frac{\text{Charge sech totale}}{\text{Cot\_cnat totale}}$$

où PP sech représente la prime pure de sécheresse et PrMoy CatNat est la prime moyenne CatNat.

Ce ratio est équivalent au quotient du montant total des sinistres dus à la sécheresse par le montant total des primes collectées pour la garantie CatNat, qui couvre non seulement la sécheresse, mais aussi d'autres types de catastrophes naturelles (inondations, tempêtes, etc.). Il ne s'agit donc pas du S/P global de la garantie CatNat, qui est le rapport entre le montant total des sinistres CatNat et le montant total des primes CatNat, mais d'un pseudo-S/P spécifique à la sécheresse.

Il peut être utilisé comme un indicateur pour évaluer la rentabilité du contrat par rapport au risque sécheresse. Cependant, ce pseudo-S/P est un indicateur partiel, et son interprétation doit être faite avec prudence. Il peut mettre en lumière la contribution relative du risque de sécheresse dans le cadre global de la garantie CatNat, mais il ne doit pas être utilisé seul pour évaluer la rentabilité ou la tarification de la garantie CatNat dans son ensemble. Pour une analyse plus complète, il serait nécessaire de considérer également les sinistres liés aux autres catastrophes naturelles couvertes par la garantie. *Pour des raisons de confidentialité, les expositions présentées ici et dans le reste du mémoire ont été retraitées !*

Zones	**Expos(%)**	**Freq(%)**	**PrMoy CatNat(€)**	**PP sech(€)**	**Pseudo-S/P**
faible (A)	43,46	0,13	23,01	17,28	0,75
sans info (B)	19,71	0,22	24,14	29,24	1,21
moyenne (C)	27,82	0,34	25,39	45,90	1,81
forte (D)	9,01	0,74	26,68	98,35	3,69

TABLE 3.5 – Pseudo S/P par zones (carte_rga)

Lorsque ce pseudo-S/P est faible, cela signifie que la part de la prime CatNat utilisée pour couvrir les sinistres liés à la sécheresse est faible par rapport au montant total des primes collectées. Cela peut indiquer que la prime CatNat est suffisante, voire surévaluée par rapport au risque de sécheresse, ou que d'autres catastrophes naturelles non liées à la sécheresse consomment la majeure partie des primes. En revanche, si ce pseudo-S/P est élevé, comme c'est le cas dans cet exemple pour 56,54% du portefeuille, cela signifie que les sinistres liés à la sécheresse consomment une part importante des primes CatNat.

b - Modèle PC (plusieurs cartes)

Pour cette étude, le modèle linéaire généralisé (GLM) a été retenu en raison de sa large utilisation en assurance dommages. De plus, cette méthode est facilement interprétable, contrairement à d'autres algorithmes d'apprentissage statistique. La distribution de Poisson a été choisie pour modéliser le nombre de sinistres, car elle est adaptée aux données de comptage, en supposant que ces événements suivent un processus de Poisson. Cependant, cette approche repose sur une hypothèse clé : l'égalité entre la moyenne et la variance du nombre de sinistres, ce qui peut ne pas être toujours respecté.

Dans le cas où la variance des sinistres est supérieure à la moyenne, on parle de surdispersion. Pour tenir compte de cette surdispersion, la distribution quasi-Poisson est souvent utilisée, car elle permet de relâcher cette hypothèse d'égalité et d'ajuster la variance indépendamment de la moyenne. Une autre alternative consiste à utiliser la loi binomiale négative, qui modélise également des données de comptage en prenant en compte la surdispersion de manière explicite. Toutefois, par souci de simplicité, la distribution de Poisson a été privilégiée dans cette analyse.

Dans l'approche standard de constitution d'un zonier, ce sont les résidus d'un modèle sans variable géographique qui sont modélisés à partir des variables géographiques. Dans le cadre de ces travaux, la construction du zonier est simplifiée et la variable réponse du GLM est la fréquence des sinistres.

Analyse univariée

Dans un premier temps, la répartition de la fréquence par modalité, telle que présentée dans le tableau 3.4, est analysée pour les autres variables des différentes cartes.

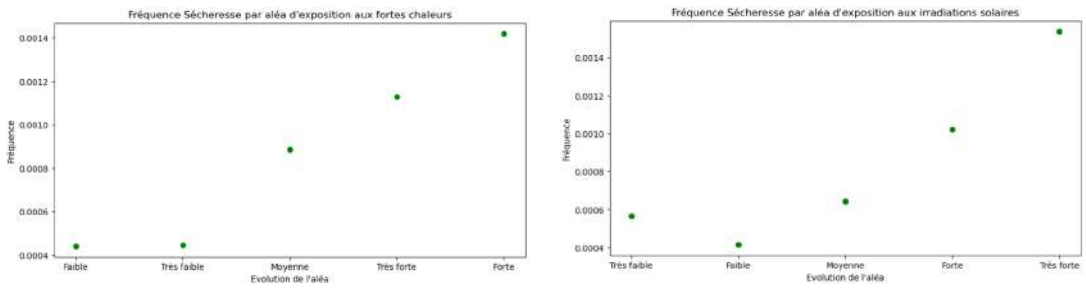


FIGURE 3.14 – Fréquence par modalité — carte_chaleur (à gauche) et carte_irrad (à droite)

Les graphiques de la figure 3.14 montrent que la fréquence des sinistres n'est pas strictement croissante en fonction du risque indiqué par les modalités. Par ailleurs, les catégories "Très faible" et "Très forte" engendre de la volatilité en raison de leur faible pourcentage d'expositions. Il est donc décidé de regrouper les modalités "Très faible" et "Faible", ainsi que les catégories "Forte" et "Très forte", afin d'obtenir une classification plus stable.

Analyse des corrélations

Dans la continuité des travaux préliminaires, il est judicieux d'examiner les corrélations entre les différentes variables de la base de données. Cette analyse permet de visualiser les interactions entre les variables ou d'éliminer potentiellement des variables redondantes en termes d'information. La matrice de corrélations présentée en figure 3.15 permet d'analyser ces relations. Il est observé que, de manière générale, la carte_rga est très peu corrélée avec les autres cartes. En revanche, la carte_chaleur et la carte_irrad apparaissent assez

corrélées entre elles, ce qui est cohérent, étant donné que l'une concerne les fortes chaleurs et l'autre l'irradiation solaire. Pour la carte_temp et la carte_pluie, les corrélations entre les variables indiquent que la température moyenne décroît avec l'augmentation des précipitations.

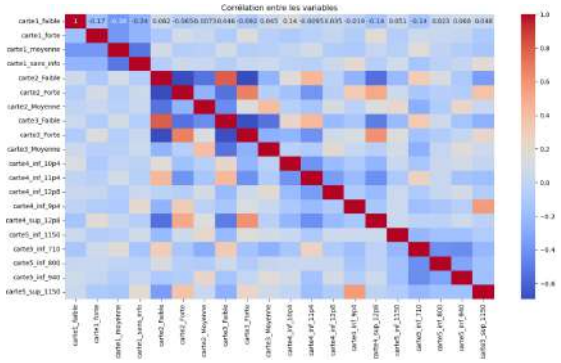


FIGURE 3.15 – Matrice de corrélations

Séparation train-test

En apprentissage statistique, il est courant de diviser la base de données en deux parties : une base d'apprentissage (*train*), sur laquelle le modèle est entraîné, et une base de test, qui permet de valider les performances du modèle. Cette pratique vise à rendre le modèle plus robuste et à éviter le surapprentissage. Dans cette étude, 70 % de la base de données a été utilisée pour constituer la base d'apprentissage, tandis que les 30 % restants ont été réservés pour la base de test. Afin de garantir que la sélection aléatoire des 70 % de données pour la base d'apprentissage ne biaise pas les résultats, une graine a été utilisée. La graine permet de contrôler le processus de sélection aléatoire, garantissant ainsi que, même si le processus est aléatoire, il peut être reproduit à l'identique lors d'exécutions ultérieures.

Dans ce contexte, l'utilisation de la graine permet de choisir une base d'apprentissage et une base de test avec des fréquences de sinistres équivalentes, assurant ainsi une meilleure représentativité des données. En effet, les bases d'apprentissage et de test doivent constituer des échantillons représentatifs de la base globale. Cela signifie que les caractéristiques des assurés et des sinistres doivent être équilibrées entre les deux bases. Si les bases ne sont pas représentatives, il existe un risque d'obtenir deux sous-ensembles de données qui reflètent des cas extrêmes. Par exemple, une base pourrait contenir principalement des assurés à faible risque tandis que l'autre pourrait contenir surtout des assurés à risque élevé. Le modèle sera alors entrainé sur un échantillon peu représentatif de la réalité, et sera testé sur un échantillon ayant le même défaut. Le modèle qui en résulte pourrait alors ne pas être pertinent.

