



Terre, Vent, Feu, Eau, Data

Surveillance des risques de feux de forêt en France



*"Les forêts précèdent les peuples, les déserts les suivent."
— François-René de Chateaubriand*

Introduction du sujet

L'Été de Tous les Dangers

L'été 2025 restera gravé dans les mémoires comme l'un des plus destructeurs pour les forêts françaises. Le 7 août 2025, l'Aude a été frappée par le plus grand incendie depuis 1949, ravageant 17 000 hectares en moins de 48 heures. Cette "catastrophe d'ampleur inédite" a mobilisé 2 100 pompiers et tous les moyens aériens nationaux, causant un mort et 13 blessés.



Source (2024) : <https://bdiff.agriculture.gouv.fr/incendies>

Mais l'Aude n'est pas un cas isolé. Juillet 2025 a été marqué par une série d'incendies exceptionnels sur tout le territoire méditerranéen, confirmant les prédictions scientifiques : la France fait face à une multiplication par trois de la période à risque d'ici 2050.

L'Intelligence des Données au Service de la Prévention

Face à cette urgence climatique, l'État français dispose depuis 2006 d'un trésor numérique : la [BDIFF](#) (Base de Données sur les Incendies de Forêts en France). Cette ressource exceptionnelle, hébergée par l'IGN, centralise chaque année les données de plus de 9000 feux avec une précision géographique au niveau communal.



Mais les données brutes ne suffisent plus. Dans un contexte où 90% des départs de feu sont d'origine humaine et où les enjeux de prévention deviennent critiques, il faut développer des outils d'intelligence prédictive capables d'anticiper les risques sur chaque commune de France.

Défis Technologiques et Opportunités

La récupération et l'usage de la BDIFF présente des défis techniques pour qui veut exploiter les données de 1973 à 2024:

- Fragmentation des données : téléchargement des fichiers CSV par paquets de 30 000 lignes
- Géocodage par code INSEE (sans précision de la latitude/longitude, qu'il faudra récupérer dans un autre dataset en open data)
- Possibilité d'analyse multi-échelles : le dataset ainsi reconstitué vous permettra une analyse spatiale, temporelle, saisonnière pour prédire le niveau de risque d'incendie.

Le présent projet n'est qu'une étape. Le croisement des données BDIFF avec celles de la météorologie (sécheresse, vent, température...) permettront de mieux qualifier le risque dans une version ultérieure du projet.



Mise en situation professionnelle

Mission Professionnelle, contexte de l'Équipe Projet

Vous êtes une équipe de 2 à 3 data scientists spécialisés en intelligence géospatiale, recrutés par la Direction Générale de la Sécurité Civile et de la Gestion des Crises dans le cadre du plan national d'adaptation au changement climatique.

A votre retour de congés, le CDO (chief data officer) de votre agence vous demande de développer rapidement un prototype d'application de suivi des incendies et de prédiction des niveaux de risques sur le territoire.

Ce prototype sera présenté à la direction de votre agence dans une semaine!

Livrable Principal

Vous devez développer un système d'intelligence prédictive complet combinant :

- **Base de données SQL consolidée** : BDIFF + référentiels géographiques (il faut consolider les codes des communes avec des coordonnées de latitude et longitude)
- **Application Streamlit interactive** : prototype de visualisation et prédiction des risques
- **Modèle de Machine Learning** : prédiction du risque par commune et période

Délai total : 5 jours ouvrés



Spécifications du prototype

Infrastructure de Données (Jour 1-2)

- Consolidation SQL : intégration BDIFF + données géographiques INSEE
- Téléchargement automatisé par départements (BDIFF limite 30k lignes)
- Géocodage : jointure codes INSEE → coordonnées lat/Ing via <https://adresse.data.gouv.fr/outils/telechargements>
- Qualité des données : validation, déduplication, gestion des valeurs manquantes

Application / Onglet 1 : Cartographie & Analyse Historique

- Cartographie interactive : heatmap des incendies avec points géolocalisés
- Filtres dynamiques :
 - Sélecteur temporel (année/mois/saison)
 - Gravité (surface brûlée)
 - Type de végétation (forêt/maquis/autres)
 - Origine (naturelle/accidentelle/malveillante)
- Statistiques descriptives :
 - Évolution temporelle des surfaces brûlées
 - Répartition des causes par région/saison
 - Hotspots géographiques et saisonnalité

