Partitionnement de données pour l'informatique climatique : Contributions à l'amélioration des méthodes d'identification automatique des régimes de temps en climat tropical insulaire.

Biabiany Emmanuel

UNIVERSITÉ DES ANTILLES

École Doctorale n°589

Milieu insulaire tropical à risques : protection, valorisation, santé et développement

Thèse

présentée par

M. **Emmanuel Biabiany**

pour obtenir le titre de

Docteur en Informatique

de l'Université des Antilles

—

Partitionnement de données pour l'informatique climatique : Contributions à l'amélioration des méthodes d'identification automatique des régimes de temps en climat tropical insulaire

—

soutenue publiquement le 25 Septembre 2020 devant le jury composé de :

Directeur de thèse :	M. Didier Bernard, MCf. HDR – Université des Antilles
Co-encadrement :	M. Vincent Pagé, MCf. – Université des Antilles
Rapporteurs :	Mme Christine Fernandez–Maloigne, Pr. – CUE LPC
	M. Dominique Béréziat, MCf. HDR – Sorbonne Université
Examinateurs :	Mme Hyewon Seo, Dir. Recherche – ICube/CNRS
	M. Clément Mallet, Dir. Recherche – IGN/CNRS
	M. Lionel Prévost, Pr. – ESCIA Paris
Invités :	M. Julien Boé, Chargé de Recherche – CERFACS/CNRS

Résumé

Ce manuscrit fait état de nouveaux travaux dans le domaine de l'informatique climatique, qui ont conduit à un ensemble de contributions aux méthodes d'identification automatique des régimes de temps en région Caraïbe. À partir de méthodes informatiques largement présentées dans la bibliographie consultée, nous avons au préalable ciblé celles de l'apprentissage non supervisé, et plus particulièrement les méthodes de clustering K-Means (KMS) et Classification Ascendante Hiérarchique (CAH). Des applications directes de ces méthodes aux problématiques des courants vecteurs de banc d'algues sargasses, puis au partitionnement de données de Géopotentiel ont été réalisées. Ces méthodes ont permis d'identifier des groupes de jours (ou clusters) ayant des caractéristiques similaires. Les barycentres (ou centroïdes) des groupes ainsi obtenus ont été analysés par les experts du climat. Cependant, cette approche ne produit pas systématiquement des résultats cohérents puisque ces barycentres ne représentent pas toujours la réalité physique des structures retenues.
Par la suite, nous avons concentré nos efforts sur la recherche et l'identification de régimes de temps caractéristiques de la zone Caraïbe. Ces régimes sont en général décrits comme des configurations spatio-temporelles récurrentes, de grande échelle, qui influencent les situations météorologiques locales. La recherche dans ce domaine pour la région Caraïbe est encore balbutiante. Pour les travaux déjà publiés, plusieurs points semblent problématiques. Trois d'entre eux ont attiré notre attention. Tout d'abord, l'absence de quantification de la qualité des clusters, rend nécessaire une grande quantité de justifications physiques, pour valider la pertinence des régimes proposés. Elle complique aussi, la comparaison entre les différents travaux existants. Ensuite, parmi les arguments présentés, certains montrent que les propositions formulées ne sont pas pleinement satisfaisantes. Enfin, selon les experts, la cohérence temporelle des clusters de certaines études ne semble pas correspondre à la saisonnalité de la région.
Pour pallier à ces difficultés, dans un premier temps, nous proposons l'usage de l'indice Silhouette. L'évaluation de la pertinence des clusters retenus mais également la comparaison des différentes méthodes utilisées, ont été réalisées par cet indice. Après vérification, il y a concordance entre l'analyse produite par l'indice et celle des experts du climat. Néanmoins, dans certains cas, l'indice indique également que les clusters constitués peuvent être améliorés. En s'intéressant plus précisément aux algorithmes de partitionnement, et en particulier à la notion de distance qu'ils utilisent, il apparaît que ces difficultés sont principalement liées à la complexité des données, mais aussi aux mesures de similarité permettant de les comparer. Après une critique des propriétés de la distance L2, utilisée par défaut, nous proposons la mise en place d'une nouvelle mesure de dissimilarité, nommée Expert Deviation (ED). Elle repose sur un découpage spatial, une quantification en histogrammes, et un traitement zonal avec la divergence de Kulback-Leibler (KL). Nous montrons que l'ED conduit à des résultats bien meilleurs, qu'il s'agisse d'évaluations numériques de la qualité des clusters par l'indice de silhouette, que

d'interprétations par les experts du domaine.
Cette nouvelle mesure est adaptative dans sa conception et son utilisation. Nous présentons son principe et passons à une application dans le domaine de la physique de l'atmosphère, en utilisant des données telles que les précipitations mesurées par satellite. Aux petites Antilles, les pluies sont connues pour leur forte variabilité spatio-temporelle et elles influencent directement le climat à ces latitudes. À l'aide d'ED, nous avons ainsi pu identifier des configurations récurrentes plus cohérentes et physiquement interprétables pour ce paramètre et pour le vent. Ces résultats ont permis d'accroître les connaissances des experts du climat sur les structures atmosphériques liées aux régimes de temps d'inter-saison et leur dynamique. L'ensemble de ces travaux et l'utilisation de la « mesure ED » ouvrent un grand nombre de perspectives pour la recherche de configurations spatio-temporelles récurrentes, mais également dans tous les domaines d'applications utilisant des images.

Disciplines

Informatique, Physique de l'atmosphère.

Mots-clés

Intelligence artificielle, Regroupement de données, Régimes de temps, Caraïbe, indice de Silhouette, divergence de Kullback-Leibler.

Abstract

THis manuscript reports on new work in the field of climate informatics, which has led to a set of contributions to methods for the automatic identification of weather patterns in the Caribbean region. Starting from computational methods widely presented in the bibliography consulted, we have previously targeted those of unsupervised learning, and more particularly the K-Means clustering (KMS) and Hierarchical Ascending Classification (HAC) methods. Direct applications of these methods to the problems of the vector currents of Sargasso algae banks, and then to the partitioning of Geopotential data have been carried out. These methods made it possible to identify groups of days (or clusters) with similar characteristics. The barycentres (or centroids) of the groups thus obtained were analysed by climate experts. However, this approach does not systematically produce consistent results since these barycentres do not always represent the physical reality of the structures selected.

Subsequently, we concentrated our efforts on researching and identifying weather patterns characteristic of the Caribbean zone. These regimes are generally described as recurrent spatio-temporal configurations, on a large scale, which influence local weather situations. Research in this area for the Caribbean region is still in its infancy. For the work already published, several points seem problematic. Three of them have attracted our attention. Firstly, the lack of quantification of the quality of the clusters, makes a large amount of physical justification necessary, to validate the relevance of the proposed regimes. It also complicates the comparison between the different existing works. Then, among the arguments presented, some show that the proposals made are not fully satisfactory. Finally, according to the experts, the temporal coherence of the clusters of certain studies does not seem to correspond to the seasonality of the region.

In order to overcome these difficulties, as a first step, we propose the use of the Silhouette index. The evaluation of the relevance of the selected clusters, but also the comparison of the different methods used, were carried out using this index. After verification, there is a concordance between the analysis produced by the index and that of the climate experts. Nevertheless, in some cases, the index also indicates that the clusters constituted can be improved. Looking more specifically at the partitioning algorithms, and in particular at the notion of distance they use, it appears that these difficulties are mainly related to the complexity of the data, but also to the similarity measures that make it possible to compare them. After a critique of the properties of the distance L2, used by default, we propose the implementation of a new dissimilarity measure, named Expert Deviation (ED). It is based on a spatial breakdown, a quantification in histograms, and a zonal treatment with the Kulback-Leibler (KL) divergence. We show that the ED leads to much better results, both in numerical evaluations of cluster quality by the silhouette index and in interpretations by experts in the field.

This new measure is adaptive in its design and use. We present its principle and move on to

an application in the field of atmospheric physics, using data such as precipitation measured by satellite. Rainfall in the Lesser Antilles is known to be highly variable in space and time and directly influences the climate at these latitudes. Using ED, we were able to identify more coherent and physically interpretable recurrent patterns for this parameter and for wind. These results have increased the knowledge of climate experts on the atmospheric structures related to inter-seasonal weather patterns and their dynamics. All this work and the use of the "ED measure" open up a large number of perspectives for the search for recurrent spatio-temporal configurations, but also in all fields of applications using images.

Disciplines

Computer science, Atmospheric Physics.

Keywords

Artificial intelligence, Clustering methods, Weather patterns, Caribbean, Expert Deviation, Silhouette index, Kullback-Leibler divergence.

Remerciements

La thèse est selon moi similaire à un voyage vers l'inconnu, comparable aux explorations spatiales, où l'on cherche du début à la fin, à se mettre dans les meilleures conditions pour réussir. Et tout comme les astronautes doivent pouvoir compter les uns sur les autres, j'ai pu compter sur ma famille et mon équipe d'encadrants.
C'est la raison pour laquelle, je souhaiterais remercier tout d'abord mon directeur de thèse, Monsieur Didier Bernard, pour la qualité de son expertise, sa rigueur dans la recherche et son engagement irréprochable qui m'ont permis de garder le cap tout au long de cette entreprise.
De la même manière, je remercie Monsieur Vincent Pagé qui a accepté d'assurer l'encadrement de mes travaux, et qui m'a apporté toute son expertise et son soutien, il m'a également communiqué son amour pour l'informatique, me permettant ainsi d'arriver jusque-là.
Je remercie également Madame Hélène Paugam-Moisy, pour avoir assurer la co-direction de ma thèse, elle a su me communiquer son sérieux et sa rigueur scientifique. C'est également à elle que je dois cette collaboration fructueuse avec Vincent.
Merci à Monsieur Lionel Prévost, pour avoir initié ces travaux de recherche et pour avoir pris part à mon jury de soutenance de thèse.
J'exprime toute ma reconnaissance à la Fondation Blandin qui a partiellement financé mes travaux, en me permettant de faire acquisition de tous les équipements avec lesquelles j'ai pu mené à bien mes travaux.
Je remercie Madame Christine Fernendez-Maloigne et Monsieur Dominique Béréziat, mes rapporteurs pour leur lecture et leurs analyses de mon manuscrit de thèse. Je remercie également les autres membres extérieurs examinateurs et invités de ma soutenance, Madame Heywon Seo, Monsieur Julien Boé et le président du jury Monsieur Clément Mallet pour leur expertise et leur participation à ce jury.
Merci à Sébastien, Jimmy, Manuel, Suzy, Andreï, Wilfried et Eric pour leur confiance et leur soutien, grâce à eux j'ai pu effectuer un large panel d'enseignements durant mes années de thèse. Merci à Raphaël, Narcisse, Richard, Romual, Gaël, Yann, Naoufal mes collaborateurs physiciens avec qui j'ai pu développer une amitié forte, c'est également grâce à eux que j'en ai appris un peu plus sur le climat dans les Antilles. Je remercie aussi Jean-Victor et Logan, mes tout premiers stagiaires, pour m'avoir fait confiance pour encadrer leur stage de fin de licence.

Ma famille et mes amis ont également joué un grand rôle dans ma vie durant ce voyage, je tiens ici à les remercier. Pour commencer, Déliex et Dany Biabiany, mes parents, pour leur soutien sans faille au quotidien. Leurs encouragements ont été pour moi une source intarissable de motivation, c'est eux qui ont fait ce que je suis, donc c'est à eux que je dois cette aboutissement.
J'ai aussi reçu l'indéfectible soutien quotidien de ma compagne Audrey. Elle a toujours su me donner la motivation et le courage pour aller de l'avant, je la remercie pour tout. Je remercie

aussi ma grande sœur, Cécilia Biabiany. Je ne le dis pas assez souvent mais elle est pour moi un exemple. J'espère que mes travaux la rendront fière.
J'ai une pensée pour mes cousins qui m'ont toujours soutenus, merci à Joris, Mathias, Mathieu, Léa, Nadège, Gladys, Géraud, Stelly, David et les autres. Bien-sûr je ne les citeraient pas tous parce que la famille est grande. Merci aussi à mes oncles et mes tantes, en particulier à Pascale, Laurent, Rogella et Marius pour tout le soutien qu'ils m'ont apporté au cours de ma vie. J'ai également une pensée forte pour mes amis Stéphane, Andrée, Christopher, Harold et Etienne. Enfin je remercie mes camarades sportifs Ruddy, Dominique, Phillipe et Mendjouka avec qui j'ai pu me changer les idées aux cours de nos sessions de football très disputées.

Je pense sincèrement que toutes ces personnes ont contribuées directement ou indirectement à la réussite de mes travaux, je leur exprime donc toute ma reconnaissance.

Équipe de recherche

Direction de thèse : M. Didier Bernard, Maitre de Conférence, en physique de l'atmosphère, au Laboratoire de Recherche en Géosciences et Énergies (LaRGE).

Co-direction *(de 2016 à 2018)* : Mme Hélène Paugam-Moisy, Professeur des Université, en informatique, au LAboratoire de Mathématiques, Informatique et Applications (LAMIA).

Co-encadrement *(2018 à 2020)* : M. Vincent Page, Maitre de Conférence, en informatique, au LAboratoire de Mathématiques, Informatique et Applications (LAMIA).

Financements

Équipements et fonctionnement : 2016 - 2020 : Bourse Claude Emmanuel Blandin

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Contexte et motivations

Au cours des dernières années, les thématiques environnementales ont pris une importance de plus en plus grande au sein de l'espace public. C'est notamment le cas des études de climatologie, comme en attestent les débats houleux qui secouent la société civile depuis une trentaine d'années, en dépit d'un consensus relativement fort parmi les experts du domaine.

Le grand public a donc pris en compte l'intérêt qu'il y a, pour nos sociétés, à s'intéresser aux évolutions du climat terrestre, qu'elles soient dues ou non aux effets du réchauffement global. Ce phénomène tend, au fil des dernières années, à s'accentuer en raison de l'intensification des activités anthropiques.

Les impacts des modifications de paramètres physiques tels les températures et précipitations, sur des échelles de temps inférieures au siècle sont désormais visibles. La recherche dans ce domaine a ainsi pu bénéficier d'un regain d'intérêt et les travaux s'y rapportant se sont multipliés, sous l'effet d'un accès facilité à des financements.

Les travaux présentés dans ce manuscrit s'inscrivent dans cette lignée, avec comme objectif principal, de contribuer à l'amélioration des connaissances sur le climat insulaire tropical maritime en fournissant des outils adaptés à cet objectif. La région géographique d'intérêt est celle des Antilles. Le climat de ces régions fait l'objet d'une réputation surfaite, car les températures, les pressions,

l'humidité et les Alizés sont d'une grande régularité. En fait, l'atmosphère est le plus souvent moite que sèche, le ciel le plus souvent couvert que dégagé et les pluies sont l'élément capricieux du climat.

Ces traits sont très différents de la dynamique atmosphérique des moyennes et hautes latitudes qui se caractérise plutôt par de brusques changements de temps qui jalonnent le climat des régions tempérées. Dans ce cas, il a été proposé une représentation conceptuelle de cette dynamique, en climat tempéré, par des régimes de temps, apparitions reconnaissables et récurrentes en altitude, au-dessus d'une zone géographique étendue, d'une situation météorologique rattachée à des caractéristiques bien précises du temps sensible et local.

Cette idée a été introduite en climatologie synoptique, car elle permet de bénéficier d'une représentation conceptuelle de la dynamique atmosphérique de ces latitudes. Elle est dans le prolongement de l'Oscillation Nord-Atlantique. Les régimes de temps permettent d'anticiper l'évolution des circulations et de représenter les grandes circulations en Europe, influencées par l'océan Atlantique et la mer Méditerranée. Ces deux surfaces maritimes modulent les contrastes de pression et de température du continent.

Qu'en est-il pour les Antilles ?

Dans cette partie du globe, ce sont les jeux d'actions de la ceinture des hautes pressions de l'Atlantique nord et l'influence de la zone intertropicale de convergence au sud, qui est susceptible de moduler les circulations des Alizés, de modifier leur structure, et amener de la variabilité au temps sensible. Il est donc nécessaire de s'intéresser aux régimes de temps à ces latitudes pour les identifier, caractériser leur récurrence et produire de la connaissance scientifique sur leurs comportements spatio-temporels. C'est l'un des objectifs de ce travail de thèse.

Dans cette région, l'année est traditionnellement divisée en deux saisons qui sont respectivement, et pour schématiser, une saison sèche (appelée "Carême") et une saison humide (appelée "Hivernage"). Toute personne vivant aux Antilles aura néanmoins participé à d'innombrables discussions au sujet de Carême qui semble de plus en plus long et sec, et d'un hivernage présentant

de plus en plus d'épisodes cycloniques. La présence d'une mer omniprésente peut-elle conduire par exemple à rendre l'influence de certains paramètres négligeable devant d'autres ?

Les régimes de temps que définissent ces centres d'action restent à ce jour très mal connus, bien que leur identification soit au cœur de toute réflexion sur la saisonnalité du climat dans nos régions et sur son évolution. Les techniques informatiques issues de l'Intelligence Artificielle et de la Reconnaissance des formes, visant de façon très générique à analyser des données, seront la base des outils que nous présenterons ici pour cette tâche.

Les répercussions de ces questionnements sont immenses, car elles peuvent donner du sens aux significations et liens physiques entre les apparitions en altitude et la répartition en surface de paramètres météorologiques à forts impacts comme par exemple les pluies extrêmes, les périodes de sécheresse et les vagues de chaleur. Ces impacts peuvent se traduire par des effets négatifs dans diverses activités socio-économiques allant de l'agriculture à la santé des personnes. Ils soulignent l'exposition et la vulnérabilité de cette zone géographique aux aléas atmosphériques.

1.2 Aspect pluridisciplinaire de ces travaux

Avant d'aller plus loin, nous souhaitons insister sur un point qui nous semble important : les travaux présentés ici relèvent d'une thèse de doctorat en informatique, réalisée au sein d'un laboratoire de physique, le LaRGE, co-encadrée pour la partie informatique par des membres d'un second laboratoire, le LAMIA.

Nos travaux se placent donc à l'intersection de la climatologie et l'informatique. Ce positionnement, dont nous pensons qu'il peut être extrêmement fécond, n'est pas sans poser un certain nombre de problèmes. En effet, ces deux disciplines se distinguent grandement dans leur méthodologie, laquelle est essentiellement due à des objectifs très différents. Trouver un point d'équilibre viable entre les deux est la caractéristique majeure des travaux présentée ici.

Il nous semble donc bénéfique de préciser un peu ces différences, pour mettre en évidence les choix que nous avons effectué et comprendre le plan de ce manuscrit. Nous prions donc le lecteur, quelle que soit sa communauté d'origine, de faire preuve d'indulgence pour les généralités qui suivent, mais qu'il semble néanmoins pertinent de rappeler ici.

1.2.1 La physique

La physique, discipline vieille d'à peu près 3000 ans, s'intéresse au monde qui nous entoure, car elle cherche à modéliser et expliquer les phénomènes naturels en proposant des lois, des principes et/ou issus de conceptualisation permettant d'évoluer vers des modélisations à partir d'observations. Pour cela, elle dispose d'outils de mesure et de techniques d'analyse de données. L'interprétation des résultats de ces analyses a ainsi permis le développement de différents modèles analytiques et numériques, dont la finesse et l'efficacité se sont accrues dans le temps.

Néanmoins, les physiciens savent que, par nature, leurs mesures sont au moins partiellement déficientes et que leurs outils d'analyse peuvent s'avérer imparfaits. Le 20e siècle et la mécanique quantique ont même remis en cause la capacité de certains modèles à simplement offrir une représentation fidèle du monde. La méthodologie en vigueur en physique le reflète très bien lors des publications des connaissances : Les chercheurs utilisent le triptyque, Introduction, puis matériel et méthodes, avant d'afficher les résultats observés. La discussion qui suit permet de les valider ou d'invalider la pertinence des modèles proposés par des observations indirectes des effets de ces derniers. Ils comparent également souvent leurs résultats à ceux présentés dans la littérature pour une co-validation partielle ou une potentielle remise en question des savoirs.

Il s'agit ainsi d'un véritable travail conceptuel de très longue haleine, marqué dans sa méthodologie par une modestie qui nous semble honorer ses acteurs, d'autant plus, encore une fois, au regard des formidables succès que cette démarche a permis d'obtenir.

1.2.2 L'informatique

À l'inverse, l'informatique et en particulier l'Intelligence Artificielle est une discipline beaucoup plus jeune, que l'on pourra estimer être née après la Seconde Guerre mondiale. C'est une discipline technologique, qui a pour objectif de fournir des outils efficaces à l'ensemble des autres sciences, ce qu'elle a fait avec des succès remarquables et remarqués.

Avec une approche beaucoup plus conquérante, les études se focalisent donc sur la mise au point de méthodes et sur le cheminement qui a conduit à leur élaboration. Pour mettre au point ces méthodes, les informaticiens partent souvent d'un cas d'école pour lequel ils trouvent une solution. Il définissent ainsi un vocabulaire qui fait totalement abstraction d'un quelconque domaine d'application. De même, la validation des méthodes est le fruit d'une réflexion ou ne figure idéalement quasiment aucune considération applicative.

1.2.3 Pour une approche pluridisciplinaire

Pour mieux expliciter ce qui précède, nous prendrons ci-dessous un exemple de traitement de données par ces deux communautés. Celui-ci est évidemment très caricatural, mais cadre tout à fait avec les travaux présentés dans ce manuscrit.

Imaginons qu'il s'agisse de regrouper des points (faire du clustering) dans un espace à 2 dimensions, correspondant à deux paramètres physiques, telles que la pression (x_1) et la température (x_2) d'un gaz, afin d'établir les grandes tendances de ces paramètres.

Cette situation est présentée figure 1.1, il s'agit d'un nuage de points non labellisés.

Les informaticiens, comme les physiciens disposent de techniques d'analyse de données pour affecter ces points à des groupes (les clusters). La figure 1.2 présente une sortie hypothétique de ces techniques. Chaque point se voit affecté à un groupe, les groupes étant représentés par des couleurs différentes.

Se pose alors la question de savoir si ces groupes sont valides ou pas ? Dans le cas présenté ici, la réponse semble évidente, car les points semblent effecti-

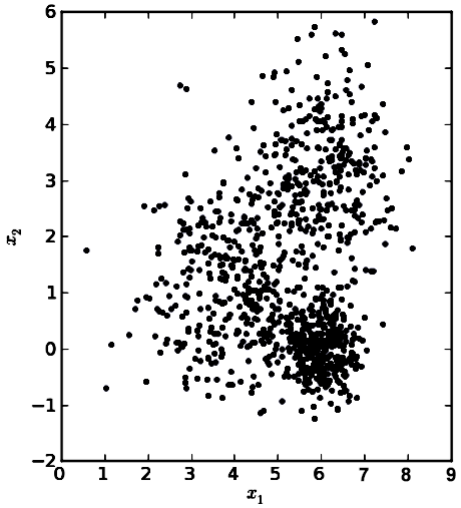


FIGURE 1.1 – Nuage de points non labellisés

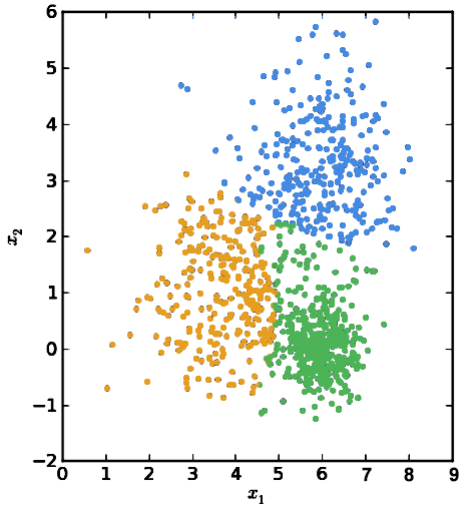


FIGURE 1.2 – Nuage de points labellisés.

vement bien groupés, de façon assez compacte autour de leur centre. Dans ce manuscrit en revanche, nous utiliserons des données de dimensions bien supérieures à deux, pour lesquelles il est impossible de visualiser le nuage de points dans son ensemble.

Selon nos études bibliographiques, voici très sommairement comment les deux communautés auraient tendance à justifier la qualité des groupements trouvés par l'algorithme de clustering.

Les physiciens pourraient par exemple n'observer que le centre des clusters

trouvés, tels qu'ils apparaissent en figure 1.3. Ils justifieront par exemple la qualité des clusters trouvés par le fait que les théories actuelles convergent vers un positionnement pertinent du centre "vert" autour du point (x, y) qui pourrait être représentatif d'un comportement moyen d'un certain type de gaz.

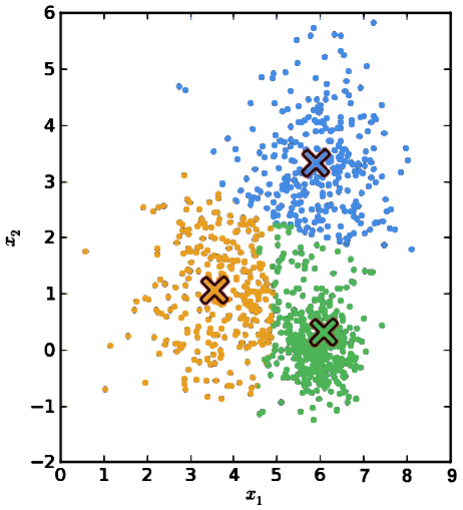


FIGURE 1.3 – Nuage de points labellisés avec centroïde.

Une telle approche réduit la perception de la dispersion des phénomènes autour de ces centres. Une expertise plus poussée des physiciens évaluerait par exemple cette dispersion en regardant comment ces points se répartissent lorsqu'on observe ce nuage de points sous l'angle du volume qu'il occupe. Et ils en concluront avec justesse que les groupements trouvés sont très pertinents, car ils peuvent prédire des comportements significatifs.

À l'inverse, des informaticiens en Intelligence Artificielle qui seraient dépourvus de contacts avec les physiciens, feraient des mesures de séparation et d'homogénéité des clusters, sans aucune considération sur ce qu'ils représentent. Aussi plaisante que soit cette approche, par sa généricité, elle pose également des problèmes. Si l'Intelligence Artificielle a connu de nombreux succès, elle a également engrangé de sérieux revers, par exemple lorsqu'elle néglige les éventuels biais présents dans ses bases de données, ou encore lorsqu'elle promettait au monde industriel, dans les années 70, de permettre de supprimer tout recours aux experts.

L'intelligence Artificielle ne fait, de façon générale, que mettre en évidence

des corrélations au sein des données. Elle ne peut en aucun cas assurer une causalité. Ceci est, selon nous, exclusivement du ressort des experts. C'est pourquoi nous pensons que ces approches pluridisciplinaires sont aussi importantes, pour peu que l'aspect pluridisciplinaire ne soit pas qu'un vernis plaqué sur des études pour qu'elles soient dans l'air du temps. La réelle confrontation des résultats selon les deux points de vue permet aux uns et aux autres de se remettre en question avec profit.

C'est en gardant ceci en tête que nous avons mis au point le plan qui va maintenant être présenté.

1.3 Plan du manuscrit

Après cette introduction, nous commencerons dans le chapitre 2 par expliciter précisément les objectifs de ces travaux concernant l'aide à l'identification de régimes de temps dans la région des Antilles. Nous y présenterons également les données utilisées et l'ensemble des méthodes d'Intelligence Artificielle qui pourraient être utilisés avant d'établir notre choix sur les deux méthodes de clustering (K-Means et CAH), déjà utilisées par les physiciens du climat.

Le chapitre 3 sera consacré à des études effectuées en début de thèse, sous la supervision quasiment exclusive de physiciens. On y verra comment ces techniques de clustering permettent d'obtenir des résultats intéressants, pour deux thématiques : l'analyse des circulations atmosphériques et océaniques avec une application pour le transport et les échouements des banc d'algues sargasses.

La suite du manuscrit prendra alors la forme d'une plongée de plus en plus profonde vers le point de vue des informaticiens. Le chapitre 4 verra le début de nos travaux sur l'identification des régimes de temps. On y trouvera une étude bibliographique dédiée à ce sujet, laquelle sera suivi de nos propres travaux sur le domaine. Notre principale contribution dans ce chapitre consistera à importer une mesure de qualité bien connue (l'indice de Silhouette) dans le domaine de l'informatique climatique. Si l'utilisation de l'indice de silhouette n'a rien de révolutionnaire pour évaluer la qualité de clusters, son application dans ce domaine à de nombreuses répercussions. En particulier, cela permet rapidement de comparer les deux algorithmes retenus. De plus, cet indice semble indiquer

que, sur les données traitées, les clusters constitués sont aux mieux médiocrement constitués (données de vent), voire totalement ineptes (données de pluie).

Dans ces deux cas, la pertinence des clusters évaluée par l'indice de silhouette est mise en regard de leur interprétation physique. Cette double analyse permettra de conclure que silhouette semble être un indice tout à fait efficace pour représenter la qualité des clusters, telle qu'elle peut être souhaitée par des physiciens.

La faible qualité des résultats obtenus dans le chapitre 4 nous a alors amenés à nous demander ce qui, au sein des données et du processus de clustering, pouvait rendre le problème aussi délicat. Le chapitre 5 présente nos réflexions sur ces points ainsi que les solutions que nous avons proposées. Cela consiste à remettre en cause le choix de L2 comme mesure de similarité entre les données, au profit d'une nouvelle mesure, nommée Expert Deviation (ED). Nous y montrerons sur quelles bases nous l'avons conçue (principalement une relaxation de la notion de positionnement spatial des structures recherchées). Nous montrerons alors que son introduction dans le processus de clustering permet d'améliorer de façon très notable les performances de l'analyse. Celles-ci seront également mesurées selon le double aspect de l'informatique et de la physique, montrant encore une fois la pertinence de l'indice de silhouette.

La fin du chapitre 5 sera sans doute assez surprenante pour le lecteur informaticien. Puisque notre objectif était de proposer des outils d'aide à l'identification des régimes de temps, et que nous l'avons partiellement atteint, il nous a semblé important d'y détailler ces résultats et leurs applications par les physiciens avec lesquels nous avons travaillé. L'objectif étant de présenter un certain nombre d'incitateurs permettant de montrer aux spécialistes du domaine le pertinence des clusters obtenus.

Enfin, la fin de ce manuscrit sera dédiée à une conclusion de ces travaux et aux multiples perspectives que nous leur envisageons.

Chapitre 2

Contexte et état de l'art

2.1 Introduction

Ce chapitre présente dans la section 2.2 le contexte général des travaux, il s'agira de présenter les données utilisées, la zone étudiée ainsi que leurs spécificités. En effet, nous verrons que ces données pourraient bien faire l'objet d'études d'analyses d'images bien qu'elles représentent une certaine complexité. Cependant, l'usage des méthodes classiques d'analyses d'images sera difficilement applicable en l'état.

L'état de l'art des méthodes classiques du domaine de l'apprentissage automatique appliquées à l'analyse climatique est présenté dans les sections 2.3 et 2.4. Nous allons donc présenter des généralités sur ces méthodes puis faire une analyse nous permettant de choisir les algorithmes à utiliser. Nous verrons que les méthodes de clustering seront celles que nous retiendrons pour effectuer cette étude, bien que plusieurs méthodes pourraient être considérées. Ce choix sera principalement motivé par les spécificités et les contraintes liées à l'étude.

La section 2.5 correspond à la présentation de la méthodologie classique appliquée pour ce type d'étude en climatologie. Il s'agira de présenter nos objectifs détaillés, les méthodes, les algorithmes et les paramétrisations retenus pour les premières applications. Cette méthodologie classique sera appliquée par la suite dans le chapitre 3.

La section 2.6 conclut ce chapitre en donnant les orientations et approches

scientifiques que nous avons retenues pour nos expérimentations.

2.2 Analyse contextuelle

Le développement continu de la recherche et de l'ingénierie en informatique durant ces dernières décennies a contribué à un grand nombre d'avancées scientifiques dans plusieurs domaines d'étude et de recherche. Cela est vraisemblablement dû au fait que les théories et méthodes de traitement informatique sont conçues pour être applicatives et adaptatives. Elles peuvent donc être utilisées pour toutes sortes de problématiques ainsi que sur tous types de données numériques. Parmi les nombreux axes de recherche compatibles à l'application de méthodes de traitements automatisés, l'analyse issue d'observations et/ou de modèles climatiques est la thématique retenue pour les travaux présentés dans ce document.

Le climat désigne les caractéristiques statistiques, critères de position et de dispersion, calculées sur au moins trente ans, d'observations de paramètres météorologiques en un lieu donné. Le climat d'un lieu est affecté par sa latitude, par l'étendue de la surface terrestre et la topographie, ainsi que par les masses d'eau océaniques ou continentales voisines de même que les courants qui s'y produisent. En somme, le climat d'une région est l'état général du système climatique à cet endroit pendant une période donnée.

Parmi les variables météorologiques couramment mesurées, on retrouve souvent des données atmosphériques locales et/ou régionales telles que la température, le géopotentiel, l'humidité, la pression atmosphérique, le vent et les précipitations. L'analyse de ces données permet de classer les climats et de mettre en évidence d'éventuelles tendances.

Les régimes de temps peuvent également être caractérisés grâce à une classification des configurations spatio-temporelles synoptiques [1] récurrentes. Ces dernières sont liées aux structures atmosphériques et à leurs interconnexions. En analysant les multiples configurations atmosphériques des masses d'air à l'aide de leurs propriétés physiques, il est donc possible de regrouper les confi-

1. de grande échelle

gurations similaires afin d'établir des tendances.

La classification des configurations atmosphériques des masses d'air est automatisable et donc compatible avec un certain nombre de méthodes de traitements informatiques. Il faut également noter que les données permettant de caractériser ces masses d'air sont numériquement quantifiables. Ces données sont multiples et correspondent aux paramètres physiques et météorologiques, elles sont définies et présentées dans la sous-section suivante.

Ensuite, nous ferrons un point sur les spécificités des données afin de mettre en évidence leur complexité. D'une part, cette analyse nous permettra de mieux choisir les méthodes applicables à ces données. D'autre part l'expérience que l'on en tirera nous permettra de proposer par la suite (cf. Chapitre 5) une nouvelle méthodologie d'analyse plus sensible à la nature des données.

Enfin, nous ferons un point sur les spécificités de la zone d'étude et l'on verra par la suite qu'il est tout à fait intéressant d'intégrer ces spécificités au processus d'identification automatique des configurations atmosphériques récurrentes. En effet, cela fera partie des principaux apports de cette thèse.

2.2.1 Données climatiques

De grandes quantités de données climatiques observées ou simulées de différents types sont mises à disposition de la communauté scientifique, offrant un large panel de déclinaisons pour les recherches scientifiques futures et pour l'exploration de données par l'apprentissage machine [59]. Ces données sont principalement des variables (ou champs) météorologiques telles que la température, les précipitations ainsi que la direction et l'intensité du vent. Elles portent également un référencement lié à leur position géographique (donnée par les longitudes et latitudes), à leur niveau d'altitude (ou d'isobare), ainsi qu'un identificateur temporel (une date).

Les données climatiques sont donc par nature spatio-temporelles puisqu'elles sont associées à une localisation spatiale, mais également à un instant ou une période définie dans le temps (Fig 2.1). Il est donc possible de chercher à établir des corrélations spatio-temporelles dans un ensemble de données liées à

un paramètre météorologique mesuré ou modélisé. Cette possibilité ainsi que l'abondance des données favorisent l'usage de méthodes automatisées, mais il faut noter tout de même que la complexité des données peut contribuer à fausser l'interprétation des résultats obtenus suite aux traitements. Nous verrons comment cela se traduit dans le chapitre suivant.

Dans la sous-section suivante, nous présentons les différents types de données climatiques utilisées dans ces travaux de thèses.

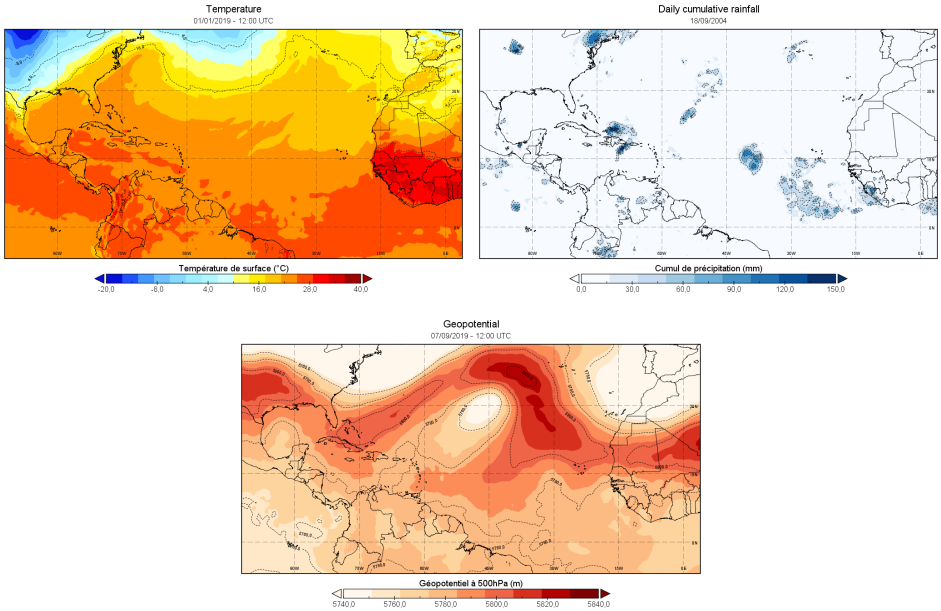


FIGURE 2.1 – Exemple de données climatiques (température de surface, cumul de précipitations et géopotentiel à 500hPa), dans la zone d'étude.

2.2.1.1 Types des données

Les données climatiques peuvent provenir de multiples méthodes d'observations et de collectes permettant d'analyser leurs variations à des échelles spatio-temporelles définies. Elles peuvent également être simulées par des modèles numériques intégrant ou non des données réelles et reproduisant la dynamique physique de l'atmosphère. Les principales catégories de données climatiques utilisées dans ces travaux sont présentées ci-après.

2.2.1.1.1 Données d'observations

Ces données que l'on peut qualifier comme étant "réelles" sont recueillies à

l'aide d'instruments de mesure et de capteurs spécifiques montés sur des dispositifs appropriés, formant des stations de mesures, ensembles ponctuels maintenus par les diverses organisations météorologiques mondiales. Ces données doivent respecter un cahier des charges précis, publié par l'Organisation Météorologique Mondiale (OMM). Elles sont utilisées par de nombreux organismes de recherche et dans certains cas peuvent être publiques.

Les scientifiques du monde entier utilisent principalement des mâts de mesures instrumentés, des ballons météorologiques, des capteurs embarqués dans des avions et des satellites, afin de collecter les valeurs des paramètres physiques à différentes altitudes leur permettant d'avoir une vue trois dimensions des phénomènes climatiques. D'ailleurs, récemment en septembre 2017, des avions spéciaux équipés d'instruments de mesure ont traversé à plusieurs reprises l'ouragan Irma afin collecter des données pour le compte de l'Agence Américaine d'Observation Océanique et Atmosphérique (NOAA). Les données satellites sont quant à elles obtenues par l'analyse des spectres de couleurs et par celle de la rayonnance des masses nuageuses. Ces valeurs sont converties par des méthodes informatiques grâce à des équations physico-chimiques, définies par les spécialistes, en valeurs numériques caractérisant un paramètre météorologique atmosphérique et une zone ciblée. Ce type de mesure permet de produire des données pour toute la planète et notamment sur les immenses zones océaniques que l'on ne peut pas couvrir de capteur.

