Classification des précipitations sur les massifs alpins français

Fanny Pagnier, Didier Coquin, Frédéric Pourraz, Hervé Verjus, Gilles Mauris

Classification des précipitations sur les massifs alpins français
Classification of precipitation in the French Alps

F. Pagnier[1] D. Coquin[1] F. Pourraz[1] H. Verjus[1] G. Mauris[1]

[1] LISTIC - Université Savoie Mont Blanc, 5 chemin de Bellevue, Annecy-le-Vieux, 74940 Annecy, France
{fanny.pagnier, didier.coquin, frederic.pourraz, herve.verjus, gilles.mauris}@univ-smb.fr

Résumé

Notre étude s'intéresse à la classification de journées de l'hiver 2017-2018 en fonction des conditions de précipitations. L'objectif est d'expliciter les étendues présentant le même type de conditions en lien avec différentes tendances météorologiques à l'échelle des Alpes françaises. Pour ce faire, nous mettons en œuvre une succession de méthodes classiques d'analyse statistique et de data-mining (Analyse en Composantes Principales, Classification Ascendante Hiérarchique et K-Moyennes). L'étude est conduite à travers une comparaison entre deux jeux de données de dimensions différentes : 90 et 15 journées. Le jeu de données de petite dimension (15 journées) a pour intérêt de confirmer les grandes tendances issues de l'analyse du jeu de données de plus grande dimension (90 journées) tout en ajoutant une nouvelle classe correspondant à une situation bien réelle mais sous-représentée et donc absorbée par le grand jeu de données. Menant cette étude dans le contexte du risque d'avalanche, nous montrons aussi qu'une corrélation peut être observée entre la localisation des étendues principalement affectées par les différentes tendances et la localisation des avalanches spontanées observées pour chaque journée correspondante.

Mots Clef

Classification, Analyse en Composantes Principales, Classification Ascendante Hiérarchique, K-moyennes, Précipitations.

Abstract

This study deals with classification conducted on winter days (from the 2017-2018 winter season) according to rainfall conditions. The aim is to detect which places, on the French Alps, receive the same amount of rainfall depending on various meteorological trends. To fulfill this objective, three methods are successively used : Principal Components Analysis, Hierarchical Ascending Classification and K-means. These methods are statistical analysis and data-mining classical methods. A comparison between two data sets, which have different dimensions (90 days and 15 days) is given. The smallest data set (15 days) confirms the overall trends given by the analysis of the biggest one (90 days) and adds a new cluster that corresponds to a real situation but which is under-represented in the 90 days data set. Working in the context of avalanche risk, we show that a correlation is observed between the places particularly affected by the various rainfall trends and the location of the spontaneous observed avalanches during the corresponding days.

Keywords

Classification, Principal Components Analysis, Hierarchical Ascending Classifcation , K-means, Precipitations

1 Introduction

En matière de gestion du risque d'avalanche, les observations d'avalanches récentes sont un très bon indicateur du niveau de danger en cours. Elles sont en effet prises en compte comme l'un des paramètres d'entrée de nombreuses méthodes d'aide à la décision. Cependant, si cet indicateur est évoqué (il est généralement question de savoir si des avalanches ont eu lieu à proximité au cours des derniers jours), il n'est jamais véritablement quantifié et aucune métrique ne lui est associé. Afin de préciser ce paramètre, il devient essentiel de connaître la zone d'influence d'une avalanche, i.e. pouvoir dire, au vu des conditions, quelles étendues sont susceptibles de réagir de la même manière en termes d'avalanches. Ainsi, l'objectif visé *in fine* est de déterminer jusqu'où une observation d'avalanche donnée a un sens et donc quelles étendues sont susceptibles de présenter le même niveau de danger. Dans ce contexte général, notre étude vise donc à connaître quelles portions des Alpes françaises présentent les mêmes conditions au même moment. Par *conditions*, nous faisons référence ici uniquement aux conditions de précipitations. Pour atteindre cet objectif, nous souhaitons réaliser une classification de journées à partir d'enregistrements de données météorologiques (mesure journalière de précipitations) sur des stations de mesures réparties sur les Alpes françaises. Nous cherchons ainsi à trier les journées similaires en termes de localisation des principaux cumuls de précipitations. Chacune des classes obtenues, donc chaque groupe de journées, peut alors être associée à une tendance météo. Nous distinguons en effet des différences dans l'orientation du flux météorologique (selon le sens d'arrivée de la perturbation, ce ne sont pas les mêmes massifs qui reçoivent la plus grande part des précipitations) ou l'intensité du phénomène observé.

Plusieurs travaux [6] [1] [5] ont abordé l'étude des précipitations dans différents pays et régions mais il n'y a pas de méthode générale qui pourrait être appliquée à notre situation. Nous présentons alors une démarche de classification adaptée pour répondre aux objectifs visés. Pour cela, nous

utilisons une succession de méthodes classiques d'analyse statistique et de data-mining : Analyse en Composantes Principales (ACP) [11], Classification Ascendante Hiérarchique (CAH) [10] et K-moyennes [4] [14] successivement (voir section 2.2).
Puisqu'il s'agissait de mettre en œuvre une méthode statistique (ACP), nous avons choisi de travailler sur un jeu de données de trois mois afin qu'il dispose d'un nombre d'individus suffisant (90 journées). De plus, nous avons aussi fait le choix d'étudier un jeu de données nettement plus réduit et constitué de journées dites *typiques*. Ceci permet de pallier à un éventuel bruit contenu dans le jeu de données de plus grande dimension et d'éviter toute disproportion entre le nombre de journées associées aux différentes tendances météo susceptibles d'être présentes dans le jeu de données. Ces deux jeux de données sont présentés en section 2.1. Ainsi, quelle différence apporte l'étude de l'un ou l'autre de ces jeux de données, et quels sont alors l'influence et l'intérêt de l'étude d'un jeu de données de relativement grande ou, à l'inverse, petite dimension ? Pour répondre à cela, nous confrontons, au cours de notre étude, les différents résultats obtenus sur chacun de ces deux jeux de données (sections 3 à 5). La section 2 présente d'abord les données et la méthode utilisée. Les sections 3 et 4 décrivent ensuite les classifications réalisées sur les jeux de données de 90 et 15 jours respectivement, puis la section 5 aborde les résultats de la classification d'un $16^{ème}$ jour.

