



Master 2 Statistiques et Sciences des Données

MÉMOIRE DE FIN D'ÉTUDES

Construction de nouvelles unités agroclimatiques dans l'Hérault à partir des évolutions climatiques observées ces dernières décennies

Auteur :
Pierre DIAS

En partenariat avec :

Université de Montpellier – Conseil Départemental de l'Hérault

Encadrante académique : Benoîte de Saporta
Encadrant professionnel : Christophe Benoît

Jury : Elodie Brunel
Maximilien Dossa
Benoîte de Saporta



Année Universitaire 2024 – 2025

Remerciements et Résumé

Remerciements

Je tiens à exprimer ma profonde gratitude à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce travail. Tout d'abord, je remercie très sincèrement l'**Hôtel du Département de Montpellier** pour m'avoir offert l'opportunité d'intégrer leurs équipes et de participer activement à leurs projets. Mon expérience au sein de l'équipe **SANOE** a été à la fois formatrice et stimulante. Elle m'a permis d'approfondir mes compétences en *Systèmes d'Information Géographique* (SIG), tout en m'offrant une meilleure compréhension des enjeux environnementaux et territoriaux propres au département de l'Hérault. J'adresse mes remerciements à mon maître d'apprentissage, **Christophe Benoît**, ainsi qu'à **Christophe Gasc**, pour leur soutien, leur disponibilité et leur encadrement bienveillant tout au long de cette période.

Je tiens également à exprimer toute ma reconnaissance à mon encadrante académique, **Benoîte de Saporta**, pour son suivi rigoureux, la pertinence de ses conseils et sa constante bienveillance.

Je remercie chaleureusement mes collègues, et en particulier **Léa Vigneau**, pour leur accueil, leur esprit collaboratif et la qualité de nos échanges. Leur expertise et leur soutien quotidien ont largement contribué à l'achèvement de ce projet.

Enfin, je remercie mes amies **Anna Girones**, **Aïcha Lahjiouj** et **Jeanne Vivier**, dont les relectures attentives et les conseils avisés ont considérablement amélioré la clarté et la qualité de ce travail.

Résumé

Cette étude propose une méthodologie actualisée pour la redéfinition des unités agroclimatiques du département de l'Hérault et de son tampon de 15 km, basée sur l'utilisation des données climatiques et météorologiques issues des stations situées dans cette zone. Plusieurs méthodes de classification non supervisée, intégrant la dimension spatiale, sont comparées afin d'optimiser l'homogénéité interne et la pertinence agroclimatique des groupes. Des techniques d'interpolation spatiale, notamment le krigage et les forêts aléatoires, permettent d'étendre les observations ponctuelles à l'ensemble du territoire étudié. Les résultats fournissent un zonage agroclimatique adapté aux besoins actuels et transposable à d'autres contextes.

Abstract

This study presents a method to redefine agroclimatic units for the Hérault department and its 15 km surrounding buffer, using meteorological station data collected throughout the area. Combining multivariate unsupervised classification with spatial constraints, the method produces agroclimatic zones that respect environmental gradients and spatial continuity. Spatial interpolation techniques, such as kriging and random forests, are used to create continuous maps from point data. The resulting classification offers an updated and reliable agroclimatic zoning framework adapted to regional agricultural and environmental needs, and potentially transferable to other regions.

Table des matières

1	Introduction	1
1.1	L'Hôtel du Département de l'Hérault et le cadre de l'alternance	1
1.2	Objectif de redéfinition des unités agroclimatiques	2
1.3	Présentation des stations et du territoire d'étude	4
1.4	Méthodologie générale et plan du rapport	5
2	Préparation des données	7
2.1	Traitements des variables dynamiques	7
2.1.1	État initial de la base de données	7
2.1.2	Description des variables utilisées	9
2.2	Extraction des variables géomorphologiques	10
2.2.1	Calcul de la pente et de l'exposition	11
2.2.2	Extraction de la latitude des fonds de vallée	12
3	Classification des stations	15
3.1	Classification non supervisée par k-means	15
3.1.1	Classification ascendante hiérarchique (CAH)	15
3.1.2	Stabilisation par <i>k-means</i>	16
3.2	Classification spatiale via la suppression des arêtes d'un arbre	17
3.2.1	Principe de construction du graphe	17
3.2.2	Résultats de la classification	19
3.3	Classification spatiale basée sur des distances combinées	21
3.3.1	Matrices de distances et choix des paramètres	21
3.3.2	Résultat de la classification	24
3.4	Discussion sur les méthodes	24
4	Interpolations	27
4.1	Interpolation par IDW (Inverse Distance Weighting)	28
4.2	Interpolation par krigeage	29
4.2.1	Krigeage ordinaire	29
4.2.2	Krigeage universel	30
4.3	Modélisation par forêts aléatoires	33
4.3.1	Prédiction d'une variable catégorielle	34
4.3.2	Prédiction par classe	36
4.4	Comparaison des modèles par validation croisée	39
4.5	Lissage de la grille d'interpolation	40
4.5.1	Transformation d'une grille en une couche de polygones	40
4.5.2	Rendu final	40
5	Conclusion	43
6	Annexe	47
I	Complements sur la classification	47
I.1	Étude complémentaire sur les anciennes périodes avec ClustGeo	47

I.2	Distribution des groupes obtenus par ClustGeo pour la période de 2010 - 2020	50
II	Compléments sur l'interpolation	52
II.1	Observations complémentaires sur le krigage	52
II.2	Observations complémentaires sur les forêts aléatoires	54
III	Protocole appliqué aux données SAFRAN	55
IV	Outil de visualisation	58
	Bibliographie	59

Chapitre 1

Introduction

1.1 L'Hôtel du Département de l'Hérault et le cadre de l'alternance

Mon alternance s'est déroulée au sein du Service Agroclimatologie et Numérique pour l'Observation de l'Environnement (SANOE), intégré dans le Pôle de l'Eau, de l'Environnement et de l'Économie du Département de l'Hérault. Ce service est une composante essentielle du SOCEEL (Service d'Observation Climatologie Eau Environnement Littoral), structure dédiée à la collecte, la gestion et l'analyse des données environnementales du territoire.

Le SOCEEL a pour mission principale le pilotage et l'animation de l'Observatoire Départemental Climatologie Eau Environnement Littoral (ODCEEL), en collaboration étroite avec les services de la Direction Générale Adjointe (DGA) tels que la DETIE (Territoires et Innovation Écologiques), la DGA AT (Agriculture et Territoires), ainsi qu'avec des partenaires externes institutionnels et scientifiques. La transversalité de ces actions garantit une coordination efficace des politiques environnementales et agricoles à l'échelle départementale.

Le service est également en lien avec la gestion du Système d'Information Géographique (SIG) dédié à l'eau et à l'environnement, outil indispensable pour visualiser, analyser et exploiter les données territoriales dans une optique d'aide à la décision. Par ailleurs, le SOCEEL joue un rôle clé dans l'Assistance à la Maîtrise d'Ouvrage des Systèmes d'Information (AMOA SI) pour plusieurs pôles métiers du Département, notamment le Pôle Économie Eau Environnement (P3E), la Direction Administration Financière et Fonds Européens (DAFFE), ainsi que la Mission Développement Durable et Prospective (MDDP). L'AMOA a pour objectif de formaliser et de clarifier les besoins numériques des utilisateurs afin de développer des solutions informatiques adaptées, performantes et en adéquation avec les exigences métier.

Le périmètre fonctionnel de l'AMOA SI s'articule autour de plusieurs phases clés : la description précise des besoins, la préparation des projets, le support au déploiement des solutions, la formation des utilisateurs, ainsi que le suivi de la validation. Cette approche méthodique garantit la bonne conduite des projets numériques et leur adéquation avec les usages de terrain.

L'ODCEEL s'appuie sur une base de données très riche et diversifiée. En climatologie, il rassemble près de 67 millions de relevés, dont 44 millions issus du réseau CD34, complétés par 21,4 millions provenant de partenaires comme Météo-France ou INRAE, et environ 1,4 million de données additionnelles en cours d'identification. La base inclut également 27,8 millions de mesures piézométriques, 1,46 million de relevés concernant la qualité des eaux souterraines, ainsi qu'1,25 million de données hydrométriques, un secteur en forte croissance. S'y ajoutent des séries spécifiques, telles que les 87 000 relevés hydrométriques des barrages du Salagou et des Olivettes, 35 000 mesures sur la qualité des eaux de surface, et plus d'1,3 million de données topographiques issues de profils bathymétriques.

Le portail web de l'ODCEEL, accessible via odee.herault.fr [9], fédère une communauté d'utilisateurs variée : syndicats de bassin, chambres d'agriculture, communes, intercommunalités. Le site propose une centaine de cartes thématiques en ligne couvrant une dizaine de grandes thématiques, ce qui en fait une ressource précieuse pour la connaissance et la gestion durable du territoire.

Dans ce contexte riche et dynamique, mon alternance m'a permis de contribuer principalement au traitement, à l'analyse et donc à la valorisation des données climatiques et environnementales, ainsi qu'à la modernisation des outils numériques destinés à soutenir les politiques départementales.

1.2 Objectif de redéfinition des unités agroclimatiques

Les unités agroclimatiques (UA) désignent des zones géographiques cohérentes sur les plans climatique, géo-morphologique et agricole. Leur première définition dans le département de l'Hérault remonte à 1997 [14], dans un contexte où les connaissances disponibles sur le milieu naturel – issues de disciplines telles que l'agroclimatologie, la pédologie, la géographie ou encore la phytosociologie – ont permis d'identifier neuf « petites régions naturelles » au sein du territoire viticole héraultais.

Cette première délimitation s'appuyait sur une méthode empirique, non formalisée de manière rigoureuse, mobilisant à la fois des données scientifiques disponibles à l'époque et une expertise de terrain locale. Les principaux critères utilisés peuvent être synthétisés dans le tableau suivant :

Type de critère	Sources ou méthodes mobilisées	Rôle dans la délimitation des unités
Agroclimatique	Étages d'humidité, variantes thermiques	Définition de grandes tendances bioclimatiques locales
Phytosociologique	Études du Centre Emberger (CNRS) sur les groupements végétaux spontanés	Identification d'associations végétales indicatrices de conditions écologiques spécifiques
Géographique	Classes d'altitude, pente, exposition	Caractérisation des reliefs (plaine, coteaux, zones de montagne)
Expérience de terrain	Observations et expertise de l'ACH (Association Climatique de l'Hérault)	Ajustements empiriques des limites de chaque unité en fonction des réalités locales

TABLE 1.1 – Principaux critères de définition des unités agroclimatiques de 1997

Les unités ainsi définies présentaient une homogénéité interne, à la fois sur le plan bioclimatique, paysager et territorial. Elles servaient également de base pour la caractérisation des terroirs viticoles, en lien avec les séries de sols dominantes et les délimitations AOC/AOP, et ont été utilisées dans plusieurs productions cartographiques techniques (ex. Infoclim, Aléachim [8]).

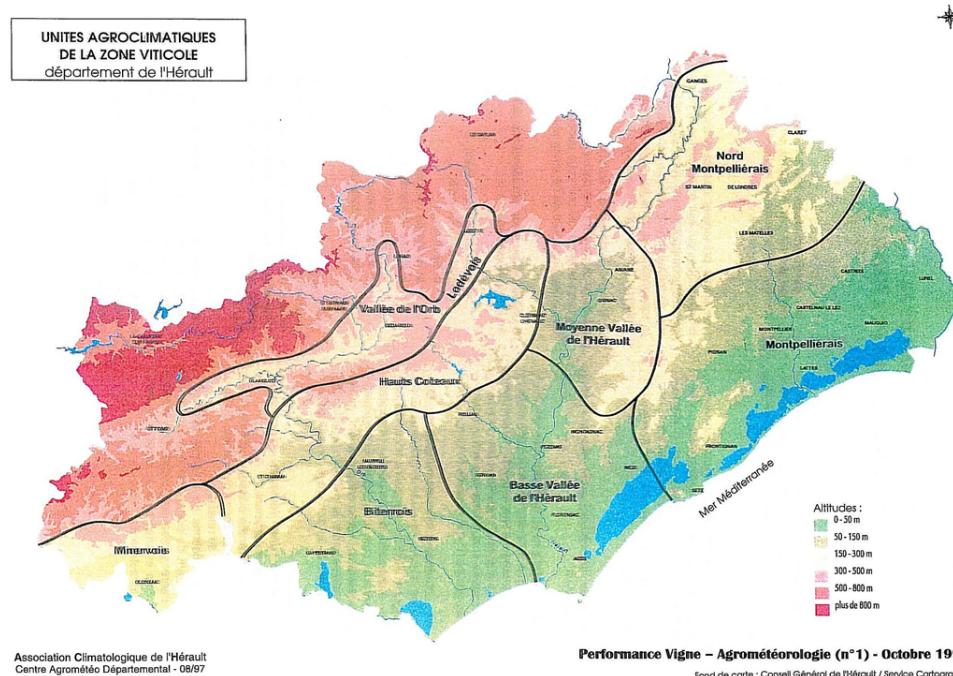


FIGURE 1.1 – Carte d'archive du tracé des unités agroclimatiques du département de l'Hérault (1997¹)

1. Étude effectuée en 1997 et adoptée en 1998.

La carte présentée (figure 1.1) illustre la répartition des neuf unités agroclimatiques historiques, s'étendant d'ouest en est à travers le département de l'Hérault. Ces unités correspondent à des zones naturelles cohérentes, identifiées selon une logique principalement qualitative mêlant critères climatiques, géographiques et agricoles.

Les neuf unités délimitées sont les suivantes : le Minervois Héraultais, les Hauts Coteaux, le Biterrois, la Vallée de l'Orb et le Lodévois, la Basse Vallée de l'Hérault, la Moyenne Vallée de l'Hérault, le MontPELLIÉRAIS, ainsi que le Nord MontPELLIÉRAIS.

Cependant, plusieurs limites rendent aujourd'hui ce zonage obsolète. D'abord, la méthode d'origine n'a jamais été formellement documentée et repose sur des critères difficilement quantifiables ou réplicables. Ensuite, les données initiales sont désormais anciennes et ne prennent pas en compte l'évolution rapide du climat régional, marquée par un réchauffement global, une variabilité accrue des précipitations et une fréquence plus élevée d'épisodes climatiques extrêmes. Enfin, certains éléments clés de l'approche originelle, tels que les observations de terrain réalisées par la chambre de l'agriculture ainsi que les données sociologiques associées aux unités territoriales, ne sont plus accessibles.

Par ailleurs, si la cohésion sociale pouvait constituer un facteur intéressant pour la structuration territoriale, elle n'a ici qu'une importance secondaire, car l'objectif principal est de définir des unités agroclimatiques homogènes du point de vue des conditions environnementales et non d'étudier les dynamiques sociales ou relationnelles entre communautés.

Parallèlement, les évolutions récentes en matière d'outils d'analyse spatiale, de traitement statistique multivarié et de modélisation territoriale offrent désormais de nouvelles perspectives pour redéfinir les unités agroclimatiques. Ces avancées permettent non seulement d'objectiver la démarche, mais aussi de garantir sa reproductibilité, en s'appuyant sur des jeux de données structurés, standardisés et régulièrement mis à jour. Dans ce contexte, il devient pertinent de proposer une refonte complète de la méthode de délimitation des UA, reposant sur une exploitation rigoureuse des données climatiques, topographiques et environnementales disponibles.

L'objectif de ce travail est donc de concevoir une méthodologie renouvelée de classification agroclimatique à l'échelle du département de l'Hérault. Cette nouvelle approche a pour objectif de représenter de manière fidèle la diversité actuelle des conditions agroclimatiques observées sur le territoire, en tenant compte des dynamiques climatiques récentes et des spécificités géo-morphologiques locales. Contrairement à l'ancienne méthode, elle s'appuiera sur des données quantitatives actualisées et des outils d'analyse robustes, permettant une délimitation plus précise et scientifiquement fondée.

De plus, le périmètre d'étude a été étendu au-delà des seules limites du département : une zone tampon de 15 kilomètres autour de l'Hérault a été intégrée à l'analyse. Cette extension permet de tirer parti de l'amélioration de la couverture des données météorologiques et topographiques disponibles, tout en assurant une meilleure continuité spatiale dans les zones en bordure du département.

L'enjeu est double. Il s'agit, d'une part, de produire un découpage territorial pertinent et représentatif des réalités agroclimatiques contemporaines, en s'affranchissant des biais subjectifs associés à l'approche originelle. D'autre part, il convient de bâtir un cadre méthodologique clair, transparent et suffisamment souple pour être actualisé dans le temps ou transposé à d'autres territoires.

Dans un contexte de dérèglement climatique avéré, il est aujourd'hui légitime de s'interroger sur sa pertinence actuelle. La question de recherche qui guide cette étude est la suivante :

« Comment concevoir une méthode fiable et actualisée pour redéfinir les unités agroclimatiques du département de l'Hérault, en assurant une meilleure homogénéité interne et une représentation fidèle des réalités climatiques et géographiques actuelles ? »

1.3 Présentation des stations et du territoire d'étude

Le département de l'Hérault, situé au cœur du bassin méditerranéen, présente une grande diversité naturelle et géographique. Il s'étend des rivages du littoral languedocien jusqu'aux contreforts méridionaux du Massif Central, en passant par un ensemble varié de plaines, de vallées intérieures et de reliefs. Cette configuration en gradient d'altitude est-ouest et sud-nord, conjuguée à des influences climatiques multiples (méditerranéenne, montagnarde, parfois océanique résiduelle à l'ouest), confère au territoire une mosaïque de conditions agroclimatiques marquées.

Historiquement tourné vers la viticulture, l'Hérault est également un territoire de forte dynamique agricole, où les enjeux de gestion des ressources, d'adaptation au changement climatique et de planification territoriale prennent une dimension croissante. À ce titre, le département constitue non seulement un terrain pertinent, mais surtout un territoire où la redéfinition des unités agroclimatiques s'impose aujourd'hui comme une nécessité, face aux mutations climatiques et aux enjeux croissants d'adaptation du secteur agricole.

Pour garantir une représentation plus cohérente des dynamiques environnementales, le périmètre d'étude ne se limite pas strictement aux frontières administratives du département. Comme annoncé dans la partie précédente, il s'étend à un tampon de 15 km au-delà de ses limites, permettant ainsi d'intégrer des zones périphériques naturellement connectées aux grands ensembles climatiques de l'Hérault, tout en renforçant la densité et la fiabilité du réseau de stations météorologiques exploitées.

Au cours de ce rapport, certaines données et analyses seront présentées en projection Lambert 93, qui correspond au système de coordonnées standard utilisé par le département. Cette projection facilite l'intégration et le traitement des données spatiales provenant des services départementaux.

Le jeu de données climatiques utilisé repose sur un ensemble de stations réparties sur l'ensemble du territoire étudié (Hérault + zone tampon). Ces stations, issues principalement du réseau du département, des réseaux Météo-France ou encore de l'INRAE, couvrent des contextes variés : zones littorales soumises à l'effet marin, plaines intérieures caractérisées par de fortes amplitudes thermiques, zones de piémont plus fraîches et arrosées, etc.

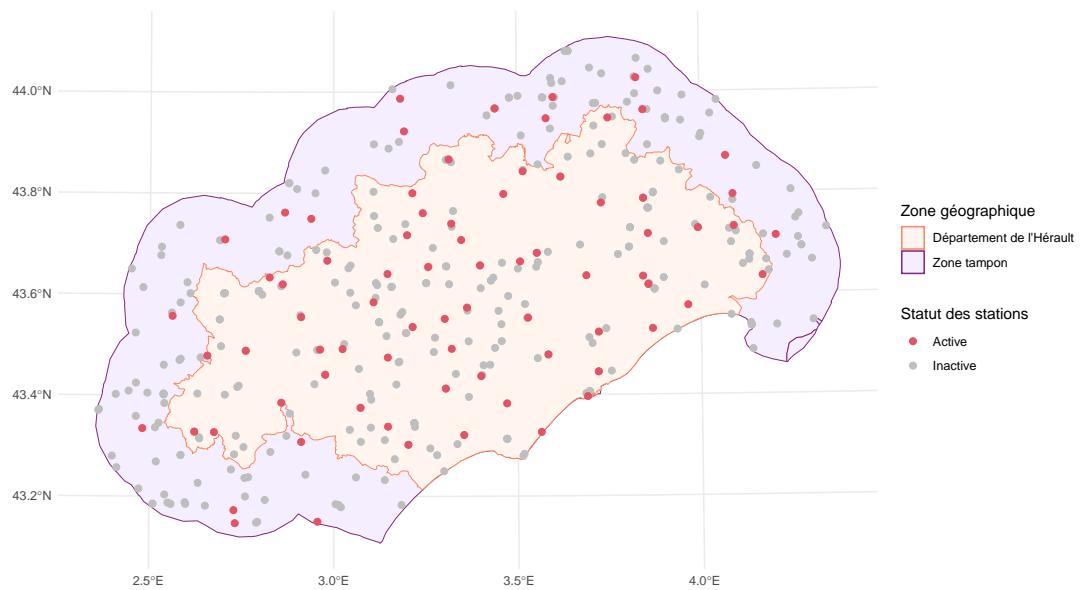


FIGURE 1.2 – Répartition des stations climatiques sur le territoire d'étude

La carte (figure 1.2) présente l'ensemble des stations météorologiques ayant, à un moment donné, enregistré des données sur le territoire étudié (qu'elles soient encore actives ou non aujourd'hui). Les stations toujours en fonctionnement sont mises en évidence afin de souligner leur statut actuel, bien qu'elles ne soient pas nécessairement retenues dans le périmètre d'analyse, notamment lorsque leur période de fonctionnement ne couvre pas suffisamment la période d'étude sélectionnée (2010–2020). De manière

générale, le maillage des stations reflète une implantation relativement homogène sur le territoire, résultant d'un positionnement initial réfléchi. Cela permet de couvrir efficacement la diversité des reliefs présents dans le département et sa zone périphérique.

1.4 Méthodologie générale et plan du rapport

L'objectif principal de cette étude est de proposer une méthode actualisée, rigoureuse et reproductible pour redéfinir les unités agroclimatiques du département de l'Hérault. Pour ce faire, la méthodologie adoptée combine des approches statistiques, spatiales et environnementales, afin de capturer la complexité et la diversité des conditions agroclimatiques locales tout en s'appuyant sur des données robustes et récentes.

Le travail débute par une phase approfondie de collecte et de préparation des données (chapitre 2). Cette étape est cruciale pour garantir la qualité et la pertinence des variables climatiques, météorologiques et géomorphologiques qui alimenteront la classification. Une attention particulière est portée à la gestion des périodes temporelles, à la cohérence spatiale des données et à l'intégration d'un tampon géographique autour du département.

Ensuite, différentes méthodes de classification non supervisée sont explorées (chapitre 3), telles que la classification ascendante hiérarchique, le k-means, et des techniques intégrant la dimension spatiale des données. Ces méthodes sont comparées afin d'identifier celle qui offre le meilleur compromis entre homogénéité interne des groupes et signification agroclimatique.

Par la suite, des méthodes d'interpolation spatiale, notamment le krigeage et les forêts aléatoires, sont mises en œuvre (chapitre 4). Elles permettent d'étendre les observations ponctuelles à l'ensemble du territoire, facilitant ainsi une visualisation fine et continue des variables agroclimatiques. Cette étape est complétée par des modèles prédictifs comme les forêts aléatoires pour améliorer la qualité des estimations.

Enfin, les résultats sont discutés en détail, avant d'être synthétisés dans une proposition de zonage renouvelé des unités agroclimatiques, destiné à répondre aux besoins actuels des acteurs agricoles et environnementaux (chapitre 5).

Cette approche méthodologique articulée garantit un équilibre entre rigueur scientifique, pertinence opérationnelle et adaptabilité à d'autres contextes territoriaux.

Chapitre 2

Préparation des données

Pour notre jeu de données, nous nous reposons sur les stations disponibles afin d'extraire des variables dites **dynamiques**, c'est-à-dire des paramètres climatiques dont les valeurs évoluent selon la période considérée. En complément, nous disposons de **variables géomorphologiques**, supposées fixes dans le temps, permettant d'intégrer une caractérisation topographique et spatiale du territoire.

2.1 Traitement des variables dynamiques

L'identification des variables climatiques les plus pertinentes pour caractériser le territoire a fait l'objet d'un travail progressif, construit en lien étroit avec le service métier. À partir des anciennes unités agroclimatiques [13] utilisées dans la région et de documents techniques retrouvés dans les archives, nous avons reconstruit une méthodologie plus rigoureuse et plus transparente. L'objectif : s'éloigner d'une ancienne approche assez subjective, fondée sur des critères difficilement quantifiables (par exemple, la « cohésion sociale » entre unités), pour tendre vers une grille d'analyse climatique plus objective.

Ce travail a abouti à la création d'un jeu de variables synthétiques, correspondant aux conditions agroclimatiques qui influencent directement le comportement des cultures : température, précipitations et sécheresse. Ces indicateurs ont été choisis pour leur lien fort avec les besoins hydriques, les périodes sensibles du cycle végétatif, et la vulnérabilité face aux stress climatiques.

2.1.1 État initial de la base de données

Sélection des paramètres

Toutes les stations ne disposent pas d'un historique complet pour l'ensemble des variables climatiques. Si les données de précipitations quotidiennes sont relativement bien couvertes (recensées dans 141 stations entre 2010 et 2020), les températures minimales et maximales ne sont disponibles que pour 104 stations. Afin de garantir la comparabilité des séries climatiques, un filtre a été appliqué pour ne conserver que les stations disposant simultanément d'un historique sur les précipitations et les températures. Cette sélection aboutit à un jeu final de 103 stations.