Plusieurs organismes de recherche recueillent des données brutes provenant des centres météorologiques régionaux du monde entier afin de concevoir des bases de données. Ces bases de données sont mises à disposition des scientifiques sur internet via des serveurs ou des sites de téléchargement (i.e. Unisys Weather [2], Wyoming weather [3], etc.) permettant à tout le monde d'y accéder grâce de multiples protocoles de connexion (tels que HTTP, FTP, torrent, etc.). Dans les travaux de thèse présentés dans ce document, nous utilisons ce type de données pour améliorer l'apprentissage automatisé en les croisant avec des observations au sol. Les données de précipitations à 6 min des stations Météo France Guadeloupe et Martinique sont comparées aux cumuls de précipitations mesurés par satellite produit dans le cadre du projet TRMM

2. http://weather.unisys.com/ → 31/12/2018
3. http://weather.uwyo.edu/upperair/sounding.html → 01/01/2020

(Tropical Rainfall Measuring Mission) de la NASA (National Aeronautics and Space Administration) [28], afin d'améliorer l'apprentissage automatique des précipitations dans les Petites Antilles (cf. Chapitre 5).

2.2.1.1.2 Sorties de modèles climatiques

Les modèles climatiques sont des simulations numériques du système climatique basées sur les équations de la physique auxquelles sont adjointes et les interactions entre les composantes telles que l'atmosphère, l'océan, la glace de mer, la surface terrestre, la végétation, les calottes glaciaires, les aérosols atmosphériques ainsi que les cycles du carbone. En d'autres termes, un modèle climatique est une modélisation mathématique de multiples paramètres physiques et météorologiques caractérisant le climat pour une zone géographique donnée [13, 54].

Du point de vue informatique, il s'agit en fait d'un ensemble de systèmes d'équations différentielles basées sur les lois fondamentales de la physique, du mouvement des fluides et de la chimie. Pour générer un modèle, les scientifiques divisent la planète en une grille tridimensionnelle, appliquent les équations de base et évaluent les résultats. Les modèles atmosphériques calculent les vents, la température, le transfert de chaleur, le rayonnement, l'humidité relative et l'hydrologie de surface dans chaque grille puis ils évaluent les interactions avec les points voisins.

Les modèles climatiques sont utilisés à des fins diverses, de l'étude de la dynamique du système terrestre passé et actuel, aux projections du climat futur. Ces modèles peuvent également être qualitatifs (c'est-à-dire non numériques) ou encore des simulations, largement descriptives, de futurs possibles. Il sont nombreux et varient plus ou moins en complexité. Les plus simples permettent d'avoir une compréhension globale du système climatique ; les plus complexes permettent d'approcher la réalité :

- **Les modèles couplés océan-atmosphère :** Ils sont constitués de plusieurs modèles, un modèle d'océan, un modèle d'atmosphère, un modèle de glace de mer, un modèle représentant les continents (végétation, ruissellement, etc.) qui échangent leurs informations (couplage). Ces modèles

sont utilisés aujourd'hui.

- **Les modèles du système Terre :** Ces modèles sont le développement des modèles couplés océan-atmosphère combinés à la simulation des cycles biochimiques et géochimiques. Ils constituent aujourd'hui les outils les plus complets pour la réalisation des projections climatiques pour lesquelles les rétroactions liées aux cycles biochimiques et géochimiques sont importantes.

  - **Les modèles du système Terre de complexité intermédiaire :** Ces modèles incluent les composantes des modèles du système Terre, mais souvent de façon idéalisée, ou à faible résolution, afin d'être moins coûteux en puissance de calcul. Ils permettent l'étude de questions spécifiques, par exemple la compréhension de certains processus de rétroactions.

- **Les modèles régionaux :** Ils sont similaires aux modèles précédents, mais leur domaine spatial ne couvre qu'une partie du globe terrestre. Leur domaine étant plus petit, il est possible d'avoir une meilleure résolution spatiale et temporelle (à maille fine) pour un même coût de calcul par rapport à un modèle global. Les informations, aux frontières sont en général fournies par les modèles globaux.

Les modèles climatiques sont entièrement configurables, chaque année les centres de recherche internationaux proposent des réglages qui simulent encore mieux le climat. Nous avons retenu des modèles du système Terre, car ce sont les plus utilisés par la communauté scientifique mondiale. Parmi eux, nous nous sommes intéressés à une sous-catégorie de modèle climatique intégrant des données d'observations. Ces modèles produisent des données dites de"réanalyses" que nous présentons ci-après. Ces données servent, dans cette thèse, à fournir une base traitement pour les méthodes informatiques.

2.2.1.1.3 Données de réanalyse

Les sorties de réanalyses météorologiques proviennent d'un modèle climatique

d'assimilation de données météorologiques qui vise à intégrer des données d'observations historiques sur une période prolongée, en utilisant un seul schéma d'assimilation (ou d'analyse) cohérent (Figure 2.2). En d'autres termes, il s'agit de faire intervenir les données d'observations à un intervalle défini dans le processus de calcul des données de réanalyse.

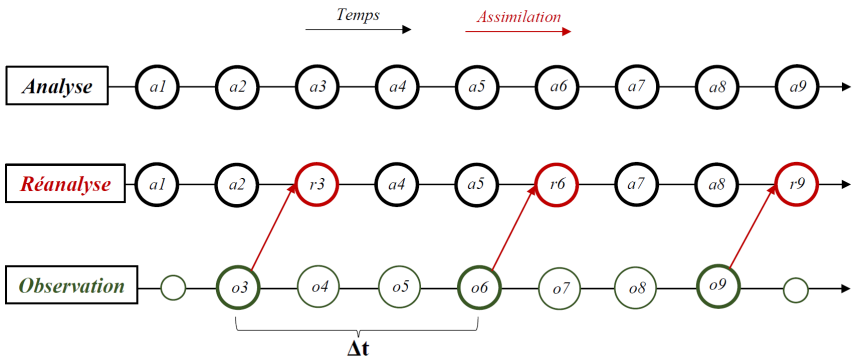


FIGURE 2.2 – Schéma d'assimilation des données d'observations aux calculs des données de réanalyses avec un pas de temps donné, Δt.

Dans les travaux de thèse présentés dans ce document, nous utilisons principalement ce type de données pour faire de l'apprentissage sur le climat passé. Les données des projets ERA Interim et ERA-5 du centre de recherche scientifique ECMWF (*European Centre for Medium-Range Weather Forecasts*) sont les principales données que nous avons utilisées [20, 9].

Il faut noter que les données modélisées sont souvent très denses comme le montre l'exemple de la figure 2.3 où l'on dispose de 35 années × 12 mois × 31 jours × 4 relevés × 360 longitudes × 180 latitudes × 37 altitudes pour un seul paramètre météorologique soit 3 374 784 000 valeurs numériques. Pour rappel, une étude climatique (comme introduit en Section 2.2) est considérée comme viable quand elle repose sur l'analyse d'au moins trente années de données.

2.2.1.2 Formats des données

Au-delà de la provenance et du type, des données climatiques se posent également la question de leur format de structuration et de stockage. Deux formats sont communément utilisés pour stocker les données géospatiales : les formats matriciels (*raster*) et vectoriels.

2.2.1.2.1 Données matricielles *(ou raster)*

Elles utilisent une matrice de zones carrées pour définir l'emplacement des éléments. Ces carrés, également appelés pixels, cellules et grilles, sont de taille uniforme et les détails qui peuvent être conservés dans l'ensemble de données. Comme les données matricielles représentent des zones carrées, elles décrivent des intérieurs plutôt que des limites, comme c'est le cas avec les données vectorielles. Les données matricielles sont bien adaptées à la capture, au stockage et à l'analyse de données telles que l'altitude, la température ou encore la couverture végétale du sol variant continuellement d'un endroit à l'autre. Les formats de données raster sont également utilisés pour stocker des images aériennes et satellites. Ce sont ces données qui sont également utilisées pour les traitements informatiques. Dans nos études, elles sont utilisées pour stocker les données modélisées de réanalyses (Fig 2.1).

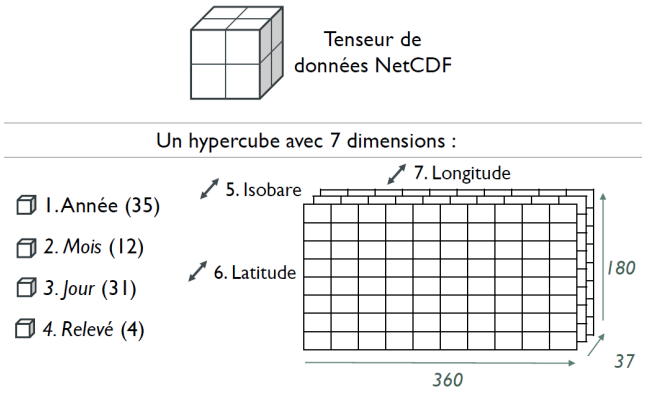


FIGURE 2.3 – Structuration d'un fichier netcdf issu de ERA-Intérim stockant les données d'un paramètre météorologique de 1979 à 2014.

Extensions de fichiers associées : Les formats de données raster couramment rencontrés dans la recherche sur le climat se répartissent en trois catégories génériques : GRIB, netCDF et HDF. Tous ces formats sont portables et auto descriptifs. Les fichiers auto descriptifs peuvent être examinés et lus par un logiciel approprié sans que l'utilisateur connaisse les détails structurels du fichier. De plus, des informations supplémentaires sur les données, appelées "métadonnées", peuvent être incluses dans le fichier. Les métadonnées typiques peuvent comprendre de l'information textuelle, sur le contenu et les unités de chaque variable (e.i. "humidité spécifique" en "g/kg"), ou de l'information numérique décrivant les coordonnées (e.i. date, niveau, latitude, longitude) qui

s'appliquent aux variables du fichier [71]. Un exemple de structuration d'un fichier netcdf issu de ERA-Intérim est présenté en figure 2.3.

2.2.1.2.2 Données vectorielles

Les données vectorielles utilisent les coordonnées X et Y pour définir l'emplacement des points, des lignes et des zones (polygones) qui correspondent à des éléments cartographiques tels que la localisation d'une station pluviométrique ou d'un capteur thermique. Les données vectorielles ont donc tendance à définir les centres et les bords des éléments. Elles sont excellentes pour la capture et le stockage de détails spatiaux. Dans nos études, elles ont été utilisées pour stocker des données mesurées en un point précis telles que la localisation des bancs d'algues sargasses (cf. sous-section 3.3).

Extensions de fichiers associées : Le format de données vecteur le plus utilisé est le shapefile. Développé par ESRI *(Environmental Systems Research Institute)* en 1990, le fichier shapefile est le format privilégié pour stocker un vecteur de données géospatiales. Utilisé généralement dans les Systèmes d'Informations Géolocalisées (SIG), il repose sur une description géométrique(point, ligne, polygone) de l'élément modélisé. Le fichier portant l'extension .shp est toujours accompagné de deux autres fichiers d'extensions : un DBF contenant les données attributaires relatives aux objets contenus dans le shapefile ; et un SHX stockant l'index de la géométrie de tous les éléments. Dans le domaine de la recherche, les informations de localisation sont souvent stockées dans des fichiers textuels (i.e. texte, CSV, etc.).

2.2.2 Spécificités des données de l'étude

Les données raster que nous utilisons dans nos études sont issues de sorties de modèles de réanalyses (tels que ERA-Interim, Mercator Ocean et ERA-5) ainsi que de mesures satellites (telle que TRMM). Ces données décrivent l'évolution au cours du temps d'une vingtaine de paramètres météorologiques (température, humidité, intensité et direction du vent, etc.) sur trente-sept altitudes disponibles, pour une zone géographique définie.

Nos données se matérialisent donc par des hypercubes dont les mailles portent les valeurs à un instant t d'un paramètre météorologique à une altitude (ou

isobare), une longitude et une latitude selon un maillage donné (ex. Fig 2.4).

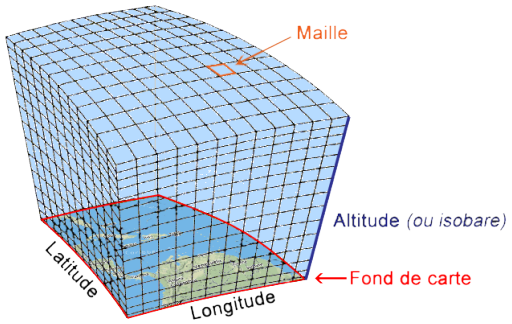


FIGURE 2.4 – Représentation du maillage : le support de l'information pour format raster, avec le fond de carte (en rouge), les différentes altitudes (en bleu) et la maille (en orange).

La résolution du maillage dépend du modèle utilisé. Ainsi, plus la résolution du modèle est faible, plus le maillage sera resserré. Tous les modèles de climatologie utilisent des mailles suffisamment grandes pour éliminer une partie des fluctuations purement locales (qui relèveraient de la météorologie), et suffisamment petites pour conserver une finesse d'analyse acceptable.

Nous avons commencé nos expériences avec les données ERA-Interim de résolution 0.5° puis nous sommes passés aux données ERA-5 (sortie en 2018) de résolution 0.25° afin d'avoir des données plus précises. Notons que même dans ce dernier modèle, l'île de Guadeloupe en son intégralité occupe seulement 4 mailles sur les 19 089 de l'ensemble de la zone d'étude.

Par exemple avec le jeu de données ERA-5, notre domaine d'étude (cf. Figure 2.9B) représente une image de 101x189 pixels soit 19 089 valeurs par niveau d'altitude. De ce fait, en comptabilisant tout cela, l'on dispose de 706 293 mailles (couvrant notre zone d'intérêt) pour un paramètre à un instant t. Ainsi si l'on considère les vingt paramètres (cf. Fig 2.5) l'on arrive rapidement à plus de 14 millions de valeurs pour un instant t.

Pour nos travaux nous avons choisi les données journalières, puisque celles horaires représenteraient une granularité trop fine pour une étude sur le climat. En effet, nous cherchons à caractériser des tendances générales et bien marquées dans les structures atmosphériques. L'évolution horaire pourrait gé-

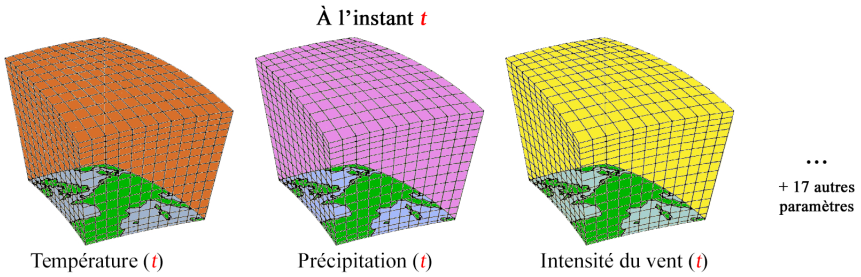


FIGURE 2.5 – Représentation des données à disposition à un instant t.

nérer du bruit non pertinent pour ce type d'étude. A contrario, l'on pourrait utiliser des données mensuelles, cependant celles-ci ne nous permettraient pas d'analyser les régimes de temps (cf.Chapitre 4) qui sont des tendances de configurations journalières.

De ce fait avec des données journalières, l'on dispose de 14 millions de valeurs par jour. Si l'on étudie un paramètre météorologique sur quarante ans, cela représente environ 14 600 jours qu'il faudra multiplier par 14 millions pour avoir le nombre de valeurs. C'est gigantesque !

C'est au sein de cet amas de mailles que nous souhaitons identifier des structures caractéristiques du climat. Même avec une granularité spatiale et temporelle adaptée à la climatologie, le volume des données pose deux problèmes :

- celui du temps de calcul, qui augmente a minima linéairement avec le nombre de données utilisées.
- celui du rapport signal sur bruit. Les éventuelles structures pertinentes (le signal) risquent d'être masquées par les fluctuations présentes (le bruit). Plus les données utilisées sont nombreuses, plus ce rapport baisse.

C'est la raison pour laquelle nous avons effectué une sélection dans les données afin d'arriver à des proportions plus abordables pour nos premiers travaux. Nous avons décidé, pour chacune de nos études, de considérer les valeurs journalières d'un paramètre météorologique à une pression (ou isobare) définie. Ainsi, la variable journalière sera représentée par une image dont chaque pixel porte la valeur du paramètre météorologique considéré à une altitude et une localisation (longitude et latitude) donnée.

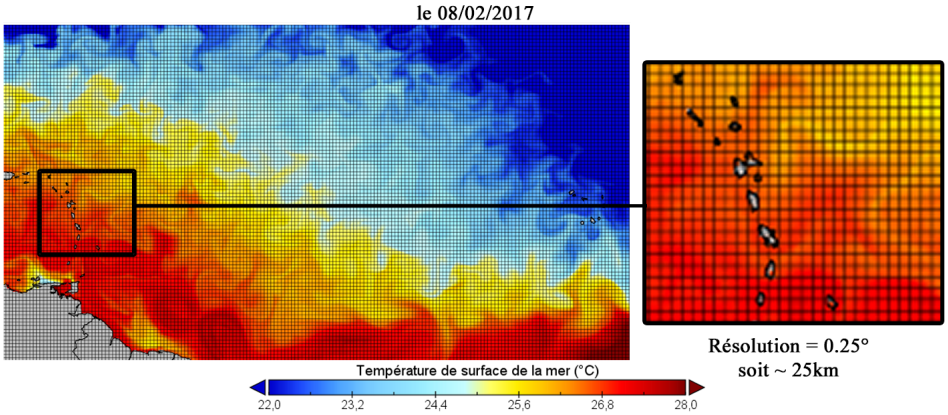


FIGURE 2.6 – Représentation d'une donnée journalière : exemple avec la température de surface de la mer provenant du projet ERA-5 (le 08/02/2017).

Les réanalyses obtenues suivent une dynamique propre aux phénomènes naturels et aux structures atmosphériques interagissant dans la région d'étude. Elle est supposée reproduire l'existence de corrélations spatio-temporelles entre les données journalières, mais également celles entre les composantes (mailles ou pixels). La recherche et la caractérisation des configurations spatio-temporelles récurrentes représentent un véritable défi. En effet, les structures atmosphériques qui les composent sont en perpétuel mouvement et de plus leur forme est extrêmement variable.

Dans le domaine de l'analyse d'image (ou *Image Processing*) [79, 46], généralement il est possible caractériser (ou classer) une image à partir les formes et motifs qui la composent. Prenons l'exemple de la détection de visage (Fig. 2.7) propre au domaine de la reconnaissance faciale, où l'enjeu est de localiser et identifier dans une image un ou plusieurs visages. Ces derniers peuvent se trouver n'importe où dans l'image avec des orientations variables et de possibles déformations liées à l'état émotionnel des figurants. Bien qu'il existe une multitude d'états possibles pour un visage, ce dernier dispose d'une logique structurelle exploitable par des méthodes informatiques. En effet, les éléments du visage (œil, bouche, nez, sourcils, etc.) ont une organisation spatiale permettant d'identifier les proportions et l'orientation d'un visage dans ce type d'étude.

Par exemple, grâce aux positionnements du nez et de la bouche l'on peut déterminer de l'axe vertical (y) du plan du visage, l'axe horizontal (x) est quant

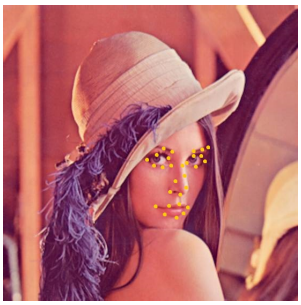
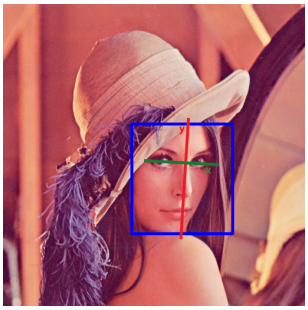
(a)  (b) 

FIGURE 2.7 – Exemple de détection de visage par analyse d'image : (a) détection des éléments du visage (en jaune), (b) le visage détecté (en bleu), l'axe horizontal du plan du visage (en vert) et l'axe vertical (en rouge).

à lui obtenu grâce à l'alignement des yeux (Fig. 2.7). Pour effectuer ce type d'analyse, les scientifiques spécialistes du traitement de l'image utilisent généralement dans des combinaisons de classifieurs permettant d'identifier les éléments du visage puis d'en déduire son état [46].

Cette cohérence structurelle pourrait bien exister dans les données climatiques dont nous disposons. Un premier problème provient du fait que nous cherchons des structures dont la forme et l'intensité précises sont beaucoup plus fluctuantes que celles des visages. En revanche, contrairement aux visages, leur positionnement au sein des images est extrêmement important, car il traduit l'existence de phénomènes physiques localisés. Notons que si la localisation de structures est importante, elle ne doit pas être établie de façon indûment précise.

On cherchera donc à identifier des structures telle qu'"une zone de forte précipitation plus ou moins circulaire dans le quart nord-ouest de la région étudié". La même zone, située à une autre position sur la carte représenterait vraisemblablement une tout autre réalité physique. Une zone de précipitation très longiligne dans cette zone serait elle aussi sans doute de nature différente. Nous cherchons donc à mettre en évidence l'existence de structures "à peu près semblables" par leur forme et leur positionnement au sein d'un jeu de données d'une taille très grande.

Comme on peut le voir, le problème est très mal défini. Ce qui compliquera un peu plus les choses, c'est que l'on ne sait pas par avance quelles configurations

atmosphériques nous souhaitons détecter. On pourrait ainsi envisager qu'une configuration atmosphérique soit "une zone de précipitation circulaire au Nord Ouest couplé à une zone longiligne de basses pressions dans le Sud".

Par exemple, la figure 2.8 montre les cumuls de précipitations mesurés par satellite de deux jours consécutifs (a) et (b). Bien que les précipitations ne représentent pas directement les structures atmosphériques, elles permettent tout de même de suivre leurs interactions (en rouge) et leur positionnement (en orange). Les jours (a) et (b) pourrait bien représenter une même configuration spatiale récurrente, ici illustrée sur un champ de précipitations.

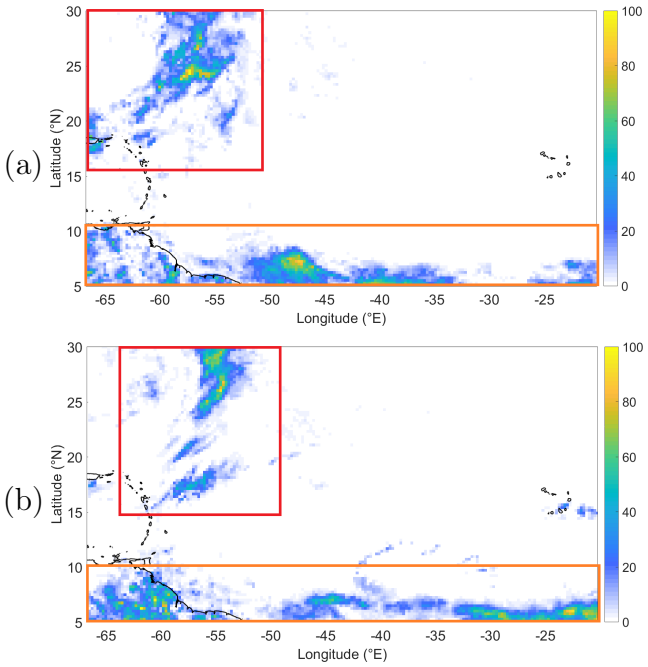


FIGURE 2.8 – Exemple de deux jours consécutifs dont le cumul de précipitation a été mesuré par satellite (TRMM), interaction entre structures atmosphériques (en rouge, en haut), il s'agit de la rencontre d'un front froid et des ondes d'est, positionnement de la Zone Intertropicale de Convergence (en orange, en bas).

En somme, nous concluons qu'une identification et un suivi des structures atmosphériques, par de leurs différentes formes grâce à des méthodes classiques d'analyse d'image sont possibles, mais laborieux. Cependant une adaptation aux spécificités de l'étude est tout à fait envisageable (cette dernière sera proposée dans le chapitre 5) et devrait être plus abordable. Nous avons donc décidé

de nous concentrer, dans un premier temps, sur les méthodes d'analyses automatiques présentées dans la bibliographie et utilisées par les spécialistes de ce type d'étude.

2.2.3 Spécificités de la zone d'étude

L'essentiel de nos travaux concerne la région comprenant la Caraïbe, l'Amérique centrale, une partie de l'océan Atlantique et de l'Afrique de l'ouest (Fig 2.9). La zone d'étude s'étend de 0 à 30°N en latitude et de -100 à -10°E en longitude. Selon les experts, il s'agit d'une zone subtropicale dont la variabilité atmosphérique est due à sa proximité avec l'équateur et la grande surface océanique dont elle est composée. De plus, le climat dans cette zone est influencé par des centres actions atmosphériques bien établis telles que la Zone Intertropicale de Convergence (ZIC) et l'anticyclone des Açores. Il est également important de noter que cette zone est régulièrement traversée par des phénomènes cycloniques. Ces derniers ont une trajectoire difficilement prévisible et sont une autre source de variabilité atmosphérique.

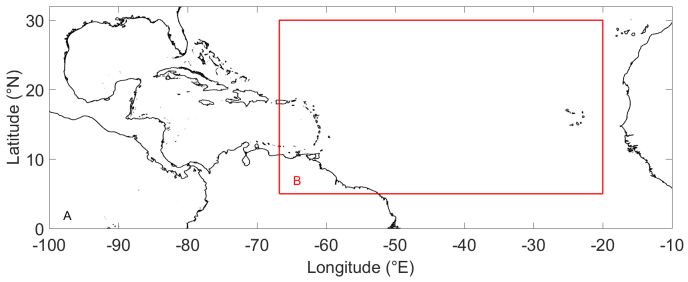


FIGURE 2.9 – A Domaine global - Amérique centrale, Caraïbe, centre de l'océan Atlantique et Afrique de l'ouest, B Domaine local - Arc des petites Antilles, une partie de l'Amérique du Sud et de l'océan Atlantique.

Cette situation géographique représente une difficulté supplémentaire, puisque jusqu'alors les experts du domaine cherchent toujours à comprendre la dynamique et donc prévoir la climatologie de cette zone. De ce fait, les spécialistes du domaine doivent faire appel à d'autres champs de compétences, dont l'expertise informatique. C'est également de cette manière qu'a été abordée la problématique de la présente thèse.

2.3 Informatique climatique

L'informatique climatique fait généralement référence à toute recherche combinant la science de l'analyse du climat à l'usage d'approches statistiques, de méthodes d'apprentissage machine et d'exploration de données. Ce terme apparaît pour la première fois dans le chapitre 4 du livre *Computational Intelligent Data Analysis for Sustainable Development* [59]. Les auteurs posent alors les bases de ce domaine de recherche, en expliquant l'intérêt d'utiliser l'apprentissage automatique pour analyser des données climatiques. Il réside dans le fait que l'emploi de méthodes d'apprentissages automatiques permet d'extraire des informations pertinentes, mais imperceptibles par l'homme et donc d'améliorer notre compréhension des structures et phénomènes liés au climat.

La section suivante vise à faire l'état de l'art des méthodes d'apprentissages automatiques ainsi que les raisons qui nous ont poussées à les utiliser dans ces travaux de thèse. Pour commencer, nous allons donc définir et présenter les concepts et méthodes propres à l'apprentissage automatique. Puis, au fur et à mesure, nous allons écarter celles qui ne sont pas appropriées à ce type d'étude et les méthodes de clustering seront retenues pour les premières applications.

2.4 État de l'art

2.4.1 Introduction

Dans cette section, nous faisons l'état des lieux des principales méthodes d'apprentissages automatiques utilisées en informatique, en précisant celles compatibles à notre sujet d'étude, celles que nous avons retenues ainsi que les raisons pour lesquelles nous les avons utilisées dans cette thèse.

2.4.2 L'apprentissage automatique

L'apprentissage automatique ou machine (*Machine Learning* en anglais) est un sous-ensemble de l'intelligence artificielle [15, 6]. Il repose sur l'usage d'algorithmes et de modèles statistiques par des systèmes informatiques (ou des ordinateurs). Cela dans le but d'effectuer une ou plusieurs tâches spécifiques sans utiliser d'instructions explicites, mais en se basant plutôt sur des modèles

et des inférences [30].

En d'autres termes, le *machine learning* donne aux machines la capacité d'apprendre à partir des données et d'améliorer leurs performances à résoudre des tâches sans être explicitement programmées pour cela. Il existe plusieurs types d'apprentissage automatique (Figure 2.10). Les sous-sections suivantes présentent les deux types d'apprentissages automatiques plus utilisés en recherche scientifique.

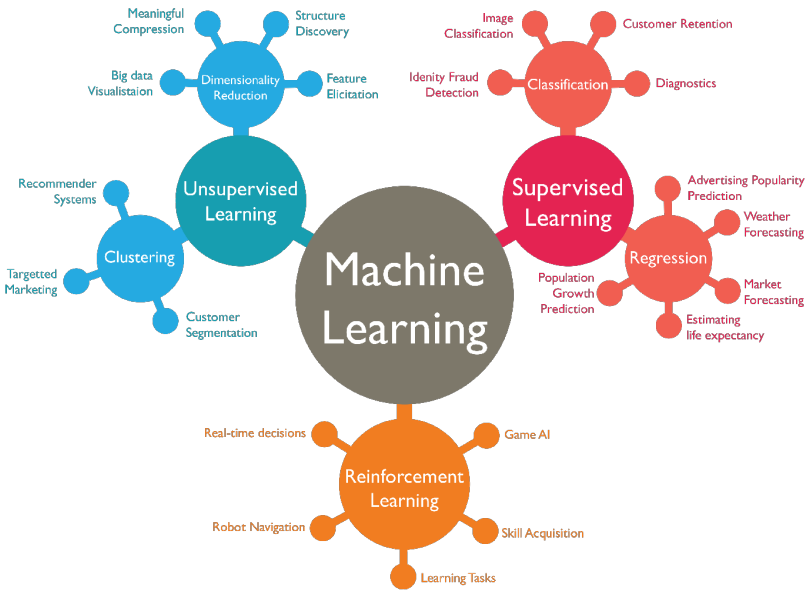


FIGURE 2.10 – Apprentissage automatique ou *Machine learning*.

2.4.3 Apprentissage supervisé

En l'intelligence artificielle et fouille de données, la méthode d'apprentissage supervisé consiste à déterminer et apprendre une fonction de prédiction à partir d'exemples labellisés. Les exemples labellisés constituent une base d'apprentissage, et la fonction de prédiction apprise peut aussi être appelée "hypothèse" ou "modèle". On suppose cette base d'apprentissage représentative d'une population d'échantillons plus large et l'objectif des méthodes d'apprentissages supervisés est donc de généraliser correctement, c'est-à-dire que la fonction apprise au cours de l'apprentissage produit des prédictions correctes sur des données non présentes dans l'ensemble d'apprentissages [58, 30].

Selon les méthodes existantes, on distingue généralement les problèmes de régression, des problèmes de classement. Ainsi, on considère que la prédiction d'une variable quantitative induit une régression tandis que les problèmes de prédiction d'une variable qualitative sont résolus par classification [15, 6]. En somme, les méthodes d'apprentissage supervisé permettent de concevoir une intelligence grâce aux connaissances préétablies concernant le jeu de données.

Il existe de nombreuses méthodes d'apprentissages supervisés [15], dont les machines à vecteurs de support (SVM) [16, 37] les arbres de décisions [66, 7], les modèles bayésiens naïfs ainsi que les réseaux de neurones artificiels [55, 5]. D'ailleurs, nous allons aborder plus en détail les réseaux de neurones dans la sous-section 2.4.3.1.

Pourtant, dans les études présentées dans le chapitre 4 dans cette thèse aucune de ces méthodes n'a été utilisée. Pourtant, dans la plupart de nos expériences, leur usage pourrait paraître évident. Notamment pour celles développées dans le chapitre 3 où l'on a d'une part des données brutes et d'autre part une ou des situations qu'il faut identifier.

Prenons la deuxième étude du chapitre 3, où il s'agit d'identifier les courants océaniques favorables aux échouements de bancs d'algues sargasses sur les côtes de des Petites Antilles. Il paraît évident pour un scientifique d'utiliser des méthodes d'apprentissages supervisés afin de construire automatiquement le modèle statistique associé à l'échouement d'algues sur les côtes à partir des configurations de courants océaniques.

Cependant, ce n'est pas notre objectif. Nous souhaitons tout d'abord identifier les configurations de courants océaniques existantes voire bien établies, puis analyser la dynamique spatio-temporelle de l'évènement échouement de bancs d'algues sargasses par rapport à ces dernières. Ce type d'analyse semble plus approprié aux méthodes d'apprentissages non supervisés présentées en sous-section 2.4.3.1.

2.4.3.1 Réseaux de neurones artificiels

Actuellement, les réseaux de neurones artificiels sont utilisés de façon très courante pour toutes sortes de tâches liées à l'intelligence artificielle. Il paraît tout à fait pertinent de chercher à utiliser les réseaux de neurones dans une thèse d'informatique, pourtant ce n'est pas l'approche que nous avons retenue pour nos travaux. En effet, ce type de méthode semble plus appropriée aux problèmes de classification supervisée, où l'on dispose d'une base d'apprentissage. Pour cela, il faudrait au moins avoir une idée précise de ce que l'on cherche à identifier. Dans notre cas, il s'agit de configurations spatiales récurrentes, cependant nous n'avons aucune idée de ce à quoi elles pourraient ressembler.

Il existe néanmoins des modèles de réseaux neurones permettant de faire de l'apprentissage non supervisé, par exemple les auto-encodeurs [3, 50]. Ces derniers s'organisent en blocs de couches successives, le premier bloc est constitué de couches réduisant au fur et à mesure les dimensions de l'information puis dans le second bloc, les couches tentent de reconstruire au fur et à mesure l'information à l'identique. L'évaluation de l'apprentissage résulte en la comparaison (en distance) de l'information d'entrée et celle de sortie reconstruite par le modèle, comme le présente le schéma de la figure 2.11.

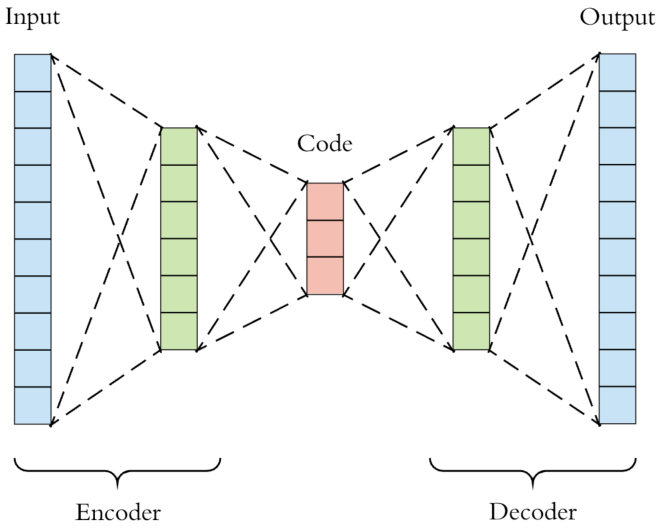


FIGURE 2.11 – Schéma représentant le fonctionnement d'un auto-encodeurs, avec la donnée d'entrée (Input), les couches d'encodage (Encoder), la donnée codée (Code), les couches de décodage (Decoder) et la donnée reconstruite (Output).

Notons que ce type de méthode relativement récent, pourrait être adapté aux données climatiques mais nous pensons que dans un premier temps il serait plus judicieux de poursuivre notre étude en utilisant des méthodes de clustering dites usuelles, puisque ce sont celles qui ont été utilisées dans la bibliographie. Par la suite, en perspective de ces travaux, nous utiliserons d'autres méthodes plus récentes.

Par conséquent, nous considérons que le contexte de notre étude ne correspond pas aux conditions optimales d'utilisation des réseaux de neurones classiques. Nous avons effectué des expérimentations avec un autre type de réseaux de neurones artificiels : les cartes de Kohonen [44]. Il s'agit d'une méthode de carte auto adaptative (*Self-Organizing Maps*), reposant sur le principe d'auto-organisation des neurones artificiels par rapport aux données d'entrées.

Après une initialisation aléatoire des valeurs de chaque neurone, on soumet une à une les données à la carte auto adaptative. Selon les valeurs des neurones, il y en a un, appelées neurone gagnant, qui répond le mieux au stimulus ; c'est celui dont la valeur est la plus proche de la donnée présentée. Ce neurone est alors gratifié d'un changement de valeur pour qu'il réponde encore mieux à un autre stimulus de même nature que le précédent. Par là même, on gratifie un peu aussi les neurones voisins du gagnant avec un facteur multiplicatif du gain inférieur à un. Ainsi, c'est toute la région de la carte autour du neurone gagnant qui se spécialise. En fin d'algorithme, lorsque les neurones ne bougent plus, ou seulement très peu, à chaque itération, la carte auto-organisatrice recouvre toute la topologie des données [45].

Les expérimentations menées avec les cartes de Kohonen au début de nos travaux, n'ont pas été retenues puisqu'elles produisaient systématiquement des résultats incohérents, que nous ne présenterons pas dans ce document. Cependant, avec du recul, nous pensons qu'une adaptation de cette méthode aux spécificités des données (comme nous l'avons fait en fin de thèse) pourrait viabiliser l'usage de cette méthode. Les méthodes d'apprentissage non supervisé sont les plus répandues pour ce type d'étude, c'est la raison pour laquelle nous les avons considérées en priorité.

2.4.4 Apprentissage semi-supervisé

L'apprentissage semi-supervisé est une catégorie de techniques de *machine-learning* qui utilise un ensemble de données étiquetées et non étiquetées. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non supervisé qui n'utilise que des données non étiquetées. L'utilisation de données non étiquetées, en combinaison avec des données étiquetées, permet d'améliorer significativement la qualité de l'apprentissage. Un autre intérêt provient du fait que l'étiquetage de données nécessite souvent l'intervention d'un expert. Lorsque les jeux de données deviennent très grands, cette opération peut s'avérer fastidieuse. Dans ce cas, l'apprentissage semi-supervisé, qui ne nécessite que quelques étiquettes, revêt un intérêt pratique certain.

Dans notre cas d'étude, nous n'avons aucune informations sur les régimes de temps dans notre zone, il nous semble donc délicat d'effectuer ne serait-ce qu'un étiquetage partiel du jeu de données. L'usage de méthodes d'apprentissage non supervisé reste la voie la plus évidente.

2.4.5 Apprentissage non supervisé

Dans le domaine informatique et de l'intelligence artificielle, l'apprentissage non supervisé est un problème d'apprentissage automatique. Il s'agit, pour un logiciel, de trouver des structures sous-jacentes à partir de données non étiquetées [30]. Puisque les données ne sont pas étiquetées, il n'est pas possible d'affecter au résultat de l'algorithme utilisé un score d'adéquation. Cette absence de labellisation (ou d'annotation) est ce qui distingue les tâches d'apprentissage non supervisé des tâches d'apprentissage supervisé. Aucune étiquette n'est donnée à l'algorithme d'apprentissage, il détermine seul la structure des données d'entrées.