Bien que plusieurs autres règles auraient pu être utilisées pour éviter les biais lors

de la constitution des bases de *train* et de test, elles n'ont pas été appliquées dans le cadre de ce mémoire. En particulier, il est essentiel que le même individu (ou le même contrat d'assurance) ne soit pas présent à la fois dans la base d'apprentissage et dans la base de test. Si un individu est utilisé pour entraîner le modèle et pour le tester, cela peut introduire un biais appelé *leakage*, où le modèle a "vu" certaines informations qu'il n'aurait pas dû voir, faussant ainsi l'évaluation de ses performances. Pour éviter cela, on utilise des méthodes de validation croisée où les données sont systématiquement réparties en plusieurs sous-ensembles sans recouvrement entre les bases d'entraînement et de validation.

Choix des modalités de référence

Par la suite, le GLM est paramétré de manière à prendre comme modalité de référence celle qui présente le plus d'expositions. En choisissant la modalité la plus fréquente dans le portefeuille comme référence, les coefficients des autres modalités s'interprètent comme des écarts par rapport à la situation la plus courante. Cela facilite la compréhension et la communication des résultats, en particulier auprès des non-statisticiens. De plus, dans le domaine de l'assurance, il est courant de comparer les risques par rapport à un profil "standard" ou typique, qui correspond souvent à la catégorie la plus fréquente.

	coef	**std err**	**z**	**P>|z|**	**[0.025, 0.975]**
Intercept	-5.9418	0.075	-79.360	0.000	[-6.089, -5.795]
C(carte_rga, (ref='moyenne'))[T.faible]	-1.1617	0.183	-6.363	0.000	[-1.520, -0.804]
C(carte_rga, (ref='moyenne'))[T.sans_info]	-0.1454	0.096	-1.515	0.130	[-0.333, 0.043]
C(carte_rga, (ref='moyenne'))[T.forte]	0.5634	0.081	6.924	0.000	[0.404, 0.723]
C(carte_chaleur, (ref='Faible'))[T.Moyenne]	0.0632	0.156	0.405	0.685	[-0.243, 0.369]
C(carte_chaleur, (ref='Faible'))[T.Forte]	0.0940	0.156	0.603	0.546	[-0.212, 0.400]
C(carte_irrad, (ref='Faible'))[T.Moyenne]	-0.3117	0.152	-2.046	0.041	[-0.610, -0.013]
C(carte_irrad, (ref='Faible'))[T.Forte]	0.4811	0.159	3.032	0.002	[0.170, 0.792]
C(carte_temp, (ref='inf_11p4'))[T.inf_9p4]	-0.8000	0.312	-2.568	0.010	[-1.411, 0.190]
C(carte_temp, (ref='inf_11p4'))[T.inf_10p4]	-0.0481	0.120	-0.402	0.688	[-0.283, 0.187]
C(carte_temp, (ref='inf_11p4'))[T.inf_12p8]	0.1039	0.124	0.837	0.402	[-0.139, 0.347]
C(carte_temp, (ref='inf_11p4'))[T.sup_12p8]	0.3786	0.132	2.875	0.004	[0.120, 0.637]
C(carte_pluie, (ref='inf_710'))[T.inf_800]	0.0160	0.090	-0.178	0.859	[-0.192, 0.160]
C(carte_pluie, (ref='inf_710'))[T.inf_940]	0.0115	0.096	0.119	0.905	[-0.177, 0.200]
C(carte_pluie, (ref='inf_710'))[T.inf_1150]	-0.3522	0.166	-2.118	0.034	[-0.678, -0.026]
C(carte_pluie, (ref='inf_710'))[T.sup_1150]	-0.3595	0.216	-1.661	0.097	[-0.784, 0.065]

TABLE 3.6 – Résumé du modèle PC (sur python)

Analyse des résultats

À un niveau de confiance de 95 %, les modalités de la carte_rga sont significatives (p-value P>|z| inférieure à 5%), à l'exception d'une seule. Cette exception justifie un regroupement de la modalité 'sans info' avec la catégorie de référence 'moyenne' pour améliorer la stabilité du modèle. De plus, comme mentionné précédemment, la carte_rga montre une faible corrélation avec les autres cartes, ce qui indique qu'elle apporte des informations distinctes. Elle est conservée dans le modèle.

En revanche, la carte_chaleur ne contient aucune modalité significative et ses modalités sont fortement corrélées avec celles de la carte_irrad. Pour éviter une redondance d'information, il est décidé de ne conserver qu'une seule des deux cartes. La carte_chaleur est donc retirée, car les deux modalités de la carte_irrad sont significatives.

La moitié des modalités de la carte_temp sont significatives et les intervalles de confiance associés à ces modalités ne contiennent pas 0, ce qui indique que leurs coefficients ne peuvent pas être nuls. Cependant, l'analyse univariée (figure 3.16) montre que la répartition de la fréquence par modalité est instable. Deux approches sont envisagées : retirer cette carte ou regrouper toutes les modalités non significatives avec la référence. Ces deux options sont explorées dans des modèles distincts, RI (sans la carte_temp) et RIT (avec la carte_temp).

Enfin, bien qu'une modalité de la carte_pluie soit significative, l'analyse univariée (figure 3.16) confirme que la carte ne présente pas de répartition cohérente des fréquences par modalité. De plus, les coefficients β de cette carte ne suivent pas une stricte monotonie (croissance ou décroissance des coefficients selon les modalités). Ces résultats montrent que la carte_pluie n'apporte pas de valeur ajoutée et que l'outil a correctement extrait les données de cette carte, qui est donc retirée.

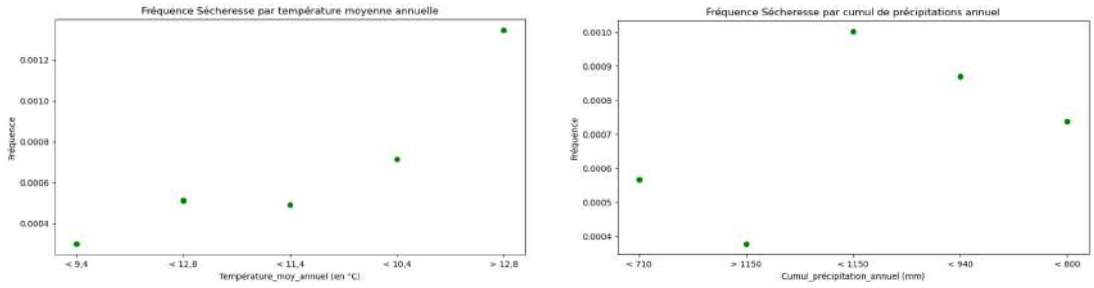


FIGURE 3.16 – Fréquence par modalité — carte_temp (à gauche) et carte_pluie (à droite)

Interprétation des résultats

L'interprétation des variables significatives du modèle PC met en évidence leur rôle clé dans la modélisation de la fréquence du risque de sécheresse. La carte_rga est déter-

minante puisqu'elle capture directement les effets liés à la nature du sol qui joue un rôle central dans la survenance de sinistres de sécheresse. La carte_irrad, en lien avec l'irradiation solaire, confirme que les zones exposées à de fortes irradiations sont particulièrement vulnérables aux mouvements des sols. On peut supposer que ces zones, davantage exposées au soleil, subissent un assèchement plus fréquent des sols, ce qui favorise l'apparition du phénomène de rétractation des sols. L'explication pourrait être la même pour la carte_temp, car les épisodes de chaleur intense favorisent la rétractation des sols.

Modèles RI et RIT. Le modèle RI (Retrait-gonflement d'argile et Irradiation solaire) conserve uniquement la carte_rga et la carte_irrad, avec de nouvelles modalités regroupées en "faible", "moyenne" et "forte". Le modèle RIT (Retrait-gonflement d'argile, Irradiation solaire et Température moyenne), quant à lui, ajoute la carte_temp aux variables du modèle RI, en réduisant le nombre de modalités de 5 à 3. Il en résulte que toutes les modalités restantes sont significatives (voir tableau 3.7 pour l'exemple du modèle RIT).

	coef	**std err**	**z**	**P>|z|**	**[0.025, 0.975]**
Intercept	-6.0000	0.055	-108.293	0.000	[-6.109, -5.891]
C(carte_rga, (ref='moyenne'))[T.faible]	-1.1473	0.179	-6.414	0.000	[-1.498, -0.797]
C(carte_rga, (ref='moyenne'))[T.forte]	0.6570	0.075	8.732	0.000	[0.510, 0.804]
C(carte_irrad, (ref='Faible'))[T.Forte]	0.4986	0.100	4.972	0.000	[0.302, 0.695]
C(carte_irrad, (ref='Faible'))[T.Moyenne]	-0.2960	0.121	-2.454	0.014	[-0.532, -0.060]
C(carte_temp, (ref='inf_12p8'))[T.inf_9p4]	-1.0819	0.269	-4.019	0.000	[-1.610, -0.554]
C(carte_temp, (ref='inf_12p8'))[T.sup_12p8]	0.3972	0.098	4.043	0.000	[0.205, 0.590]

TABLE 3.7 – Résumé du modèle RIT (sur python)

Les coefficients du modèle RIT associés aux modalités de la carte_rga reflètent l'ordre croissant du risque de sécheresse suivant le risque de retrait-gonflement d'argiles. Le coefficient de -1.15 pour la modalité "faible" traduit une fréquence de sinistres plus faible que celle associée à la modalité de référence "moyenne" (qui a un coefficient de 0). À l'inverse, le coefficient positif 0.65 pour la modalité "forte" indique une fréquence de sinistres plus élevée, confirmant ainsi que la fréquence des sinistres augmente avec le niveau de risque indiqué par la carte_rga.
Pour la carte_temp, les coefficients du modèle RIT suivent une logique similaire, indiquant que les zones à température moyenne élevée sont plus susceptibles de connaître des sinistres liés à la sécheresse.
La carte_irrad, quant à elle, présente un résultat plus nuancé. Il faut rappeler que le modèle de Poisson utilisé emploie une fonction de lien logarithmique, ce qui signifie que la fonction exponentielle doit être appliquée aux coefficients pour interpréter les fréquences. Bien que la modalité "Forte" soit associée à une fréquence plus élevée de sinistres (exp(0.4986) = 1.646), la modalité "Moyenne" présente paradoxalement une fréquence inférieure à celle de "Faible" (exp(-0.2960) = 0.744). Cela suggère que la relation entre l'irradiation solaire et la fréquence des sinistres est plus complexe que ne le montre le

modèle actuel.

