Méthodes de classification des séries temporelles : application à un réseau de pluviomètres

Mohamed Djallel Dilmi

Sorbonne Université

Ecole doctorale science de l'environnement (129)

Laboratoire ATmosphères, Milieux, Observations Spatiales / Equipe SPACE

Méthodes de classification des séries temporelles

Application à un réseau de pluviomètres

Par Mohamed Djallel DILMI

Thèse de doctorat de Data science

Dirigée par Cécile Mallet et Laurent Barthès

Présentée et soutenue publiquement le 18 juin 2019

Devant un jury composé de :

M. RAVETTA François	Professeur	Président
Mme. DORIZZI Bernadette	Professeur	Rapporteur
M. BOUDEVILLAIN Brice	Physicien-adjoint	Rapporteur
M. ROQUET Hervé	Directeur-adjoint	Examinateur
M. HANCZAR Blaise	Professeur	Examinateur
Mme. MALLET Cécile	Maitre de Conférences, HDR	Directrice de thèse
M. BARTHES Laurent	Maitre de Conférences, HDR	Directeur de thèse

Dédicace

Je dédie ce travail …

A mes très chers parents Lakhdar et Souraya

qu'ils trouvent en ce travail l'accomplissement de leurs vœux et l'expression de ma profonde gratitude.

A ma chère femme Fadhila

Que notre vie soit pleine de bonheur et de joie

A mes frères bien-aimés

Chouaib, Younes et Yasser Idris

A qui je souhaite plein de bonheur et de réussite

A tous ceux qui me sont chers amis et collègues

Qu'ils trouvent tous ici l'expression de mon profond respect.

Remerciements

La thèse de doctorat est présentée comme le travail d'une seule personne mais c'est loin d'être vrai…. elle est le fruit d'un travail collectif et la liste des personnes à remercier est très longue :

Je commence par mes encadrants :

Cécile Mallet, la femme qui a cru en moi. Comme elle l'a mentionné lors de la soutenance c'est grâce à elle que je suis au LATMOS, elle avait les mots pour me convaincre, me séduire scientifiquement et me recruter alors que j'étais un étudiant qui suivait le master TRIED car elle manifestait une réelle passion. Pendant la thèse, elle exprimait sa confiance lors des réunions, elle était toujours là à écouter mes propositions, m'encourager et à me diriger. Une phrase qui m'a marqué de sa part : « Djallel, fais confiance à ton intuition, et fais-le, en plus c'est original, c'est bon pour être publié». Deux ans après on était à deux publications. En plus, d'être ma directrice de thèse, elle me forme au métier d'enseignant-chercheur, ses conseils m'ont d'une grande utilité, merci pour tous.

Laurent Barthès, l'homme qui a 40 heures par jour. Je ne trouve pas d'autre explication pour justifier sa grande disponibilité, sa flexibilité scientifique et sa vitesse d'assimilation malgré sa charge d'enseignement. Quand je proposais une idée basée sur une théorie quelconque, ça lui prenait quelques jours pour se familiariser avec et venir discuter la proposition et en débattre. De plus, la thèse est une expérience qui présente des défis sur les deux plan scientifique et psychique (pour nombreux c'est le premier projet long sans garantit de convergence), Laurent a su me coacher tout au long de cette expérience alors un grand merci.

Ensuite, mon jury : les deux rapporteurs Brice Boudevillain et Bernadette Dorizzi,les deux examinateurs : Blaise Hanzcar et Hervé Roquet et le président François Ravetta. Le débat avec vous était très intéressant et vos remarques étaient très constructives.

Je remercie aussi les membres de l'équipe Space et spécialement, Yvon, Nicola, Ruben, Renaud, Meriem, Quitterie, Justine , Constantino, Pierre, Aymeric et Richard et Sylvie du locean pour l'accueil, les conseils et le soutien.

Sans oublier de remercier ma famille, mes parents Lakhdar et Souraya, ma chère épouse Fadhila et mes frères Chouaib, Younes et Yasser ainsi qu'une amie très proche Yamina Bencheikh que je compte comme membre de la famille.

Je remercie aussi mes amis Khalil, Yannick et Beyrem.

Merci à tous ceux qui avaient participé à l'élaboration de ce travail.

Table des matières

Introduction

L'intérêt porté sur la variabilité des précipitations à des échelles infra-journalière voire infra-horaire conduit à s'interroger sur l'utilisation des réseaux pluviométriques existants. En effet, les études concernant l'évolution des précipitations dans le contexte du changement climatique montrent que l'utilisation d'observations à des échelles de plus en plus fines est nécessaire pour progresser dans la compréhension de la relation entre l'augmentation de la température atmosphérique et les changements dans les précipitations. Ces études montrent que les précipitations extrêmes observées à des échelles horaires s'intensifient plus rapidement que les précipitations mesurées à des échelles de temps quotidiennes (Westra et al., 2014). Les analyses climatiques doivent donc pouvoir prendre en compte l'intermittence et la distribution infra-journalière des précipitations. Les événements extrêmes étant rares par définition, il s'avère nécessaire d'analyser de longues périodes d'observations des précipitations échantillonnées à « haute fréquence ». En hydrologie urbaine la nécessité de disposer de statistiques de pluie à des échelles intra-évènements a été mise en évidence dans un article de Berne (Berne et al., 2004) *« les applications hydrologiques pour les bassins versants urbains de l'ordre de 1000 ha nécessitent une résolution temporelle d'environ 5 min et une résolution spatiale d'environ 3 km. Pour les bassins versants urbains de l'ordre de 100 ha, une résolution d'environ 3 min et 2 km est nécessaire».* Dans l'article de Susana Ochoa-Rodriguez et al. (2015) ont quantifié l'impact sur les sorties des modèles hydrologiques de drainage en milieu urbain de la résolution des intensités de pluie mises en entrée. Concernant le suivi des cellules orageuses ces auteurs montrent qu'une résolution inférieure à 5 minutes est nécessaire pour bien saisir la variabilité observée dans les données pluviométriques. Les résolutions spatiales théoriquement requises semblent moins contraignantes, de l'ordre du kilomètre. L'auteur souligne cependant que les ruptures d'échelle révélées par les analyses multifractales suggèrent que les précipitations devraient être mesurées à des échelles inférieures à 500 m pour saisir des structures et des extrêmes qui ne peuvent être extrapolés à partir de mesures à des résolutions plus grossières. Les réseaux de pluviomètres sont également couramment utilisés pour évaluer les produits de pluie basés sur la télédétection spatiale ou terrestre. Dans l'article de Villarini et al. (2008), les auteurs considèrent que « *les réseaux de pluviomètres fournissent des mesures de précipitations avec un haut degré de précision à des emplacements spécifiques* », ils les utilisent pour évaluer les erreurs résultant de lacunes temporelles dans les observations pluviométriques fournies par télédétection. Ces auteurs mentionnent l'impact de la mesure par auget basculant notamment pour les faibles valeurs de précipitation fréquentes dans certaines régions *« Comme indiqué dans la littérature (par exemple, Habib et al., 2001; Ciach, 2003) ces erreurs tendent à diminuer pour les*

taux de précipitations et les temps d'accumulation importants. … il est probable que les résultats pour les corrélations spatiales à court terme (jusqu'à 15 min) peuvent être corrompus par des incertitudes de mesure… De la même manière, il est probable que nos résultats sur la variabilité à petite échelle des précipitations est affectée par ces erreurs instrumentales ».

Les travaux présentés dans ce manuscrit concernent l'étude de la variabilité temporelle et spatiale des précipitations de la région parisienne. Les observations utilisées sont d'une part celles d'un disdromètre et d'autre part les observations du réseau de pluviomètres de Météo-France. Le disdromètre présente l'avantage d'une observation continue mais nous ne disposons que de mesures effectuées sur le site du SIRTA situé au sud de l'Ile de France ce qui ne permet pas d'analyser la variabilité spatiale. La mesure pluviométrique du réseau de pluviomètre de Météo-France est basée sur une technique qui induit une erreur de quantification qui doit être prise en compte pour des analyses à fine résolution. La variabilité des précipitations est caractérisée par l'intermittence d'évènements pluvieux et de périodes sèches mais aussi par la variabilité de l'intensité de la pluie au sein d'une cellule de pluie.

Les travaux présentés s'appuient sur des études menées depuis une dizaine d'année par l'équipe SPACE sur la variabilité des précipitations (De Montera et al., 2009). Dans la thèse de Sébastien Verrier (Verrier, 2011), les propriétés multifractales des précipitations ont été analysées, différents régimes d'invariance d'échelle liés à l'alternance pluie/non pluie ont été mis en évidence. L'analyse sur des périodes ou des espaces continus de précipitation a permis de mettre en évidence le fait que les relations d'échelles spatiales étaient identiques aux relations d'échelles temporelles (Verrier et al., 2010, 2011). Les travaux de thèse de Nawal Akrour (Akrour et al., 2015) qui ont suivi, reposent sur ces constats, ils ont permis de réaliser un simulateur multi-échelles de précipitation. Le simulateur réalisé permet de générer des séries chronologiques ou des cartes de précipitations qui reproduisent les propriétés statistiques à l'échelle des observations utilisées mais également après agrégation à d'autres échelles plus grossières. Ces travaux ont montré que la variabilité de l'alternance pluie/non pluie et la variabilité interne des périodes ou des espaces continus de pluie ne sont pas liées. La modélisation indépendante de ces deux processus permet d'obtenir des simulations qui reproduisent correctement les différents régimes d'invariance d'échelle, alors qu'il est difficile à partir d'une modélisation globale de la variabilité des précipitations à une échelle particulière de reproduire correctement ces différents régimes.

Les études des régimes pluviométriques sont généralement réalisées à partir de moyennes annuelles ou saisonnières. De telles échelles d'agrégation ne permettent pas de distinguer l'évolution de l'alternance pluie/non pluie de celle des périodes ou des espaces

continus de pluie (évènements). Pour prendre en compte les résultats obtenus lors des travaux antérieurs cités ci-dessus, l'approche utilisée pour l'étude de la variabilité temporelle et spatiale des précipitations de la région parisienne, est basée sur la notion d'événement de pluie.

Le manuscrit est structuré en six sections. Le chapitre 1 présente les données sur lesquelles repose cette étude et introduit les approches utilisées pour leur analyse. Le chapitre 2 est consacré à la définition, la description par caractéristiques et la classification des évènements de pluie à partir d'observations réalisées par un disdromètre au pas de temps d'une minute. Au chapitre suivant nous nous sommes intéressés à la possibilité d'étendre les résultats obtenus à des observations réalisées par des pluviomètres à auget. Ces instruments sont en effet beaucoup plus fréquents et les séries chronologiques observées par les pluviomètres sont beaucoup plus longues et plus nombreuses. Le chapitre 4 présente une mesure de dissimilarité entre événements précipitants pour palier à la trop forte sensibilité de la description par caractéristiques au moyen d'observation. Le chapitre 5 est consacré à l'impact du moyen d'observation sur la mesure de dissimilarité et sur l'utilisation de cette dernière pour la classification des évènements de pluie à partir d'observations réalisées par un pluviomètre. Le chapitre 6 est consacré à l'étude de la variabilité spatiale et temporelle des précipitations en Île de France à partir des données d'observations du réseau de pluviomètres à auget mis à notre disposition par Météo-France.

Chapitre 1 : observation des précipitations, grandeurs, instruments et notations

L'objectif de ce chapitre est de présenter les différentes grandeurs caractérisant les précipitations, les instruments utilisés pour les observer, les contraintes relatives à ces mesures et le type de données fourni par ces instruments. Nous présenterons également les différentes approches possibles pour la classification de séries chronologiques de précipitations.

Nous nous intéressons dans ce mémoire uniquement aux précipitations sous forme liquide. Dans ce qui suit le mot précipitations renvoie donc à sa forme liquide.

1.1. Origine de l'observation des précipitations

La civilisation humaine au cours de son évolution s'est intéressée à plusieurs caractéristiques des précipitations en fonction des questions et de l'approche de chaque époque. La question : « pourquoi les populations se sont intéressées à l'observation des précipitations ? », nous aide à comprendre pourquoi telle ou telle grandeur a été choisie pour la mesure des précipitations et pourquoi tel ou tel instrument a été développé. Dans un premier temps nous commençons par présenter brièvement l'histoire et l'origine de l'observation des précipitations.

Depuis l'antiquité, le besoin en eau a conduit les populations à s'installer le long des rives des fleuves comme le tigre et l'Euphrate en Mésopotamie, le Nil en Egypte, l'Indus en Inde et le Fleuve Jaune en Chine. Les préoccupations de l'antiquité étaient le stockage et la distribution de l'eau. Il existait à cette époque un certain niveau de connaissances empiriques qui a permis la construction d'ouvrages hydrauliques comme par exemple le barrage de Marib Yémen 750 ans av. J.-C. Même si les populations avaient remarqué l'existence d'un lien entre les précipitations et les réserves d'eau disponibles, la compréhension du cycle de l'eau n'a pas été une tâche prioritaire de l'antiquité, Lloyd (un psychohistorien, 1974) avance comme théorie que les populations de l'époque adoptaient toujours le miracle comme explication.

La première tentative d'explication et de compréhension du cycle de l'eau qui marque la rupture avec l'explication du « miracle », a été menée par Thalès au 6^{e} siècle av. J.-C., en Grèce. Il a décrit ainsi les précipitations : « Le ciel et la pluie dépendent de l'action de Zeus, qui très anciennement est le maître des nuages ». Cette explication bien que mythique forge les prémisses d'une explication causale et surtout la nécessité de cette compréhension. Après

Thalès, Anaxagore a proposé une explication philosophique de la formation de la grêle, la neige et la pluie. Un siècle plus tard, Aristote propose différentes explications concernant le phénomène de condensation/évaporation dans sa grande encyclopédie du savoir. Mis à part Aristote qui a laissé d'impressionnants travaux sur les processus, le géographe Strabon (64 av. J.-C), a lui aussi étudié les mouvements des eaux et la météorologie. Il a cherché à comprendre ce qui cause les élévations du niveau de l'eau dans les fleuves ; et a souligné par exemple les mécanismes d'alimentations des fleuves par les pluies et par conséquent l'intérêt d'étudier la pluie comme une source d'eau.

On note que les grecs avaient une bonne approche descriptive du cycle de l'eau et principalement des précipitations, mais faute de mesures, aucune progression quantitative n'était possible. Toutefois, ces premières explications des éléments du cycle de l'eau ont été à l'origine de l'intérêt suscité pour l'observation et la compréhension des impacts induits par sa forte variabilité.

1.2. Points de vue macro-physique et microphysique des précipitations

L'impact des précipitations sur les crues des fleuves et les réserves d'eau, souligné par Strabon (64 av. J.-C) a posé implicitement le cumul d'eau comme la grandeur principale caractérisant les précipitations. Ce dernier a souligné qu'il n'était pas le seul à avoir fait ce lien; des écrits religieux mentionnent la mesure du cumul d'eau pour des besoins agricoles en Palestine, à partir du 2ème siècle av J.-C. D'après un manuscrit rédigé en sanscrit le cumul d'eau était mesuré dans plusieurs régions de l'Inde dès le quatrième siècle av. J.-C. D'un autre côté, les travaux d'Aristote et d'Anaxagore relatifs à la compréhension des processus à l'origine des précipitations ont traité de la formation des gouttes et leurs caractéristiques comme grandeurs principales. Ainsi, le premier point de vue (le cumul d'eau) définit un point de vue que nous appellerons par la suite « point de vue macrophysique » car il est observé à des échelles temporelles relativement grande (journée, mois, année ou durée de l'évènement) et s'intéresse plutôt à des grandeurs intégrées tandis que le second point de point s'intéresse plutôt aux échelles temporelles « fines » (minutes, heure) et à des paramètres directement reliés aux caractéristiques des gouttes d'eau (diamètre, vitesse de chute). Ce dernier point de vue définira « l'approche microphysique des précipitations ».

1.3. Mesure du cumul d'eau

La mesure du cumul d'eau des précipitations peut s'effectuer à l'aide d'un pluviomètre. Il s'agit de l'instrument historique permettant de mesurer le cumul de précipitation traversant une petite surface de collecte. La mesure est donc considérée comme ponctuelle. Le cumul d'eau est noté H son unité de mesure est le millimètre et correspond à un volume d'eau recueilli par unité de surface : $1\ mm\ =\ 1\ litre\ d'eau\ /\ m^2$.

1.3.1. Les pluviomètres

De nombreuses versions de pluviomètre ont été conçues et utilisés depuis le 2ème siècle av. J.-C. Les améliorations ont principalement concernées :

1- la précision
2- la méthode de relevé (manuelle ou automatique). Dans ce dernier cas on parle de pluviographe
3- le mécanisme de mesure

La première version (figure 1.1) était un simple collecteur circulaire gradué, délimitant la zone de collecte, la lecture se faisait manuellement. En 1622 Christopher Wren mis au point le premier pluviographe à augets (bien que le terme pluviomètre à augets soit plus couramment employé). Les pluviomètres à augets modernes se composent en plus du collecteur circulaire, d'un entonnoir qui canalise la pluie recueillie vers un mécanisme de mesure lui-même relié à une centrale d'acquisition. Une autre façon de mesurer le volume d'eau est de mesurer son poids, c'est le cas notamment des pluviomètres à pesée qui mesure le cumul d'eau avec une très bonne précision grâce à une balance permettant de mesurer l'écart de poids entre deux instants d'échantillonnage. Cependant, ce système reste relativement cher et par conséquent il est nettement moins répandu que la version dite à augets basculants.

Figure1.1 (a) le premier pluviomètre connu datant 1441.