2 Données et démarche mise en œuvre

2.1 Données

Stations de mesure. Nos travaux s'appuient sur les données issues des relevés des stations nivo-météorologiques réalisés par les partenaires conventionnés avec Météo-France [8]. Ce réseau contient 89 stations sur les Alpes françaises. Pour notre étude, nous avons sélectionné 23 stations de mesures (dont les localisations sont visibles sur les figures 2 et 4) de façon à :

- conserver une disposition la mieux répartie possible sur l'ensemble des massifs alpins français ;

- faire en sorte que les stations conservées soient concernées par un minimum de données manquantes ;

- intégrer des stations situées très à l'est de la zone d'étude pour s'assurer de détecter l'effet d'un retour d'est (qui reste très localisé).

Comme évoqué en introduction, nous ne nous intéressons alors qu'aux mesures de précipitations. Les valeurs (hauteur en mm) sont communiquées quotidiennement et correspondent au cumul de précipitations sur 24 heures, mesuré aux alentours de 8 heures le jour considéré. Chaque jour est théoriquement associé à une valeur, mais nous verrons par la suite que les données sont cependant entachées de nombreuses valeurs manquantes.

Journées. Les individus étudiés sont des journées de l'hiver 2017-2018. Nous avons constitué deux jeux de données d'étude comme suit :

- un premier jeu de données contenant les jours de janvier à mars 2018, soit 90 journées ;

- un second jeu de données contenant les journées comprises entre le 15 décembre 2017 et le 31 mars 2018 associées à des pics d'avalanches ($\geq$ 5 avalanches) d'après la base de données de data-avalanche.org. L'objectif de cette sélection est de ne travailler que sur des journées caractéristiques en termes d'avalanches observées, et donc potentiellement associées à des précipitations marquées, le tout sans recourir à une connaissance experte. Ces critères ont engendré la sélection de 15 journées.

Traitement des données manquantes. Les jeux de données ainsi constitués comportent de nombreuses données manquantes. Les méthodes mises en œuvre dans cette étude ne pouvant l'être avec des données manquantes, un traitement a été réalisé pour les combler : nous avons remplacé chaque donnée manquante par la valeur de la moyenne de la station de mesure sur l'ensemble des journées du jeu de données (90 jours ou 15 jours selon les cas).

2.2 Démarche mise en œuvre

La démarche mise en œuvre est composée de trois méthodes utilisées successivement : ACP, CAH et K-moyennes [12] [9]. La figure 1 illustre le processus de cette démarche.
En partant d'un jeu de données initial, composé de x critères d'analyse appelés variables (23 dans notre cas, les stations de mesure correspondant à nos critères d'analyse), le rôle de l'ACP est de transformer ces variables en composantes principales, dont seules les premières sont véritablement informatives (expliquant chacune la plus grande part d'inertie du nuage de points initial). Chaque composante principale correspond alors à une combinaison linéaire des variables initiales. Le rôle de l'ACP est donc de réduire les x variables initiales à un nombre de composantes restreint qui facilite l'interprétation des résultats de classification.
Le choix du nombre de composantes à conserver à l'issue de l'ACP est essentiel. Plusieurs règles permettent d'estimer le nombre optimal de composantes à conserver (règle de Kaiser-Guttman, règle de Karlis-Spinaki, test des bâtons brisés, ...). Une fois appliquées à nos deux jeux de données, elles suggèrent de conserver entre 2 et 5 composantes pour le premier et entre 2 et 6 pour le second. Pour trancher, il est aussi essentiel de conserver uniquement les composantes pour lesquelles nous savons expliquer la réalité physique qu'elles traduisent. C'est pourquoi nous n'avons retenu que les 3 premières composantes principales, dont l'inertie initiale s'élève à 68,1 % pour le jeu de données de 90 jours et 71 % pour les jeux de données de 15 jours.
La figure 2 représente la contribution des différentes variables (donc des différentes stations) aux composantes 1 à 3 de l'ACP ainsi que les valeurs prises par les stations sur chacun des 3 axes obtenus en sortie. En regardant en parallèle la position des individus sur l'axe 1 et leurs précipitations associées (figure 3), nous pouvons en déduire