L'utilisation d'autres variables, comme l'évapotranspiration potentielle (ETP), a été envisagée. Toutefois, le nombre de stations mesurant cet indicateur étant très limité (trois pour les stations du département et seulement une de celles de Météo France), leur intégration aurait diminué la robustesse statistique de l'analyse. Nous avons donc fait le choix d'exclure ces variables ponctuellement disponibles, pour garantir un socle cohérent et homogène.

Les héritages

Le jeu de données utilisé présente une caractéristique précieuse : la notion d'**héritage**. En effet, la maintenance d'un réseau de stations météorologiques s'accompagne inévitablement de pannes, de déplacements ou de renouvellements de capteurs. Ces interruptions peuvent générer des trous dans les séries temporelles, problématiques pour les analyses de tendance ou de classification.

Le système mis en place permet de limiter cette perte d'information en autorisant, sous certaines conditions, le *report de mesures* entre stations proches géographiquement. Lorsqu'une station est temporairement inopérante, les données peuvent être interpolées ou directement remplacées par celles

d'une station voisine assez proche. De même, en cas de déplacement léger ou de remplacement matériel, un mécanisme d'« héritage » assure la continuité de la série de mesures sur le long terme.

Ce fonctionnement garantit une densité et une profondeur temporelle suffisantes pour établir des profils climatiques exploitables à l'échelle annuelle ou saisonnière.

Requêtes SQL et préparation des indicateurs

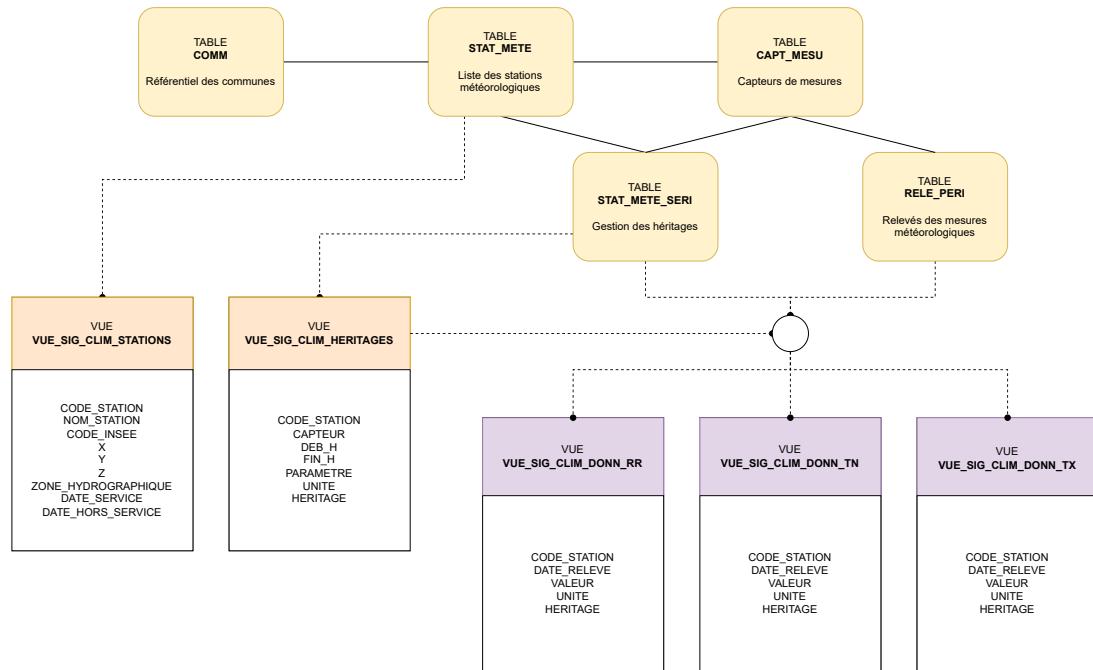


FIGURE 2.1 – Schéma de la base de données climatologique de l'Hérault

La constitution des variables climatiques a nécessité un important travail de préparation des données, reposant sur des requêtes SQL directement exécutées sur les vues agrégées du système d'information climatologique du département. L'extraction des données s'est concentrée sur une décennie complète (2010–2019), permettant une analyse temporelle stable tout en assurant une représentativité climatique récente. Une même étude a été faite parallèlement sur les deux décennies précédentes (soit pour la période de 1990 à 2000 et pour la période de 2000 à 2010) afin d'observer les évolutions des unités que nous construisons.

Une requête SQL a été élaborée pour regrouper les valeurs mesurées par station et par mois, fusionnant trois jeux de données principaux (figure 2.1) : les températures maximales (TX), les températures minimales (TN) et les précipitations (RR). Ces données ont été agrégées à l'échelle mensuelle afin d'obtenir, pour chaque station et chaque mois, les moyennes de températures (minimales et maximales), les extrêmes, ainsi que le cumul mensuel de précipitations.

Cette phase d'agrégation a permis de construire un tableau intermédiaire comportant pour un maximum de stations un enregistrement mensuel avec les principales variables climatiques de base. Ces données brutes ont ensuite été retravaillées (sous-section 2.1.2), afin de calculer des indicateurs plus synthétiques et agronomiquement pertinents.

L'objectif de cette phase de traitement est double : (1) garantir une qualité minimale des données en filtrant les stations incomplètes, et (2) produire des indicateurs agrégés exploitables pour la classification. Ainsi, les stations comportant des valeurs manquantes ou un nombre insuffisant de relevés (moins de 100 mois de données sur les 120 attendus) ont été exclues du jeu final¹, afin d'assurer la robustesse des résultats. Nous avons donc un jeu de données de 77 stations réparties dans l'Hérault et ses alentours.

Chaque station a été enrichie avec ses coordonnées géographiques pour permettre une spatialisation ultérieure des résultats. L'ensemble de ces traitements a conduit à la génération d'un tableau final, consolidant pour chaque station un profil climatique moyen sur la décennie étudiée.

1. Écartées en raison de données manquantes sur de longues périodes, rendant l'imputation inadéquate.

2.1.2 Description des variables utilisées

À partir des données mensuelles nettoyées, plusieurs indicateurs ont été calculés et moyennés sur les 10 années : température moyenne annuelle, cumul annuel des précipitations, précipitations saisonnières (été/hiver) cumulées, températures extrêmes par saison et fréquence des mois secs ou très secs. La détermination des mois secs repose sur une règle agroclimatique : un mois est considéré comme sec si le cumul des précipitations est inférieur à deux fois la moyenne des températures du mois ; il est qualifié de très sec si cette condition est plus stricte ($1,5 \times$ précipitations $< 2 \times$ température moyenne).

Le tableau 2.1 présente l'ensemble des variables climatiques retenues, regroupées selon leur thématique principale (température, précipitations, sécheresse). Chaque variable est accompagnée de son intérêt agronomique, défini en collaboration avec le service métier du département, ainsi que l'étude effectuée par l'observatoire de l'environnement de Bretagne [1, 4].

Thématique	Variable	Intérêt agronomique
Température	Température moyenne annuelle	Indication du niveau thermique général du territoire. Influence la durée des cycles végétatifs, les étages bioclimatiques, ainsi que le choix des espèces et variétés cultivables.
	Température maximale estivale et hivernale	Identification des risques de stress thermique pendant les périodes sensibles (floraison, maturation). Les pics de chaleur peuvent affecter la photosynthèse, la qualité des fruits, ou provoquer l'avortement des fleurs.
	Température minimale estivale et hivernale	Identification des risques de gel hivernal (dommages aux cultures pérennes, mortalité des jeunes plants) et des températures nocturnes trop élevées en été (réduction de la respiration, impact sur le rendement).
Précipitations	Cumul annuel de précipitations	Reflète la disponibilité globale en eau pour les cultures. Essentiel pour estimer les besoins en irrigation et les conditions de recharge des nappes.
	Précipitations estivales et hivernales	La répartition saisonnière est déterminante : un déficit estival peut entraîner un stress hydrique majeur ; un excès hivernal peut engendrer des pertes par lessivage ou saturation des sols.
Sécheresse	Proportion de mois secs et très secs	Indicateur direct de la fréquence et de la sévérité des épisodes de sécheresse. Informe sur la contrainte hydrique subie par les cultures, avec des implications sur la productivité et la résilience agricole.

TABLE 2.1 – Variables climatiques retenues et leur intérêt agronomique

Les variables climatiques présentées dans le tableau 2.1 permettent de définir un ensemble d'indicateurs pertinents pour caractériser les conditions agroclimatiques d'un territoire. Les variables retenues couvrent à la fois des aspects classiques, comme les températures moyennes annuelles ou les cumuls de précipitations, et des indicateurs plus ciblés, tels que les extrêmes saisonniers ou la fréquence des mois secs. Cette combinaison permet de mieux appréhender la variabilité intra-annuelle et les épisodes climatiques pouvant affecter la production agricole. Par exemple, les températures maximales estivales peuvent renseigner sur les risques de stress thermique en période de floraison, tandis que les précipitations hivernales influencent la recharge hydrique des sols. L'objectif est de disposer d'un ensemble de variables à la fois synthétiques, interprétables d'un point de vue agronomique, et suffisamment robustes pour être exploitées dans un cadre de classification ou de modélisation spatiale. Le choix final des indicateurs a été guidé par les contraintes de disponibilité des données, la stabilité des séries temporelles, ainsi que par les recommandations issues de la littérature [19] et des échanges avec les acteurs du domaine.

2.2 Extraction des variables géomorphologiques

Afin de caractériser finement la topographie de la zone d'étude, nous nous appuyons sur un Modèle Numérique de Terrain (MNT). Ce type de donnée représente la surface terrestre sous forme d'une grille régulière, chaque cellule contenant l'altitude à un point donné. Le MNT constitue ainsi une base essentielle pour l'extraction de variables géomorphologiques telles que la pente, l'exposition ou encore divers indices topographiques plus élaborés.

Dans cette étude, nous utilisons un MNT à résolution de 25 mètres, ce qui correspond à une taille de cellule raisonnable pour concilier précision spatiale et efficacité de traitement. Pour limiter les problèmes en bordure lors du calcul des dérivées morphométriques (qui impliquent des fenêtres de voisinage), le MNT a été étendu au-delà des limites strictes de la zone d'étude. Nous limiterons cette grille à l'Hérault et une zone tampon de 15 km lorsque nous travaillerons sur l'interpolation.

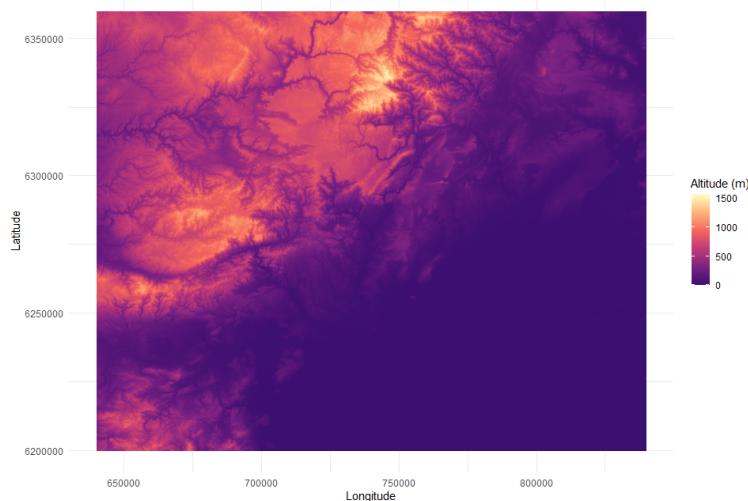


FIGURE 2.2 – Modèle Numérique de Terrain utilisé pour l'extraction des variables géomorphologiques.

Ce MNT (figure 2.2) constitue le support principal pour l'extraction des variables géomorphologiques nécessaires aux étapes suivantes de traitement et de modélisation.

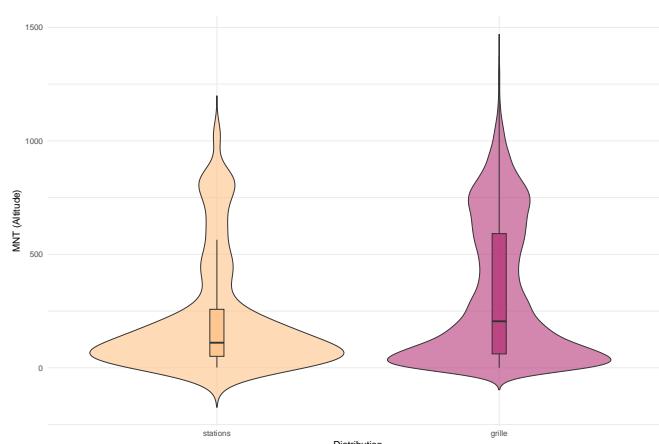


FIGURE 2.3 – Comparaison des distributions d'altitude entre les stations météorologiques (en jaune) et l'ensemble des points de la grille MNT (en rose).

L'analyse comparative des distributions d'altitude (figure 2.3) met en évidence une légère surreprésentation des stations en zones de basse altitude et sous-représentation en haute altitude par rapport à la distribution globale issue de la grille MNT. Cette observation s'explique par le choix réfléchi et stratégique de l'implantation des stations, qui tend à privilégier les zones habitées et accessibles. Toutefois, cette légère disparité reste modérée et ne devrait pas avoir d'impact significatif sur les analyses spatiales à venir. Par conséquent, nous ne considérerons pas cette différence comme un problème dans nos interpolations.

2.2.1 Calcul de la pente et de l'exposition

Les variables géomorphologiques telles que la pente et l'exposition jouent un rôle crucial dans de nombreuses analyses environnementales et territoriales. Elles permettent notamment de mieux comprendre les processus d'érosion, la dynamique hydrologique, ainsi que les conditions microclimatiques influençant la végétation et l'occupation du sol.

La pente, exprimée en degrés, correspond à l'inclinaison locale du terrain et indique la rapidité avec laquelle l'altitude change autour d'un point donné. L'exposition, ou orientation, quant à elle, renseigne sur l'azimut de la pente, c'est-à-dire la direction vers laquelle la surface s'incline, généralement mesurée en degrés à partir du nord.

Pour calculer ces variables, nous avons utilisé la fonction `terrain()` du package `raster` en R, qui applique des algorithmes standards issus de la littérature géomorphologique. Deux méthodes principales sont disponibles : l'algorithme de Fleming-Hoffer [10] et Ritter [17] pour un voisinage à 4 pixels, ainsi que celui de Horn [12] pour un voisinage à 8 pixels. Le choix de l'algorithme dépend du type de surface étudiée, Horn étant généralement recommandé pour des terrains rugueux, tandis que Fleming et Hoffer conviennent mieux aux surfaces plus lisses.

Dans notre cas, la méthode employée tient compte d'un voisinage de 8 pixels, privilégiant ainsi la précision sur les reliefs complexes.

Le calcul repose sur l'estimation des dérivées partielles de l'altitude selon les directions horizontale (x) et verticale (y), notées respectivement p et q . Ces dérivées sont approximées à partir des altitudes des pixels voisins dans une fenêtre 3×3 centrée sur le pixel étudié.

Soient les altitudes des pixels de la fenêtre organisées comme sur la figure 2.4 :

z_1	z_2	z_3
z_4	z_5	z_6
z_7	z_8	z_9

FIGURE 2.4 – Disposition des pixels autour du pixel central z_5 pour le calcul des dérivées partielles

Les dérivées partielles p et q pour z_5 sont calculées selon Horn par :

$$p = \frac{(z_3 + 2z_6 + z_9) - (z_1 + 2z_4 + z_7)}{8r}, \quad q = \frac{(z_7 + 2z_8 + z_9) - (z_1 + 2z_2 + z_3)}{8r},$$

avec r la résolution spatiale de la grille (taille d'une cellule).

La **pente** S_d en degrés s'obtient via :

$$S_d = \arctan\left(\sqrt{p^2 + q^2}\right) \times \frac{180}{\pi}. \quad (2.1)$$

L'**exposition** A , exprimée en degrés, est définie par :

$$A = \arctan 2(q, -p) \times \frac{180}{\pi},$$

où `arctan2` est la fonction arc tangente à deux arguments, permettant de déterminer correctement le quadrant de l'angle. Afin d'obtenir une valeur dans $[0^\circ, 360^\circ]$, on normalise A par :

$$\text{si } A < 0, \quad A \leftarrow 360 + A.$$

Par convention, si la pente est nulle (surface plane), l'exposition est fixée à 90° .



FIGURE 2.5 – Carte de la pente du territoire étudié, extraite du MNT.

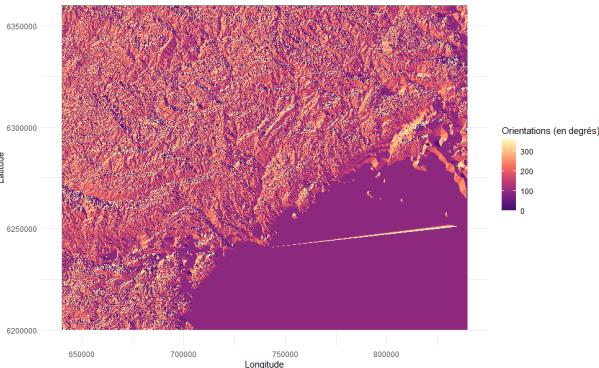


FIGURE 2.6 – Carte de l'exposition (orientation) du territoire étudié, extraite du MNT.

Ces cartes illustrent la variabilité spatiale des pentes et des orientations du terrain sur l'ensemble de la zone d'étude. La carte de pente (figure 2.5) met en évidence les secteurs fortement inclinés, potentiellement sujets à l'érosion ou à des phénomènes spécifiques liés au relief. La carte d'exposition (figure 2.6) permet quant à elle d'identifier les versants orientés dans différentes directions, ce qui influence notamment l'ensoleillement, les conditions microclimatiques et la répartition de la végétation. À noter qu'une anomalie visuelle apparaît sous la forme d'une ligne d'erreur en mer Méditerranée, résultant de l'algorithme d'extraction appliqué à des zones sans relief réel. Cette irrégularité n'aura toutefois aucune incidence sur nos analyses, puisque seules les zones terrestres incluses dans le buffer de 15 km autour du département de l'Hérault seront conservées pour la suite de l'étude.

L'intégration de ces variables dans les étapes ultérieures de modélisation et d'interpolation apportera ainsi une meilleure compréhension des interactions entre le relief et les phénomènes étudiés.

2.2.2 Extraction de la platitude des fonds de vallée

L'indice MRVBF (*Multi-Resolution Valley Bottom Flatness* ou l'indice de platitude des fonds de vallée multi-résolution) [11] permet de détecter automatiquement les fonds de vallée à travers une approche multi-échelle. Le principe repose sur l'analyse conjointe de deux attributs topographiques à plusieurs résolutions : la pente et le percentile d'altitude (*elevation percentile*). À chaque échelle, ces deux indicateurs sont combinés, transformés non linéairement, puis intégrés dans une chaîne de pondérations progressives. Le résultat final donne une carte continue où des valeurs élevées (>1.5) désignent des zones de fonds de vallée.

Étape 1 : Calcul à la résolution native

Deux variables sont d'abord calculées à partir du MNT d'origine :

- **La pente S** , exprimée en % comme $S = 100 \cdot \tan(S_d)$, avec S_d , la pente en degrés de la partie précédente (équation 2.1).
- **Le percentile d'altitude $PCTL$** , qui évalue la position relative d'un pixel dans le paysage. Il est défini comme le ratio du nombre de pixels d'altitude inférieure au total des pixels dans un voisinage circulaire (rayon : 3 cellules).

Ces deux valeurs sont ensuite transformées en indices de « platitude » et de « bassesse » à l'aide de la fonction sigmoïde suivante :

$$f(x) = \frac{1}{1 + \left(\frac{x}{t}\right)^p}$$

où :

- x est la variable d'entrée (pente ou percentile),
- t est un seuil (par exemple $t = 0.4$ pour les percentiles),
- p est un paramètre de forme (plus il est grand, plus la transition est brutale).

La combinaison de la **platitude** F_1 et de la **bassesse** donne un indice intermédiaire de « caractère fond de vallée » :

$$VF_1 = f_{\text{slope}}(S_1) \cdot f_{\text{percentile}}(PCTL_1)$$

Cet indice est ensuite transformé à nouveau (via la même fonction sigmoïde) pour obtenir une valeur comprise entre 0 et 1.

Étape 2 : Première généralisation

On recommence le processus avec un seuil de pente plus bas (typiquement divisé par 2), et un voisinage plus large (rayon : 6 cellules). Cela permet de détecter des vallées plus larges.

Le résultat donne un nouvel indice VF_2 , qui est ensuite combiné au résultat précédent pour former le premier niveau multi-résolution :

$$MRVBF_2 = (1 - w) \cdot VF_1 + w \cdot (1 + VF_2)$$

où le poids w est défini comme :

$$w = \frac{1}{1 + \left(\frac{VF_2}{t}\right)^{-p}}$$

avec des valeurs typiques $t = 0.4$ et $p \approx 6.68$, choisies pour garantir une transition fluide grâce à la documentation du module de **SAGA** [18].

Étapes suivantes : analyse multi-échelle

Les étapes ultérieures appliquent le même schéma, mais :

- le MNT est d'abord **lissé** (filtrage gaussien 11x11),
- puis **agrégé** (coarsening) pour augmenter la taille des pixels d'un facteur 3 à chaque fois,
- la pente est recalculée à partir du MNT lissé,
- le percentile est recalculé sur ce MNT agrégé,
- les résultats sont combinés avec les indices précédents.

À chaque niveau L , un indice de platitude cumulée CF_L est calculé :

$$CF_L = CF_{L-1} \cdot f_{\text{slope}}(S_L)$$

Puis un nouvel indice de caractère « fond de vallée » est obtenu :

$$VF_L = CF_L \cdot f_{\text{percentile}}(PCTL_L)$$

Cet indice est encore une fois pondéré et combiné à l'accumulation précédente :

$$MRVBF_L = (1 - w_L) \cdot MRVBF_{L-1} + w_L \cdot (L - 0.5 + VF_L)$$

où w_L est la pondération issue de VF_L , avec des paramètres $t = 0.4$ et p_L ajusté pour que :

$$MRVBF_L = L \quad \text{quand } VF_L = 0.6$$

L'indice MRVBF combine une série d'analyses morphométriques hiérarchiques pour extraire des formes de terrain associées aux vallées, en tenant compte de leur position relative et de leur platitude à plusieurs résolutions. Ce traitement permet une caractérisation continue, robuste et multi-échelle des fonds de vallée.

Cette méthode a été implémentée à l'aide du module MRVBF de la librairie `ta_morphometry` (analyse de terrain morphométrique) dans **SAGA GIS**, via l'interface R `RSAGA`. Les paramètres permettent d'ajuster la sensibilité de l'algorithme en fonction du type de relief étudié. Dans notre cas, les seuils ont été paramétrés pour favoriser la détection de fonds de vallée larges mais également de talwags (ligne de plus grande pente d'une vallée) plus discrets.

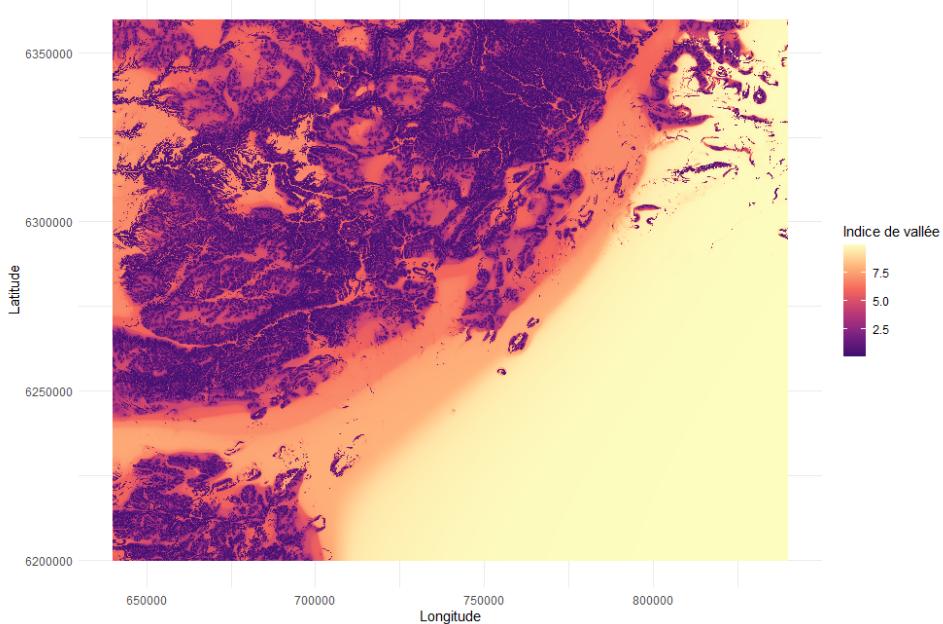


FIGURE 2.7 – Indice MRVBF (platitude des fonds de vallée) extrait à partir du MNT 25 m. Les valeurs élevées indiquent une forte propension à se situer en fond de vallée.

La carte présentée en figure 2.7 permet de visualiser les zones du territoire présentant une forte propension d'appartenir à un fond de vallée. On y distingue clairement les grandes vallées principales mais aussi de nombreux talwegs secondaires, soulignant la sensibilité du MRVBF à la structure du réseau hydrographique.