Le pluviomètre à auget basculant

Le pluviomètre à auget basculant (TBRG : Tipping bucket rain gauges) est l'instrument le plus répandu pour la mesure des précipitations au sol (Fig. 1.2). Il fonctionne à l'aide d'un système à deux augets. Le principe est simple : l'eau recueillie par le système de collecte est dirigée vers un auget. Lorsque ce dernier est plein, il bascule pour se vider et provoque de façon fugitive la fermeture d'un contact électrique. Dans le même temps, l'eau recueillie est dirigée vers le second auget qui va pouvoir ainsi se remplir. Le cycle se répète lorsque le second auget sera plein à son tour. La mesure du volume se résume donc à la détection des fermetures du contact électrique (figure 1.2-b). Un système d'enregistrement permet d'horodater ces fermetures et de calculer des cumuls d'eau pour une période donnée.

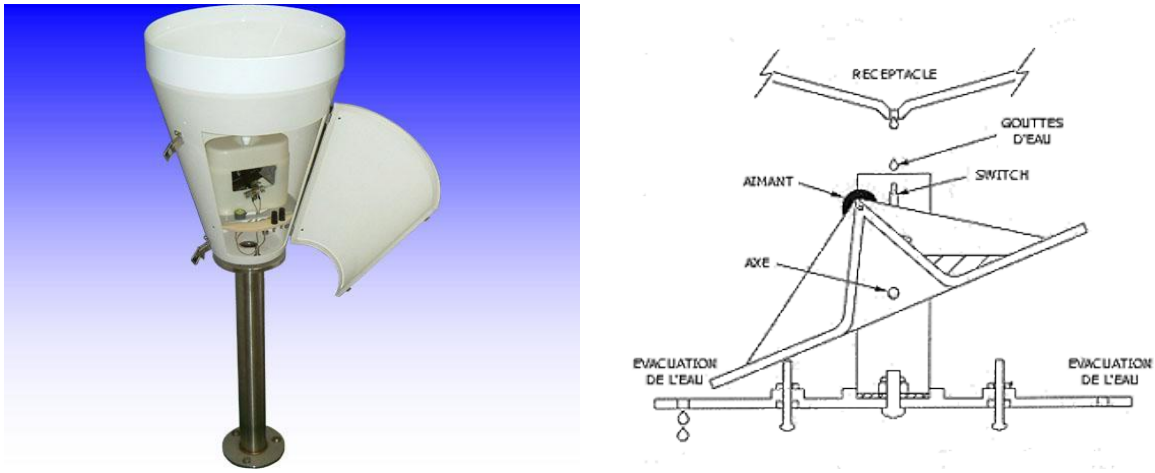


Figure 1.2 (a)à gauche pluviomètre enregistreur de type Précis mécanique. A droite(b) schéma représentatif du principe de fonctionnement d'un pluviomètre à auget basculants

La réponse de ce pluviomètre est discontinue, les incréments de mesures sont équivalentes aux volumes d'auget noté v. En fonction des modèles d'appareils, v prend différentes valeurs, les plus répandues étant 0,1, 0,2 ou 0,5 mm. Deux formats d'enregistrements sont classiques: Le premier est d'enregistrer le cumul d'eau pour un intervalle de temps fixé à l'avance (1 heure par exemple), Sadler et Busscher (1989) associent la popularité de ce format au fait que l'utilisation d'un intervalle de temps fixé produit une série de mesure de longueur prédictible. C'est de ce format qu'il s'agit dans ce mémoire lorsque cela n'est pas précisé. Le deuxième format existait bien avant et a été repris par Costello et Williams (1991), il consiste à enregistrer les temps de basculements produisant ainsi une série de mesure de longueur variable.

Il est à noter que tous les modèles de pluviomètre sont sujets à des erreurs de mesures systématiques. Elles se produisent notamment en présence de fortes précipitations et/ou des vents forts. Les erreurs dépendent du modèle et sont typiquement de l'ordre de 10% pour des précipitations de l'ordre de 200mm/h (Leroy, 2000). Lorsqu'elles sont clairement quantifiées par le constructeur, ces erreurs peuvent dans une certaine mesure être corrigées par logiciels.

Par ailleurs, les pluviomètres à auget basculant nécessitent un entonnoir avec un orifice suffisamment étroit pour guider l'eau vers les augets. Cela rend les pluviomètres sujets à un bouchage éventuel par des déchets végétaux ou animaux. Malgré une grille de protection un entretien régulier reste donc nécessaire. Une revue détaillée sur les sources d'erreurs de mesure est présentée dans le chapitre trois du document éducatif de Météo-France[1].

Les pluviomètres à auget sont cependant extrêmement répandus et forment des réseaux avec un maillage plus ou moins dense. La figure 1.3 montre le réseau des pluviomètres de Météo-France. Certaines zones, notamment l'Ile de France est particulièrement bien couverte.

[1] http://education.meteofrance.fr/observer-et-mesurer/les-precipitations/la-mesure-des-precipitations-par-pluviometre#

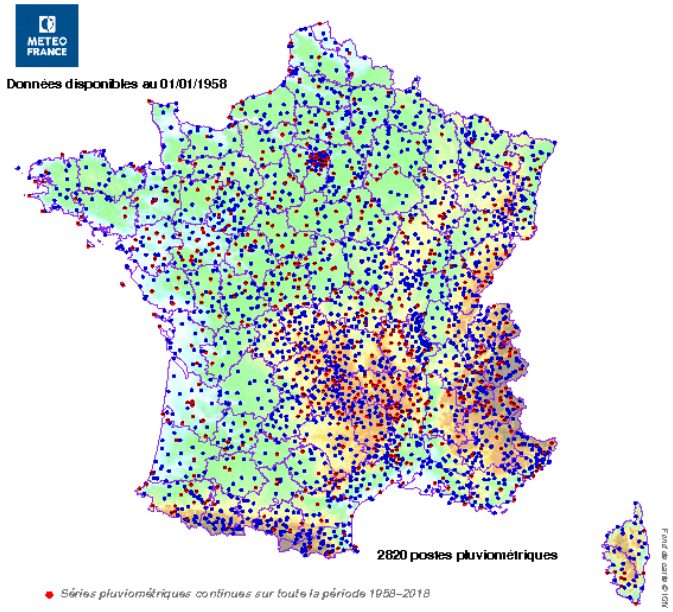


Figure 1.3 le réseau des pluviomètres météo-France déployé en France entre 1958 et 2018 (source : meteo.fr)

Afin d'estimer les précipitations en tout point à partir d'un nombre limité d'observations ponctuelles, différentes méthodes d'interpolation déterministes et/ou probabilistes sont utilisées, la plus utilisée étant l'interpolation optimale par krigeage (Drogue et al. 2010). Cette opération introduit une incertitude supplémentaire d'autant plus importante que la variabilité sera forte, ce qui est particulièrement vrai durant des épisodes de type convectif.

1.4. Résolution temporelle – intensité des précipitations

Les pluviomètres permettent de mesurer la hauteur d'eau H cumulée sur un intervalle de temps donné fixe T. Une série temporelle de durée L est donc subdivisée en $\lambda = \frac{L}{T}$ enregistrements. Notons H_i^λ , le i^{ème} enregistrement correspondant au cumul d'eau sur le i^{ème} sous-intervalle. Selon cette notation la série de mesure de hauteur d'eau est la série temporelle H^λ de longueur λ et peut être écrite sous la forme :

$$H^\lambda = H_1^\lambda, H_2^\lambda, \dots, H_i^\lambda, \dots, H_{\lambda-1}^\lambda, H_\lambda^\lambda \quad i = 1 \dots \lambda \qquad (1.1)$$

Le nombre d'enregistrements λ sera appelé résolution temporelle de la série temporelle H^λ sur l'intervalle de temps étudié L et T est appelé temps d'intégration (ou temps d'agrégation).

A partir de la série de base H^λ (que nous noterons H pour des raisons de simplicité) enregistrée avec un temps d'intégration T, on peut définir une série agrégée H^{λ_m} obtenue pour

un temps d'agrégation supérieur à T : $T_m = mT$ avec m nombre entier compris entre 1 et $\frac{L}{T}$, le i[ème] terme de la série H^{λ_m} s'exprime :

$$H_i^{\lambda_m} = \sum_{j=m(i-1)+1}^{mi} H_j^{\lambda} \quad i = 1 \dots \lambda_m \qquad (1.2)$$

Avec

$$\lambda_m = \frac{L}{T_m} = \frac{L}{mT} = \frac{\lambda}{m} \qquad (1.3)$$

Ainsi $H^{\lambda_m=\lambda}$ correspond à la série initiale tandis que H^1 contient une seule mesure correspondant au cumul d'eau total sur tout l'intervalle L étudié.

Intensité des précipitations

Même si la hauteur d'eau a été la grandeur utilisée durant deux millénaires, il est souvent utile d'exprimer cette grandeur par unité de temps. La grandeur ainsi obtenue est appelée intensité de pluie ou taux précipitant et nous la noterons RR (Rain Rate) dans la suite. En général, on mesure la hauteur d'eau en millimètre et on se réfère à une unité de temps d'une heure. L'unité de RR est le millimètre par heure et correspond à 1 $mm/h = 1$ *litre d'eau / m² / heure*

La série H^{λ} à la résolution temporelle λ est convertible en une série de taux précipitants RR^{λ} à la même résolution temporelle λ, et le ième taux précipitant RR_i^{λ} vaut :

$$RR_i^{\lambda} = \frac{H_i^{\lambda}}{T} \qquad (1.4)$$

Avec T exprimé en heures (ex. si le temps d'agrégation est de 6min alors $T = 0.1$ h).

La série temporelle RR^{λ} de longueur λ, s'écrit sous la forme :

$$RR^{\lambda} = RR_1^{\lambda}, RR_2^{\lambda}, \dots, RR_i^{\lambda}, \dots, RR_{\lambda-1}^{\lambda}, RR_{\lambda}^{\lambda} \qquad i = 1 \dots \lambda \quad (1.5)$$

De même que pour les hauteurs d'eau, on peut définir une série agrégée RR^{λ_m} obtenue pour un temps d'agrégation plus grand $T_m = mT$ avec m entier compris entre 1 et $\frac{L}{T}$, le ième terme de la série RR^{λ_m} vaut:

$$RR_i^{\lambda_m} = \frac{1}{m} \sum_{j=m(i-1)+1}^{mi} RR_j^{\lambda} \quad i = 1 \dots \lambda_m \qquad (1.6)$$

Les météorologues utilisent plus communément les séries de taux précipitants RR alors que les hydrologues utilisent plutôt les séries de hauteur d'eau H. Quelle que soit la résolution λ_m on peut passer de l'une à l'autre par la relation :

$$RR^{\lambda_m} = \frac{H^{\lambda_m}}{mT}$$

Avec mT en heure.

La figure 1.4 illustre une série temporelle de taux précipitants de durée $L = 1409\ min$ et un temps d'agrégation $T = 1\ min$ pour 3 résolutions temporelles $m = 1, 32, 64$.

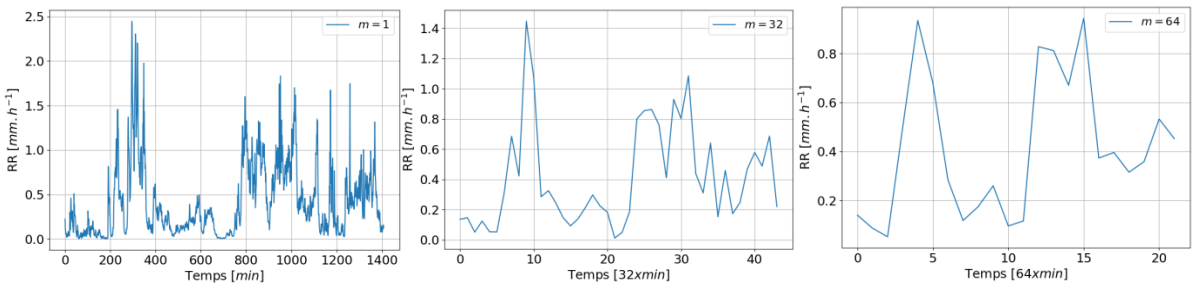


Figure 1.4. Série temporelle d'intensité de pluie du 22 décembre 2012 pour 3 résolutions temporelles $m = 1, 32, 64$

1.5. Mesure de la microphysique des précipitations

Deux caractéristiques principales des gouttes sont d'une part leurs diamètres et leurs vitesses de chute. La connaissance de la microphysique de la pluie revient donc à estimer ces grandeurs ou plus précisément leur distribution spatiotemporelle.

La mesure in situ des diamètres des gouttes et de leurs vitesses de chute s'effectue à l'aide de disdromètres. Le disdromètre que nous avons utilisé est un spectropluviomètre bi-faisceaux (figure 1.5) développé au LATMOS (DBS : Dual-Beam Spectropluviometer). Cet instrument enregistre le temps d'arrivée[2]/détection des gouttes de pluie à la milliseconde près ainsi que leur diamètre D et leur vitesse de chute verticale V.

Le principe de mesure du DBS repose sur l'utilisation de deux faisceaux infrarouges juxtaposés. Lorsqu'une goutte traverse les faisceaux, les signaux reçus par les deux capteurs sont atténués, la variation des signaux dépendent directement du diamètre et de la vitesse de chute de la goutte. L'utilisation d'un double faisceau permet de réduire notablement le taux de fausse

[2] L'instant où la goutte traverse les deux fuseaux

détection : une goutte est considérée comme étant une « vraie » goutte si est seulement si elle est « vue » par les deux faisceaux avec des caractéristiques relativement similaires. De plus, le temps écoulé entre les deux détections permet d'estimer directement la vitesse verticale de la goutte avec une grande précision. Pour une description plus détaillée de l'instrument voir Delahaye et al. (2006).

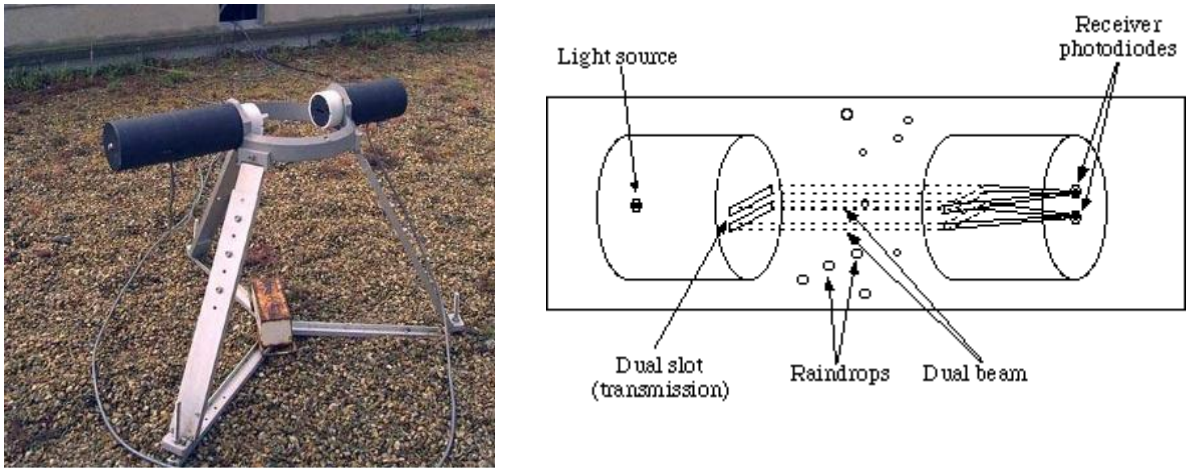


Figure 1.5. *à gauche le spectropluviomètre bi-faisceaux DBS. A droite(b) schéma représentatif du principe de fonctionnement du DBS*

Comme pour le pluviomètre, deux formats d'enregistrements sont utilisés: Le premier consiste à enregistrer les données brutes (i.e. la date d'arrivée de chaque goutte, sa taille et sa vitesse), ce format initial garde toute l'information mais fournit des séries de mesures de longueur quelconque. Le second format –plus couramment utilisé- enregistre une distribution de taille de gouttes $N(D)$ sur un intervalle de temps T fixé à l'avance (typiquement $T = 1\ min$). En pratique, les diamètres sont discrétisés, pour le i^{ème} intervalle de temps, T_i, on décrit la densité de gouttes par sa distribution volumique des diamètres par :

$$N_{T_i}(D_k) \quad k = 1 \dots n \qquad (1.7)$$

Avec $\{D_k\}_{k=1\dots n}$ classes de diamètres, $N_{T_i}(D_k)$ correspond donc au nombre de gouttes enregistrées à la période T_i dont le diamètre est compris entre D_k et $D_{k+1} = D_k + dD$ pour un volume de 1 m^3. La connaissance de $N(D)$ permet le calcul de diverses grandeurs intégrées comme le diamètre volumique moyen qui sera introduit au prochain chapitre.

Même si la fonction principale du DBS est l'étude de la microphysique de la pluie et non la mesure du cumul d'eau ou du taux précipitant[3], il est facile d'estimer le taux précipitant :

$$RR \propto \int N(D)V(D)D^3 dD \quad (1.8)$$

L'intégration sur un intervalle de temps T_i, des volumes des gouttes recueillies permet d'estimer le taux précipitant RR_i^{λ} sur cet intervalle de temps :

$$RR_i^{\lambda} = \frac{600\pi}{ST} \sum_{\substack{D\ collectées \\ dans\ l'intervalle T_i}} D^3 \quad (1.9)$$

Avec $S = 100\ cm^2$: surface de collecte du DBS, exprimée en mm^2 dans l'équation (1.9) ,
$D[mm]$: diamètre de la goutte, T exprimé en heures et RR_i^{λ} en $mm.h^{-1}$.

1.6. Variabilité spatio temporelle – notion d'événement de pluie

La conscience humaine de la dimension temporelle des précipitations a fait un long trajet en histoire : d'abord la nature épisodique « il pleut/il ne pleut pas », puis une différence annuelle (les précipitations peuvent différer d'une année à une autre, la saisonnalité....) En voyageant, l'homme a eu conscience de la dimension spatiale de cette variabilité : s'il pleut, il ne pleut pas partout (on souligne alors la question : « où s'arrête la pluie ? »).