Indicateurs	**Modèle RI**	**Modèle RIT**	**Observée**
Déviance	10 143	10 084	
AIC	11 554,98	11 505,65	
BIC	-3 338 271	-3 338 300	
Fréquence(%)	0,323	0,323	0,323
RMSE	0,06508	0,06507	

TABLE 3.8 – Tableau de synthèse des indicateurs de performance sur la base de *train*

Le tableau 3.8 montre que le modèle RIT (incluant la carte_temp) surpasse légèrement le modèle RI (sans la carte_temp) selon plusieurs indicateurs de performance. La déviance, l'AIC et le BIC sont légèrement inférieurs pour le modèle RIT, suggérant un meilleur ajustement au coût d'une complexité raisonnable. Le RMSE est également un peu plus bas, ce qui indique une précision de prédiction améliorée.
La carte_temp, qui capte l'effet des températures moyennes, apporte donc des informations suffisamment pertinentes vis-à-vis du phénomène de sécheresse, pour rendre le modèle RIT plus précis que le modèle RI dans la modélisation la fréquence des sinistres.

Constitution des zones. Dans l'approche standard de constitution d'un zonier, ce sont les résidus d'un modèle sans variable géographique qui sont modélisés à partir des variables géographiques. Dans le cadre de ces travaux, la construction du zonier est simplifiée. Une fois les coefficients β obtenus, il devient possible de calculer la fréquence prédite pour chaque combinaison des modalités des variables explicatives. Par exemple, pour un contrat couvrant un logement situé dans une zone classée "faible" par la carte_rga, "faible" par la carte_irrad et "inf_9p4" par la carte_temp, la fréquence prédite peut être estimée en utilisant les coefficients correspondants :

$$\text{fréquence} = exp(\text{Intercept} + \beta_{\text{carte\_rga.faible}} + \beta_{\text{carte\_irrad.Faible}} + \beta_{\text{carte\_temp.inf\_9p4}})$$

$$\text{fréquence} = exp(-6 + (-1.1473) + 0 + (-1.0819)) = 0.0002667$$

Suivant ce raisonnement, un tableau regroupant toutes les combinaisons de modalités possibles et les fréquences correspondantes est établi. En pratique, les combinaisons de modalités absentes de la base de risque sont retirées. À partir de ce tableau, un algorithme de classification ascendante hiérarchique est utilisé pour regrouper les combinaisons de modalités en groupes homogènes de fréquence. Cela permet d'obtenir cinq groupes, nommés A, B, C, D et E, par ordre de fréquence prédite moyenne croissante.
Pour des raisons de confidentialité, les expositions présentées ici et dans le reste du mémoire ont été retraitées !

Zones	Expos(%)	Freq(%)	PrMoy CatNat(€)	PP sech(€)	Pseudo-S/P
A	28,97	0,10	24,02	12,48	0,52
B	48,93	0,32	24,20	40,29	1,66
C	17,5	0,54	26,49	67,71	2,56
D	1,55	0,75	26,68	94,28	3,53
E	3,05	1,17	28,17	147,36	5,23

TABLE 3.9 – Pseudo-S/P par zone - Base d'apprentissage (modèle RIT)

Par analogie au tableau 3.5, le tableau 3.9 permet d'observer les fréquences de chaque zone, les primes pures (calculées en multipliant les fréquences par un coût moyen unique) et les coefficients (pseudo-S/P). Il en ressort que la zone B est la plus représentée dans le portefeuille de contrats, et que la fréquence ainsi que la prime pure obtenue suivent une progression croissante selon les zones. La figure 3.17 ci-dessous suggère que cette domination de la zone B dans le portefeuille, vient de la présence de l'Île-de-France dans cette zone.

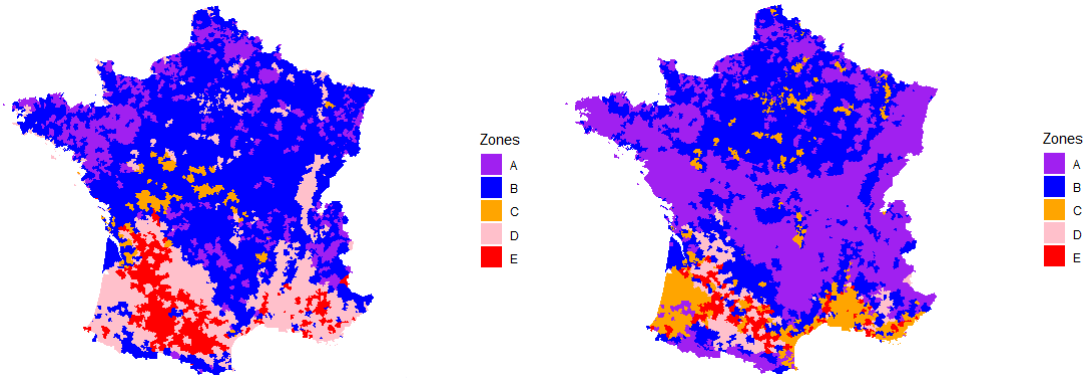


FIGURE 3.17 – Zoniers RI (à gauche) et RIT (à droite)

Le modèle RI ne prend en compte que la carte des sols argileux (carte_rga) et la carte d'irradiation solaire (carte_irrad). Cela explique pourquoi les zones les plus à risque (D et E) se concentrent principalement dans le sud-est et le sud-ouest de la France, où l'exposition au risque de retrait-gonflement des argiles est la plus forte. La carte_rga montre un risque élevé dans ces régions, et influence fortement le modèle RI, ce qui se traduit par des zones D et E au sud-ouest et au sud-est.
Le modèle RIT, quant à lui, ajoute la carte de la température moyenne (carte_temp) à la modélisation. Le sud-est de la France, qui était principalement en zones D et E dans le modèle RI, est désormais beaucoup moins marqué dans le modèle RIT. Cela pourrait signifier que les températures moyennes élevées, caractéristiques du sud-est de la France, n'ont pas apporté d'informations nouvelles ou discriminantes au modèle. En d'autres termes, les zones qui étaient déjà classées comme à haut risque dans le modèle RI n'ont pas vu leur classification en zones C ou D renforcée par l'ajout de la carte_temp. Cela a

conduit à une réévaluation et à une redistribution du risque, avec certaines zones étant reclassées dans des niveaux de risque moins élevés (par exemple, zones C ou B).
Une autre observation est la migration des zones C plus au sud dans le modèle RIT. Dans le modèle RI, les zones C étaient largement présentes dans le centre de la France. Avec l'ajout de la carte_temp, ces zones ont été repoussées plus au sud. Cela s'explique par le fait que le centre de la France est moins affecté par des températures extrêmes, ce qui a probablement réduit le risque modélisé pour ces régions.
Enfin, la zone A (risque le plus faible) a connu une expansion significative dans le modèle RIT, notamment dans le nord et le centre de la France. Cela pourrait s'expliquer par le fait que ces régions, historiquement moins exposées à des températures élevées ou à des niveaux élevés d'irradiation solaire, se trouvent davantage renforcées par l'ajout de la carte_temp, qui réduit encore leur niveau de risque global.

Comparaison des modèles RI et RIT. Sur la figure 3.18 ci-dessous, le graphique de gauche est issu du modèle RI et celui de droite du modèle RIT.