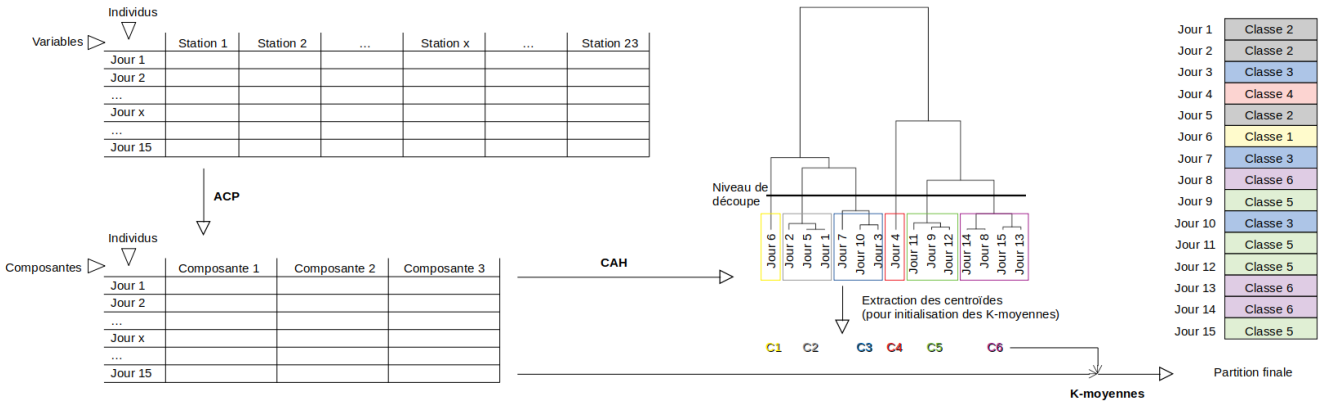


FIGURE 1 – Processus de la démarche mise en œuvre

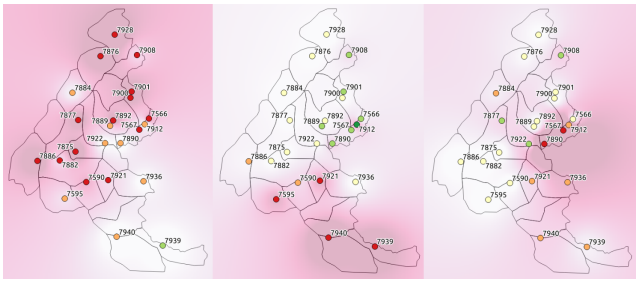


FIGURE 2 – Contribution des stations aux 3 premières composantes de l'ACP (de gauche à droite, composante 1 à composante 3) et valeurs associées (cas du jeu de données de 15 jours). Contribution des différentes stations : dégradé du rose (forte contribution) au blanc (contribution nulle) ; Valeurs prises par les stations sur l'axe : dégradé du rouge (valeurs négatives) vers le vert (valeurs positives).

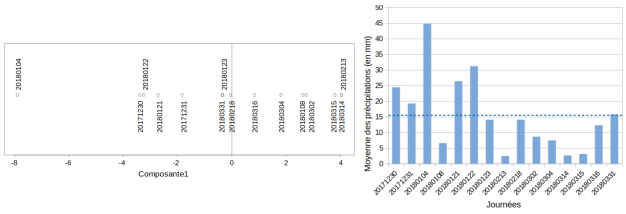


FIGURE 3 – Représentation des 15 journées selon la composante 1 de l'ACP et précipitations enregistrées. La ligne pointillée correspond à la valeur moyenne de l'histogramme.

qu'il correspond à l'opposition journées sèches / journées humides (ou encore à une tendance météo générale de type flux de nord ouest qui arrose la majeure partie des Alpes françaises). Les axes 2 et 3 marquent respectivement une opposition sud / nord et est / ouest. La figure 4 représente la position des stations selon les composantes 2 et 3 de l'ACP permettant d'ores et déjà de distinguer 3 zones principales.

Notons que les données sont centrées réduites (i.e. soustraction de la moyenne et division par l'écart-type) avant l'ACP afin de donner la même importance à chacune des

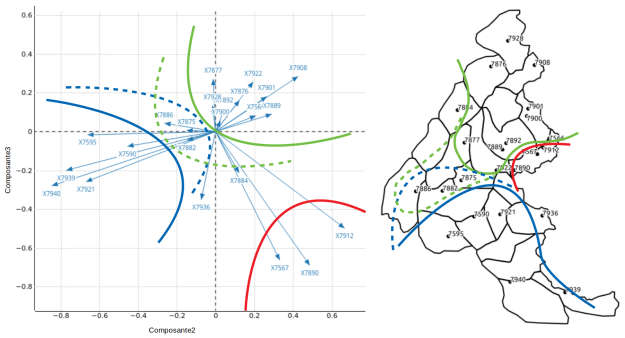


FIGURE 4 – Représentation des stations selon les composantes 2 et 3 de l'ACP (cas du jeu de données de 15 jours)

variables initiales, quel que soit leur ordre de grandeur.

À l'issue de l'ACP, nous réalisons successivement une **CAH** et les **K-moyennes** pour tirer les avantages de chacune des deux méthodes.

La CAH permet de construire des classes par agrégation d'éléments deux à deux. Au cours des itérations successives, elle construit donc de n-1 classes à une seule classe contenant l'ensemble des n individus. Nous utilisons la CAH avec la distance euclidienne et le critère d'aggrégation de Ward [13], qui consiste à minimiser la perte d'inertie inter-classe lors de l'agrégation de deux classes.

Grâce à la fonction HCPC du package FactoMineR de R, qui suggère un niveau de découpe tel que la partition soit celle avec la plus grande perte d'inertie relative [3], nous obtenons automatiquement le nombre de classes. En revanche, avec la CAH, si un individu est mal classé au départ il le reste jusqu'à la fin. C'est pourquoi, afin d'affiner la classification, nous mettons ensuite en œuvre l'algorithme des K-moyennes (puisqu'il permet à un individu mal classé de changer de classe au cours des itérations successives). L'algorithme des K-moyennes est initialisé sur la base des k classes obtenues à l'issue de la CAH, ce qui évite de laisser une initialisation aléatoire. Sans connaissance experte il est difficile de fixer la valeur de k. Souhaitant développer une méthode générique, duplicable sur d'autres données (par exemple à l'échelle des Alpes Suisses), nous avons

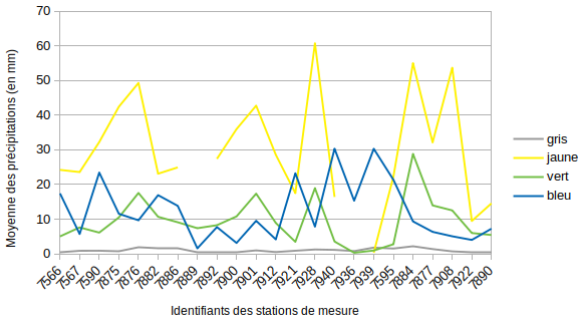


FIGURE 5 – Profils des 4 classes obtenues
Certains profils sont discontinus du fait des valeurs manquantes (annotation valable pour les figures 5, 7 et 11).

cherché à proposer une démarche permettant une première estimation de la valeur de k. Nous utilisons donc la CAH comme première classification non supervisée suggérant automatiquement un niveau de découpe adapté aux données. Notons que la valeur obtenue à l'issue de la CAH va différer d'un jeu de donnée à un autre, mais permet de dégager une première valeur utile à l'analyse. Ensuite, nous verrons que certaines classes obtenues en sortie (ie. à l'issue de la consolidation par les K-moyennes) peuvent présenter les mêmes types de profils de précipitations.