Application / Onglet 2 : Prédiction des Risques

- Modèle de clustering géo-temporel : prédiction du risque par commune
- Variables prédictives :
 - Historique des incendies de la commune et communes voisines



- Jour et mois dans l'année (saisonnalité)
- Surface de végétation par type (selon définitions BDIFF)
- Interface prédictive : sélection commune + période → score de risque

Votre feuille de route indicative (5 Jours)

Jour 1 : Infrastructure et Ingestion

Matin – Veille technologique :

1. Comprendre la structure de la BDIFF et ses évolutions depuis 2006
2. Analyser les définitions officielles
3. Explorer l'API et les formats de téléchargement BDIFF
4. Étudier les référentiels géographiques INSEE disponibles
5. Comprendre les enjeux de consolidation multi-sources

Après-midi – Architecture de données :

6. Conception du schéma SQL consolidé
7. Développement du pipeline d'ingestion BDIFF (limitation 30k lignes par fichier CSV)
8. Téléchargement des données géographiques par département
9. Premiers tests de géocodage et jointures

Jour 2 : Consolidation et EDA

Matin – Intégration :

- Finalisation de la base SQL consolidée
- Scripts de mise à jour automatisée
- Tests d'intégrité et de cohérence

Après-midi – Exploration :

- Analyse exploratoire des données (EDA) complète



- Identification des patterns spatio-temporels
- Documentation des biais et limitations

Jour 3 : Développement de l'application

Matin – Architecture Streamlit :

- Structure de l'application multi-onglets
- Connexion base de données
- Interface cartographique de base

Après-midi – Onglet Historique :

- Implémentation des filtres dynamiques
- Visualisations statistiques
- Tests d'interactivité

Jour 4 : Machine Learning

Matin – Modélisation :

- Préparation des features géo-temporelles
- Implémentation clustering spatial (communes voisines)
- Clustering temporel (saisonnalité jour/mois)

Après-midi – Prédiction :

- Entraînement du modèle de risque
- Validation croisée géographique
- Interface prédictive dans Streamlit

Jour 5 : Finalisation et Tests

Matin – Intégration finale :

- Tests end-to-end de l'application
- Optimisation des performances
- Documentation utilisateur



Après-midi – Préparation soutenance :

- Rédaction du README technique
- Préparation des slides de présentation
- Tests de déploiement local

Méthodologie de Prédiction des Risques

Approche par Clustering Géo-Temporel

Le modèle de prédiction, pour notre prototype, et en l'absence de données météorologiques, repose sur l'hypothèse que le risque d'incendie dans une commune dépend :

1. De son historique propre : fréquence et intensité des incendies passés
2. De son voisinage spatial : risque des communes limitrophes (contagion géographique)
3. De la saisonnalité : jour et mois dans l'année (patterns climatiques récurrents)

Variables Prédictives Recommandées

Variables spatiales :

- Nombre d'incendies historiques dans la commune (5/10/20 dernières années)
- Surface totale brûlée historique par type de végétation
- Densité d'incendies dans un rayon de 10/20/50 km
- Moyenne pondérée par distance des incendies voisins

Variables temporelles :

- Jour julien (1-365) pour capturer la saisonnalité fine
- Mois de l'année (1-12)
- Saison météorologique (été=juin-août prioritaire)



- Variables cycliques : $\sin/\cos(2\pi \times \text{jour}/365)$, $\sin/\cos(2\pi \times \text{mois}/12)$
- Weekends, jours fériés, périodes de vacances...

Variables de contexte (optionnel):

- Surface communale par type de végétation (forêt/maquis/autres)
- Distance à la côte méditerranéenne (effet climatique)
- Altitude moyenne de la commune (si disponible)

Pipeline de Modélisation

1. Agrégation spatiale : calcul des features par commune et période
2. Clustering spatial : regroupement des communes par profil de risque similaire
3. Clustering temporel : identification des périodes à risque homogène
4. Modèle de prédiction : régression ou classification sur les clusters identifiés
5. Validation géographique : test sur départements non vus à l'entraînement, ou/et sur une période non vue à l'entraînement.