L'apprentissage non supervisé peut être un but en soi (découvrir des modèles cachés dans les données) puisqu'il consiste à apprendre sans superviseur. Il permet d'identifier des classes ou groupes d'individus présentant des caractéristiques communes. La qualité de ce type de méthode est mesurée par sa capacité à découvrir certains ou l'ensemble des motifs cachés.

Il faut bien distinguer l'apprentissage supervisé et non supervisé. Dans le premier apprentissage, il s'agit d'apprendre à classer un nouvel individu parmi un ensemble de classes prédéfinies : on connaît les classes à priori. Tandis que dans l'apprentissage non supervisé, le nombre et la définition des classes ne sont pas donnés à priori.

Plus appropriées au contexte de cette étude, les méthodes d'apprentissages non supervisés sont diverses et variées. Les méthodes de clustering ou *clustering* sont celles que nous avons retenues pour la suite.

2.4.5.1 Méthodes de clustering

Le partitionnement de données (ou data *clustering* en anglais) est une méthode en analyse des données. Elle vise à diviser un ensemble de données en différents groupes homogènes nommés "cluster", en ce sens que les données de chaque sous-ensemble partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité (similarité informatique) que l'on définit en introduisant des mesures et classes de distance entre objets [30].

Pour obtenir un bon partitionnement, il convient à la fois de : minimiser l'inertie intra-classe pour garantir des clusters les plus homogènes possible ; maximiser l'inertie inter-classe afin de garantir des sous-ensembles bien différenciés. Nous utiliserons plus loin les termes classiques d'homogénéité et de séparation pour désigner ces propriétés recherchées. Le centre de gravité d'un cluster est nommé centroïde (Figure 2.12).

Les algorithmes de clustering peuvent être catégorisés en fonction de leur "modèle de cluster" (Figure 2.13). Les chercheurs utilisent différents modèles de clusters, et pour chacun de ces modèles de clusters, des algorithmes différents peuvent être donnés. Tous ne fournissent pas de modèles pour leurs clusters et ne peuvent donc pas être facilement catégorisés.

La notion de cluster, telle qu'elle est définie par différents algorithmes, varie considérablement dans ses propriétés. La compréhension de ces modèles de cluster est essentielle pour comprendre les différences entre les divers algo-

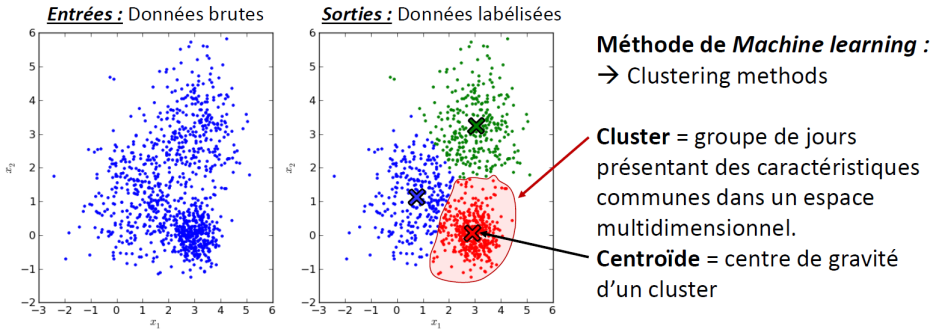


FIGURE 2.12 – Schéma explicatif du fonctionnement du clustering et de la définition du cluster et du centroïde.

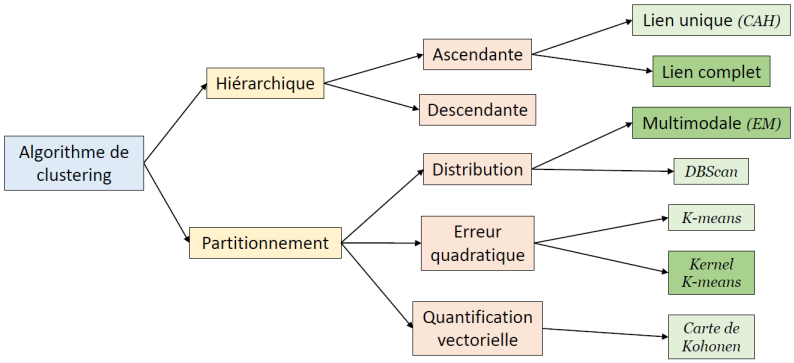


FIGURE 2.13 – Arbre des algorithmes de clustering les utilisées.

rithmes.

2.4.5.1.1 Clustering hiérarchique

Cette méthode construit le cluster en partitionnant récursivement les instances. Le clustering hiérarchique est une méthode d'analyse de cluster qui cherche à construire une hiérarchie de clusters. Il existe deux principales stratégies de regroupement hiérarchique [68] :

- ***Agglomerative* - Il s'agit d'une approche "ascendante" :** chaque observation constitue un cluster au début puis des paires de clusters sont fusionnées au fur et à mesure que l'on remonte dans la hiérarchie.
- ***Divisive* - Il s'agit d'une approche "descendante" :** toutes les observations commencent dans un *cluster*, et les fractionnements sont effectués de façon récursive au fur et à mesure que l'on descend dans la hiérarchie.

Le résultat des méthodes hiérarchiques est un dendrogramme, représentant le regroupement imbriqué d'objets et les niveaux de similarité auxquels les groupes changent. Le résultat du clustering des objets est obtenu en coupant le dendrogramme au niveau de similarité désirée.

Nous avons choisi l'approche ascendante du clustering hiérarchique (CAH) afin d'étudier les successions de regroupement au cours du processus de traitement. CAH est modulable, puisque l'on peut choisir la distance à utiliser (euclidienne, manhattan, mahalanobis, etc.) pour comparer les éléments entre eux. Le choix de la distance appropriée influencera la forme des clusters, car certains éléments peuvent être proches les uns des autres en fonction d'une distance et plus éloignés en fonction d'une autre. Il est également possible de définir un critère de jonction (en anglais *linkage criteria*), celui-ci donne la condition de jonction entre deux éléments ou groupes d'éléments[62, 63].

2.4.5.1.2 Clustering par partitionnement

Les méthodes de partitionnement sont des méthodes de clustering utilisées pour regrouper des données issues d'un ensemble, en plusieurs clusters, en fonction de leur similarité. Le fonctionnement de ces méthodes repose principalement sur le déplacement des centres calculés (ou centroïde) des *clusters* d'un *cluster* à l'autre, à partir d'un positionnement initial. De telles méthodes exigent généralement que le nombre de *clusters* soit prédéfini par l'utilisateur [35, 32].

K-Means (KMS)est la méthode de clustering par partitionnement la plus populaire. C'est une méthode de quantification vectorielle, issue du traitement du signal. KMS vise à partitionner n éléments en k clusters dans lesquelles chaque observation appartient au cluster ayant la moyenne la plus proche, servant de centroïde du cluster. KMS vise à minimiser les variances à l'intérieur des clusters en utilisant la distance euclidienne pour comparer les éléments. Nous avons retenu cette méthode pour effectuer la plupart des clustering présentés dans cette thèse.

2.4.5.1.3 Clustering basée sur la distribution

La méthode de Clustering la plus étroitement liée aux statistiques est basée

sur des modèles de distribution. Les *clusters* peuvent être définis comme des objets appartenant le plus souvent à la même distribution. Une propriété pratique de cette approche est que cela ressemble beaucoup à la façon dont les ensembles de données artificielles sont générés : en échantillonnant des objets aléatoires d'une distribution [68]. L'une des plus célèbres est l'algorithme *Espérance-Maximisation* [21, 22].

Bien que le fondement théorique de ces méthodes soit excellent, elles souffrent d'un problème clé connu sous le nom de "suradaptation" à moins que des contraintes ne soient imposées à la complexité du modèle. C'est l'une des raisons pour lesquelles, nous n'avons pas utilisé ce type de méthode.

2.4.5.1.4 Clustering basée sur la densité de probabilité

Dans les méthodes de Clustering liées à la densité de probabilité, les *clusters* sont définis comme des zones de plus forte densité que le reste de l'ensemble de données. Les éléments de ces zones clairsemées, qui sont nécessaires pour séparer les *clusters*, sont généralement considérés comme du bruit et des points frontaliers [52].

L'algorithme DBSCAN pour Density-Based Spatial Clustering of Application with Noise [25], est le plus connu. Il permet d'identifier des clusters à forte densité. Dans nos travaux, ce type de méthode de clustering n'a pas été retenu, car il compare les données sur l'ensemble de leurs composantes mélangeant ainsi des phénomènes atmosphériques de nature différente [25].

2.4.5.1.5 Clustering non paramétrique

Il existe des méthodes de clustering permettant de déterminer automatiquement le nombre optimal de clusters à retenir pour un jeux de données. Par exemple, la méthode *mean-shift* est une technique d'analyse non paramétrique de l'espace des caractéristiques permettant de localiser les maxima d'une fonction de densité, c'est en fait un algorithme de recherche de mode [81, 14]. Nous n'avons pas utilisé cette méthode dans ces travaux, mais nous considérons son application dans des études ultérieures. Nous verrons plus loin que nous allons définir une méthode pour choisir le nombre de cluster à retenir.

2.5 Méthodologie classique et objectifs détaillés

Les méthodes de clustering étant diverses et variées, dans ces travaux nous nous sommes focalisés sur les méthodes les plus utilisées, KMS et CAH. Pour effectuer le clustering, il était nécessaire que les données cartographiques soient vectorisées (Fig 2.14). Dans cette étude, les données étaient complexes et leurs dimensions étaient relativement grandes puisqu'il s'agit souvent d'étudier de longues périodes et de vastes zones [70]. La vectorisation s'applique bien aux données de réanalyses (de type raster) qui par définition, portent des variables temporelles spatialisées en pixel couvrant les zones géographiques d'intérêt.

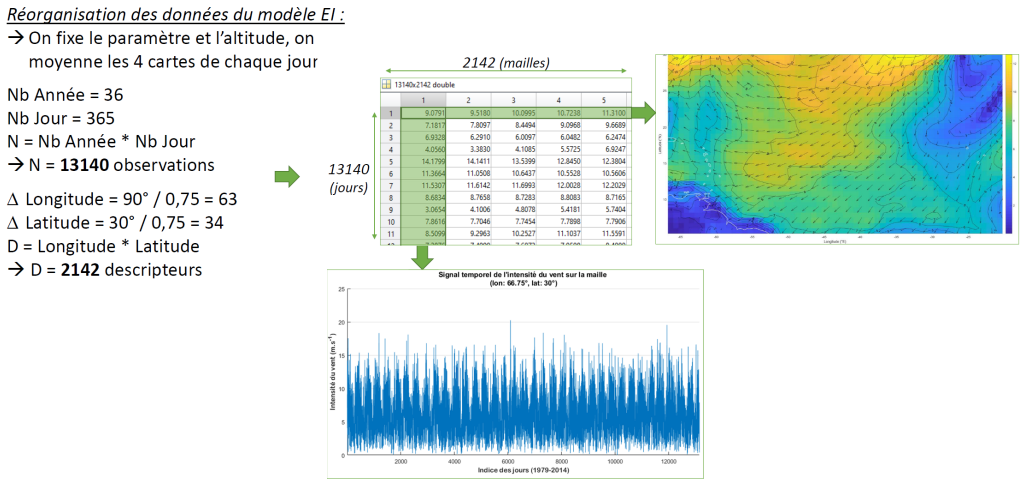


FIGURE 2.14 – Exemple de réorganisation des données pour l'intensité du vent à 850hPa de Era Interim (EI) de 1979 à 2014 pour la zone des petites Antilles.

Nous allons donc appliquer la méthodologie introduite dans la sous-section 2.2.2, consistant à sélectionner un paramètre météorologique à une altitude, pour notre période d'étude. Ces données seront réorganisées puis utilisées comme entrées des algorithmes de clustering (KMS et CAH). En suivant cette méthodologie classique, nous allons nous focaliser sur l'analyse spatiale des configurations atmosphériques, en procédant à l'identification automatique des motifs récurrents par clustering.

Durant le processus, les données journalières seront comparées entre elles grâce à la distance euclidienne (ou norme L2) et les éléments similaires entre eux seront regroupés en clusters. Nous montrerons par la suite que la norme L2 n'est pas tout à fait adaptée à la nature de nos données d'étude. En effet, elle a

tendance à affaiblir les résultats du clustering.

Ensuite, puisque chaque cluster est construit autour de son centroïde, ce dernier pourra être analysé par les experts. Il est généralement interprété comme étant une configuration spatiale récurrente. Cette approche sera également remise en cause par la suite. Nous verrons que le centroïde est un élément artificiel qui ne représente en rien une configuration atmosphérique réelle.

Par la suite, la dimension temporelle sera étudiée indirectement, il s'agira d'utiliser les références (ou dates) des données journalières pour étudier la dynamique temporelle des clusters obtenus. En effet, le clustering produit une labellisation des données d'entrées. Chaque donnée journalière, donc chaque jour, est associée à un cluster. En déterminant, la répartition des effectifs des clusters par mois (intra-annuelle) et par année (inter-annuelle), il est possible d'analyser de manière indirecte l'évolution temporelle des clusters. Cela fait partie des principaux indicateurs utilisés par les experts du climat pour analyser les tendances temporelles des configurations spatio-temporelles.

L'intégration de l'aspect temporel directement dans le clustering n'est pas considérée dans la méthodologie classique, néanmoins cela représenterait une vision plus complète de l'analyse des données. Nous considérons cette approche comme une perspective notoire et complémentaire aux innovations apportées dans cette thèse.

2.6 Conclusion

Dans ce chapitre l'on comprend que l'informatique climatique apporte de nouvelles perspectives dans la recherche sur l'analyse du climat. L'utilisation de méthodes d'apprentissages automatisées pour comprendre le passé climatique et prévoir les changements futurs permet vraisemblablement de faire progresser la recherche dans ce domaine.

Parmi les méthodes de *machine learning*, nous avons choisi l'apprentissage non supervisé et plus précisément les algorithmes K-Means (KMS) et Classification Ascendante Hiérarchique (CAH), qui sont les plus utilisés. Ces méthodes sont capables de regrouper des éléments spatio-temporels ayant des caractéristiques similaires. Elles seront utilisées pour déterminer et comprendre l'organisation interne des données climatiques dont nous disposons.

Les données climatiques de cette étude étaient de type raster, elles s'apparentent donc à des images où chaque pixel porte une valeur d'un paramètre météorologique modélisé pour une zone précise. Ces données sont souvent denses et relativement complexes. Cette complexité réside surtout dans le fait que par nature, il y a des corrélations spatio-temporelles entre elles. Les paramètres météorologiques caractérisés par ces données seront donc passés en entrée des algorithmes de clustering.

Les clusters obtenus à la suite du clustering correspondent à des groupes de jours ayant des caractéristiques similaires pour un paramètre météorologique donné. Leur centroïde est calculé et utilisé pour visualiser et comparer les tendances spatiales des différents clusters. En ce qui concerne la dynamique temporelle de ces derniers, il est possible de l'étudier en utilisant la date associée à chaque élément de la base, puisqu'il s'agit de données quotidiennes.

En somme, cela correspond à la première approche méthodologique que nous présentons dans cette thèse et dont des applications sont exposées dans le chapitre qui suit.

Chapitre 3

Premières applications de la méthodologie classique

3.1 Introduction

Dans ce chapitre, deux études préliminaires mettant en application des méthodes classiques de clustering (KMS et CAH) sont présentées. Il s'agit dans les deux cas de procéder à un clustering de données climatiques, afin d'identifier des configurations spatio-temporelles prédominantes.

Dans les deux cas, le jeu de données couvrira une période réduite (pour la première étude) ou correspondant à un évènement (pour la seconde étude). Les deux algorithmes de clustering, cités précédemment, seront mis en correspondance afin d'établir des tendances (ou motifs) similaires et d'associer leurs résultats pour confirmer la relative stabilité des configurations trouvées.

Nous allons également procéder, au fil de ces études, à une analyse comparative des méthodes utilisées : Celles des experts du climat pour interpréter les résultats et celles du domaine informatique que nous chercherons à promouvoir.

L'objectif est de monter que les méthodes d'analyses automatisées peuvent être tout aussi pertinentes que celles des experts. L'organisation de chacune de ces études suivra peu ou prou le schéma suivant. Pour commencer, nous présenterons le contexte de l'étude, puis viendra l'évaluation visuelle de l'expert. Celle-ci sera confirmée ou remise en cause par les méthodes d'analyse que

nous utiliserons. Pour finir, nous conclurons sur l'intérêt de ces analyses et les résultats obtenus.

Ces études nous permettront ainsi de mettre en évidence l'application d'une méthodologie classique en climatologie et les résultats qu'en tirent les physiciens. Cela permettra également de pointer un certain nombre de problèmes liés à ces méthodologies, ces derniers pourront de ce fait être résolus dans le chapitre suivant (cf. Chapitre 4).

Dans la section 3.2, des clusters du paramètre météorologique "géopotentiel" sont étudiés durant la période humide afin d'identifier des configurations prédominantes dans les circulations (les vents) des couches supérieures de l'atmosphère. Alors que dans la section 3.3, des clusters de courants océaniques de surfaces sont comparés à un historique de date d'échouement d'algues sargasses afin de déterminer les configurations spatio-temporelles vectrices de bancs de sargasses vers la Guadeloupe.

Avec le recul acquis durant cette thèse, ces travaux présentent un certain nombre de points d'amélioration. Il s'agissait en effet des premières applications des méthodes de clustering sur les données climatiques, bien que nous n'avions pas encore saisi toute la complexité de ces données, nous avons tout de même été en mesure de produire une analyse cohérente et ainsi identifier des points d'améliorations que nous intégrerons dans les chapitres suivants.

3.2 Analyse du géopotentiel

3.2.1 Motivations et introduction

Cette étude correspond à nos premières expérimentations sur des données climatiques en utilisant les méthodes de clustering que nous avons longuement présentées précédemment dans le chapitre 2.

Pour commencer, nous avons choisi d'utiliser un paramètre classique qui fait sens dans le domaine de physique : le géopotentiel. Exprimé en mètre (m), ce paramètre donne pour une isobare définie, la valeur de l'altitude à atteindre pour égaler ce niveau de pression. Généralement, les physiciens de l'atmosphère utilisent ce paramètre pour identifier les zones de dépressions et d'anticyclones en surface.

Nous allons, dans cette étude, nous focaliser sur l'analyse de ce paramètre à $500h$Pa ($\approx 5\,600m$ à nos latitudes), avec les réanalyses ERA-Intérim. Le niveau de pression $500h$Pa correspond, selon les physiciens, à une altitude garantissant l'abstraction des phénomènes et activités de surface. L'objectif physique étant d'analyser les circulations atmosphériques de grande échelle.

Pour effectuer le clustering, nous utiliserons plus précisément l'anomalie du géopotentiel, son mode de calcul sera précisé dans la section suivante. D'après les experts en climatologie, les anomalies de ces niveaux, calculées sur de longues périodes, permettent de trouver des précurseurs et/ou inducteurs de la dynamique météorologique de grande échelle.

L'objectif informatique est d'effectuer une analyse de la correspondance entre les résultats de deux méthodes de clustering (KMS et CAH) sur le même jeu de données, afin d'identifier au travers de ces algorithmes des motifs similaires.

Pour arriver à nos objectifs, nous procéderons en trois temps. Pour commencer, nous présenterons le contexte de l'étude, en précisant les données utilisées, la zone et la période d'intérêt (cf. Sous-section 3.2.2).

Ensuite, dans un second temps, nous présenterons les résultats obtenus pour le

clustering, ils seront par la suite interprétés par un expert du domaine. Nous effectuerons également une analyse de correspondance des clusters obtenus en s'appuyant sur une mesure de proximité entre les centroïdes et des indicateurs statistiques (cf. Sous-section 3.2.3).

Enfin, nous conclurons sur les principaux résultats de cette première expérimentation (cf. Sous-section 3.2.5).

3.2.2 Matériels et méthodes

Dans cette application, nous avons choisi de nous focaliser temporellement sur la saison humide définie pour la zone géographique entière, appelée domaine A (cf. Fig 2.9) correspondant à notre zone géographique. Elle représente une image de 40 x 120 pixels. À ces latitudes la saison pluvieuse s'étend principalement sur trois mois : Août, Septembre et Octobre.

Les données journalières que nous avons utilisées pour cette étude décrivent le géopotentiel à $500h$Pa provenant du projet ERA-Intérim, pour trente-six années (de 1979 à 2014), à raison de quatre-vingt-douze jours par année. Avant d'effectuer le clustering, nous avons procédé à un pré-traitement permettant, selon les physiciens, de mieux analyser ce paramètre dans le temps [43].

Il consiste en l'utilisation d'un écart à la moyenne du jour dans l'année, pour ce paramètre. Il se calcule de la manière suivante : pour un jour associé à un mois (ex. le 25 décembre), l'on effectue une moyenne sur les trente-six années. De ce fait, cela produit autant de moyennes que de jours dans une année. Dans notre cas, nous disposerons de quatre-vingt-douze moyennes de références, puisque nous travaillons sur trois mois. C'est à partir de ces moyennes que nous allons calculer l'anomalie de chaque donnée journalière de la base.

Les méthodes de clustering K-Means (KMS) et Classification Ascendante Hiérarchique (CAH), présentées dans le chapitre précédent, seront utilisées pour l'analyse. Leurs résultats seront décrits par les experts dans la section suivante, ensuite nous allons chercher à mettre en correspondance les clusters et leurs centroïdes.

Les physiciens de l'atmosphère nous proposent de fixer le nombre de clusters à retenir à $k = 5$, afin d'identifier les cinq motifs (ou tendances) les plus marqués en terme de circulation atmosphérique à cette altitude.

3.2.3 Analyse visuelle de clusters résultants

L'inspection visuelle des clusters résultants a été effectuée à partir de la figure 3.1. Les centroïdes des deux méthodes sont disposés de gauche vers la droite, de C1 à C5. Les centroïdes en partie (a) sont issus de KMS, la partie (b) correspond à ceux issus de CAH.

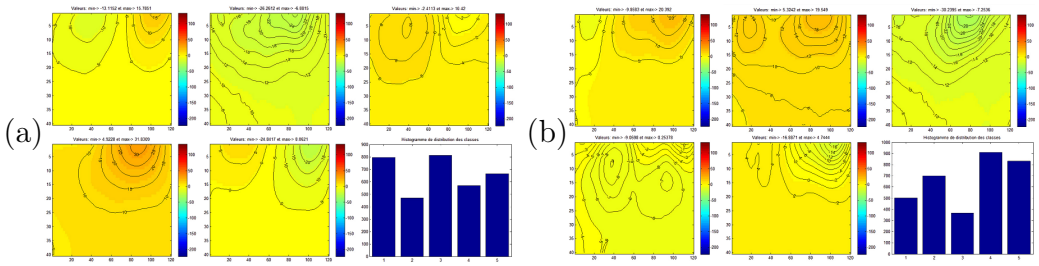


FIGURE 3.1 – Centroïdes d'anomalies du géopotentiel à 500hPa produits (a) KMS et (b) CAH, avec les lignes de d'altitude (en noir) et la distribution des éléments par cluster.

D'après les experts, chacune des méthodes produit des clusters qui reflètent assez fidèlement la situation.

Les clusters KMS-C2, KMS-C3, KMS-C4, CAH-C2, CAH-C3 et CAH-C5 montrent une évolution (ou un gradient) méridionale (du Nord au Sud), qui peut être positive ou négative, avec une anomalie faible au Sud. Il s'agit d'une configuration classique de vent d'alizés.

En somme, selon l'expert, quatre configurations peuvent être retenues, voici les clusters qui y sont associés :

- Gradient méridional positif au Nord (KMS-C4, CAH-C2) ;
- Gradient méridional négatif au Nord (KMS-C2, CAH-C3) ;
- Bipôle zonal positf-négatif (KMS-C5, CAH-C5) ;

- Bipôle zonal négatif-positf (KMS-C1, CAH-C1) ;

Les centroïdes KMS-C3 et CAH-C4 sont assez particuliers, selon l'expert, ils ne semblent pas correspondre à la même situation et ne sont pas exactement similaires aux autres configurations. Nous verrons par la suite que, contrairement à l'analyse visuelle experte, les méthodes d'analyse informatiques permettent de compléter ces évaluations visuelles.

Notons également que le choix du nombre de configurations ne correspond pas au nombre de clusters produits. En effet, bien que les physiciens s'attendaient à voir cinq clusters différents, ces derniers au travers de leur analyse n'ont identifié que quatre configurations.

Dans la sous-section suivant, nous allons procéder à une analyse plus orientée informatique, afin de vérifier si l'interprétation des spécialistes du domaine correspond à celle des outils d'analyses que nous avons utilisés.

3.2.4 Mise en correspondance

Dans cette section, nous faisons une brève analyse numérique des motifs, des distributions comparatives et des variances internes obtenues pour les deux méthodes de clustering. Pour cela nous utilisons des tables de comparaisons (cf. Tableaux 3.1, 3.2 et 3.3).

Les valeurs de la table 3.1 représentent la distance euclidienne entre les centroïdes des clusters issue des méthodes KMS et CAH. Cette méthode vise à remplacer l'évaluation visuelle de l'expert par une analyse automatisée des centroïdes, pixel par pixel. Notons que plus les valeurs sont faibles et plus les centroïdes représentatifs des clusters sont proches, voire quasi semblables. Les valeurs en gras sont les distances euclidiennes les plus faibles par cluster et celles en vert représentent les minimums en commun entre les deux méthodes.

On constate que quatre des cinq centroïdes sont similaires aux deux méthodes puisque la distance qui les séparent est relativement faible. Cela confirme en partie l'avis de l'expert, néanmoins les couples constitués ne sont pas similaires. La méthode permet d'associer directement les couples (KMS-C1, CAH-C4) et

TABLEAU 3.1 – Distances euclidiennes entre les centroïdes des clusters de K-Means et CAH, avec une accentuation des meilleures valeurs par cluster (en gras), et les concordances pour les deux algorithmes (en vert).

CAH		C1	C2	C3	C4	C5
KMS	**C1**	5.98	12.76	11.06	**4.09**	7.54
	C2	16.93	23.75	**0.77**	7.86	14.76
	C3	**5.52**	5.03	18.79	10.87	**3.95**
	C4	7.83	**2.25**	24.83	16.90	9.99
	C5	8.80	14.06	9.99	**4.76**	5.35

(KMS-C2, CAH-C3), reconnus par l'expert, mais également les groupes (KMS-C3, CAH-C5, CAH-C1) et (KMS-C4, CAH-C2, KMS-C5).

Il est donc important de noter, que la méthode mesure la similarité entre les motifs. Son usage permet donc d'effectuer une analyse plus fine des centroïdes, fournissant ainsi une meilleure évaluation de la correspondance visuelle de ces derniers. La nécessité d'une telle méthode sera plus longuement expliquée dans le chapitre 5.

La table 3.2 donne le pourcentage de correspondance entre les clusters de KMS et CAH. Ce pourcentage est calculé à partir du nombre d'éléments en commun entre les clusters des différentes méthodes. Les valeurs en gras dans le tableau sont les pourcentages de correspondance les plus forts par cluster. Celles en vert représentent les pourcentages forts en communs pour les deux méthodes.

Cette fois-ci l'objectif est différent, la méthode permet d'analyser la constitution des clusters. Celle-ci semble beaucoup plus fiable que les précédentes, puisqu'il ne s'agit plus de considérer uniquement les motifs des centroïdes.

Les groupes obtenus sur la base de la correspondance des éléments constitutifs des clusters semblent quasiment correspondre au résultat des analyses précédentes. A peu de chose près que le cluster CAH-C1 se retrouve cette fois-ci associé au couple (KMS-C1, CAH-C4).

Il faut tout de même noter, que le pourcentage de correspondance maximal est de 51.3% et correspond au couple (KMS-C3, CAH-C5), qui pourtant n'avait

TABLEAU 3.2 – Pourcentages de correspondance entre les clusters de K-Means et de CAH, avec une accentuation des meilleurs valeurs par cluster (en gras), et les concordances pour les deux algorithmes (en vert).

CAH		C1	C2	C3	C4	C5
KMS	C1	**18.6%**	2.3%	1.5%	**49.8%**	9.5%
	C2	0%	0%	**37.4%**	13.3%	1.3%
	C3	12.5%	28.7%	0%	1.5%	**51.3%**
	C4	11.8%	**37.6%**	0%	0.1%	0.1%
	C5	0.7%	0.1%	2.1%	**27.8%**	24.7%

pas du tout été identifié par l'évaluation experte.

Cela révèle deux choses, d'une part l'analyse visuelle du centroïde est insuffisante. D'autre part il est absolument nécessaire de la compléter par une analyse numérique permettant de valider la mise en correspondance des clusters de différentes méthodes.

Pour aller plus loin, nous avons commencé à nous pencher sur l'évaluation des clusters. En effet, bien que la mise en correspondance semble fonctionner nous pensons qu'une évaluation directe de la qualité des clusters nous permettrait de choisir quelles méthodes produiraient de meilleurs résultats.

Pour cela nous avons déterminé la variance interne de chacun des clusters de KMS et de CAH, pour vérifier l'homogénéïté des clusters, elle est présentée dans la table 3.3. Notons que plus la variance interne est faible et meilleur sera le placement du centroïde par rapport à son cluster.

TABLEAU 3.3 – Comparaison de la variance intra-clusters pour KMS et CAH, avec une accentuation des meilleurs valeurs par cluster (en gras), et les concordances pour les deux algorithmes (en vert).

Méthodes	C1	C2	C3	C4	C5
KMS	12.9e+4	**8.9e+4**	10.2e+4	**8.8e+4**	9.9e+4
CAH	4.1e+3	6.1e+3	**3.5e+3**	8.3e+3	7.3e+3

De façon global, on constate que les centroïdes des clusters de CAH sont mieux

placés que ceux de KMS, il y a donc une meilleure cohérence interne au sein de ces derniers.

Nous relevons également que la variance cumulée du couple (KMS-C2, CAH-C3) est la plus faible. Selon les expert, cela correspond aux jours où l'anticyclone des açores s'affaiblit. Il s'agit donc de la configuration atmosphérique la mieux identifiée par les méthodes de clustering utilisées.

3.2.5 Conclusion

Dans cette section, nous avons montré que les méthodes de clustering sont tout à fait capables d'identifier des configurations atmosphériques liées à un paramètre physique, qui ont du sens pour les spécialiste du domaine.

Le paramètre choisi pour cette étude, le géopotentiel à $500h$Pa, et le prétraitement proposé nous ont semblé tout à fait pertinents, puisqu'ils permettent de produire des clusters cohérents. Ces derniers peuvent être exploités par les spécialistes afin d'identifier les fluctuations du géopotentiel (à environ $5\ 600m$) et contribuer à la détermination de possibles régimes de temps qui impactent directement la météorologie de surface. Pourtant cette relative "simplicité", soulève tout de même un questionnement : ces motifs sont viable ou même réelle ? Ce questionnement reviendra assez régulièrement dans ce manuscrit, le chapitre 4 y apportera un certain nombre de réponses.

Néanmoins, pour tenter d'y voir plus clair, dans cette étude nous avons cherché à mettre en correspondance les résultats de nos deux algorithmes (KMS et CAH) afin d'assoir une fiabilité toute relative. Nous avons donc usé d'un raisonnement à la fois basé sur l'expertise physique et les méthodes d'analyse de cluster, propre au domaine informatique, pour permettre d'identifier des configurations spatiales bien établies.

Enfin, nous pouvons déjà affirmer que l'analyse effectuée, à partir à nos méthodes automatisées, est complémentaire visuelle de l'expert. Néanmoins, celle-ci à tout de même permis d'identifier les tendances générales, ce qui est tout à fait remarquable.

Il est également important de noter qu'il s'agit de notre première approche des données, celle-ci nous a permis de constater qu'il existe certaines fois des correspondances entre les résultats des algorithmes de clustering. En effet, bien que la taille et la composition des clusters de deux méthodes varient, les motifs de leurs centroïdes peuvent être relativement similaires.

Avec du recul, cette étude pourrait être enrichie sous plusieurs aspects. Il faudrait en priorité une évaluation plus poussée au sens informatique de la qualité des clusters, ici présentée sous la forme d'une recherche de stabilité entre les méthodes et de la mesure de la variance intra-classe. Nous chercherons donc à enrichir cet aspect dans le chapitre 4.

La section suivante présente les résultats obtenus dans le cadre d'une étude annexes aux régimes de temps. Nous avons jugé intéressant de l'intégrer au manuscrit, car là encore nous utiliserons les méthodes de clustering pour identifier dans ce cas des configurations spatiales récurrentes de courants océaniques. Ces configurations nous permettront, par la suite, d'étudier le processus d'échouement d'algues sargasses sur les côtes de la Guadeloupe.

3.3 Identification des courants océaniques favorables à l'échouement de sargasses

3.3.1 Motivations et introduction

En 2011, et de 2015 à 2018, l'afflux sans précédent d'algues sargasses qui a atteint les Antilles Françaises (AF), à plusieurs reprises, a constitué une menace majeure pour l'écologie, la santé et l'économie de ces îles. De faible surface, ces deux îles font partie des territoires gravement touchés par le transport via les courants marins favorisant l'échouement de ces dernières dans la caraïbe (cf. Figure 3.2).

FIGURE 3.2 – Exemple d'échouement massif en Martinique en 2018, photographié par un drone (Mad'InAir).

Lors d'invasions majeures, les volumes à ramasser et à ressuyer, à proximité des zones habitées, ou à forts enjeux économiques ou environnementaux, pour les dépôts constatés suites aux échouements, ont été estimés en moyenne à 150 $000 m^3$.

L'imprévisibilité des échouements est une difficulté récurrente souvent marquée par une augmentation de la durée et des volumes sur la période de l'année 2011 à 2018. Un effort de recherche devient donc nécessaire pour réduire ces incertitudes, et ainsi de mieux appréhender les cycles à venir et les possibles échouements afin de mieux réagir.

Pour cette seconde étude, nous proposons d'identifier l'ensemble discret des régimes spécifiques de circulations océaniques amenant à des échouements à

partir de données quotidiennes de courant de surface correspondant aux jours d'échouements. Elles sont issues du modèle Mercator Océan. Nous avons sélectionné les données correspondant aux dates d'échouements référencés par la DEAL Guadeloupe.

Nous utiliserons des méthodes de clsutering (KMS et CAH), pour identifier les régimes de circulation océanique, à partir de données maillées, centrées sur la partie orientale des Antilles (Fig 3.3(a) (55-66°W et 8-17°N).

Les experts physiciens de l'équipe s'appuieront sur les courants généraux (cf. Fig 3.3) prédéfinis dans la littérature scientifique, pour analyser ceux obtenus par clustering.

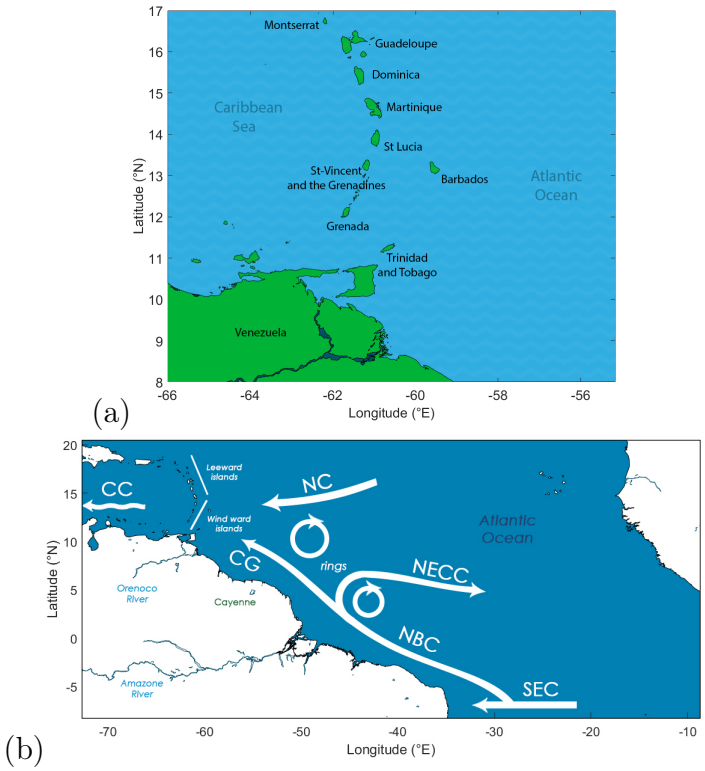


FIGURE 3.3 – (a) Zone d'intérêt de l'étude sargasse. (b) Représentation schématique de la circulation dans l'océan Atlantique. Courant Atlantique sud Équatorial (SEC), courant Nord Brésil (NBC), Contre Courant Nord Equatorial (NECC), Courant Atlantique Nord Équatorial (NC), Courant des Guyanes (CG), Courant de la Caraïbe ou des Antilles (CC).

La méthodologie est donnée dans la sous-section 3.3.2 et les résultats sont présentés dans la sous-section 3.3.3. Nous examinerons également les possibles voies de transport en surface à l'aide d'une simulation de lâchés de particules dans l'océan Atlantique proche des Antilles Françaises. Les discussions et la conclusion sont en sous-section 3.3.6.

3.3.2 Matériels et méthodes

Pour trouver les possibles clusters nous avons utilisé des données quotidiennes de prévisions océaniques opérationnelles du système d'analyse PSY4V3R1 Mercator, à 1/12 de degré en 3D[29]. Ces données comprennent l'assimilation des données provenant aussi bien des observations multi-instruments in situ que d'observations issues de la télédétection par satellite. Il s'agit donc de données réanalysées. De plus ce modèle intègre l'influence du vent de surface dans le calcul du courant de surface.

Nous disposons également de séries chronologiques débutant en mai 2017 et se terminant en mars 2019, et correspondant à 158 dates d'échouements (cf. Tableau 3.4). Les observations des jours d'échouements sont référencées par la Direction de l'Environnement, de l'Aménagement et du Logement de la Guadeloupe (voir http://www.guadeloupe.developpement-durable.gouv.fr/actualite ssargasses-r989.html).

Tableau 3.4 – Nombre d'échouements pour les trimestres Janvier-Février-Mars (JFM), Avril-Mai-Juin (AMJ), Juillet-Août-Septembre(JAS) et Octobre-Novembre-Décembre (OND). Au premier trimestre 2017, (-) pas d'occurrences.

Années	**JFM**	**AMJ**	**JAS**	**OND**
2017	-	12	16	3
2018	19	55	16	16
2019	17	-	-	-

Nous utiliserons alors les champs horaires moyens des courants journaliers, notées les composantes U (zonale) et V (méridionale) obtenues à $50cm$ de profondeur, pour la zone comprise entre 8 à 17°N et -66 à -55°E.

Pour explorer la sensibilité des sorties quotidiennes donnant les vitesses de courant en surface et trouver des motifs récurrents nous utiliserons des méthodes de clustering. L'objectif étant de regrouper en k clusters les champs spatio-temporels similaires en utilisant respectivement la distance euclidienne pour KMS et CAH.

Dans ces travaux, nous avons retenu les résultats de CAH utilisés la distance de Ward [78, 62] qui produit de meilleurs résultats que la norme L2 classique. Cette première modification de la distance nous a permis de constater que la norme L2 ne produit pas toujours de bons résultats. Nous confirmerons cette hypothèse par la suite dans le chapitre 5. La norme L2 n'est pas toujours adaptée au clustering exploitant des données climatiques.