Récapitulatif des variables géomorphologiques

Les variables géomorphologiques retenues dans cette étude incluent l'altitude, la pente, l'orientation (ou exposition), ainsi que l'indice MRVBF, utilisé pour caractériser les zones d'accumulation potentielles. À cela s'ajoute la distance à la mer Méditerranée, qui constitue un facteur pertinent dans l'interprétation des gradients climatiques, notamment en lien avec l'humidité de l'air ou les effets thermiques modérateurs du littoral.

Le calcul de cette distance a été réalisé à partir d'un masque vectoriel de la mer Méditerranée, en mesurant pour chaque cellule de la grille régulière la distance euclidienne au trait de côte.

Pour extraire ces données au niveau des stations météorologiques, il suffit d'utiliser leurs coordonnées géographiques (X, Y) et d'interroger la cellule correspondante dans les couches raster. Cela permet d'associer à chaque station un ensemble cohérent de variables géomorphologiques, qui pourront ensuite être utilisées comme covariables dans un modèle d'interpolation spatiale.

Chapitre 3

Classification des stations

L'objectif de cette section est de regrouper les stations climatiques en zones homogènes à partir d'un ensemble de variables environnementales. Cette classification permet d'identifier des régions présentant des conditions agroclimatiques similaires, facilitant ainsi leur analyse et interprétation agronomique. Dans ce chapitre, l'accent sera mis uniquement sur les variables dynamiques, les variables géomorphologiques étant réservées à une étape ultérieure afin d'affiner la définition des groupes. Deux critères principaux ont guidé cette approche :

- une certaine contiguïté spatiale entre stations d'un même groupe (éviter les groupes dispersés),
- une taille de groupe suffisante pour que les divisions aient du sens.

Nous avons tout d'abord appliqué une méthode de classification non supervisée classique en deux étapes : une **classification ascendante hiérarchique (CAH)** permettant d'explorer la structure naturelle des données, puis une **classification par *k-means*** visant à stabiliser la partition obtenue. Les détails méthodologiques sont exposés ci-dessous.

3.1 Classification non supervisée par *k-means*

Avant toute classification, l'ensemble des variables climatiques sélectionnées ont été **centrées et réduites**, c'est-à-dire transformées selon la formule suivante :

$$x' = \frac{x - \mu}{\sigma}$$

où x est une valeur brute, μ la moyenne de la variable et σ son écart-type. Cette standardisation est indispensable ici, car les variables utilisées (températures, précipitations, fréquences) sont exprimées dans des unités différentes et avec des amplitudes de variation très hétérogènes.

3.1.1 Classification ascendante hiérarchique (CAH)

La CAH est une méthode de partitionnement qui construit de manière récursive une hiérarchie d'agrégation entre les observations, sans supposer a priori le nombre de groupes. On commence avec chaque station dans un groupe séparé, puis on fusionne itérativement les groupes les plus proches, jusqu'à obtenir une seule classe globale.

Distance utilisée : critère de Ward

Le critère de fusion utilisé est la méthode de **Ward**, qui cherche à minimiser la *variance intra-groupe*. Plus précisément, à chaque étape, on fusionne les deux clusters A et B qui engendrent l'augmentation minimale de l'inertie intra-classe :

$$\Delta(A, B) = \frac{|A||B|}{|A| + |B|} \|\bar{x}_A - \bar{x}_B\|^2$$

où $|A|$ et $|B|$ sont les tailles des deux clusters, et \bar{x}_A , \bar{x}_B leurs centres de gravité respectifs. Cette distance est équivalente à un critère de somme des carrés, ce qui en fait un choix naturel pour des données numériques standardisées.

Choix du nombre de classes

Le dendrogramme issu de la CAH (figure 3.1) permet d'identifier des ruptures dans la structure d'agrégation. Un seuil de **5 classes** a été retenu ici, en analysant la courbe d'inertie et en tenant compte des contraintes d'interprétation cartographique.

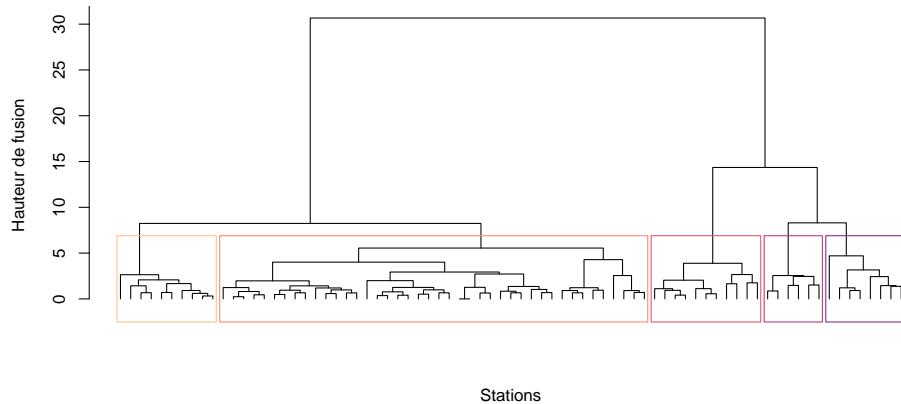


FIGURE 3.1 – Dendrogramme de la classification hiérarchique (méthode de Ward)

Ce seuil de 5 classes a été retenu en concertation avec les services métier. Une partition en 3 classes a été jugée trop grossière pour capturer la diversité des profils, tandis qu'un découpage en plus de 5 classes produisait des groupes trop fins, difficilement interprétables dans le cadre de l'analyse cartographique. Ce compromis permet donc un bon équilibre entre lisibilité des résultats et l'interprétation.

3.1.2 Stabilisation par *k-means*

La classification par *k-means* est une méthode de partitionnement itérative qui vise à minimiser la variance intra-classe en ajustant les centres de chaque groupe. Le principe est le suivant :

1. Initialisation des centres (ici, à partir des barycentres issus de la CAH) ;
2. Affectation de chaque observation au centre le plus proche (selon la distance euclidienne) ;
3. Mise à jour des centres comme moyenne des observations dans chaque cluster ;
4. Répétition des étapes 2–3 jusqu'à convergence (stabilité des clusters).

Ce couplage CAH–*k-means* permet de bénéficier d'une bonne initialisation (issue de la CAH), tout en stabilisant la partition finale, ce que la CAH seule ne garantit pas.

Résultats cartographiques

Pour représenter la répartition spatiale des groupes obtenus, nous avons appliqué deux méthodes de classification, à savoir la classification ascendante hiérarchique (CAH) et le *k-means*, chacune produisant cinq groupes. Afin de mieux visualiser la contiguïté spatiale des clusters, nous avons choisi de représenter les polygones de Voronoï associés à chaque station. Ces polygones définissent une partition de l'espace en attribuant à chaque station la région géographique dont elle est la plus proche, ce qui facilite une interpolation simple.

Formellement, soit un ensemble de stations avec des coordonnées géographiques $\{(x_i, y_i)\}_{i=1}^n$, où x_i et y_i représentent respectivement la longitude et la latitude de la station i .

Le polygone de Voronoï V_i associé à la station i est défini par :

$$V_i = \{(x, y) \in \mathbb{R}^2 \mid d((x, y), (x_i, y_i)) \leq d((x, y), (x_j, y_j)), \quad \forall j \neq i\}$$

où d est la distance euclidienne dans le plan :

$$d((x, y), (x', y')) = \sqrt{(x - x')^2 + (y - y')^2}.$$

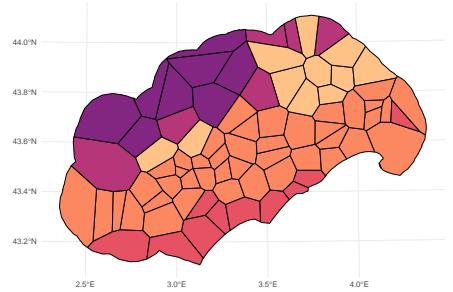
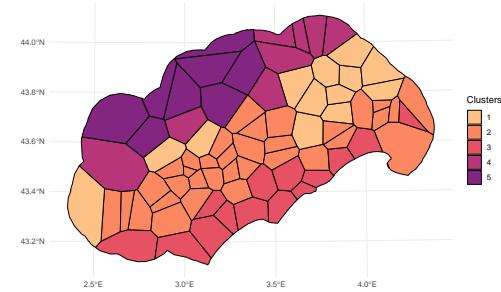


FIGURE 3.2 – Classification par CAH

FIGURE 3.3 – Classification par *k-means*

Les résultats obtenus montrent une similarité globale dans la structuration des stations, les deux méthodes identifiant des groupes relativement homogènes. Cependant, la classification par *k-means* figure 3.3 semble offrir une meilleure continuité spatiale que la classification par CAH figure 3.2, notamment en zone littorale où une région distincte se dessine plus nettement. Cette différence s'explique par le fait que le *k-means* optimise directement la partition dans l'espace des variables, tandis que la CAH, qui repose sur des regroupements hiérarchiques successifs, peut engendrer des clusters plus fragmentés.

Malgré ces différences, les cartes restent assez proches dans leur représentation générale. Cette première approche fournit donc un découpage objectif des stations selon leurs caractéristiques climatiques. Néanmoins, l'absence de contrainte spatiale explicite dans ces algorithmes peut générer des groupes peu réalistes sur le plan géographique, avec des zones parfois disjointes.

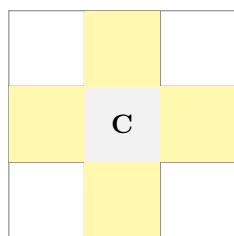
Pour pallier cette limite, la suite de ce travail portera sur des méthodes intégrant la dimension spatiale, afin de produire des zones homogènes tout en assurant leur contiguïté géographique, condition essentielle pour une interprétation agroclimatique pertinente.

3.2 Classification spatiale via la suppression des arêtes d'un arbre

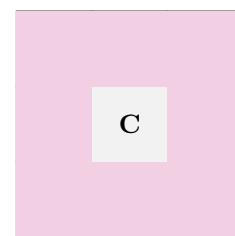
Pour partitionner l'espace de manière contiguë, nous pouvons utiliser l'algorithme SKATER (*Spatial 'K'luster Analysis by Tree Edge Removal*), introduit par Assunção et al. (2006) [2]. Cette méthode s'appuie sur un graphe de voisinage pondéré et effectue des coupures successives dans un arbre de poids minimal pour créer des groupes homogènes sur des variables attributaires, tout en maintenant la connectivité spatiale.

3.2.1 Principe de construction du graphe

Le voisinage entre polygones est défini à partir de la structure de Voronoï construite autour des stations météorologiques.



Voisinage rook



Voisinage queen

FIGURE 3.4 – Comparaison entre les voisinages **rook** (adjacents par côté uniquement) et **queen** (adjacents par côté ou sommet) autour d'une cellule centrale.

Dans notre cas, les stations météorologiques sont associées à des polygones de Voronoï, qui définissent des zones s'étendant jusqu'à mi-distance entre stations voisines. Par construction, ces polygones ne peuvent se toucher que par une arête, et non par un simple sommet. En effet les stations de l'Hérault n'ont pas été placées dans une grille régulière. Ainsi, les voisinages de type `rook` (adjacence par côté uniquement) et `queen` (adjacence par côté ou sommet) sont équivalents dans notre configuration. Cette distinction théorique est illustrée en figure 3.4. Nous optons donc pour une contiguïté de type `rook` (`queen = FALSE`).

La figure 3.5 illustre le graphe de voisinage initial entre les stations. Ce graphe non orienté permet de modéliser les relations spatiales sous-jacentes, chaque sommet représentant une station, et chaque arête représentant une frontière commune entre polygones.

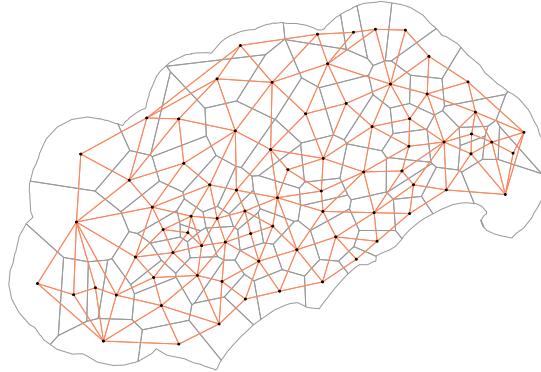


FIGURE 3.5 – Graphe de voisinage basé sur la contiguïté des polygones de Voronoï

À partir du graphe de voisinage, nous construisons un arbre de poids minimal (MST : *Minimum Spanning Tree*) à l'aide de l'algorithme de *Prim* [16]. L'objectif est de relier toutes les stations météorologiques (les noeuds du graphe) avec le coût total minimal, sans créer de cycles.

Chaque poids d'arête est défini par une dissimilarité entre deux stations voisines, calculée via la distance euclidienne dans l'espace des variables climatiques standardisées. Pour deux stations i et j , cette dissimilarité est donnée par :

$$d_{ij} = \sqrt{\sum_{k=1}^p (z_{ik} - z_{jk})^2}$$

où p est le nombre de variables, et z_{ik} la variable k standardisée pour la station i .

On considère un graphe non orienté, connexe et pondéré noté $G = (V, E)$, où :

- V est l'ensemble des sommets ;
- E est l'ensemble des arêtes, chaque arête étant associée à un poids (ou coût).

Algorithme 1 : Algorithme de Prim

Entrée : Un graphe pondéré connexe $G = (V, E)$

Sortie : Un arbre couvrant minimal T

Initialisation : Initialiser l'arbre T avec un sommet arbitraire $v_0 \in V$;

Tant que il existe un sommet $v \in V \setminus T$ Trouver l'arête (u, v) de poids minimal telle que $u \in T$ et $v \notin T$;

Ajouter v et l'arête (u, v) à T ;

Retourner T ;

Formellement, on cherche un sous-ensemble $T \subseteq E$ tel que :

$$T = \arg \min_{T \subseteq E} \sum_{(i,j) \in T} w_{ij}, \quad \text{sous la contrainte que } (V, T) \text{ est connexe et sans cycles.}$$

Le résultat est illustré en figure 3.6, où seules les arêtes retenues par l’arbre couvrant minimal sont représentées.

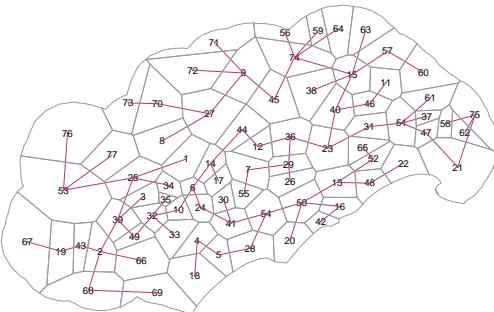


FIGURE 3.6 – Arbre de poids minimal sur les stations

Cet arbre est, par construction, connexe et sans cycle. Toute coupure d’une arête entraîne la division de l’arbre en deux composantes connexes distinctes. L’enjeu est désormais d’identifier les coupures pertinentes à effectuer. Par exemple, supprimer l’arête reliant les stations 24 et 41 permettrait d’isoler une zone de stations littorales de toutes les autres stations.

3.2.2 Résultats de la classification

Découpe automatique par SKATER

L’algorithme **SKATER** procède à une suppression itérative des arêtes les plus coûteuses dans l’arbre, c’est-à-dire celles dont la dissimilarité entre les nœuds est la plus élevée. Pour obtenir K clusters, l’algorithme effectue $K - 1$ coupures.

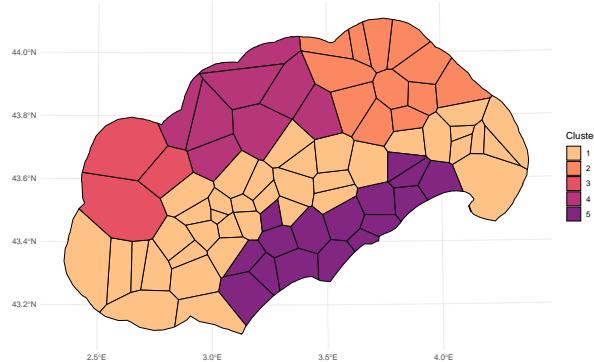


FIGURE 3.7 – Découpe automatique en 5 clusters (sans contrainte de taille)

La figure 3.7 présente le résultat d’une classification en $K=5$ clusters sans contrainte de taille. L’organisation spatiale paraît cohérente : chaque cluster forme une zone contiguë, sans discontinuité visible. Cependant, cette apparente homogénéité masque un déséquilibre notable en termes de nombre de stations. Certains groupes couvrent de vastes zones géographiques (en raison de l’utilisation des polygones de Voronoï) mais ne regroupent en réalité qu’un nombre très restreint de stations, ce qui limite leur représentativité. C’est notamment le cas du groupe 3, qui ne comprend que trois stations.

Classification équilibrée

Afin de garantir une meilleure homogénéité en termes de nombre d'individus à l'intérieur des clusters, nous introduisons une contrainte sur le nombre minimal d'éléments par groupe, via l'argument `crit` de la fonction `skater`.

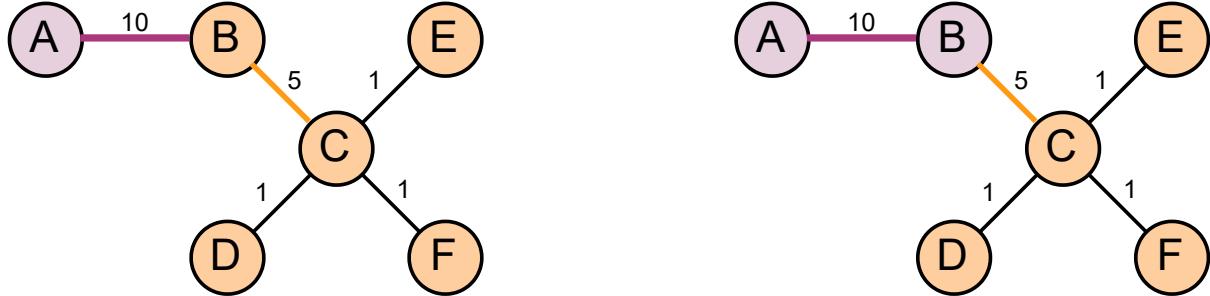


FIGURE 3.8 – Exemple de classification sans l'argument `crit` et avec `crit=2` pour $K = 2$

Dans cet exemple (figure 3.8), sans l'argument `crit`, l'algorithme décide de couper l'arbre entre les sommets A et B, à l'endroit où l'arête a le poids maximal (10), créant ainsi un cluster composé d'un seul noeud. Lorsque l'on utilise l'argument `crit`, l'algorithme parcourt l'arbre à la recherche de l'arête de poids maximal et s'arrête également entre A et B. Cependant, la condition `crit = 2` n'est pas respectée : cette arête est donc exclue du champ des possibles. L'algorithme continue alors à itérer en suivant l'ordre décroissant des poids des arêtes. Dans cet exemple, la coupe se fera finalement entre B et C. On perd ainsi un peu d'homogénéité dans les groupes (les sommets A et B étant bien plus dissemblables que B et C) mais la condition de nombre minimal d'individus par cluster est respectée.

Pour notre cas concret, nous imposons un plancher à 10 stations par cluster. Au-delà (par exemple 11), la contrainte devient trop forte pour permettre la création de $K = 5$ groupes contigus et l'algorithme ne couperait l'arbre que 3 fois (donc ne créerait que 4 groupes).

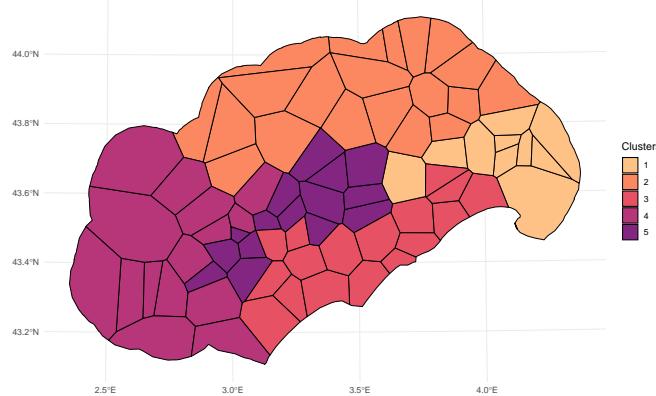


FIGURE 3.9 – Classification spatiale équilibrée (minimum 10 stations par groupe)

On constate de cette classification (figure 3.9) un meilleur équilibre entre les tailles des groupes, au prix d'une légère perte de compacité spatiale. Certains clusters deviennent un peu plus étendus ou moins compacts, mais la contrainte sur le nombre d'individus est respectée.

Comparaison et analyse

Les figures figure 3.7 et figure 3.9 illustrent les résultats de l'algorithme SKATER dans deux configurations : une version sans contrainte (automatique) et une version avec contrainte, imposant un nombre minimal de stations par groupe.

Cluster	Découpe automatique	Découpe avec nombre d'individus minimal
1	48.37	10.82
2	19.92	48.66
3	6.06	17.43
4	12.17	44.23
5	16.49	14.46

TABLE 3.1 – Sommes des poids des arêtes internes par cluster pour les deux cas

Dans le cas non contraint, l'algorithme privilégie l'homogénéité interne des groupes, ce qui peut entraîner la formation de clusters de taille très inégale. En revanche, l'introduction d'une contrainte sur la taille minimale des groupes permet un meilleur équilibre entre clusters, mais se fait au détriment de leur homogénéité interne. Comme le montre le tableau 3.1, la somme des poids des arêtes internes est globalement plus élevée dans le cas contraint, traduisant une plus grande hétérogénéité intra-cluster.

Dans les deux cas, l'algorithme produit des groupes spatialement contigus, ce qui constitue un avantage important par rapport à des méthodes comme le k-means, qui ne prennent pas en compte la structure spatiale des données.

3.3 Classification spatiale basée sur des distances combinées

L'objectif de l'approche **ClustGeo** [6] est de réaliser une classification ascendante hiérarchique (CAH) en intégrant à la fois des données agroclimatiques et une contrainte de proximité spatiale.

3.3.1 Matrices de distances et choix des paramètres

Matrices de distances

Pour commencer, on utilise deux matrices de dissimilarités normalisées et de même dimension $n \times n$:

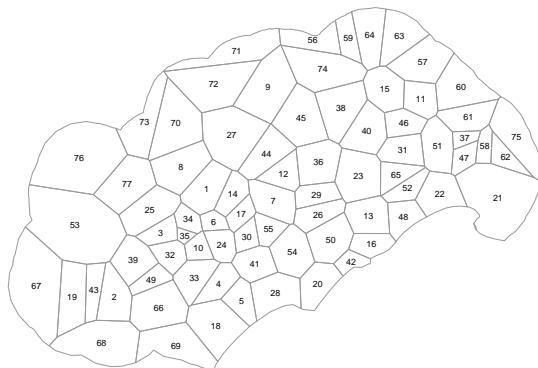


FIGURE 3.10 – Carte des index des stations

La figure 3.10 nous indique les index des stations qui seront utiles pour la lecture des matrices de distances suivantes :

- La matrice D_0 , qui représente les dissimilarités agroclimatiques entre les n stations, calculée à partir d'une matrice de données $X \in \mathbb{R}^{n \times p}$ (avec p variables quantitatives) via des distances euclidiennes.
- La matrice D_1 , qui contient les distances spatiales entre individus, ici fondée sur des distances géographiques ou topologiques.

Les matrices obtenues sont illustrées ci-dessous :

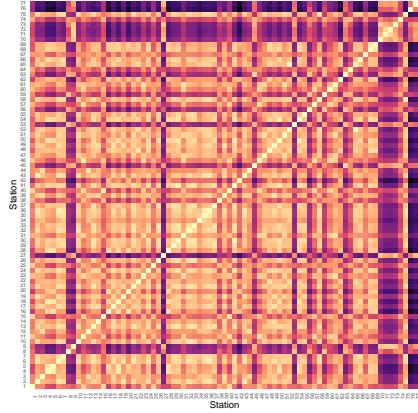


FIGURE 3.11 – Matrice D_0 : dissimilarités agroclimatiques (euclidiennes)

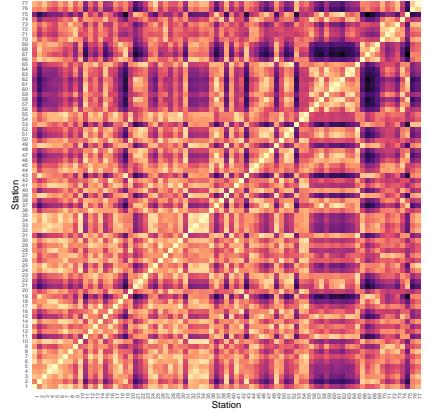


FIGURE 3.12 – Matrice D_1 : distances spatiales

La figure 3.11 montre le dissimilarités agroclimatiques entre les stations. Pour regarder plus en détail, les stations d'index 3 et 77 sont très différentes au sens de la matrice D_0 malgré le fait qu'elles soient géographiquement relativement proches comme on peut le voir dans la figure 3.12 avec la distance D_1 . Cela s'explique par le climat beaucoup plus montagneux dans le nord du département tandis que la station 3 se trouve beaucoup moins en altitude et ce, malgré le fait qu'elles soient à moins de 25 kilomètres l'une de l'autre.

Choix des paramètres

Pour chaque cluster C_k , la méthode définit un critère d'hétérogénéité pondéré [7] :

$$H(C_k) = \alpha \cdot I(C_k, D_0) + (1 - \alpha) \cdot I(C_k, D_1)$$

où $I(C_k, D_i)$ désigne l'inertie intra-classe de C_k calculée à partir de la matrice D_i ($i = 0, 1$). Ce critère permet d'ajuster l'équilibre entre homogénéité thématique et proximité spatiale à l'aide du paramètre $\alpha \in [0, 1]$.