L'historien Alain Corbin (Corbin, 2013) a exploré la sensibilité de l'homme aux phénomènes météorologiques depuis l'antiquité. Il a parlé de l'influence des précipitations ou du soleil sur les activités des populations ou simplement sur l'humeur des personnes repérable en particulier à travers les correspondances et les textes anciens. Ce sont les expressions du langage courant telles que: « Après la pluie, le beau temps », «il pleut, on part quand ça s'arrête »… qui ont répertorié la conscience de l'homme ancien de la nature épisodique et d'états successif (pluvieux, sec) des précipitations à l'échelle de l'observation. Ce changement d'état a motivé P. S. Eagleson (1970) à utiliser le terme d'« événement » pour parler des précipitations.

Un évènement (de pluie) peut être défini qualitativement comme une continuité spatiotemporelle de la chute des précipitations à l'échelle de la perception humaine.

[3] À la base les spectromètres ont été développés pour étudier la distribution des gouttes pour l'étalonnage des radars car la relation $Z - R$ dépend de $N(D)$

Cette perception/définition de la notion d'événement est difficilement transposable pour une étude rigoureuse des précipitations car elle implique plusieurs dimensions (spatiale, temporelle). En effet, selon l'échelle et la grandeur mesurée, la continuité perçue par l'humain est détectée ou perdue.

La nature alternative des précipitations entre état sec et pluvieux permet en un lieu donné de définir la variable indicatrice P_0 des événements de précipitation, définie par :

$$P_0(précipitations) = 1_{précipitations} = \begin{cases} 1 & quand\ il\ pleut \\ 0 & quand\ il\ ne\ pleut\ pas \end{cases} \quad (1.10)$$

La représentation temporelle la variable P_0 définit une variable aléatoire X_t^P à deux états 0 et 1 dont une réalisation est représentée à fig. 1.6.

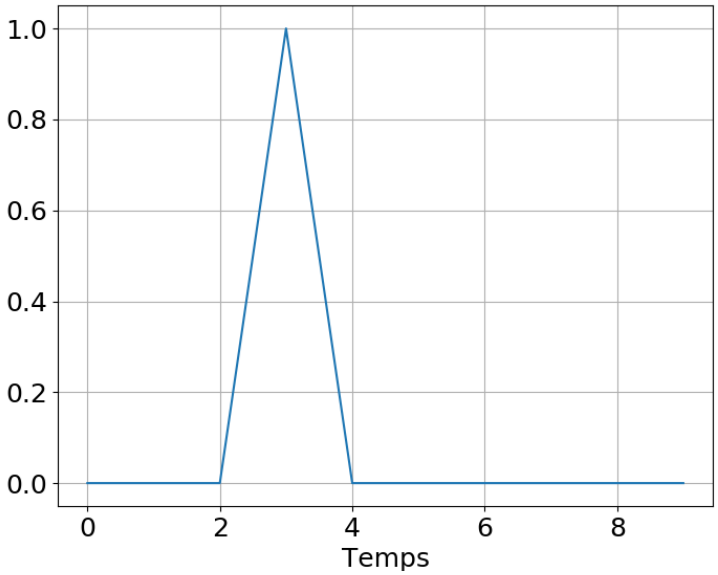


Figure 1.6 série temporelle qui représente une réalisation de la variable P_0 (support des précipitations)

La série temporelle de mesures des précipitations permet de définir une variable aléatoire X_t^d également à deux états 0 et 1 où l'état 1 définit une détection de pluie et l'état zéro une non-détection. En combinant l'information des deux variables aléatoires X_t^P et X_t^d , quatre cas sont possibles :

- cas 1 : $X_t^P = 1$ et $X_t^d = 1$ il pleut et l'instrument détecte la présence de pluie ;
- cas 2 : $X_t^P = 0$ et $X_t^d = 0$ il ne pleut pas et l'instrument ne détecte pas de pluie ;
- cas 3 : $X_t^P = 0$ et $X_t^d = 1$ il ne pleut pas et l'instrument détecte de la pluie ;
- cas 4 : $X_t^P = 1$ et $X_t^d = 0$ il pleut et l'instrument ne détecte pas de pluie.

Les deux premières situations représentent des cas de bonne détection. Le troisième cas représente le cas de faux positifs, i.e. des détections intempestives dont les causes peuvent être multiples (poussière, insectes, opération de maintenance, …). Le cas 4 représente celui des faux négatifs : il pleut mais l'instrument ne détecte pas de pluie. Plusieurs raisons peuvent engendrer ce cas, par exemple en cas d'obturation de l'embout du collecteur d'un pluviomètre, ou en présence de gouttes dont le diamètre est inférieur au seuil de sensibilité du disdromètre, ou tout simplement lorsque l'appareil est hors service !

1.6.1. Durée minimum inter-événement

Nous nous appuyons sur la définition du temps inter évènement minimum (minimum inter event : MIT) attribué à Coutinho et al. (2014) et défini de la façon suivante : chaque valeur de taux de pluie non nulle estimée avec un temps d'intégration d'une minute est affectée à un événement de pluie donné, c'est-à-dire à l'événement en cours ou à l'événement ultérieur considéré comme indépendant et «nouveau». Le MIT peut également être défini comme la durée d'une période sèche à l'issue de laquelle la prochaine occurrence de pluie non nulle marque le début d'un nouvel événement. Pour les périodes sèches inférieures au MIT, les taux de pluie des deux côtés de cette période sont considérés comme appartenant au même «événement composite». Divers auteurs ont proposé différentes valeurs du MIT « garantissant » l'indépendance des événements. Llasat (2001) a noté que: «La définition d'un épisode est assez subjective. Dans ce cas, il a été jugé possible de distinguer deux épisodes différents, lorsque le temps qui s'écoule entre eux sans précipitations dépasse 1 heure, ce qui garantit que les deux épisodes proviennent de «nuages» différents. Bocquillon et Moussa (2014) ont écrit: «les observations de pluie continue sur moins de trente minutes ne représentent que 5% de toutes les périodes de pluie. Le seuil représentatif de la discrétisation des données est de 30 minutes à une heure ».

Dunkerley (2008 a, b) a effectué une analyse du temps inter-événements (IET) afin de vérifier l'influence de cette variable sur la définition des événements pluvieux et son influence sur le taux pluviométrique moyen. Comme souligné dans cette étude, lors de la détermination d'une valeur pour le MIT, il est essentiel de trouver un compromis approprié entre l'indépendance des événements de pluie et la variabilité intra-événement des taux de pluie. Le choix du MIT a donc un impact direct sur les caractéristiques macrophysiques. D'autres chercheurs ont proposé d'utiliser des valeurs de MIT de 20 minutes, 1 heure ou même 1 jour (voir Dunkerley, 2008a pour une liste détaillée).

Dans la présente étude, nous avons fixé le MIT à 30 minutes. Ceci est en accord avec la valeur utilisée par Coutinho et al. (2014), Haile et al. (2011), Dunkerley (2008a, b), Balme et al. (2006) et Cosgrove et Garstang (1995).

Il est à noter qu'il existe d'autres critères pour définir un évènement, ainsi Driscoll et al. (1989) définissent le volume d'eau minimal requis pour détecter un événement de précipitations (Minimum Event Volume, MEV) avec une valeur typique $MEV = 1mm$.

1.6.2. Données pluviométriques utilisées

L'étude présentée dans ce manuscrit repose sur les séries temporelles de pluie décrites dans le tableau ci-après. Les séries mesurées par le DBS sont utilisées principalement aux chapitres 2 et 3. Dans les chapitres 3 à 5, les travaux relatifs à la sensibilité au type d'instrument, des séries temporelles de « pseudo-pluviomètres » ont été simulées à partir de séries temporelles issues du disdromètre. Les deux séries mesurées par des pluviomètres sur le site instrumental de l'IPSL (Pluvio_Z1, Pluvio_Z2) sont essentiellement utilisées pour valider les séries temporelles de « pseudo-pluviomètres ». Nous disposons d'un nombre important de séries temporelles mesurées par des pluviomètres à auget qui seront utilisées dans le chapitre 6 consacré à la variabilité spatiale et temporelle des précipitations en île de France. Le tableau 1.1 présente les disponibilités des différentes données.

Nom	Début	fin	disponibilités des données [% de temps]	occurrence de pluie [% de temps]		%de temps de non pluie
Disdromètre	2008	2015	86	$T = 1min$	4,28	95,72
Pluvio_Z1 ($v = 0.1mm$)	18/04/2005	02/09/2015	97,56	$T = 1min$	0,91	99,09
Pluvio_Z2 ($v = 0.2mm$)	19/07/2002	01/06/2016	98,52		0,56	99,44
Stations Météo-France	01/01/2006	31/12/2015	97,23	$T = 6\ min$	2,8	97,2

Tableau 1.1 Données pluviométriques utilisées

1.7. Analyser la variabilité des précipitations

L'étude de la variabilité des précipitations implique l'inter-comparaison de plusieurs périodes du temps, plusieurs zones spatiales ou les deux en même temps. Par conséquent, on est amené à comparer :

1- Des séries temporelles mesurées sur des périodes différentes dans le cas temporel.
2- Des séries temporelles mesurées à des endroits différents (stations différentes).
3- Des séries temporelles mesurées à différents endroits sur des différentes périodes.

La classification non supervisée (le clustering) présente une solution pour réduire la dimensionnalité des données sans perdre leur variabilité en les regroupant par similitude (similarité) quand il n'y a aucune connaissance à priori des classes. Dans notre situation, la classification **non supervisée (abrégée classification dans ce manuscrit)** des évènements de précipitations est donc un moyen pour découvrir des motifs intéressants dans les séries temporelles.

Les classes formées permettent de caractériser des situations types. Elles permettent aussi d'inférer des paramètres non observables en les associant à chacune des classes.

En dépit du type de méthode de classification (K-means, cartes auto-organisatrices SOM, classification ascendante hiérarchique CAH,… ou autres) employée, la classification des séries temporelles a été utilisée dans divers domaines scientifiques adoptant des approches différentes en fonction de la communauté. Saeed Aghabozorgi et al. (2015) présente une revue générale de ces approches communément utilisées.

1.7.1. Différentes approches de classification

En général, il existe trois différentes approches pour classer les séries temporelles, à savoir, l'approche basée sur **les caractéristiques**, l'approche basée sur **les formes** et l'approche basée sur **les modèles**. Le tableau 1.2 montre une typologie de ces approches.

Dans l'approche basée sur **les caractéristiques**, les séries temporelles sont converties en un vecteur de caractéristiques de dimension inférieure. Un algorithme de classification classique est appliqué aux vecteurs de caractéristiques extraits. Généralement, dans cette approche, le vecteur de caractéristiques est de longueur fixe et la distance euclidienne est la plus souvent utilisée. Les caractéristiques utilisées pour décrire les séries temporelles sont le plus souvent définies par les experts du domaine.

Dans l'approche basée sur les formes, les formes de deux séries sont appariées au mieux par un étirement et une contraction non linéaires des axes temporaux. Cette approche a également été qualifiée d'approche basée sur les données brutes car elle fonctionne souvent directement avec les séries temporelles dans l'état brut. Les algorithmes basés sur les formes utilisent les méthodes de classification classiques (K-means, cartes auto-organisatrices SOM ou classification ascendante hiérarchique CAH …), compatibles avec les données statiques

(représentations vectorielles de longueurs fixes), tandis que leur mesure de distance / similarité est remplacée par une méthode appropriée pour les séries temporelles.

Dans les méthodes basées sur des modèles, une série temporelle est transformée en paramètres d'un modèle (un modèle paramétrique pour chaque série temporelle), puis une distance de modèle appropriée et un algorithme de classification (en général, des algorithmes de classification classiques) sont choisis et appliqués aux paramètres extraits du modèle.

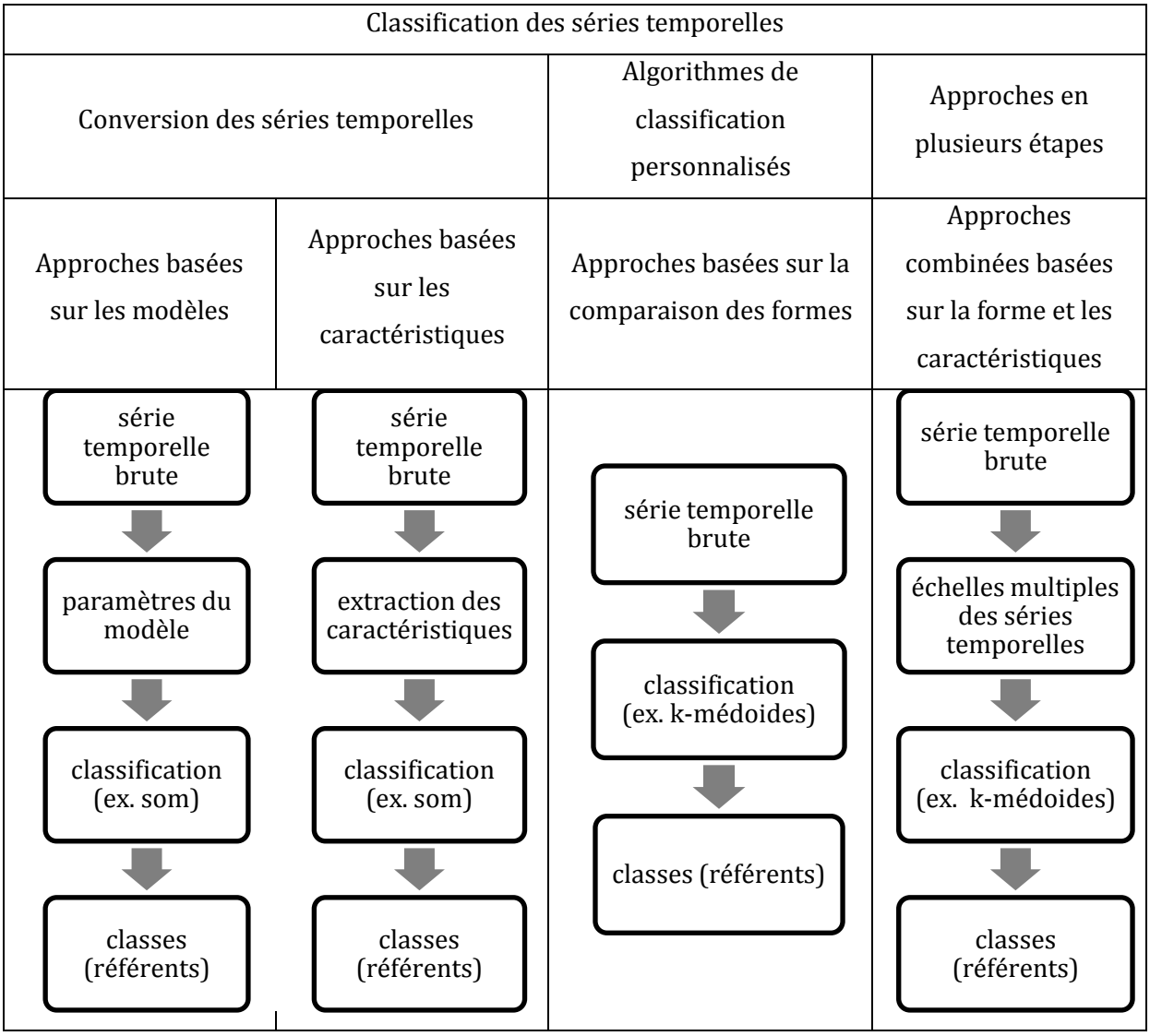


Tableau 1.2. Les approches de classification non supervisée des séries temporelles (Aghabozorgi et al. 2015)

L'approche la plus commune dans le domaine de la géophysique est l'approche basée sur les caractéristiques (Coutinho et al. 2014, Dunkerley 2008a). Dans ce cas, le choix de la bonne caractérisation est primordial, c'est un élément clé sur les performances et l'efficacité de la classification résultante. La difficulté majeure est donc de choisir la bonne caractérisation.

Cette tâche est loin d'être triviale car elle dépend de nombreux paramètres, eux même dépendants de l'instrument d'observation utilisé. Ainsi, comme nous le verrons par la suite, observer la pluie avec un pluviomètre au pas de temps d'une minute, six minutes ou une heure ne fournit pas la même quantité d'information et modifie quelque peu la statistique des paramètres utilisés. Concernant les autres approches, il est à noter une quasi absence dans la littérature de l'utilisation de l'approche basée sur la forme.

1.8. Conclusion

Les données mesurées par des pluviomètres à auget sont les seules qui sont suffisamment longues pour étudier la variabilité temporelle et suffisamment nombreuses pour étudier la variabilité spatiale à l'échelle de l'île de France. Le principe de mesure des pluviomètres implique une quantification plus ou moins importante suivant le volume de l'auget. **Dans le contexte de l'observation de l'évolution du climat, l'intérêt actuel de la communauté scientifique pour la variabilité des précipitations à des échelles infra-journalière voire infra-horaire conduit à s'interroger sur la possibilité d'utiliser les données issues des réseaux de pluviomètres**. Il s'agit de développer des méthodes adéquates et de quantifier l'impact éventuel du volume d'auget et du temps d'agrégation sur les résultats obtenus.

Pour cela, le site instrumental de l'IPSL dispose de mesures co-localisées de plusieurs types d'instrument (notamment le DBS au pas de temps $T = 1min$), ce qui nous donne l'opportunité d'évaluer les différentes méthodes proposées pour différents volumes d'augets et différentes résolutions temporelles.