Le jeu de données étant de grande taille, le choix du nombre de cluster k n'est pas intuitif. Pour éviter d'obtenir de clusters trop généralistes, (k petit), ou d'avoir un partitionnement trop fragmenté des données (k grand) nous tracerons la variance de la distance intra-classe aux centroïdes correspondant à la division successive en k clusters.

De plus, afin d'éviter les problèmes des optimums locaux, KMS a été initialisé 100 fois avec des valeurs prises au hasard et un nombre maximal de 10 000 itérations maximales pour stabiliser le résultat obtenu. Le minimum trouvé et la stabilisation par itération nous permettent de sélectionner, en première analyse, la valeur k=5 comme nombre de cluster utilisé pour regrouper et représenter les types de circulation océanique en surface dans cette partie de la Caraïbe.

Il s'agit d'un début d'approche de l'évaluation des clusters. Bien qu'elle soit peu élaborée, elle a le mérite d'intégrer d'avantage de robustesse dans notre analyse.

Les clusters trouvés n'ont pas la prétention de capturer toute la variabilité du continuum océanique mais ils peuvent être considérés comme des attracteurs du système de la circulation envisagée.

3.3.3 Évaluation numérique

L'évaluation numérique de la qualité des clusters produits dans ces travaux, s'appuie essentiellement sur le tracé représenté sur la figure 3.4. Elle montre, pour les deux algorithmes de clustering, l'évolution de la moyenne des variances intra-classes des clusters en fonction du nombre de clusters.

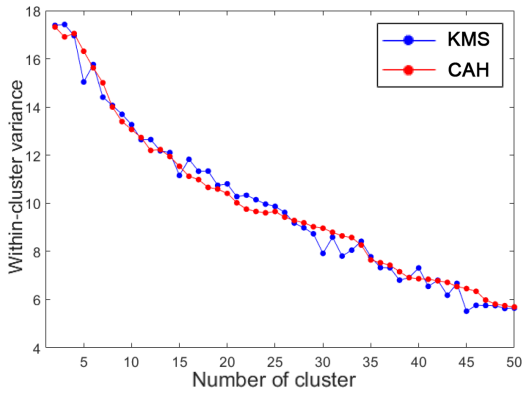


FIGURE 3.4 – Évolution du rapport variance intraclasse sur variance interclasse en fonction du nombre k de clusters. KMS (courbe bleue) et CAH (courbe rouge)

Pour la méthode KMS, la courbe bleue, lue de gauche à droite le long de l'abscisse, montre un ensemble de minima locaux pour $k = \{5, 15, 30, 32, 35\}$. En accord avec l'intuition de nos collaborateurs physiciens nous choisissons de garder $k = 5$.

La courbe rouge de la méthode CAH suit une décroissance plutôt monotone, n'indiquant pas de minimum en particulier.

Poursuivons avec l'analyse visuelle établie par les physiciens, des centroïdes produit par clustering. Les configurations des deux algorithmes seront mises en correspondance comme dans la section précédente (cf. section 3.2).

3.3.4 Évaluation visuelle

Les cinq types de circulations ainsi obtenus sont représentés sur la figure 3.5. C'est la situation réelle et quotidienne la plus proche du centroïde calculé qui est donnée. Cette approche sera également retenue pour les résultats des tra-

vaux présentés en chapitre 5.

Afin d'effectuer une analyse complète, les physiciens se basent sur les courants définis dans la littérature scientifique, ils sont représentés sur la figure 3.5. A partir de là commence l'analyse par les experts des configurations retenues.

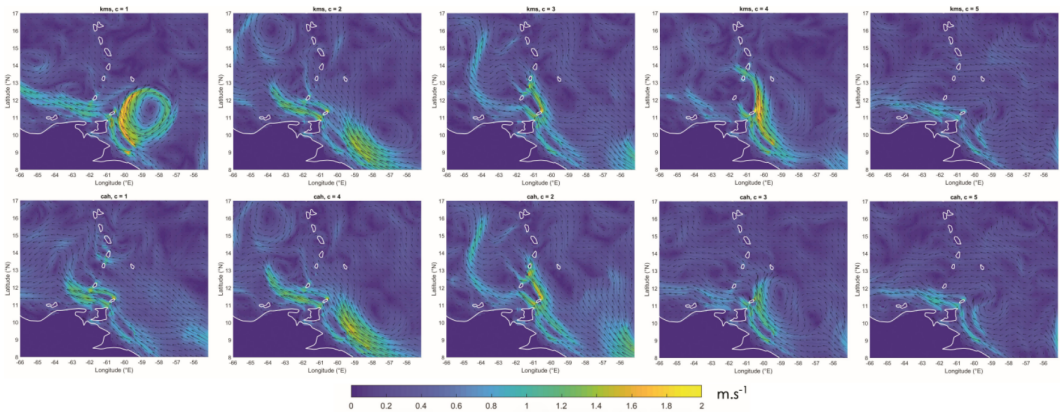


FIGURE 3.5 – Clusters de régimes de courants océanique de surface, les cinq clusters de KMS en haut sont mises en correspondance avec ceux de CAH, formant ainsi des couples.

Les clusters KMS-C3 et CAH-C5 ont respectivement les effectifs les plus importants par méthode. D'ailleurs, le cluster CAH-C5 représente plus du tiers des occurrences. L'analyse de la configuration spatiale pour les courants de surface nous permet de rassembler les clusters par couples. Les couples identifiés sont représentés sur la figure 3.5.

Trois motifs sont similaires. Il s'agit des couples (KMS-C2,CAH-C4), (KMS-C3,CAH-C2) et (KMS-C5,CAH-C5). L'élément de gauche de chaque couple indique le numéro de la classe obtenue par la méthode KMS et celui de droite celui de la méthode CAH.

Les couples trouvés se différencient par l'emplacement et l'extension des courants de surface, ainsi que par la présence ou l'absence des anneaux de rétroflexion dus au courant du nord Brésil (NBC). Ainsi, l'écoulement des eaux de surface provenant de l'Atlantique équatorial par le Sud-Est vers les Caraïbes, contient trois structures principales :

- celle qui décroche à l'est de l'Arc des Antilles et qui prend une direction nord, en restant dans l'Atlantique,

- les deux autres qui pénètrent dans la mer des Caraïbes par le sud, soit en :

  - restant collée au talus continental,
  - traversant au nord de l'île de Trinidad, pour s'incurver à l'intérieur du bassin des Caraïbes.

Le premier type de circulation correspond au couple (KMS-C5,CAH-C5), en dernière colonne de la Figure 3.5. Il apparaît un courant quasiment plaqué sur le plateau continental gardant une composante ouest. Dans le reste du domaine, les eaux de surface traversent l'archipel presque perpendiculairement aux îles, transportées vers l'ouest par les courants de surface de la gyre Atlantique.

Le deuxième type de circulation correspond au couple (KMS-C2,CAH-C4). La deuxième colonne de la Figure 3.5 montre qu'au nord de Trinidad, le flux se décolle du talus continental, s'infléchit et se divise en deux parties en prenant une orientation nord à nord ouest. La partie inférieure plus rapide passe au sud des îles de Tobago et de Grenade. La partie supérieure, plus large, transporte les eaux de surface marines en passant au nord, entre Saint-Vincent et les Grenadines pour atteindre les autres îles de Antilles. À l'ouest de la Guadeloupe et de la Dominique apparaissent des circulations fermées.

Le dernier type de circulation, couple (KMS-C3,CAH-C2), se différencie du cas précédent par l'arrivée d'eaux de surface marines provenant d'une zone géographique relativement large située entre -60 à -55°E, et représente la troisième colonne de la Figure 3.5. Elles sont transportées par un courant ayant une forte composante méridionale qui se divise en trois vers 11°N. Les eaux de surface marines traversent les Antilles comme auparavant mais avec des intensités plus marquées, soit des vitesses supérieures à 1.2 $m.s^{-1}$. De part et d'autres de la branche centrale coexistent deux branches quasi parallèles et méridionales situées à l'Est et à l'Ouest des îles. La branche centrale rejoint

celle de la mer des Caraïbes au niveau du 14°N.

Enfin, les anneaux dus à la rétroflexion du NBC apparaissent sur les clusters C1, C2, C4 pour KMS et C4, C5 pour CAH. La méthode KMS attribue un centroïde pour le cluster C1, identifiant clairement cette structure juste à l'est de Trinidad. Le cluster C1 de la méthode CAH montre la prédominance jusqu'à 13°N du courant Guyanais.

3.3.5 Mise en correspondance

La fréquence des cinq types de circulation à l'origine des échouements sur les côtes de l'archipel de la Guadeloupe est présentée dans le Tableau 3.5. Au delà de leur répartition par classe les types de circulation décrits dans la partie précédente varient annuellement.

Nous avons donc fait la correspondance temporelle entre les clusters trouvés en identifiant les jours communs par cluster et par méthode. Les résultats sont regroupés dans les tableaux 3.6 et 3.7.

Tableau 3.5 – Fréquence en pourcentage de chaque régime de circulation océanique obtenus par KMS et CAH

Algorithmes	**C1**	**C2**	**C3**	**C4**	**C5**
KMS (%)	10	13	25	16	36
CAH (%)	3	18	22	17	40

En analysant le Tableau 3.5, nous trouvons que 22 cas (14.3%) d'échouements correspondent au couple (KMS-C2,CAH-C4), 20 cas (13%) pour le couple (KMS-C3,CAH-C5) et 56 cas (36%) pour le couple (KMS-C5,CAH-C5). Ce sont les éléments des deux clusters C3 qui amènent le plus de variabilité car ces derniers se répartissent dans les autres clusters des deux méthodes appliquées.

La correspondance temporelle indique également une forte proportion d'éléments du cluster C4 de KMS avec ceux du cluster C3 de CAH soit 17% des cas d'échouements. Cette répartition confirme l'existence des trois premiers

régimes et met en évidence un nouveau type de circulation océanique.

TABLEAU 3.6 – Pourcentage de correspondance des dates d'échouement entre les clusters de KMS et de CAH, avec une accentuation des meilleurs valeurs par cluster (en gras), et les concordances pour les deux algorithmes (en vert).

	CAH	C1	C2	C3	C4	C5
KMS	**C1**	0%	0%	**10%**	0%	6%
	C2	0%	0%	**20%**	0%	0%
	C3	**4%**	6%	2%	**26%**	0%
	C4	0%	**22%**	2%	0%	0%
	C5	0%	0%	0%	0%	**56%**

Lorsqu'ils sont rassemblés par couple, les occurrences d'apparition des circulations par trimestre, présentées dans le Tableau 3.7 montrent que celles-ci varient en fréquence au cours de l'année.

TABLEAU 3.7 – Nombre d'occurrences trimestrielles des motifs spatio-temporels détectés par trimestre.(-) pas d'occurrences.

Couples	JFM	AMJ	JAS	OND
(KMS-C2,CAH-C4)	3	19	-	-
(KMS-C3,CAH-C2)	-	20	-	-
(KMS-C4,CAH-C3)	15	10	1	-
(KMS-C5,CAH-C5)	13	2	21	19
Total	31	51	22	13

Les éléments du motif (KMS-C5,CAH-C5) se distribuent toute l'année avec un minimum marqué à la période Avril-Mai-Juin (AMJ) et un maximum en Juillet-Août-Septembre (JAS). C'est le seul motif dont les éléments apparaissent uniquement en Octobre-Novembre-Décembre (OND).

Par contre, les éléments du motif (KMS-C3,CAH-C2) n'apparaissent que pendant la période AMJ et dans une moindre mesure ceux du motif (KMS-C2,CAH-C4). Les éléments du motif (KMS-C4,CAH-C3) sont observés majoritairement au premier trimestre soit Janvier-Février-Mars (JFM).

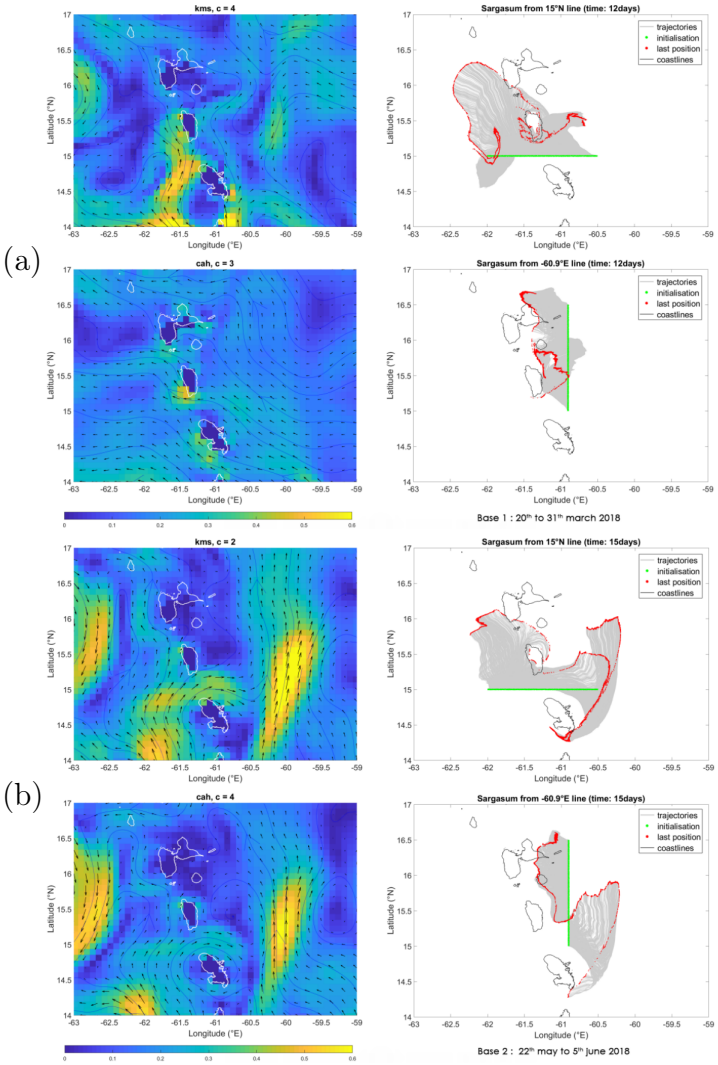


FIGURE 3.6 – (a) échouements du 22 mai au 05 juin 2018 pour le couple (KMS-C2,CAH-C4). A gauche, zoom sur les Antilles Françaises. À droite, trajectoires de 5000 particules atteignant le littoral, (b) échouements du 20 au 31 mars 2019 pour le couple (KMS-C4,CAH-C3). A gauche, zoom sur les Antilles Françaises. À droite, trajectoires de 5000 particules atteignant le littoral

3.3.5.1 Analyse locale et simulation de dérive

À l'aide des sorties modèles nous avons regroupé les multitudes de circulations océaniques de surface possibles en quatre grands types afin d'identifier les voies océanographiques utilisées par d'éventuels bancs de sargasses arrivant aux AF.

Dans cette partie nous analyserons le lien entre ces types avec les échouements. Nous avons simulé un lâcher de particules passives représentant des Sargasses

au plus à 150 km des côtes de la Guadeloupe au sud et à l'est des AF, respectivement entre -62 et -60,9°E et de 15 à 16,5°N.

Nous avons utilisé les régimes de circulations (KMS-C2,CAH-C4), (KMS-C4,CAH-C3) et (KMS-C5,CAH-C5). Ils présentaient une forte récurrence sur une plage temporelle suffisante nous permettant de réaliser un transport depuis les points sources énumérés auparavant. Les résultats obtenus sont donnés sur la Figure 3.6 (a) et (b) présentée à la page précédente.

3.3.6 Conclusion

A l'issue de cette seconde étude applicative, nous avons identifié par deux méthodes de clustering plusieurs configurations de circulations océaniques de surface, spatialement définies et récurrentes, permettant de résumer les courants principaux favorisant le transport et l'échouement de Sargasses sur le littoral des Antilles Françaises.

Quatre configurations trouvées à partir du jeu d'observations de ré-analyses Mercator sont quasi similaires tandis que deux autres sont différentes.

L'analyse détaillée des types trouvés nous apprend que d'autres facteurs de variabilité entrent en jeu. Présent le long de la pente continentale américaine, le courant nord brésilien peut être considéré comme le premier facteur de variabilité observé pour les types de circulation océanique de surface.

Au printemps boréal, il circule vers le Nord Ouest en s'étendant sur zone plus ou moins étendue, au large du Venezuela et des Guyanes,alors que le reste de l'année il se sépare des côtes sud-américaines pour se rétro fléchir vers l'est. Ce résultat est confirmé par [26, 27, 38].

L'apparition spatio-temporelle des différentes branches issues du courant Guyanais permet d'inférer qu'elles constituent le second facteur de variabilité des types de circulation trouvés. Ces branches assurent une transport méridien des eaux de surfaces vers le nord des Caraïbes entraînant les éventuels bancs présents à proximité.

Le troisième facteur s'explique par la circulation persistante et quasi-stationnaire des anneaux de NBC. D'une durée de vie de plusieurs mois, ils sont observés pour les quatre types de clusters "courant de surface" trouvés, mais se localisent à différentes latitudes . Ces différences en latitudes laissent suggérer que les Petites Antilles modifient la translation des anneaux NBC, orientée initialement vers le nord-ouest, vers un tracé plus au nord parallèle à l'arc insulaire. Ce constat a été fait par [38].

De plus, l'un des types de circulation trouvé dans nos travaux contient une

branche située plus à l'est assurant conjointement un transport méridien supplémentaire des eaux de surface d'origine Sud-Atlantique. Ces deux voies croisent, aux latitudes des AF, la composante sud du flux géostrophique à grande échelle, qui est suffisamment différente pour être capable de favoriser la translation des bancs vers le littoral de ces îles.

Notre attente initiale était de retrouver les types de circulation de surface susceptibles d'être à l'origine des échouements observés sur le littoral des Antilles françaises. Le jeu des ré-analyses Mercator qui fournit la description de la circulation quotidienne des couches océaniques de surface a été utilisé. La région située au sud-est des Antilles orientales a été choisie pour son rôle important dans le transport des eaux de surface du sud atlantique vers les Caraïbes. Ces eaux marines contiennent une grande partie des bancs de sargasses qui s'échouent sur les Antilles.

La forte variabilité spatio-temporelle de cette région océanique a pu être résumée par un nombre fini de type de circulation océanique de surface. Nous avons établi les premiers liens entre ces types et les impacts comme les échouements sur le littoral. Par des algorithmes de clustering, nous avons donc suscité l'intérêt de développer et de tester des méthodes alternatives pour trouver des processus d'aide à la décision rapides. Les résultats présentés sont une première partie de ce travail qui avait pour objectif global de proposer des outils potentiels permettant d'indiquer les conditions favorables aux échouements voire de les anticiper.

Les résultats de cette étude croisé ont donné lieu à un article et un poster au Congrès Français de Mécanique en Août 2019.

Pour revenir à la présente, il est intéressant de noter qu'avec du recul, cette étude montre l'importance de comparer les clusters sur la base de l'élément le plus proche du centroïde. L'évaluation visuelle semble bien meilleure. Notons également que l'usage de le distance de Ward associé à CAH a donné des résultats pertinents.

3.4 Synthèse des premières applications

Les différentes applications mises en œuvre dans ce chapitre grâce aux méthodes classiques de clustering ont permis d'apporter de l'expertise sur l'identification des récurrences dans les circulations atmosphériques et océaniques dans la région des Petites Antilles mais également d'améliorer les connaissances sur le phénomène d'échouement de banc d'algues sargasses sur les côtes des îles de la zone.

Ces études ont également permis de constater un certain nombre de particularités propres aux processus d'analyse climatique par clustering de données modélisant des paramètres météorologiques et océaniques. Plus particulièrement la nécessité de renforcer l'analyse visuelle de l'expert par des méthodes automatisées d'analyse d'image et d'évaluation de la qualité des clusters produits.

En effet, nous avons également effectué des mises en concordance des résultats d'algorithmes différents. L'objectif étant d'associer les clusters de ces méthodes entre eux en se basant sur leurs centroïdes (ou l'élément le plus proche de celui-ci) et leurs éléments constitutifs. Cette approche s'est révélée concluante pour les deux études.
Notons cependant que l'évaluation de l'homogénéïté interne nous permet de comparer les méthodes entre elles, afin déterminer laquelle est plus adaptée au jeu de données utilisé.

L'interprétation des résultats par des experts du climat a été grandement améliorée en ne considérant que l'élément réel le plus proche du centroïde. Celle-ci permet dans un sens de vérifier la cohérence des résultats obtenus par clustering puis par analyse.

Nous avons vu également que les clusters issus des deux méthodes de clustering doivent être caractérisés spatialement et temporellement afin de comprendre leurs impacts sur le climat. L'aspect temporel n'avait encore été évoqué jusque là, mais par la suite, dans les chapitres suivants nous allons systématiquement chercher à l'analyser.

Avec du recul, il est tout de même important de noter que dans ces deux études l'évaluation de la qualité des clusters n'est pas assez rigoureuse, car réalisée de manière imparfaite (ratio variance intra/extra) pour la première étude ou de façon indirecte dans la deuxième étude. C'est un défaut que nous retrouverons dans les études traitant des régimes de temps et auquel nous apporterons une solution partielle en fin du chapitre 4 suivant.

Dans ce chapitre, nous allons faire un état de l'art des travaux effectués dans la Caraïbe par les experts du climat. Dans toutes ces études, la méthodologie classique d'analyse par clustering (comme présenté en chapitre 2) a été appliquée sur des données climatiques (de la même manière que ce présent chapitre) afin de détecter des configurations spatio-temporelles récurrentes appelées "Régimes de Temps", pour un autre type de champ.

Il s'agit du cœur cette thèse, nous allons donc prendre le temps d'analyser les résultats des scientifiques ayant travaillé sur notre zone d'étude. Puis nous allons mettre en évidence des points d'amélioration que nous expérimenteront par la suite. Cela nous permettra d'entrevoir la principale difficulté de la méthodologie classique, la cohérence physique des clusters obtenus. Nous serrons alors donc en mesure (dans le chapitre 5) de proposer une nouvelle méthodologie produisant des résultats plus pertinent au point de vue physique.

Chapitre 4

Identification des régimes de temps par clustering

4.1 Introduction

L'objectif principal de ces travaux de thèse est de mettre en œuvre des méthodes d'apprentissage automatique afin d'identifier les régimes de temps prédominants dans la zone Caraïbe en utilisant des données historiques. Ces régimes de temps sont déterminants à long terme dans l'étude des saisons et du climat, mais également à court moyen terme dans une dynamique opérationnelle, pour :

- comprendre et anticiper les situations météorologiques à venir puisque les conditions météorologiques régionales sont guidées par ces structures persistantes.

- compléter l'apport en connaissances météorologiques et climatiques localisés des zones géographiques comprenant de vastes étendues maritimes et de faibles surfaces terrestres. En effet, le nombre de stations d'analyses météorologiques est très faible dans la Caraïbe. L'accès aux données satellitaire est donc primordial pour initier toute étude dans cette région.

- comprendre leurs possibles rôles dans le déplacement et l'intensification des phénomènes cycloniques, aléas atmosphériques coutumiers de cette région du globe.

Ce chapitre s'appuie donc sur la littérature scientifique pour faire dans un premier temps l'état des lieux des principales méthodes informatiques utilisées dans les travaux visant à identifier des régimes de temps dans cette région du globe. Nous verrons que bien que les algorithmes changent, toutes ces études suivent la méthodologie classique présentées en chapitre 2.

Puis, dans un second temps, les principaux résultats obtenus pour ces études sont présentés. Ils sont utilisés comme socle d'une analyse scientifique permettant de faire émerger des pistes d'améliorations que nous présentons. En particulier une évaluation numérique de la qualité sera proposée en sous-section 4.4.1.1.

4.2 Définition du régime de temps

Les régimes de temps sont les briques élémentaires permettant de décrire les saisons et plus largement le climat. En d'autres termes, il s'agit de configurations spatio-temporelles des circulations atmosphériques produisant ensemble de situations météorologiques dans une région géographique [57].

Ils sont donc relativement différents selon l'endroit où l'on se situe, même s'il existe des interconnexions entre eux. Ils sont liés à des circulations atmosphériques intra saisonnières de grandes échelles et ils produisent des situations météorologiques spécifiques, et peuvent être étudiés en utilisant des méthodes informatiques.

4.2.1 Régimes de temps en zones tempérées

Dans les zones tempérées, les deux plus connus sont le régime zonal et le blocage. Le blocage en Europe est responsable de grandes vagues d'air froid venant de Scandinavie vers l'Europe de l'Ouest. Le régime zonal est celui amenant le plus de dépressions sur l'Europe et notamment est à l'origine de formation des tempêtes hivernales sur cette région du globe [36].

Ces régimes de temps influencent donc le comportement des dépressions synoptiques à nos latitudes et des tempêtes hivernales en particulier. En retour, celles-ci rétroagissent sur l'évolution du régime de temps qui les a fait naître.

Bien que cette rétroaction soit l'objet d'une intense recherche depuis les années 80 du fait de son rôle dans la dynamique du climat, sa nature exacte est encore mal connue [8, 36, 56]. De nombreuses études ont d'abord montré que l'action des dépressions permet de maintenir le régime de temps en place [76, 75, 57, 8, 51].

Par exemple, en Europe durant la période hivernale, selon le régime de temps il y a plus ou moins de tempêtes au Nord et au Sud de l'Europe [43], avec une alternance sèche (froid) et humide (cf. Fig 4.1). Ce phénomène est appelé l'Oscillation Nord Atlantique (NAO). Dans sa configuration (+) les centres d'action sont plus grands et dans la (-), les centres d'action occupent un plus faible surface par rapport à la moyenne.

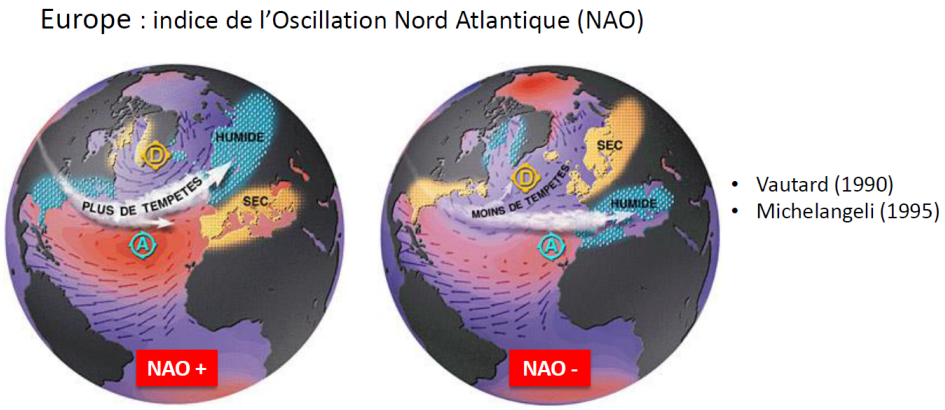


FIGURE 4.1 – Régimes de temps interagissant en hiver en Europe, liée à l'Oscillation Nord Atlantique (NAO), [D] : Dépression d'Islande et [A] : Anticyclone des Acores

NAO+ : La circulation atmosphérique est forte et le contraste de température entre le nord et le sud est élevé.
NAO- : Dans cette situation, la position de l'anticyclone des Açores et de la dépression islandaise peut varier grandement, même s'inverser, permettant l'invasion d'air arctique en Europe. Les vents sont faibles et les fortes dépressions peu nombreuses.

Ces régimes de temps ont été identifiés grâce à l'utilisation des fonctions orthogonales empiriques (EOF). Ces dernières sont extraites des données afin d'analyser le comportement des données.

En fait, le terme EOF est principalement utilisé en géophysique, en informatique l'on parle d'Analyse en Composantes Principales (ACP). Cette méthode d'analyse des données et plus généralement de statistique multivariée consiste à transformer des variables "corrélées" en variables "décorrélées". Ces nouvelles variables sont nommées "composantes principales", ou axes principaux.

D'autre part, l'ACP est utilisée pour réduire le nombre de variables et ainsi rendre l'information moins redondante. Les composantes principales retenues traduisent le plus souvent au moins 95% de la variabilité globale du nuage de point.

Généralement, les EOF sont déterminés en calculant les valeurs propres et les vecteurs propres d'une matrice de covariance d'anomalie pondérée dans l'espace d'un champ. Le plus souvent, en Géophysique, les pondérations spatiales sont le $cos(latitude)$ ou, mieux pour l'analyse des EOF, la $\sqrt{cos(latitude)}$.

Les valeurs propres dérivées fournissent une mesure du pourcentage de variance expliqué par chaque mode. Ce sont ces modes qui peuvent être alors considérés comme des régimes de temps. C'est grâce à cette méthode que les régimes de temps en période d'hiver ont été identifiés en Europe et cela en utilisant des champs de Géopotentiel et de vent [76, 75, 57].

Dans nos travaux nous n'utiliserons pas les EOF puisqu'il s'agit de configurations artificielles représentants les répartitions spatiales les plus fréquentes. De la même manière, la réduction de dimension par ACP ne sera pas utilisée puisque nous considérons qu'il s'agit d'une perte importante d'informations.

Dans la section suivante, nous présentons la littérature scientifique décrivant les travaux menés dans les zones tropicales et plus particulièrement dans la Caraïbe.

4.3 État de l'art en zone Caraïbe

Pour l'heure, les régimes de temps dans les régions tropicales ne sont pas toujours clairement identifiés pour toutes les régions. C'est notamment l'un des objectifs de ces travaux que de proposer des pistes permettant de le faire. À ces latitudes, l'identification des régimes de temps et leurs influences sur le climat, est un processus fastidieux et pas toujours concluant.

Une recherche exhaustive des travaux réalisés sur la zone Caraïbe nous a permis de dénombrer six articles à ce sujet : [74, 61, 12, 41, 53, 77]. Ils font état d'un nombre de régimes de temps variant entre quatre et onze types. Ces études s'appuient principalement sur des paramètres largement utilisés dans ce domaine (Géopotentiel, Précipitations ou encore Température de surface). Ils présentent un panel de méthodes semblables à celles utilisées dans les régions tempérées pour identifier les régimes de temps (EOF, ACP et clustering).

Parmi ces travaux, nous en avons sélectionné quatre [41, 74, 61, 12] quatre d'entre eux seront détaillés dans cette section, qui présentera les résultats obtenus par les auteurs, directement extraits de leurs articles. Cette section est la seule du manuscrit où certaines des figures présentées ne sont pas le fruit de nos travaux.

Ces quatre articles nous permettront de pointer ce qui nous apparait comme les faiblesses de l'ensemble des six références citées auparavant. Nommément, ces faiblesses sont les suivantes :

- le ou les paramètres étudiés sont-ils pertinents ?
- le nombre de régimes de temps retenus est-il justifié et raisonnable ?
- La qualité des régimes de temps retenus est-elle évaluée de façon rigoureuse ?

4.3.1 Étude de Jury et Malmgren

Commençons par cette étude [41] qui est la seule n'utilisant pas le clustering. Il faut également noter que les données utilisées dans cette étude sont des moyennes mensuelles, contrairement aux autres travaux de cette section qui

utilisent des données journalières. Elle s'articule autour sur d'une utilisation élégante des EOF comme indicateurs représentant les principales tendances. Les paramètres utilisés sont la température de surface (T), la pression de surface (SLP) et les champs de vent zonal (U) issus du modèle de réanalyse NCEP/NCAR.
Pour chacun des paramètres physiques, les auteurs choisissent les 3 premiers modes (EOF). En moyenne, ils représentent à eux seuls environ 75% (notons que c'est <95%) de la variabilité de leur ensemble respectif.

Sur la base d'une interprétation purement physique, les auteurs associent ensuite les modes de chaque paramètre. Ils obtiennent ainsi trois grandes configurations (SEAS1, SEAS2, SEAS3), chacune représentée par un triplet de modes (T, SLP, U) (cf. Fig 4.2).

L'évaluation de la qualité de ces modes est fondée essentiellement sur des réflexions physiques concernant la pertinence des triplets présentés en figure 4.2. Autrement dit, ces cartes semblent faire sens pour un physicien.

Une corrélation avec des données de précipitation est également présentée, qui, de façon indirecte, montre que ces modes peuvent être pertinents en terme d'action sur un autre paramètre. Nous retrouverons cette approche dans les autres articles. On peut parler de l'influence d'un (ou plusieurs) paramètre(s) par une évaluation en termes d'impact (les précipitations).

Comme dans l'étude de Chadee et Clarke, présentée par la suite (cf. sous-section 4.3.4, les auteurs présentent à l'appui de leurs résultats un graphe de la fréquence d'apparition de leurs modes au cours de l'année qui semble montrer que ces trois saisons sont bien réparties annuellement.

Tout d'abord, il nous semble que les trois paramètres traités sont tout à fait pertinents. De plus, les auteurs font ici une analyse prenant en compte conjointement ces trois paramètres, contrairement aux études présentées plus loin, qui se concentrent chacune sur un paramètre unique.

Le nombre de modes retenus, trois, semble faible comparativement aux propositions que nous verrons plus loin, ce qui s'explique vraisemblablement par la

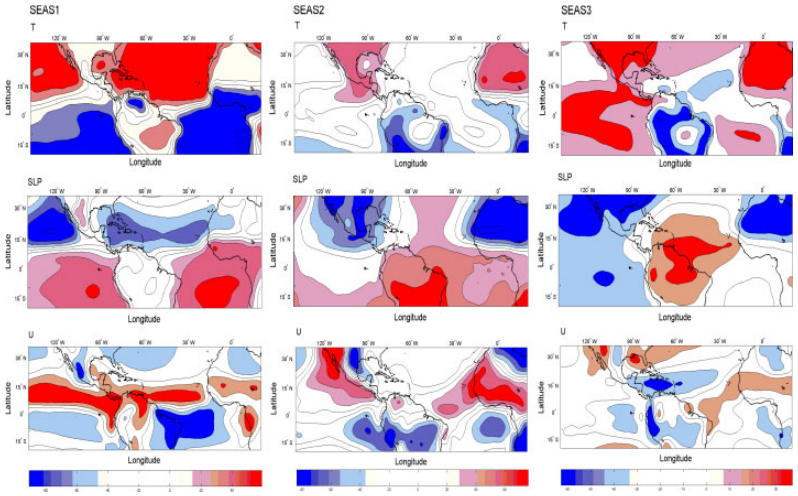


FIGURE 4.2 – Les triplets de modes retenues par Jury et Malmgren (SAES1, SEAS2, SEAS3), avec la température de surface (T), la pression de surface (SLP) et le vent zonal (U).

fréquence mensuelle utilisée par les auteurs.

Le point le plus problématique concerne selon nous l'utilisation des ACP (ou EOF). Ces méthodes, nées en mathématiques un peu avant la Seconde Guerre mondiale ont essaimé dans toutes les disciplines scientifiques avec des succès notables, ce qui explique qu'on les retrouve, encore aujourd'hui, dans de très nombreuses publications. Appliquées à notre problème, elles posent néanmoins plusieurs problèmes.

Le premier est que ce que nous cherchons est un ensemble de régimes de temps auquel une donnée appartienne (1) ou pas (0). Éventuellement, cette appartenance peut être floue et une donnée pourrait être plus ou moins associée à un régime de temps, sur la base d'une distance ou d'une probabilité. Dans le cas des ACP, on projette les données sur les modes, ce qui fournit une appartenance qui peut être positive comme négative, rendant l'interprétation des résultats très délicate. De plus, par nature, les modes obtenus ne correspondent pas exactement à notre conception des régimes de temps.

Le second problème est que ces méthodes effectuent une combinaison linéaire des valeurs trouvées à différentes mailles. Elles regroupent ainsi spatialement les mailles entre elles, de façon automatique et adaptée aux données, mais cette

association n'est guidée par aucun principe physique (de proximité spatiale, par exemple). Dans le chapitre 5, nous proposerons une solution partielle à ce problème.

Enfin les auteurs des études présentées dans les sous-sections suivantes ont écarté ces solutions pour leur préférer des techniques de clustering, plus adaptées selon eux à la grande variabilité des données climatologique de notre zone.

4.3.2 Étude de Sáenz et Durán-Quesada

Les auteurs de l'article [74] ont effectué une analyse du géopotentiel à $925h$Pa de 1998 à 2012, sur la zone des grandes îles de la Caraïbe et l'Amérique centrale, plus particulièrement sur Puerto-Rico. Pour ce faire, ils procèdent à une réduction de dimensions grâce à une ACP, puis effectuent un clustering par KMS des données dans leur nouvelle base.

Les auteurs testent ainsi la recherche de k de régimes de temps, k variant de 2 a 15. L'usage de la mesure de stabilité "indice de Rand", utilisée entre les différentes itérations de KMS, a permis de choisir le nombre de cluster k à retenir. Le nombre de clusters donnant les résultats les plus stables est de 5, mais suite à une inspection visuelle des clusters, les auteurs retiennent finalement le second choix le plus stable (k=11). Ils définissent ainsi onze régimes temps obtenus nommés et listés ci-dessous :

1 Vents hivernaux du nord-est (WNEW)

2 Fronts froids hivernaux du nord (WNCS)

3 Fronts froids hivernaux du Golfe (WGCS)

4 Tempêtes hivernales du nord-ouest (WNWS)

5 Vents de printemps anticycloniques dans le Golfe (SPAG)

6 Printemps NASH Ouest (SPNW)

7 Jet d'été à basse altitude (SLLJ)

8 Régime des vents de mousson d'été (SMWR)

9 Vents d'automne du sud-ouest (ASWW)

10 Anomalies de dipôle Géopotentiel d'automne (AGAD)

11 Anticyclone de l'Atlantique Nord Est (ENAH)

Les figures 4.3, 4.4, 4.5 présentent les onze régimes de temps obtenus. Les auteurs ont regroupé ces onze régimes en trois saisons qui correspondent respectivement à chacune des saisons proposées.

Notons que dans ces figures, le domaine géographique des données (ERA-Interim) utilisées dans l'algorithme de clustering est délimité par l'encadré blanc. Les données sont clusterisées sur la base du géopotentiel et les figures présentent ce géopotentiel à l'isobare 925hPa (les contours noirs).

Ce géopotentiel est extrêmement délicat à interpréter directement. Pour faciliter l'évaluation des modes trouvés, les auteurs présentent donc en fond de carte le champ de vent correspondant, directement impacté par ce géopotentiel et qui lui, témoigne des circulations d'air.