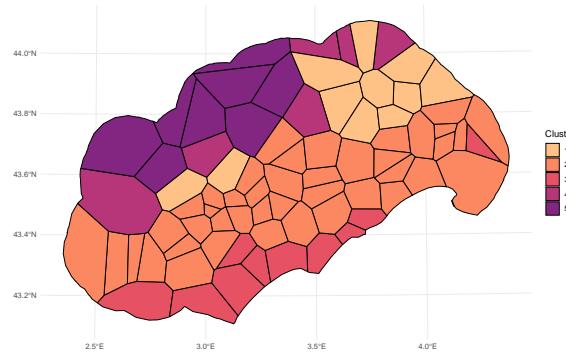
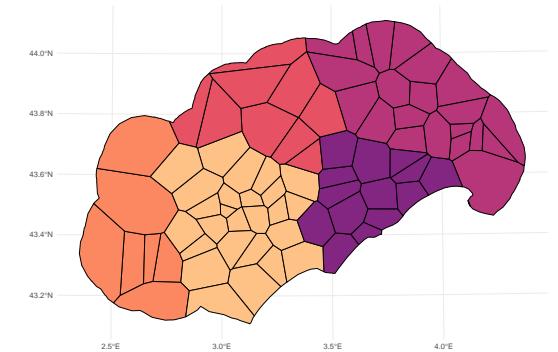
Le critère d'hétérogénéité total pour une partition $\mathcal{P}_K = \{C_1, \dots, C_K\}$ est alors défini comme la somme des critères $H(C_k)$ calculés pour chaque cluster :

$$H(\mathcal{P}_K) = \sum_{k=1}^K H(C_k).$$

Ce critère global sert de fonction objective que l'algorithme de classification hiérarchique cherche à minimiser afin d'obtenir une partition optimisée. Le paramètre α joue un rôle central dans l'équilibre entre la prise en compte des distances thématiques et spatiales.

Choix du paramètre α :

Pour mieux comprendre l'impact du paramètre α sur la classification, nous présentons ici deux partitions extrêmes obtenues par CAH avec $\alpha = 0$ et $\alpha = 1$ respectivement :

FIGURE 3.13 – Classification avec $\alpha = 0$ FIGURE 3.14 – Classification avec $\alpha = 1$

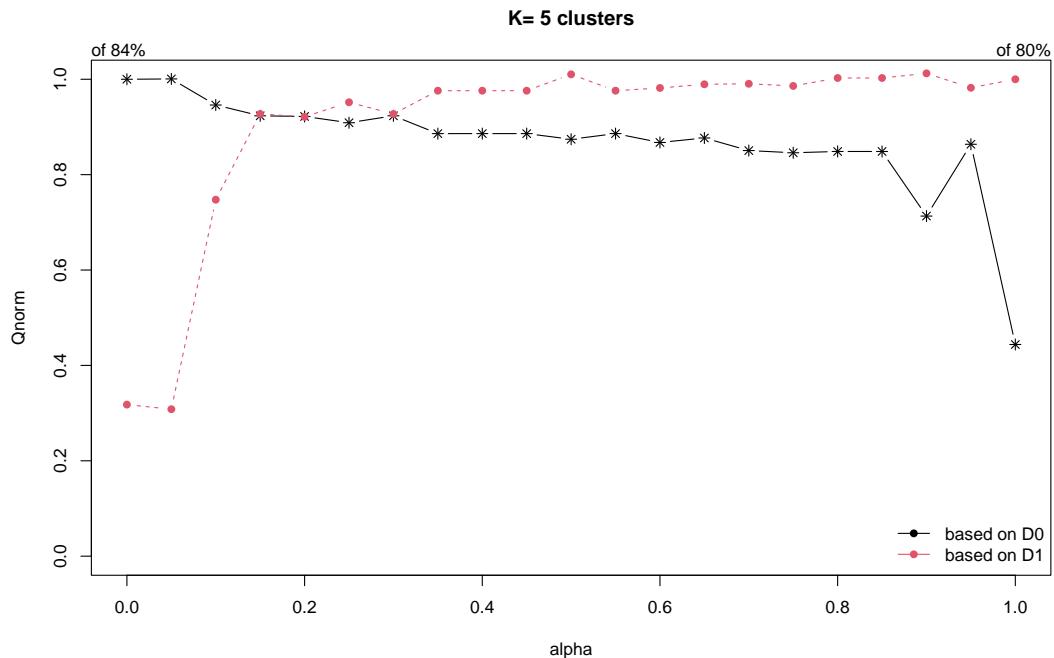
D'une part, en choisissant $\alpha = 0$ (figure 3.13), la classification par CAH repose uniquement sur les variables agroclimatiques et nous obtenons de nouveau la classification vue précédemment dans ce chapitre. D'autre part, pour $\alpha = 1$ (figure 3.14), la classification est entièrement basée sur les distances géographiques entre stations, ce qui aboutit à des groupes contigus.

Le choix du α optimal repose donc sur un compromis entre la qualité thématique et la cohérence spatiale. Pour cela, on s'appuie sur deux critères de qualité définis par :

$$Q_\alpha^{(D_0)} = 1 - \frac{I_\alpha(D_0)}{I(D_0)}, \quad Q_\alpha^{(D_1)} = 1 - \frac{I_\alpha(D_1)}{I(D_1)}$$

où $I_\alpha(D_i)$ désigne l'inertie intra-classe pour la partition obtenue avec un certain α , mesurée dans l'espace défini par la matrice de dissimilarité D_i , et $I(D_i)$ est l'inertie totale (avant partition). Ces coefficients mesurent donc respectivement la proportion d'inertie expliquée par la classification dans les espaces thématique (D_0) et spatial (D_1).

En faisant varier α , on peut suivre l'évolution conjointe de $Q_\alpha^{(D_0)}$ et $Q_\alpha^{(D_1)}$, ce qui permet d'identifier une valeur de compromis.

FIGURE 3.15 – Évolution des critères d'homogénéité thématique $Q_\alpha^{(D_0)}$ et spatiale $Q_\alpha^{(D_1)}$

La valeur optimale de α est choisie comme un compromis visuel sur la figure 3.15. Nous voulons quand même prioriser la distance D_0 car c'est celle de nos variables agroclimatiques. Le meilleur compromis, qui maximise la distance D_1 , sans trop sacrifier la cohérence sur D_0 , est ici obtenu pour $\alpha = 0.15$.

3.3.2 Résultat de la classification

La carte finale, présentée ci-dessous (figure 3.16), illustre la partition des stations en $K = 5$ groupes, obtenue avec un $\alpha = 0.15$. Cette classification respecte à la fois la proximité spatiale et l'homogénéité agroclimatique des stations.

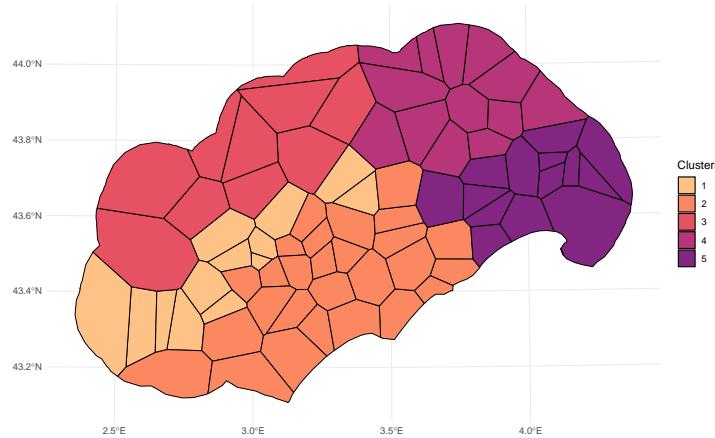


FIGURE 3.16 – Partition finale obtenue par ClustGeo avec $\alpha = 0.15$

Cette méthode permet donc d'obtenir des clusters à la fois pertinents du point de vue thématique et cohérents géographiquement, en évitant les déséquilibres ou discontinuités constatés dans la classification sans contrainte.

On note toutefois que la contiguïté avec les polygones de Voronoï n'est pas respectée dans le cas optimal. Cette observation est due à l'interpolation trop simple de nos stations qui ne prend pas en compte les variables géomorphologiques. En réalité, et comme nous le verrons dans la partie suivante, le groupe 1 représente une zone structurante du département qui sépare le littoral sud des plateaux du nord-ouest.

3.4 Discussion sur les méthodes

Les trois approches testées ont chacune apporté des informations complémentaires sur la structuration spatiale et climatique des stations.

Le k-means, appliqué aux seules variables agroclimatiques, fournit une classification thématique efficace, mais ignore totalement la géographie. Il en résulte des groupes parfois discontinus, peu cohérents sur le plan spatial, malgré une bonne séparation climatique.

L'algorithme SKATER, basé sur les relations de voisinage, introduit la dimension spatiale. Sans contrainte, il favorise des groupes très homogènes mais déséquilibrés en taille. En imposant un seuil minimal de stations par cluster, la répartition devient plus équilibrée, au prix d'une perte d'homogénéité interne. Cette méthode permet néanmoins une segmentation contiguë et opérationnelle du territoire.

La méthode ClustGeo combine ces deux logiques. En ajustant le paramètre α , elle équilibre la contribution des distances spatiales et des différences thématiques. Le résultat obtenu avec $\alpha = 0.15$ permet une partition géographiquement cohérente tout en conservant des groupes agroclimatiquement homogènes. Ce compromis la rend particulièrement pertinente pour une interprétation cartographique robuste.

Au vu des résultats, et après discussion avec les services concernés, nous avons décidé de garder la répartition obtenue par la méthode de classification spatiale basée sur des distances combinées. L'étude en parallèle sur les deux autres décennies utilisant cette méthode est donc disponible en annexe (sous-section I.1).

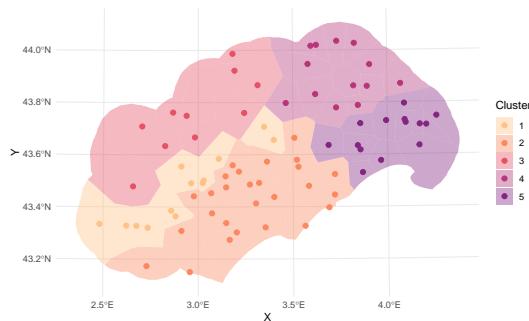


FIGURE 3.17 – Partition spatiale des stations selon ClustGeo

La carte de la figure 3.17 présente cette classification finale, mettant en évidence cinq groupes à la fois spatialement cohérents et agroclimatiquement distincts à l'échelle du territoire étudié. L'information clé réside dans la localisation des stations et leur appartenance aux groupes, plutôt que dans le fond coloré, qui peut être biaisé par l'utilisation des polygones de Voronoï.

Afin de mieux caractériser les profils agroclimatiques associés à chacun de ces groupes, nous analysons la distribution de plusieurs variables clés. Les figures suivantes présentent, à titre d'exemple, la répartition des précipitations estivales et de la fréquence des mois secs dans les cinq clusters identifiés. Celles-ci ont été choisies car elles sont les plus pertinentes.

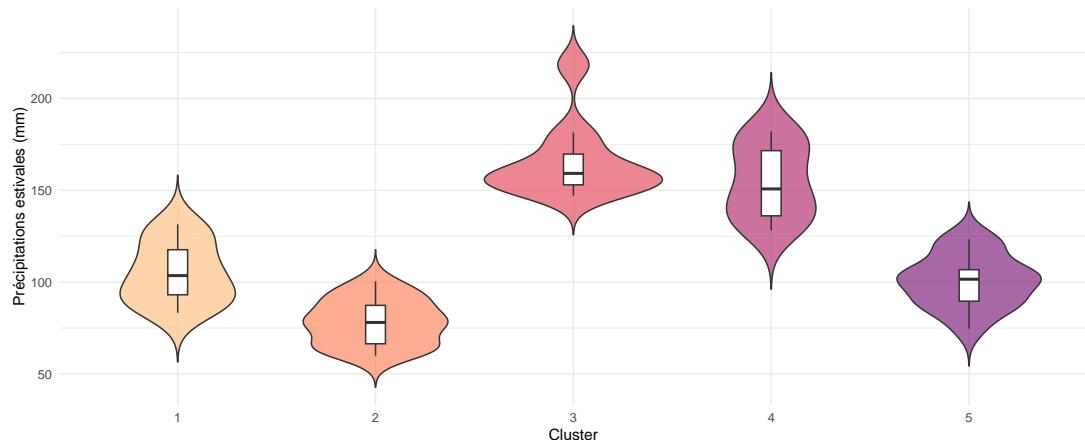


FIGURE 3.18 – Distribution des précipitations estivales dans chaque groupe

Pour les précipitations estivales (figure 3.18), on observe une faible pluviométrie dans les groupes 1, 2 et 5, situés principalement au sud du territoire. Le groupe 3 reçoit plus de précipitations estivales que le groupe 4, tandis que le groupe 2 se distingue par une quasi-absence de pluie durant cette saison. La différenciation entre les groupes 1 et 5 reste cependant à approfondir.

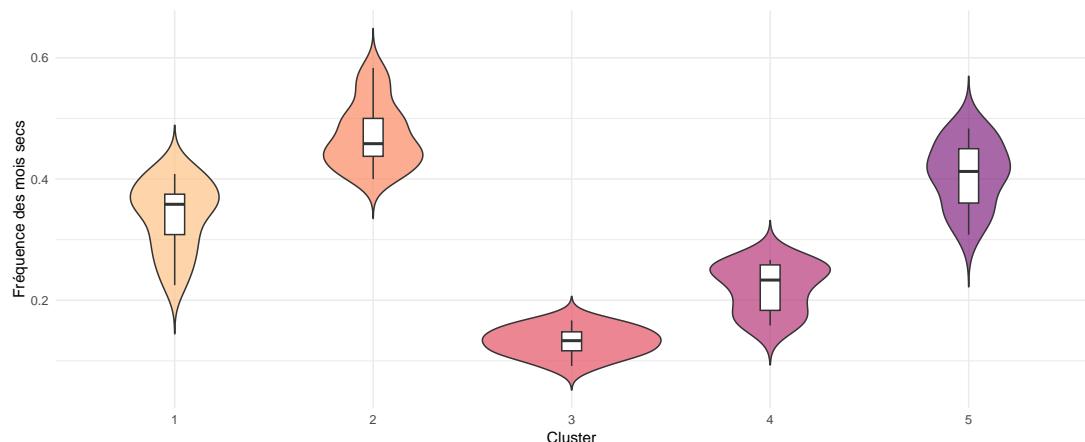


FIGURE 3.19 – Distribution de la fréquence des mois secs dans chaque groupe

La fréquence des mois secs (figure 3.19), calculée en lien avec la température, suggère un climat légèrement plus doux dans la zone 1 comparée à la zone 5, bien que ces deux zones restent globalement sèches. Le groupe 2 connaît quant à lui un nombre élevé de mois secs, tandis que le groupe 3 enregistre en moyenne seulement un mois et demi de sécheresse, et le groupe 4 présente également peu de mois secs. Les distributions des autres variables se trouvent en annexe (sous-section I.2).

Ainsi, pour tirer pleinement parti de cette typologie agroclimatique et en permettre une application territoriale, il convient de passer à une étape d'interpolation spatiale. Celle-ci fera l'objet du chapitre suivant.

Chapitre 4

Interpolations

Après avoir réalisé une classification des stations météorologiques en unités agroclimatiques , il est crucial d'aller au-delà des points discrets pour construire une représentation continue et cohérente de ces classes sur l'ensemble du territoire. En effet, les classifications ponctuelles, bien que précises au niveau des stations, ne rendent pas compte de la variabilité spatiale fine et des transitions progressives qui existent naturellement dans les phénomènes agroclimatiques. L'interpolation spatiale joue ici un rôle fondamental : elle permet de transformer des données ponctuelles en surfaces continues, offrant ainsi une cartographie fluide et réaliste des unités agroclimatiques.

Le principal enjeu de cette étape est de respecter à la fois l'hétérogénéité intrinsèque des données et la dépendance spatiale, c'est-à-dire la similarité attendue entre observations proches. Pour ce faire, plusieurs méthodes sont envisageables, allant de techniques simples et déterministes à des approches plus sophistiquées et probabilistes.

L'Interpolation Inverse Distance Weighting (IDW - pondération inverse à la distance) [20] est une méthode très répandue pour sa simplicité et sa rapidité. Elle repose sur l'hypothèse intuitive que l'influence d'un point d'observation décroît avec la distance. Elle attribue ainsi à chaque localisation un poids inversement proportionnel à sa distance aux stations observées. Bien que cette approche soit efficace pour générer des surfaces continues, elle peut parfois produire des transitions abruptes et ne tient pas compte de la structure statistique des données.

Les méthodes géostatistiques, notamment le krigage, apportent une dimension supplémentaire en modélisant explicitement la variabilité spatiale via le semi-variogramme, qui quantifie la corrélation entre valeurs en fonction de leur séparation géographique. Le krigage ordinaire suppose une moyenne constante inconnue, tandis que le krigage universel intègre une tendance spatiale générale, permettant de mieux capter les variations globales. Le krigage par classes permet quant à lui de prendre en compte des structures hétérogènes en segmentant l'espace selon des critères spécifiques. Ces méthodes permettent non seulement d'obtenir des prédictions optimales au sens de la variance minimale, mais aussi d'estimer les incertitudes associées, élément clé pour interpréter la fiabilité des cartes produites.

Par ailleurs, l'émergence des méthodes d'apprentissage automatique, telles que la forêt aléatoire spatiale, propose une alternative intéressante. Ces techniques sont capables de modéliser des relations non linéaires complexes et d'intégrer un large éventail de variables explicatives, tout en prenant en considération les corrélations spatiales. Elles sont particulièrement adaptées aux grands jeux de données hétérogènes et offrent un compromis entre précision et flexibilité.

Quelle que soit la méthode utilisée, la validation des modèles d'interpolation reste un point crucial. Les méthodes de validation croisée permettent d'évaluer la qualité des estimations en comparant les prédictions aux observations réelles laissées hors du processus d'apprentissage. Ces procédures garantissent que les surfaces interpolées reflètent fidèlement la réalité et permettent d'orienter le choix de la méthode la plus adaptée en fonction des objectifs spécifiques de l'étude.

En résumé, cette section présente les principales méthodes d'interpolation utilisées pour créer des unités agroclimatiques continues. Nous expliquerons comment fonctionnent ces méthodes, quels paramètres sont importants, et comment interpréter les résultats obtenus.

4.1 Interpolation par IDW (Inverse Distance Weighting)

L'interpolation IDW par classes est une méthode simple et intuitive pour estimer la distribution spatiale des classes à partir d'un jeu de points classés. C'est la technique utilisée par les géomaticiens à l'hôtel du département pour interpoler les données de température et de pluviométrie. Elle consiste à appliquer l'IDW de manière indépendante à chaque classe, en considérant la présence ou l'absence de la classe comme une variable binaire.

Pour une classe donnée C_j , on définit la variable indicatrice suivante :

$$I_{C_j}(x_i) = \begin{cases} 1 & \text{si la station } x_i \text{ appartient à la classe } C_j, \\ 0 & \text{sinon.} \end{cases}$$

L'estimation IDW de la probabilité d'appartenance à la classe C_j en un point x_0 s'écrit alors :

$$\hat{P}_{C_j}(x_0) = \frac{\sum_{i=1}^N \frac{I_{C_j}(x_i)}{d(x_0, x_i)^p}}{\sum_{i=1}^N \frac{1}{d(x_0, x_i)^p}},$$

où $d(x_0, x_i)$ désigne la distance entre x_0 et la station x_i , et p est le paramètre de puissance contrôlant la décroissance de l'influence avec la distance.

La classe assignée au point x_0 correspond à celle dont la probabilité est maximale :

$$\hat{C}(x_0) = \arg \max_j \hat{P}_{C_j}(x_0).$$

Cette approche produit des cartes continues reflétant la probabilité d'appartenance à chaque classe, tout en restant simple et rapide à mettre en œuvre. Elle est particulièrement adaptée pour obtenir une représentation fluide des classes issues d'une classification ponctuelle.

Cependant, l'IDW ne prend pas en compte la structure spatiale complexe des données, notamment les relations géomorphologiques ou environnementales, ce qui limite sa pertinence dans des contextes hétérogènes.

La figure 4.1 présente les erreurs issues de la validation croisée Leave-One-Out (LOO), calculées sur l'ensemble des stations. Ces résultats ont guidé le choix du paramètre $p = 7$, en minimisant le taux d'erreur global.

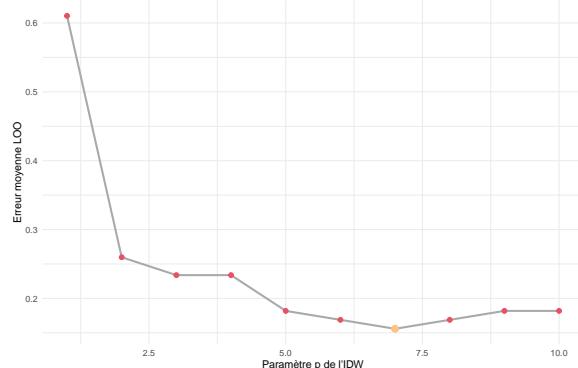


FIGURE 4.1 – Erreurs de validation croisée Leave-One-Out (LOO) pour différentes stations.

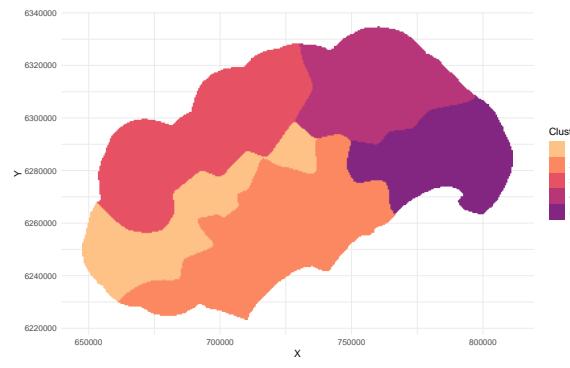


FIGURE 4.2 – Carte finale des classes interpolées par IDW avec $p = 7$.

La carte finale des classes interpolées, obtenue avec ce paramètre, est illustrée en figure 4.2. L'aspect général évoque une partition de type Voronoï aux contours lissés. Cependant, l'absence de prise en compte des variables géomorphologiques ou environnementales limite la pertinence de cette approche pour une classification agroclimatique fine.

Face à ces limites, nous explorons à présent des méthodes plus robustes, capables d'exploiter pleinement la structure spatiale des données et les covariables disponibles : le krigage constitue une première étape dans cette direction.

4.2 Interpolation par krigeage

Le krigeage [15] est une méthode d'interpolation géostatistique qui permet d'estimer une variable spatiale en s'appuyant sur la structure de dépendance spatiale observée entre les données. Cette approche repose sur la modélisation du variogramme, qui quantifie la variance des différences entre observations en fonction de la distance.

Le variogramme expérimental $\gamma(h)$ est défini par :

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2,$$

où $Z(x_i)$ et $Z(x_i + h)$ sont les valeurs observées en des points séparés par le vecteur de distance spatiale h , et $N(h)$ est le nombre de paires de points à cette distance.

Dans un contexte agroclimatique, la variable d'intérêt (par exemple une mesure climatique ou une classe agroclimatique) peut être influencée par différents facteurs environnementaux tels que le modèle numérique de terrain (MNT), la distance à la mer, l'indice de vallée, la pente ou encore l'orientation. Ces covariables, disponibles sur une grille spatiale, peuvent être intégrées dans le cadre du krigeage universel afin de modéliser explicitement la tendance spatiale.

Le krigeage universel combine ainsi un modèle déterministe basé sur ces covariables avec un terme aléatoire spatialement corrélé. L'estimation en un point x_0 s'écrit alors comme :

$$\hat{Z}(x_0) = \sum_{k=0}^K \beta_k f_k(x_0) + \sum_{i=1}^N \lambda_i \left[Z(x_i) - \sum_{k=0}^K \beta_k f_k(x_i) \right],$$

où :

- $f_k(x)$ sont les fonctions représentant les covariables (par exemple, altitude, distance à la mer, pente...),
- β_k sont les coefficients du modèle déterministe,
- λ_i sont les poids du krigeage déterminés par la structure spatiale du résidu,
- N est le nombre d'observations.

Cette approche permet d'améliorer la précision des estimations en tenant compte à la fois des tendances liées aux facteurs environnementaux et des variations locales non expliquées, modélisées par la structure spatiale résiduelle.

Le krigeage, notamment dans sa version universelle, constitue donc un outil puissant pour produire des cartes agroclimatiques continues, intégrant à la fois les données climatiques ponctuelles et les variables environnementales spatialisées, essentielles à la compréhension et à la représentation fine des unités agroclimatiques.

4.2.1 Krigeage ordinaire

Pour interpoler des données catégorielles telles que des classes agroclimatiques, nous utilisons le krigeage ordinaire par classes. Cette méthode consiste à modéliser séparément chaque classe C_j à l'aide de variables indicatrices binaires définies par :

$$I_{C_j}(x_i) = \begin{cases} 1 & \text{si } x_i \text{ appartient à } C_j, \\ 0 & \text{sinon.} \end{cases}$$

Chaque variable indicatrice est interpolée indépendamment par krigeage ordinaire, en supposant une moyenne locale constante mais inconnue et une structure spatiale des résidus caractérisée par un variogramme adapté.