3 Jeu de données de 90 jours : mise en œuvre de la démarche ACP - CAH - K-moyennes

Dans cette section, nous travaillons sur le jeu de données constitué des 90 journées de janvier à mars 2018 (soit 90 individus), avec les informations des précipitations enregistrées sur les 23 stations de mesure étudiées (soit 23 variables). À l'issue de l'ACP, la dimension du jeu de données est réduite à 3 composantes (voir section 2.2). Sur ce résultat, nous avons mis en œuvre la CAH, qui préconise alors une classification en 4 classes. Nous avons ensuite mis en œuvre les K-moyennes pour obtenir la partition finale ; l'algorithme des K-moyennes étant initialisé avec les 4 centroïdes obtenus à partir des 4 classes issues de la CAH.

Résultats. Nous obtenons ainsi 4 classes distinctes (résultat de la classification, k=4). Pour chacune des classes, nous avons pu établir un profil traduisant les moyennes de précipitations reçues en chaque station de mesure. Ces 4 classes présentent 3 profils qui diffèrent dans leur forme : 1) *jaune* et *vert*, 2) *bleu*, 3) *gris* (figure 5). Ces 3 profils correspondent à 3 tendances météorologiques différentes. Le groupe 1 (*jaune* et *vert*) correspond à une même tendance mais avec une différence d'intensité (en termes de cumul de précipitations reçues). Le groupe 2 (*bleu*) est caractérisé par un profil dont la forme se distingue du précédent. Enfin, le groupe 3 (*gris*) correspond aux jours n'ayant reçu que peu ou pas de précipitations (avec 1 mm en moyenne). En reportant les valeurs des profils (figure 5) sur une carte,

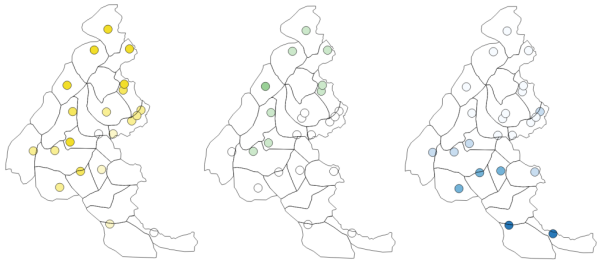


FIGURE 6 – Identification des étendues principalement affectées par les précipitations pour chacun des groupes *jaune*, *vert* et *bleu*. (Dégradé de couleur réalisé avec un pas de 10 mm de précipitations.)

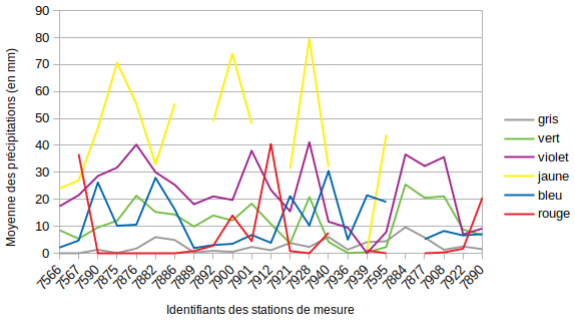


FIGURE 7 – Profils des 6 classes obtenues

on peut alors identifier, pour chaque groupe de journées et donc pour chacune des trois tendances météorologiques, les étendues qui sont principalement affectées (figure 6). Ceci met à nouveau en évidence que les étendues affectées par les précipitations dans le cas des groupes *jaune* et *vert* coïncident.

4 Jeu de données de 15 jours issus des pics d'avalanches

4.1 Mise en œuvre de la démarche ACP - CAH - K-moyennes

Dans cette section, nous allons voir si le jeu de données de 15 jours issus des pics d'avalanches permet de faire ressortir des profils similaires à ceux obtenus sur le jeu de données de 90 jours. À l'issue de l'ACP, la dimension du jeu de données est réduite à 3 composantes (voir section 2.2). Sur ce résultat, la CAH préconise une classification en 6 classes. Nous avons alors mis en œuvre l'algorithme des K-moyennes pour obtenir la partition finale, en l'initialisant avec les 6 centroïdes des 6 classes obtenues à l'issue de la CAH.

Résultats. Nous obtenons ainsi 6 classes distinctes présentant 4 profils qui diffèrent dans leur forme et correspondent à 4 tendances météorologiques différentes (figure 7). A l'instar de ce que nous obtenions en section 3, le groupe *gris* correspond aux jours n'ayant reçu que peu ou

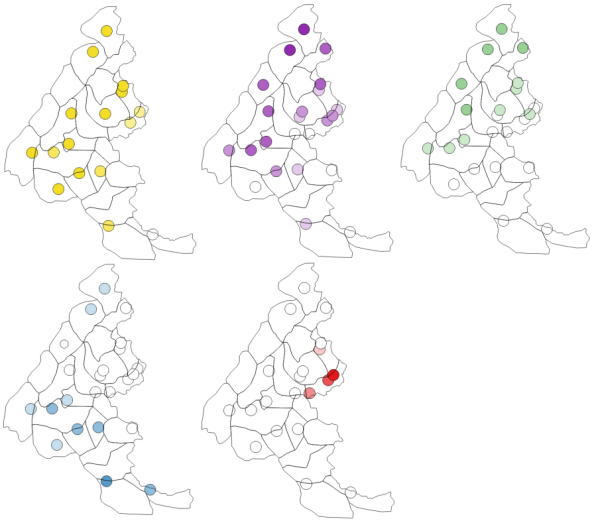


FIGURE 8 – Identification des étendues principalement affectées par les précipitations pour chacun des groupes *jaune*, *vert*, *violet*, *bleu* et *rouge*. (Dégradé de couleur réalisé avec un pas de 10 mm de précipitations.)