Chapitre 2 : description parcimonieuse par caractéristiques des évènements de pluie

2.1. Introduction

En raison de la nature turbulente des processus en jeu dans l'atmosphère terrestre nombres de séries temporelles météorologiques (température, humidité …) ont la particularité de présenter des propriétés d'invariance d'échelle. Les caractéristiques d'une série temporelle dépendent du pas d'intégration utilisé, et elles obéissent à des relations d'échelle. Pour plus d'informations sur l'analyse et la modélisation des relations d'échelle voir les travaux de (Schertzer et Lovejoy, 1987 et Verrier et al., 2011). Les séries temporelles de précipitation possèdent également ce type de caractéristiques, mais ont une spécificité supplémentaire liée à l'intermittence du processus de pluie. Les travaux menés ces dernières années notamment au LATMOS ont mis en évidence l'impact de l'intermittence sur les propriétés d'invariance d'échelle des précipitations (Verrier et al., 2011) et l'intérêt pour modéliser les séries temporelles de précipitation de séparer la modélisation du support de la pluie à celle des « évènements de pluie » (Akrour et al., 2015). Ces travaux ont montré que la séparation événement/support n'est vraiment pertinente que si l'on dispose de données à très fine résolution (pas de temps T = 1 min).

Partant de ce constat, l'étude qui fait l'objet de l'article ci-dessous, a cherché à réaliser non pas la comparaison ou la classification de séries temporelles mais d'« évènements de pluie » observés avec un pas de temps $T = 1\ min$. Comme indiqué au chapitre précédent, nous avons retenu comme définition d'une série temporelle d'un événement de pluie une série de taux précipitant contenant des périodes sans pluie de durée inférieure à 30 min.

Il existe cependant peu de méthodes pour comparer entre elles des séries temporelles de longueurs variables. L'utilisation des méthodes d'analyse multidimensionnelles des données nous a semblé être une approche pertinente. Ces méthodes nécessitent de définir préalablement des descripteurs pertinents pour un événement de pluie. Si on se réfère au tableau 1.2 du chapitre 1 la méthodologie développée dans ce chapitre correspond donc à une approche par caractéristiques (feature-based approach). L'objectif est de réaliser une typologie des événements de pluie observés au pas de temps 1 minute avec une approche par caractéristiques.

Après avoir fait le constat qu'il existe dans la littérature un nombre très important de descripteurs pour les évènements de pluie, une analyse en composantes principales, qui est un

des moyens couramment utilisé en analyse multidimensionnelles des données pour synthétiser l'information à un nombre réduit de facteurs, a été réalisée. Les 5 premiers axes principaux contiennent 90 % de l'information et montrent qu'il est possible de décrire les évènements de pluie à partir d'un nombre réduit de facteurs. Cependant les facteurs obtenus par l'ACP présentent deux défauts majeurs pour notre étude. D'une part les facteurs issus de l'ACP ne sont pas des caractéristiques physiques facilement interprétables des évènements de pluie, et d'autre part la réduction opérée n'utilise que les relations linéaires entre les 23 variables initiales alors que des relations non linéaires existent également entre certaines d'entre elles.

Une méthodologie a donc été proposée (cf. section 3 de l'article ci-dessous), basée sur l'utilisation combinée d'un algorithme génétique (GA) (Holland, 1975) et de cartes auto-organisatrices de Kohonen (SOM) (Kohonen, 1982) afin d'établir une caractérisation parcimonieuse des événements de pluie, en définissant un ensemble minimal de variables choisies parmi les 23 variables initiales. Les cartes auto-organisatrices de Kohonen constituent une méthode neuronale d'analyse multidimensionnelle des données qui permet de visualiser sur une carte de dimension 2 un ensemble de données initialement en grande dimension (23 dans notre cas) tout en conservant, de manière non supervisée, la plupart des informations contenues dans la topologie spatiale initiale (Kohonen, 2001). Les algorithmes génétiques sont utilisés ici pour leur capacité à sélectionner un sous-groupe de variables optimales sans avoir à tester toutes les combinaisons possibles. Ainsi par exemple en supposant un sous-groupe composé de 5 à 10 variables, il faudrait apprendre environ 2,8 millions cartes de Kohonen si on souhaitait tester toutes les combinaisons.

Cette analyse a permis de déterminer un jeu optimal de 5 variables qui sont : la durée de l'événement (D_e), l'écart type des taux precipitants (σ_R), le taux precipitant maximal (R_{max}), la variation absolue du taux précipitant (P_{C1})[4], et la hauteur d'eau cumulée sur l'évènement (R_d). Cette première étape permet de remplacer l'analyse d'une série temporelle de longueur quelconque (dans notre cas 6 ans à pas de temps $T = 1min$) par celle des N évènements (545 évènements observés dans notre cas) de pluie qui la constitue, chaque évènement étant décrit par ces 5 variables.

Nous montrons ensuite les résultats de la classification ascendante hiérarchique appliquée à la partition des 64 neurones de la carte de Kohonen optimale obtenue par notre algorithme[5]. Dans un premier temps, l'objectif a été de vérifier que l'analyse basée sur une approche statistique permet de retrouver les types d'évènements tels que classiquement défini par les experts. En effet, la partition en deux classes permet effectivement de retrouver (sans

[4] Voir tableau 2.2
[5] Plus de détails dans l'article

aucune introduction d'information à priori) la distinction stratiforme/convective communément utilisée. Cependant, il est possible d'affiner cette classification, ainsi d'après la classification hiérarchique, des classes d'évènements plus homogènes peuvent être obtenues en sub-divisant la classe stratiforme en 2 sous-groupes et celle des convectifs en 3 sous-groupes.

La dernière partie de l'article s'attache à montrer que les 5 variables retenues contiennent bien, de manière sous-jacente, les informations caractéristiques de la physique des processus en jeu. Il s'agit de s'assurer que la classification, réalisée à partir de la similitude des 5 variables retenues, permet de constituer des groupes d'évènements issus de processus physiques similaires. La série temporelle de pluie utilisée ayant été obtenue par un disdromètre, il est possible de déterminer les propriétés microphysiques de chaque événement de pluie. On montre que les classes d'évènements de pluie réalisées, à partir des 5 variables macro-physiques sélectionnées, ont bien des propriétés microphysiques homogènes.

Data-driven clustering of rain events: microphysics information derived from macro-scale observations

Dilmi M. D., Mallet C., Barthès L., Chazottes A. Atmospheric Measurement Techniques, European Geosciences Union, 2017, **10** (4), pp.1557-1574.

Atmos. Meas. Tech., 10, 1–18, 2017
www.atmos-meas-tech.net/10/1/2017/
doi:10.5194/amt-10-1-2017

Atmospheric Measurement Techniques
Open Access
EGU

2.2. Article: Data-driven clustering of rain events: microphysics information derived from macro scale observations

Mohamed Djallel Dilmi, Cécile Mallet, Laurent Barthes, Aymeric Chazottes

LATMOS-CNRS/ UVSQ/ UPSay, 11 boulevard d'Alembert, 78280 Guyancourt, France

Correspondence to: Laurent Barthes (laurent.barthes@latmos.ipsl.fr)

Received: 25 November 2016 – Discussion started: 29 November 2016
Revised: 21 March 2017 – Accepted: 29 March 2017 – Published: 25 April 2017

Abstract. Rain time series records are generally studied using rainfall rate or accumulation parameters, which are estimated for a fixed duration (typically 1 min, 1 hour or 1 day). In this study we use the concept of "rain events". The aim of the first part of this paper is to establish a parsimonious characterisation of rain events, using a minimal set of variables selected among those normally used for the characterization of these events. A methodology is proposed, based on the combined use of a Genetic Algorithm (GA) and Self Organising Maps (SOM). It can be advantageous to use a SOM, since it allows a high dimensional data space to be mapped onto a two-dimensional space while preserving, in an unsupervised manner, most of the information contained in the initial space topology. The 2D maps obtained in this way allow the relationships between variables to be determined and redundant variables to be removed, thus leading to a minimal subset of variables. We verify that such 2D maps make it possible to determine the characteristics of all events, on the basis of only five features (the event duration, the peak rain rate, the rain event depth, the standard deviation of the rain rate event, and the absolute rain rate variation of the order of 0.5). From this minimal subset of variables, hierarchical cluster analyses were carried out. We show that clustering into two classes allows the conventional convective and stratiform classes to be determined, whereas classification into five classes allows this convective / stratiform classification to be further refined. Finally, our study made it possible to reveal the presence of some specific relationships between these five classes and the microphysics of their associated rain events.

1. Introduction

The analysis of "precipitation events" or "rain events" can be used to obtain information concerning the characteristics of precipitation at a particular location, and for a specific application. This is a convenient way to summarize precipitation time series in the form of a small number of characteristics that make sense for particular applications.
The concept of a precipitation event is not new, and has been used for many years (Eagleson, 1970; Brown et al., 1984). A wide variety of definitions, varying according to the context of each study, has been reported in the literature (Larsen and Teves, 2015). Moreover, when a rain rate time series (generally based on rain gauge records) is broken down into individual rainfall events, a wide variety of their characteristics, such as average rainfall rate, rain event duration and rainfall event distribution (known as hydrological information), can be computed for each event. Our analysis of the literature has led to the identification of seventeen features used to characterize rainfall, which makes it quite difficult to compare different studies. The first goal of the present study is to select a reduced set of features characterizing rainfall events, through the use of a data-driven approach, without taking *a priori* knowledge of the field of application into account, thereby characterizing rainfall events in the most parsimonious and efficient manner.
The second goal is to assess, without using any *a priori* criteria, whether the rain events are still correctly clustered by the most relevant observed features. Indeed, atmospheric process specialists distinguish between stratiform and

convective events, arguing that the physical processes involved in their evolution are different. The goal here is to check that a small sample of variables, derived from spot measurements to describe rain events, can allow this distinction to be made, and ultimately be used to refine it. Hydrological (hereafter referred to as "macrophysical") information makes use of rain gauge measurements to characterize rain events. This information is defined in order to characterize the features of global events, but not to provide any information concerning the raindrop microphysics of the event. Nevertheless, in many applications such as remote sensing, knowledge of the microphysics is essential. One key parameter in remote sensing is the raindrop size distribution, noted as N(D), which is defined by the number of raindrops per unit volume and per unit raindrop diameter (D). Information related to the raindrop size distribution is often derived from its proxies, as explained in section 5. Such features are not currently accessible through rain-gauge measurements, which provide macrophysical information only. However, more expensive devices referred to as disdrometers can provide both hydrological and microphysical information. There are currently several tens of thousands of rain gauges operated throughout the world, in locations equipped with a far smaller number (if any) of disdrometers. As described later in this paper, it is possible to retrieve some microphysical information from the hydrological data. As a consequence, rain gauge data could provide valuable information in microphysics studies, through the use of a statistical approach to indirectly infer the missing microphysics information. In the following, the terms "macrophysical" or "hydrological" information are associated with characteristics related to rain rates or rain accumulation, whereas the term "microphysical" is associated with the characteristics of the raindrop size distribution.

In the present study, we use a data-driven approach to study the relationships between different rain properties. As disdrometers provide drop size distributions, they allow one-minute (or shorter) rain rates to be estimated, and in the present study these can be used to derive the hydrological information of interest, which is coherent with the data that would be provided by standard rain gauges. Through the combined interpretation of microphysical and hydrological information, we are also able to analyse the microphysical properties of the rain-event clusters provided by our algorithm. This makes it possible to retrieve (unobservable) microphysical information from rain gauge measurements.

From a single rain-rate times series, observed with a one minute time resolution, we seek to answer the following questions:

- Among the large number of hydrological variables described in the literature, which are the most significant?
- Does the resulting description of rain events allow different types of rain event to be discriminated?
- What (unobserved) microphysical properties of an event, or type of rain event, can be inferred from its macrophysical description?

Our paper is structured as follows. Section 2 presents the data used in our study, and lists various hydrological parameters that are commonly found in the literature. Seventeen macrophysical variables are identified, requiring appropriate normalisation. Section 3 presents our methodology, which is based on the use of a genetic algorithm (GA) implementing a self-organizing map (SOM also referred to as a topological map). This unsupervised approach is used to select a small subset of variables from the 17 identified variables, allowing a parsimonious characterisation of rainfall events to be applied. An exploratory statistical analysis of rainfall events is provided. In section 4, the rainfall events are grouped in clusters, and are divided into two classes. It is then shown that this grouping of the dataset corresponds to the standard convective / stratiform classification. We then propose a five-subclass classification, which corresponds to a refinement of the two initial groups. In section 5 we include some additional microphysical features of rainfall events, allowing the microphysical properties of the five previously defined event classes to be studied. Our conclusions are presented in section 6.

2. The disdrometer datasets - data processing methodology

This research relies on the analysis of raindrop measurements obtained with a Dual-Beam Spectropluviometer (DBS) disdrometer, first described by Delahaye et al. (2006). This instrument allows the arrival time, diameter and fall velocity of incoming drops to be recorded. As the capture area of the sensor is 100 cm^2 its observations can be considered, in spatial terms, to be "point-like". In the present study the integration time T_{int} was set to one minute, and the raindrop measurements were used to estimate the corresponding one-minute rain-rate time series $RR_t(t)$. In order to eliminate false raindrop detections that could be generated by dust or insects, a threshold T_0=0.1 $mm.h^{-1}$ was applied. Rain rates lower than T_0 are thus set to zero. This conventional threshold is also chosen to ensure coherency with previous studies (Verrier et al., 2013; Llasat et al., 2001). In the present study, we worked with two datasets recorded during the period between July 2008 and July 2014, at the "Site Instrumental de Recherche par Télédétection Atmosphérique" (SIRTA[6]) in Palaiseau, France.

2.1 Rain event definition

In everyday life, it is common knowledge that rain starts at a certain moment, and stops some time later. However, due to its discreet nature, rain (which generally consists in a very large number of raindrops) is not an easy concept to define. Indeed, the exact definition of a rain event will depend on the sensor's characteristics (specific surface capture, detection threshold, instrumental noise), as well as the spatial or temporal resolution chosen for the study. This definition may also depend on the purpose of the study, and thus on the scientific community behind it. There is thus a wide range of criteria used to break down precipitation

[6] http://sirta-dev.ipsl.jussieu.fr/joomla/index.php/85-article-sans-categorie/71-sirta-home-page

records into rain events. For this reason, it is important to define and apply an unambiguous definition of a "rain event".

In this study, the pattern produced by the one-minute rain rate time series $RR_t(t)$ can be simplified by grouping non-null rain rates into a set of separate "primitive events" (Brown et al., 1985). On the basis of an assigned Minimum Inter-event Time (MIT) (Coutinho et al., 2014) each rain rate value, corresponding to a specific one-minute period of observation, is assigned to a given rainfall event, i.e. either the rainfall event in progress, or a subsequent event that is considered to be independent and "new". The MIT could also be defined as the duration of a dry period D_{dry} following which the next occurrence of non-null rainfall marks the beginning of a new event. For dry periods shorter than the MIT, rain rates from either side of this period are considered to belong to the same "composite event". Various authors have proposed different values of MIT that ensure event independence. Llasat (2001) noted that: "*The definition of an episode is quite subjective. In this case it was felt possible to distinguish between two different episodes, when the time which elapses between them without rainfall exceeds 1h, which ensures that, the two episodes come from different 'clouds'*". Bocquillon and Moussa (2014) wrote: "*the constant rain observations on less than thirty minutes represent only 5% of all the rainy periods. The representative threshold of the discretization of the data is 30 minutes to an hour*."

Dunkerley (2008 a, b) carried out an analysis of the Inter-Event Time (IET) in order to check the influence of this variable on the definition of rainfall events, and its influence on the average rainfall rate. As emphasized in this study, when determining a value for the MIT, it is crucial to find an appropriate compromise between the independence of rain events and the intra-event variability of rain rates. The choice of MIT thus has a direct impact on the macrophysical characteristics that are ultimately determined by the analysis. Other researchers have proposed to use MIT values of 20 minutes, 1hour or even 1day (see Dunkerley, 2008a for a detailed list). In the present study it was decided to set the MIT to 30 minutes. This is in agreement with the value used by Coutinho et al. (2014), Haile et al. (2011), Dunkerley (2008a, b), Balme et al. (2006) and Cosgrove and Garstang (1995).

Table 1. Observation periods and availability of DBS observations, and numbers of rain events for the learning and test datasets

	Observation period	Availability (%)	Number of rain events
Learning data set	01/01/2013-12/31/2014	96.4%	234
Test data set	04/16/2008-01/31/2012	60%	311

When applied to our dataset, this choice leads to the identification of 545 rain events, which can be divided up into two subsets, i.e. one for learning and the other for testing (Tab. 1). The learning dataset is composed of observations collected over a two-year period between 2013 and 2014, with an availability of 96.4%, whereas the test dataset collected during the 2008 – 2012 period contains periods with missing data, due to a malfunction of the recording device.

2.2 Macrophysical description of rain events

Rain events contain a wealth of information, which generally needs to be condensed into a limited set of well-chosen features. However, there is no conventional or commonly accepted list, or specific set of macrophysical features that can be used to accurately describe and summarize an event. In the present study it was thus decided to consider a large number of features, allowing the macrophysical rain event information described in the literature to be correctly represented. Seventeen characteristics were selected and identified (Llasat, 2001; Moussa, 1991), and are listed in Tab. 2. Some of these are parameter-dependent, such as Pc which uses 3 values of the parameter c. These 3 values lead to 3 Pc indices, namely Pc1, Pc2, Pc3. Finally, a total of 23 descriptors were defined and numbered from 1 to 23 (column 1 in Table 2)

Among the 23 indicators (hereafter referred to as variables) corresponding to the previously defined features, some are very well known. These include the event duration (De), the quartile (Qi), the mean event rain rate (Rm) and the standard rain rate deviation (σ_R). Other less traditional parameters such as the parameter βL (indicator for the convective nature of the rain, see Llasat (2001)), the absolute rain rate variation of order c (Pc), or the absolute rain rate variation (Ps,c). Some variables that are usually used to describe time series, such as the fractal dimension, multi-fractal parameters, trend, seasonality, and autocorrelation, require a long series of data and are not well suited to an event-by-event analysis. This set of 17 features is not exhaustive, and some other features could also be included, depending on the application. One example is the case of hydrology, for which the positions of the intensity peaks inside the event could be a relevant feature. Although, for events comprising a very small number of samples (very low value of variable De), the computation of some indicators (σ_R, Qi) is questionable, in the present study the 23 variables were computed for each of the 545 rain events.