L'estimation de la probabilité d'appartenance à la classe C_j en un point non observé x_0 s'écrit alors :

$$\hat{P}_{C_j}(x_0) = \sum_{i=1}^N \lambda_i^{(j)} I_{C_j}(x_i),$$

avec la contrainte :

$$\sum_{i=1}^N \lambda_i^{(j)} = 1.$$

Les poids $\lambda_i^{(j)}$ sont obtenus en résolvant le système linéaire du krigage ordinaire, basé sur le variogramme, qui minimise la variance de l'erreur d'estimation tout en assurant une estimation non biaisée.

La classification finale est obtenue en attribuant au point x_0 la classe correspondant à la probabilité maximale :

$$\hat{C}(x_0) = \arg \max_j \hat{P}_{C_j}(x_0).$$

Cette approche spatiale pure, sans intégration de covariables, permet de produire une carte continue des classes à partir des données observées, mais peut conduire à une classification fragmentée en l'absence d'une forte structure spatiale.

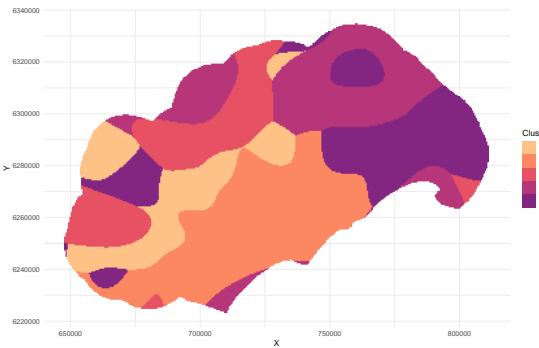


FIGURE 4.3 – Carte des classes agroclimatiques interpolées par krigage ordinaire

La carte interpolée par krigage ordinaire (figure 4.3) montre une classification relativement fragmentée, avec une continuité spatiale limitée. Cela reflète une structure spatiale peu marquée des variables indicatrices, rendant l'ajustement des variogrammes délicat.

Pour améliorer la modélisation spatiale, nous intégrerons ensuite des variables géomorphologiques explicatives via le krigage universel. Cette approche permettra d'introduire une tendance déterministe liée aux covariables, offrant une meilleure continuité spatiale et une meilleure représentation des processus environnementaux.

4.2.2 Krigage universel

Pour dépasser les limites observées du krigage ordinaire, nous avons recours au krigage universel par classes, qui intègre une composante déterministe fondée sur des covariables environnementales. Cette approche est particulièrement adaptée lorsque la probabilité d'appartenance à une classe varie en fonction de facteurs explicatifs spatialisés tels que le modèle numérique de terrain, la distance à la mer, la pente ou l'orientation.

Pour chaque classe C_j , la variable indicatrice est décomposée en une tendance déterministe $m_j(x)$ liée aux covariables, et un résidu spatialement corrélé $\varepsilon_j(x)$:

$$I_{C_j}(x) = m_j(x) + \varepsilon_j(x),$$

avec

$$m_j(x) = \sum_{k=0}^K \beta_{jk} f_k(x),$$

où les fonctions $f_k(x)$ représentent les covariables spatialisées et β_{jk} les coefficients associés à la classe C_j .

Les coefficients β_{jk} du modèle déterministe sont estimés préalablement par une régression linéaire des variables indicatrices $I_{C_j}(x_i)$ sur les covariables $f_k(x_i)$. Cette étape permet de modéliser la tendance spatiale liée aux facteurs environnementaux, séparément de la composante aléatoire. Formulée de manière matricielle, cette régression s'écrit :

$$I_{C_j} = F\beta_j + \varepsilon_j,$$

où $I_{C_j} = (I_{C_j}(x_1), \dots, I_{C_j}(x_N))^T$ est le vecteur des observations pour la classe C_j , F est la matrice des covariables $(f_k(x_i))$, $\beta_j = (\beta_{j0}, \dots, \beta_{jK})^T$ le vecteur des coefficients, et ε_j les résidus.

Les coefficients $\hat{\beta}_j$ sont alors estimés par moindres carrés ordinaires :

$$\hat{\beta}_j = (F^T F)^{-1} F^T I_{C_j}.$$

Les résidus correspondants

$$r_j(x_i) = I_{C_j}(x_i) - \sum_{k=0}^K \hat{\beta}_{jk} f_k(x_i)$$

sont ensuite utilisés pour modéliser la structure spatiale par un variogramme, ce qui permet d'estimer les poids $\lambda_i^{(j)}$ du krigeage universel.

L'estimation par krigeage universel de la probabilité d'appartenance à la classe C_j en un point x_0 s'écrit alors :

$$\hat{P}_{C_j}(x_0) = \sum_{k=0}^K \hat{\beta}_{jk} f_k(x_0) + \sum_{i=1}^N \lambda_i^{(j)} \left[I_{C_j}(x_i) - \sum_{k=0}^K \hat{\beta}_{jk} f_k(x_i) \right].$$

La classification finale est obtenue en attribuant la classe C_j pour laquelle la probabilité estimée $\hat{P}_{C_j}(x_0)$ est maximale :

$$\hat{C}(x_0) = \arg \max_j \hat{P}_{C_j}(x_0).$$

Cette méthode permet d'exploiter explicitement les liens entre la distribution spatiale des classes et les facteurs environnementaux, tout en prenant en compte les variations locales non expliquées par ces covariables. Le krigeage universel par classes offre ainsi une interpolation plus précise et plus cohérente des unités agroclimatiques dans un contexte spatial complexe.

Cependant, la qualité des résultats dépend de la modélisation des variogrammes des résidus et de l'estimation fiable des coefficients de tendance pour chaque classe, ce qui requiert des données explicatives de bonne qualité.

Voici les deux premiers variogrammes obtenus :

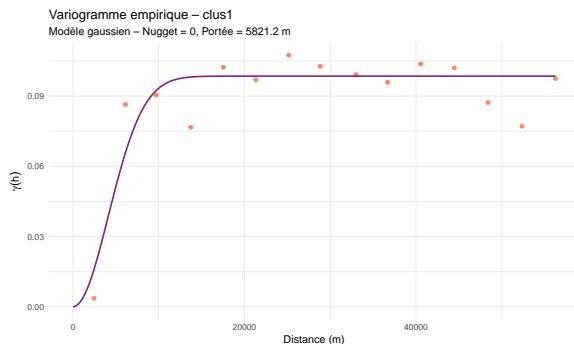


FIGURE 4.4 – Variogramme des résidus pour la classe 1

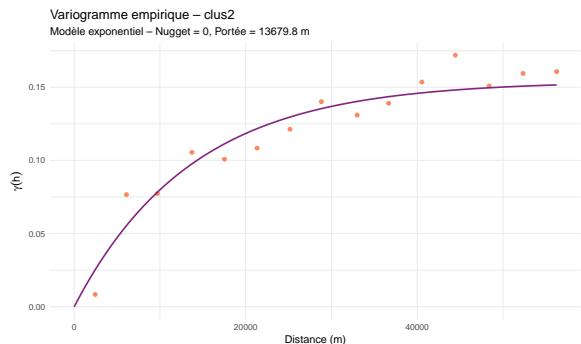


FIGURE 4.5 – Variogramme des résidus pour la classe 2

Les figures 4.4 et 4.5 présentent les variogrammes expérimentaux des résidus pour les classes 1 et 2, ainsi que les modèles variogrammes ajustés. Les trois autres variogrammes ainsi qu'une observation complémentaire sur les clusters sont disponible en annexe (sous-section II.1).

Les points sur ces graphes représentent les semivariances expérimentales, calculées à partir des différences entre les valeurs des résidus observées aux stations de mesure pour différentes distances h . Concrètement, pour chaque intervalle de distance, la semivariance est la moyenne des demi-carrés des écarts entre les paires de stations situées à cette distance. Ces semivariances expriment la similarité spatiale : des valeurs proches aux stations voisines impliquent de faibles semivariances pour les courtes distances, tandis que les semivariances augmentent avec la distance lorsque la dépendance spatiale diminue.

La courbe continue correspond au modèle variogramme ajusté, qui décrit cette structure spatiale sous une forme mathématique paramétrée. Les paramètres des variogrammes ajustés sont les suivants :

- Nugget (c_0) : représente la variabilité à très courte distance, souvent due à des erreurs de mesure ou à une variabilité microscopique non spatialement structurée. Un nugget nul indique une absence de cette variabilité non spatialisée.
- Sill (c) : correspond à la valeur maximale de la semivariance, qui reflète la variance totale expliquée par la structure spatiale. Au-delà de la portée, la semivariance se stabilise à cette valeur.
- Portée (a) : distance maximale à laquelle les données sont spatialement corrélées. Au-delà de cette distance, la semivariance atteint la sill et les points ne présentent plus de dépendance spatiale.

Nous sommes alors confrontés à deux types de variogrammes :

Classe 1 (figure 4.4)

Modèle gaussien avec nugget nul et portée 5,8 km :

$$\gamma(h) = c_0 + c \left[1 - \exp \left(- \left(\frac{h}{a} \right)^2 \right) \right],$$

où c_0 est le nugget (ici nul), $c = 0.099$ la sill, et $a = 5,8$ km la portée.

Classe 2 (figure 4.5)

Modèle exponentiel avec nugget nul et portée 13,6 km :

$$\gamma(h) = c_0 + c \left[1 - \exp \left(- \frac{h}{a} \right) \right],$$

où c_0 est le nugget (ici nul), $c = 0.154$ la sill, et $a = 13,6$ km la portée.

Ces différences de portée indiquent que la dépendance spatiale des résidus est plus locale pour la classe 1, tandis qu'elle s'étend sur une plus grande distance pour la classe 2. L'absence d'effet nugget suggère que la variabilité non spatialisée (bruit ou erreurs de mesure) est faible dans les deux cas.

Ces modèles variogrammes sont essentiels pour paramétriser le krigage universel, car ils quantifient la corrélation spatiale des résidus après prise en compte des covariables géomorphologiques, permettant ainsi une meilleure interpolation spatiale des classes.

Nous pouvons maintenant nous intéresser à l'importance des variables. Pour construire le modèle de prédiction, on utilise la formule suivante :

```
clus_i ~ poly(mnt, 2) +
         poly(slope, 2) +
         poly(aspect, 2) +
         poly(distmer, 2) +
         poly(mrvbf, 2)
```

Chaque variable environnementale est modélisée à l'aide d'un polynôme d'ordre 2 pour capturer d'éventuels effets non linéaires. L'ajustement de ce modèle pour chaque classe nous permet d'identifier quelles covariables influencent le plus la probabilité d'appartenance à une classe donnée.

Les tableaux suivants présentent les coefficients estimés pour les covariables les plus significatives dans les classes 1 et 2, accompagnés de leurs intervalles de confiance (IC à 95 %). Les coefficients ont été obtenus après standardisation implicite via la base orthogonale générée par `poly()`.

Variable	Coef.	Erreur	p-valeur
(Intercept)	0.169	0.039	< 0.001
mnt (lin)	-2.343	1.154	0.046
distmer (lin)	1.711	1.069	0.114
mrvbf (lin)	-1.073	0.717	0.140

TABLE 4.1 – Effet des variables principales sur la classe 1

Variable	Coef.	Erreur	p-valeur
(Intercept)	0.351	0.045	< 0.001
distmer (lin)	-2.911	1.231	0.021
mrvbf (lin)	-1.708	0.825	0.042
mrvbf (quad)	-2.226	0.573	< 0.001

TABLE 4.2 – Effet des variables principales sur la classe 2

On remarque dans les tableau 4.1 et tableau 4.2 que la classe 1 est particulièrement influencée par l'altitude (`mnt`), dont le premier terme du polynôme présente un effet négatif significatif. La classe 2, quant à elle, est fortement marquée par la distance à la mer (`distmer`) et la morphologie du terrain (`mrvbf`), toutes deux avec des coefficients significativement négatifs. Cela indique que cette classe est associée à une zone plutôt proche de la mer avec un indice de vallée très faible.

Ces résultats confirment que l'appartenance à une classe ne dépend pas uniquement d'un gradient géographique brut, mais bien d'un ensemble de facteurs géomorphologiques que le modèle prend en compte. Ces covariables renforcent la capacité du modèle à structurer l'espace agroclimatique de manière plus cohérente et informative.

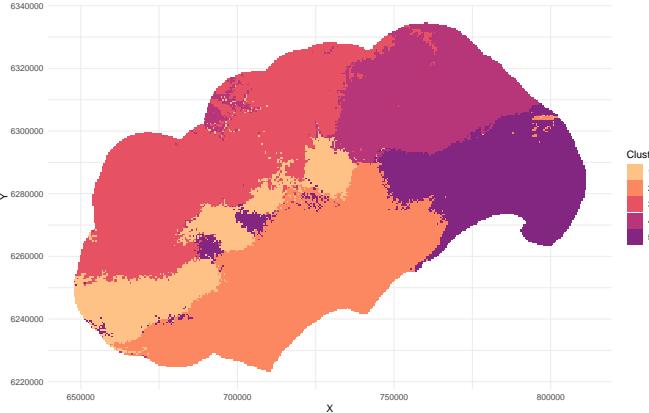


FIGURE 4.6 – Carte des classes agroclimatiques interpolées par krigeage universel

La carte issue du krigeage universel (figure 4.6) montre une nette amélioration par rapport au krigeage ordinaire, avec une distribution spatiale plus lisse et des classes plus cohérentes. L'ajout des covariables explicatives (comme l'altitude ou la distance à la mer) permet de mieux capturer les grandes tendances spatiales. On observe cependant quelques artefacts au sein même du cluster 1, notamment des petites zones isolées appartenant à d'autres classes. Ces discontinuités peuvent résulter d'une modélisation imparfaite des résidus, ou d'une influence locale trop forte de certaines covariables. Cela souligne les limites du krigeage universel lorsqu'il s'appuie sur des relations linéaires ou quadratiques avec les variables explicatives, et invite à explorer des approches plus souples pour capturer la complexité des interactions environnementales.

Dans cette optique, nous proposons dans la section suivante une méthode d'interpolation fondée sur les forêts aléatoires, qui permet de modéliser des relations non linéaires et potentiellement plus adaptées à la distribution complexe des classes agroclimatiques.

4.3 Modélisation par forêts aléatoires

Afin de cartographier la distribution spatiale des groupes agroclimatiques identifiés précédemment, nous pouvons aussi avoir recours à une approche de classification supervisée basée sur les forêts aléatoires (*random forests*). Cette méthode [5] repose sur l'agrégation d'arbres de décision construits sur des sous-échantillons aléatoires de données. Elle permet de modéliser des relations non linéaires complexes entre des variables explicatives (environnementales et spatiales) et une variable cible catégorielle.

Cadre mathématique

Soit un ensemble d'apprentissage $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, où $\mathbf{x}_i \in \mathbb{R}^p$ est le vecteur des covariables pour la station i (comprenant notamment l'altitude, la pente, l'exposition, la distance à la mer, l'indice MRVBF), auquel on ajoute les coordonnées spatiales (x_i, y_i) . La variable cible $y_i \in \{1, \dots, 5\}$ correspond au groupe agroclimatique auquel appartient la station i .

Les covariables environnementales utilisées ici sont uniquement les variables géomorphologiques, car elles sont disponibles pour l'ensemble des cellules de la grille spatiale. Les variables agroclimatiques, issues des séries météorologiques, ne sont en revanche disponibles qu'en station et ne peuvent donc être utilisées directement pour l'interpolation.

Mathématiquement, l'objectif est d'approximer une fonction de classification f telle que :

$$y_i = f(\mathbf{z}_i, \mathbf{x}_i, y_i) + \varepsilon_i,$$

où \mathbf{z}_i désigne le vecteur des covariables environnementales, (x_i, y_i) les coordonnées spatiales, et ε_i un terme d'erreur aléatoire.

L'estimateur final de la forêt aléatoire est défini comme une moyenne (régression) ou un vote majoritaire (classification) sur un ensemble d'arbres $\hat{f}_1, \dots, \hat{f}_B$ construits indépendamment sur des échantillons bootstrap :

$$\hat{f}_{RF}(\mathbf{x}) = \arg \max_{k \in \{1, \dots, 5\}} \frac{1}{B} \sum_{b=1}^B \mathbb{1}_{\hat{f}_b(\mathbf{x})=k}.$$

Chaque arbre partitionne l'espace \mathbb{R}^p en régions homogènes, selon un critère d'impureté (indice de Gini) minimisé à chaque noeud.

Algorithme de la forêt aléatoire

L'algorithme de construction d'une forêt aléatoire pour la classification des stations se décrit de la manière suivante :

Algorithm 2 : Algorithme de la forêt aléatoire pour la classification spatiale

Entrée : Ensemble d'apprentissage $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, nombre d'arbres B , nombre de variables testées m_{try}

Sortie : Prédicteur \hat{f}_{RF}

for $b = 1$ to B **do**

Tirer un échantillon bootstrap $\mathcal{D}_n^{(b)}$ de taille n à partir de \mathcal{D}_n ;

Construire un arbre de décision \hat{f}_b :

- À chaque noeud, sélectionner aléatoirement m_{try} variables parmi les p disponibles ;
- Choisir la variable et le seuil de partition qui minimisent l'impureté ;
- Répéter récursivement jusqu'à atteindre une condition d'arrêt (pureté, profondeur maximale, taille minimale des feuilles, etc.).

end

return $\hat{f}_{RF}(\mathbf{x}) = \arg \max_k \frac{1}{B} \sum_{b=1}^B \mathbb{1}_{\hat{f}_b(\mathbf{x})=k}$

Utilisation spatiale

Dans notre application, l'apprentissage est réalisé sur les stations météorologiques pour lesquelles on connaît les classes agroclimatiques (`clus`) ainsi que les valeurs des covariables géomorphologiques. Pour chaque station, on extrait :

- L'altitude (`mnt`),
- La pente (`slope`),
- L'exposition (`aspect`),
- La distance à la mer (`distmer_lowres`),
- L'indice MRVBF (`mrvbf`),
- Les coordonnées spatiales (x, y).

Une fois le modèle ajusté, il est appliqué à la grille régulière couvrant le territoire. Chaque cellule de cette grille possède les mêmes covariables géomorphologiques que les stations, ce qui permet de prédire la classe agroclimatique par vote majoritaire des arbres.

Pour le modèle de forêt aléatoire, nous utilisons la formule suivante (plus simple que pour le krigage) :

```
clus ~ mnt + slope + aspect + distmer_lowres + mrvbf + coord_X + coord_Y
```

Chaque observation (station) est ainsi décrite par ses caractéristiques environnementales et sa position, permettant d'intégrer les structures spatiales directement dans l'apprentissage. Une fois le modèle ajusté, il est appliqué à la grille régulière couvrant le territoire. Chaque cellule de cette grille possède les mêmes covariables géomorphologiques que les stations, ce qui permet de prédire la classe agroclimatique par vote majoritaire des arbres.

4.3.1 Prédition d'une variable catégorielle

Dans un premier temps, nous avons choisi de prédire directement la variable catégorielle `clus`, correspondant aux groupes agroclimatiques issus de la classification précédente. Il s'agit ici d'un problème de classification supervisée à cinq classes.

Le modèle de forêt aléatoire est entraîné sur les données en station, en utilisant uniquement les covariables géomorphologiques disponibles pour toutes les cellules de la grille, ainsi que les coordonnées spatiales. La prédiction est ensuite effectuée sur la grille complète, nous servant à produire une carte continue des classes agroclimatiques estimées.

Un arbre de décision indépendant a été ajusté à l'aide de l'algorithme CART (rpart) sur les mêmes données que celles utilisées pour la forêt aléatoire, dans le but d'illustrer de manière plus lisible les règles de classification issues des covariables environnementales. Bien que cet arbre ne fasse pas partie des arbres constituant la forêt aléatoire, il fournit une approximation interprétable de la manière dont les variables explicatives influencent l'appartenance aux différentes classes agroclimatiques.

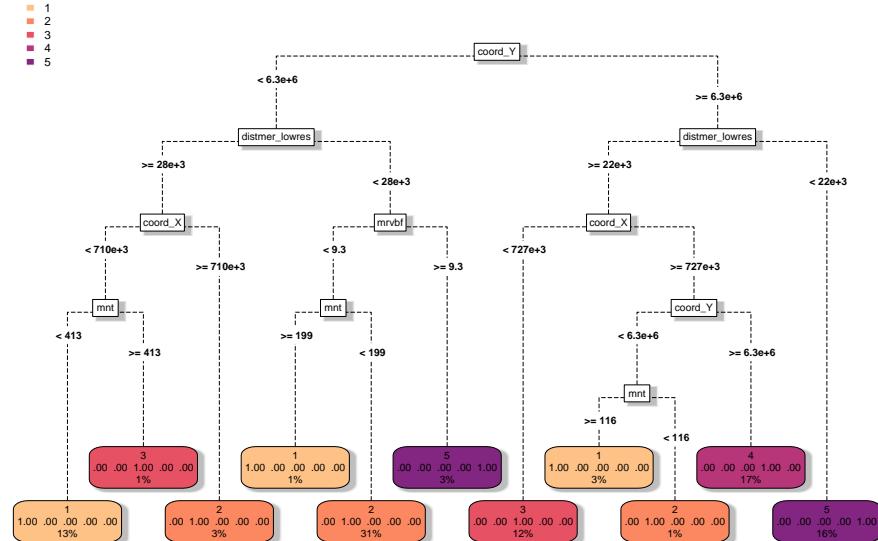


FIGURE 4.7 – Arbre de décision ajusté sur les données d'apprentissage pour la classification directe (variable `clus`)

L'arbre représenté en figure 4.7 met en évidence les principales variables de segmentation utilisées par le modèle. On observe que certaines covariables géomorphologiques, par exemple la distance à la mer (`distmer_lowres`) ou l'indice de vallée (`mrvbf`), interviennent précocement dans la structure de l'arbre, ce qui suggère leur importance dans la différenciation des classes. Les feuilles de l'arbre correspondent aux prédictions de classes, et les seuils sur les covariables déterminent les règles de passage d'un nœud à l'autre.

Cette visualisation permet donc de mieux comprendre les relations entre les caractéristiques spatiales du terrain et la classification agroclimatique, en complément des résultats plus complexes et moins interprétables issus de la forêt aléatoire.

La carte de prédiction obtenue par vote majoritaire est présentée ci-dessous. Chaque cellule de la grille est associée à la classe prédite par la majorité des arbres.

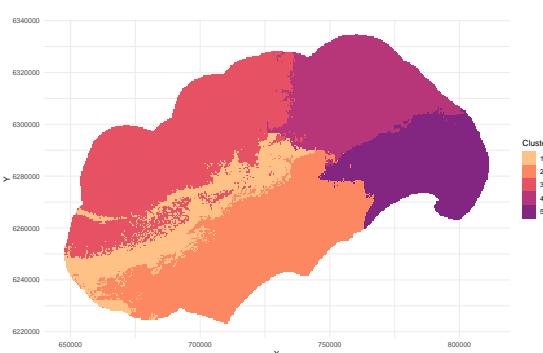


FIGURE 4.8 – Carte de prédiction des classes agroclimatiques par vote majoritaire

La carte de la figure 4.8 illustre la prédiction des classes agroclimatiques obtenue par forêt aléatoire selon un vote majoritaire sur chaque pixel. Comparée aux cartes issues du krigage ordinaire (figure 4.3) ou même universel (figure 4.6), cette approche produit une partition spatiale nettement plus homogène, avec des zones bien définies et une cohérence spatiale accrue. Cela s'explique par le fait que les forêts aléatoires s'appuient directement sur les covariables environnementales structurantes, et non uniquement sur la dépendance spatiale des observations. Toutefois, cette carte ne reflète pas l'incertitude du modèle, ni les zones de transition, qui peuvent être importantes dans le contexte agroclimatique. Ces aspects seront explorés par la suite au travers d'une modélisation probabiliste des classes.

4.3.2 Prédition par classe

Dans un second temps, nous affinons l'analyse en procédant à une prédition binaire pour chaque classe de manière indépendante. Pour chaque classe $k \in \{1, \dots, 5\}$, un modèle de forêt aléatoire est ajusté afin d'estimer la probabilité conditionnelle $P(y = k | \mathbf{x})$. L'intérêt de cette approche est double :

- Elle permet d'obtenir une carte continue des probabilités d'appartenance à chaque classe, révélant les zones de transition ou d'ambiguïté.
- Elle offre une base pour une classification « par maximum de probabilité », en associant à chaque cellule la classe ayant la plus forte probabilité estimée.

Les cinq cartes de probabilités sont représentées ci-dessous.

À partir de ces cinq cartes, il est ensuite possible de reconstruire une carte de classification en attribuant à chaque cellule de la grille la classe k pour laquelle la probabilité est maximale :

$$\hat{y}(\mathbf{x}) = \arg \max_{k \in \{1, \dots, 5\}} P(y = k | \mathbf{x}).$$

Afin de mieux comprendre la structure des règles apprises par la forêt aléatoire, nous pouvons extraire et visualiser un arbre de décision représentatif pour chacune des classes 1 et 2. Ces arbres simplifient la complexité du modèle en mettant en évidence les covariables et seuils clés utilisés pour segmenter l'espace d'entrée.