Données et Sources

- BDIFF : <https://bdiff.agriculture.gouv.fr/incendies>
 - Format : CSV par paquets de 30 000 lignes maximum
 - Période : 1973–2024 (données 2025 en cours de saisie)
 - Granularité : Niveau communal avec code INSEE
- Référentiel géographique :
<https://adresse.data.gouv.fr/outils/telechargements>
 - Coordonnées lat/Ing par commune (code INSEE)
 - Téléchargement par département



- Format : CSV avec colonnes standardisées

Compétences visées

Alignement avec les compétences du Bachelor Data-IA (3 Années)

Prérequis Techniques

Python, SQL, Git, algorithmique de base, jointures SQL, versioning.

Pandas, Flask, analyse exploratoire, ML supervisé, EDA avancée, interface web, clustering.

Géopandas, MLOps, projets complexes, Power BI, Cartographie, déploiement, dashboard.

Correspondance RNCP38616

BC01 - Conception et développement Architecture de données consolidées, pipeline robuste

BC02 - Analyse de données, EDA multi-échelle, insights géo-temporels

BC03 - Machine Learning, Modèle prédictif clustering, validation géographique

BC06 - Interface décisionnelle, Application Streamlit opérationnelle

Rendu

NB: Ce projet de consolidation est ambitieux dans beaucoup de dimensions liées à la science des données. Chaque groupe d'étudiants est appelé à fournir un travail avec les compétences acquises au niveau Bachelor (ou équivalent). Le succès du projet ne suppose pas la fourniture exhaustive de tous les livrables, mais sera plus largement dépendant des démarches effectuées et de la communication sur ces démarches.

Livrables Techniques



1. Base de données SQL consolidée et fonctionnelle en local

- Script de création des tables
- Pipeline d'ingestion automatisé
- Documentation du schéma de données

2. Application Streamlit complète avec :

- Onglet cartographie historique (filtres + statistiques)
- Onglet prédiction des risques (modèle ML intégré)
- Interface utilisateur intuitive et responsive

3. Repository GitHub terre-vent-feu-eau-data contenant :

- Code source complet et reproductible
- Notebook EDA détaillé
- Scripts SQL et pipeline ETL
- README technique avec guide d'installation
- Requirements.txt et instructions de déploiement

4. Documentation d'analyse :

- Rapport méthodologique (markdown)
- Insights sur les patterns d'incendies
- Limitations et recommandations d'amélioration

Soutenance Professionnelle (20 min)

1. Démonstration technique (10 min) :

- Présentation de l'application en fonctionnement
- Navigation dans les deux onglets
- Test de prédiction sur cas d'usage concrets

2. Présentation méthodologique (8 min) :

- Architecture de données consolidées



- Approche de modélisation ML
- Principaux insights découverts

3. Questions/Réponses (2 min) : discussion technique et perspective d'amélioration

Base de connaissances

Documentation Officielle

- BDIFF : <https://bdiff.agriculture.gouv.fr/aide/generalites>
- GeoPandas : <https://geopandas.org/en/stable/gallery/index.html>
- Streamlit : <https://docs.streamlit.io/>
- Folium cartographie : <https://python-visualization.github.io/folium/latest/>

Références Techniques (tous les liens ne sont pas nécessaires à la réalisation du projet)

- Clustering géospatial :
<https://scikit-learn.org/stable/modules/clustering.html#clustering>
- Étude du clustering spatial DBSCAN pour les feux de forêt : [Wildfire Risk Map Based on DBSCAN Clustering](#)
- Étude de cas DBSCAN + UMAP pour feux : [A Study on Enhanced Spatial Clustering Using Ensemble Dbscan and Umap to Map Fire Zone in Greater Jakarta, Indonesia](#)
- Approche hybride spatio-temporelle : [Wildfire Prediction Model Based on Spatial and Temporal Characteristics: A Case Study of a Wildfire in Portugal's Montesinho Natural Park](#)



- Machine Learning géographique :

<https://geographicdata.science/book/intro.html>

- PostGIS spatial : <https://postgis.net/workshops/postgis-intro/>