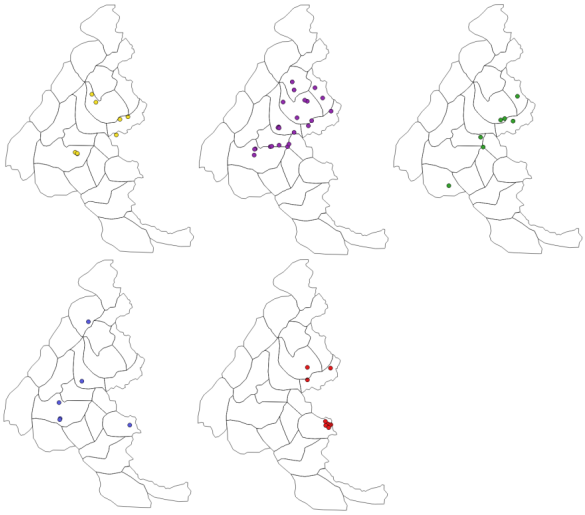


FIGURE 9 – Localisation des avalanches spontanées observées sur chaque journée constituant les 5 groupes *jaune*, *vert*, *violet*, *bleu* et *rouge*

pas de précipitations. Les groupes *jaune*, *violet* et *vert* présentent des profils similaires (dans leur forme) mais se différencient par leur intensité. Les groupes *bleu* et *rouge* présentent deux profils différents et qui se distinguent des trois autres précédemment évoqués.
Comme ceci a été fait en section 3, en reportant les valeurs de la figure 7 sur une carte, nous pouvons alors identifier, pour chaque groupe de journées et donc pour chaque tendance météorologique associée, les étendues qui sont principalement affectées (figure 8).
Nous avons ensuite cherché à identifier s'il y avait ou non une corrélation avec les avalanches observées à ces mêmes périodes. Pour ce faire, nous nous sommes intéressés uniquement aux avalanches spontanées dont le principal facteur de déclenchement est lié aux dernières précipitations reçues, induisant une surcharge récente. Les autres avalanches (déclenchées accidentellement) sont quant à elles plus complexes et font intervenir, de façon non négligeable, de nombreux autres paramètres. Nous avons alors mis en correspondance les avalanches observées le jour J et les précipitations enregistrées le jour J à 8 heures, correspondant donc au cumul des 24 heures précédentes, pour chacun des jours des 5 classes obtenues. Les avalanches sont ainsi observées dans les heures qui suivent la surcharge étudiée et il y a donc adéquation entre les deux temporalités. La figure 9 montre que les localisations des avalanches spontanées qui se sont déclenchées pour chacune des journées considérées conïncident avec les zones identifiées pour chacun des groupes de la figure 8. Notons que ces cartes représentent les avalanches spontanées *observées*, ce qui, nécessairement, ne traduit pas l'ensemble des avalanches qui se sont produites. En particulier, nous savons qu'un défaut d'observation est envisageable dans les massifs situés au nord ouest (affectant principalement les observations des groupes *jaune*, *vert* et *violet*) et au sud (affectant principalement les observations du groupe *bleu*). Nous pouvons aussi donner une explication du nombre d'avalanches observées dans les trois cas liés à la tendance affectant le nord ouest (groupes *jaune*, *violet* et *vert*). Le groupe *vert* est associé à une plus faible intensité, ce qui produit moins de départs spontanés. Le groupe *jaune* est associé à un phénomène particulièrement marqué (qui a pu durer plusieurs jours avec des intensités différentes). De ce fait : 1) certaines avalanches ont pu être recouvertes et donc non visibles après la perturbation et 2) les observations n'ont certainement pas été faites le jour même, leurs dates d'occurence ont donc parfois pu être estimées, ce qui a pu générer un décalage d'un à plusieurs jours. Le groupe *violet*, intermédiaire, est quant-à lui le cas le plus propice pour avoir des observations en nombre et fiables.
Quoi qu'il en soit, nous pouvons distinguer l'existence d'une corrélation entre les zones principalement affectées par les précipitations (figure 8) et les zones de localisation des avalanches spontanées (figure 9) : pour chaque groupe, ces deux étendues géographiques coïncident en effet. Ceci est particulièrement visible dans le cas de la tendance rouge, correspondant à un phénomène très localisé, condensé à l'extrémité est des Alpes françaises.

4.2 Confrontation avec le jeu de données de 90 jours

La figure 10 présente les deux classements obtenus sur les 15 journées issues des pics d'avalanches dans le cas de leur étude au sein du jeu de données de 90 jours (qui a donné 4 classes) ou de 15 jours (qui a donné 6 classes).
On constate que :

- les journées classées dans les groupes *gris* et *bleu* le sont dans les deux cas,
- les journées classées dans le groupe *jaune* sur les 90

	Jeu de données de 90 jours (4 classes)	Jeu de données de 15 jours (6 classes)
20171230	/	
20171231	/	
20180104		
20180108		
20180121		
20180122		
20180123		
20180213		
20180218		
20180302		
20180304		
20180314		
20180315		
20180316		
20180331		

FIGURE 10 – Tableau comparatif des résultats obtenus sur les 15 jours issus des pics d'avalanche. (Les deux journées des 30 et 31 décembre 2017 (/) ne sont pas contenues dans le jeu de données de 90 jours.)

jours sont soit *jaune* soit *violet* sur les 15 jours,

- les journées classées dans le groupe *vert* sur les 90 jours sont soit *vert* soit *rouge* sur les 15 jours.