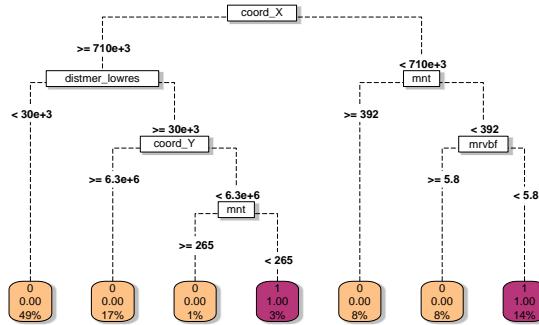


FIGURE 4.9 – Arbre de décision extrait pour la classe 1

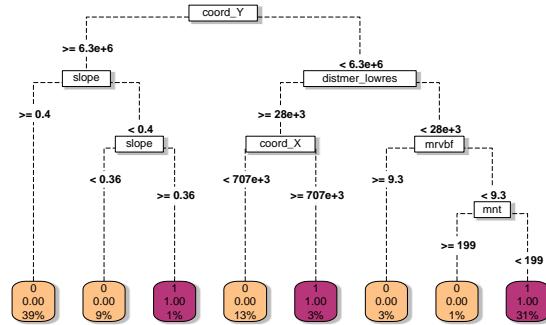


FIGURE 4.10 – Arbre de décision extrait pour la classe 2

Pour les deux arbres analysés, la première division s'effectue selon une coupe géographique, distinguant par exemple les zones est/ouest ou nord/sud.

Dans le cas de la classe 1 (figure 4.9), la classification privilégie les cellules situées en basse altitude et localisées dans des fonds de vallée, caractérisés par une faible valeur de la variable MRVBF (modèle numérique de relief). Pour la classe 2 (figure 4.10), les critères retenus mettent en avant les cellules proches du littoral, également en basse altitude, avec des valeurs faibles de MRVBF.

Les arbres des autres classes ainsi que les cartes pour les deux autres périodes étudiées sont disponibles en annexe (sous-section II.2).

La forêt aléatoire permet également d'estimer la probabilité d'appartenance à chaque classe pour chaque pixel, fournissant ainsi une information sur la vraisemblance des prédictions. Les cartes suivantes présentent, pour chacune des cinq classes, la distribution spatiale des probabilités prédites :

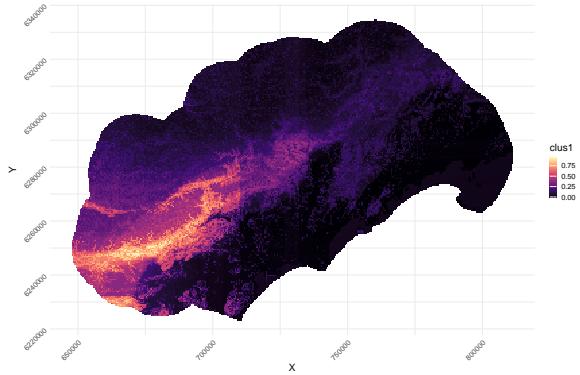


FIGURE 4.11 – Probabilité d'appartenance à la classe 1

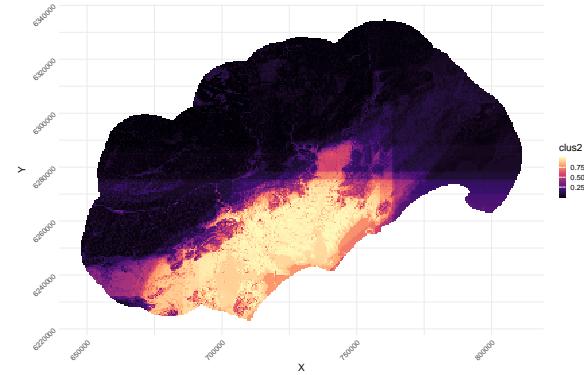


FIGURE 4.12 – Probabilité d'appartenance à la classe 2

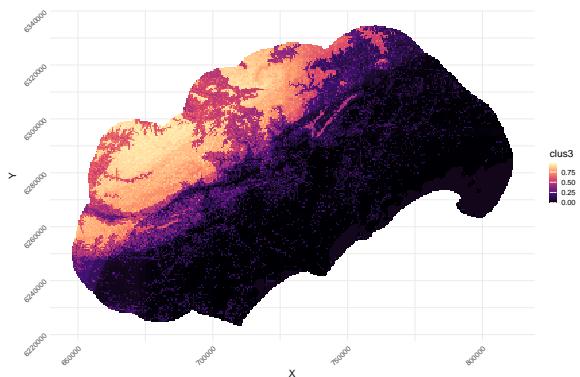


FIGURE 4.13 – Probabilité d'appartenance à la classe 3

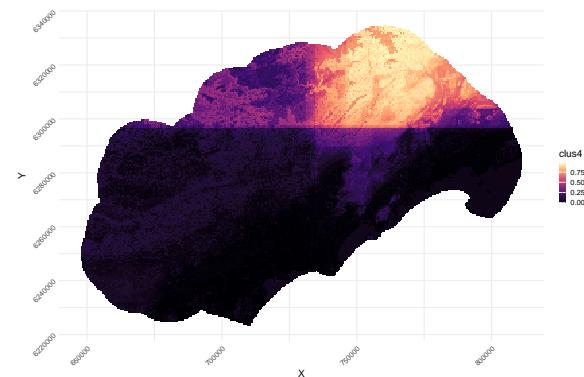


FIGURE 4.14 – Probabilité d'appartenance à la classe 4

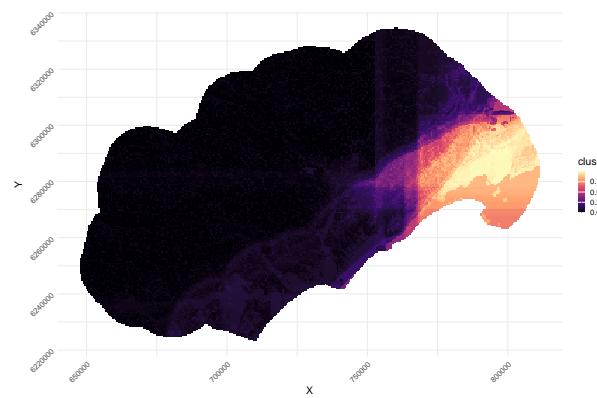


FIGURE 4.15 – Probabilité d'appartenance à la classe 5

Ces cartes mettent en évidence les zones de transition entre classes, souvent marquées par des probabilités intermédiaires (autour de 0.5), ainsi que les noyaux de classes où le modèle exprime une forte confiance (valeurs proches de 1). Elles offrent ainsi une lecture probabiliste bien plus nuancée que la carte issue du vote majoritaire (figure 4.8) et permettent de mieux cerner l'incertitude dans les zones frontières.

On peut par exemple observer que la carte de probabilité de la classe 1 (figure 4.11) suit très nettement les talwegs situés dans l'ouest du département, traduisant une forte association de cette classe avec les fonds de vallée. Par ailleurs, l'inclusion explicite des coordonnées spatiales dans le modèle se manifeste

par l'apparition de découpages nets, voire artificiels, dans les probabilités des classes 4 et 5 (figure 4.14 et figure 4.15), suggérant une structuration liée à la position géographique plutôt qu'aux covariables environnementales seules.

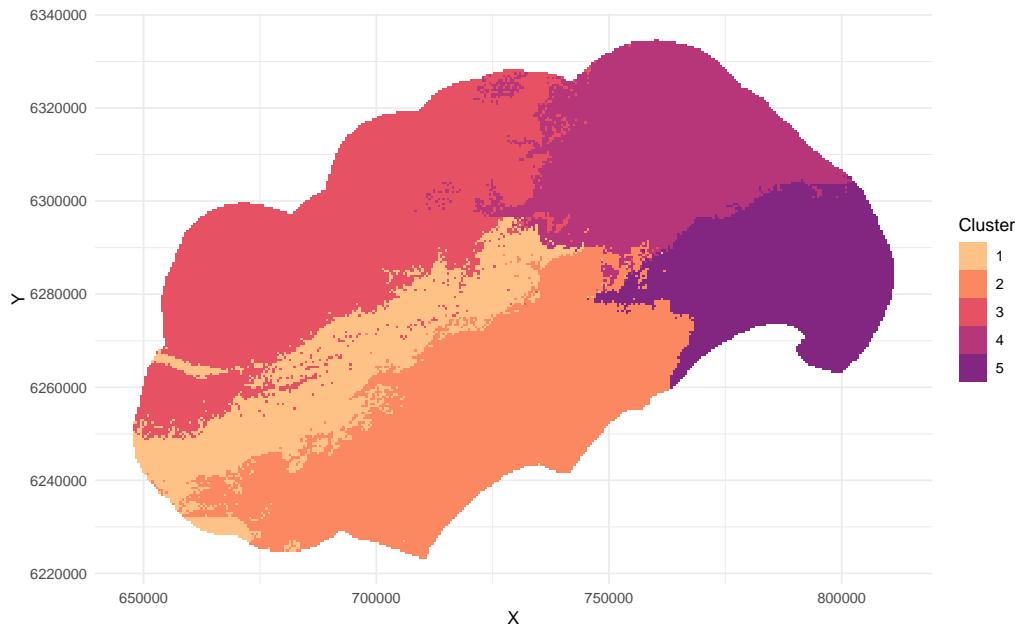


FIGURE 4.16 – Grille finale de classification par probabilité maximale

La figure 4.16 présente la carte finale obtenue à partir de l'affectation par probabilité maximale. Cette approche en deux temps (d'abord l'estimation des probabilités d'appartenance à chaque classe, puis la sélection de la classe la plus probable) permet de croiser les résultats avec la classification directe, d'évaluer la confiance du modèle dans ses prédictions, et d'apporter une lecture plus fine des dynamiques spatiales.

Par rapport à la carte obtenue par vote majoritaire, cette version attribue une surface légèrement plus large à la bande étroite de la classe 1. On note également que les transitions entre classes suivent davantage les structures du relief, ce qui témoigne d'une meilleure prise en compte des discontinuités environnementales par le modèle. Cette interpolation par probabilité maximise ainsi la cohérence spatiale tout en enrichissant l'analyse géographique.

La carte obtenue par cette approche probabiliste montre une meilleure adéquation avec les structures environnementales et une représentation plus réaliste des zones de transition. Pour confirmer ces observations qualitatives et comparer objectivement les performances des différentes méthodes, nous procédons désormais à une validation croisée. Cette analyse permet d'évaluer la robustesse et la précision des modèles sur des jeux de données indépendants.

La section suivante détaille ces résultats comparatifs.

4.4 Comparaison des modèles par validation croisée

Afin d'évaluer la robustesse et la performance prédictive des différents modèles de classification spatiale, nous procémons à plusieurs méthodes de validation croisée. Cette étape est essentielle, car les stations d'échantillonnage sont dispersées et le jeu de données est relativement restreint, ce qui peut influencer la qualité des prédictions.

Méthodes de validation croisée

Nous utilisons principalement deux types de validation croisée :

- **Validation croisée Leave-One-Out (LOO)** : chaque station est retirée à tour de rôle du jeu d'apprentissage, et sa classe est prédite à partir du modèle entraîné sur les autres points. Cette méthode est particulièrement adaptée à un faible nombre de points (ici 77 stations), car elle maximise l'usage des données disponibles pour l'entraînement.
- **Validation croisée par k-fold** : le jeu de données est divisé en k sous-ensembles (ici 5 et 10 plis ont été testés). À chaque itération, un sous-ensemble est utilisé pour la validation et les autres pour l'apprentissage. Cette méthode permet de contrôler la variance de l'estimation des performances et d'éviter le surapprentissage.

Dans les deux cas, l'erreur est calculée comme la proportion de points pour lesquels la classe prédite diffère de la classe observée, puis normalisée par le nombre total de points.

Ces méthodes de validation croisée ont été choisies car elles reflètent correctement les situations réelles rencontrées dans le suivi agroclimatique. En effet, une station peut devenir indisponible temporairement (cas simulé par la validation Leave-One-Out), mais il est aussi fréquent que plusieurs stations manquent simultanément, notamment selon la période ou les conditions d'observation (cas modélisé par la validation k-fold). Ainsi, selon la décennie étudiée, les stations ne sont pas du tout les mêmes.

Résultats et analyse

Les tableaux 4.3 et 4.4 synthétisent les taux d'erreur moyens et leurs variances pour chaque modèle selon les différentes procédures de validation.

Modèle	Taux erreur	Variance erreur
Kriging Ordinaire	0.0597	6.80e-04
Kriging Universel	0.0279	4.81e-04
Random Forest	0.0211	3.80e-05
Random Forest par classe	0.0173	2.20e-05

TABLE 4.3 – Taux d'erreur et variance obtenus par validation croisée Leave-One-Out (LOO) pour les différents modèles

Modèle	CV k-fold (5 plis)		CV k-fold (10 plis)	
	Taux erreur	Variance erreur	Taux erreur	Variance erreur
Kriging Ordinaire	0.2162	1.68e-04	0.1917	4.95e-03
Kriging Universel	0.1279	1.37e-03	0.0720	1.20e-03
Random Forest	0.0598	6.34e-04	0.0394	3.91e-04
Random Forest par classe	0.0709	8.21e-04	0.0388	5.16e-04

TABLE 4.4 – Taux d'erreur et variance obtenus par validation croisée k-fold (5 et 10 plis) pour les différents modèles

On observe que les méthodes basées sur les forêts aléatoires, en particulier le modèle par classe, présentent systématiquement les taux d'erreur les plus faibles, témoignant de leur capacité à capturer des relations complexes entre les covariables et la classification.

Les krigeages, quant à eux, montrent des taux d'erreur plus importants, surtout en validation k-fold, ce qui peut s'expliquer par une moindre flexibilité dans la modélisation des classes et une sensibilité aux choix des paramètres variogrammes. Le krigeage universel, intégrant des covariables, surpassé toutefois le krigeage ordinaire, confirmant l'intérêt d'utiliser les informations environnementales.

Enfin, la cohérence des résultats entre les différentes validations (LOO, 5 plis et 10 plis) valide la robustesse de ces conclusions. Le choix du type de validation croisée doit cependant rester adapté à la taille de l'échantillon et à la problématique, la LOO étant privilégiée ici pour maximiser les données d'apprentissage. Le modèle d'interpolation que nous utilisons pour notre protocole de définition des unités agroclimatiques est donc le *Random Forest par classe*.

4.5 Lissage de la grille d'interpolation

4.5.1 Transformation d'une grille en une couche de polygones

La grille résultant de l'interpolation par forêt aléatoire attribue à chaque cellule x_i une classe dominante $c_i \in \{1, \dots, K\}$. Afin de produire une représentation cartographique exploitable, cette grille est convertie en une couche vectorielle de polygones.

Dans un premier temps, la classification discrète est soumise à un filtrage spatial par fenêtre glissante. Ce filtre repose sur la statistique locale majoritaire dans un voisinage carré de taille $w \times w$, où w est impair. Pour chaque cellule centrale x_0 , on définit :

$$\hat{c}(x_0) = \arg \max_{k \in \{1, \dots, K\}} \sum_{x_j \in \mathcal{N}_w(x_0)} \mathbb{1}(c_j = k),$$

où $\mathcal{N}_w(x_0)$ désigne l'ensemble des cellules dans la fenêtre centrée en x_0 . Cette opération atténue les petites taches isolées et améliore la cohérence spatiale des zones.

Le raster ainsi nettoyé est ensuite converti en polygones vectoriels en agrégeant les cellules contiguës de même classe. Les polygones obtenus subissent une correction topologique pour garantir leur validité (notamment face aux artefacts liés à la grille), puis sont simplifiés géométriquement en réduisant le nombre de sommets tout en préservant la topologie.

4.5.2 Rendu final

Un lissage supplémentaire est ensuite appliqué pour adoucir les contours abrupts hérités de la structure raster. Ce lissage utilise une approximation fondée sur le lissage de type kernel, via la fonction `ksmooth`. À partir d'un ensemble de sommets (x_i, y_i) décrivant un polygone, une courbe lisse $\tilde{y}(x)$ est estimée à l'aide d'un noyau gaussien :

$$\tilde{y}(x) = \frac{\sum_{i=1}^n K_h(x - x_i)y_i}{\sum_{i=1}^n K_h(x - x_i)}, \quad \text{avec } K_h(u) = \exp\left(-\frac{u^2}{2h^2}\right),$$

où h contrôle le degré de lissage. Ce traitement est appliqué indépendamment aux coordonnées x et y , ce qui produit des contours visuellement plus naturels.

Enfin, pour éviter des artefacts tels que des petits trous entre polygones, les espaces résiduels sont identifiés (via la différence entre l'enveloppe spatiale globale et l'union des polygones), puis comblés si leur aire reste inférieure à un seuil donné (ici 1000 m^2). Ces zones sont redistribuées vers le polygone voisin le plus proche, déterminé à l'aide de la distance minimale entre centroïdes.

Le résultat final est une carte vectorielle lissée et topologiquement propre, prête à être intégrée dans les SIG pour analyse ou diffusion.

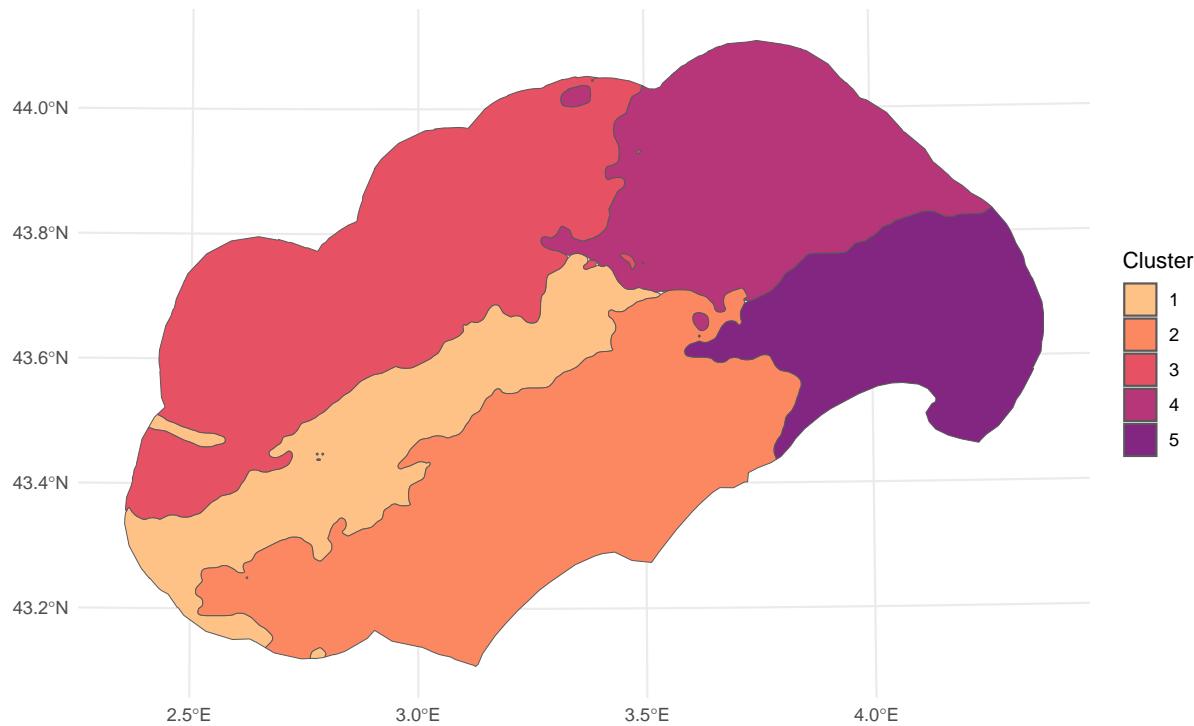


FIGURE 4.17 – Carte finale des zones interpolées après lissage et vectorisation.

La figure 4.17 illustre le rendu final obtenu après lissage des unités agroclimatiques. Cette étape de transformation, qui consiste à convertir des milliers de cellules raster en polygones vectoriels continus et à lisser leurs contours, facilite grandement l'exploitation des résultats par les services de SIG du département. En effet, cette représentation vectorielle offre une meilleure lisibilité et une intégration plus aisée dans les environnements SIG, où les analyses spatiales et les applications cartographiques sont courantes.

Bien que le processus automatisé permette d'éliminer une grande partie des artefacts et des irrégularités, un nettoyage manuel complémentaire des polygones pourrait être envisagé par les opérateurs SIG afin de garantir une qualité optimale, notamment en supprimant les discontinuités résiduelles. Cependant, le cœur de ce travail n'a pas été centré sur cette étape de post-traitement, mais plutôt sur la définition et la validation d'une méthodologie robuste pour redéfinir des unités agroclimatiques homogènes sur le territoire.

Ainsi, la méthode développée ici vise à identifier des zones présentant des similitudes agroclimatiques fortes, en s'appuyant sur des données spatiales et des techniques d'interpolation puis de lissage adaptées. Le passage à une couche vectorielle structurée constitue une base solide pour des analyses spatiales approfondies, mais aussi pour des applications opérationnelles comme la planification agricole, la gestion des ressources naturelles ou la modélisation environnementale. Ce cadre méthodologique ouvre la voie à des traitements complémentaires par les équipes SIG, qui pourront affiner et adapter les unités selon leurs besoins spécifiques.

Chapitre 5

Conclusion

Face aux bouleversements climatiques actuels, la nécessité de disposer d'outils adaptés à l'échelle locale devient cruciale pour mieux comprendre, anticiper et accompagner les évolutions agricoles. Cette étude s'est donnée pour objectif de redéfinir les unités agroclimatiques du territoire héraultais, en prenant en compte à la fois les réalités climatiques récentes et les contraintes géographiques du terrain.

Protocole

De toute cette étude, il en découle une méthode reproductible et adaptable, permettant de redéfinir les unités agroclimatiques à l'échelle d'un territoire dans un contexte de changement climatique. Cette méthode repose sur une approche statistique hybride, mobilisant à la fois des techniques d'apprentissage non supervisé et supervisé, et s'articule en plusieurs étapes :

- **Préparation et sélection des stations météorologiques** : les données brutes issues des stations (températures maximales, minimales et précipitations) sont d'abord fusionnées, puis agrégées à l'échelle mensuelle pour construire un jeu de variables climatiques de base. Afin de garantir la robustesse des analyses, seules les stations présentant une couverture temporelle suffisante et des données complètes sont conservées. Cette sélection réduit légèrement le nombre de stations disponibles, mais permet de travailler sur un jeu de données plus homogène et plus fiable pour la suite des traitements.
- **Classification des stations via ClustGeo** : une classification ascendante hiérarchique est ensuite effectuée à l'aide de la méthode ClustGeo, qui combine deux types de distances : la distance dans l'espace des variables climatiques et la distance géographique. Le paramètre α , qui pondère l'influence de la composante spatiale, est optimisé afin de garantir un compromis entre homogénéité climatique interne des groupes et cohérence spatiale.
- **Interpolation spatiale par forêts aléatoires** : une fois les stations regroupées en classes agroclimatiques, un modèle supervisé de type forêt aléatoire est entraîné pour spatialiser ces classes à l'ensemble du territoire. Les variables explicatives utilisées sont des données géomorphologiques continues (altitude, pente, indice de vallée, distance à la mer, etc.). Le modèle permet ainsi d'étendre les résultats ponctuels à une échelle spatiale fine et complète.

Cette méthodologie permet ainsi d'objectiver la structuration agroclimatique d'un territoire et d'en suivre l'évolution dans le temps.

Observations

Appliqué à plusieurs périodes temporelles, ce protocole met en évidence une dynamique climatique marquée à l'échelle du département de l'Hérault. En exploitant les données issues de trois décennies consécutives, il devient possible d'observer l'évolution spatiale des unités agroclimatiques, révélant ainsi les effets du changement climatique à l'échelle territoriale.



FIGURE 5.1 – Anciennes unités agroclimatiques vectorisées avec zone tampon de 15 km (1997)

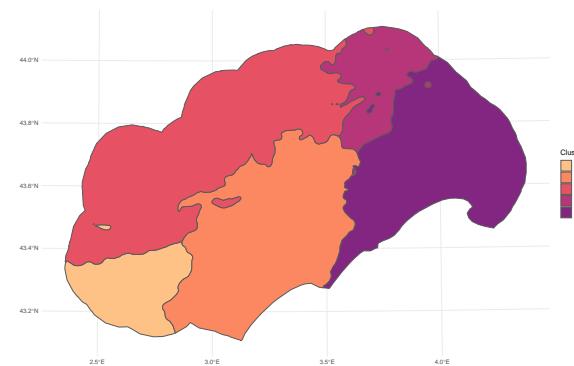


FIGURE 5.2 – Période 1990-2000

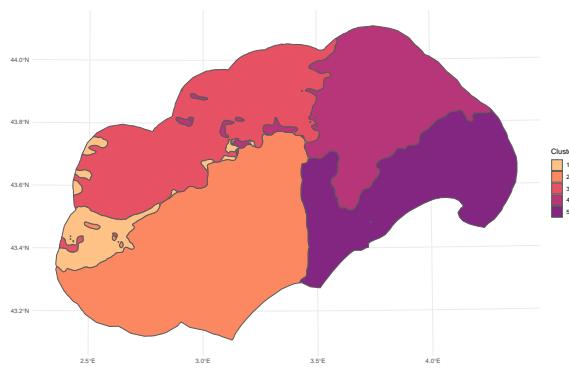


FIGURE 5.3 – Période 2000-2010

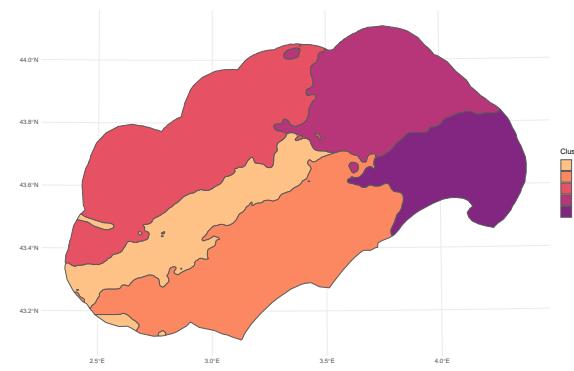


FIGURE 5.4 – Période 2010-2020

Pour des raisons d’homogénéité méthodologique et de comparabilité temporelle, il a été décidé de conserver un même nombre de classes (cinq) pour toutes les périodes étudiées. Les cartes obtenues pour les périodes 1990–2000, 2000–2010 et 2010–2020 (figure 5.2, figure 5.3, et figure 5.4) illustrent plusieurs tendances :

Pour la période 1990–2000 (figure 5.2), cinq zones bien distinctes sont observables. Nous pouvons établir une comparaison avec les unités agroclimatiques de 1997 (figure 5.1), car nous avons fait en sorte que les données historiques utilisées soient similaires à celles de notre carte couvrant 1990–2000.