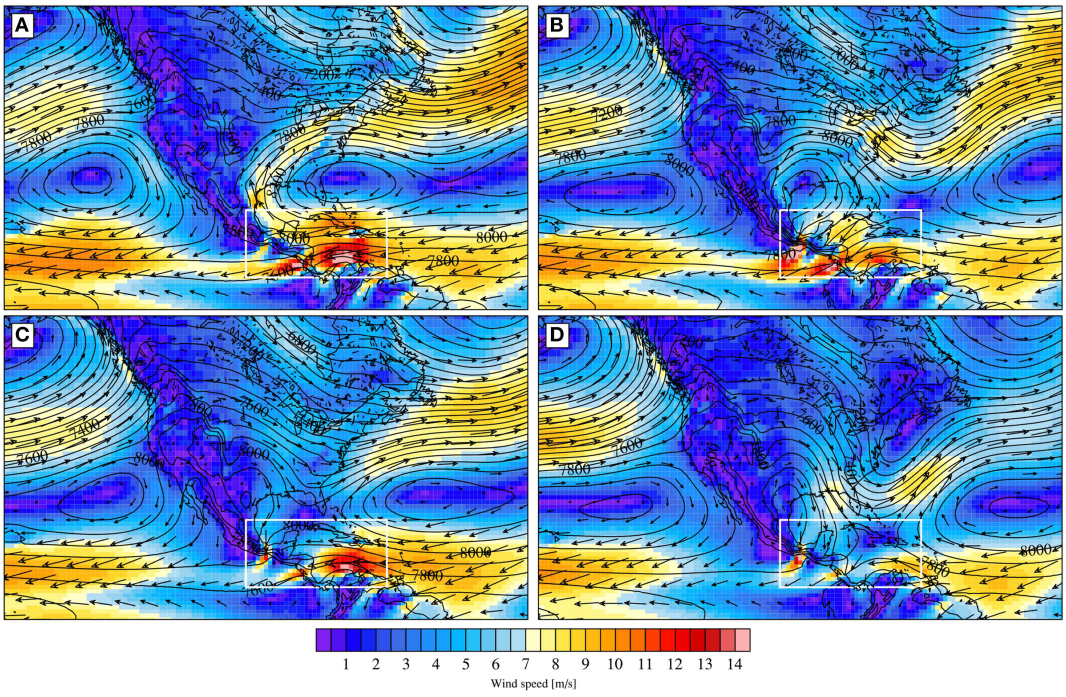


FIGURE 4.3 – Saison hivernale, composée de quatre régimes de temps : (A) WNEW, (B) WNCS, (C) WGCS, et (D) WNWS. Chaque mode est représenté par la hauteur du géopotentiel 925hPa [m^2/s^2] (contours noirs) et le champ de vent [m/s] (pixels colorés et vecteurs).

La première de nos remarques porte sur le nombre important de régimes de temps trouvés, qui nous semble montrer une recherche trop précise compte tenu des connaissances actuelles sur le sujet. De plus, l'examen des figures semble

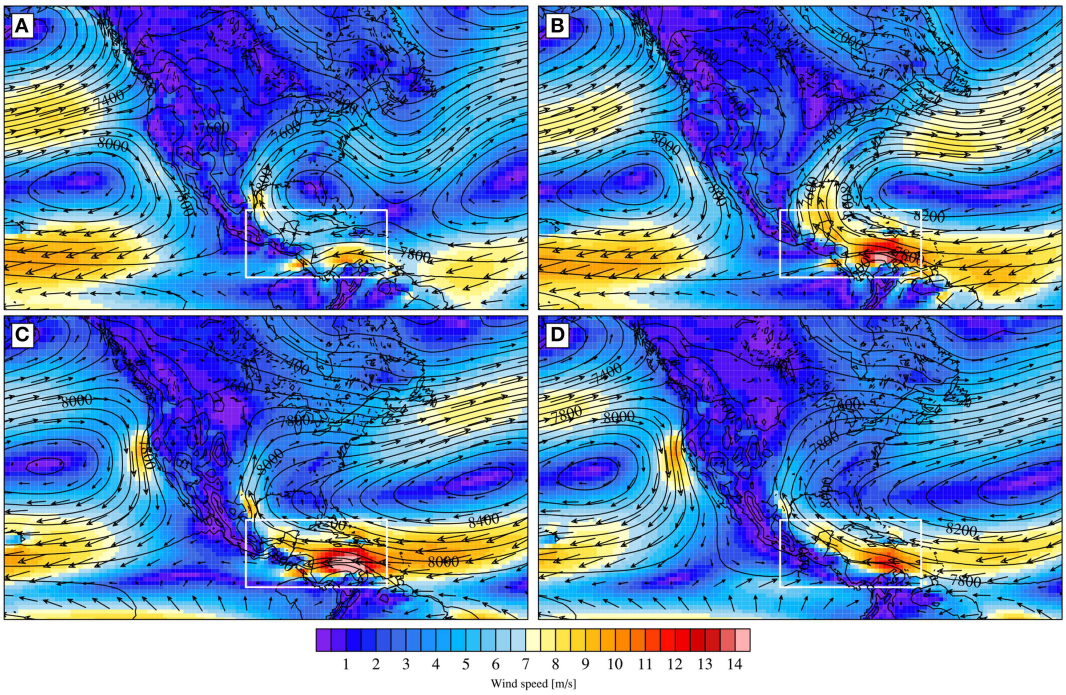


FIGURE 4.4 – Saison printanière : (A) SPAG, (B) SPNW, (C) SLLJ, et (D) SMWR. Chaque mode est représenté par la hauteur du géopotentiel 925hPa [m^2/s^2] (contours noirs) et le champ de vent [m/s] (pixels colorés et vecteurs).

montrer de grandes similitudes entre certains modes (ex. SPNW, SLLJ, SMWR dans la figure 4.4). La mesure de stabilité seule pour justifier ce choix de k nous semble également relativement légère.

Il faut aussi souligner que le choix du paramètre géopotentiel à 925hPa n'est peut-être pas judicieux quand on sait que ce paramètre est plutôt stable à ces latitudes. Néanmoins, c'est un bon marqueur des centres d'action classiques (anticyclones et dépressions). Cependant, la météorologie dans notre zone d'étude dépend beaucoup plus directement des circulations atmosphériques et des précipitations, bien qu'il existe un lien entre ces paramètres et les centres d'action.

Enfin, l'évaluation des modes trouvés est délicate, selon les auteurs. Ne disposant pas d'un critère numérique, validant la cohérence physique des clusters, ils s'appuient sur une inspection visuelle des modes trouvés pour les évaluer. Cette démarche, très physique et nécessaire pour valider les résultats trouvés, nous parait néanmoins très fastidieuse voire incertaine.

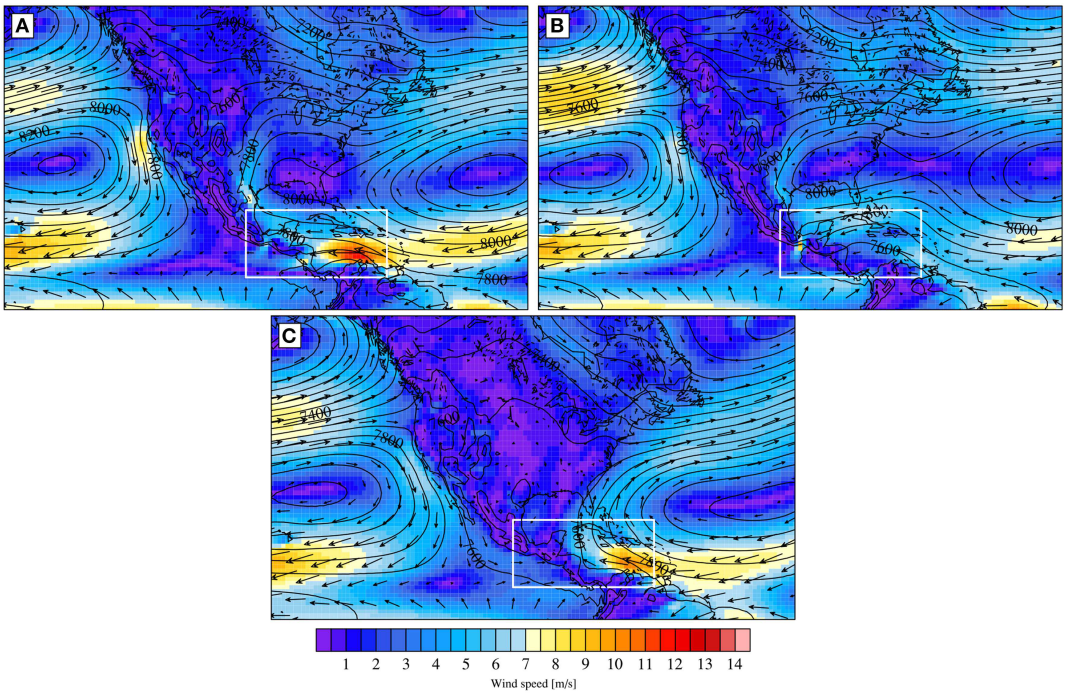


FIGURE 4.5 – Saisons automnale : (A) ASWW, (B) AGAD et (C) ENAH. Chaque mode est représenté par la hauteur du géopotentiel 925hPa [m^2/s^2] (contours noirs) et le champ de vent [m/s] (pixels colorés et vecteurs).

Conscients de ces difficultés, les auteurs appuient leur propos par une évaluation indirecte de leurs modes : les journées ayant été labellisées, ils observent, pour chaque cluster, le champ moyen de précipitations des journées correspondantes en utilisant des données satellites issues du projet TRMM. Selon les auteurs, ces onze régimes de temps produisent ainsi onze types de champs de précipitations présentés en Figure 4.6.

Un des points les plus problématiques est que la seule représentation des régimes de temps proposée reste le centroïde des clusters. Nous avons déjà noté des similarités entre ces centroïdes, témoignant d'une mauvaise séparation des clusters. En clustering, l'autre aspect important de la qualité des clusters concerne leur homogénéité ou toute mesure témoignant de la variabilité interne des clusters. Celle-ci n'est absolument pas évoquée dans ces articles. Cette remarque vaut pour les clusters initiaux (de géopotentiel) comme pour les clusters d'impacts (concernant le vent et les précipitations) et obtenus indirectement.

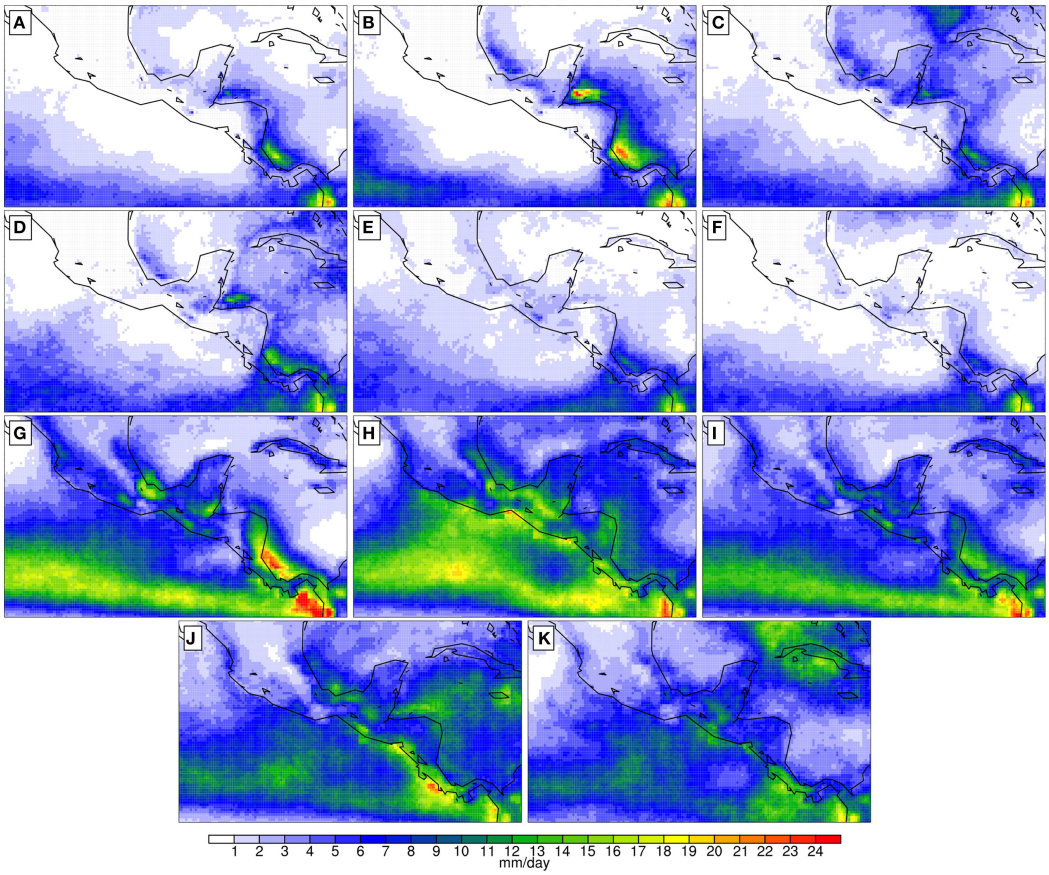


FIGURE 4.6 – Champs des cumuls à la surface des précipitations journalières TRMM-3B42 (mm/jour) pour les régimes (A) WNEW, (B) WNCS, (C) WGCS, (D) WNWS, (E) SPAG, (F) SPNW, (G) SLLJ, (H) SMWR, (I) ASWW, (J) AGAD, et (K) ENAH.

Les autres études ont donc procédé à des analyses par clustering sur les paramètres influençant directement la météorologie de la zone. L'étude suivante propose de huit régimes de temps grâce à un clustering sur l'intensité du vent et des données radars modélisées à 925hPa [61].

4.3.3 Étude de Moron et Gouriand

L'étude de Moron et Gouriand [61] s'intéresse à des données de précipitations obtenues par télédétection radar (OLR) dans le bassin caribéen. Les auteurs réduisent leurs données par ACP avant d'appliquer l'algorithme KMS. Ils testent également un nombre variable de clusters possibles. Le même critère de stabilité que dans l'étude précédente est utilisé pour justifier le nombre de clusters

retenu.

Ils définissent ainsi huit régimes de temps qui ne sont pas présentés ici. Comme dans l'article précédent, on trouve dans l'article une illustration des centroïdes qui est analysée d'un point de vue physique pour justifier l'intérêt de la méthode.

L'évaluation des clusters est également partielle, puisqu'elle repose sur une étude indirecte des clusters en observant leurs impacts (sur le vent et la température de la mer). Les auteurs s'appuient également sur les similarités de leurs résultats avec ceux de l'étude précédente (cf. sous-section 4.3.2) pour assoir la validité de leurs régimes de temps. Compte tenu des proximités méthodologiques de ces deux études, les conclusions que nous avons présentées dans la sous-section précédente restent valables.

Le point intéressant de cette étude est l'utilisation de paramètres directement liés à la météorologie considérés donc potentiellement comme de bons marqueurs des régimes de temps [56]. On peut constater que le nombre de régimes de temps est également important, mais inférieur à l'étude précédente.

4.3.4 Étude de Chadee et Clarke

Enfin, l'une des études les plus récentes porte sur les types de circulations atmosphériques (le vent). Les auteurs effectuent un pré-traitement de leurs données pour combiner intensité et direction du vent avant de les réduire par ACP. Ils appliquent alors un algorithme CAH qui permet d'initialiser les centroïdes pour l'algorithme KMS utilisé par la suite pour fournir les clusters attendus.

Pour sélectionner le nombre de clusters, les auteurs observent la somme des variances intra-classes et cherchent un coude dans la courbe classique des éboulis. Cette méthode du coude leur permet d'extraire sept configurations récurrentes (cf. Fig 4.7) influençant ou étant directement les régimes de temps dans la Caraïbe [12].

Un des ajouts notables de cette étude concerne l'évaluation des clusters. Les

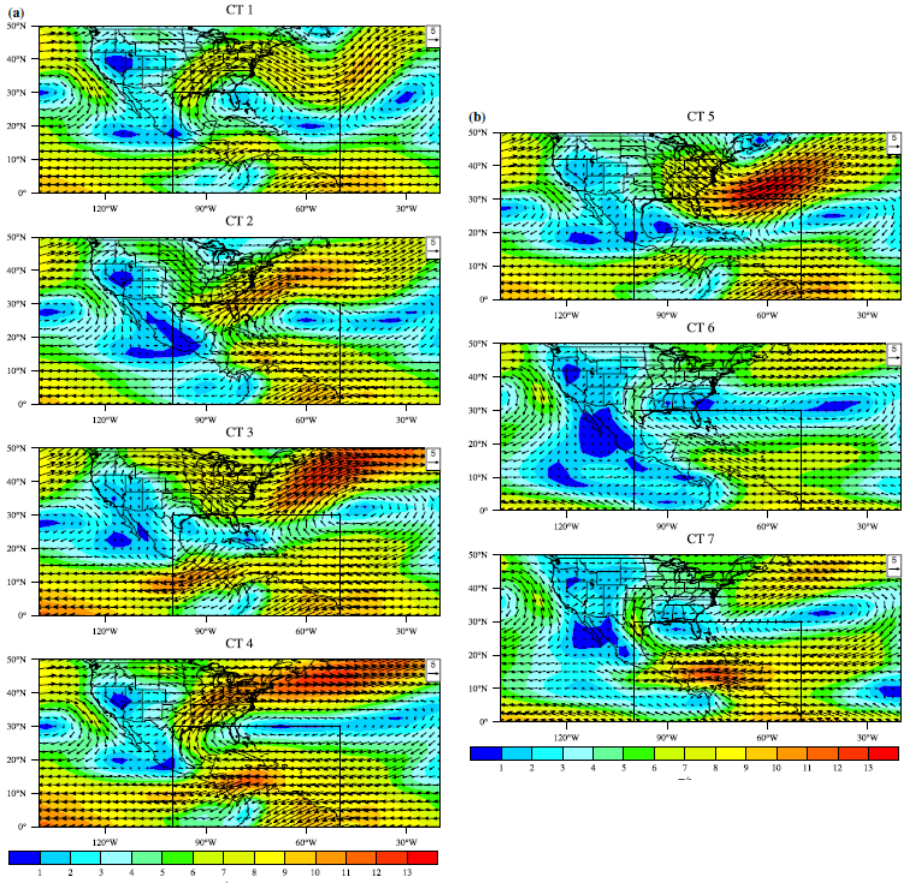


FIGURE 4.7 – Les sept types de circulations atmosphériques retenus dans l'étude de Chadee et Clarke

auteurs présentent en effet la répartition temporelle des journées de chaque cluster comme indicateur de qualité. Celle-ci, extraite de leur article, est présentée en figure 4.8.

Cette présentation de la dynamique temporelle des clusters nous semble d'autant plus méritante qu'elle montre très clairement les limites de l'étude. En effet, certains clusters ont à quelques détails près la même dynamique temporelle et la même fréquence d'apparition (cf. Figure 4.8).

4.3.5 Conclusion sur l'état de l'art

L'ensemble des travaux présenté dans cette partie témoigne d'une extrême rigueur dans l'analyse physique des configurations trouvées avec des résultats présentés sont ayant une grande utilité pour la communauté des climatologues.

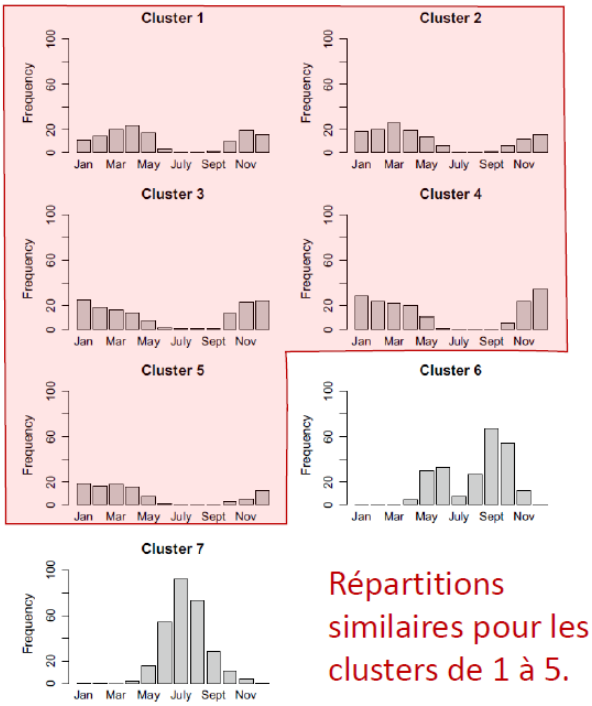


FIGURE 4.8 – Répartition temporelle de chaque cluster dans l'étude de Chadee et Clarke. Elle permet de voir l'étalement des clusters dans le temps, et la fréquence avec laquelle ces clusters apparaissent.

Force est de constater, qu'il n'existe pas encore de consensus sur les régimes de temps. C'est une préoccupation que nous tenterons de solutionner dans ces travaux.

Néanmoins, selon nous ces études présentent les défauts patents qui avaient été annoncés dès l'introduction de cette section, le plus critique étant selon nous l'évaluation de la qualité des clusters.

Tous ces articles observent principalement les modes ou les centroïdes retenus, sans se poser de questions sur l'homogénéité interne des clusters. Elle n'est abordée qu'indirectement par le biais d'études d'impact. Cette méthodologie nous semblent problématique. En effet, comme la montre la figure 4.9 (à la page suivante), à l'issue d'un clustering, même si l'on obtient une bonne homogénéité interne et une bonne séparation des clusters pour un paramètre (à gauche), cela ne garantit pas forcement que cela soit le cas dans l'espace de l'impact (à droite).

Si nous pouvions disposer d'une évaluation de la qualité des clusters, quantifiée numériquement, un grand nombre de ces problèmes serait réglé. La section suivante est consacrée à ce point.

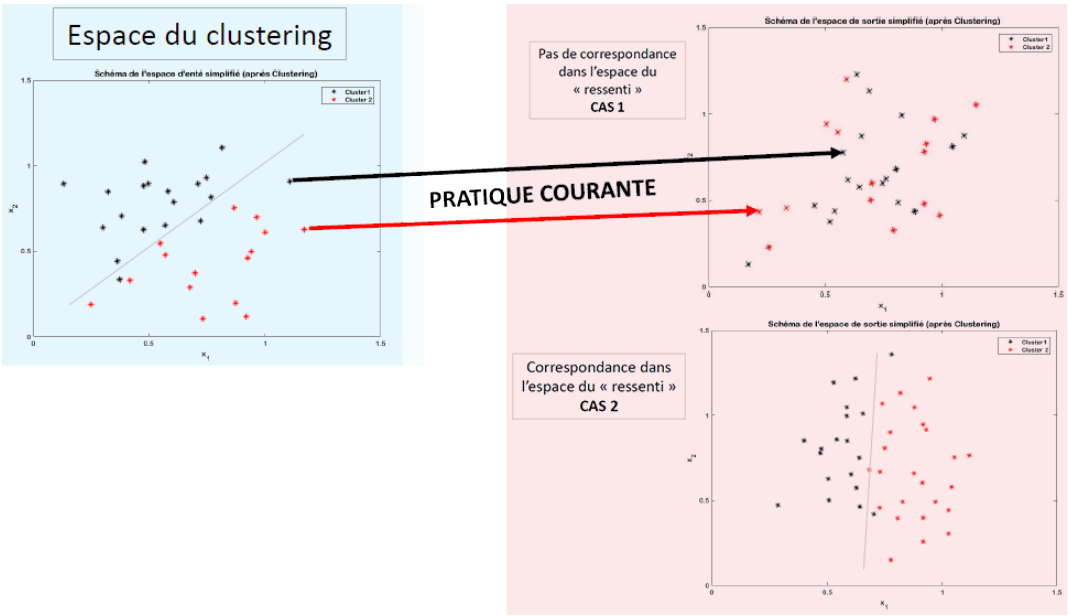


FIGURE 4.9 – Schéma de l'analyse classique des impacts d'un clustering effectué avec un paramètre (espace de clustering) sur un autre paramètre (espace de ressenti ou d'impact), cas1 : inexistence d'une séparation dans l'espace des impacts, cas2 : existence d'une séparation dans l'espace des impacts

4.4 Nos travaux

Ici, nous présenterons les résultats que nous avons obtenus par application d'algorithmes de clustering (CAH, KMS). Nous souhaitions néanmoins, avant cela, apporter une solution au moins partielle au problème de l'évaluation de la qualité des clusters.

Disposer d'une évaluation numérique du clustering nous semble primordial. Cela permettrait en effet :

a. de s'assurer que les clusters constitués sont pertinents et de garantir qu'il existe bien une organisation interne des données,

b. de valider le nombre de clusters obtenus,

c. de comparer différentes méthodes entre elles, sur une base numérique.

Dans les études précédentes, ces comparaisons et évaluations reposent exclusivement sur les interprétations des experts physiciens du domaine. Nous souhaitons ici promouvoir une mesure quantitative de cette évaluation.

Cette section présentera nos travaux en ce sens. Nous allons donc choisir un indicateur de qualité dans la sous-section 4.4.1. Nous appliquerons alors les algorithmes de clustering à des données de vent (cf. sous-section 4.4.2) , puis à des données de précipitations (cf. sous-section 4.4.3).

Dans ces sous-sections, l'organisation est semblable : dans un premier temps, nous évaluerons numériquement les clusters obtenus par les deux méthodes de clustering. Dans un second temps, il s'agira de montrer que l'évaluation numérique correspond bien à un niveau de qualité des clusters telle qu'elle serait évaluée par un expert. Nous procèderons donc à une inspection visuelle détaillée des clusters.

4.4.1 Évaluation du *clustering*

Il existe un grand nombre de méthodes et d'indices d'évaluation du clustering [49]. En l'absence d'informations sur la vérité du terrain, voici une liste de quelques mesures internes de la qualité qui peuvent être utilisées : l'indice

de silhouette [69], l'indice de Calinski–Harabasz [10] et l'indice de Davies–Bouldin [17].

Tous ces indices combinent les deux qualités essentielles suivantes attendues des clusters représentant différentes situations physiques : compacité interne des clusters (homogénéité) et distance séparant les clusters entre eux (séparation).

Nous avons choisi l'indice de silhouette parce qu'il présente les avantages suivants :

- Il peut fournir une mesure de qualité pour chaque jour au sein de son cluster, mais aussi une mesure de qualité de chaque cluster, ainsi qu'une mesure générale de qualité pour la méthode.
- Il est facile à interpréter en termes de pertinence des clusters.

4.4.1.1 Indice de silhouette

L'indice de silhouette désigne une méthode d'interprétation et de validation de la cohérence au sein de groupes de données. Cette technique fournit une représentation succincte de la façon dont chaque objet est bien intégré dans son groupe. Elle a été décrite pour la première fois par [69, 19, 42].

La valeur d'indice silhouette est une mesure de la similarité d'un élément avec son propre groupe (homogénéité) par rapport à d'autres groupes (séparation). L'indice varie de -1 à 1, une valeur élevée indiquant que l'élément est bien adapté à son propre groupe et mal adapté aux autres. L'indice est défini comme suit :

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i)} \tag{4.1}$$

ou encore

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases} \tag{4.2}$$

où $a(i)$ représente la distance moyenne entre un point et son cluster. Cette valeur est calculée via l'équation 4.3. La variable $b(i)$ représente la distance minimale entre un point et un autre cluster, cette valeur est calculée via l'équation 4.4, $s(i)$ est donc l'expression de l'indice de silhouette, avec $-1 \leq s(i) \leq 1$.

$$a(i) = \frac{1}{n} \sum_{j=1}^{n} \sqrt{(p_j - p_i)^2} \qquad i \neq j, \quad \forall p_i, p_j \in \mathcal{G} \tag{4.3}$$

où p_i et p_j sont deux-points du même groupe $\mathcal{G}$, n l'effectif de ce groupe, i et j varie de 1 à n.

$$b(i) = min(\sqrt{(p_j - p_i)^2}) \qquad i \neq j, \quad \forall p_i \in \mathcal{G}, \quad p_j \notin \mathcal{G} \tag{4.4}$$

où p_i et p_j sont deux points de groupes différents, $\mathcal{G}$ étant le groupe de p_i.

Deux autres critères, construit à partir de l'indice silhouette, peuvent être utilisés pour analyser les résultats du clustering. Le premier indique la qualité moyenne d'un cluster C_i, et peut être calculé en utilisant la formule suivante :

$$Sc(C_i) = \frac{1}{|C_i|} \sum_{j \in C_i} s(j). \tag{4.5}$$

Afin d'évaluer tous les clusters obtenus en appliquant une méthode de clustering $\mathcal{M}_k$ à nos données, nous pouvons considérer la moyenne des coefficients de chaque cluster, comme étant :

$$Sa(\mathcal{M}_k) = \frac{1}{k} \sum_{i=1}^{k} Sc(C_i). \tag{4.6}$$

Par définition, de la même manière que $s(i)$, chacun de ces coefficients est compris entre -1 et 1 .
Voici quelques valeurs de référence communément acceptées que nous utiliserons plus tard :

- Les valeurs supérieures à 0.20 indiquent une bonne performance (existence des clusters pertinents et bien séparés),
- Les valeurs inférieures à 0.10 indiquent le contraire,
- Les valeurs négatives indiquent que de nombreux points sont attribués à des clusters qui ne représentent pas le meilleur choix possible.

À partir de là, nous allons systématiquement utiliser l'indice de Silhouette pour produire une évaluation numérique de la qualité des clusters.

4.4.2 Clustering des données de vent

La première série de données provient de la réanalyse des sorties quotidiennes du modèle ERA-5 du ECMWF, il s'agit de l'intensité du vent à 850hPa. La direction n'a pas été considérée dans le clustering, mais elle sera utilisée pour la visualisation des configurations présentées.

La zone géographique est comprise entre -66,25 et -20,25°E et entre 5 et 30°N (Fig 5.5). Avec une résolution de 0.25°, chaque jour est donc représenté par un champ de 101 x 189 valeurs, qui ensuite sont transformés en un vecteur de 19 089 composantes. Les données couvrent une période représentant une base de 13 140 jours.

Comme annoncé précédemment, nous comptons nous appuyer exclusivement sur le coefficient de Silhouette pour évaluer la qualité des clusters trouvés (cf. sous-section 4.4.2.1). Nous vérifierons ensuite que cet indice est pertinent d'un point de vue physique par une inspection détaillée des clusters (cf. sous section 4.4.2.2).

4.4.2.1 Évaluation numérique

La figure 4.10 présente ainsi l'évolution de l'indice de silhouette, en fonction du nombre de clusters k cherché, obtenu par les algorithmes CAH et KMS sur les données de vent.

On observe que l'indice de KMS est systématiquement et notablement supérieur à celui de CAH, car des valeurs inférieures à 0.1 sont les plus fréquentes. Elles signalent l'absence de structures pertinentes au sein des données. Nous privilégierons donc par la suite, l'utilisation de KMS, rejoignant ainsi les chercheurs de la discipline, à ceci près que nous étayons ce choix par une mesure aisée à comprendre. C'est l'un des aspects primordiaux de cette évaluation que de permettre de comparer les méthodes entre elles.

Si l'on se concentre sur la courbe de KMS-L2, entre k=2 et k=8 les valeurs

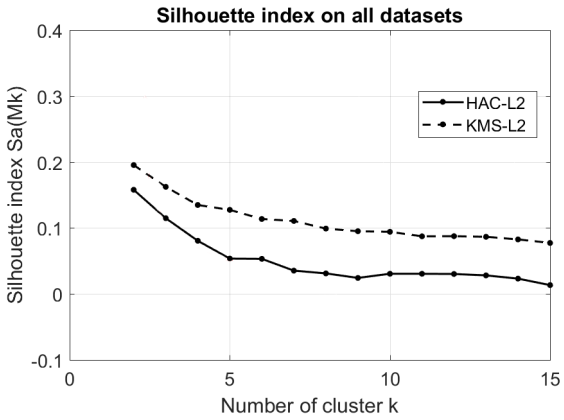


FIGURE 4.10 – Évolution de l'indice de silhouette en fonction du nombre de clusters k - HAC ou CAH (ligne continue), KMS (ligne discontinue), en utilisant L2 (noir), données ERA-5.

de l'indice sont comprises entre 0.2 et 0.1. Comme vu précédemment (cf. sous-section 4.4.1.1), cela témoigne d'une certaine cohérence dans les données et donc dans la composition des clusters, bien que celle-ci ne soit pas très bien marquée.

L'indice de Silhouette nous permet également de choisir le nombre clusters à retenir. Bien que la courbe montre une décroissance plutôt monotone, on peut noter, pour $k = 5$ une légère inflexion qui nous a conduit à retenir ce nombre comme pour la suite. Nous allons donc poursuivre avec l'inspection visuelle des 5 centroïdes de KMS.

4.4.2.2 Évaluation visuelle

La figure 4.11 présente les centroïdes des clusters. Ces derniers semblent plutôt différents, ce qui traduit une bonne séparation des différents clusters. Nous relevons tout de même que ces centroïdes donnent un aperçu, plutôt lisse, de la configuration spatiale réelle captée par chaque cluster. Cette idée sera développée plus tard.

La figure 4.12 présente la répartition par mois de ces clusters permettant d'étudier partiellement et indirectement la dynamique temporelle des clusters. Notons que contrairement aux centroïdes, certains clusters (C3, C4 et C5) ont une répartition quasi similaire. Les experts du climat de la région précisent que l'on ne voit pas clairement les effets des saisons (saisons sèche et pluvieuse).

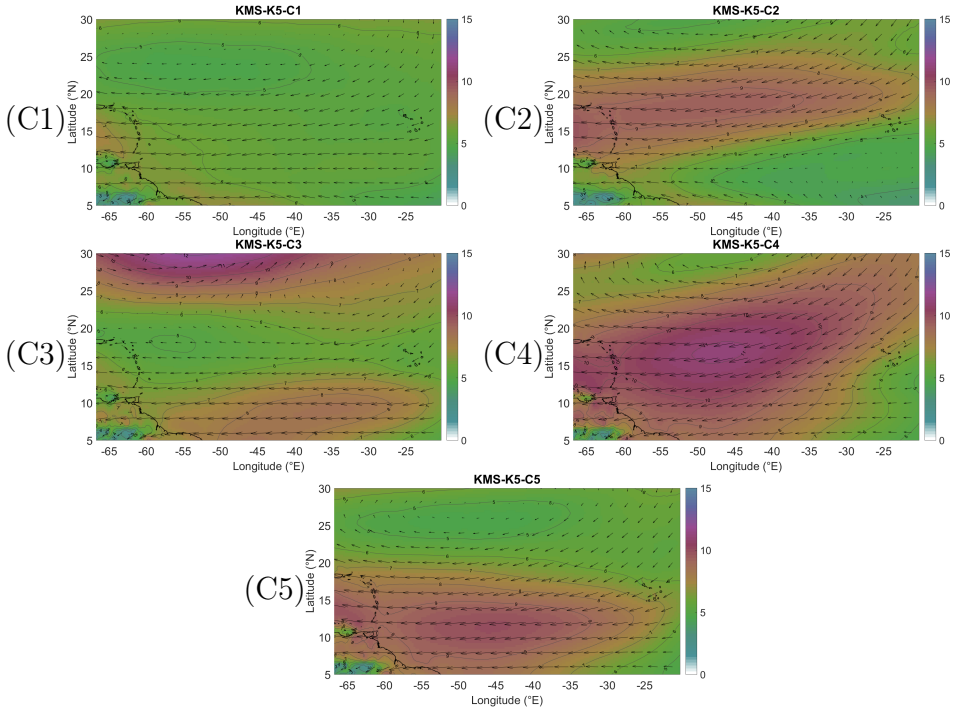


FIGURE 4.11 – Centroïdes obtenus par KMS-L2, pour l'intensité du vent à 850hPa.

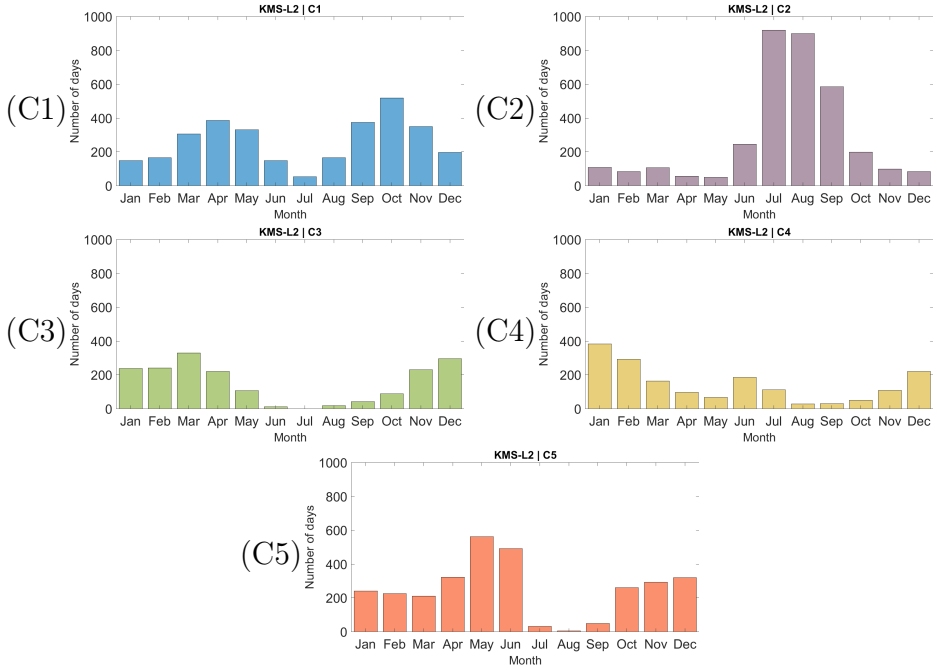


FIGURE 4.12 – Répartition par mois de l'effectif des clusters de KMS, pour l'intensité du vent à 850hPa (de 1979 à 2014).

L'inspection visuelle confirme les résultats indiqués préalablement par l'indice de silhouette. Les clusters semblent bien témoigner de structures présentes dans les données. Ces structures sont néanmoins très peu marquées et leur distribution temporelle est assez mauvaise. Nous rejoignons ici complètement les résultats des études présentées dans l'état de l'art, en ayant ajouté une évaluation numérique. Voyons maintenant ce qu'il se passe pour un autre paramètre : les précipitations.

4.4.3 Clustering des données de précipitations

Le second jeu de données provient de la réanalyse des précipitations cumulées quotidiennes mesurées par satellite par le projet TRMM de la NASA, de 2000 à 2014, pour la même zone géographique et la même résolution que le premier jeu de données. Les données couvrent une période représentant une base de 5 415 jours.

Ces données sont intéressantes, car elles présentent une différence majeure par rapport aux données concernant le vent : elles sont par nature beaucoup plus discontinues, temporellement comme spatialement. Nous verrons que ceci influe drastiquement sur les résultats obtenus. Nous appliquerons peu ou prou le même plan de présentation que dans le cas des données de vent.

4.4.3.1 Évaluation numérique

La figure 4.13 montre une décroissante rapide de l'indice de Silhouette. Les résultats de CAH sont complètement négatifs, ce qui disqualifie ici encore cet algorithme pour le clustering de nos données.

Pour KMS, seul le point obtenu pour $k = 2$ donne une valeur supérieure à 0.1, ce qui est problématique, sauf à retenir comme régimes de temps exclusivement "saison humide" et "saison sèche". Nous pouvons donc considérer que, selon l'indice de Silhouette, KMS, sur les données de précipitations, ne trouve aucun cluster pertinent.