2.3 PCA analysis and normalization step

It is important to note that very few of these 23 variables are compatible with the probabilistic assumptions generally associated with exploratory statistical methods. They are often highly variable, with highly skewed distributions, and therefore do not have normal distributions. It is thus more difficult to make direct use of standard statistical methods with this data, as it may lead to misleading interpretations (Daumas, 1982). It is thus necessary to introduce an additional step in order to transform the original distributions into *quasi* normally distributed distributions. The most suitable type of normalizing transformation for each of these variables was selected empirically, by testing 7 different possible transformations (Tab. 3). For each variable, the retained transformation is that leading to a distribution having the strongest similarity to a normal distribution, i.e. with a kurtosis close to 3, and a skewness close to 0. For each indicator, the selected transformation is provided in the last column of Table 2.

Table 2. The 23 variables identified in the literature, used for the characterization of rain events

Number (#)	Feature name	symbol	Formula	Normalisation
1	Event duration	D_e	$D_e = T_{end} - T_{begin} + 1$ [min] With T_{begin}: Event start time and T_{end}: Event end time	1
2	Mean event rain rate	R_m	$R_m = \frac{1}{D_e}\sum_{t=T_{begin}}^{t=T_{end}} RR_t$ [mm h-1]	2
3	Intra-dry duration	D_d	$D_d = \sum_{t=T_{begin}}^{t=T_{end}} I_t$ [min] With $I_t = \begin{cases} 1 & if\ RR_t = 0\ [\text{mm } h^{-1}] \\ 0 & else \end{cases}$	0
4	First quartile	Q_1	The 25th percentile [mm h-1]	0
5	Median	Q_2	the 50th percentile [mm h-1]	0
6	Third quartile	Q_3	The 75th percentile [mm h-1]	2
7	Previous IET	IET_p	$IET_p = T_{begin}(current\ event) - T_{end}(previous\ event) + 1$ [min]	0
8	Mean rain rate over the rainy period	$R_{m,r}$	$R_{m,r} = \frac{1}{(D_e - D_d)}\sum_{t=T_{begin}}^{t=T_{end}} RR_t$ [mm h-1]	3
9	Event Rain rate std.	σ_R	$\sigma_R = \sqrt{\frac{1}{D_e}\sum_{t=T_{begin}}^{t=T_{end}}(RR_t - R_m)^2}$ [mm h-1]	2
10	Mode	M_0	M_0 =the most frequent RRt	0
11	Rain rate peak	R_{max}	$R_{max} = \max(RR_t)$	2
12	Dry Percentage in event	$D_{d\%e}$	$D_{d\%e} = \frac{D_d}{D_e}$	5
13	Rain event depth	R_d	$R_d = R_m * D_e / 60$ [mm]	0
14	Absolute rain rate variation of order c	P_{c1}	$P_{c_i} = \sum_{t=T_{begin}}^{t=T_{end}-1} \lvert RR_{t+1} - RR_t \rvert^{c_i}$ For $c_i = 0.5, 1, 2$	6
15		P_{c2}		3
16		P_{c3}		2
17	Normalized	$P_{C_{N1}}$	$P_{C_{Ni}} = \frac{P_{c_i}}{D_e}$ For $i = 1..3$	3
18	Absolute rain rate variation of order ci	$P_{C_{N2}}$		2
19		$P_{C_{N3}}$		0
20	Absolute rain rate variation of order C and threshold S	$P_{S,C}$	$P_{S,C} = \sum_{t=T_{begin}}^{t=T_{end}-1} \lvert \max[(RR_{t+1} - S), 0] - \max[(RR_t - S), 0] \rvert^C$ With $s = 0.3\ and\ c = 2$	6
21	β_L parameter	β_{L1}	$\beta_{L_i} = \frac{\sum_{i=T_{begin}}^{T_{end}} RR_t\ \theta(RR_t - L_i)}{\sum_{i=T_{begin}}^{T_{end}} RR_t}$ For Li = 0.3 , 1 , 3 $mm\ h^{-1}$ With $\theta(RR_t - L_i)$ is the Heaviside function defined as $\theta(RR_t - L_i) = 1\ if\ RR_t \geq L_i$ $\theta(RR_t - L_i) = 0\ if\ RR_t < L_i$	5
		β_{L2}		0
22				
23		β_{L3}		0

Following the normalisation step, Principal Component Analysis (PCA) was carried out on the learning data set (cf. end of section 2.1 and Tab.1). It follows that the two principal axes contain 73% of the total information, whereas the first 5 principal axes are needed to represent 90% of the total information. The IET_p variable (#7) is very well correlated with axis 5, whereas the other variables are not.

This means that there is no linear relationship between IET_p and the other variables. For this reason, this variable was not considered as a possible candidate, in the variable selection process, during the remainder of the study. The results obtained in Section 4.2 confirm that there is no relationship between this variable and the other 22 variables. The correlation circle on axes 1 & 2 (Fig. 1.a.) shows that among the 23 variables, 16 are well correlated with the axis (close to a unit circle) and are distributed in approximately 5 groups (hereafter referred to as PCA groups). A first PCA group (G_1) can be identified by the variables, which are grouped close to the first axis, and are well correlated with it. As an example, this is the case for the variables σ_R (#9), P_{C_N} (#17 – 18) and β (#21 to 23). A second PCA group (G_2) comprises the variables R_{max} (#11) and P_{C3} (#16), just above axis 1. The third PCA group (G_3) is formed by the variable P_{C2} (#15) only. The fourth PCA group (G_4) comprises the variables P_{c1} (#14) and *Ps,c* (#20) and is well correlated with axis 2. The last PCA group (G_5) is formed by the variable D_e (#1). The correlation circle on axis 1 & 3 (fig. 1.b.) shows that the variables Q_1 (#4), Q_2 (#5) and M_0 (#10) are quite well represented by these two axes. A similar remark can be made for variables D_d (#3) and β_{L1} (#21) on axes 1 & 4 (not shown).

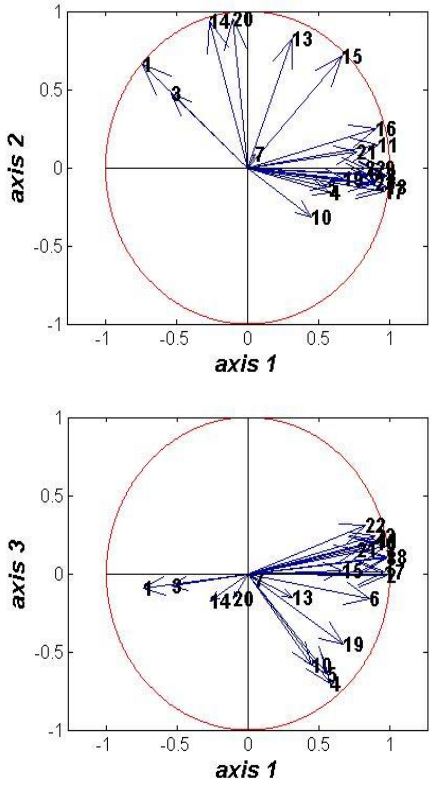


Figure 1. PCA on the learning data set based on the 23 variables described in Tab. 3. Left: correlation circle on axes 1 & 2. Right: correlation circle on axes 1 & 3. All of the variables are normalised according to the last column of Table 2.

Finally, PCA analysis clearly shows that, within each PCA group, many variables are highly inter-correlated, i.e. linearly dependant on each other. This means that several variables could be removed with no substantial loss of information. This leads to the following question: which variables can be removed in order to retain the most parsimonious subset of variables representative of the full dataset? The PCA extracts summary variables, which are a linear combination of original variables, but do not allow for the selection of variables. To answer to this question, we propose a method for the global selection of variables, which seeks to identify the relevant variables in a dataset. As it appears to be intuitively more advantageous to select variables with a physical sense, rather than using dimension reduction methods (e.g. PCA which is more suitable for the detection of linear relationships), the proposed method is based on the use of a genetic algorithm. The following section provides a brief introduction into the concept of genetic algorithms, and shows how they can be advantageously used for the selection of variables in the context of the present study.

3. Variable selection using a genetic algorithm

Computer-assisted variable selection is important for several reasons. Indeed, the selection of a subset of variables in a high-dimensional space can improve the performance of the model or its statistical properties, but also provides more robust models and reduces their complexity. In practice, it is not generally possible to try all potential combinations of variables, and to select the best of these, as a consequence of the enormous computational cost associated with such an approach. Among the many different variable-selection techniques described in the literature (Guyon and Elisseeff, 2003), we chose to develop a model based on the use of genetic algorithms, to search for an optimal subset of variables. Genetic algorithms (GAs) (Holland, 1975) are stochastic optimization algorithms based on the mechanics of natural selection, and the genetics described by Charles Darwin. In our study, a chromosome is defined as a subset made up from our 23 variables. A first generation composed of a population of 60 potential chromosomes is arbitrarily chosen. The performance of each chromosome (i.e. for each corresponding subset of 60 variables) is evaluated through a fitness function f. This fitness function is defined in such a way that the higher its value, the greater the fitness function's ability to represent the full dataset (of dimension 23), using the smallest possible number of variables. On the basis of the performance of these 60 chromosomes, we create a new generation of 60 potential-solution chromosomes, using classical evolutionary operators: selection, crossover and mutation. The performance of this new generation is then evaluated. This cycle is repeated until a predefined stop criterion is satisfied. The best chromosome from the current generation then provides the optimal subset of variables.

3.1 Methodology

We define by x^k the chromosome number k: $\boldsymbol{x^k} = (x_1, x_2, \dots, x_{23})$. $\boldsymbol{x^k}$ is a binary vector in $\{0,1\}^{23}$ space such that each component has the following meaning:

$$for\ all\ i\ in\ \{1,\dots,23\},\ \begin{cases} x_i = 1 & \text{The variable number i in Tab 2 column 1 is selected} \\ x_i = 0 & \text{The variable number i in Tab 2 column 1 is not selected} \end{cases} \quad (1)$$

Table 3. Transformations used to normalize the variables listed in Table 2.

Transformation number	Transformation name	Formula f(x)	Notes
0	Standardisation	$\frac{x - mean(x)}{std(x)}$	-
1	Power	x^n	$n = 0.05$
2	Boxcox	$\frac{(x^\gamma - 1)}{\gamma}$	$\gamma = -0.1$
3			$\gamma = -0.2$
4			$\gamma = -0.3$
5	Arc-sin of square	$\arcsin\sqrt{x}$	Data are between 0 and 1
6	decimal Logarithm	$\mathrm{Log}(x + c)$	$c = 0.1$

The word "selected" in eq. 1 means that the corresponding variable will be used, both in the learning step described in "Step 2" below, and for performance evaluation. Otherwise, if the corresponding variable is not selected it will be used only for performance evaluation.

As previously stated, the fitness function allows a measure to be provided of how well a minimal subset of variables can represent the entire data space (in dimension 23). The fitness function *f* is thus defined as follows:

$$f(\boldsymbol{x^k}) = \frac{1}{nb(\boldsymbol{x^k})\ te(\boldsymbol{x^k})} \quad (2)$$

where x_k is chromosome *k*, $n(x_k)$ is the number of selected variables in chromosome x_k and $t_e(x_k)$ is the topological error associated with chromosome *k*

As the aim of this approach is to minimise the number of selected variables *nb* and the topological error *te*, we seek to maximize the fitness function. The estimation of the topological error made from a Self-Organizing Map (SOM) is somewhat complicated, and requires some explanation. The notion of a Self-Organizing Map, introduced by Teuvo Kohonen (Kohonen, 1982, 2001), makes use of a popular clustering and visualization algorithm. SOM is a neural network algorithm based on unsupervised learning, derived from the technique of competitive learning (Kohonen, 1982, 2001; Vesanto and Alhoniemi, 2000). It may be considered as a nonlinear generalization, which has many advantages over the conventional feature extraction techniques such as Empirical Orthogonal Functions (EOF), or Principal Component analysis PCA (e.g., Liu et al., 2006). SOM applications are becoming increasingly useful in geosciences (e.g., Liu and Weisberg, 2011). As stated by Uriarte and Martín (2008): "The SOM provides a nonlinear, ordered, smooth mapping of high-dimensional input data manifolds onto the elements of a regular, low-dimensional array. The main characteristic of the projection provided by this algorithm is the preservation of neighborhood relationships; as far as possible, nearby data vectors in the input space are mapped onto neighboring locations in the output space". This property makes it straightforward to compute a topological error (see Uriarte and Martín, 2008, eq. 2). For each of the $\boldsymbol{x^k}$ chromosomes, a Self-Organizing Map $M(\boldsymbol{x^k})$ is learned on the learning dataset. Only the selected variables are used during the learning process. Finally, for each Map $M(\boldsymbol{x^k})$, the topological error $te(\boldsymbol{x^k})$ can be computed in accordance with eq. 2 in Uriarte and Martín (2008). Section 4 provides additional information concerning Self-Organising Maps.

The Genetic Algorithm is based on the following five steps (Fig.2.):

Step 1- Initialization: **(**initial population)

Randomly generate a population $\{\boldsymbol{x^k}$, k=1, … , 60$\}$ of 60 chromosomes of dimension 23.

Step 2- Evaluation: For each of the x^k chromosomes, a SOM $M(\boldsymbol{x^k})$ is learned. Only the selected variables are used for learning. Once the learning has been completed, the test dataset and the 23 variables are used on each of the 60 maps to estimate their topological error $te(\boldsymbol{x^k})$ allowing their fitness score $f(\boldsymbol{x^k})$ to be computed.

Step 3- Select the best chromosome $\boldsymbol{x^{Best}}$ from the full set of 60 chromosomes according to the fitness score previously computed with the test dataset. If $\boldsymbol{x^{Best}}$ remains unchanged over a period of 50 generations, stop the procedure and select the most relevant variables, i.e. those for which the corresponding components are equal to 1 in $\boldsymbol{x^{Best}}$. Otherwise, go to step 4.

Step 4- Selection: Create a new population of 60 chromosomes from the current population, by randomly sampling with replacement chromosomes based on their probabilities, determined using the formula:

$$\Pr(\boldsymbol{x^k}) = \frac{f(\boldsymbol{x^k})}{\sum_{i=1}^{60} f(\boldsymbol{x^i})} \quad (3)$$

Step 5- Reproduction: Mutation and Crossover possibilities in the new population.

Mutation: This consists in modifying (or not) certain components of the chromosomes. The probability of mutation is in general very low, and is commonly set to $p = 10^{-7}$. In the present case, the number of generations needed to reach the objective is less than a few hundred, such that the probability of a mutation is very low.

Crossover: In an initial step $\frac{60}{2} = 30$ pairs of chromosomes are randomly drawn from the population. Then, for each pair ($\boldsymbol{x}^k$, $\boldsymbol{x}^l$) (called parents) one crossover point, noted I_c, is randomly drawn over the range [1, 23], using a discrete uniform law. Two new chromosomes ($x^{k'}$, $x^{l'}$) are created as follows:

$$\begin{cases} \boldsymbol{x}^{k'} = (x_1^k, x_2^k, \dots, x_{Ic}^k, x_{I_c+1}^l, x_{I_c+2}^l, \dots, x_{23}^l) \\ \boldsymbol{x}^{l'} = (x_1^l, x_2^l, \dots, x_{I_c}^l, x_{I_c+1}^k, x_{I_c+2}^k, \dots, x_{23}^k) \end{cases} \quad (4)$$

Thus, from two parents, two children are generated, allowing a new generation to be produced with the same number of chromosomes. Finally, the algorithm returns to step 2.