Premièrement, la zone de Piémont (classe 3) est restée stable à l’intérieur de l’Hérault tout en s’étendant vers les montagnes situées dans la zone tampon. Une autre zone identifiée grâce au protocole est celle du Minervois (classe 1). Quant à la classe 2, elle regroupe l’ensemble des unités centrales du département, incluant le Bitterois, les Hauts Coteaux, la vallée de l’Orb-Lodévois ainsi que toute la vallée de l’Hérault. Enfin, la surface occupée par le Nord MontPELLIERais a diminué au profit de l’aire du MontPELLIERais qui, elle, s’est propagée au sud, au niveau du littoral.

Globalement, ces classes ne s’éloignent pas trop des anciennes unités agroclimatiques, bien que nous ne disposions plus aujourd’hui des données précises utilisées auparavant. On peut donc supposer que le sous-découpage observé, notamment pour la classe 2, a été réalisé à partir d’informations désormais manquantes ou par l’expertise des agents de la chambre d’agriculture.

Durant la période 2000–2010 (figure 5.3), l’unité 1 disparaît presque complètement, absorbée par l’expansion de l’unité 2. Cette dernière progresse vers l’ouest, colonisant des secteurs où les conditions deviennent plus sèches. Le centre du département se caractérise alors par une plus grande homogénéité, dominée nettement par l’unité 2. La zone des piémonts reste relativement stable, tandis que le MontPELLIERais (classe 5) s’étend en profondeur dans la vallée de l’Hérault. Par ailleurs, le Nord MontPELLIERais gagne également en importance au sud.

Enfin, durant la période 2010–2020 (figure 5.4), l’unité 1 réapparaît et s’étend cette fois sur un bandeau au centre du département, englobant l’ancien Minervois, la vallée de l’Orb-Lodévois ainsi que les Hauts Coteaux. Elle délimite ainsi une zone de piémont qui reste stable, tandis que la classe 2 s’étend désormais vers l’est. Par ailleurs, l’unité 5 n’est plus présente dans la vallée de l’Hérault et perd ainsi du terrain.

Les tendances générales observées révèlent une division climatique d’ouest en est. Au cours de la

dernière décennie, la zone sud-est se distingue par un climat très chaud tout en restant relativement humide, tandis que la zone sud (classe 2) est plutôt caractérisée par un climat aride. La zone de piémont demeure stable au fil des décennies, en raison de son climat montagnard.

Discussion et ouverture

Dans le cadre de cette étude, nous avons également exploré d'autres sources de données, notamment le produit SAFRAN[3] (présenté en annexe : section III), qui propose des variables climatiques sous forme de grilles régulières d'environ 8 km de résolution. Bien que cette approche permette de disposer d'un plus grand nombre de points de mesure, elle présente l'inconvénient de perdre l'expertise de terrain apportée par le choix des emplacements des stations, souvent positionnées dans des sites particulièrement pertinents.

Cette méthode pourrait toutefois constituer une piste intéressante pour des analyses futures. Cependant, nous avons préféré nous appuyer sur nos propres données, plus directement adaptées à notre territoire. De plus, le modèle SAFRAN a été calibré principalement sur un relief alpin, ce qui ne correspond pas complètement aux caractéristiques topographiques de notre zone d'étude, limitant son intérêt.

Par ailleurs, l'imputation des données manquantes s'est avérée particulièrement complexe. En effet, lorsqu'une station météorologique tombe en panne, la reprise des mesures n'est généralement pas immédiate, ce qui crée des lacunes importantes et prolongées dans les séries temporelles. Ces interruptions rendent délicate toute tentative de reconstitution fiable des données, limitant ainsi leur exploitation pour des analyses fines sur la variabilité spatiale et temporelle.

Afin de faciliter l'accès aux résultats de cette étude et de rendre les dynamiques spatiales plus accessibles aux acteurs locaux, nous avons développé un outil simple de visualisation interactive basé sur la bibliothèque **Leaflet**. Cet outil permet d'explorer les classifications agroclimatiques sur différentes périodes à travers une interface intuitive, offrant une meilleure compréhension des évolutions observées.

Cette approche interactive constitue une étape importante vers une diffusion plus large des connaissances et ouvre la voie à des applications concrètes, notamment en appui à la prise de décision agricole et à la gestion territoriale. L'outil est présenté en détail en annexe (section IV).

Chapitre 6

Annexe

I Complements sur la classification

I.1 Étude complémentaire sur les anciennes périodes avec ClustGeo

Comme pour la période 2010–2020, nous avons préparé les données nécessaires à l’application de la méthode ClustGeo pour les périodes plus anciennes. Les stations utilisées diffèrent donc selon les périodes, afin de répondre au mieux à nos critères d’analyse.

Période de 1990 à 2000

Pour cette période, nous disposons de 55 stations. Nous réalisons d’abord une classification préliminaire par la méthode des k-means, dont le résultat est présenté en figure 1.

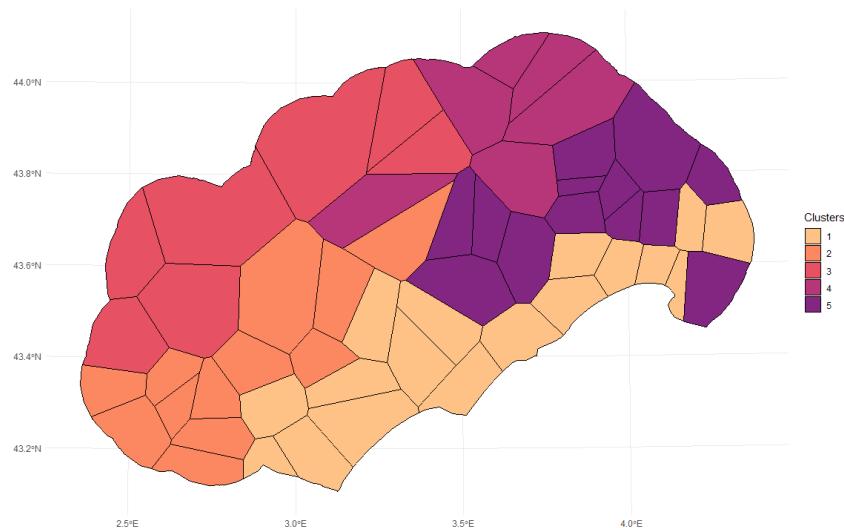


FIGURE 1 – Classification par k-means des données 1990–2000.

Le choix du paramètre α , qui équilibre l’importance des distances géographiques et statistiques dans la méthode ClustGeo, est déterminé à partir de la courbe présentée en figure 2. Nous retenons finalement $\alpha = 0.45$ pour cette période.

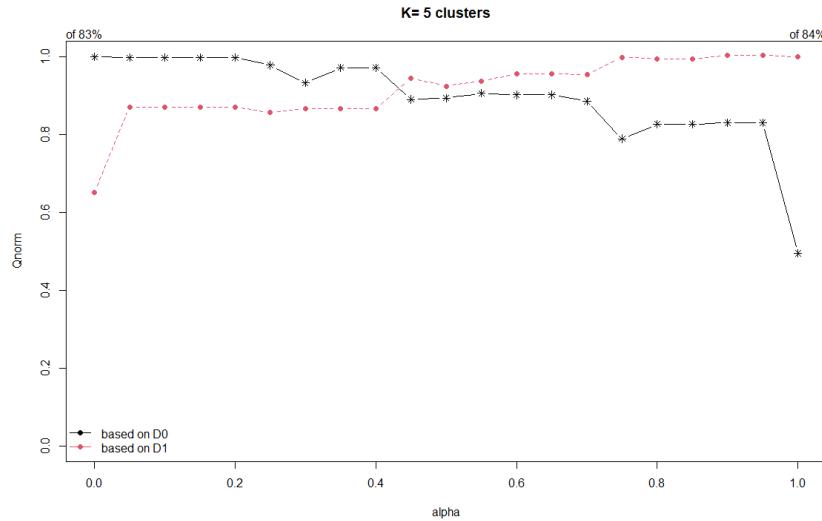


FIGURE 2 – Détermination du paramètre α pour la période 1990–2000.

L’application de la méthode ClustGeo avec ce paramètre conduit à la classification finale illustrée en figure 3.

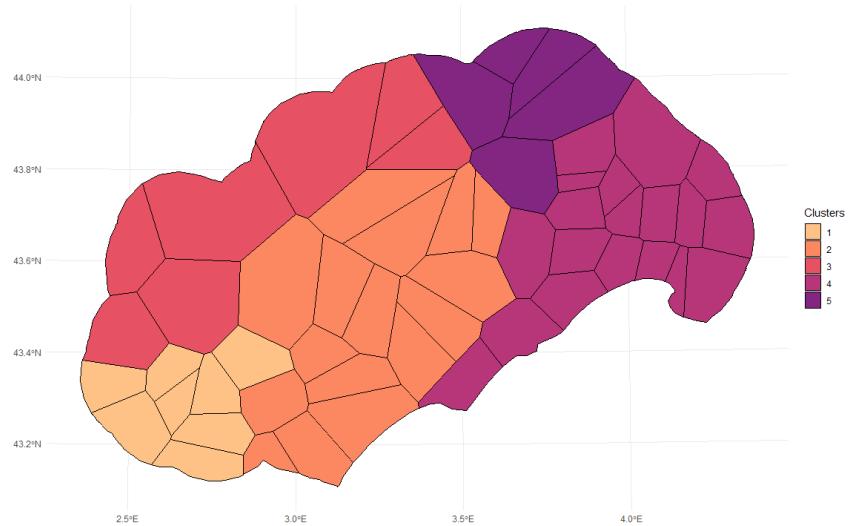


FIGURE 3 – Résultat de la classification ClustGeo pour la période 1990–2000 avec $\alpha = 0.45$.

Cette classification semble cohérente avec les anciennes unités agroclimatiques définies sur un périodes similaire.

Période 2000–2010

Pour cette période, nous disposons de 75 stations. La première étape consiste en une classification par k-means, dont le résultat est présenté en figure 4.

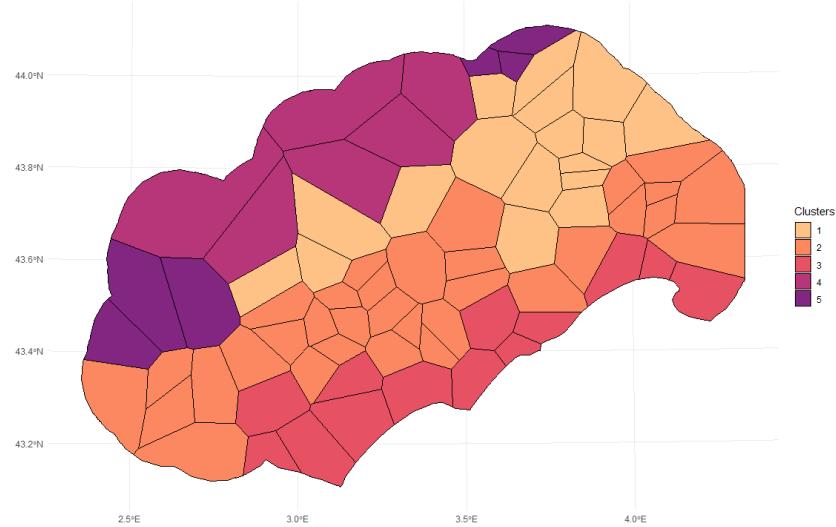


FIGURE 4 – Classification par k-means pour la période 2000–2010.

Le choix du paramètre α est effectué en comparant différentes matrices de distance, comme illustré en figure 5. Nous retenons finalement $\alpha = 0.4$, préférant la matrice de distance D_0 pour cette période.

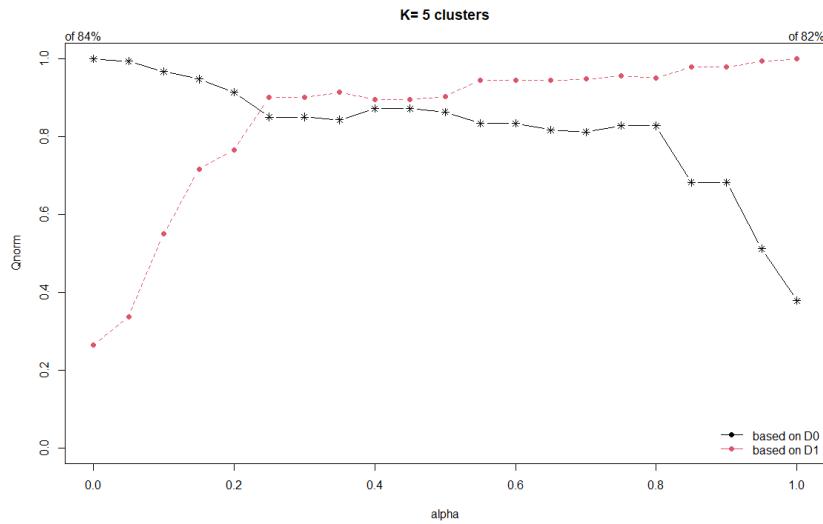


FIGURE 5 – Choix du paramètre α pour la période 2000–2010.

La classification finale obtenue avec la méthode ClustGeo est présentée en figure 6.

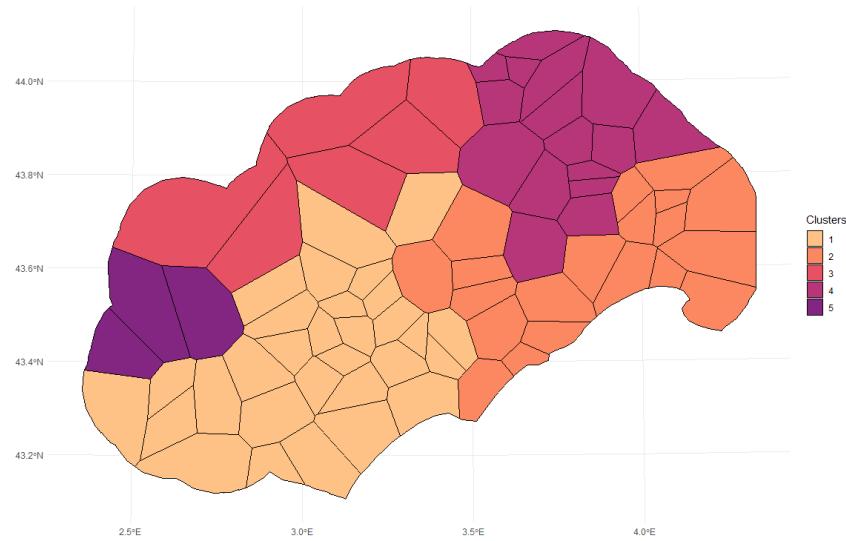


FIGURE 6 – Résultat de la classification ClustGeo pour la période 2000–2010 avec $\alpha = 0.4$.

Cette classification ne nous satisfait pas pleinement, notamment en raison de la présence d'une classe très sous-représentée. Néanmoins, par souci d'homogénéité et pour faciliter la comparaison entre périodes, nous conservons un découpage en cinq classes.

I.2 Distribution des groupes obtenus par ClustGeo pour la période de 2010 - 2020

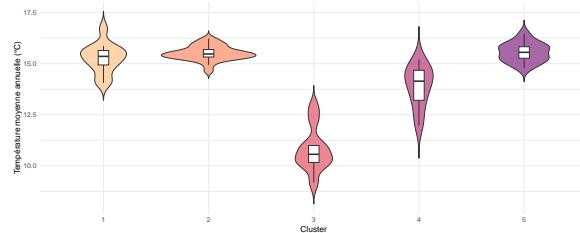


FIGURE 7 – Température moyenne annuelle par groupe

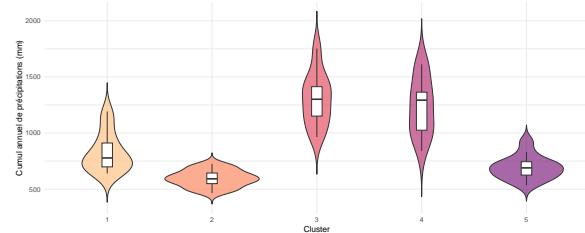


FIGURE 8 – Cumul annuel de précipitations par groupe

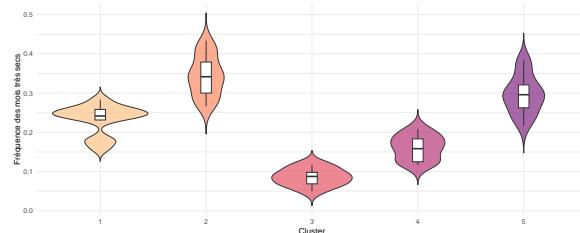


FIGURE 9 – Fréquence des mois très secs par groupe

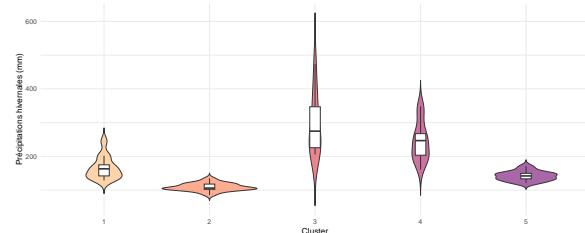


FIGURE 10 – Cumul des précipitations hivernales par groupe

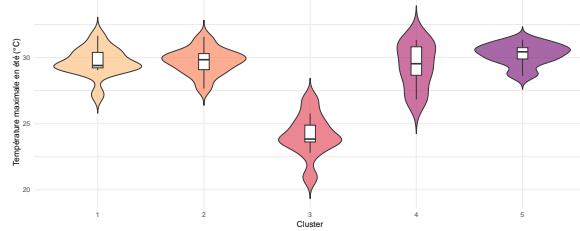


FIGURE 11 – Température maximale estivale par groupe



FIGURE 12 – Température maximale hivernale par groupe

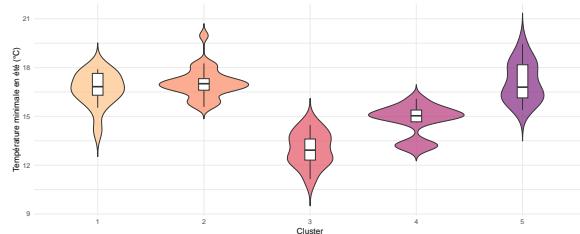


FIGURE 13 – Température minimale estivale par groupe

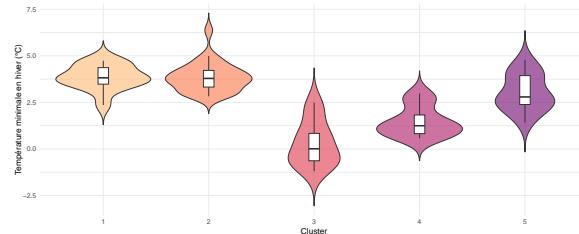


FIGURE 14 – Température minimale hivernale par groupe

La classe 3 se distingue quasiment tout le temps des autres. On voit clairement que c'est le manque de pluie et donc la fréquence de mois secs qui différencie la classe 2 des autres.

II Compléments sur l'interpolation

II.1 Observations complémentaires sur le krigage

Variogrammes et krigages pour la période 2010–2020

Les variogrammes expérimentaux, présentés en figure 15, figure 16 et figure 17, ont été calculés pour les différentes classes afin d'analyser la structure spatiale des données et d'ajuster les modèles nécessaires au krigage.

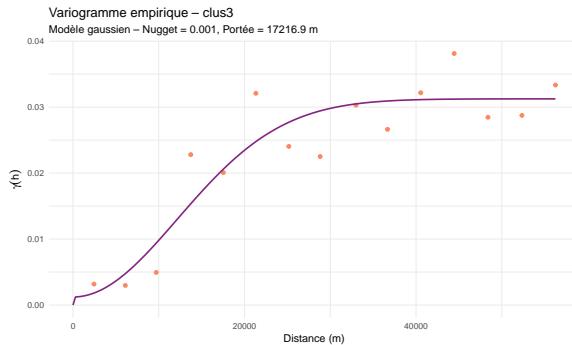


FIGURE 15 – Variogramme expérimental pour la classe 3.

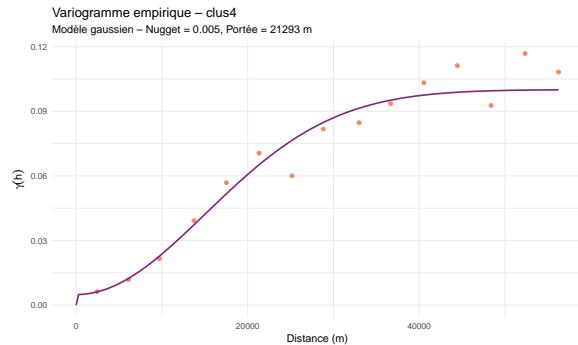


FIGURE 16 – Variogramme expérimental pour la classe 4.

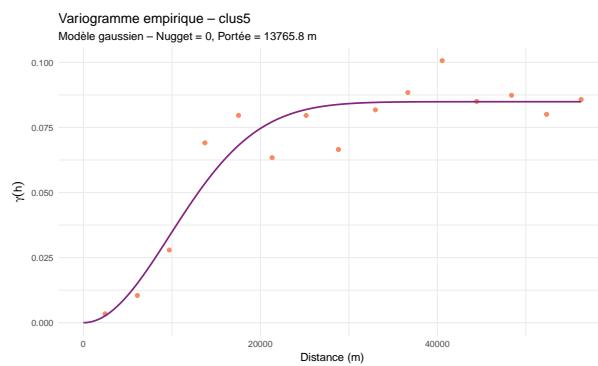


FIGURE 17 – Variogramme expérimental pour la classe 5.

Nous obtenons trois variogrammes suivant un modèle gaussien.

À partir de ces modèles, nous réalisons le krigage qui permet d'estimer pour chaque cellule de la grille un indicateur représentatif de l'appartenance à chaque classe. Les cartes de ces indicateurs sont illustrées en figure 18.

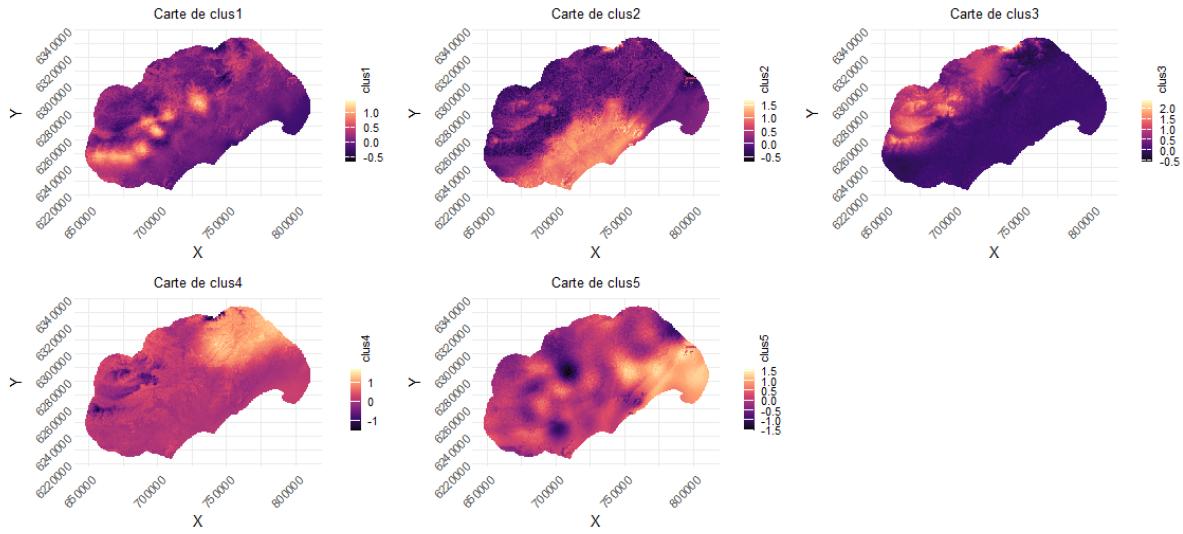


FIGURE 18 – Cartes des indicateurs issus du krigage pour chaque classe sur la période 2010–2020.

La carte finale présentée dans le corps du rapport est obtenue en assignant à chaque cellule de la grille la classe correspondant à la valeur maximale parmi ces indicateurs (*argmax*).

Krigage pour les autres périodes

En parallèle, nous pouvons une nouvelle fois observer les résultats obtenus pour les deux autres périodes.