Pour s'assurer néanmoins que ces conclusions sont valides, nous allons également procéder à une inspection détaillée des clusters. Ici encore, le nombre de clusters retenu sera de $k = 5$, correspondant à un changement de pente bien

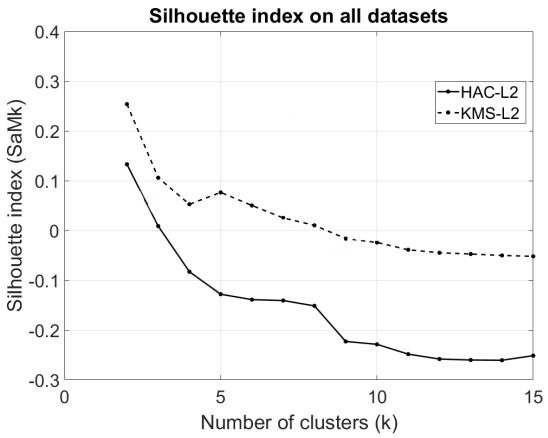


FIGURE 4.13 – Évolution de l'indice de silhouette en fonction du nombre de clusters k - HAC ou CAH (ligne continue), KMS (ligne discontinue), en utilisant L2 (noir), données TRMM.

marqué.

4.4.3.2 Évaluation visuelle

Comme précédemment, nous présentons les 5 centroïdes résultants de l'application de KMS, ce qui fait l'objet de la figure 4.14.
L'inspection visuelle de ces centroïdes montre pourtant une séparation plutôt bien établie puisque ces derniers présentent des motifs relativement différents. Par déduction les mauvais résultats de l'indice de silhouette sont vraisemblablement dus à une forte variabilité interne dans les clusters.

La figure 4.15 a pour but d'évaluer rapidement l'homogénéité des clusters. Par souci de concision, nous ne présentons qu'un cluster en particulier, il s'agit du cluster, dont la valeur, de $S(C_i)$ est la plus grande, le cluster C2.
En analysant la figure 4.15, l'on comprend assez rapidement que ce cluster semble constitué de motifs plutôt différents, incluant tous peu ou prou une zone de fortes précipitations au Sud Est, de forme très variable, a laquelle vient parfois s'adjoindre une seconde zone au Nord. Le cluster C2 est donc composé d'éléments très disparates. Les autres clusters (non présentés ici) souffrent des mêmes défauts. De plus, notons dès maintenant qu'aucune des journées du cluster ne ressemble à ce centroïde. Celui-ci est une moyenne, qui en lissant les phénomènes nous semble problématique. C'est sur cette base que nous remettrons en cause son utilisation dans le chapitre 5.

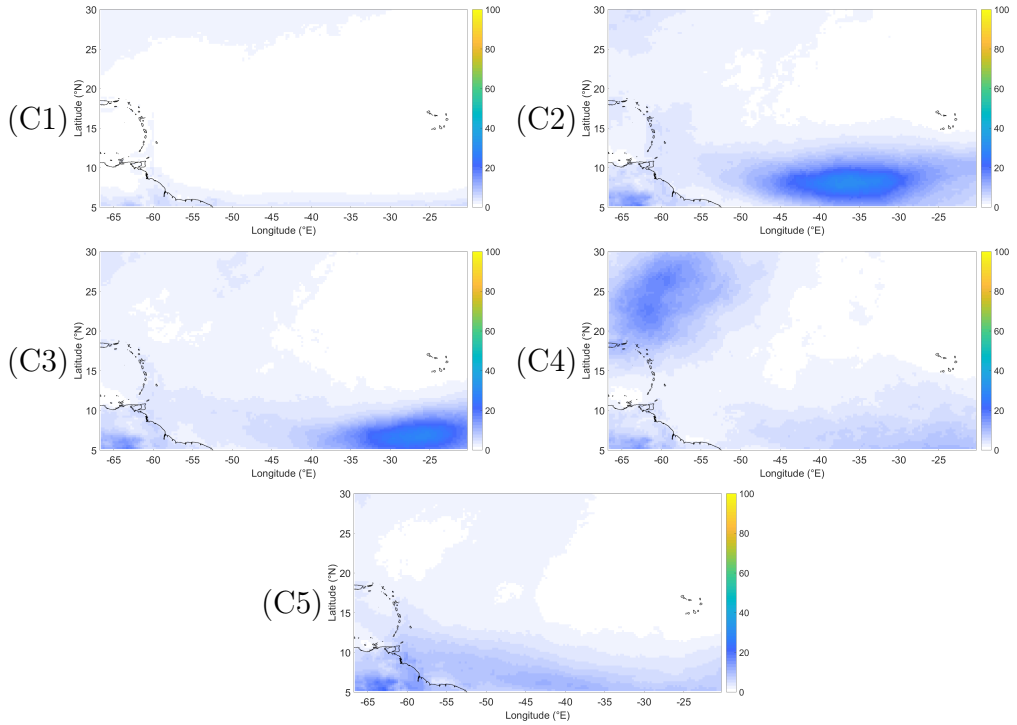


FIGURE 4.14 – Centroïdes obtenu par KMS-L2, pour le cumul de pluie journalier en surface.

Observons maintenant la répartition par mois de ces clusters. Celle-ci est présentée dans la figure 4.16. Comme pour le vent, on observe des répartitions assez similaires pour certains clusters (C2 et C3), laissant penser que ces clusters sont peu pertinents.

Pour allez plus loin dans l'analyse de la dynamique temporelle, prenons la figure 4.17. Elle montre l'évolution inter-annuelle de l'effectif des clusters. Ici encore, on peut donc observer que les clusters ne présentent aucune tendance clairement et statistiquement identifiable. Nous verrons plus loin dans ce manuscrit que les tendances (assèchement général, décalage de la saison humide) bien connues des spécialistes peuvent apparaitre sur ce type de courbe, sous réserve de disposer d'algorithmes suffisamment efficaces. Pour alléger ce manuscrit, les conclusions générales sur ces deux études seront présentées en conclusion du chapitre.

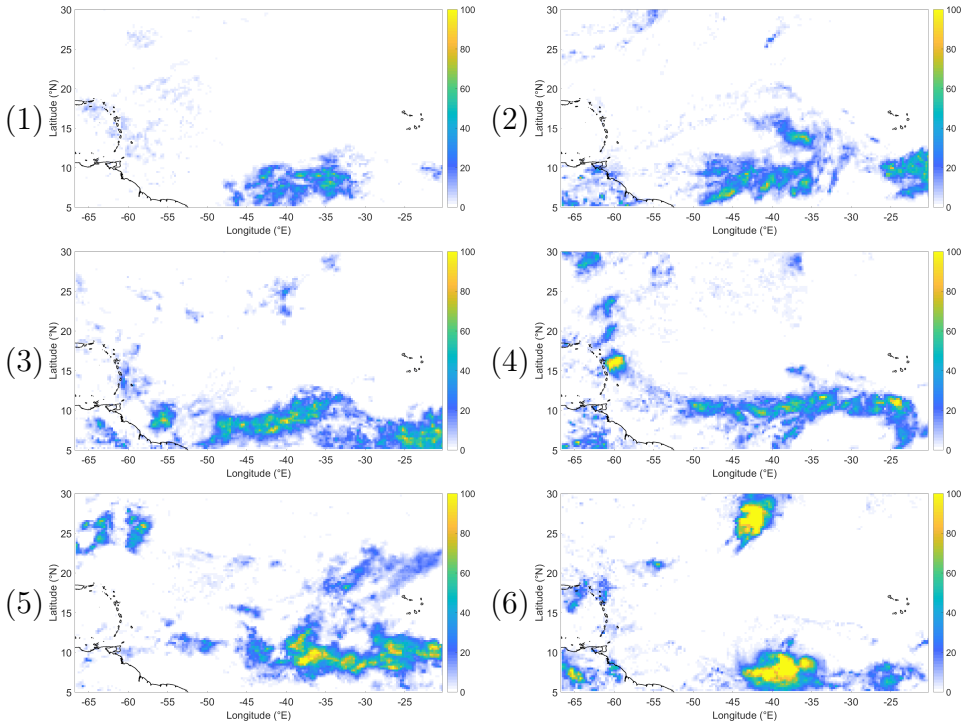


FIGURE 4.15 – Variabilité interne du cluster (C2) d'après la méthode *KMS-L2*. Six jours (1-6) de ce cluster sont présentés dans l'ordre croissant de la distance L2 par rapport au centroïde, du plus proche au plus éloigné avec un pas constant.

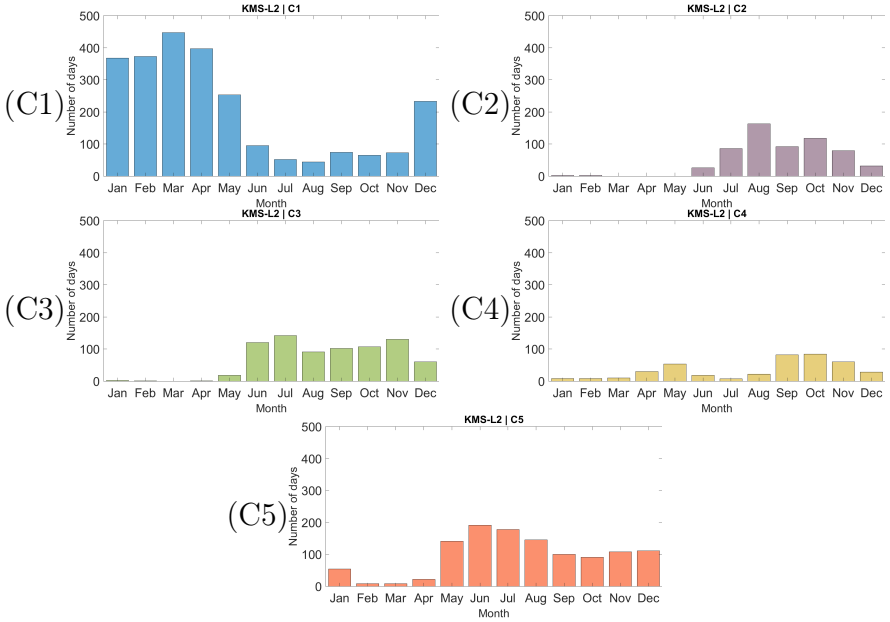


FIGURE 4.16 – Répartition par mois de l'effectif des clusters de KMS, pour le paramètre cumul journaliers de pluies (de 2000 à 2014).

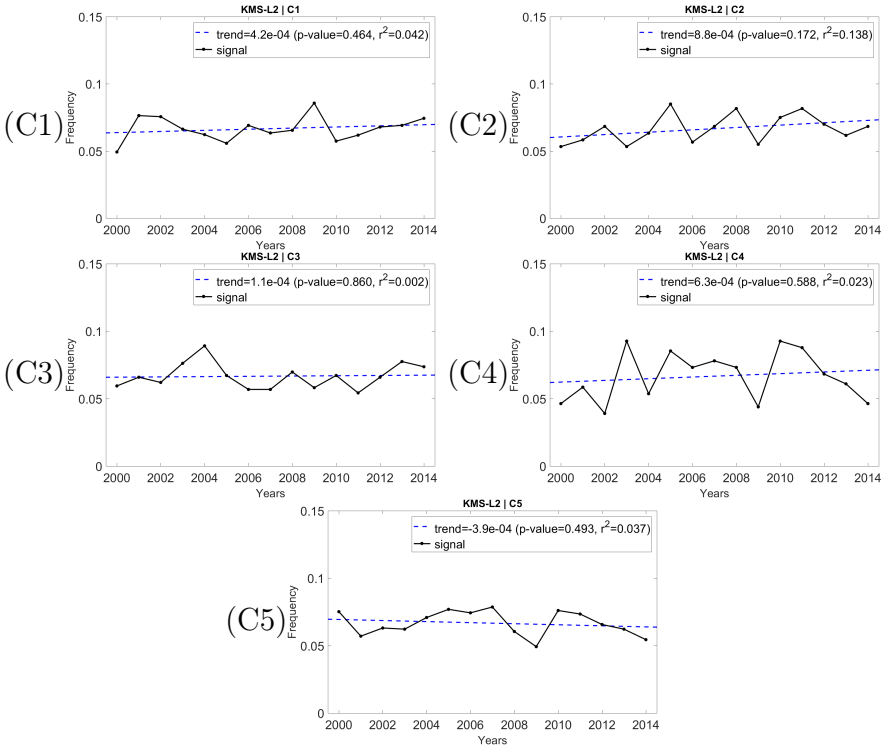


FIGURE 4.17 – Répartition par inter-annuelle (de 2000 à 2014) de l'effectif des clusters de KMS, pour le cumul journalier de pluies.

4.5 Conclusion

Ce chapitre nous a permis de présenter et d'analyser les travaux réalisés sur la recherche des régimes de temps dans la Caraïbe. Chacune de ces approches mène à des résultats permettant d'étoffer nos connaissances des données, des méthodes ainsi que des applications réalisables.

Au travers cette analyse, nous avons pu mettre en lumière un certain nombre de difficultés rencontrées par les auteurs et que nous avons également constatées lors de nos expérimentations.

Pour commencer, les paramètres sélectionnés par les auteurs ne sont pas toujours pertinents. Nous pensons que le choix du paramètre utilisé dans le clustering doit être directement impactant. Nous pensons que le géopotentiel utilisé dans plusieurs études de la bibliographie, n'impacte pas directement la climatologie dans la zone. Le nombre de clusters retenus pour ces travaux varie entre 7 et 11 ce que nous trouvons un peu sur évalué. D'ailleurs, l'on observe assez

fréquemment de fortes similitudes entre certains centroïdes.

Enfin la principale difficulté résidait dans l'absence d'une évaluation numérique pertinente de la qualité des clusters. D'ailleurs, les auteurs des études précédentes éludent systématiquement cette étape, remplacée par un amoncellement de considérations physiques, certes pertinentes, mais ne pouvant être faites que par un expert. Fort de ce constat, il nous a semblé intéressant de tester nous même des algorithmes de clustering (KMS, CAH) sur des données (vent et précipitations) non traitées et non réduites.

Nous avons proposé l'utilisation d'un indice bien connu en Intelligence Artificielle et Reconnaissance des Formes pour les problèmes de clustering : l'indice de Silhouette. Celui-ci permet entre autres de synthétiser en une seule valeur la séparation et l'homogénéité des clusters. Il offre de plus l'avantage d'être très facilement interprétable : la question de la pertinence des clusters trouvés se réduit à une comparaison à des valeurs de référence.

Ceci nous a permis de comparer les deux méthodes de clustering étudiées et ainsi d'écarter CAH dont les résultats sont toujours moins bons que KMS, et qui ne produit jamais de clusters pertinents. L'indice de Silhouette nous a aussi fourni une justification au choix du nombre de clusters retenu ($k = 5$) dans les deux jeux de données traitées, bien que ce choix ne soit pas aussi marqué que nous aurions pu le souhaiter.

Enfin, nous avons montré que si KMS fournit des résultats presque acceptables sur les données de vent, dans le cas des données de précipitations le résultat est très décevant. Nous avons attribué ceci au fait que les précipitations sont fortement discontinues dans le temps et dans l'espace.

Ces conclusions, tirées de la seule analyse des courbes de silhouette a longuement été étayée par une inspection visuelle détaillée des clusters obtenus. Cette inspection confirme en tout point que les résultats indiqués par l'indice de Silhouette donnent une indication physiquement valable de la qualité des clusters.

Considérant que nous disposons maintenant d'une mesure de qualité fiable, nous avions alors de nombreuses possibilités : La première aurait consisté à

reproduire ces expériences sur différents paramètres pour choisir ceux qui présentent les meilleures cohérences. De tels travaux seront sans aucun doute nécessaires, mais nous ont semblé relever de la simple application des techniques plus que de la recherche qui nous tenait à cœur.

Une autre possibilité consisterait à mesurer l'intérêt de combiner plusieurs paramètres (comme l'ont fait Jury et Malmgren). En clusterisant des données composées de couples ou de triplets de paramètres, il devrait être possible d'avancer un peu plus dans la compréhension des paramètres qui, conjointement, impactent le temps de nos régions.

Une troisième idée consisterait à s'intéresser à ces études d'impact qui parsèment les différentes études du domaine. Il suffirait d'utiliser la labellisation obtenue par clustering sur un paramètre (vent) comme référence pour le calcul de l'indice de silhouette du jeu de données d'impact (précipitations). Ceci permettrait par exemple de voir quel paramètre (vent, température de surface, etc.) explique le mieux tel ou tel impact (précipitation ou autre).

Ces idées, bien que très attrayantes, n'ont pas été développées dans cette thèse. Nous avons choisi de plutôt nous concentrer sur un point essentiel de nos résultats : les méthodes de clustering telles qu'elles ont été utilisées dans ce chapitre donnent des résultats au mieux médiocres. Le faible indice de silhouette des méthodes basées sur la distance L2 expliquent probablement que les études précédentes n'ont pas présenté une telle évaluation numérique.

Le chapitre suivant sera ainsi consacré à une tentative d'amélioration de cette situation.

Chapitre 5

Expert Deviation

5.1 Introduction

Ce chapitre a pour objectif d'analyser les difficultés mises en lumière dans le chapitre précédent afin d'apporter des propositions d'améliorations et ainsi de dégager une nouvelle approche plus robuste.

Il s'agit donc, dans un premier temps, de comprendre pourquoi les algorithmes de clustering les plus courants (KMS et CAH), utilisés avec un paramétrage par défaut, produisent des clusters peu pertinents physique pour des données décrivant des paramètres météorologiques. Nous nous intéressons ici plus particulièrement à la distance L2 utilisée pour comparer les configurations météorologiques lors du processus de clustering dont nous montrerons les faiblesses (cf. section 5.2.1).

Dans un second temps, nous présentons les pistes que nous avons étudiées pour pallier à ces difficultés et ainsi introduire la nouvelle méthodologie que nous avons retenue pour l'identification automatique des configurations atmosphériques récurrentes.

Elle repose sur la conception et l'usage d'une mesure de dissimilarité nommée Expert Deviation (ED) dans le processus de clustering. Cette mesure intègre l'usage d'une analyse par patch, commune en analyse d'image. Le découpage de ces patchs s'appuie ici sur une expertise physique. La dissimilarité ED synthétise alors les similarités des distributions de chaque patch.

En intégrant ED à la place de la L2 dans les méthodes de clustering, nous montrons que les clusters ainsi constitués présentent une pertinence beaucoup plus grande. Celle-ci est évaluée tant par l'indice de Silhouette vu au chapitre précédent, que par une inspection physique.

Les perspectives d'usage ou d'amélioration de cette méthode ainsi que des possibilités d'applications dans d'autres domaines sont présentées en fin de chapitre.

5.2 Difficultés du clustering de données climatiques

Les données climatiques sont relativement complexes puisqu'elles décrivent des paramètres météorologiques dans l'espace et dans le temps. Pourtant, durant le processus de clustering, ces données sont comparées entre elles, en supposant l'indépendance inter-composantes.

D'autre part, dans la majorité des algorithmes de clustering, des moyennes d'occurrences sont calculées et utilisées pour représenter des groupes ayant des caractéristiques spatiales similaires. Ces moyennes sont purement artificielles, elles ne représentent pas des situations réelles, pourtant elles sont systématiquement utilisées, par les scientifiques, pour décrire des régimes de temps.

Afin de mieux comprendre les difficultés du clustering, nous avons analysé les effets des distances utilisées pour comparer des données. En effet, celles-ci permettent d'exprimer une mesure de la similarité ou la dissimilarité entre les occurrences. Pour ce faire, nous avons choisi d'étudier la norme L2, puisqu'il s'agit de la distance la plus fréquemment utilisée pour le clustering.

5.2.1 Difficultés provoquées par l'usage de la distance L2

La plupart des études effectuées précédemment dans ce domaine (cf. section 4.3) utilisent la même distance pour comparer deux champs : la distance associée à la norme L2. Cette distance entre deux vecteurs $V = (v_1, v_2, \ldots, v_n)$ de données quotidiennes pour deux jours d_1 et d_2 est calculée par :

$$d_{L2}(V(d_1), V(d_2)) = \sqrt{\sum_{i=1}^{n}(v_i(d_1)) - v_i(d_2))^2} \tag{5.1}$$

où $v_i(d_j)$ est la i-ème valeur du vecteur $V(d_j)$ de données du jour d_j et n est le nombre de données journalières. Bien que la norme L2 soit couramment utilisée dans les méthodes de clustering, nous pensons qu'elle est en partie responsable des difficultés rencontrées dans l'étude automatisée de l'analyse du climat.

Pour illustrer ce propos et se rapprocher de notre cas d'étude, la figure 5.1 montre deux exemples schématiques en 2D, d'un champ de référence et deux autres situations. La distance L2 entre l'image de référence et chaque situation est la même, pourtant l'une est "physiquement" beaucoup plus proche de la référence que l'autre. Deux raisons sous-tendent un tel comportement.

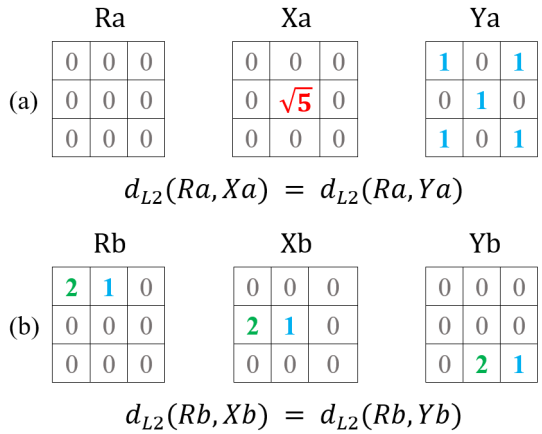


FIGURE 5.1 – Représentation des caractéristiques de la distance L2 sur des données 2D : (a) une forte fluctuation localisée Xa produit la même distance L2 qu'une multitude de petites variations Ya par rapport à la référence Ra, (b) un petit décalage spatial Xb, ou un grand Yb, produit la même distance L2 par rapport à la référence Rb.

Premièrement, lorsque les données sont décrites dans un grand espace vectoriel, une multitude de petites fluctuations réparties dans l'espace sur ce champ, peut être considérée comme aussi importante qu'une seule grande fluctuation très localisée (Fig 5.1a).

Deuxièmement, une situation (Xb) qui présente la même structure spatiale mais légèrement décalée par rapport à la référence (Rb) a la même norme L2

qu'une situation (Yb), où le décalage spatial est important (Fig 5.1b). Cette situation est probablement plus sensible lorsque le clustering est appliqué à des variables, telles que les précipitations, qui sont des champs spatio-temporels intermittents. Deux champs, légèrement translatés, n'ont pas beaucoup de pixels en commun, même s'ils sont similaires.

Dans le processus de mise en cluster, ces effets secondaires ont tendance à fausser la comparaison entre les modèles spatiaux quotidiens et affectent donc la qualité des clusters.

5.3 Proposition d'une nouvelle méthodologie

Les effets de la distance L2 dans les méthodes de clustering ont déjà été soulignés dans de nombreuses études comparatives dans d'autres domaines [83, 65, 31, 72]. Dans cette section, une mesure de dissimilarité originale basée sur une approche d'analyse d'images est proposée.

La conception de la mesure comprend trois phases : premièrement, subdiviser le domaine d'intérêt en zones selon les connaissances des spécialistes ; deuxièmement, construire des histogrammes pour quantifier la variable d'intérêt tout en réduisant l'influence de la localisation spatiale ; troisièmement, appliquer une mesure de dissimilarité aux histogrammes. Dans l'exemple choisi pour cette étude, la segmentation géographique en zones est effectuée en fonction des structures atmosphériques (Fig 5.5).

5.3.1 Gestion partielle de la spatialisation

Pour un certain nombre d'applications de vision par ordinateur, l'image peut être analysée par "patch" plutôt que par pixel individuel [2, 23, 1, 34]. Les patchs d'image contiennent des informations contextuelles et présentent des avantages en termes de calcul et de généralisation [33].

De ce point de vue, la décomposition d'une image en parcelles ou zones (se chevauchant ou pas), offre un moyen simple mais efficace de surmonter "the curse of dimensionality" (i.e. la malédiction de la dimensionnalité) [67, 80].

Il serait donc envisageable de définir des découpages simples, avec des patchs de même surface, comme présentés dans la figure 5.2. Il serait également possible d'ajuster le niveau de précision souhaité, en modifiant le nombre de patchs, par exemple de quatre à seize.

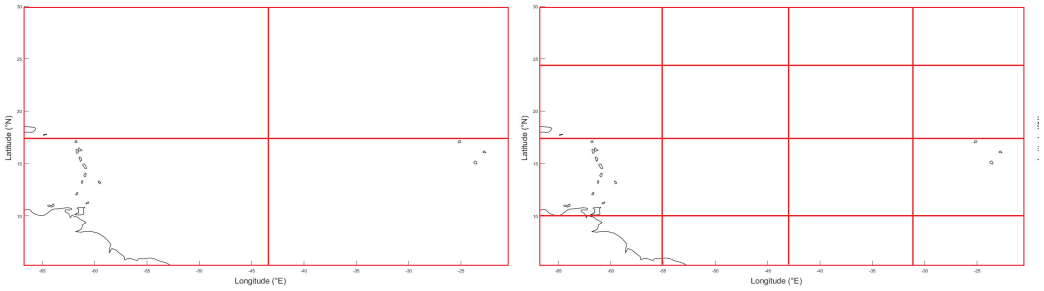


FIGURE 5.2 – Découpage simple (en rouge) du domaine d'étude, en quatre (a) ou en seize (b) zones (ou patch).

L'un des points essentiels, propre au concept de patch, est la possibilité d'utiliser différentes méthodes d'analyse selon le patch considéré. Il devient alors possible d'intégrer les spécificités d'une zone dans l'analyse d'une image. Cet aspect nous a paru tout à fait pertinent pour notre cas d'étude.

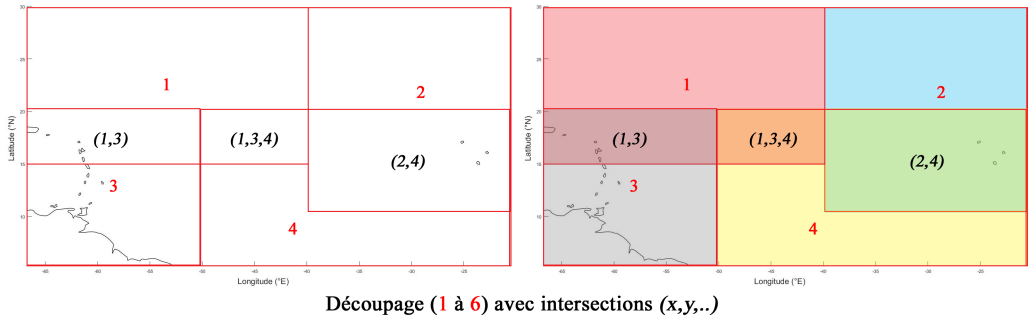


FIGURE 5.3 – Découpage en quatre patchs (en rouge) comportant des zones d'intersections (notées en noir).

Pour aller plus loin, nous pourrions considérer plusieurs autres approches. En effet, il est possible de définir des patchs qui se chevauchent, comme le montre la figure 5.3. Dans cette illustration, plusieurs cas de chevauchement sont visibles, par exemple, les îles de l'archipel du Cap–Vert se trouvent dans la zone d'intersection des patchs 2 et 4.

Il est important de noter que ces zones d'intersections sont considérées plusieurs

fois dans le processus d'analyse de ces patchs. Elles résolvent en revanche partiellement les problèmes d'effets de seuil pour des structures qui se trouveraient à la limite entre deux patchs. Ne voulant pas intégrer de la redondance dans notre analyse, nous n'avons pas tenté cette approche.

Une autre approche intéressante est présentée en figure 5.4, elle consisterait à utiliser un ensemble hiérarchisé de patchs, qui interviennent au fur et à mesure dans le processus d'analyse, à différents niveaux de résolution. Les dimensions des patchs ainsi que les méthodes d'analyse appliquées sur ces derniers peuvent varier selon le niveau [82].

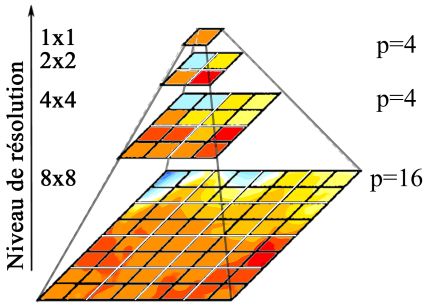


FIGURE 5.4 – Schéma représentant une analyse à résolutions multiples avec des découpages en patch (p) allant de 16 à 4, produisant ainsi plusieurs niveaux de résolution.

Évaluer le nombre de patch optimal, l'intérêt d'un chevauchement et du multi-échelle serait tout à fait pertinent. Mais dans un premier temps, nous souhaitions surtout démontrer l'intérêt de cette démarche. Afin d'en limiter le temps de calcul sans chercher à l'optimiser, nous avons préféré fixer ce nombre de patch à quatre.

Nous avons donc subdivisé la zone Atlantique tropicale et subtropicale, proche des Antilles, en quatre zones (Fig 5.5). Pour comparer deux jours sur un paramètre météorologique, nous avons découpé les champs en fonction de ces quatre zones qui sont utilisées pour les comparer.

Notons que ces zones ne sont pas de forme et de surface identiques. Celles-ci ne sont définies que pour prendre en considération les connaissances des experts. Ainsi, trois zones correspondent à des centres d'action spécifiques et connus.

Sur la figure 5.5, A1 est la zone d'arrivée des fronts froids ; A2 est la zone d'action de l'anticyclone subtropical de l'Atlantique Nord (NASH) ; A3, comprend la zone continentale et l'arc des Petites Antilles ; et A4 est la zone de basse pression liée à la zone de convergence intertropicale (ZCIT).

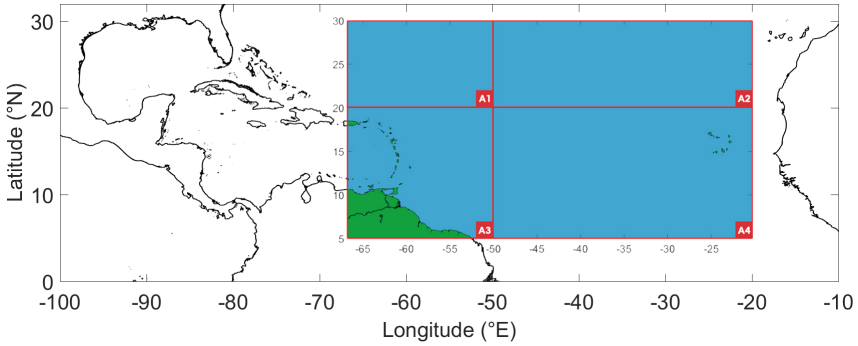


FIGURE 5.5 – Zone géographique d'intérêt : Les surfaces terrestres se trouvent dans la zone A3 (en bas à gauche) : Petites Antilles avec une partie nord-est de l'Amérique du Sud. Les zones A1, A2 et A4 sont principalement maritimes : une partie de l'océan Atlantique central et l'archipel du Cap-Vert.

Notons que cette approche a le mérite de réduire drastiquement la dimension des caractéristiques traitées, car celle-ci ne considère qu'un nombre réduit de patchs, permettant ainsi de surmonter "the Curse of dimensionality" [67].

Dès lors, nous allons pouvoir mettre en place une méthode d'analyse exploitant ces patchs, celle-ci sera détaillée dans la sous-section suivante.

5.3.2 Comparaison de l'intensité des distributions

Une fois que chaque champ a été subdivisé dans l'espace, il n'est plus nécessaire de localiser exactement les structures atmosphériques dans chaque zone. Il semble relativement raisonnable d'ignorer leur position jusqu'au maillage exact, et d'examiner plutôt la distribution de chaque champ des ensembles de données, en ignorant ainsi la notion de localisation spatiale.

Nous avons opté pour une représentation discrète des données par des histogrammes de fréquence, afin d'estimer la distribution des intensités des champs physiques. En outre, la quantification pourrait aider à réduire les effets des petites fluctuations signalées dans la section 5.2.1, même si les effets de bord

autour des limites des classes d'intensité sélectionnées subsisteraient.

Pour le calcul des histogrammes de l'ensemble des données journalières utilisées dans les exemples d'applications en sous-sections 5.4 et 5.5, nous avons procédé comme suit :

- pour le vent, l'échelle de Beaufort est utilisée pour définir les classes d'histogramme afin de représenter la distribution de la vitesse du vent.

- pour les précipitations, les classes sont déterminées à partir des données pluviométriques recueillies par les stations météorologiques au sol dans la zone. Á partir de ces seuils, qui reflètent plutôt la dynamique réelle des phénomènes, nous avons ajusté les choix des centiles afin de tendre vers une distribution des classes d'histogrammes plus ou moins équirépartie. Nous avons également pris le soin de prévoir des classes captant les valeurs extrêmes possiblement synonymes d'évènements importants (cyclones et tempêtes). Nous avons sélectionné huit classes d'intensités possibles (Tableau 5.1).

TABLEAU 5.1 – Bornes des classes d'histogramme utilisées pour quantifier les données sur les précipitations quotidiennes. Ces limites sont déterminées à partir des relevés pluviométriques de la zone d'étude.

Centiles (%)	**Rainfall** (mm)
0	0
0.35	]0,1.2]
0.5	]1.2,2.2]
0.7	]2.2,5.2]
0.8	]5.2,8.7]
0.9	]8.7,16.4]
0.95	]16.4,26.9]
0.99	]26.9,59.2]
1	]59.2,+∞[

5.3.2.1 Distances entre histogrammes

Á ce niveau du processus nous disposons pour une donnée journalière de quatre histogrammes représentant la dynamique de chaque zone. Pour comparer deux données journalières, nous proposons de quantifier la dissimilarité entre leurs histogrammes.

Nous disposons de plusieurs distances permettant de comparer ou mesurer la proximité de deux histogrammes. Les distances les plus fréquentes sont les normes L1 et L2 ou encore la distance de Wasserstein [64], pourtant nous ne les avons pas retenues.

Nous avions l'intuition qu'au delà d'une simple distance, pour ce type de données, il nous faudrait une mesure de l'information mutuelle entre ces histogrammes. C'est pour cela que nous nous sommes intéressés à la mesure de l'entropie relative. En effet, dans nos travaux nous avons expérimenté la divergence de Kullback–Leibler [47].

5.3.2.1.1 Divergence symétrisée de Kullback-Leibler

Pour comparer des histogrammes sans définir une distribution spécifique, la divergence symétrisée de Kullback-Leibler [47, 48] semble être un choix judicieux, voici son expression mathématique :

$$\begin{aligned} D_{KL}(P,Q) &= \sum_{c=1}^{m} D_{KL}(P(c),Q(c)) \\ &= \sum_{c=1}^{m} P(c) log \frac{P(c)}{Q(c)} \\ D_{KLS}(P,Q) &= D_{KL}(P,Q) + D_{KL}(Q,P) \end{aligned} \tag{5.2}$$

où P et Q sont deux distributions de probabilités discrètes, c est l'indice d'une classe pour chaque distribution et m le nombre de classes. Bien qu'il soit possible d'utiliser la divergence symétrisée de Kullback-Leibler avec des champs continus, nous avons pensé qu'il était plus intéressant de quantifier les données.

Les distributions d'intensité distinctes obtenues sont ensuite utilisées pour calculer la divergence de Kullback-Leibler dans chaque zone. La moyenne des divergences par zone fournit la dissimilitude entre deux jours. Nous avons

nommé "Expert Deviation" (ED), la quantité définie par la moyenne des D_{KLS} des p zones :

$$ED(d_1, d_2) = \frac{1}{p} \times \sum_{i=1}^{p} D_{KLS}(Z_i(d_1), Z_i(d_2)) \tag{5.3}$$

où d_1 et d_2 sont deux jours, $Z_i(d_j)$ est l'histogramme de la zone avec la référence i et p le nombre de zones.

Il est important de noter que les valeurs des classes de l'histogramme peuvent avoir un effectif nul, auquel cas D_{KL} prend une valeur infinie. Pour pallier à cette difficulté, nous avons ajusté nos valeurs de la façon suivante. Lors du calcul de $\hat{P}(c)$, la fréquence d'occurrence de la classe c parmi les m possibles, comprenant n_c évènements parmi N, nous utiliserons l'estimateur suivant :

$$\hat{P}(c) = \frac{n_c + \epsilon}{N + m\epsilon} \tag{5.4}$$

ϵ a été choisi de façon à ce que son influence soit plus faible que celle du moindre pixel potentiellement différent entre les deux images. Nous avons retenu $\epsilon = 10^{-6}$.

La figure 5.6 résume brièvement les opérations énumérées ci-dessus permettant d'aboutir à la mesure ED caractérisant la dissimilarité existant entre deux données journalières. Les EDs seront conçues de cette manière pour chaque ensemble de données que nous utiliserons.

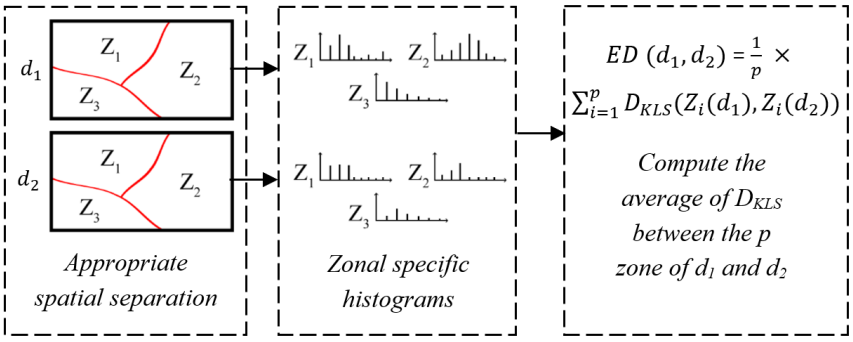


FIGURE 5.6 – Schéma montrant le processus de calcul de ED : quantification zonale à l'aide de classes d'histogrammes prédéfinies, utilisation de la divergence symétrisée de Kullback-Leibler (D_{KLS}) sur chaque zone pour obtenir quatre valeurs et calcul de la moyenne pour obtenir $ED(d_1, d_2)$.

5.3.3 Évaluation du clustering

Afin d'évaluer et de comparer les résultats produits par la méthode ED avec ceux produits par L2, nous utilisons de nouveau l'indice de silhouette dont les conclusions seront appuyées par une analyse physique des clusters trouvés.

Néanmoins, avant de présenter ces courbes et analyses, il convient de comprendre plus précisément le processus de clustering, car le centroïde utilisé jusqu'ici va prendre une forme tout à fait nouvelle, qui nous amènera à changer la façon dont on visualise le représentant d'un cluster.

5.3.4 Intégration dans le processus d'analyse

Par défaut, dans KMS, à la fin de chaque itération, les centroïdes de chaque grappe sont recalculés. Ces centroïdes sont le champ moyen calculé sur chaque cluster.

Nous pensons que ces centroïdes sont artificiels et non représentatifs d'une configuration spatiale réaliste puisque la moyenne lisse chaque petite structure spatiale, ce qui donne une structure énorme qui n'apparaît dans aucun champ observé. Ainsi, le processus de clustering conduit à regrouper les jours autour d'un champ non pertinent. Cette méthode n'est pas efficace pour les champs intermittents tels que les précipitations cumulées.