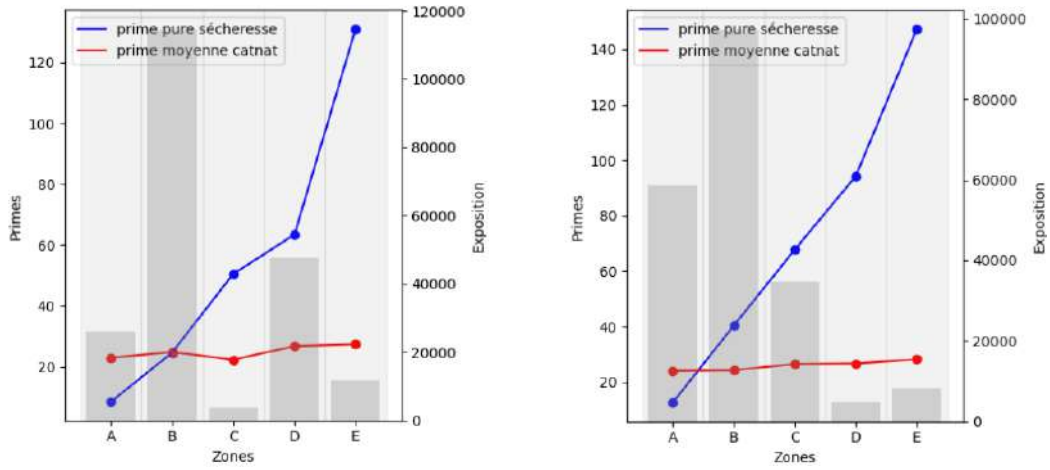


FIGURE 3.18 – Prime pure sécheresse et Prime moyenne CatNat par zones sur la base d'apprentissage

Les graphiques montrent la prime pure sécheresse (en bleu) et la prime moyenne CatNat (en rouge) par zones pour la base d'apprentissage des modèles RI et RIT, avec les expositions représentées par des barres grises. Dans les deux modèles, la prime CatNat reste stable à travers les zones, reflétant son caractère non segmenté. Cependant, la prime pure sécheresse augmente dans les deux cas, mais de manière plus marquée dans le modèle RIT, notamment dans les zones D et E, où elle dépasse 140. Cela suggère que le modèle RIT capture mieux les risques dans les zones à haut risque que le modèle RI, en prévoyant des primes plus élevées. Les expositions restent similaires dans les deux modèles, avec une forte concentration dans la zone B, mais des primes plus élevées dans les zones D et E, où l'exposition est plus faible.

Par ailleurs, le modèle RI, qui n'intègre pas la carte_temp, présente un coefficient de Gini de 0,319 sur la base d'apprentissage. Ce coefficient reflète une capacité de discrimination modérée, signifiant que le modèle parvient à segmenter les risques, mais de manière limitée. Lorsque la carte_temp est ajoutée dans le modèle RIT, le coefficient de Gini s'élève légèrement à 0,327, indiquant une meilleure segmentation des risques. Cette augmentation montre que le modèle RIT possède un pouvoir discriminant supérieur, en capturant plus finement les différences de risque entre les zones. Ainsi, l'intégration de la carte_temp a renforcé la capacité du modèle à identifier les zones à risque plus élevé.

Validation des modèles RI et RIT. Après avoir calibré et étudié les modèles RI et RIT à partir de la base de *train*, il est important d'analyser les résultats de ces modèles sur la base de test. Cette analyse a pour but de valider la robustesse des modèles, notamment en vérifiant qu'il n'y a pas eu de surapprentissage.
Pour des raisons de confidentialité, les expositions présentées ici et dans le reste du mémoire ont été retraitées !

Indicateurs	**Bases**	**Modèle RI**	**Modèle RIT**	**Observée**
Fréquence(%)	*Train*	0,323	0,323	0,323
Fréquence(%)	Test	0,3249	0,3254	0,3347
RMSE	*Train*	0,06508	0,06507	
RMSE	Test	0,06633	0,06632	

TABLE 3.10 – Tableau de synthèse des indicateurs de performance sur la base de test

Le tableau 3.10 montre qu'il n'y a pas d'écarts significatifs entre les principales métriques (fréquence des sinistres et RMSE) des bases d'apprentissage et de test, validant donc un premier critère de stabilité et de robustesse du modèle, ainsi que le découpage des données en base de *train* et de test.

Zones	**Expos(%)**	**Freq(%)**	**PrMoy CatNat(€)**	**PP sech(€)**	**Pseudo-S/P**
A	27,59	0,14	17,87	22,96	0,78
B	49,35	0,30	38,32	24,85	1,54
C	17,72	0,47	58,79	27,08	2,17
D	1,83	1,16	145,95	25,65	5,69
E	3,51	1,36	170,95	28,19	6,06

TABLE 3.11 – Pseudo-S/P par zone - Base de test (modèle RIT)

Les proportions de contrats par zone restent cohérentes avec la base d'apprentissage, avec la Zone B représentant près de 49% des contrats, suivie par la Zone A à environ 28%. Cela confirme que la répartition des expositions reste similaire entre les bases d'apprentissage et de test, et que la constitution des bases d'apprentissage et de test a été réalisée de manière efficace.
Par contre, les fréquences des sinistres observées sur la base de test montrent des écarts par rapport à la base d'apprentissage. En Zone A, la fréquence est plus élevée sur la base de test, indiquant un nombre de sinistres supérieur aux prévisions. Les Zones B et C affichent des fréquences légèrement inférieures, tandis que les Zones D et E montrent une augmentation notable, signalant un risque plus élevé que prévu dans ces zones.
Les primes pures sécheresse, calculées en multipliant les fréquences modélisées par le coût moyen unique, suivent la même tendance. En ce qui concerne les primes moyennes CatNat, elles sont assez stables à travers les zones, avec de légères variations par rapport à la base d'apprentissage. Cependant, ces variations sont relativement mineures et n'ont pas d'impact significatif sur l'analyse générale.
En résumé, l'analyse du tableau pour la base de test montre que les fréquences des sinistres et les primes pures observées suivent globalement la tendance attendue, avec une augmentation régulière du risque selon les zones. Toutefois, certaines zones, comme D et E, montrent des fréquences beaucoup plus élevées que celles prévues par le modèle d'apprentissage, ce qui conduit à des primes pures plus élevées.

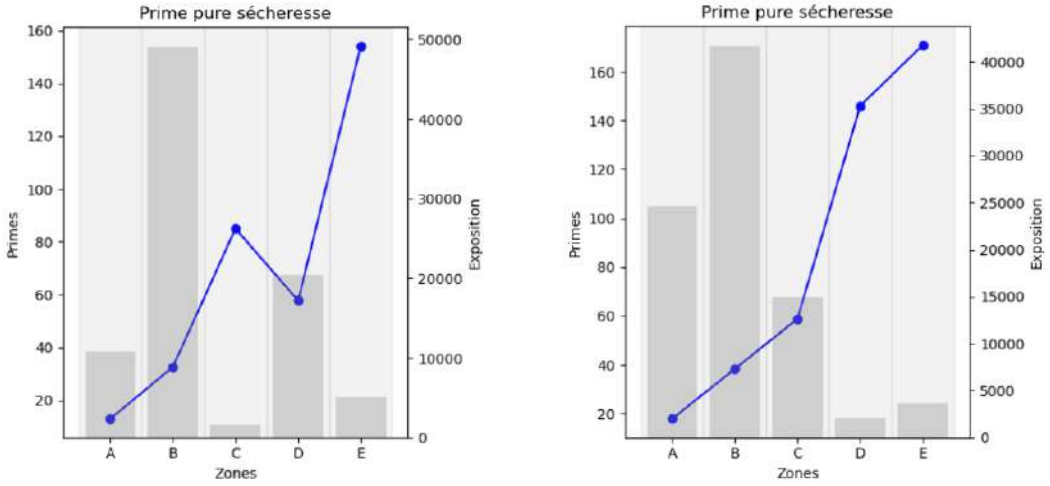


FIGURE 3.19 – Prime pure sécheresse et Prime moyenne CatNat par zones sur la base de test

Les graphiques des modèles RI et RIT sur la base de test présentent des tendances similaires à celles observées sur la base d'apprentissage. Dans le modèle RI, la prime pure sécheresse suit une progression croissante entre les zones A et E, mais un point atypique est la baisse de la prime entre les zones C et D. Cela indique que le modèle RI a du mal à capturer correctement la progression du risque entre ces zones spécifiques, probablement

en raison du fait que la zone C soit sous-représentée (très faible exposition). Une solution à ce problème serait d'associer les zones C et D. Le modèle RIT, en revanche, présente une progression plus régulière et continue des primes pures entre les zones. Cette cohérence sur la base de test démontre que le modèle RIT est plus robuste et stable dans sa capacité à capturer le risque à travers les différentes zones.
Pour faire une comparaison avec le modèle 1C (voir figure 3.13), qui est basé sur une seule carte géographique, on remarque que ce dernier montre des primes pures plus faibles dans les zones à risque élevé (C et D) par rapport aux modèles RI et RIT. Bien qu'il suive une tendance générale d'augmentation du risque entre les zones, il manque de précision pour les zones les plus risquées, ce qui peut être attribué à la limitation d'utiliser une seule variable explicative. Les modèles RI et RIT, avec l'ajout de variables supplémentaires, capturent mieux les variations de risque et produisent des primes pures mieux distribuées, couvrant une plus grande étendue de valeurs, particulièrement dans les zones à risque élevé comme D et E.