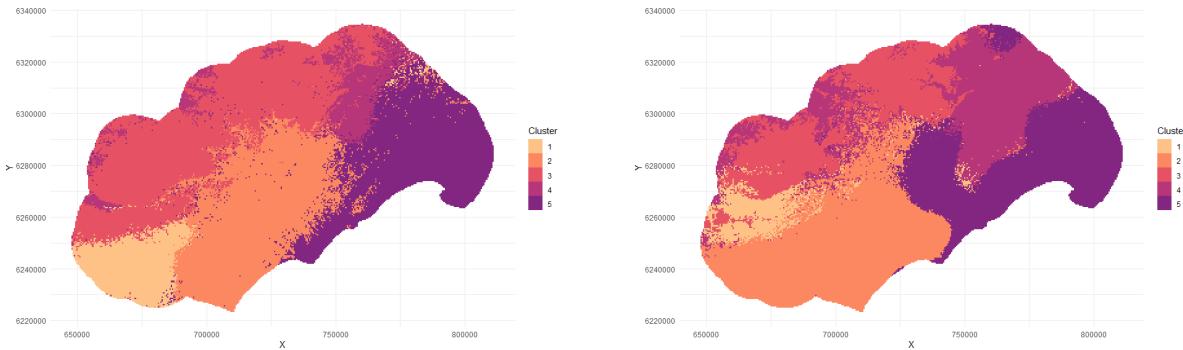


FIGURE 19 – Résultat du krigage pour la période 1990–2000.

FIGURE 20 – Résultat du krigage pour la période 2000–2010.

Nous voyons, surtout pour la figure 20 que les classes ont tendance à se fragmenter. C'est aussi le cas pour la figure 19 au niveau des bordures de la zone de Piémont.

II.2 Observations complémentaires sur les forêts aléatoires

Arbres de décision pour la période de 2010 à 2020

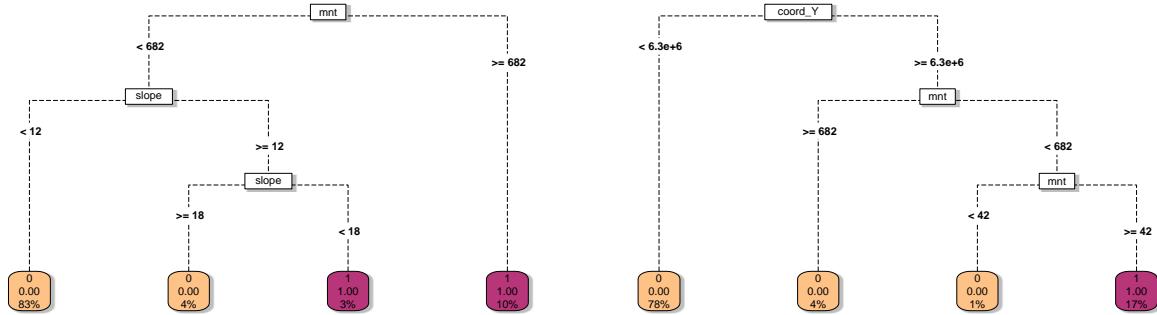


FIGURE 21 – Arbre de décision extrait pour la classe 3

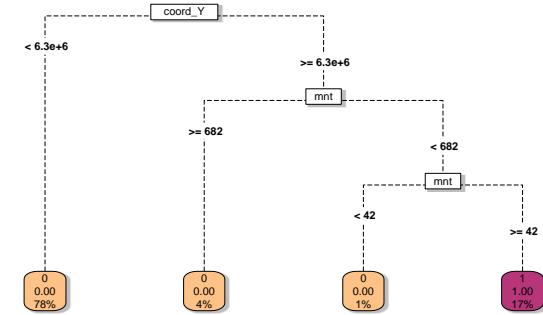


FIGURE 22 – Arbre de décision extrait pour la classe 4

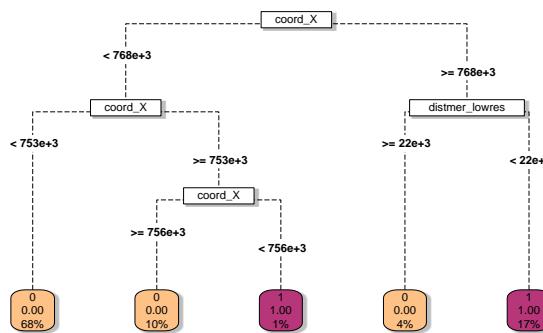


FIGURE 23 – Arbre de décision extrait pour la classe 5

Forêts aléatoires pour les autres périodes

En parallèle, nous présentons ci-dessous les résultats de classification spatiale obtenus par les forêts aléatoires pour les deux autres périodes temporelles :

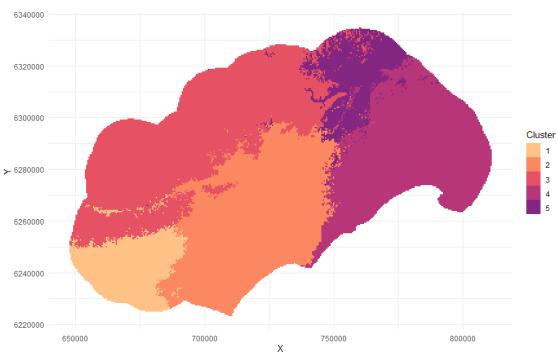


FIGURE 24 – Classification par forêt aléatoire — période 1990–2000

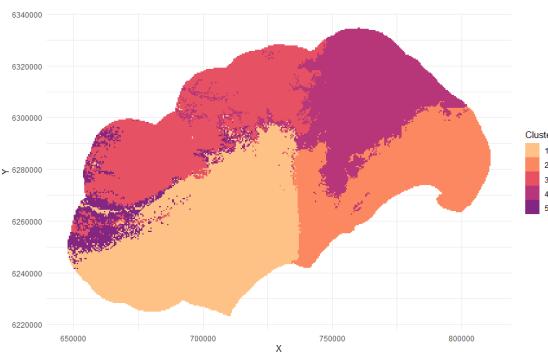


FIGURE 25 – Classification par forêt aléatoire — période 2000–2010

III Protocole appliqué aux données SAFRAN

Après préparation des données issues de la grille SAFRAN, nous avons appliqué un algorithme de classification non supervisée par *k-means* :

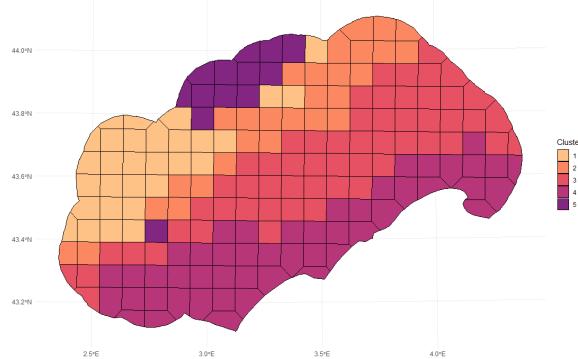
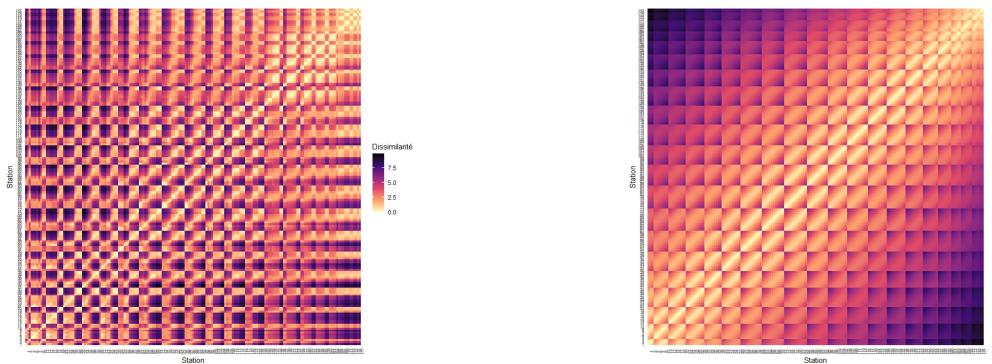


FIGURE 26 – Résultat du *k-means* sur la grille SAFRAN pour la période 2010–2020. Chaque point correspond à une maille de 8 km, représentant le barycentre d'une cellule de Voronoï.

La résolution régulière de la grille (8 km) fait que l'utilisation des polygones de Voronoï pour visualiser les classes donne une représentation très géométrique, où chaque point SAFRAN est au centre de sa cellule (figure 26).

Classification

Nous avons ensuite utilisé la méthode **ClustGeo** afin de déterminer le paramètre α optimisant l'équilibre entre la prise en compte des distances statistiques et géographiques. Les matrices de distances correspondantes sont illustrées ci-dessous, côte à côte (figure 27) :



(a) Matrice des distances statistiques

(b) Matrice des distances géographiques

FIGURE 27 – Matrices des distances utilisées dans la méthode **ClustGeo** pour le jeu de données SAFRAN.

Ces matrices sont beaucoup plus régulières, c'est bien sûr dû au fait que les données proviennent d'une grille et ont été estimées plutôt que récoltées.

Le choix de α est effectué en fonction de la pondération optimale entre ces deux types de distances. La figure suivante présente le critère d'optimisation permettant de sélectionner la valeur retenue (figure 28) :

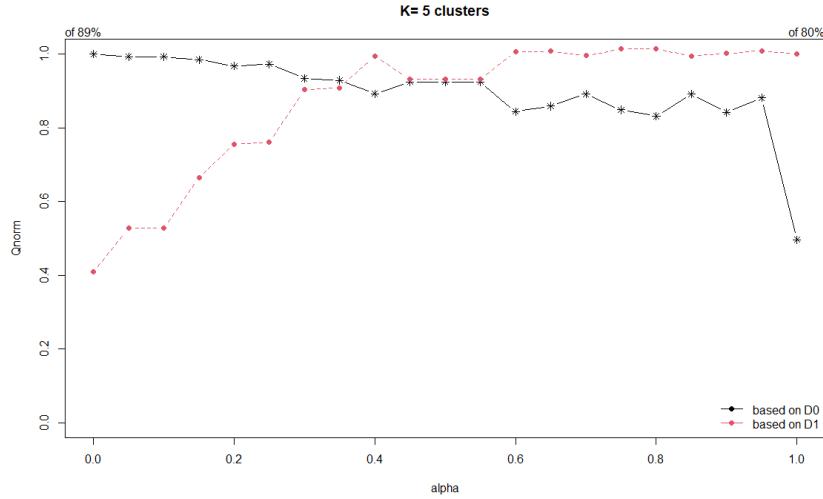


FIGURE 28 – Critère d'optimisation pour le paramètre α

Nous retenons donc $\alpha = 0.35$ pour la suite de l'analyse. Le résultat de la classification issue de la méthode **ClustGeo** avec ce paramètre est présenté en figure 29.

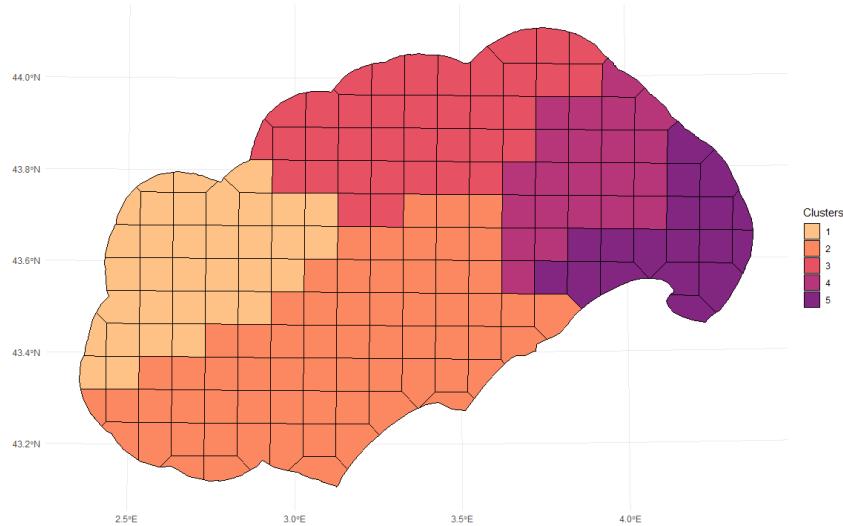


FIGURE 29 – Résultat de la classification **ClustGeo** sur les données SAFRAN avec $\alpha = 0.35$.

Toutefois, cette partition ne nous satisfait pas pleinement, notamment en raison de la division de la zone de Piémont, qui est pourtant la zone la plus homogène et la plus stable selon nos observations précédentes. Cette fragmentation paraît donc peu cohérente avec les dynamiques agroclimatiques attendues dans cette région.

Interpolation

Nous poursuivons le protocole en utilisant la même grille de données géomorphologiques, puisque le territoire étudié reste identique.

Dans un premier temps, nous appliquons un algorithme de forêt aléatoire (random forest) afin d'interpoler les données ponctuelles issues de la classification. Le résultat de cette étape, donnant pour chaque cluster la probabilité d'appartenance, est illustré en figure 30.

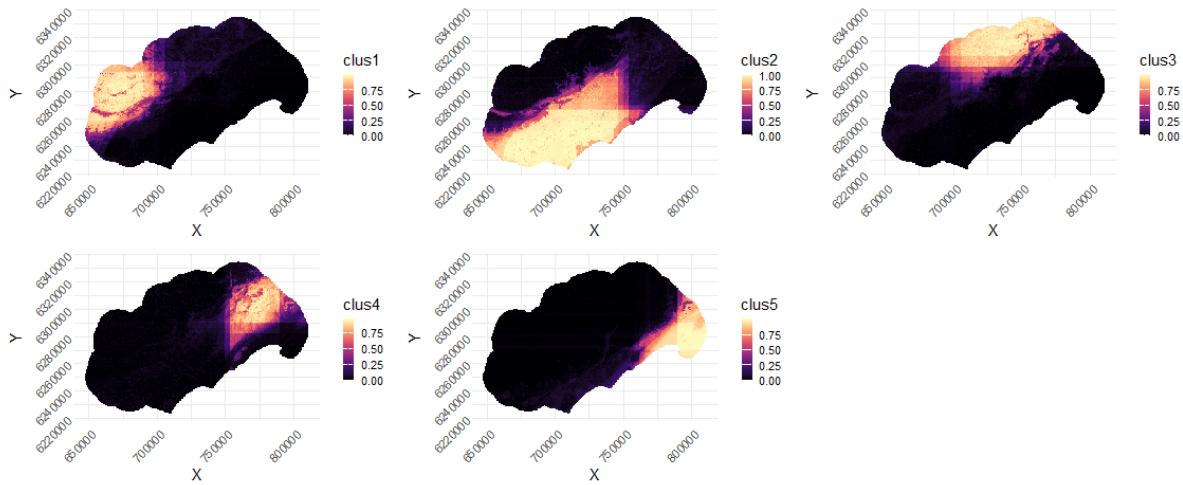


FIGURE 30 – Probabilités estimées par forêt aléatoire pour chaque cluster.

Ensuite, pour chaque point de la grille, nous sélectionnons la classe associée à la probabilité la plus élevée, comme montré en figure 31. Puis, afin d'éliminer les pixels résiduels isolés et de garantir une continuité spatiale cohérente, nous appliquons un lissage sur la carte finale, dont le rendu est présenté en figure 32.

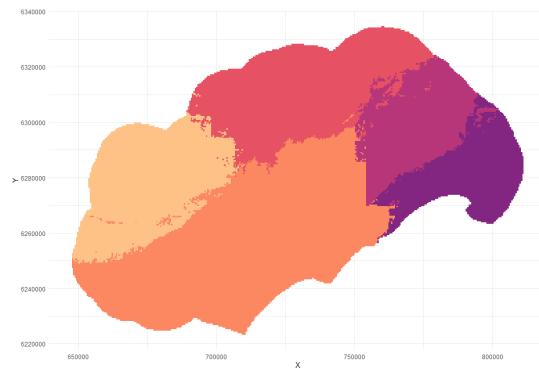


FIGURE 31 – Classe retenue à chaque point de la grille, correspondant à la probabilité maximale issue de la forêt aléatoire.

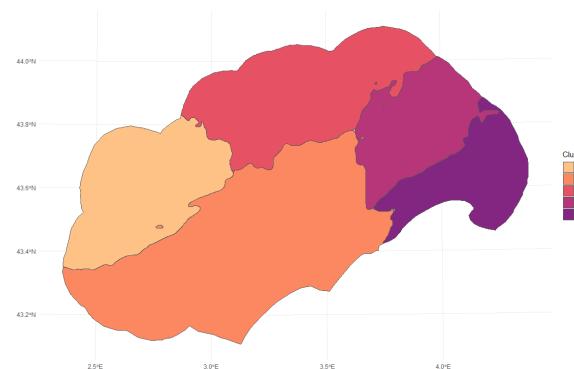


FIGURE 32 – Carte finale lissée après retrait des pixels résiduels, obtenue à partir de l'interpolation par forêt aléatoire.

Ces résultats diffèrent beaucoup de ceux obtenus à partir des données issues de vraies stations. Pour cette raison, nous privilégierons l'utilisation de notre modèle basé sur les observations terrain.

IV Outil de visualisation

Afin de permettre aux agents de l'Hôtel du Département de visualiser les évolutions des unités agroclimatiques, une carte interactive a été développée. Elle est mise à leur disposition pour une consultation et une utilisation autonome.

Cette carte rassemble une grande partie des informations dont les acteurs auraient besoin.

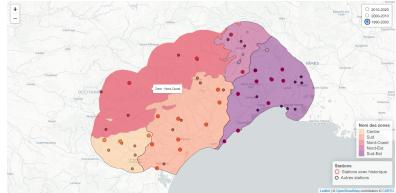


FIGURE 33 – Vue 1 de la carte interactive

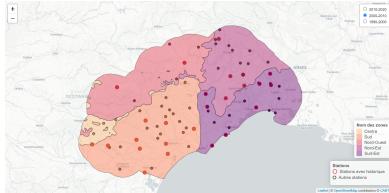


FIGURE 34 – Vue 2 de la carte interactive

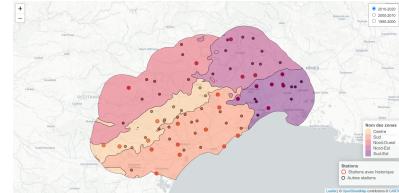


FIGURE 35 – Vue 3 de la carte interactive

L'outil nous permet de naviguer entre les différentes cartes d'unités agroclimatiques soit pour la période 1990-2000 (figure 33), pour la période 2000-2010 (figure 34) et pour la période (figure 35). Le nom des zones, si on leur en donne un à terme, reste encore à définir.

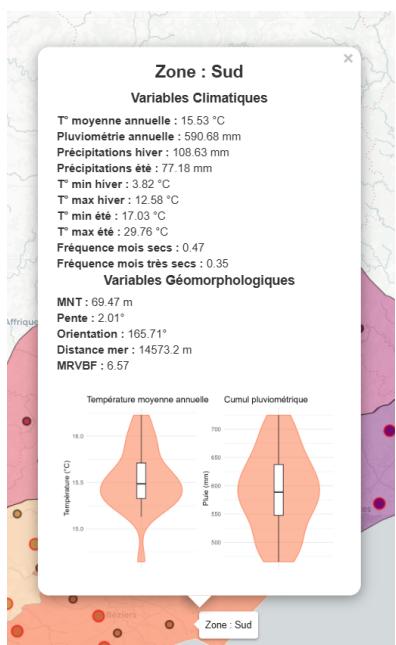


FIGURE 36 – Vue 4 de la carte interactive

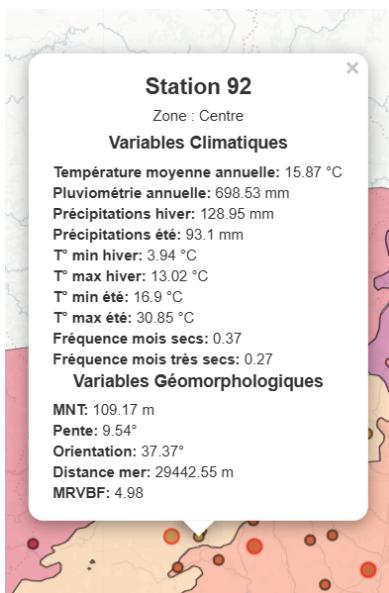


FIGURE 37 – Vue 5 de la carte interactive

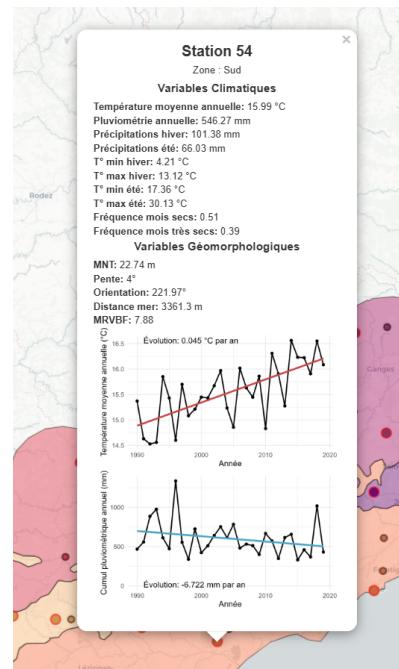


FIGURE 38 – Vue 6 de la carte interactive

L'utilisateur peut aussi cliquer sur les zones, les stations avec un long historique et les autres stations pour en obtenir les informations. Comme cet outil est destiné à tout public, les informations renseignées sont accessibles.

Lorsqu'un utilisateur clique sur une zone (figure 36), la moyenne des valeurs de toutes les stations situées dans cette zone est affichée, accompagnée de deux graphiques représentant la distribution des variables climatiques les plus pertinentes : la température moyenne annuelle et le cumul pluviométrique moyen de la zone.

En cliquant sur une station quelconque, l'utilisateur obtient toutes les informations associées à cette station (figure 37). Si la station est entourée d'un liseré rouge (figure 38), ce qui signifie qu'elle a servi à définir les unités agroclimatiques pour les trois périodes étudiées, des droites de régression sur 30 ans apparaissent alors, illustrant les tendances des températures moyennes annuelles et des cumuls pluviométriques.

Bibliographie

- [1] Louis Amiot, Vincent Dubreuil, Elisabeth Colnard, Laurence Ligneau, and Valerie Bonnardot. Méthode de zonage agroclimatique en Bretagne et Pays de la Loire. In *36e Colloque de l'Association Internationale de Climatologie.*, pages 21–24, Bucarest (Roumania), Romania, July 2023.
- [2] Renato M. Assuncao and Elias T. Krainski. Spatial 'k'luster analysis by tree edge removal.
- [3] Patrick Bertuzzi, Philippe Clastre, and Maël Aubry. Information sur les mailles safran, 2022.
- [4] V. Bonnardot, L. Amiot, T. Petitjean, and V. Dubreuil. Zonage agroclimatique en bretagne : Programme fermadapt, July 2024.
- [5] Leo Breiman. Statistical modeling : The two cultures. *Statistical Science*, 2001.
- [6] Marie Chavent, Vanessa Kuentz, Amaury Labenne, and Jérôme Saracco. Clustgeo : Hierarchical clustering with spatial constraints. <https://CRAN.R-project.org/package=ClustGeo>, 2021. R package version 2.1.
- [7] Marie Chavent, Vanessa Kuentz-Simonet, Amaury Labenne, and Jérôme Saracco. Clustgeo : Classification ascendante hiérarchique (cah) avec contraintes de proximité géographique. In *47èmes Journées de Statistique de la SFdS*, Lille, France, 2015. Société Française de Statistique.
- [8] Conseil départemental de l'Hérault. Bulletins infoclim et aléacli climatologie. <https://odee.herault.fr/index.php/thematiques/climatologie/167-bulletins-infoclim>, 2025.
- [9] Conseil départemental de l'Hérault. Portail de l'odceel - observatoire départemental climatologie eau environnement littoral. <https://odee.herault.fr/>, 2025.
- [10] R. W. Fleming and R. M. Hoffer. A digital terrain analysis algorithm for landform classification. *Photogrammetric Engineering*, 1979.
- [11] John C. Gallant and Timothy I. Dowling. A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research*, 2003.
- [12] B. K. P. Horn. Mountain building by the process of erosion. *Geological Society of America Bulletin*, 1981.
- [13] Hérault Data. Unités agroclimatiques de l'hérault. <https://www.herault-data.fr/explore/dataset/unites-agroclimatiques-de-lherault/information/>, 1997.
- [14] C. JEAN, G. GARAPIN, JL. TONDUT, JJ. MONLEZUN, D. BARIDA, and MH. MOUBAYED-BREIL. *Valorisation des ressources naturelles - Bioclimats et séries de sols - Application au vignoble*.
- [15] Polytechnique Montréal. Cours sur le krigage et les variogrammes, 2023. Disponible en ligne : <https://moodle.polymtl.ca/course/view.php?id=1118>.
- [16] Robert C Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 1957.
- [17] D. F. Ritter. Algorithm for determining slope and aspect. *Journal of Geographical Systems*, 1987.
- [18] SAGA GIS. Multi-resolution valley bottom flatness (mrvbf). https://saga-gis.sourceforge.io/saga_tool_doc/2.1.4/ta_morphometry_8.html, 2015.
- [19] Techniloire. Variables et indicateurs du climat — indicateurs agroclimatiques (ia).
- [20] Wikipédia contributors. Pondération inverse à la distance. https://fr.wikipedia.org/wiki/Pond%C3%A9ration_inverse_%C3%A0_la_distance, 2024.