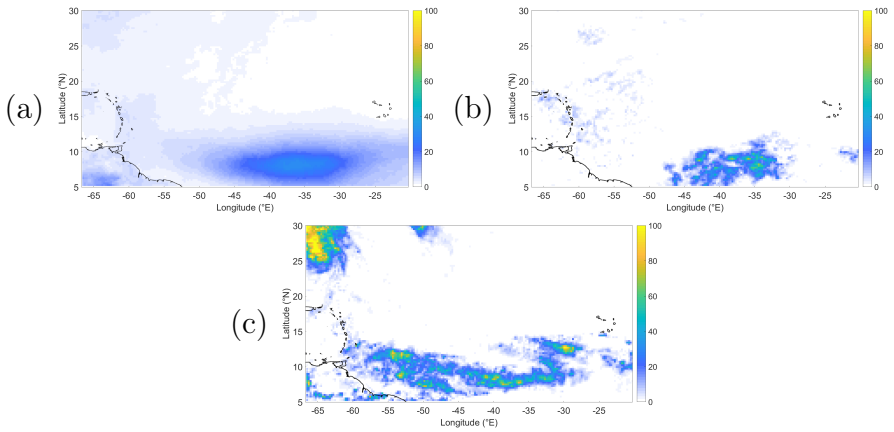


FIGURE 5.7 – Représentativité des centroïdes : centroïde de précipitation (a), comparé à l'élément le plus proche selon L2 (b) et un autre élément de la grappe pris au hasard (c).

La figure 5.7 présente ainsi un centroïde et deux journées associées au même

cluster. Clairement, le centroïde ne ressemble aucunement à ces deux journées.

Un autre point concerne la méthodologie vue dans les articles de la section 4.3 : la plupart des auteurs se fondent sur le fait que le centroïde leur semble pertinent physiquement pour justifier la qualité de leur clustering.

Nous pensons que cette idée est invalidée par les considérations suivantes : Compte tenu des fortes corrélations spatio-temporelles existant au sein de nos données, quasiment n'importe quelle moyenne peut être interprétée comme physiquement plausible même si elle a été calculée à partir de données regroupées aléatoirement.

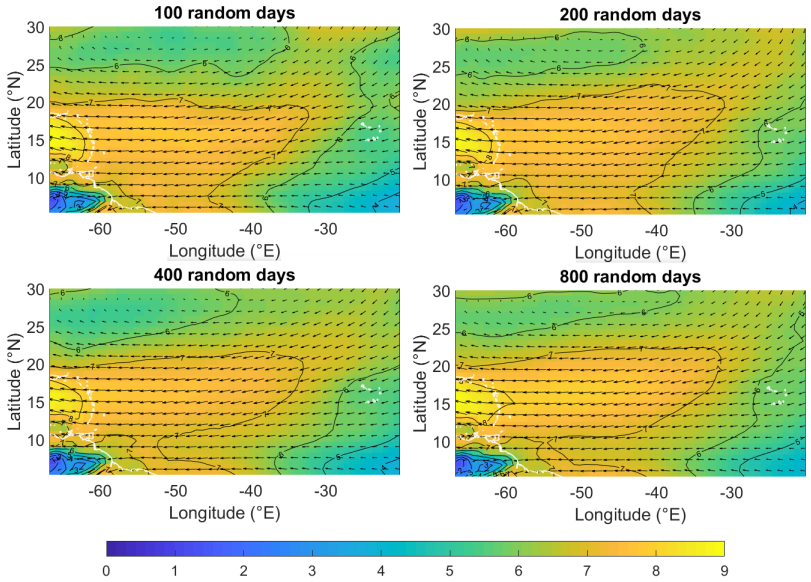


FIGURE 5.8 – Exemple de centroïdes d'intensités du vent de surface provenant de ERA-Interim, et qui sont interprétables en régimes de circulations atmosphériques, mais qui pourtant résultent d'un clustering hasardeux

C'est ce que nous avons testé dans la figure 5.8. Les quatre pseudo centroïdes présentés sont la moyenne d'un nombre variable de journées de vent. Les quatre exemples présentent bien une cohérence spatiale qui semble interprétable (une zone centrale de vents forts, étalée selon un faible gradient, et une zone de vents faibles au Sud Ouest.

Dans le cas de ED, les données journalières sont transformées en groupes de quatres histogrammes qui correspondent aux zones géographiques précédem-

ment introduites dans la section 5.3.1. Le centroïde devient donc, également un groupe de quatre histogrammes, correspondant aux histogrammes moyens de chaque zone pour le cluster. On regroupe ainsi les jours autour d'un groupe d'histogrammes moyens, ce qui selon nous participe à l'amélioration des résultats.

Se pose néanmoins alors le problème de la visualisation de l'élément représentatif du cluster, dont les physiciens ont un réel besoin pour évaluer au moins partiellement la qualité des cluster.

Nous avons déjà écarté l'idée d'utiliser le champ moyen (le centroïde) comme représentant. Afin d'obtenir une vue significative de chaque centre de cluster, nous proposons de sélectionner l'élément le plus proche du centroïde. "Le plus proche" est, évidemment, entendu au sens de la dissimilarité ED. Ce centre est maintenant un champ existant.

Bien entendu, le calcul de l'indice de silhouette intègre également ED pour effectuer une évaluation efficace des résultats du clustering selon ED. Les notations CAH-ED, CAH-L2, KMS-ED et KMS-L2 sont retenues pour désigner les algorithmes et la distance qu'ils utilisent.

Dans la section suivante, nous allons présenter les résultats de nos travaux portant sur les données de vents et de précipitations utilisées auparavant dans le section 4.4.

5.4 Application sur les données de vents

Dans cette section, nous allons présenter les résultats obtenus par clustering des données de vent. Les résultats de ces méthodes seront notés KMS-ED$_{WIND}$ et CAH-ED$_{WIND}$. Les résultats préalablement obtenus dans en section 4.4, seront quand à eux notés KMS-L2$_{WIND}$ et CAH-L2$_{WIND}$.

Comme dans la section 4.4, nous allons analyser les résultats de l'évaluation de la qualité des clusters pour nos deux algorithmes. Dans cette section, chaque algorithme existe en deux versions intégrant l'une ou l'autre des deux distance utilisées (L2 et ED).

Puis nous effectuerons une évaluation de ces clusters du point de vue du physicien du climat. Cette étape nous permettra de confronter l'évaluation numérique et l'avis expert. Elle commencera par une inspection visuelle des clusters. Nous regarderons ensuite les différentes répartitions temporelles des clusters, afin d'évaluer indirectement la dynamique temporelle de ces derniers. Enfin, d'autres indicateurs d'analyse seront présentés afin de juger de la pertinence et la robustesse qu'offre cette nouvelle méthode.

5.4.1 Évaluation numérique

L'évaluation de la qualité des clusters est présentée en figure 5.9. Celle-ci montre l'évolution de l'indice de $Sa(\mathcal{M}_k)$ en fonction du nombre de clusters k. Elle est le pendant de la figure 4.10 du chapitre précédent, à laquelle nous avons ajouté les résultats obtenus pour KMS-ED et CAH-ED.

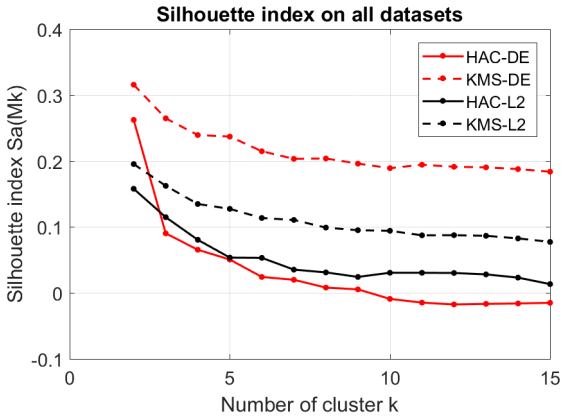


FIGURE 5.9 – Évolution de l'indice de Silhouette en fonction de k, le nombre de cluster - HAC ou CAH (ligne continue), KMS (ligne discontinue), en utilisant L2 (noir), en utilisant ED (rouge) - Résultats pour le clustering de l'intensité du vent à 850hPa.

Ce résultat met en évidence deux points intéressants. Premièrement, CAH produit des coefficients nettement inférieurs à ceux obtenus par KMS, et cela indépendamment de la distance choisie pour le clustering du vent. CAH produit de plus des résultats systématiquement inférieurs à 0.1 (pour $k > 3$), indiquant des clusters non pertinents. KMS sera donc utilisé pour l'inspection visuelle. Cela correspond tout à fait au constat préalablement établi en chapitre 4.

Deuxièmement, KMS-ED$_{WIND}$ présente des résultats bien meilleurs que toutes les études que nous avons pu aborder ou effectuer jusque-là. C'est la seule courbe dont la quasi intégralité des points sont supérieurs à la valeur de référence 0.2. L'indice de silhouette indique enfin des clusters pertinents.

De plus, nous pouvons fixer le nombre de clusters à $k = 5$. Ici encore, l'optimalité de ce nombre n'est pas aussi franche que nous aurions pu le souhaiter, mais l'on observe toujours une légère inflexion autour de cette valeur. Maintenant passons à l'évaluation visuelle, en précisant que seuls les 5 éléments représentatifs KMS-ED, seront présentés.

5.4.2 Évaluation visuelle

La figure 5.10 présente les éléments les plus représentatifs de chaque cluster. Comme précisé auparavant, il s'agit du champ le plus proche (au sens de ED) du centroïde du cluster.

La séparation des clusters semble bonne. En effet, les configurations sont toutes différentes et les structures atmosphériques habituelles sont clairement identifiables. Il est maintenant possible d'interpréter ces situations pour les experts.

Ainsi, selon les spécialistes du domaine, dans le KMD-ED$_{WIND}$ C1, les alizés sont du sud-ouest et de faibles intensités, répartis sur une étroite bande maritime qui s'étend au sud des Petites Antilles. On observe également un phénomène de divergence dans la circulation atmosphérique, qui entraîne des vents alizés plus forts, notamment dans la partie nord des Petites Antilles.

C2 montre plusieurs bandes de divergence qui produisent des vents forts très localisés sous l'anticyclone subtropical de l'Atlantique Nord (NASH), situé vers l'ouest entre 25 et 30°N. La configuration favorise un affaiblissement des alizés avec une composante sud-est à l'approche de l'arc des Petites Antilles. Elle pourrait indiquer les passages d'ondes d'est au-dessus des Petites Antilles.

En C3, les alizés sont forts et ont une composante d'est en nord-est. C4 présente le NASH donnant des vents de plus en plus forts du nord des Petites Antilles

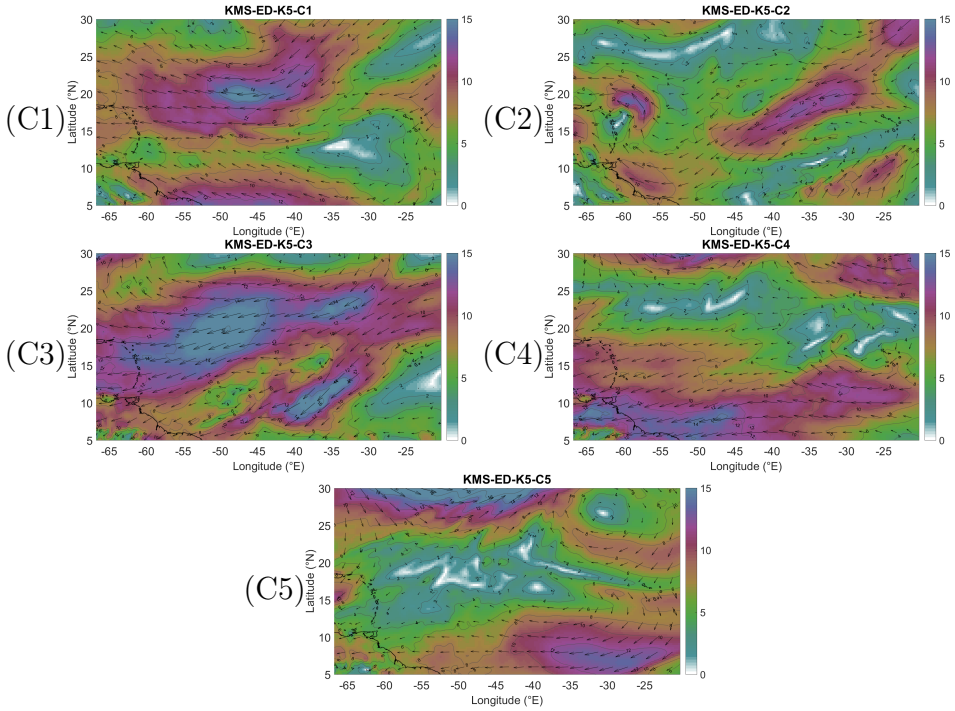


FIGURE 5.10 – Éléments représentatives des clusters d'intensité de vents 850 hPa obtenus par la méthode KMS-ED$_{WIND}$ pour $k = 5$.

au continent sud-américain. Enfin, C5 montre une situation de panne d'alizés dans les Petites Antilles. Observons maintenant la répartition mensuelle des clusters. Celle-ci est donnée par la figure 5.11.

Cette répartition temporelle correspond également assez bien aux attentes des experts. Par exemple, C1, contenant les jours de faibles intensités de vent, est principalement centré sur juillet-aout, bien qu'on le retrouve également tout au long de l'année. Tout expert le confirmera.

L'adéquation entre la description des clusters et leur répartition temporelle nous a été confirmée pour chaque cluster mais ne sera pas détaillée plus longuement ici. Ainsi, si les résultats obtenus par KMS-L2 n'étaient pas trop mauvais, l'introduction de ED dans le clustering a permis de les améliorer de façon notable.

Voyons maintenant ce que nous obtenons pour les précipitations, dont nous avions noté que les discontinuités spatio-temporelles posaient de gros problèmes à KMS-L2.

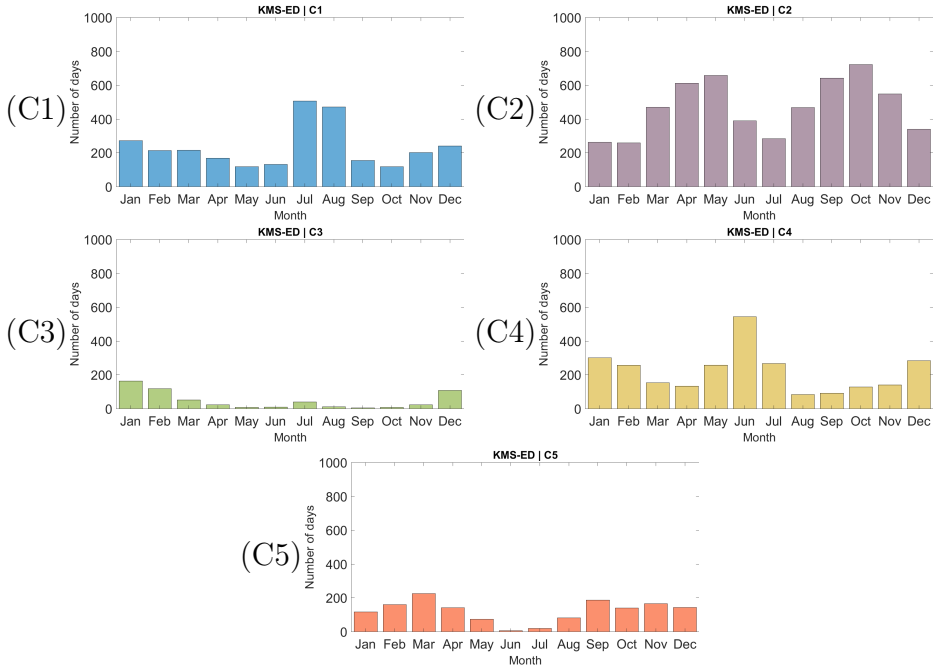


FIGURE 5.11 – Répartitions mensuelles des clusters obtenus par la méthode KMS-ED$_{WIND}$ pour la période 1979 à 2014.

5.5 Application sur les données de précipitations

Ici, les résultats des méthodes de clustering seront notés KMS-ED$_{RAINFALL}$ et CAH-ED$_{RAINFALL}$. Les résultats préalablement obtenus en section 4.4 seront quand à eux notés KMS-L2$_{RAINFALL}$ et CAH-L2$_{RAINFALL}$. Nous allons également suivre le même plan de présentation qu'auparavant.

5.5.1 Évaluation numérique

L'analyse de la qualité des clusters présentée ici en figure 5.12 devrait nous permettre de renforcer nos précédents résultats. Ce graphe, présenté en page suivante, confirme un certain nombre d'observations que nous avons faites pour les données d'intensité du vent, traitées plus tôt. Les clusters produit par CAH ne semblent toujours pas exploitables. CAH-L2 présente des valeurs négatives, constituant des clusters quasiment aléatoirement affectés. Si la situation s'améliore pour CAH-ED$_{RAINFALL}$, l'indice de Silhouette avoisine zéro, cela signifie qu'il ne trouve aucune structure vraiment pertinente. Ici encore, on pourra donc écarter CAH pour la suite de l'analyse.

Néanmoins, en ce qui concerne KMS, la différence entre les deux distances

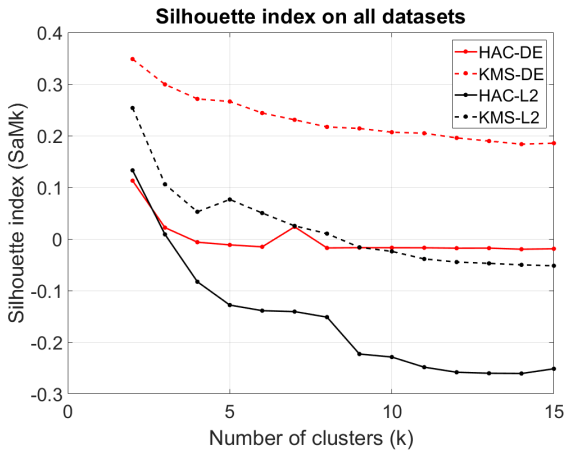


FIGURE 5.12 – Évolution de l'indice de Silhouette en fonction de k, le nombre de cluster - HAC ou CAH (ligne continue), KMS (ligne discontinue), en utilisant L2 (noir), en utilisant ED (rouge) - Résultats pour le clustering des cumuls journaliers précipitations.

est beaucoup plus marquée que dans le cas du vent. KMS-L2 présente des performances indiquant des résultats non pertinents. Á l'inverse, pour KMS-ED$_{RAINFALL}$ la quasi totalité de la courbe se situe au dessus de la valeur de référence 0.2, garantissant une structuration interne bien marquée des données.

On peut également noter que ces résultats sur les données de précipitations sont tout à fait à la hauteur de ceux présentés pour les données du vent, bien que les premières soient censément plus difficiles à traiter. L'inflexion sur laquelle nous nous basons pour choisir le nombre de clusters indique $k = 5$ comme étant optimal, avec un indice de 0.26. Les éléments représentatifs des cinq clusters de KMS-ED$_{RAINFALL}$ seront donc inspectés, dans la sous-section suivante.

5.5.2 Évaluation visuelle

L'intérêt de notre approche est renforcé par les figures 5.13 à 5.15 qui présentent les éléments les plus représentatifs de chaque cluster des données de précipitation.

Selon les experts, les résultats sont tout aussi significatifs puisque les éléments représentatifs de KMS-ED$_{RAINFALL}$ décrivent différentes configurations des précipitations cumulées quotidiennes (Fig 5.13), tandis que ceux de KMS-

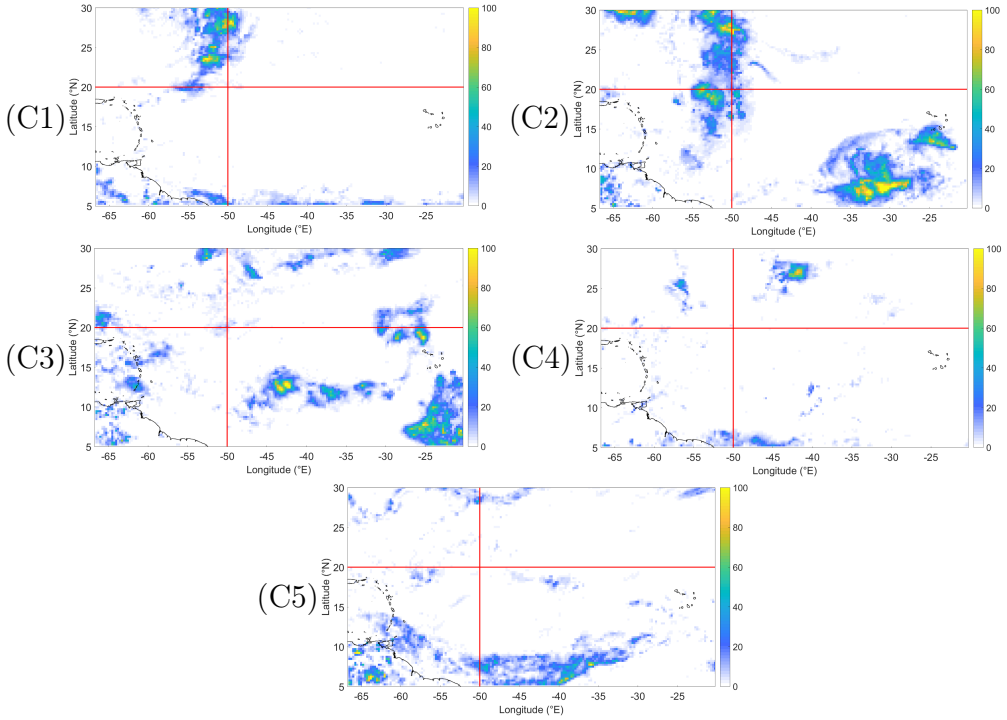


FIGURE 5.13 – Éléments représentatives des clusters des cumuls journaliers de précipitations obtenus par la méthode KMS-ED$_{RAINFALL}$ pour $k = 5$.

L2$_{RAINFALL}$ montraient au chapitre 4 des situations sèches avec une ZIC active pour certains clusters et l'arrivée de fronts froid en provenance du nord-ouest pour un autre.

Nous utilisons ensuite la figure 5.14 pour avoir un aperçu l'homogénéité interne des clusters, produit par KMS-ED. Pour ce faire, dans cette figure, nous présentons quelques éléments du cluster C4. Il s'agit de l'un des clusters ayant les meilleures valeurs de $Sc(C_i)$ (cf. tableau 5.2). La même expérience a été réalisée avec KMS-L2$_{WIND}$ (cf. Fig 4.15).

On observe que les éléments du cluster C4 présentent une certaine constance physique bien qu'ils puissent différer dans leur aspect visuel. Ce cluster est ainsi constitué de manière relativement homogène de champs à faible pluviométrie générale. Si cette homogénéité est toute relative, elle reste néanmoins bien meilleure que celle observée pour KMS-L2.

Les résultats obtenus par KMS-ED sur les données de précipitations consti-

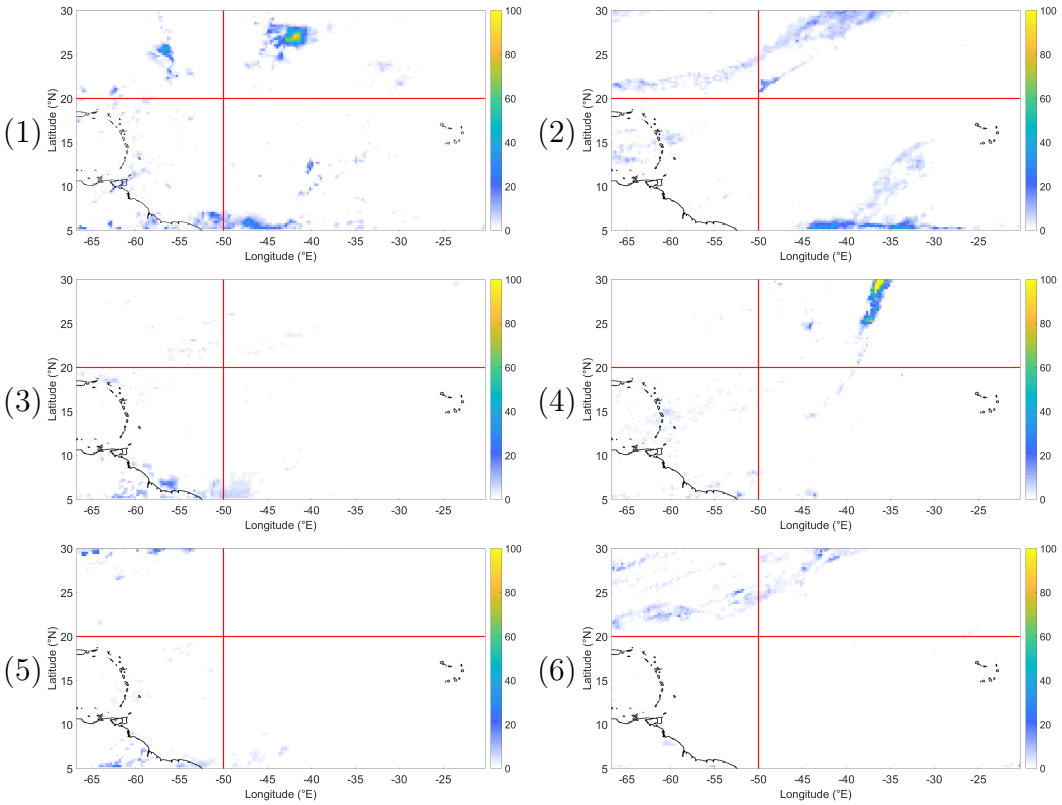


FIGURE 5.14 – Variabilité interne du cluster (C4) d'après la méthode *KMS-ED*. Six jours (1-6) de ce cluster sont présentés dans l'ordre croissant de ED par rapport au centroïde, du plus proche au plus éloigné avec un pas constant.

tuent les plus fiables et les plus intéressants pour la communauté de climatologie, obtenus au cours de ces travaux de thèse. Nous ferons donc l'hypothèse qu'ils peuvent avoir des applications importantes.

Pour cette raison, la suite de cette section sera consacrée à l'analyse physique de ces clusters, sous de nombreux aspects. Nous présenterons donc les graphiques correspondant, leur analyse combinée étant déportée en fin de section. Ici encore, la répartition temporelle des clusters nous fournira une indication indirecte de la pertinence des clusters.

Il est également intéressant d'observer l'évolution, pour la période 2000-2014, de la fréquence d'apparition de chaque cluster, permettant de voir si ces clusters se développent ou ont tendance à disparaitre. Des mesure de significativité statistique des tendances relevées sont fournies (p-value).

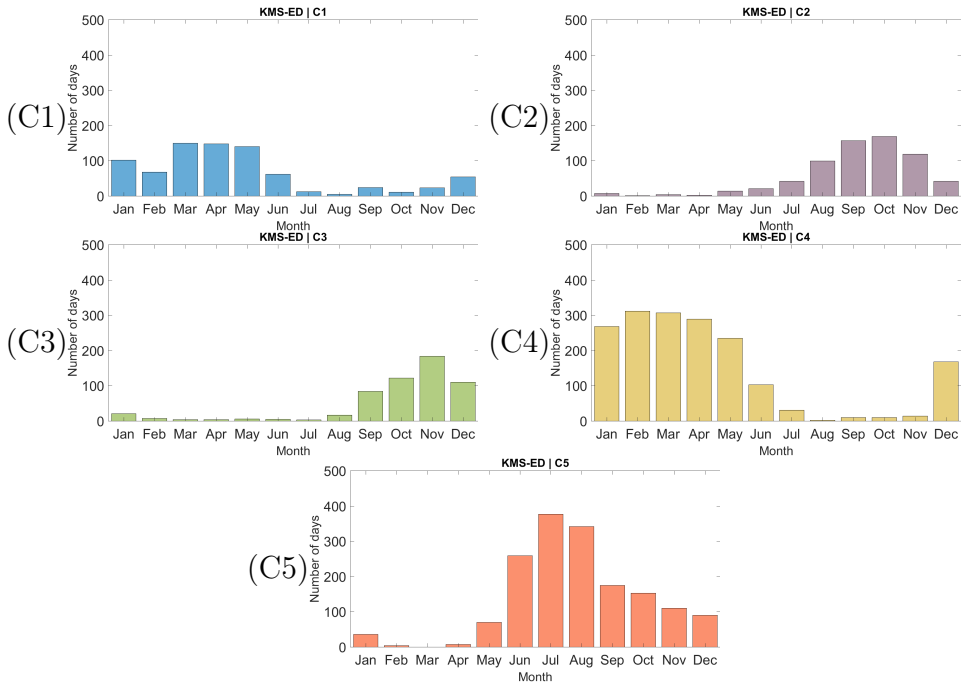


FIGURE 5.15 – Répartitions mensuelles des clusters obtenus par la méthode KMS-ED$_{RAINFALL}$ pour la période 2000 à 2014.

En considérant l'indice de Silhouette de chaque cluster, noté $Sc(C_i)$ et présenté dans le tableau 5.2, nous pouvons établir une analyse globale des clusters de KMS-DE$_{RAINFALL}$.

TABLEAU 5.2 – Comparaison de l'indice de Silhouette $Sc(C_i)$ pour les cinq clusters de KMS-ED$_{RAINFALL}$ et de KMS-L2$_{RAINFALL}$, avec une accentuation des meilleurs valeurs par algorithme (en gras) et de la plus pertinente (en vert).

$Sc(C_i)$	**C1**	**C2**	**C3**	**C4**	**C5**
KMS-ED$_{RAINFALL}$	0.23	0.28	0.25	0.26	**0.33**
KMS-L2$_{RAINFALL}$	0.18	0.02	-0.04	**0.22**	6e-3

Globalement, tous les indices des clusters de KMS-ED$_{RAINFALL}$ sont positifs et supérieurs à ceux de KMS-L2$_{RAINFALL}$. Le cluster C5 de KMS-ED$_{RAINFALL}$ est le plus cohérent, puisque sa valeur est supérieure à 0.3. Il est tout à fait intéressant de noter que la valeur d'indice de KMS-ED-C4$_{RAINFALL}$ n'est pas éloigné de celle de KMS-L2-C2$_{RAINFALL}$. Il s'agit de l'indice médian à la méthode KMS-ED$_{RAINFALL}$, c'est pour cela que nous les avons choisis pour l'évaluation visuelle de l'homogénéité interne. L'indice de silhouette nous permet

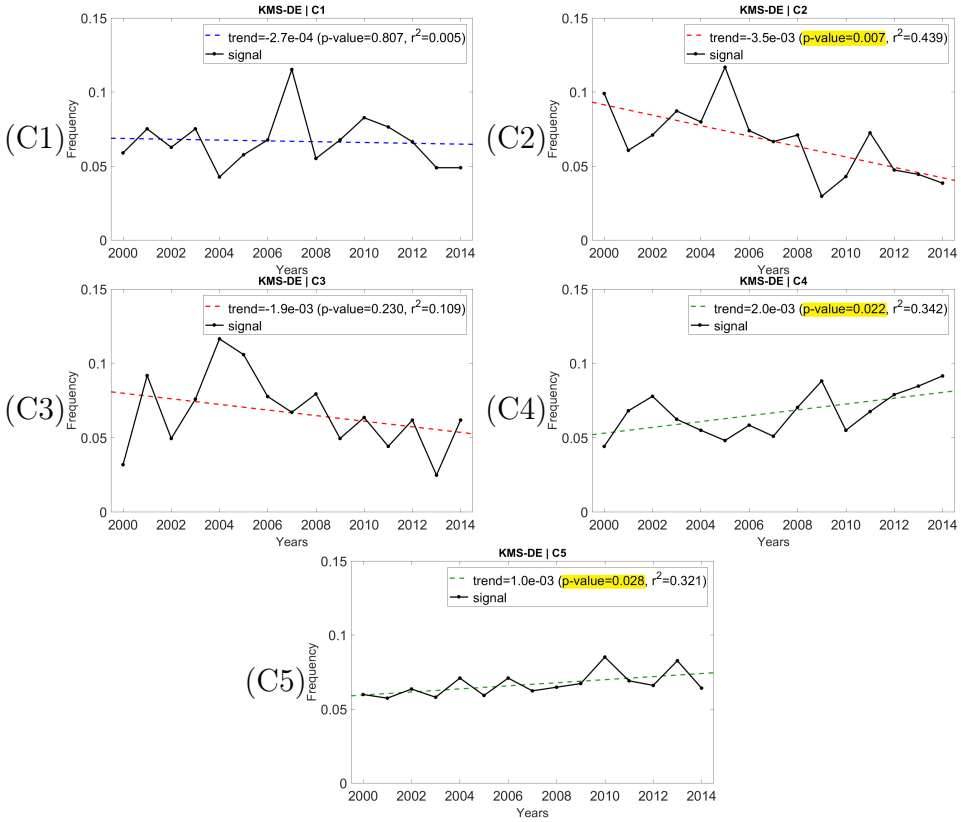


FIGURE 5.16 – Évolution inter-annuelle des fréquences d'apparition des clusters obtenus pour la méthode KMS-ED$_{RAINFALL}$, pour la période 2000 à 2014.

ici de confirmer l'analyse visuelle de l'expert. En effet, bien que leurs indices soient proches, l'on peut conclure, au vu de ces résultats, que la séparation et l'homogénéité produites par KMS intégrant la ED est largement meilleure que celle que nous avons obtenus avec L2.

Les paragraphes qui suivent regroupent l'ensemble des observations que nous avons pu produire concernant les clusters KMS-ED$_{RAINFALL}$. Ils correspondent à nos meilleurs résultats. En analysant les figures 5.15 et 5.16 l'on peut en tirer les propos suivants :

C1, représente 14.8% des jours analysés, avec des jours répartis sur l'ensemble des 12 mois, mais les données restent plutôt groupées sur la première partie de l'année. Le maximum est centré sur les mois d'avril et mai. Ce cluster a été associé aux CT 1 à 5 trouvés dans la [12] et aux WT 1 à 3, 7, et 8 trouvés dans la [61]. Aucune tendance significative n'a été constatée pour les fréquences an-

nuelles.

C2 et C3, respectivement 12.5 et 10.4%, sont relativement proches l'un de l'autre, en effet ils sont localisés à la fin de l'année. Cependant, pour C2, le maximum correspond à septembre-octobre, alors que pour C3, le maximum est en novembre. Des diminutions des fréquences annuelles sont constatées pour C2 (valeur p=0.02 et r^2=0.439), alors qu'aucune tendance significative est observée pour C3. Ces clusters peuvent être associés aux WT 4,5,6, et 8 dans [61].

Le C4 est le plus répandu, il représente 32% des jours de la base. Les éléments de celui-ci sont principalement répartis de décembre à avril avec un pic en février. Il représente l'ensemble de la saison sèche. La distribution annuelle montre une "baisse" marquée entre juillet et novembre (Fig 5.15). Pour l'analyser plus en détail, nous avons réalisé une étude de la colonne d'air des jours de ce cluster. La figure 5.17 montre les observations de la haute atmosphère à la station 78897-TFFR Le Raizet.

Les champs de précipitations quotidiennes de C4 montrent une fine couche humide, qui s'assèche à partir de 800hPa. Dans les niveaux inférieurs, les alizés soufflent du nord-est, tandis que pour les niveaux inférieurs à 500hPa, le flux provient principalement du nord-ouest. Dans ce cas, les moteurs, qui sont la NASH et la SST, limitent l'intensité des précipitations observées [18, 24, 60].

Le NASH et l'anticyclone nord-américain sont reliés entre eux et provoquent de forts vents alizés divergents et des affaissements dans les Caraïbes. L'ITCZ migre vers le sud, favorisant le développement du flanc sud de la NASH [53]. Une analyse détaillée de la C4 indique la présence quelques fois de fortes précipitations. Ces conditions sont liées aux passages de fronts froids qui réussissent à atteindre les basses latitudes des Petites Antilles (de janvier à mai).

En effet, à l'interface entre les deux masses d'air, les alizés transportent un flux d'air équatorial chaud et humide, qui intensifie la convection et le développement de bandes nuageuses frontales actives produisant de fortes précipitations [4].

C'est l'un des deux clusters pour lesquels, l'évolution du nombre de jours par

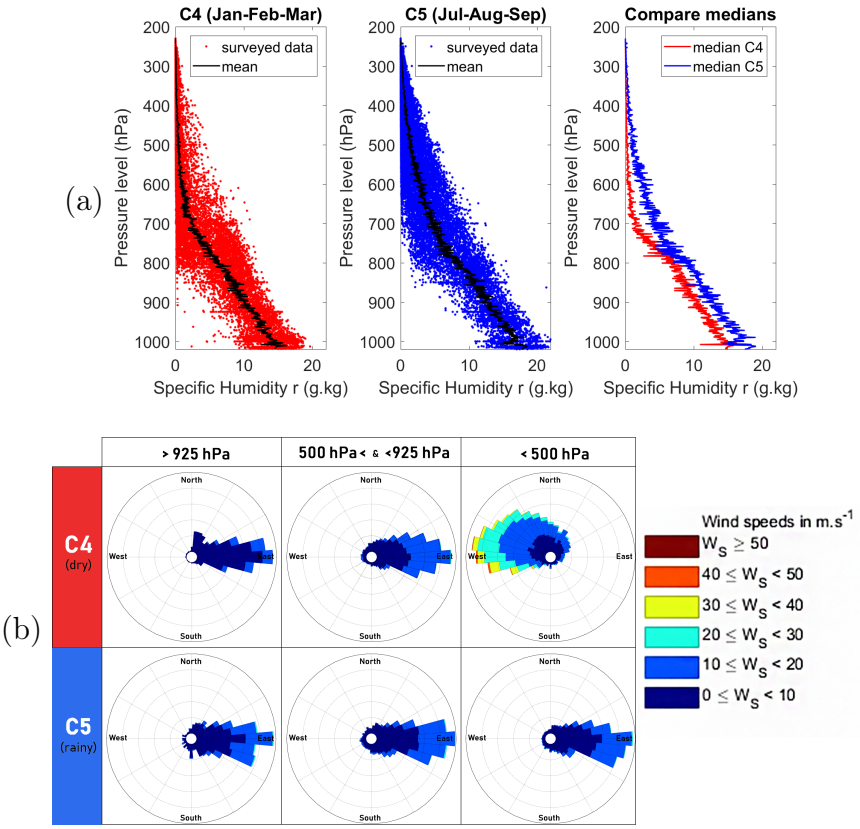


FIGURE 5.17 – (a) Diagramme de l'évolution de l'humidité dans les couches d'air en fonction du niveau de pression : les données recueillies par radiosonde appartenant au cluster C4 dans la période de janvier-février-mars (en rouge) et celles du cluster C5 dans la période de juillet-août-septembre (en bleu), avec leurs moyennes respectives (en noir). (b) Diagramme de l'évolution de la direction et de la vitesse du vent dans les couches d'air en fonction du niveau de pression : les données recueillies par radiosonde appartenant au cluster C4 (en rouge) et celles du cluster C5 (en bleu).

cluster au cours des 15 années d'observations TRMM montre une corrélation positive. Le nombre d'éléments de ce cluster est en augmentation, de 2000 à 2014, avec une $p-value$<0,022 et un r^2=0,342, lorsqu'on applique une régression linéaire. Ce groupe peut être associé aux CT 1 à 5 trouvés dans la [12] et aux WT 1 à 3, 7 et 8 trouvés dans la [61].

Pour C5, le dernier cluster, il peut être reliée aux CT 6–7 et aux WT 4–6 trouvés respectivement dans [12] et [61]. Les alizés sont conflictuels car ils présentent de nombreuses lignes de cisaillement, générées pendant cette période par les changements de direction et de vitesse du vent, passant du secteur nord-est au secteur sud-est, entrainant une forte couverture nuageuse accompagnée

de fortes précipitations. C5 regroupent les champs spatio-temporels avec une forte épaisseur de la couche humide des alizés d'Est, consécutifs à un NASH positionné au delà de 25°N et dans la partie orientale de l'Atlantique Nord.