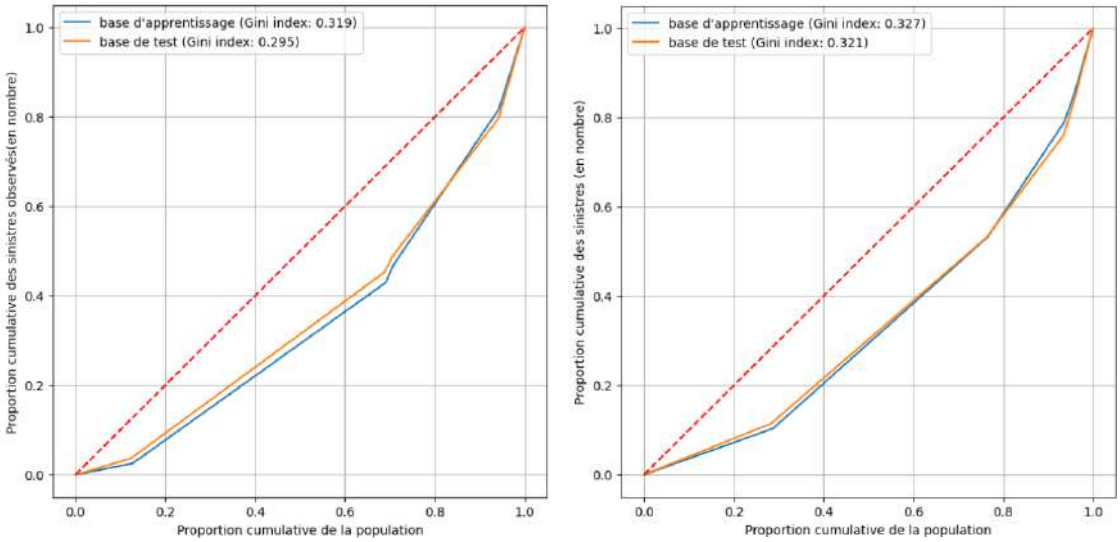


FIGURE 3.20 – Courbe de Lorenz et indicateur de Gini - Modèles RI (à gauche) et RIT (à droite)

Les graphiques des courbes de Lorenz des modèles RI et RIT, sur les bases d'apprentissage et de test, montrent une faible déperdition de performance discriminante entre les deux bases. Pour le modèle RI, le coefficient de Gini passe de 0.319 sur la base d'apprentissage à 0.295 sur la base de test, tandis que pour le modèle RIT, il passe de 0.327 à 0.321. Cette faible variation suggère que les deux modèles sont robustes. Le modèle RIT, en particulier, montre une stabilité accrue, avec une moindre déperdition de performance discriminante entre les deux bases, confirmant ainsi sa robustesse.

3.5 Synthèse des différents modèles

En comparant les modèles RI et RIT, on observe que l'ajout de la carte_temp dans le modèle RIT a amélioré la capacité du modèle à capturer les variations de risque. En effet, le coefficient de Gini plus élevé du modèle RIT (0.327 sur la base d'apprentissage) par rapport au modèle RI (0.319) indique un meilleur pouvoir discriminant. Cette amélioration est en grande partie due à la carte_temp, qui fournit des informations supplémentaires sur les températures moyennes, un facteur clé influençant le phénomène de sécheresse. La carte_temp permet de mieux identifier les zones à risque élevé, entraînant une répartition plus inégale des sinistres, mais également une segmentation plus fine du portefeuille.
Par rapport au modèle 1C, qui repose uniquement sur la carte_rga (retrait-gonflement des argiles), les modèles RI et RIT montrent une meilleure capacité à discriminer les risques grâce à l'intégration de plusieurs variables explicatives. Le coefficient de Gini de 0.275 du modèle 1C montre une répartition plus homogène des sinistres, mais cela reflète une segmentation moins précise. Le modèle 1C, en s'appuyant uniquement sur une carte géographique, manque de profondeur pour capturer les complexités du phénomène de sécheresse. En revanche, les modèles RI et RIT, en incluant la carte_temp et la carte_irrad, offrent une meilleure segmentation, capturant des informations supplémentaires liées à la fréquence des sinistres de sécheresse, comme l'effet de la chaleur et de l'irradiation solaire sur la contraction des sols argileux.
En prenant en compte les tableaux de fréquences, les courbes de Lorenz, et les coefficients de Gini, le modèle RIT se distingue comme le plus performant. Son coefficient de Gini légèrement plus élevé que celui des autres modèles reflète une meilleure capacité à segmenter les zones à risque élevé, permettant ainsi des primes pures plus étendues et mieux distribuées. Le modèle RIT propose une modélisation plus précise et cohérente du risque de sécheresse, tout en offrant la meilleure segmentation. En intégrant la carte_temp, il parvient à capturer les nuances géographiques avec plus de précision, en faisant le modèle le plus performant pour ce cas d'étude

Codes postaux	**Zones**
01000	B
01090	B
01100	C
01110	B
01120	B
⋮	⋮
95850	B
95870	B
95880	A

TABLE 3.12 – Aperçu du zonier obtenu avec le meilleur modèle

3.6 Discussions

Dans cette partie du mémoire, il est question de répondre à plusieurs interrogations concernant les travaux effectués et d'apporter une réflexion plus approfondie aux interprétations déjà fournies.

Intérêt d'une segmentation interne (risque sécheresse). Bien que la prime CatNat ne soit pas segmentée en raison de la réglementation, empêchant ainsi toute segmentation interne directe de cette prime, une répartition du portefeuille par zones de risque pourrait présenter de nombreux avantages pour une compagnie d'assurances, en plus de ceux évoqués en début de chapitre.
On pourrait s'imaginer un assureur proposant des garanties complémentaires spécifiques à la CatNat. Par exemple, une garantie "Sécheresse étendue" pourrait être introduite et proposée aux contrats dans les zones les plus à risque. Cela offrirait aux assurés une protection supplémentaire adaptée aux risques locaux non couverts par la garantie CatNat standard.
De plus, la segmentation facilite une meilleure gestion des sinistres. En identifiant les zones à haut risque, la compagnie peut sensibiliser ses conseillers sur les spécificités de ces zones, leur permettant d'adapter leur accompagnement en fonction des risques particuliers liés à la sécheresse. Cette sensibilisation proactive permet d'optimiser les processus de traitement des sinistres, tout en améliorant la satisfaction des clients dans les zones les plus exposées.
Cette segmentation est également bénéfique dans les relations avec les réassureurs. En ayant une compréhension approfondie des zones à risque au sein du portefeuille, l'assureur peut démontrer aux réassureurs sa maîtrise des risques encourus. Cela renforce la crédibilité de l'assureur et facilite les négociations lors de l'achat de couvertures de réassurance, en montrant qu'il est en pleine compréhension des spécificités de son portefeuille, ce qui est un atout précieux aux yeux des réassureurs.

En résumé, même si la prime CatNat reste non segmentée, une segmentation interne basée sur le risque sécheresse permettrait à l'assureur d'améliorer sa gestion du risque, tant en matière de garanties complémentaires, de gestion des sinistres, que de réassurance. Ces actions concrètes contribuent à une meilleure maîtrise du risque sécheresse.

Biais engendré par l'extraction de variables. L'extraction de variables géographiques à partir de cartes peut introduire des biais dans la modélisation des sinistres. En effet, aucun algorithme d'extraction n'est parfait à 100 %, ce qui signifie qu'il y a toujours un risque d'erreurs dans les données obtenues. Ces erreurs, même si elles sont faibles, peuvent avoir un impact sur la manière dont les risques sont modélisés, en influençant par exemple les estimations des sinistres pour certaines zones géographiques.

Cependant, ce risque peut être atténué en utilisant un grand nombre de variables géographiques. En combinant plusieurs sources d'information, les erreurs d'une variable

spécifique peuvent être compensées par la fiabilité d'autres variables. Cela réduit l'effet des biais individuels et renforce la robustesse du modèle global. En d'autres termes, plus le modèle intègre de variables, plus il est capable de minimiser l'impact des éventuelles erreurs provenant de l'extraction de données géographiques.
Cet avantage est illustré dans le cas pratique avec les modèles 1C, RI, et RIT, où une multitude de cartes ont été transformées en bases de données. Bien que seules cinq cartes aient finalement été retenues pour la modélisation, l'ajout progressif de variables a démontré son utilité en permettant de construire des modèles de plus en plus performants. L'inclusion de la carte_temp et de la carte_irrad dans le modèle RIT a non seulement permis de mieux capturer la complexité du risque de sécheresse, mais a également abouti à une meilleure segmentation des zones à risque. Cela montre que la multiplicité des variables, malgré le risque initial de biais, peut finalement conduire à une modélisation plus précise et robuste.