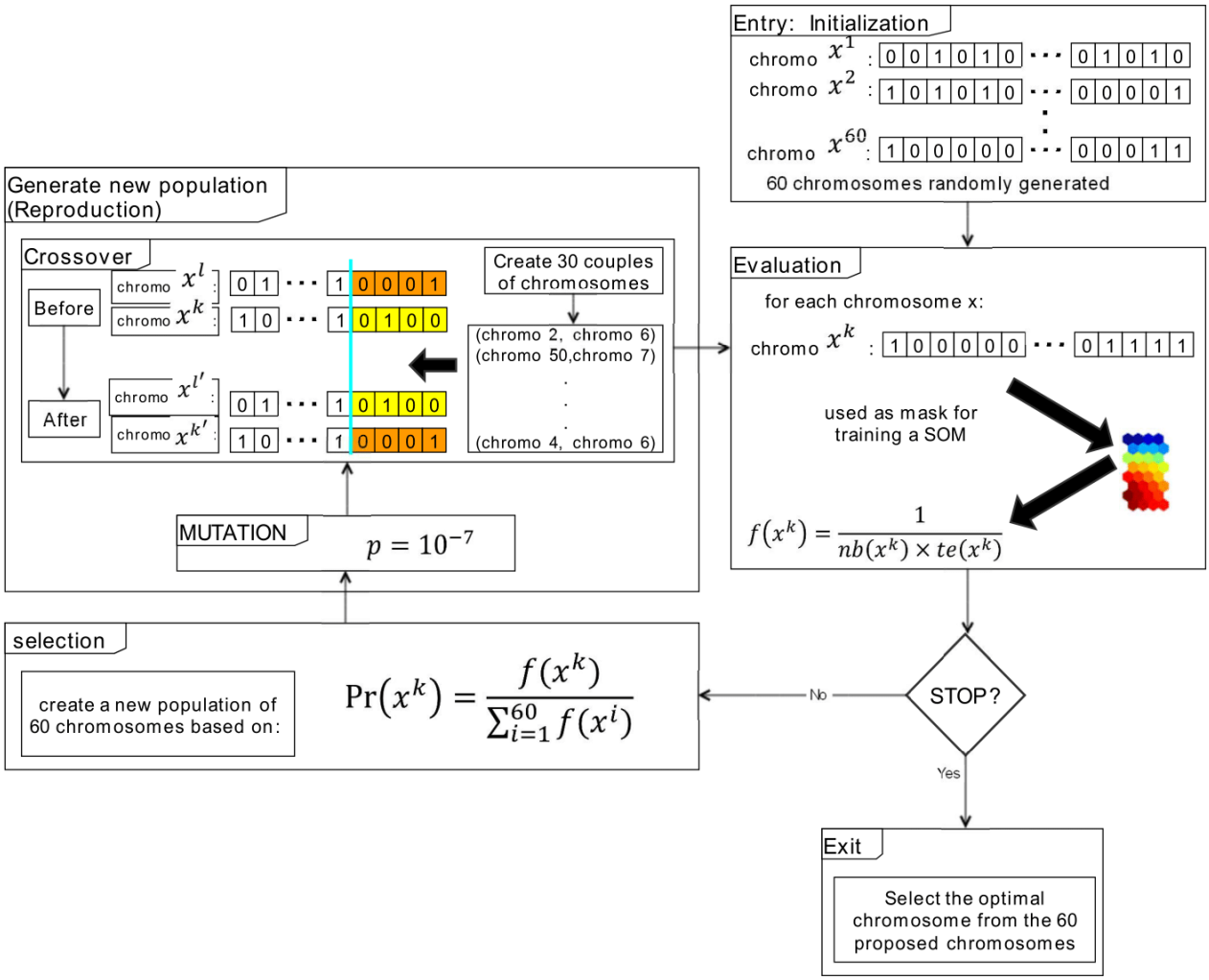


Figure 2. Diagram for the selection of variables based on a Genetic Algorithm associated with Kohonen Maps

3.2. Parsimonious description of a rain event

The genetic algorithm is applied to our datasets in order to obtain an optimal subset of variables forming a subspace, which can (in a certain sense) provide relatively accurate information concerning the global space, whilst having the particularity of containing non-redundant information. At the 187th generation the algorithm produces a subspace comprising 5 variables, namely: Event duration D_e (#1), Standard deviation σ_R (#9), maximum rain rate during event R_{max} (#11), Rain event depth R_d (#13), and Absolute rain rate variation P_{c1} (#14).

The 3 variables (D_e, R_{max}, R_d) selected using this data-driven approach are commonly used in the study of hydrological processes (Haile and al., 2011). Moreover it should be noted that the commonly used variable R_m, which is computed simply by dividing the Rain event depth (R_d) by the duration (D_e), was not selected by the algorithm. This result could be expected, since it is correlated with the latter variables, and the algorithm provides a parsimonious description. Concerning the Absolute rain rate variation (P_{c1}), this variable was proposed by Moussa and Bocquillon (1991). It tends to provide information on the structure of the events, more specifically related to smooth events with a small number of sharp peaks. In fact, this variable promotes low variations of RRt because

PC1 is in a certain sense a structure function of order c1 of the variable RRt (see #14 column 4 in Tab. 2), with a low value for the exponent (c1 = 0.5). Finally, the standard deviation variable (σ_R), which is a second-order moment, is the most commonly used indicator to describe the variability of the precipitation rate within the rain event.

4. SOM learned with the five selected variables

A SOM is a topological map composed of neurons. In the present case, a neuron is a vector of dimension 23 containing the 23 variables defined in Tab. 2. Each neuron has 6 neighboring neurons. SOM is an unsupervised neural network trained by a competitive learning strategy that performs two tasks: vector quantization and vector projection. The SOM, which is different to k-means, uses the neighborhood interaction set to learn the topological structure hidden in the data. In addition, in order to achieve optimal referent vector (neuron) matching, its neighbors on the map are updated, leading to the generation of regions in which neurons located in the same neighborhood are very similar. The SOM can thus be considered as an algorithm that maps a high-dimensional data space onto a two-dimensional space called a map. A map can be used both to reduce the amount data by means of clustering, and to project the data in a nonlinear manner onto a regular grid (the map grid).

In the present study we used the toolbox developed by "the SOM Toolbox Team", which is available at the following site: . A SOM with 8×8 = 64 neurons is considered here. This choice corresponds to a compromise, since a smaller map would not be able to distinguish fine details whereas, in view of the number of observations, and a larger map would not be meaningful.

After learning by the GA algorithm described in the previous section, the resulting map $M(\boldsymbol{x^{Best}})$ can be used to assign to any event the best matching reference vector (neuron), in accordance with the 5 selected variables associated with the chromosome $\boldsymbol{x^{Best}}$. The $M(\boldsymbol{x^{Best}})$ map obtained with this procedure can be considered as an optimal representation of the initial data set.

Fig. 3 shows the distance matrix. For each neuron, the color indicates the mean distance between a neuron and its neighbors. The value at the center of each neuron represents the number of rain events of the learning data set, captured by the corresponding neuron. All neurons capture rain events and slightly more than half of these capture between 3 and 5 rain events, which is close to the value that would be obtained ($234 / 64 \cong 4$) if the rain events were uniformly distributed over the map.

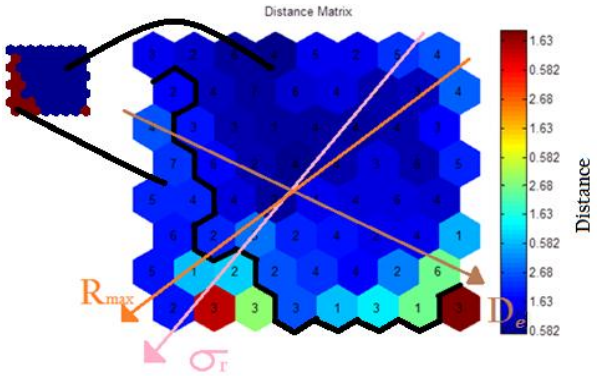


Figure 3. Distance matrix for the $M(x^{Best})$ map: The colour of each neuron represents the average distance between itself and its neighbouring neurons. The value inside each neuron indicates the number of rain events that it has captured inside the learning dataset. The black line separates the neurons into 2 classes, using the Hierarchical Ascendant Classification (see section 4.1). The arrows represent the gradients of the variables R_{max}, σ_R and D_e

4.1 Projection of the selected and unlearned variables onto the SOM

The five variables D_e, σ_R, R_{max}, R_d and P_{c1} used for learning are referred to as 'selected' variables, whereas the remaining 18 variables are referred to as 'unlearned' variables. In order to study the relationship between these variables, Fig. 4 shows the projections for each of the variables in the $M(\boldsymbol{x^{Best}})$ map obtained with the aforementioned GA selection algorithm. The variables are discussed individually, by considering their structure, as well as the relationships between them. We note that the map is well structured for the majority of variables. This advantageous structuration of most of the variables confirms the ability of the selected variables to summarize all of the significant characteristics of rain events. Only a small number of characteristics are not adequately represented. It should be noted that almost all variables are structured according to the first or second diagonal. Among these, one may consider an initial subset comprising variables that are more or less structured according to the first diagonal. This is the case for the unlearned variable D_d, as well as for the selected variables P_{c1} and D_e. A second subset comprising variables that are structured in approximate accordance with the second diagonal can be identified. This is the case for the unlearned variables $R_{m,r}$, $R_{m,}$, $P_{C_{Ni}}$ which are very similar to the selected variable σ_R. The unlearned variables Q_3, P_{C3}, $P_{S,C}$ also belong to the second subset and have a structure close to that of the selected variable R_{max}.

The map can be related to the previously implemented PCA (Fig.1). As can be seen in Fig. 4, the variables P_{C3} (#16) and R_{max} (#11), which have a similar structure, also belong to the same PCA group, namely group G_1 (see section 2.3, Fig. 1.a). It is interesting to note that the variables $P_{S,C}$ (#20) and R_{max} (#11), which also have a similar structure, do not belong to the same PCA group (groups G_4 and G_2 respectively) and are uncorrelated (they are orthogonal in Fig.1a). This remark means that the topological map reveals a relationship that cannot be detected using PCA. As the Rain event depth (R_d) depends on both the duration and the

intensity of the events, the corresponding map has a top-down structure. Two distinct situations thus occur:

- Those events which contribute the greatest quantities of water (Fig. 4., brown neuron at the bottom right of R_d) are among the longest (see corresponding neuron of D_e), but do not have an extremely high peak rain rate (see corresponding neuron of R_{max}) and are quite smooth (see corresponding neuron of P_{c1} and σ_R).
- Other events which contribute large amounts of water (but less than previously) (Fig. 4. red neuron at the bottom left of R_d) have short durations (see corresponding neuron of D_e), but are violent (see corresponding neuron of R_{max}) and are less smooth (see corresponding neuron of P_{c1} and σ_R). The latter case reflects situations that are typical of convective storms.

The resulting map confirms the dependence structure of the two hydrological variables R_d and D_e studied by Gargouri and Chebchoub (2010).
The variable IET_p (Previous *IET*): the map is not structured, reflecting the independence of the characteristics of a rain event with respect to the drought period preceding the event. This corroborates the results of several previous studies (Lavergnat and Gole, 1998, 2006; Akrour et al., 2015) dealing with rain support simulations. When studying temperate mid-latitudes for relatively short periods, these authors noticed that successive rain and no-rain periods are uncorrelated, such that a rain time series could be considered as an independently drawn, alternating series of rain events and periods without rain. This is equivalent to an inter-event time (IET) that does not characterize the rain events. The same effects are not necessarily observed at other locations, and under different climatological conditions. Brown et al. (1983) also investigated a possible correlation between IETp and the intra-event characteristics, and concluded that their data provided no evidence of this.

Since the variable β_L: β_L (Llasat, 2001) is considered to represent a measure of the convective nature of the rain, it makes sense that the three variables β_{L1}, β_{L2}, β_{L3} are structured similarly, with the peak rain rate variable R_{max}. This relationship is clearly visible on the maps.

Several other relationships, which are not described in detail here, can be observed. These include the correlation between the Normalized Absolute rain rate variation ($P_{C_{Ni}}$) and the standard deviation of the intensity (σ_R). We conclude that the combination of the five selected variables provides a relatively accurate summary of the information needed to describe the rain events. The poor structuring of some variables is justified by the independence of these variables with respect to the properties of the rain events; this is the case for the variable Dry Percentage in event D_d or the variable IET_P.

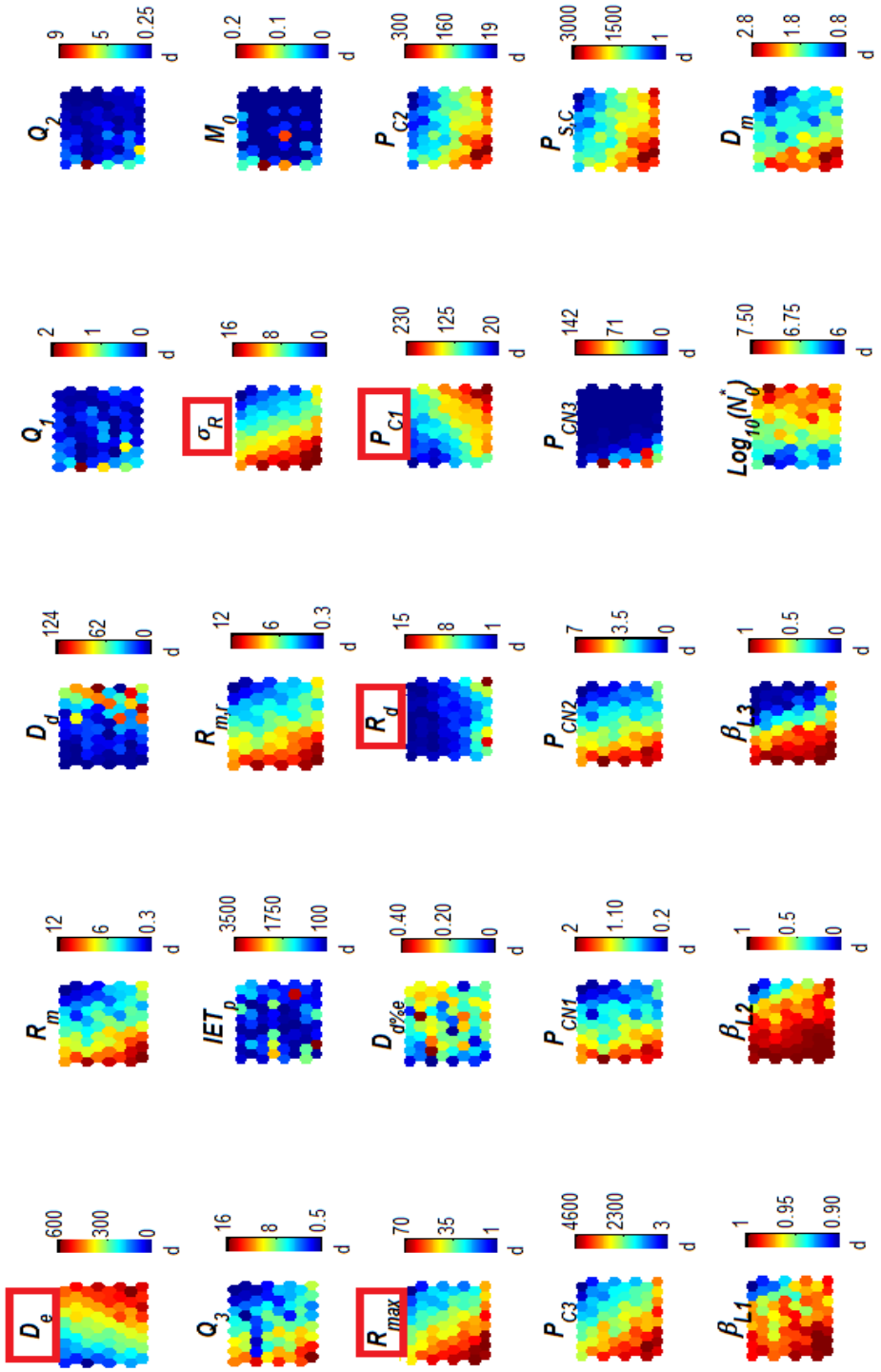


Figure 4: Projection of the $M(x^{Best})$ map according to the 23 variables. The red-framed variables are those selected by the GA algorithm. The last two variables D_m and N_0^* are defined in section 5

4.2 Representation of rain events on SOM

In an effort to provide additional information for validation of the map, we compared each of the 23 variables with their corresponding value given by the SOM, for the learning dataset and the test dataset. For each of the 311 events of the test dataset, the best matching unit of the SOM, i.e. the neuron that is the closest to the event, is determined with respect to the five selected variables. As an example, for each event Fig. 5 shows the current value of the unlearned variable β_{L3} as a function of the corresponding value given by the best matching unit of the event. A spread can be seen, in particular in the central zone, whereas the spread is relatively small for values located near to the edges (which are more numerous). A linear regression leads to a relatively good determination coefficient (R-square) (0.96 and 0.89 respectively, for the learning and test data sets). Table 4 lists the value of R-square for the 23 variables obtained with the learning and test data sets. As expected, the coefficient of

determination of the variable IET_p is very poor (*0.31/0.26*), since this variable is not related to the 5 selected variables, and as a consequence cannot be well represented by the SOM (Fig. 4). The selected variables have good determination coefficients, with both the learning and the test datasets; this confirms the quality of the learning and the generalization ability of the SOM. The quality of the learning step is confirmed by the fact that the R-square values of the selected variables obtained on the test set are close to those obtained on the learning set. The R-squares corresponding to the unlearned variables obtained on the learning data set emphasize the ability of the selected variables to provide the information contained in the unlearned variables; in the case of the test data set it denotes the ability of the SOM to derive all event characteristics from the selected variables only.

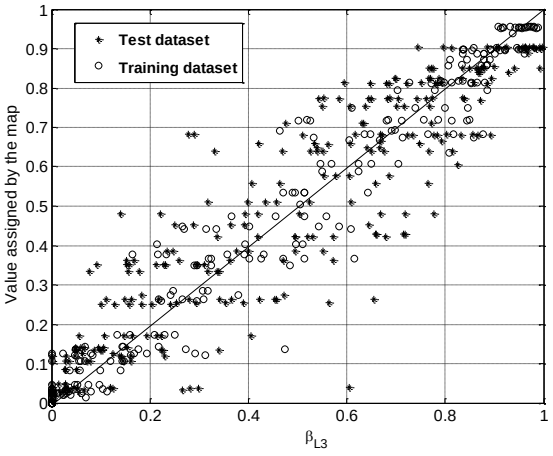


Figure 5. The variable β_{L3} versus its corresponding value, given by the best matching unit: from the learning data set (circles), and on the test dataset (stars). The solid line corresponds to the first diagonal

4.3 Hierarchical clustering of rain events

We have shown that the distance matrix (Fig. 3) confirms the successful deployment of the map. Based on the distance between neurons, it appears that neurons can be grouped to obtain a limited number of classes, each with its own characteristics. In order to group the 64 neurons into a small number of classes, a hierarchical cluster analysis was carried out (Everitt, 1974). Only the five selected variables were used for the classification, and a Euclidian distance was selected for the hierarchical algorithm. Fig. 6 shows the resulting dendrogram, applied to the 64 neurons.

Depending on the physical processes involved, experts tend to separate rain events into two different classes: stratiform and convective events. Although this classification is relatively crude, since stratiform and convective events can sometimes exist inside the same rain event, it is very commonly used. Concerning the times series, most authors use a very simple scheme to distinguish between stratiform and convective rain types. For reasons of simplicity, rain classification is sometimes defined using the instantaneous rain rate and the standard deviation estimated over consecutive samples. As an example, Bringi et al. (2003) defined stratiform rain samples when the standard deviation of the rain rate, taken over five consecutive 2-min samples, is less than 1.5 $mm.h^{-1}$, the convective rain samples are defined for a rain rate greater than or equal to 5 $mm.h^{-1}$, and the standard deviation of the rain rate over five consecutive 2-min samples is greater than 1.5 $mm.h^{-1}$.