En ce qui concerne les groupes *jaune*, *violet*, *vert*, *bleu* et *gris*, l'ensemble est parfaitement cohérent avec les profils précédemment établis (figure 11). En effet :

- les profils *jaune* et *vert* des 90 jours sont similaires aux profils *jaune*, *violet* et *vert* des 15 jours (seule une nuance sur l'intensité des précipitation apparaît) : le nouveau profil *jaune* présente des valeurs plus élevées que l'ancien (puisque ce groupe ne contient plus que la journée du 04 janvier 2018, journée ayant reçu les plus gros cumuls de précipitations) et le nouveau profil *violet* (constitué des autres journées initialement classées dans le groupe *jaune*) s'inscrit entre les anciens profils *jaune* et *vert*,

- le profil *bleu* des 90 jours correspond au profil *bleu* des 15 jours.

- les deux groupes *gris* établis correspondent aux jours avec peu ou pas de précipitations,

En revanche, le groupe *rouge* est un nouveau groupe apporté par l'étude des journées extraites des pics d'avalanches, donc d'un petit jeu de données. Ce groupe présente en effet un profil nettement différent des autres, tandis que la journée du 08 janvier 2018 (seule représentante de ce groupe *rouge*) est classée avec les journées du groupe *vert* dans le cas de l'étude sur les 90 jours.

Ainsi, le jeu de données de 15 jours permet de faire ressortir les mêmes profils de journées que celui de 90 jours, tout en apportant une nuance supplémentaire sur l'un des 3 profils initialement détectés. Le groupe *violet* ajouté permet en effet de distinguer un troisième groupe de même tendance que les *jaune* et *vert* mais avec une nouvelle différence d'intensité. Le groupe *rouge* émergeant du jeu de données de 15 jours permet ensuite de distinguer une nouvelle tendance (associée à un profil différent). Dans ce cas, travailler sur un tel jeu de données (i.e. de petite dimension, contenant des journées bien ciblées et caractéristiques) permet

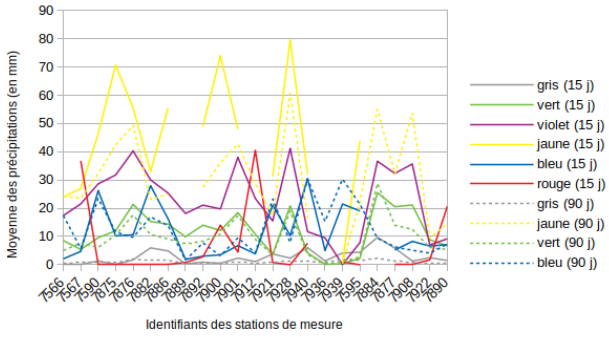


FIGURE 11 – Superposition des profils des 4 classes obtenues sur 90 jours et 6 classes obtenues sur 15 jours (figures 5 et 7)

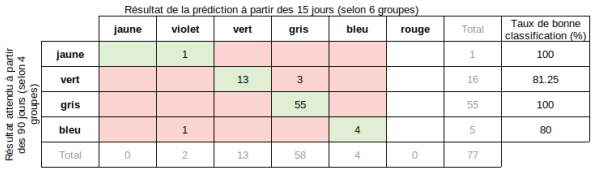
de faire émerger une tendance représentée par peu d'individus (1 seule journée sur les 15, mais aussi 1 seule sur les 90). Cette tendance, dite de "retour d'est", sous-représentée dans le jeu de données de 90 jours, n'est pas détectée dans ce dernier alors qu'elle l'est sur un jeu de données réduit.

5 Classification sur une 16ème journée grâce au jeu de données de 15 jours

Nous avons vu dans la section précédente l'intérêt de travailler sur le jeu de données de 15 jours. En l'utilisant, nous pouvons estimer le groupe auquel appartiendra une 16ème journée. Pour cela nous définissons la journée à étudier selon les 3 composantes explicitées par l'ACP (combinaisons linéaires des 23 variables existantes) et relançons l'algorithme des K-moyennes, initialisé sur les 6 centroïdes des 6 classes obtenues par la CAH mise en œuvre sur les 15 jours du jeu de données (voir section 4.1). De cette manière, nous pouvons ainsi analyser toute nouvelle journée.

Dans cette étude nous avons choisi d'estimer de cette façon le classement de chacune des journées du jeu de données étudié en section 3. Nous pourrons ainsi vérifier si les résultats obtenus sont cohérents ou non. Nous avons donc testé les 77 journées du jeu de données de 90 jours n'appartenant pas au jeu de données de 15 jours.

Résultats. Le tableau présenté en figure 12 donne le résultat de la classification des 77 journées par rapport au résultat attendu (obtenu sur le classement direct des 90 journées, section 3). Pour permettre la comparaison des résultats, il est cependant nécessaire d'établir une correspondance entre les 6 classes (présentant 4 profils différents) obtenues sur 15 jours et les 4 classes (présentant 3 profils différents) obtenus sur 90 jours (figure 11). Notons que la mise en correspondance entre les classes obtenues selon les 2 jeux de données ne peut être réalisée automatiquement (ie. à partir de la position des centroïdes). Leurs valeurs ne sont en effet pas comparables puisqu'elles diffèrent du fait de la réalisation préalable de l'ACP, propre à chaque



	jaune	violet	vert	gris	bleu	rouge	Total	Taux de bonne classification (%)
jaune		1					1	100
vert			13	3			16	81.25
gris				55			55	100
bleu		1			4		5	80
Total	0	2	13	58	4	0	77	

FIGURE 12 – Résultat des classifications des 77 journées. Vert, bonne classification ; Rouge, mauvaise classification ; Blanc, aucune conclusion possible.

jeu de données. Ainsi, au vu des constats exposés en section 4.2 et de la superposition des différents profils obtenus (figure 11), nous pouvons établir que la classification sera bonne lorsqu'une journée attendue dans le groupe *jaune* sera classée *jaune* ou *violet* et lorsqu'une journée attendue dans les groupes *vert*, *bleu* ou *gris* sera obtenue dans le même groupe respectif. Par contre, le profil *rouge* en trait continu n'ayant aucun homologue dans les profils initiaux, aucune conclusion ne pourra être faite sur les jours qui seront classés dans le groupe *rouge*.