Autres applications potentielles de l'outil. L'outil d'extraction de variables géographiques à partir de cartes offre de nombreuses possibilités au-delà de la modélisation du risque sécheresse pour la segmentation de la prime CatNat. Il peut être utilisé pour modéliser divers autres risques naturels tels que les inondations, les tempêtes et les glissements de terrain, en analysant des facteurs comme la proximité des cours d'eau, l'altitude, la topographie et l'exposition au vent. Cet outil peut permettre également de mieux calibrer les couvertures de réassurance en identifiant les zones géographiques à fort risque, facilitant ainsi l'achat de couvertures spécifiques pour protéger les portefeuilles exposés à des événements climatiques récurrents. Par ailleurs, il peut aider à simuler des scenarii catastrophes, en évaluant l'impact potentiel de ces événements sur l'ensemble du portefeuille d'assurances et en testant la robustesse des réserves et des stratégies de réassurance. En somme, cet outil polyvalent permet aux assureurs d'affiner leur compréhension des risques, d'améliorer leur tarification et de développer des stratégies plus ciblées dans divers domaines de l'assurance.

L'outil développé n'est pas limité aux seules cartes de la France. Il pourrait également être utilisé avec des cartes d'autres pays ou spécifiques à une région particulière. Bien qu'un travail préalable soit nécessaire pour reconstruire une base de données et une carte de référence adaptées, cette flexibilité élargit considérablement les possibilités d'utilisation. De plus, les applications de l'outil ne se restreignent pas uniquement au domaine de l'actuariat, mais peuvent également être exploitées par toute entité travaillant sur des données géographiques.

Conclusion

Dans un contexte de changement climatique et d'intensification des catastrophes naturelles, il devient nécessaire pour les compagnies d'assurance d'accéder à des données de qualité pour mieux modéliser ces risques. De cette problématique est né ce sujet de mémoire qui propose la conception d'un outil algorithmique d'extraction de variables géographiques à partir d'images de cartes. L'objectif principal de cet outil est de fournir un moyen supplémentaire d'obtenir des données géographiques, essentielles pour améliorer la modélisation des risques naturels, et dans ce mémoire, en particulier, le risque de sécheresse.

L'outil développé a été conçu avec l'idée de transformer des images de cartes géographiques en données exploitables par des modèles actuariels. Il repose sur plusieurs étapes clés : la reconnaissance des éléments cartographiques, leur conversion en variables catégorielles, puis leur intégration dans des bases de données utilisables pour la modélisation des sinistres. Les performances de l'outil ont été testées à travers des méthodes de classification et de segmentation, montrant une classification correcte des communes dans plus de 85 % des cas. Bien que des erreurs d'extraction subsistent, ces résultats prometteurs démontrent l'efficacité de l'outil pour enrichir les bases de données actuarielles avec des variables géographiques supplémentaires. Cette capacité à fournir des données fines et contextualisées se révèle particulièrement utile dans la modélisation du risque de sécheresse, comme en témoigne le cas pratique réalisé au cours de cette étude, où l'ajout de ces variables a permis d'améliorer la précision des prévisions et d'affiner la segmentation des zones à risque.
Ce cas pratique s'inscrit dans la stratégie globale de l'entreprise, qui vise à développer un modèle de risque complet, intégrant des variables techniques et des données externes pour estimer la prime pure technique d'un contrat. L'outil contribue à cet objectif en fournissant une méthode innovante d'extraction de données géographiques, particulièrement pertinente pour la modélisation des risques liés à la sécheresse. Bien que la prime CatNat ne puisse pas être modifiée, l'outil permet de créer un zonier simplifié pour la sécheresse, utile pour une meilleure gestion des risques.

Cependant, des limitations subsistent principalement en raison des erreurs d'extraction liées à l'algorithme. Bien que ces erreurs soient généralement mineures, elles peuvent introduire des biais dans la modélisation des sinistres. L'amélioration continue de l'algo-

rithme est donc essentielle pour minimiser ces biais et renforcer la fiabilité des données extraites. Pour ce faire, des techniques plus avancées, comme les réseaux de neurones appliqués à la reconnaissance des éléments cartographiques, pourraient être explorées afin d'accroître la précision de l'extraction et d'enrichir encore davantage les bases de données utilisées en assurance. L'extension de cet outil à d'autres types de catastrophes naturelles, comme les inondations, constituerait également une piste intéressante pour élargir son champ d'application et renforcer son utilité pour Cardif IARD.

En somme, ce mémoire démontre la pertinence et le potentiel de cet outil d'extraction de données géographiques, tout en soulignant les défis techniques à relever et les nombreuses possibilités d'évolution future. Il constitue une contribution prometteuse à l'amélioration de la gestion des risques.

Annexe A

Système Lambert 93 : théorie mathématique

A.1 Définitions

A.1.1 Lambert 93

Lambert 93 est une projection cartographique conique conforme utilisée principalement en France métropolitaine. Il s'agit d'une projection conique conforme sécante, ce qui signifie que la surface de l'ellipsoïde est projetée sur un cône qui coupe l'ellipsoïde en deux parallèles. Cette projection a été adoptée pour remplacer les anciennes projections Lambert en raison de sa meilleure précision pour les applications géographiques modernes.

A.1.2 Latitude géographique

La latitude géographique est l'angle formé entre l'équateur et un point donné sur la surface de la Terre. Elle est mesurée en degrés, minutes et secondes au nord ou au sud de l'équateur. La latitude géographique d'un point est une coordonnée essentielle dans les systèmes de géolocalisation.

A.1.3 L'excentricité e de l'ellipsoïde

L'excentricité e d'un ellipsoïde est une mesure de la déviation de l'ellipsoïde par rapport à une sphère parfaite. Elle est définie par la formule :

$$e = \sqrt{1 - \left(\frac{b^2}{a^2}\right)}$$

où a est le demi-grand axe et b est le demi-petit axe de l'ellipsoïde. L'excentricité est une valeur sans dimension qui varie entre 0 (sphère parfaite) et 1 (parabole).

A.1.4 Projection conique conforme sécante

Une projection conique conforme sécante est un type de projection cartographique où la surface de l'ellipsoïde est projetée sur un cône qui coupe la surface en deux lignes parallèles. Ce type de projection préserve les angles, ce qui en fait une projection conforme, mais ne conserve pas les distances ou les surfaces. En Lambert 93, cette projection est utilisée pour minimiser les distorsions sur l'ensemble du territoire français.

A.2 Différences entre projection de Lambert 93 dans le cas tangent et dans le cas sécant

Caractéristiques	**Cas sécant**
Interaction avec l'ellipsoïde	Le cône coupe l'ellipsoïde en deux parallèles
Parallèles standards	Deux parallèles standards
Distorsion	Nulle aux deux parallèles standards, minimisée entre eux
Utilisation	Zones géographiques larges
Caractéristiques	**Cas tangent**
Interaction avec l'ellipsoïde	Le cône est tangent en un seul parallèle
Parallèles standards	Un seul parallèle standard
Distorsion	Nulle au parallèle standard, augmente en s'éloignant
Utilisation	Zones étroites en latitude

TABLE A.1 – Lambert 93 : cas sécant et tangent

A.3 Méthodologie pour passer des coordonnées géographiques à Lambert 93

A.3.1 Cas sécant

Pour convertir des coordonnées géographiques (latitude ϕ et longitude λ) en coordonnées Lambert 93 (X, Y), les étapes ci-dessous sont suivies :

1. **Calcul de la latitude isométrique** $L(\phi)$:
La latitude isométrique est une transformation de la latitude géographique tenant compte de l'excentricité de l'ellipsoïde.
Formule :

$$L(\phi) = \ln\left(\tan\left(\frac{\pi}{4} + \frac{\phi}{2}\right) \cdot \left(\frac{1 - e \cdot \sin(\phi)}{1 + e \cdot \sin(\phi)}\right)^{\frac{e}{2}}\right)$$

2. **Détermination des paramètres de projection** :
Les constantes de projection (exposant n, constante c, et coordonnées du pôle X_s, Y_s)

sont calculées en fonction des parallèles standard et du méridien central.
La projection Lambert 93 étant une projection conique conforme sécante, elle nécessite plusieurs paramètres calculés à partir des parallèles standards ϕ_1 et ϕ_2 et du méridien central λ_0.
Les parallèles standards ϕ_1 et ϕ_2 sont deux lignes de latitude spécifiques utilisées dans les projections coniques conformes, comme la projection Lambert 93, pour définir où le cône de projection coupe l'ellipsoïde (la surface de la Terre modélisée).
Dans une projection conique conforme sécante, le cône de projection n'est pas simplement tangent à l'ellipsoïde (ce qui se produirait en un seul parallèle) mais le coupe en deux parallèles distincts. Ces parallèles sont appelés parallèles standards ϕ_1 et ϕ_2. Le long de ces parallèles, il n'y a aucune distorsion des distances ou des surfaces dans la projection. Et entre ces deux parallèles, la distorsion est minimale, c'est pourquoi cette configuration est souvent utilisée pour minimiser les erreurs sur une zone géographique étendue, comme un pays.
Les paramètres de projection sont ainsi :

1. L'exposant de la projection n :

$$n = \frac{\ln\left(\frac{N(\phi_2,a,e)\cdot\cos(\phi_2)}{N(\phi_1,a,e)\cdot\cos(\phi_1)}\right)}{L(\phi_1,e) - L(\phi_2,e)}$$

2. La constante de la projection c :

$$c = \frac{N(\phi_1,a,e)\cdot\cos(\phi_1)}{n}\cdot\exp\left(n\cdot L(\phi_1,e)\right)$$

3. Les coordonnées du pôle en projection X_s, Y_s :

$$X_s = X_0$$

$$Y_s = Y_0 + c\cdot exp(-n\cdot L(\phi_0,e))$$

où :

— $N(\phi,a,e)$ est la grande normale au point ϕ, définie par :

$$N(\phi,a,e) = \frac{a}{\sqrt{1-e^2\cdot sin^2(\phi)}}$$

— a est le demi-grand axe de l'ellipsoïde.
— ϕ_0 est la latitude de l'origine.