Firstly, we separate the dendrogram into two classes. The first class contains 51 neurons and 79% of the observations, whereas the second class contains 13 neurons and 21% of the observations. The solid black line in Fig. 3 corresponds to the dividing line between these two classes. The first class, containing the greatest number of neurons, is in most cases characterized by relatively low rain rates. This can be seen by examining the structure of the map, according to the mean rain rate variable (R_m). Moreover, analysis of the standard deviation (small values of σ_R), absolute rain rates P_c (high values of P_{c1} , and low values *of* P_{c3}) shows that this class is more or less characterised by quiet, homogeneous events. Our analysis of event durations (D_e) shows that this class contains both short and long durations, but is dominated by the latter. These characterizations are relatively well matched to a description involving stratiform and stable precipitations, which are often the consequence of the slow, large-scale uprising of a large mass of moist air which then condenses uniformly.

Table 4: Coefficient of determination obtained on the learning and test data sets. The values with a dark grey background correspond to the 5 selected variables

Variables	D_e	R_m	D_d	Q_1	Q_2	Q_3	IET_p	$R_{m,r}$	σ_R	M_0	R_{max}	$D_{d\%e}$
R^2 learning data set	0.96	0.91	0.58	0.57	0.48	0.77	0.31	0.93	0.97	0.50	0.96	0.52
R^2 Test data set	0.93	0.84	0.55	0.57	0.50	0.74	0.26	0.84	0.86	0.54	0.82	0.50

Variables	R_d	P_{c1}	P_{c2}	P_{c3}	P_{cN1}	P_{cN2}	P_{cN3}	$P_{S,C}$	β_{L1}	β_{L2}	β_{L3}	-
R^2 learning data set	0.97	0.97	0.93	0.94	0.91	0.95	0.78	0.94	0.70	0.89	0.96	
R^2 Test data set	0.94	0.99	0.91	0.83	0.83	0.85	0.71	0.82	0.61	0.76	0.89	-

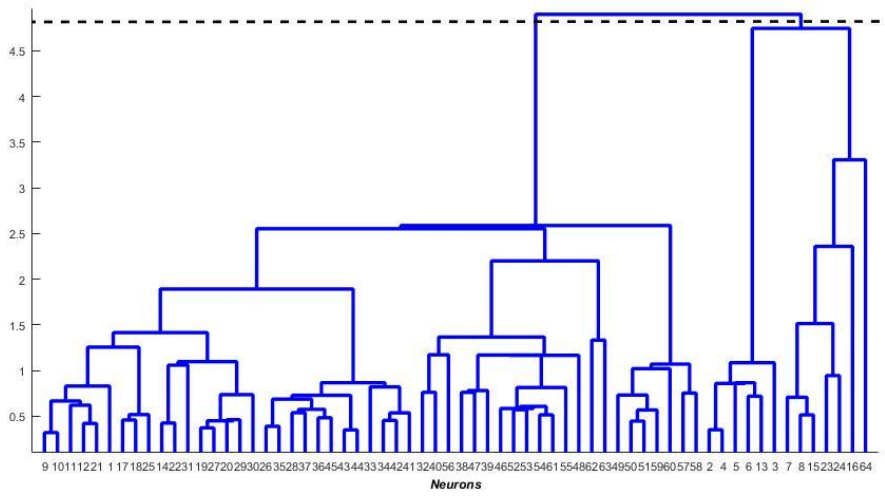


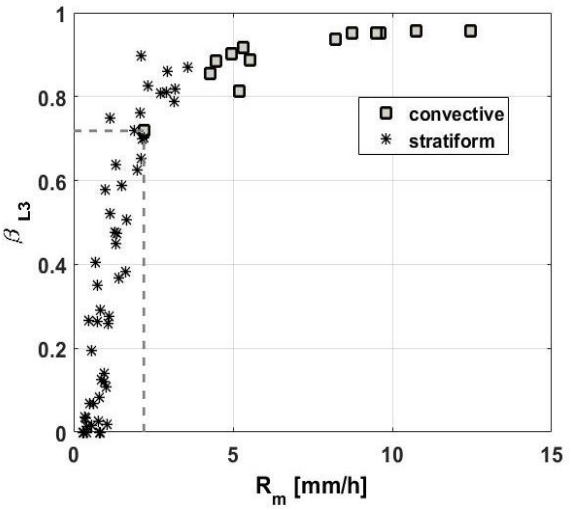
Figure 6. Dendrogram obtained from the Hierarchical Cluster Analysis of the 64 neurons in the SOM. The horizontal dashed line represents the threshold between the two classes.

The second group is characterized by a smaller number of neurons. This corresponds to the higher values of the mean rain rates (R_m) and peak rain rates (R_{max}). The variables σ_R, P_c have the opposite values with respect to those of the previous group. Most of the event durations (D_e) in this group are short, with the exception of neuron #64 (bottom right on the maps). This group fits well with the definition of convective events resulting from the rapid rise of air masses loaded with moisture, for buoyancy. This convective moist air can lead to the development of cumulus clouds up to an altitude in excess of 10 km, and to heavy rain.

Our analysis of the structure of the variables $\beta_{L1}, \beta_{L2}, \beta_{L3}$ in Fig. 4 confirms the previous interpretation of the two groups. These three variables, which are representative of convective rain, have high values for the neurons belonging to this group.

Figs. 7.a and 7.b show the neurons in the R_m, $\beta_{L3}$and P_{C2} subspace. These 3 variables were not used in the learning step. Nevertheless, the two classes are well separated, although an overlap does occur in Fig 7a due to neuron #64 (bottom right on the map, Fig. 4). We checked although it belongs to the convective class, this neuron nevertheless has some characteristics of the stratiform class.

The hypothesis that the two categories of precipitation events corresponding to different dynamic regimes can be identified solely on the basis of hydrometeorological variables is in agreement with the findings of Molini et al. (2011). These authors have shown that there is a strong agreement between the hydro-meteorological classification (based on the duration and extent of events from rain gauge network data), and dynamic classifications (the convective adjustment time-scale identified to distinguish between equilibrium and non-equilibrium convection derived from ECMWF analysis). We conclude that this unsupervised automatic clustering, based on the five selected variables, makes it possible to correctly implement a classification with these two well-known classes (stratiform and convective). It should be noted that, unlike other classifications described in the literature, this was established without making use of *a priori* information, since it is produced by an unsupervised process.



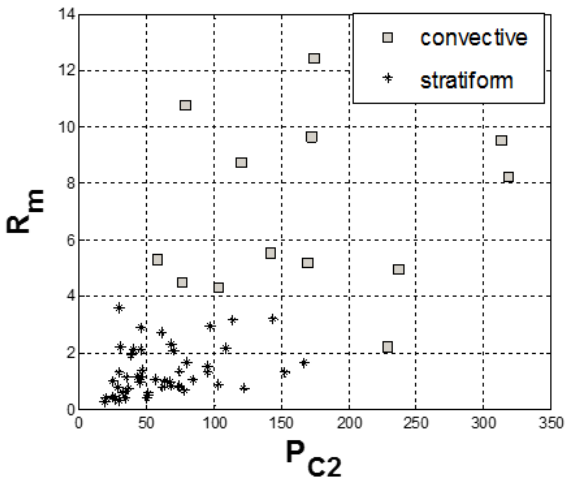


Figure 7. Representation of the neurons in the Rm, $\beta_{L3}$and R_m, and P_{C2} subspaces. The stars represent neurons from group 1 (stratiform), and the squares correspond to neurons from group 2 (convective). Dashed lines indicate the neuron #64

4.4 Classification of events into several classes

From the stratiform and convective classification described above, it is interesting to refine the two classes into a set of subclasses. The synoptic rainfall associated with mid-latitude depressions provides an example of stratiform precipitation, which forms in depressions in the vicinity of warm and cold fronts. The very light type of rainfall (drizzle) associated with stratus or stratocumulus is included in the class of stratiform precipitation. This can occur under anticyclonic conditions, or in the warm region of a depression. The associated rain depths (R_d) are minimal, and usually have no hydrological impact other than superficial wetting. In order to identify relevant subclasses, our classification was broken down into a number of unknown subclasses, such that $n > 2$.

An important step in hierarchical clustering is the selection of an optimal number of partitions (n_{opt}) in the dataset (Grazioli et al., 2015). Many indices can be used to evaluate each partition, from the point of view of data similarity only. Most of these evaluate the scattering inside each cluster, with respect to the distance between clusters, and assign relatively favourable scores to partitions with compact and well-separated clusters. Although different indices were tested, these did not provide the same number of subclasses (between 2 and 32 with the indices tested in this study). It should be noted that these did not take the physical meaning of each class into account. Finally, we chose $n_{opt} = 5$, since higher values led to classes with the same physical sense. The new classification based on the use of five subclasses is shown in Fig. 8.

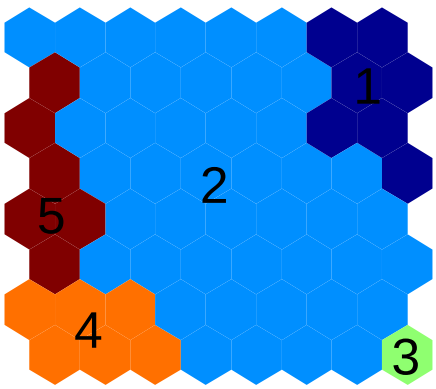


Figure 8: Hierarchical Clustering of the map into five subclasses. The colours represent the subclass numbers: Subclass 1: Dark blue, Subclass 2 : blue, Subclass 3 : Green, Subclass 4 : Orange, Subclass 5 : Red

From these five subclasses, two belong to the stratiform class and the other three belong to the convective class. In the learning dataset, the first subclass represents 12% of all events, and respectively 68%, 1.2%, 6.8% and 12% for subclasses 2 to 5. The characteristics of these five subclasses are summarized below and in Tab. 5. The five selected variables are remarkably heterogeneous between classes, meaning the accuracy of these variables for clustering:

- **Subclass 1** (drizzle and very light rain): the main feature of this class is the very low mean value (R_m) and standard deviation σ_R of rain rate events, in addition to the features of the superclass. The mean rain rate events lie in the range [0, 0.5] mm.h^{-1} with a mean value of 0.36 mm.h^{-1} and $\sigma_R$in the range [0, 3] mm.h^{-1}, with a mean value equal to 0.1 mm.h^{-1}. Although this event has a significant duration, the corresponding subclass, which corresponds to drizzle, involves only small quantities of water. It can also be noted that a low value of β_{L3} is a good indicator (< 0.01) for drizzle.

- **Subclass 2** ("normal" events): this is a relatively broad class containing 68 % of all events, with a mean event rain rate (R_m) in the range [0.5 , 6] mm.h^{-1} and a mean value of 1.48 mm.h^{-1}. The standard deviation σ_R lies in the range [1, 10] mm.h^{-1} with a mean value of 2. This subclass is characterised by a significant relative variation of some parameters (D_e, R_m, P_{c1} for instance), together with dry periods (D_d), which may be sufficiently long.

The three remaining subclasses correspond to convective classes of events, which are characterized by a strong temporal heterogeneity and significant intensities. Depending on the depth of rain events, this convective class is subdivided into:

- **Subclass 3:** containing relatively long events (D_e) with high values for the rain event depth (R_d) variable and P_{c1}. This class represents events with a very small likelihood of occurrence (1.2%).

- **Subclass 4:** containing relatively short events (D_e) with peak rain rate R_{max} > 50 mm.h^{-1}, in addition to strong heterogeneities (σ_R, P_{C2} and P_{C3} are high) and large values for the convective indicator (β_{L3}).

- **Subclass 5**: the events in this subclass are characterised by relatively low values for the rain event depth (R_d). This is due to the short duration of the events (D_e). The variables σ_R and P_{C3} remain high. Another feature of this subclass is that it includes continuous events only, with no short, embedded dry periods (low values of D_d in Fig. 4 and Tab. 5).

To conclude this section, this new classification allows the conventional definition for stratiform events to be refined. The convective classification can be subdivided into five different subclasses, each of which is homogeneous. This classification is obtained for mid-latitude climates. As the dataset used in this study is representative of only one specific region and topography (i.e. the temperate climate encountered in the Ile de France region, France), its analysis cannot reveal information related to different processes, i.e. those which are not sampled in the dataset. Such processes could lead to the identification of additional specific clusters of events. In particular, there are no orographic rainfall events or oceanic observations. The final step in this study involves assessing whether the homogeneous character of each class is preserved at the microphysics scale, and

attempting to identify any relationships between the information present at the scale of both the microphysics and the macrophysics of these events (hydrological information).

5. Microphysical point of view

Our study of the microphysical properties of rain is based on a comprehensive analysis of its drop size distribution *N(D)*, corresponding to the number of raindrops per unit volume and per interval of diameter *D*. The shape of *N(D)* reflects the microphysical processes involved. The identification of various features of the drop size distribution, as well as the type of precipitation, is very useful for many applications. As an example, this information is used in the calculation of heating profiles in the precipitation parameterization of atmospheric models, to gain a more detailed understanding of microphysical processes, as well as for the development of rain retrieval algorithms applied to remote sensing observations. The microphysical characteristics of rainfall act as hidden variables that affect the relationship between microwave remote sensing measurements and the volume of water in a rainfall event (Ulaby, 1981; Iguchi, 2009). It can thus be very useful to use conventional rain gauges to determine the microphysical characteristics of rainfall events, thereby improving the quality of active or passive remote sensing observations, and the spatial properties of rainfall events in particular.

Table 5: Summary of rain event subclasses computed with the learning dataset.

	Stratiform events		**Convective events**		
	Subclass 1	**Subclass 2**	**Subclass 3**	**Subclass 4**	**Subclass 5**
Variables	Mean	Mean	mean	mean	mean
$\boldsymbol{D_e(min)}$	321	149	464	75	49
$\boldsymbol{\sigma_R}$	0.36	2.01	3.62	11.7	9.64
$\boldsymbol{R_{max}(mm\ h^{-1})}$	2.08	10	22	52.7	36.06
$\boldsymbol{R_d(mm)}$	1.99	2.62	11.24	6.9	2.72
$\boldsymbol{P_{c1}}$	75.7	64.5	193	78.2	40.94
$\boldsymbol{R_m(mm\ h^{-1})}$	0.37	1.48	2.35	7.85	7.11
$\boldsymbol{D_d(min)}$	80	31	75	11	1
$\boldsymbol{\beta_{L3}}$	0.01	0.42	0.48	0.89	0.86

A general expression for the drop size distribution defined by Testud et al. (2001) is commonly used in the literature. This allows a distinction to be made between the stable shape function f and the variability induced by rain. This variability is represented by two microphysical parameters, namely the mass-weighted volume diameter (D_m) and the parameter N_0^*. In some studies, the term N_w is used rather than N_0^*. Not all authors use exactly the same units, in particular Bringi et al. (2003) and Suh et al. (2016) use $mm^{-1}m^{-3}$ for the units of N_w, rather than the unit m^{-4} which is used in this study for N_0^*.

$$N(D) = N_0^* f\left(\frac{D}{D_m}\right) \quad [m^{-4}] \quad (5)$$

where D_m and N_0^* are defined as:

$$D_m = \frac{M_4}{M_3}\ [mm], \quad N_0^* = \frac{4^4}{\Gamma(4)}\frac{M_3^5}{M_4^4}\ [m^{-4}] \quad (6)$$

and M_i is *i*th-order moment of the drop size distribution *N(D)*:

$$M_i = \int_0^{+\infty} N(D) D^i dD \quad (7)$$

Rain samples are usually analysed by computing the microphysical parameters (D_m and N_0^*) for each rain sample obtained over a given time scale. In the present study, *N(D)* is obtained by considering the entire raindrop collection corresponding to each rain event of (variable) duration D_e. This approach leads to one pair (D_m, N_0^*) of microphysics variables per rain event, whereas most other authors rely on values computed over a fixed time scale.

Projections of the learned map, according to D_m and N_0^*, are shown in Fig. 4 (bottom right). It can be seen that the two maps are well structured, and that these two parameters have opposite influences on the map projection. Although these two microphysical parameters were not learned, the relationship between them is clearly accounted for by the information used to structure the map (the 5 selected variables). Moreover, the existence of a relationship between the microphysical and macrophysical features of the rainfall is also confirmed in this figure, since both of the macrophysical variables used to learn the SOM, i.e. σ_R and R_{max}, have patterns similar to those revealed on the D_m map.

Many authors, including Atlas et al., 1999; Bringi et al., 2003; Marzuki et al., 2013; Suh et al. 2016, have endeavoured to associate specific microphysical properties

with each type of precipitation (convective or stratiform). In view of the maps shown in Fig. 4 and the convective/stratiform classification developed in section 4.3, we are able to confirm that precipitation events classified as stratiform express small values for D_m and large values for N_0^*. In the case of the convective class, the opposite trend is observed (i.e. larger values for D_m and smaller values for N_0^*). Similar observations have been reported by Testud et al. (2001). It can also be noticed that the two microphysical variables are relatively homogeneous in the convective class, whereas in the stratiform class they are characterised by a higher level of variability.

In order to improve our analysis of the microphysical information embedded in the dataset, we analysed the relationship between the two microphysical parameters using the reference vectors (neurons) from the map, which include information related to the original rain events.

Fig. 9 shows the variable D_m as a function of N_0^* for the 64 neurons on the map. This relationship is indicated through the use of distinct markers to identify the five subclasses defined in section 4.4, thus facilitating the discussion of the microphysics associated with stratiform and convective rain. The two solid lines show the linear regressions computed for these two classes.