Contraindre la méthode à retourner 6 classes sur 90 jours ne permet pas de s'affranchir de cette étape de correspondance entre les classes, la classe rouge n'émergeant pas sur 90 jours, y compris avec un forçage à 6 classes.

Nous obtenons ainsi (figure 12) : 100 % de bonne classification sur les journées attendues dans le groupe *jaune*, 81.25 % sur les journées attendues dans le groupe *vert*, 100 % sur les journées attendues dans le groupe *gris*, et 80 % sur les journées attendues dans le groupe *bleu*.

Dans le cas du groupe *bleu*, la journée dont la classification est mal estimée (groupe *violet* au lieu de *bleu*) peut être due à une valeur erronée dans le jeu de données. La station 7566 a une valeur de 120.1 mm alors que la valeur moyenne sur l'ensemble des autres stations pour ce jour est de 19.3 mm (avec des valeurs allant de 0 à 46.8 mm) et, la veille, aucune donnée n'était communiquée pour cette station ; nous pouvons donc supposer qu'une erreur s'est glissée dans le jeu de données. En traitant cette valeur (potentiellement fausse) comme une valeur manquante (section 2.1) alors ce jour attendu dans le groupe *bleu* est correctement classé. Le taux de bonne classification est alors ramené à 100 % dans ce cas.

De plus, sans ce traitement correctif, le mauvais classement de cette journée engendre aussi un changement de classe pour 2 des 15 journées utilisées pour la classification : la journée du 31 décembre 2017 passe du groupe *violet* au groupe *vert* et la journée du 04 mars 2018 passe de *vert* à *gris*. Toutefois, ces changements de classe ne sont pas un réel problème puisqu'ils correspondent seulement à des décalages entre des groupes correspondant à une même tendance mais présentant des différences d'intensité.

La classification des journées des 11 et 12 mars 2018 (bien que correcte, les classant comme attendu dans le groupe *bleu*) génère un changement de classe pour l'une des 15 journées : le 31 mars 2018 passe alors du groupe *bleu* au groupe *vert*. Ce changement ne correspond pas seulement à un décalage mais à un changement de tendance (les groupes *bleu* et *vert* n'ayant pas les mêmes profils). Ce problème constitue une limite de la méthode pour les jours situés en bordure de leur classe selon la position de la journée ajoutée.

6 Synthèse et discussion

La démarche mise en œuvre dans cette étude (i.e. l'enchaînement ACP - CAH - K-moyennes) permet d'atteindre l'objectif visé *in fine* qui consiste à pouvoir estimer, pour une observation d'avalanche donnée, quelles portions des Alpes françaises, affectées par le même type de conditions de précipitations sont susceptibles d'être sujettes à des avalanches. Cette démarche permet en effet d'estimer à quel type de tendances appartient une journée donnée sachant que la zone principalement affectée par les précipitations, à l'échelle des Alpes françaises, coïncide globalement avec la localisation des avalanches spontanées observées.

Le jeu de données de 90 jours permet de dégager 4 grands groupes de journées : d'abord celles associées à des précipitations très faibles voire nulles, ensuite deux groupes associés à une même tendance météo (de type flux de nord ouest) mais avec deux intensités différentes, enfin un dernier groupe plutôt associé à un flux de sud. Le jeu de données de 15 jours permet quant à lui de dégager 6 groupes. Il fait alors ressortir une nuance supplémentaire concernant le flux de nord ouest en ajoutant un troisième groupe. Le jeu de données issu des pics d'avalanches permet aussi de distinguer un nouveau groupe, lié aux retours d'est. Ce groupe, représenté par un faible nombre de journées (un seul individu sur les 90), n'émergeait pas lors de l'analyse du jeu de données de 90 jours. Le jeu de données de petite dimension permet donc de mieux prendre en compte les tendances sous-représentées dans le jeu de données de grande dimension.

Retrouver, en travaillant sur 15 jours uniquement, un résultat similaire à celui obtenu sur les 90 jours (tout en apportant des informations supplémentaires et davantage de nuances) démontre une certaine robustesse de la méthode.

De plus, les trois zones qui émergent à l'issue de la démarche globale (liées aux trois tendances météo : *bleu* pour le sud, *rouge* pour l'est ainsi que *jaune*, *vert* et *violet* pour le nord et l'ouest) coïncident bien avec les trois groupes de stations que l'on distingue dès l'ACP (figure 4).

L'estimation de la classe à laquelle appartiendrait une journée, obtenue à partir du jeu de données de 15 jours, est très bonne, pouvant atteindre 100 % selon les groupes. Elle atteint cependant sa limite pour les jours situés en bordure de classe qui peuvent, selon la position du jour à étudier, passer d'une classe à une autre. Bien que ceci ne se produise que dans de rares cas, il serait intéressant, pour ces journées, de travailler sur une classification floue [2] ou évidentielle [7], de façon à établir si leur position est typiquement associée à un groupe ou si, au contraire, elles ont une tendance naturelle à se situer entre deux voire plu-

sieurs classes.