3. **Transformation des coordonnées géographiques en coordonnées Lambert 93** :
Les coordonnées X et Y sont obtenues en utilisant les formules suivantes :

$$X = X_s + c\cdot\exp\left(-n\cdot L(\phi)\right)\cdot\sin\left(n\cdot(\lambda-\lambda_c)\right)$$

$$Y = Y_s - c \cdot \exp\left(-n \cdot L(\phi)\right) \cdot \cos\left(n \cdot (\lambda - \lambda_c)\right)$$

où :

— λ_c est la longitude d'origine par rapport au méridien d'origine

A.3.2 Cas tangent

Paramètres de la projection tangentielle

Dans le cas tangent, le cône de projection est tangent à l'ellipsoïde en un seul parallèle appelé le parallèle standard ϕ_0. Les paramètres principaux sont :

— ϕ_0 : Latitude du parallèle standard, où le cône est tangent à l'ellipsoïde.
— λ_0 : Longitude du méridien central, qui est la référence pour mesurer les longitudes.
— n : Exposant de la projection, qui dépend de ϕ_0.
— c : Constante de la projection, qui dépend de ϕ_0 et du demi-grand axe de l'ellipsoïde.
— X_0 et Y_0 : Coordonnées projetées de l'origine (souvent prises comme $X_0 = 0$ et $Y_0 = 0$).

Calcul des paramètres de projection

1. L'exposant n est calculé à partir de la latitude ϕ_0 du parallèle standard :

$$n = \sin(\phi_0)$$

2. La latitude isométrique $L(\phi)$ est une transformation de la latitude géographique ϕ_0 pour tenir compte de l'excentricité e de l'ellipsoïde. Elle est calculée par :

$$L(\phi) = \ln\left(\tan\left(\frac{\pi}{4} + \frac{\phi}{2}\right) \cdot \left(\frac{1 - e \cdot \sin(\phi)}{1 + e \cdot \sin(\phi)}\right)^{\frac{e}{2}}\right)$$

3. La constante de projection c est donnée par :

$$c = \frac{N(\phi_0, a, e) \cdot \cos(\phi_0)}{n} \cdot \exp\left(n \cdot L(\phi_0)\right)$$

Pour un point de latitude ϕ et de longitude λ, les coordonnées X et Y dans le système Lambert sont toujours données par :

$$X = X_s + c \cdot \exp\left(-n \cdot L(\phi)\right) \cdot \sin\left(n \cdot (\lambda - \lambda_c)\right)$$

$$Y = Y_s - c \cdot \exp\left(-n \cdot L(\phi)\right) \cdot \cos\left(n \cdot (\lambda - \lambda_c)\right)$$

Mais dans ce cas :

$$X_s = X_0$$

$$Y_s = Y_0 + k_0 \cdot \frac{N(\phi_0, a, e)}{\operatorname{cotan}(\phi_0)}$$

où k_0 est le facteur d'échelle à l'origine.

Annexe B

Optimisation : méthode de Powell

B.1 Introduction

La méthode de Fletcher et Powell est une méthode d'optimisation numérique sans contrainte qui fait partie de la famille des méthodes quasi-Newton. Elle est utilisée pour minimiser des fonctions différentiables sans calculer explicitement la matrice hessienne, mais en construisant progressivement une approximation de l'inverse de celle-ci.

B.2 Principe de la méthode

1. **Initialisation :** On commence avec un point initial x_0 et une matrice H_0, qui est une approximation de l'inverse de l'hessienne (souvent, on prend H_0 comme la matrice identité).

2. **Étape itérative :** À chaque itération i, on détermine la direction de recherche s_i en utilisant la matrice H_i actuelle :

$$s_i = -H_i \nabla f(x_i)$$

où $\nabla f(x_i)$ est le gradient de la fonction f au point x_i.

3. **Mise à jour du point :** On trouve ensuite un scalaire α_i qui minimise la fonction le long de la direction s_i, c'est-à-dire qu'on résout :

$$\alpha_i = \arg\min_{\alpha} f(x_i + \alpha s_i)$$

Le nouveau point est mis à jour en utilisant :

$$x_{i+1} = x_i + \alpha_i s_i$$

4. **Mise à jour de la matrice hessienne :** La matrice H_i est mise à jour pour obtenir H_{i+1}. La mise à jour suit une règle spécifique qui garantit que la nouvelle matrice respecte certaines propriétés et améliore l'approximation de l'inverse de l'hessienne. La mise à jour est donnée par :

$$H_{i+1} = H_i + A_i + B_i$$

où les matrices A_i et B_i sont définies par des relations spécifiques basées sur les vecteurs de déplacement σ_i et y_i.

B.3 Détails mathématiques supplémentaires

Les directions s_i générées par cette méthode sont **conjuguées**, ce qui signifie qu'elles sont adaptées à la géométrie de la fonction, en particulier pour les fonctions quadratiques. Cela permet de minimiser efficacement la fonction en un nombre fini d'itérations dans le cas quadratique.

Une autre caractéristique importante est que la matrice H_i converge vers l'inverse de l'hessienne de la fonction au point de minimum.

Supposons que nous ayons une fonction quadratique simple :

$$f(x) = \frac{1}{2} x^\top A x + b^\top x$$

Si la matrice A est définie positive et symétrique, la méthode de Fletcher et Powell, avec la mise à jour itérative de la matrice H_i, convergera vers le minimum global en un nombre d'itérations égal à la dimension de l'espace.

Annexe C

Coefficient de Gini et courbe de Lorenz

Le coefficient de Gini et la courbe de Lorenz sont des outils statistiques utilisés pour mesurer l'inégalité dans une distribution. Ils sont largement utilisés en économie pour analyser la répartition des revenus ou des richesses, mais peuvent également être appliqués à d'autres domaines, comme l'assurance, pour évaluer la répartition des sinistres ou des primes.

C.1 Principe de la courbe de Lorenz

1. **Construction de la courbe :** La courbe de Lorenz est construite en représentant graphiquement la proportion cumulée de la variable d'intérêt (par exemple, le revenu ou les sinistres) en fonction de la proportion cumulée de la population. Sur l'axe horizontal, on place la part cumulative de la population, classée par ordre croissant de richesse ou de sinistres, et sur l'axe vertical, on place la part cumulative de la richesse ou des sinistres correspondante.

2. **Courbe d'égalité parfaite :** La diagonale de la courbe de Lorenz représente l'égalité parfaite, c'est-à-dire une situation dans laquelle chaque membre de la population reçoit exactement la même part de la variable d'intérêt. Plus la courbe de Lorenz s'éloigne de cette diagonale, plus l'inégalité est grande.

C.2 Coefficient de Gini

1. **Définition :** Le coefficient de Gini est un indicateur synthétique d'inégalité basé sur la courbe de Lorenz. Il est défini comme le rapport de la surface entre la courbe de Lorenz et la diagonale, par rapport à la surface totale sous la diagonale. Formellement, si A représente la surface entre la courbe de Lorenz et la diagonale, et B la surface sous

la courbe de Lorenz, alors :

$$G = \frac{A}{A + B}$$

Le coefficient de Gini varie entre 0 (égalité parfaite) et 1 (inégalité maximale).

2. **Calcul :** En pratique, le coefficient de Gini peut être calculé en utilisant la formule :

$$G = 1 - 2 \int_0^1 L(x)dx$$

où $L(x)$ est la courbe de Lorenz. Cette formule reflète la différence entre la distribution réelle et la distribution parfaitement égalitaire.

C.3 Application en assurance et interprétation

Le coefficient de Gini et la courbe de Lorenz sont des outils puissants pour évaluer la segmentation du portefeuille d'assurés en fonction des sinistres. Un coefficient de Gini élevé indiquerait que certains segments du portefeuille concentrent une part disproportionnée des sinistres. Cette situation révélerait un fort pouvoir discriminant des variables utilisées pour la segmentation du risque, suggérant une différenciation marquée entre les assurés. Cela pourrait également inciter à ajuster les primes pour refléter plus précisément le risque associé à chaque segment. À l'inverse, un coefficient de Gini faible traduirait une segmentation moins discriminante, avec une répartition plus homogène des sinistres entre les assurés. Dans ce cas, la capacité des variables à distinguer les différents niveaux de risque pourrait être mise en question, nécessitant potentiellement une révision des critères de segmentation pour mieux capturer les différences de risque au sein du portefeuille.

Table des figures

Liste des tableaux

Bibliographie

[1] A. Paglia et M. V. Phélippé-Guinvarc'h. *Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique.* Bulletin français d'actuariat, volume 12, juillet-décembre 2011, 24p.