En outre, les passages rapprochés des ondes d'est déclenchent de puissants courants ascendants avec de grandes zones de convergence en surface favorisant des pluies plus abondantes sur les Antilles orientales [4]. Pour aller plus loin, ils nous a paru intéressant de regarder quelques applications d'impacts en ne considérant que les phénomènes cycloniques.

5.5.2.1 Liaison avec les phénomènes cycloniques

Les tempêtes tropicales (TS) et les ouragans (H), recensés sur le site de la NOAA de 2000 à 2014, nous ont permis de rattacher les différentes proportions de ces types de phénomènes atmosphériques aux clusters retenus dans le KMS-ED$_{RAINFALL}$. Ce travail n'avait pas été fait par [77, 61, 74, 39, 41].

Les proportions sont présentées dans le tableau 5.3. Les clusters C2, C3 et C5 présentent presque tous ces risques. Le plus grand nombre de jours avec TS ou avec H a été trouvé pour C5, suivi par les clusters C2 et C3. Ils comprennent environ 80% des H et plus de 90% des TS.

Par rapport à la taille des clusters, $P_{Cx}(H)$ dans le tableau 5.3 montre une proportion deux fois plus importante (13%) de H dans C2 que dans C5 (5,8%) et C3 (6,9%). C1 et C4 présentent peu de risques.

Cependant, dans la sous-section précédente, ces clusters sont identifiés comme des groupes de jours secs, mais nous pouvons observer la présence de quelques jours TS ou H. Cela semble être dû au fait que la base de données de la NOAA couvre l'océan Atlantique central et toute la région des Caraïbes, donc ces phénomènes peuvent se trouver à la limite de notre domaine.

Nous avons ensuite jugé intéressant de considérer nos clusters à une l'échelle plus locale. Cela nous permet de comprendre comment les configurations spatiales de grandes influences la météorologie des îles de Petites Antilles, où nous vivons.

TABLEAU 5.3 – Statistiques descriptives et probabilités : analyse de la répartition des ouragans et des tempêtes tropicales dans les cinq clusters de la méthode KMS-ED$_{RAINFALL}$. $P_{TS}(Cx)$ exprime la probabilité qu'un TS soit en Cx, $P_H(Cx)$ exprime la probabilité qu'un H soit en Cx, $P_{Cx}(TS)$ exprime la probabilité que Cx produise un TS et $P_{Cx}(H)$ exprime la probabilité que Cx produise un H.

Clusters	**TS**	**H**	**Cluster sizes**	$P_{TS}(Cx)$	$P_H(Cx)$	$P_{Cx}(TS)$	$P_{Cx}(H)$
C1	1	12	799	0.013	0.049	0.001	0,015
C2	20	88	677	0.253	0.362	0.029	0.130
C3	11	39	567	0.139	0.160	0.019	0.069
C4	3	9	1749	0.038	0.037	0.002	0.005
C5	44	95	1623	0.557	0.391	0.027	0.058
Total	79	243	5415				

5.5.2.2 Influences sur les Petites Antilles

Pour cette étude nous avons établi le tableau 5.4, disposé plus bas, qui présente une analyse par île de l'impact des clusters de KMS-ED$_{WIND}$. Dans ce tableau, nous avons calculé les précipitations cumulées moyennes sur toutes les mailles occupant au moins 20% de la surface terrestre de chaque île (MSS), d'une part, et celles obtenues par pixel (MSM), d'autre part. Nous avons également enregistré le pourcentage de zéros inclus dans chaque cas (DWR).

Les valeurs MSS de l'île de la Dominique, de la Guadeloupe et de la Martinique sont les plus élevées. Les chaînes de montagnes de ces îles sont de véritables barrières imposantes aux alizés, créant une ascension orographique de la masse d'air humide et des précipitations. Le relief est le facteur géographique régional fondamental de la climatologie des Caraïbes. En revanche, les îles "basses", comme la Barbade, ont des totaux plus faibles.

En moyenne, les groupes C5 et C2 produisent des jours très pluvieux dans toutes les Petites Antilles. Le groupe C3 correspond aux précipitations "normales" (ou tropicales). Les clusters C1 et C4 produisent des jours de faibles précipitations. Les pourcentages de jours de sécheresse confirment cette répartition.

L'histogramme de la Fig 5.18 nous permet de comparer TRMM aux observa-

TABLEAU 5.4 – Valeurs détaillées des précipitations mesurées par satellite pour les îles des Petites Antilles *(avec **MSS**=Moyenne de la somme spatiale [mm/jour], **MSM**=Moyenne de la moyenne spatiale [mm/jour] et **DWR**=Pourcentage de jours sans précipitations [%])*, pour les clusters KMS-ED (de C1 à C5).

Clusters	Islands	MSS [mm/day]	MSM [mm/day]	DWR [%]
C1	**Guadeloupe**	8.98	1.50	37
	Dominica	16.24	1.80	39
	Martinique	13.21	1.47	34
	St-Lucia	10.79	1.80	43
	Barbados	4.14	2.07	57
	St-Vincent	7.21	1.80	50
	Flat islands	7.40	3.38	34
C2	**Guadeloupe**	38.68	6.45	22
	Dominica	57.78	6.42	24
	Martinique	45.88	5.10	18
	St-Lucia	33.48	5.58	26
	Barbados	12.12	6.06	42
	St-Vincent	24.55	6.14	37
	Flat islands	17.49	8.52	22
C3	**Guadeloupe**	12.45	2.07	34
	Dominica	19.24	2.14	31
	Martinique	19.29	2.14	24
	St-Lucia	13.64	2.27	33
	Barbados	6.89	3.44	49
	St-Vincent	13.55	3.39	50
	Flat islands	6.51	3.21	32
C4	**Guadeloupe**	6.27	1.04	38
	Dominica	10.67	1.19	36
	Martinique	9.84	1.09	31
	St-Lucia	7.40	1.23	40
	Barbados	4.14	2.01	57
	St-Vincent	7.31	1.83	55
	Flat islands	3.31	1.67	36
C5	**Guadeloupe**	33.48	5.58	24
	Dominica	57.74	6.19	21
	Martinique	46.69	5.19	14
	St-Lucia	30.78	5.13	19
	Barbados	14.60	7.30	41
	St-Vincent	26.46	6.61	35
	Flat islands	9.07	4.65	27

tions des stations météorologiques de la Guadeloupe et de la Martinique.

Les données de TRMM surestiment les pixels sans pluie de même que les classes

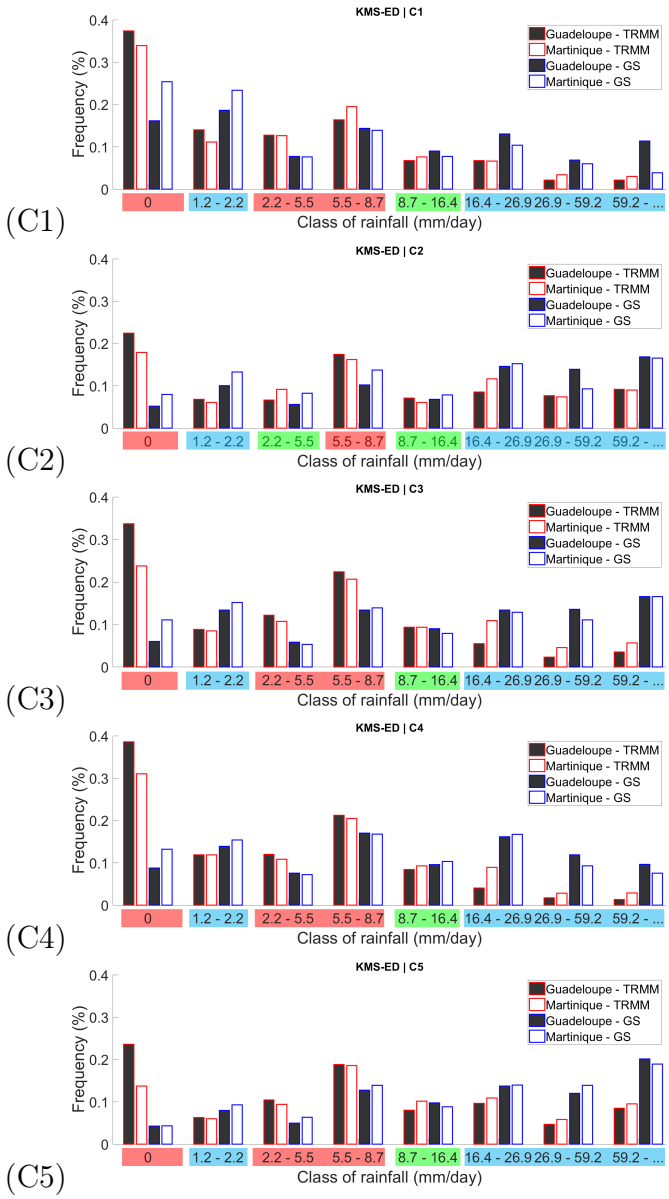


FIGURE 5.18 – Distribution des précipitations TRMM (contour rouge) par rapport aux précipitations des stations au sol (GS) (contour bleu) observées en Guadeloupe (en noir) et en Martinique (en blanc) pour les clusters KMS-ED (de C1 à C5). Les classes qui sont surestimées par TRMM sont mises en évidence en rouge, celles qui sont sous-estimées sont mises en évidence en bleu, et lorsque TRMM est presque similaire à GS, les classes sont mises en évidence en vert.

2.2 à 8.7, et sous-estiment les fortes précipitations. Ces résultats ont été trouvés par [40] pour les pics nuageux de cette région. Les biais signalés ici (par exemple, de faibles précipitations) pourraient être davantage liés aux particularités du microclimat dans une topographie abrupte qu'aux performances de la

physique des modèles et de l'assimilation des données. Cependant, il convient de noter que les précipitations entre 8,7 et 16,4 mm/jour sont relativement bien mesurées par les données du TRMM.

En considérant que la classe 8,7–16,4 mm/jour représente un jour de pluie modérée (bien détectée par TRMM), la Fig 5.19 montre la distribution de fréquence intra-annuelle (axe y) pour la période analysée (axe x).

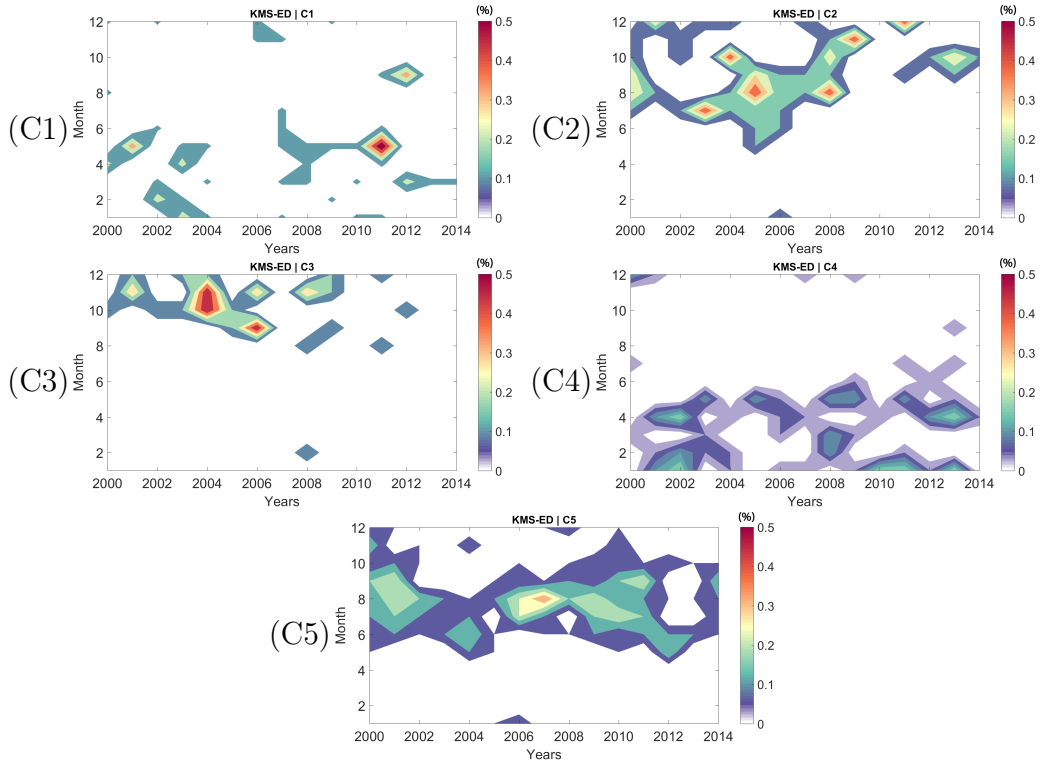


FIGURE 5.19 – Variation de la fréquence d'apparition intra-annuelle (axe y) de la fréquence des pluies modérées en Guadeloupe (8,7–16,4 mm/jour) pour la période analysée (axe x) de 2000 à 2014 dans les cinq différents clusters de KMS-ED (de C1 à C5).

L'évolution intra-annuelle de cette classe de précipitation dans le temps montre une bonne concordance entre les échelles locale et régionale. Pour les saisons sèches (C4) et humides (C5), la classe des précipitations modérées est distribuée presque uniformément avec des fréquences basses. Dans les trois autres groupes, des pics avec des fréquences plus élevées sont apparus, mais la distribution est plus dispersée (éparpillée). De 2002 à 2011, C1, C2 et C3 montrent des maximums ponctuellement isolés de fréquences de pluies modérées, qui

sont observés d'avril à décembre. C1 semble être le groupe de l'inter-saison avril-mai, avec sa transition sèche à pluvieuse, et C3 semble être le groupe de la transition inverse, en novembre-décembre.

5.6 Conclusion

Dans ce chapitre, nous avons soulevé certaines questions théoriques qui pourraient apparaître lors de l'utilisation de la distance L2 comme dissimilarité pour le clustering de champs spatio-temporels météorologiques. Ces questionnements ont pu être établis sur la base d'une analyse de son fonctionnement et de l'usage de l'indice de Silhouette (cf. Chapitre 4).

Il en ressort que L2 est fortement sensible à localisation spatiale des motifs. Dans le cas des précipitations, caractérisées par une forte discontinuité spatio-temporelle, les champs "proches" au sens de la L2 sont rares. D'ailleurs, cette distance tend à agglomérer des champs ayant une structure spatiale commune, bien qu'ils soient par ailleurs différents, en termes d'interprétations physiques.

Afin de pallier à ces difficultés, nous avons proposé une nouvelle mesure de dissimilarité (ED) pour surmonter ces problèmes. A cette fin, il nous a semblé logique d'utiliser des mesures liées à la théorie de l'information, telles que la divergence Kullback-Leibler. D'autres mesures de similarité entre histogrammes existent [64, 11], mais cette dernière a été choisie que nous avons opté pour une une analyse sans "*a priori*" sur la distribution des données.

À partir des patchs des champs de précipitations quotidiennes, nous avons compilé tous les histogrammes et les avons comparé, en utilisant cette divergence. La méthode qui en résulte est toujours une méthode de classification non supervisée. L'innovation concerne l'intégration d'une déviation d'expert pour mieux quantifier la similitude des champs entre eux.

Nous avons également expliqué que le centroïde utilisé par KMS ne représente pas toujours une situation physique réaliste. De plus, il en convient de dire que les centroïdes ne devraient pas toujours être interprétés étant des régimes de temps. Ces derniers donnent une vision trop lissée à notre sens.

La nouvelle mesure ED donne de bien meilleurs résultats que la distance L2. D'un point de vue numérique, l'utilisation de la ED produit systématiquement de meilleurs résultats que la L2. Ainsi, parmi les clusters obtenus par KMS-ED, cinq d'entre eux ont été retenus. La conception de cette nouvelle métrique

est plus sensible aux connaissances des structures météorologiques issues des données d'observations et de la littérature scientifique [61, 12, 73]. De plus, ces résultats ont été analysés par des physiciens de l'atmosphère qui ont confirmé la qualité des informations qui peuvent être extraites de ces clusters.

Contrairement à la L2, la DE est capable de produire différentes configurations qui rendent les structures atmosphériques habituelles clairement identifiables. Les physiciens de l'atmosphère peuvent interpréter les impacts de chaque cluster sur une zone spécifique en fonction de l'emplacement des structures atmosphériques. KMS-L2$_{WIND}$ ne fournit pas cela, les situations représentées sont spatialement assez lisses, les structures habituelles ne sont pas clairement visibles. Elle s'applique à la fois à un champ spatio-temporel complexe et intermittent, telle que la pluie et à un champ classique comme les vents.

La ED proposée fournit des résultats plus fiables et plus interprétables que les méthodes de clustering traditionnelles. C'est le moyen le plus efficace de produire des résultats qui soient vraiment fiables et interprétables. Elle nécessite une séparation spatiale préalable et une quantification du champ pour améliorer la pertinence physique des clusters et ainsi révéler les changements possibles dans les transitions entre les saisons.

Chapitre 6

Conclusion & perspectives

6.1 Conclusion générale

Les travaux présentés dans cette thèse ont principalement cherché à apporter de l'expertise informatique et technique dans les études du domaine de la physique de l'atmosphère, en particulier celles concernant l'analyse du climat. Par analogie aux propos de début de manuscrit, le climat est d'ores et déjà une préoccupation mondiale, de ce fait entre le début et la fin de cette thèse, plus d'une centaine d'études ont été menées par des scientifiques par delà le monde.

En tant qu'informaticiens, nous avons contribué à cet effort collectif, à notre échelle, en produisant une analyse des méthodes "classiques" utilisées dans ce domaine. Ainsi, nous avons identifié deux difficultés majeures. Premièrement, l'évaluation de la qualité des clusters est insuffisante à notre sens, elle est remplacée par une discussion s'appuyant essentiellement sur les travaux préliminaires.

L'un des principaux apports de cette thèse réside en la promotion, dans le domaine de la physique, du renforcement de l'évaluation de la qualité des clusters. Cette évaluation est effectuée en utilisant des indices ou mesures, bien connues dans le domaine informatique, qui permettent de garantir la cohérence structurelle des clusters obtenus. Nous pensons que les physiciens n'utilisent pas ces méthodes d'évaluations, car celles-ci produisent souvent des valeurs très faibles pouvant être en désaccord avec leur analyse experte.

Nous l'avons compris en effectuant, nous aussi, à l'issue de plusieurs clustering, des évaluations avec un indice pertinent d'évaluation de la qualité des clusters, l'indice de Silhouette. Les résultats se sont révélés non concluants. Néanmoins nous en avons tiré un enseignement, les résultats de la mesure semblent dépendre directement de la nature des données. En effet, les indices calculés après clustering de l'intensité du vent, champ plutôt continu, étaient meilleurs que ceux des précipitations, comportant de fortes discontinuités spatiales et temporelles. De ce fait, nous attribuons ces difficultés d'une part, à la complexité et la spécificité des données et d'autre part, aux méthodes classiquement utilisées pour mesurer la proximité entre deux entités.

Ce qui nous mène à la deuxièmement difficulté, nous pensons que les méthodes utilisées pour mesurer la similarité entre les données ne sont pas toujours adaptées à leurs spécificités. Conscients de cette situation, les physiciens utilisent des pré-traitements (filtre, anomalie, etc.) afin d'améliorer leurs résultats [43, 41]. Nous avons alors identifié la norme L2, utilisée dans bon nombre d'algorithmes, comme étant problématique quand il s'agit de déterminer la proximité de deux situations météorologiques.

Ainsi nous avons pu fournir une nouvelle méthodologie d'analyse du climat, reposant essentiellement sur l'usage d'une mesure de dissimilarité, nommée Expert Deviation (ED) et intégrant l'expertise du physicien directement dans les méthodes de clustering. Cette expertise réside essentiellement dans la pertinence du découpage en patch (issu du domaine de l'analyse d'image) visant à cibler ou isoler des structures connues, mais également dans la détermination de la représentation des données issues des patchs (dans nos études, nous avions choisi l'histogramme).

En somme, cette approche reposant techniquement sur l'analyse par patch, l'usage d'une métrique appropriée au sein d'une méthode de clustering, se veut totalement adaptative et donc perfectible. Nous avons donc pu démontrer la pertinence de notre approche, en effectuant des expérimentations avec ED et en évaluant la qualité des clusters produits avec l'indice de Silhouette. Ces travaux ont porté sur deux paramètres météorologiques impactant directement le climat de nos latitudes, le vent et les précipitations.

Notons que lors de nos expériences nous avons principalement utilisé les algorithmes K-Means et de Classification Hiérarchique Ascendante, néanmoins, puisqu'il s'agit d'une mesure, cette dernière peut être intégrée dans tout autre traitement numérique utilisant une comparaison d'image.

Avec du recul, cette entreprise fut tout de même complexe et fastidieuse puisqu'il s'agissait de faire concorder deux méthodologies d'analyse relativement différentes, celle de l'informaticien et celle du physicien. Pour ces deux conceptions, la phase d'apprentissage langage métier et d'appropriation de la thématique primordiale était longue, celle-ci fût garante d'une meilleure coordination des efforts.

En conclusion, nous pensons que la recherche, d'aujourd'hui et de demain, doit tendre vers une conception pluridisciplinaire des sciences, et ainsi fédérer les scientifiques de tout horizon au sein de projet permettant de mieux répondre aux questionnements de notre monde.

Dans la section suivante, nous évoquons nos perspectives pour étude intégrant l'usage de la nouvelle méthodologie apportée dans cette thèse.

6.2 Perspectives

Ces travaux nous ont permis de concevoir un nouvel outil d'analyse des données climatiques qui se révèle plutôt adaptatif dans sa conception, mais également dans son usage, ouvrant ainsi un large panel de possibilités. Dans cette section, nous présentons quelques-unes pour leur caractère innovant.

Si ces travaux ont permis de mettre en avant des configurations de vents et de précipitations qui s'avèrent pertinentes, nous n'avons procédé à aucune analyse conjointe de celles-ci. Une première piste consisterait donc à observer si les clusters de vent et de précipitations obtenus se correspondent, et dans quelles mesures ? Il s'agit là de travaux que nous effectuerons de façon imminente.

Également, il nous semble intéressant de concevoir des mesures de dissimilarité telles que ED, présentée ici, pour chaque paramètre météorologique caractérisant le climat. Nous pourrions alors nous pencher sur l'application de

méthodes de clustering prenant d'emblée en compte plusieurs paramètres météorologiques.

Dans le même ordre d'idée, si l'on sépare les paramètres physiques en deux groupes causaux (effecteurs et impacts), il serait intéressant d'étudier les possibles corrélations entre paramètres. Ceci est possible du fait de l'introduction d'une mesure numérique de qualité des clusters telle que Silhouette : en effectuant un clustering pour un paramètre (a), il est possible d'évaluer la cohérence de la séparation ainsi produite sur un second paramètre (b).

Afin d'apporter plus de robustesse aux résultats obtenus, il serait également intéressant d'intégrer une mesure ou un indice permettant d'évaluer la stabilité des clusters obtenus (exemple Indice de Rand), ce type de méthodes a été utilisée par certains auteurs de la bibliographie.

De façon plus lointaine, nous souhaiterions étudier plus précisément la dynamique temporelle de l'analyse. Si la cohérence temporelle des configurations retenues semble maintenant établie de façon globale, il serait intéressant de déterminer dans quelles mesures une configuration peut mener à une autre. Des mesures de transitions de types Markoviennes offriraient sans doute des résultats intéressants.

L'usage de méthodes de clustering plus récentes est aussi une piste que l'on nous a proposé, telles que MeanShift, Espérence-Maximisation, ClusTree ou encore simpleTS.

Pour terminer, nous proposons d'assembler les couches de neurones suivant le modèle des auto-encodeurs dans la phase d'apprentissage, pour intégrer toute la base de données disponible. Par la suite, nous utiliserons les couches d'encodage pour réduire de manière efficace (mieux qu'une ACP) les dimensions des données. Ce nouveau jeu de données sera passé en entrée des méthodes de clustering usuelles et leurs résultats seront évalués avec silhouette intégrant ED. Notons que des réseaux de neurones convolutifs peuvent être intégrés dans les auto-encodeurs d'effectuer une analyse zonale, similaire à celle que nous réalisons avec la ED.

Comme on peut le voir, les perspectives ne manquent pas et s'il ne fait aucun doute que nous allons persévérer sur ce sujet, il est également clair qu'au cours de ce long et tortueux chemin, nous serons sans doute amenés à réviser nos priorités. Nous espérons que le lecteur aura trouvé autant d'intérêt à la lecture de ce manuscrit, parfois aride, que nous avons pris plaisir à réaliser les travaux qui ont amenés à sa rédaction.

Bibliographie

[1] Alessia Amelio and Clara Pizzuti. A patch-based measure for image dissimilarity. *Neurocomputing*, 171 :362–378, 2016.

[2] Connelly Barnes, Dan B. Goldman, Eli Shechtman, and Adam Finkelstein. The patchmatch randomized matching algorithm for image manipulation. *Commun. ACM*, 54(11) :103–110, November 2011.

[3] Yoshua Bengio. Learning deep architectures for ai. *Foundations and Trends® in Machine Learning*, 2(1) :1–127, 2009.

[4] F. Beucher. *Météorologie tropicale : des alizés au cyclone*. Number vol. 2 in Cours et manuels - Direction de la météorologie. La Documentation Française, 2010.

[5] H. K. D. H. Bhadeshia. Neural networks in materials science. *ISIJ International*, 39(10) :966–979, 1999.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[7] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2) :123–140, Aug 1996.

[8] M. Burlando. The synoptic-scale surface wind climate regimes of the mediterranean sea according to the cluster analysis of era-40 wind fields. *Theoretical and Applied Climatology*, 96(1) :69–83, April 2009.

[9] Copernicus Climate Change Service (C3S). Era5 : Fifth generation of ecmwf atmospheric reanalyses of the global climate. *Copernicus Climate Change Service Climate Data Store (CDS)*, 2017.

[10] T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1) :1–27, 1974.

[11] Sung-Hyuk Cha and Sargur N. Srihari. On measuring the distance between histograms. *Pattern Recognition*, 35(6) :1355 – 1370, 2002.

[12] Xsitaaz T. Chadee and Ricardo M. Clarke. Daily near-surface large-scale atmospheric circulation patterns over the wider caribbean. *Climate Dynamics*, 44(11) :2927–2946, Juin 2015.

[13] M. Claussen, L. Mysak, A. Weaver, M. Crucifix, T. Fichefet, M.-F. Loutre, S. Weber, J. Alcamo, V. Alexeev, A. Berger, R. Calov, A. Ganopolski, H. Goosse, G. Lohmann, F. Lunkeit, I. Mokhov, V. Petoukhov, P. Stone, and Z. Wang. Earth system models of intermediate complexity : closing the gap in the spectrum of climate system models. *Climate Dynamics*, 18(7) :579–586, Mar 2002.

[14] D. Comaniciu and P. Meer. Mean shift : a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5) :603–619, 2002.

[15] Antoine Cornuéjols, Laurent Miclet, Yves Kodratoff, and Tom Mitchell. *Apprentissage artificiel : Concepts et algorithmes (EYROLLES)*. Eyrolles, 2002.

[16] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[17] David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1(2) :224–227, February 1979.

[18] Robert E. Davis, Bruce P. Hayden, David A. Gay, William L. Phillips, and Gregory V. Jones. The north atlantic subtropical anticyclone. *Journal of Climate*, 10(4) :728–744, 1997.

[19] Renato Cordeiro [de Amorim] and Christian Hennig. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, 324 :126 – 145, 2015.

[20] D. P. Dee, S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. Van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geea, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Holma, L. Isaksen, P. Kallberg, M. Köhler, M. Matricardi, M. A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart. The era-interim reanalysis : configuration and performance of the data assimilation system. *Quartely Journal of Royal Meteorological Society*, 137 :553–597, April 2011.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38, 1977.

[22] S. M. Diffey, A. B. Smith, A. H. Welsh, and B. R. Cullis. A new reml (parameter expanded) em algorithm for linear mixed models. *Australian & New Zealand Journal of Statistics*, 59(4) :433–448, 2017.

[23] Liviu Petrisor Dinu, Radu-Tudor Ionescu, and Marius Popescu. Local patch dissimilarity for images. In Tingwen Huang, Zhigang Zeng, Chuandong Li, and Chi Sing Leung, editors, *Neural Information Processing*, pages 117–126, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[24] Jason P. Dunion. Rewriting the climatology of the tropical north atlantic and caribbean sea atmosphere. *Journal of Climate*, 24(3) :893–908, 2011.

[25] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[26] David M. Fratantoni and Deborah A. Glickson. North brazil current ring generation and evolution observed with seawifs. *Journal of Physical Oceanography*, 32(3) :1058–1074, 2002.

[27] David M. Fratantoni and Philip L. Richardson. The evolution and demise of north brazil current rings. *Journal of Physical Oceanography*, 36(7) :1241–1264, 2006.

[28] Huffman G., Bolvin D., Braithwaite D., Hsu K., Joyce R., and Xie P. Integrated multi-satellite retrievals for gpm (imerg). *NASA's Precipitation Processing Center*, version 4.4., March 2015.

[29] Florent Gasparin, Eric Greiner, Jean-Michel Lellouche, Olivier Legalloudec, Gilles Garric, Yann Drillet, Romain Bourdallé-Badie, Pierre-Yves Le Traon, Elisabeth Rémy, and Marie Drévillon. A large-scale view of oceanic variability from 2007 to 2015 in the global high resolution monitoring and forecasting system at mercator océan. *Journal of Marine Systems*, 187 :260 – 276, 2018.

[30] Hinton Geoffrey and Sejnowski Terrence. *Unsupervised Learning : Foundations of Neural Computation (Computational Neuroscience)*. A Bradford Book, 1999.

[31] Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3) :419–435, 2002.

[32] S. Gokila, K. Ananda Kumar, and A. Bharathi. Different versions of k-mean clustering in complete set of numerical data points. *International Journal of Scientific Engineering and Applied Science*, 2, July 2016.

[33] G. Guo and C. R. Dyer. Patch-based image correlation with rapid filtering. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, June 2007.

[34] Le Hou, Dimitris Samaras, Tahsin M. Kurç, Yi Gao, James E. Davis, and Joel H. Saltz. Efficient multiple instance convolutional neural networks for gigapixel resolution image classification. *CoRR*, abs/1504.07947, 2015.

[35] Anil K. Jain. Data clustering : 50 years beyond k-means. In *International Conference on Pattern Recognition (ICPR)*. ICPR, December 2008.

[36] P. M. James. An assessment of european synoptic variability in hadley centre global environmental models based on an objective classification of weather regimes. *Climate Dynamics*, 27(2) :215–231, Aug 2006.

[37] Thorsten Joachims. Text categorization with support vector machines : Learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Machine Learning : ECML-98*, pages 137–142, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[38] Kerstin Jochumsen, Monika Rhein, Sabine Hüttl-Kabus, and Claus W. Böning. On the propagation and decay of north brazil current rings. *Journal of Geophysical Research : Oceans*, 115(C10), 2010.

[39] Mark R. Jury. An intercomparison of observational, reanalysis, satellite, and coupled model data on mean rainfall in the caribbean. *Journal of Hydrometeorology*, 10(2) :413–430, 2009.

[40] Mark R. Jury and Didier Bernard. Climate trends in the east antilles islands. *International Journal of Climatology*, 40(1) :36–51, 2019.

[41] Mark R. Jury and Björn A. Malmgren. Joint modes of climate variability across the inter-americas. *International Journal of Climatology*, 32(7) :1033–1046, 2012.

[42] Leonard Kaufman and Peter J. Roussew. Finding groups in data - an introduction to cluster analysis. *Journal of Applied Meteorology*, pages 1131–1147, August 1990.

[43] Pirmin Kaufmann and C. David Whiteman. Cluster-analysis classification of wintertime wind patterns in the grand canyon region. *Journal of Applied Meteorology*, 38(8) :1131–1147, 1999.

[44] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biol. Cybern*, 43 :59—69, 1982.

[45] Teuvo Kohonen and Timo Honkela. Kohonen network. *Scholarpedia*, 2(1) :1568, January 2007.

[46] Mineichi Kudo and Jack Sklansky. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33(1) :25 – 41, 2000.

[47] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22 :79–86, 1951.

[48] R. Kwitt and A. Uhl. Image similarity measurement by kullback-leibler divergences between complex wavelet subband statistics for texture retrieval. In *2008 15th IEEE International Conference on Image Processing*, pages 933–936, Oct 2008.

[49] Stéphane Lallich and Philippe Lenca. Indices de qualité en clustering. In *Journée thématique : clustering et co-clustering*, Issy Les Moulineaux, France, Octobre 2015. Société française de classification.

[50] Cheng-Yuan Liou, Wei-Chen Cheng, Jiun-Wei Liou, and Daw-Ran Liou. Autoencoder for words. *Neurocomputing*, 139 :84 – 96, 2014.

[51] Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2) :130–141, 1963.

[52] Ester Martin, Kriegel Hans-Peter, Sander Jörg, and Xu Xiaowei. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226–231. Knowledge Discovery and Data Mining, AAAI Press, 1996.

[53] Carlos Martinez, Lisa Goddard, Yochanan Kushnir, and Mingfang Ting. Seasonal climatology and dynamical mechanisms of rainfall in the caribbean. *Climate Dynamics*, Jan 2019.

[54] Adam C. Martiny, Chau T. A. Pham, Francois W. Primeau, Jasper A. Vrugt, J. Keith Moore, Simon A. Levin, and Michael W. Lomas. Strong latitudinal patterns in the elemental ratios of marine plankton and organic matter. *Nature Geoscience*, 6(4) :279–283, Apr 2013.

[55] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4) :115–133, Dec 1943.

[56] R. Mehrotra and Ashish Sharma. A nonparametric nonhomogeneous hidden markov model for downscaling of multisite daily rainfall occurrences. *Journal of Geophysical Research : Atmospheres*, 110(D16), 2005.

[57] Paul-Antoine Michelangeli, Robert Vautard, and Bernard Legras. Weather regimes : Recurrence and quasi stationarity. *Journal of the Atmospheric Sciences*, 52(8) :1237–1256, 1995.

[58] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[59] C. Monteleoni, G. A. Schmidt, F. Alexander, A. Niculescu-Mizil, K. Steinhaeuser, M. Tippett, A. Banerjee, M. B. Blumenthal, A. R. Ganguly, J. E. Smerdon, and M. Tedesco. *Climate informatics*, pages 81–126. Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC, 2013.

[60] Vincent Moron, Romain Frelat, Pierre Karly Jean-Jeune, and Cédric Gaucherel. Interannual and intra-annual variability of rainfall in haiti (1905–2005). *Climate Dynamics*, 45(3) :915–932, Aug 2015.

[61] Vincent Moron, Isabelle Gouirand, and Michael Taylor. Weather types across the caribbean basin and their relationship with rainfall and sea surface temperature. *Climate Dynamics*, 47(1) :601–621, Jul 2016.

[62] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method : Which algorithms implement ward's criterion ? *Journal of Classification*, 31(3) :274–295, Oct 2014.

[63] Frank Nielsen. *Hierarchical Clustering*, pages 195–211. Springer International Publishing, Cham, 2016.

[64] I. Olkin and F. Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48 :257 – 263, 1982.

[65] Shraddha Pandit and Suchita Gupta. A comparative study on distance measuring approches for clustering. *International Journal of Research in Computer Science*, 2 :29–31, 2011.

[66] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1) :81–106, Mar 1986.

[67] Bellman R.E. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.

[68] Lior Rokach and Oded Z. Maimom. *Clustering Methods*, chapter 15, pages 321–352. Springer, 2010.

[69] Peter J. Rousseeuw. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20 :53–65, 1987.

[70] Jung-Hee Ryu and Katharine Hayhoe. Understanding the sources of caribbean precipitation biases in cmip3 and cmip5 simulations. *Climate Dynamics*, 42(11) :3233–3252, Jun 2014.

[71] Dennis Shea. The climate data guide : Common climate data formats : Overview. National Center for Atmospheric Research Staff (Eds), December 2013. https ://climatedataguide.ucar.edu/climate-data-tools-and-analysis/common-climate-data-formats-overview.

[72] Pandit Shraddha and Gupta Suchita. Computer science and technology. computing. data processing. *International journal of research in computer science*, pages 29–31, 2011.

[73] Tannecia S. Stephenson, Lucie A. Vincent, Theodore Allen, Cedric J. Van Meerbeeck, Natalie McLean, Thomas C. Peterson, Michael A. Taylor, Arlene P. Aaron-Morrison, Thomas Auguste, Didier Bernard, Joffrey R. I. Boekhoudt, Rosalind C. Blenman, George C. Braithwaite, Glenroy Brown, Mary Butler, Catherine J. M. Cumberbatch, Sheryl Etienne-Leblanc, Dale E. Lake, Delver E. Martin, Joan L. McDonald, Maria Ozoria Zaruela, Avalon O. Porter, Mayra Santana Ramirez, Gerard A. Tamar, Bridget A. Roberts, Sukarni Sallons Mitro, Adrian Shaw, Jacqueline M. Spence, Amos Winter, and Adrian R. Trotman. Changes in extreme temperature and precipitation in the caribbean region, 1961–2010. *International Journal of Climatology*, 34(9) :2957–2971, 2014.

[74] Fernán Sáenz and Ana M. Durán-Quesada. A climatology of low level wind regimes over central america using a weather type classification approach. *Frontiers in Earth Science*, 3 :15, April 2015.

[75] Robert Vautard. Multiple weather regimes over the noth atlantitc : Analysis of precutsors and successors. *Monthly Weather Review*, 118 :2056–259, 1990.

[76] Robert Vautard and Bernard Legras. On the source of midlatitude low-frequency variability. part ii : Nonlinear equilibration of weather regimes. *Journal of the Atmospheric Sciences*, 45(20) :2845–2867, 1988.

[77] N. Vigaud and A.W. Robertson. Convection regimes and tropical-midlatitude interactions over the intra-american seas from may to november. *International Journal of Climatology*, 37(S1) :987–1000, Mars 2017.

[78] Jr. Ward, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58 :236–244, 1963.

[79] Howard W.R. Pattern recognition and machine learning. *Kybernetes*, 36(2) :275–275, Jan 2007.

[80] Li Xin. *Perceptual Digital Imaging : Methods and Applications*, volume 1, chapter Patch-Based Image Processing : From Dictionary Learning to Structural Clustering. CRC Press, March 2017.

[81] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8) :790–799, 1995.

[82] F. Zhang, Y. Song, W. Cai, M. Lee, Y. Zhou, H. Huang, S. Shan, M. J. Fulham, and D. D. Feng. Lung nodule classification with multilevel patch-based context analysis. *IEEE Transactions on Biomedical Engineering*, 61(4) :1155–1166, 2014.

[83] Qin Zhang, Huug Van Den Dool, Suru Saha, Malaquias Peña, Emily Becker, Peitao Peng, and Jin Huang. Preliminary evaluation of multi-model ensemble system for monthly and seasonal prediction. In NOAA's National Weather Service, editor, *Science and Technology Infusion Climate Bulletin*, pages 124–131, 2011.