Discussion. Nous pouvons tout d'abord soulever la question du nombre de journées nécessaire et optimum pour la constitution de chacun des deux jeux de données de dimensions différentes (90 et 15 journées). De plus, le jeu de données utilisé est affecté par de nombreuses données manquantes (et de potentielles valeurs erronées). Ceci est susceptible de fausser ou affaiblir la véracité des résultats. La figure 8 par exemple est affectée par ce défaut, la station 7936 (située dans le Queyras) n'apparaissant pas en rouge du fait de l'absence de valeur le jour considéré. Ceci affaiblit alors la corrélation entre les étendues affectées en rouge (figure 8) et la localisation des avalanches associées (figure 9). Il serait alors intéressant, dans un premier temps, de retravailler sur un jeu de données exempt de ce type de problème et, dans un second temps, de voir comment mieux approximer les données manquantes.
Ensuite, puisque les journées des 04 et 08 janvier 2018 sont respectivement les seules qui constituent les groupes *jaune* et *rouge*, on peut se demander si ces deux individus n'ont pas une contribution trop importante lors de l'ACP pour la constitution des axes 1 (axe décrivant le mieux le groupe *jaune*) et 3 (axe décrivant le mieux le groupe *rouge*). Ainsi, afin de valider encore davantage les résultats de la classification, d'autres journées, choisies pour leurs caractéristiques sur la base d'une connaissance experte, devront donc être testées par la suite.
De plus, il serait intéressant de réaliser un travail similaire avec d'autres méthodes afin de comparer les résultats. En particulier, la méthode SOM (Self Organizing Map) ou carte auto-organisatrice de Kohonen, qui permet d'une part une réduction non linéaire des variables initiales et d'autre part de tenir compte de la dimension spatiale des variables, pourra être testée sur nos données.
Enfin, une piste d'amélioration serait de construire des jeux de données avec un rythme basé sur les perturbations plutôt que sur un pas quotidien. Ce dernier en effet est susceptible de diviser une perturbation (si celle-ci s'étend sur plusieurs journées) et donc réduire l'intensité qui lui est associée ou, à l'inverse, de concaténer deux tendances de flux différents (si deux tendances successives se sont produites au cours de la même journée du fait d'une évolution rapide du flux). L'objectif serait alors de mieux synchroniser les données aux phénomènes naturels.
Pour aller plus loin, nous supposons que cette démarche peut être transférée quel que soit le massif. La seule étape susceptible de varier réside dans le choix du nombre de composantes à conserver à l'issue de l'ACP et à leur interprétation de façon à assurer l'explicabilité des résultats.
Pour finir, le développement d'un système *temps réel*, capable d'analyser en permanence les données au regard des observations est une réelle perspective d'évolution.

Remerciements

Le projet CIME est soutenu par le programme européen de coopération transfrontalière Interreg France-Suisse 2014-2020 et a bénéficié à ce titre d'une subvention européenne (Fonds Européen de Développement Régional) couvrant 60% du coût total français.

Références

[1] M.J. Casado, M.A. Pastor, F.J. Doblas-Reyes, Links between circulation types and precipitation over Spain, Physics and Chemistry of the Earth 35 (2010), pp. 437-447.

[2] J. Bezdek, R. Ehrlich, W. Full, 1984. FCM - the Fuzzy C-Means clustering-algorithm. Computers & Geosciences 10, 191 - 203. https ://doi.org/10.1016/0098-3004(84)90020-7

[3] HCPC function - RDocumentation [WWW Document], n.d. URL https ://www.rdocumentation.org/packages/FactoMineR /versions/2.4/topics/HCPC (accessed 5.18.21).

[4] M. Huan, R. Lin, S. Uang, T. Xing, A novel approach for precipitation forecast via improved K-nearest neighbor algorithm, Advanced Engineering Informatics 33 (2017), pp. 89-95.

[5] M. Irannezhad, A.K. Ronkanen, S. Kiani, D. Chen, B. Klove, Long-term variability and trends in annual snowfall/total precipitation ratio in Finland and the role of atmospheric circulation patterns, Cold Regions Science and Technology 143 (2017), pp. 23-31.

[6] M. Lemus-Canovas, J. A. Lopez-Bustins, L. Trapero, J. Martin-Vide, Combining circulation weather types and daily precipitation modelling to derive climatic precipitation regions in the Pyrenees, Atmospheric Research 220 (2019), pp. 181-193.

[7] M.-H. Masson, T. Denœux, Algorithme évidentiel des c-moyennes ECM : Evidential c-means algorithm. LFA 2007

[8] Météo-France, `https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=94&id_rubrique=32`

[9] F. Murtagh, P. Legendre, Ward's Hierarchical Agglomerative Clustering Method : Which Algorithms Implement Ward's Criterion ?, *Journal of Classification,* 31, p.274-295, 2014. DOI : 10.1007/s00357-014-9161-z

[10] J.P. Praene, B. Malet-Damour, M.H. Radanielina, L. Fontaine, G. Rivière : GIS-based approach to identify climatic zoning : A hierarchical clustering on principal component analysis, Building and Environment 164 (2019), 106330.

[11] M. B. Richman, I. Adrianto : Classification and regionalization through kernel principal component analysis, Physics and Chemistry of the Earth 35 (2010), pp. 316-328.

[12] S. Tufféry, *Data mining et statistique décisionnelle : L'intelligence des données, 4ème édition,* Editions Technip, Paris, 2012.

[13] J.H. Ward, Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, 1963. https ://doi.org/10.1080/01621459.1963.10500845

[14] B. Zahraie, A. Rooszbahani, SST clustering for winter precipitation prediction in southeast of Iran : comparison between modified K-means and genetic algorithm-based clustering methods. Expert Systems with Application 38 (2011), pp. 5919-5929.