[2] ACPR. *Les assureurs français face au risque de changement climatique.* n°102-2019 Analyses et synthèses, 2019.

[3] Adam Boughrara. *Les tendances du marché de l'assurance en 2023*, 16 avril 2024.

[4] Alexandre Laverdure. *Risques climatiques : comment l'assurance se réinvente face à cette tempête ?*. Precisely, janvier 2024.

[5] Arthur Charpentier et Michel Denuit. *Mathématiques de l'assurance non-vie - tome 1 & 2.* Economica, 2004.

[6] Autres cartes de la population de France. *Carte de la densité de population par commune en 2009.* `https://www.cartes-de-france.fr/population.html#ixzz8iIwMDvPY`.

[7] B. Antonia. *Augmentation des primes d'assurance habitation.* L'Assurance en mouvement. 5 mai 2024.

[8] *Carte de France.* Quiz géographiques et cartes de France. Disponible à : `https://www.cartes-de-france.fr/`.

[9] Catalina Sepulveda. *Modélisation du risque géographique en Santé, pour la création d'un nouveau Zonier. Comparaison de deux méthodes de lissage spatial.* Mémoire présenté pour l'obtention du diplôme de Statisticien Mention Actuariat et l'admission à l'Institut des Actuaires.

[10] CCR. *The French CatNat system : Post-flood recovery and resilience issues.* CCR Report, 2024.

[11] CCR. *Les catastrophes naturelles en France Bilan 1982-2023.* Rapport de la CCR, 2024.

[12] Chantal Bichler. *Nombre de reconnaissances Cat Nat au titre des inondations (1982 - 2014).* Documents d'urbanisme et risque inondation, 18 septembre 2018.

[13] Cheikh Sene. *Adéquation au profil de risque et besoin global de solvabilité*. Mémoire d'actuariat ISFA, 2021.

[14] Christine Lavarde. *Régime d'indemnisation des catastrophes naturelles*. Rapport d'information n°603 (2023-2024), 15 mai 2024.

[15] Claudine Ferrier. *Le zonier en tarification IARD : approche comparative de deux techniques de construction d'un critère de segmentation géographique en assurance habitation*. Mémoire d'actuariat ISFA, 2016.

[16] Colin Brändén and Kfir Yehuda. *Generalized Principal Component Analysis*. arXiv :1907.02647, 2019.

[17] Conseil économique, social et environnemental. *Climat, cyber, pandémie : le modèle assurantiel français mis au défi des risques systémiques*. Journal Officiel de la République Française Mandature 2021-2026, Séance du 13 avril 2022.

[18] Daniel Joly, Thierry Brossard, Hervé Cardot, Jean Cavailhes, Mohamed Hilal et Pierre Wavresky. *Les types de climats en France, une construction spatiale*, 2010.

[19] Data.gouv. *Base officielle des codes postaux*. `https://www.data.gouv.fr/fr/datasets/base-officielle-des-codes-postaux/`.

[20] Delphine Bardou. *Que couvre l'assurance habitation multirisque ?*, 24 mars 2024.

[21] Eglantine Tamet. *Que couvre l'assurance habitation multirisque (MRH) ?*, 11 avril 2024.

[22] EITCA. *Comment l'algorithme de descente de gradient met-il à jour les paramètres du modèle pour minimiser la fonction objectif, et quel rôle le taux d'apprentissage joue-t-il dans ce processus ?*. 22 mai 2024.

[23] Equipe Blent. *Algorithme k-means : comment ça marche ?*, janvier 2022. `https://blent.ai/blog/a/k-means-comment-ca-marche`.

[24] Expert du Droit. *Le guide complet de l'assurance habitation*, MonExpertduDroit, 2024.

[25] Exposition au RGA 2019. *indicateurs-rga 2019-communes-departements.xlsx*. BRGM, 2019 ; Fideli, 2017. Traitements : SDES, 2021.

[26] François Bafoil. *Les financements (1/2) : Les CATNAT*. 23 mai 2022. `https://www.caissedesdepots.fr/blog/article/les-financements-12-les-catnat`

[27] Gabrielle Ross and Liam Walters. *Analysis of the Lambert Projection System for Geospatial Data*. arXiv :1903.10573, 2019.

[28] Groupe CDC Habitat. *Le plan d'adaptation au changement climatique du groupe CDC Habitat*, décembre 2022.

[29] Guillaume Beraud-Sudreau. *Construction d'un zonier en MRH à l'aide d'outils de data-science*. Mémoire présenté pour l'obtention du Master professionnel Sciences de gestion, mention finances de marché Spécialité Actuariat du CNAM et l'admission à l'Institut des Actuaires.

[30] Hellowatt. *Production annuelle en kWh/kWc par région en France*. Panneaux solaires photovoltaïques. `https://www.hellowatt.fr/panneaux-solaires-photovoltaiques/`.

[31] Institut géographique national. *Projection cartographique conique conforme de Lambert*. Notes techniques NT/G 71, 1ère édition, 1995.

[32] James, M. and Barry, L. *Insurance against Natural Catastrophes : Balancing Actuarial Fairness and Social Solidarity*. The Geneva Papers on Risk and Insurance, 2022.

[33] James Thomson and Kai Young. *Principal Component Analysis and its Applications to Image Compression*. arXiv :1609.08247, 2016.

[34] Jennifer Pariente. *Modélisation du risque géographique en assurance habitation*. Mémoire d'actuariat Université Paris Dauphine, 2017.

[35] Jérémy Rouot et Pierre Ailliot. *Analyse de données*. Cours EURIA, 2021.

[36] Julien. *Qu'est-ce que le retrait-gonflement des argiles ?*. Blog dangers-habitat, 2023. `https://blog.dangers-habitat.fr/retrait-gonflement-des-argiles/`.

[37] Kaiming He, Georgia Gkioxari, Piotr Dollar and Ross Girshick. *Mask R-CNN*. arXiv 1703.06870v3, 2018.

[38] La rédaction Meilleurtaux. *Hausse des primes d'assurance habitation à prévoir face aux catastrophes naturelles*, 10 avril 2024.

[39] Le Comparateur Assurance. *Les perspectives pour l'assurance en 2024*, 25 octobre 2023.

[40] Les sols métropolitains. *Répartition des grands types de sols en France métropolitaine*. `https://www.donnees.statistiques.developpement-durable.gouv.fr/lesessentiels/essentiels/sol-diversite-metropolitain.htm`.

[41] Li Wei and Zhang Ming. *Research and Practice of Image Processing Based on Python*. arXiv :1505.07263, 2015.

[42] Louis Zhang and Raphael Fortunato. *PDFO : A Cross-Platform Package for Powell's Derivative-Free Optimization Solvers*. arXiv :2302.13246, 2023.

[43] Manal Bourachid. *Création d'un nouveau zonier de la garantie Tempête, Grêle et Neige*. Mémoire d'actuariat ISUP, 2017.

[44] Michel Fromenteau, Vincent Ruol et Laurence Eslous. *Sélection des risques : où en est-on ?*, Les Tribunes de la Santé, 25 juillet 2011, volume 31, pages 63 - 71.

[45] Mohamed Naouai. *Filtrage d'image (Cours 7)*. Faculté des sciences de Tunis, Université de Tunis El Manar.

[46] MonCoachData. *Machine Learning : l'algorithme des k plus proches voisins*, décembre 2023. `https://moncoachdata.com/blog/algorithme-des-k-plus-proches-voisins/`.

[47] Mulah Moriah. *Mesure et mitigation des biais : vers une tarification non-vie réellement équitable*. Mémoire d'actuariat EURIA, 2022.

[48] Nicolas Langevin. *Modélisation de la sinistralité tempête, apport de l'Open Data et du Machine Learning*. Mémoire d'actuariat ENSAE ParisTech, 2019.

[49] Nussbaum, R. *The Role of Public Guarantees in France's CatNat Insurance*. Springer, 2015.

[50] Patrick Soulignac. *En 2023, le secteur de l'assurance dommages devra poursuivre sa dynamique d'innovation face à l'instabilité du monde*, 22 décembre 2022. `https://www.wesur.fr/guides/les-tendances-du-marche-de-lassurance-en-2023`.

[51] Pierre Ailliot. *Modèles linéaires*. Cours EURIA, 2021.

[52] Pulkit Sharma. *Computer Vision Tutorial : Implementing Mask R-CNN for Image Segmentation*, 2024.

[53] Revital Assurances. *Les augmentations des assurances habitation en 2024 : Ce que vous devez savoir*, 30 décembre 2023.

[54] *Risque retrait-gonflement des argiles*. Les services de l'État dans le Bas-Rhin, 2021. `https://www.bas-rhin.gouv.fr/Actions-de-l-Etat/Prevention-des-risques-naturels-et-technologiques/Presentation-des-differents-risques/Risque-retrait-gonflement-des-argiles/Risque-retrait-gonflement-des-argiles`.

[55] Thierry Langreney, Gonéri Le Cozannet et Myriam Merad. *Adapter le système assurantiel français face à l'évolution des risques climatiques*, décembre 2023.