In the case of the stratiform subclasses (1 and 2) a clear relationship can be observed between the two variables. The microphysics characteristics of these two subclasses are clearly distinct. Indeed, subclass 1 (drizzle and light rain) has the smallest D_m and the highest N_0^*, and varies over just a small range. Conversely, as in the case of the macrophysical variables (see section 4.4), the microphysical characteristics of subclass 2 (normal events) are considerably more heterogeneous. Knowledge of D_m makes it straightforward to identify the corresponding subclass. As a consequence, an event with D_m lying in the range [0.5, 1] millimetre belongs to subclass 1. Similarly, it is very likely that an event with D_m lying in the range [1, 1.7] millimetre belongs to subclass 2.

For the convective events (subclasses 3, 4, 5), small differences can be noticed with respect to N_0^*. In the range [1.7 , 2.5] mm, two neurons belonging to subclass 4 are close to a neuron belonging to subclass 5, and therefore have similar microphysics. Although they are located far from all other subclass 2 neurons, three isolated neurons belonging to subclass 2 (stratiform) can be noted. These are characterised by relatively strong values of D_m (2 mm) and low values of N_0^*. The corresponding events are a mixture of stratiform and convective rain. A typical case is given by convective rain associated with strong rain rates occurring at the beginning of an event, whereas the remainder of the event is stratiform with low rain rates and small variations.

Following our classification, Fig. 9 indicates that there are real relationships between the macrophysical and microphysical variables. Nevertheless, knowledge of the variables (D_m, N_0^*) does not allow the correct subclass to be determined in all cases.

Researchers who study microphysical features and their association with specific types of precipitation use simple schemes, based on rain rate estimations over a fixed period of integration (a few minutes), in order to separate stratiform and convective rain types. They also use these simple schemes to label D_m and N_0^* as stratiform or convective (Testud et al., 2001). This approach is significantly different to the method presented here, which assumes that all of the samples in a given event belong to the same class. Our values for N_0^* and D_m are thus computed for the time scale of a given event, rather than for a fixed integration time. Thus, although in the present study a good agreement is found for the range of values covered by D_m, those determined for N_0^* do not cover the same range as in the case of the previously cited studies.

Many previous authors have observed that the drop size distribution is closely related to processes controlling rainfall development mechanisms. In the case of stratiform rainfall, the residence time of the drops is relatively long, and the raindrops grow by the accretion mechanism. In convective rainfall, raindrops grow by the collision–coalescence mechanism, associated with relatively strong vertical wind speeds. Numerous studies have been published concerning the variability of N_0^* and D_m: Bringi et al. (2003) studied rain samples from diverse climates and analysed their variability in stratiform and convective rainfall; Marzuki et al. (2013) investigated the variability of the raindrop size distribution through a network of Parsivel disdrometers in Indonesia; and Suh et al. (2016) investigated the raindrop size distribution in Korea using a POSS disdrometer. In the case of stratiform rain, all of these authors observe that N_0^* and D_m are nearly log-linearly related, with a negative slope. This is consistent with the trend shown in Fig. 9 for the two stratiform subclasses (1 and 2). Even the three distinct neurons, which are isolated from the others, appear to be governed by the same relationship.

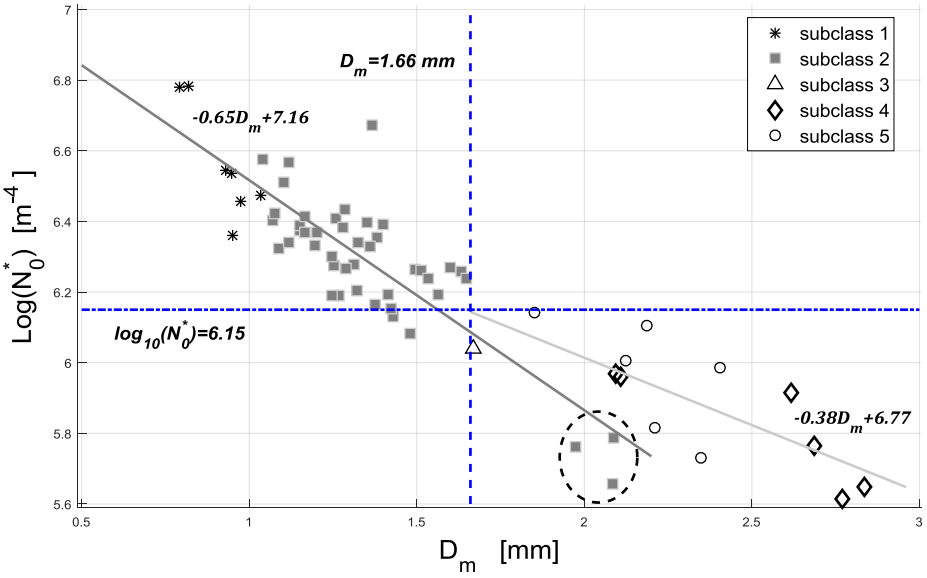


***Figure 9**:* Microphysical variable N_0^* versus D_m for the five rainy event subclasses. The three neurons corresponding to mixed events are circled. The dashed lines indicate the limits defined by D_m > 1.66 and Log(N_0^*)>6.15

Marzuki et al. (2013) noted that during convective rain the increase in value of N_0^* with decreasing D_m is nearly log-linear, with a flatter slope. In the present case, the dependence is also log-linear, with a slope that is slightly flatter for convective events than for stratiform events. In the aforementioned studies, the data was aggregated over time, campaign or site, on the basis of a criterion computed over a fixed period of time. We believe that this process is weakly suited to determining the properties of convective events, as a consequence of their strong variability and shorter characteristic time. In this study we were able to retrieve the log-linear relationship between N_0^* and D_m without having to learn it directly.

When applying our algorithm to the various macroscopic properties by rain event, we also take into account the variability of rain within an individual rain event. Fig. 9 clearly shows that the spreading of parameters N_0^* and D_m inside each subclass has the same magnitude as the distance between subclasses. This remark confirms the hypothesis of Tapiador et al. (2010): the intra-event variability can exceed the inter-event variability, due to events arising from different precipitation systems. It is thus preferable to examine the properties of events with a more general approach, rather than using individual samples to study the distinction between stratiform and convective processes. The three isolated neurons in subclass 2 described above (circled in Fig. 9) have the same properties as the other events of their subclass (i.e. the same slope for the log-linear relationship between N_0^* and D_m). This example confirms the ability of our methodology to preserve the macroscopic information needed to cluster rain events, thus allowing the intra-event variability as well as microphysical information to be (partially) retrieved.

Suh et al. (2016) also compare log(N_0^*) and D_m pdf, for the case of stratiform and convective samples over a 4-year period. On the basis of the D_m pdf of both stratiform and convective classes, they compute a threshold value for D_m, such that when D_m > 1.66 mm the rainfall samples are mainly convective, and when D_m < 1.66 mm they are mainly stratiform. This finding is consistent with the results of Atlas et al. (1999), who also found a threshold value for D_m, distinguishing between convective and stratiform rainfall. In Fig. 9, it can be seen that this threshold is confirmed (vertical solid line), with D_m smaller than 1.6 mm corresponding to stratiform events, whereas higher values correspond to mainly convective events. When we consider the events analysed in the present study, there are also three neurons corresponding to a "mixed event" beyond this threshold.

Suh et al. (2016) show in Fig. 4c of their study that the pdf for convective rainfall is higher than that corresponding to stratiform rainfall, when log(N_0^*) > 6.2 (N_w = 3.2 in their figure). As described above, by considering the data corresponding to rain events, rather than to samples recorded over fixed periods of time, our range of values for N_0^* is smaller than that used in other publications. In addition, log(N_0^*) < 6.15 for all neurons labelled as convective in our study, which is very close to the value of 6.2 determined by Suh et al. (2016).

In view of the generally satisfactory retrieval of microphysical information from macrophysical parameters, we are of the opinion that the topological map successfully restores some of the information implicitly embedded in the dataset. It is thus interesting to note that the macrophysical parameters of rainfall are related to its microphysical properties. Firstly, the map collects similar events, whilst ensuring, through the minimization of topological errors, that the unfolding of the map is correct. A neuron is thus closer to its neighbours than to any other neuron on the map. This criterion ensures that the data space is optimally partitioned into connected subparts, such that the neurons on the map can be related to the underlying processes governing rainfall.

6. Conclusion

Although the definition of a 'rain xc event' is relatively subjective, this study underlines the advantages of using event analysis rather than sample analysis. This data-driven analysis of events shows that rain events exhibit coherent features. As a consequence of the discrete and intermittent nature of rainfall, some of the features commonly used to describe rain processes are inadequate, in particular when they defined for a fixed duration. Excessively long integration times (hours or days) can lead to the mixing of observations that correspond to distinct physical processes, and also to the mixing of rainy and clear air periods, within the same sample. An excessively short integration time (seconds, minutes) leads to noisy data, which is sensitive to the sensor's characteristics (sensor area, detection threshold and noise). By analyzing entire rain events, rather than short individual samples of fixed duration, it is possible to clearly identify certain relationships between the different features of rain events, in particular the influence of the microphysical properties of rain on its macrophysical characteristics. This approach allows the intra-event variability caused by measurement uncertainties to be reduced, thus improving the accuracy with which physical processes can be identified.

Once an event has been clearly identified, it is possible to choose a small number of variables to describe it. We present a new data-driven approach, which can be used to select the most relevant variables for this characterization. This approach has generic properties and can be adapted to many multivariate applications. A genetic algorithm, when combined with Self-Organizing Map (SOM) clustering, can allow the unsupervised selection of an optimal subset of five macrophysical variables. This is achieved by minimizing a score function, which depends on the topology error of the SOM and the number of variables. This score provides a parsimonious description of the event, whilst preserving as much as possible the topology of the initial space.

Numerous variables derived mainly from rain rate recordings are used to describe precipitation in the context of rain time series studies, and a wide variety of topics of interest, including hydrology, meteorology, climate, and weather forecasting. The algorithm proposed in this study produces a subspace formed by only 5 of the 23 rain features described in the literature. We show that these five features can be selected by the algorithm in an unsupervised manner and, from the macrophysical point of view, can provide an adequate description of the main characteristics of rainfall events. These characteristics are: the event duration, the peak rain rate, the rain event depth, the standard deviation of the event rain rate, and the absolute rain rate variation of order 0.5.

In order to confirm the relevance of the five selected features, we analyze the corresponding SOM and are able to clearly reveal the presence of relationships between these features. This approach also reveals the independence of the inter-event time (IET_p) characteristic, and the weak dependence of the Dry percentage in event ($D_{d\%e}$) characteristic, thus confirming that a rain time series can be considered as an alternating series of independent rain events, interrupted by periods without rain. Hierarchical clustering allows the well-known separation between stratiform and convective events to be clearly identified. This dual classification is then refined into a set of five relatively homogeneous subclasses. The stratiform class is divided into 2 subclasses: a drizzle / very light rain subclass, and a normal event subclass. The convective class is divided into 3 subclasses, characterized by a strong temporal heterogeneity and significant rain rates.

As this research was based on the analysis of observations made in mid-latitude plains in France, the relevance of this classification remains to be confirmed through the analysis of datasets recorded in different climatic zones, and under different meteorological conditions, such as those encountered in mountainous or coastal areas. If the SOM described in the present study were learned with a more exhaustive dataset, a larger map would be produced, and this could reveal new types of rainfall behavior, which remained undetected in the current dataset. This point will be addressed in future studies.

The data-driven analysis of entire rain events (rather than the analysis of fixed-length samples) is relevant to the study of interactions between the macrophysical (based on the rain rate) and microphysical (based on raindrop) properties of rain. In the present study, several strong relationships were identified between these microphysical and macrophysical characteristics, and we show that some of the five subclasses identified in this analysis have specific microphysical characteristics. When a relationship between the microphysical and macrophysical properties of rain is identified, this can have many practical implications, especially for remote sensing. In the context of weather radar applications, the microphysical properties of rain are needed in order to estimate rain rates, through the use of the Z–R relationships. The estimation of microphysical rain characteristics, based on easily observable rain gauge measurements, could play a significant role in the development of the quantitative precipitation estimation (QPE).

Data availability. The dataset is available on request from the authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors wish to thank the team of SIRTA (Site Instrumental de Recherche par Télédétection Atmosphérique) as well as ACTRIS-FR for financial support.

Edited by: G. Vulpiani
Reviewed by: D. Dunkerley, A. Parodi, and three anonymous referees

References

Akrour, N., Chazottes, A., Verrier, S., Mallet, C., and Barthes, L.: Simulation of yearly rainfall time series at microscale resolution with actual properties: Intermittency, scale invariance, and rainfall distribution. Water Resources Research, 51(9), 7417–7435, 2015.

Atlas, D., Ulbrich, C.W., Marks, F. D., Amitai, E., and Williams, C. R.: Systematic variation of drop size and radar-rainfall relations, J. Geophys. Res.-Atmos., 104, 6155–6169, 1999.

Balme, M., Vischel, T., Lebel, T., Peugeot, C., and Galle, S. : Assessing the water balance in the Sahel: impact of small

scale rainfall variability on runoff Part 1: rainfall variability analysis. Journal of Hydrology 331:336–348, 2006.

Bringi, V. N., Chandrasekar, V., Hubbert, J., Gorgucci, E., Randeu, W. L., and Schoenhuber, M.: Raindrop size distribution in different climatic regimes from disdrometer and dual-polarized radar analysis, J. Atmos. Sci., 60, 354–365, 2003.

Brown, B.G., Katz, R. W., and Murphy, A.H.: Statistical analysis of climatological data to characterize erosion potential: 1. Precipitation Events in Western Oregon. Oregon Agricultural Experiment Station Spec. Rep. No. 689, Oregon State University (1983).

Brown, B.G., Katz, R. W., and Murphy, A.H.: Statistical analysis of climatological data to characterize erosion potential: 4. Freezing events in eastern Oregon/Washington. Oregon Agricultural Experiment Station Spec. Rep. No. 689, Oregon State University, 1984.

Brown, B.G., Katz, R. W., and Murphy, A.H.: Exploratory Analysis of Precipitation events with Implications for Stochastic Modeling. Journal of Climate and Applied meteorology(57-67), 1985.

Cosgrove, C.M. and Garstang, M.: Simulation of rain events from rain-gauge measurements. International Journal of Climatology 15, 1021–1029, 1995.

Coutinho, J.V., Almeida, C. Das, N., Leal. A.M. F., Barbarosa, L. R.: Characterization of sub-daily rainfall properties in three rain gauges located in northeast Brazil. Evolving Water Resources Systems: Understanding, Predicting and Managing Water–Society Interactions Proceedings of ICAR 2014, Bologna, Italy, 345-350, 2014.

Daumas, F. : Méthodes de normalisation de données, Revue de statistique appliquée, 30(4), 23-38, 1982.

Driscoll, E. D., Palhegyi, G. E., Strecker, E. W., & Shelley, P. E. (1989). Analysis of storm events characteristics for selected rainfall gauges throughout the United States. *US Environmental Protection Agency, Washington, DC*.

Dunkerley, D.: Rain event properties in nature and in rainfall simulation experiments: a comparative review with recommendations for increasingly systematic study and reporting, Hydrological Processes, 22(22), 4415–4435, 2008a.

Dunkerley, D.: Identifying individual rain events from pluviograph records: a review with analysis of data from an Australian dryland site, Hydrological Processes, 22(26), 5024–5036, 2008.

Eagleson, P. S.: Dynamic Hydrology, McGraw-Hill, 1970.

Galmarini, S., Steyn, D. G., and Ainslie, B.: The scaling law relating world point-precipitation records to duration. Int. J. Climatol., 24, 533–546, 2004.

Everitt, B.: Cluster Analysis. London: Heinemann Educ. Books, 1974.

Delahaye, J.-Y., Barthès, L., Golé, P., Lavergnat J., and Vinson, J.P.: a dual beam spectropluviometer concept, Journal of Hydrology, 328(1-2), 110-120, 2006.

Gargouri, E., Chebchoub, A.: Modélisation de la structure de dépendance hauteur-durée d'événements pluvieux par la copule de Gumbel. Hydrological Sciences–Journal–des Sciences Hydrologiques, 53(4), 802-817, 2010.

Grazioli, J., Tuia, D., and Berne, A.: Hydrometeor classification from polarimetric radar measurements: a clustering approach, Atmos. Meas. Tech., 8, 149-170, 2015.

Guyon, I. and Elisseeff, A.: An Introduction to Variable and Feature Selection (Kernel Machines Section), 3, 1157--1182, 2003.Haile, A. T., Rientjes, T. H. M., Habib, E., Jetten, V., and Gebremichael, M.: Rain event properties at the source of the Blue Nile River, Hydrol. Earth Syst. Sci., 15, 1023-1034, 2011.

Iguchi, T., Kozu, T., Kwiatkowski, J., Meneghini, R., Awaka, J., and Okamoto, K.: Uncertainties in the rain profiling algorithm for the TRMM precipitation radar. J. Meteor. Soc. Japan, 87A, 1-30, 2009.

Holland, J. H.: Adaptation In Natural And Artificial Systems, University of Michigan Press, 1975.

Kohonen, T.: Self-organizing formation of topologically correct feature maps. Biological Cybernetics, 46, 59-69, 1982.

Kohonen, T. (2001). Self-Organizing Maps. Springer-Verlag, ISBN 3-540-67921-9, New York, Berlin, Heidelberg, 2001

Larsen, M. L. and Teves, J. B.: Identifying Individual Rain Events with a Dense Disdrometer Network, Advances in Meteorology, 2015, ID582782, 2015.

Lavergnat, J. and Golé, P.: A Stochastic Raindrop Time Distribution Model. Journal of Applied Meteorology, 37, 805-818, 1998.

Lavergnat, J. and Golé, P.: A stochastic model of raindrop release: Application to the simulation of point rain observations, Journal of Hydrology, 328(1), 8-19, 2006.

Liu, Y., Weisberg, R. H. and Mooers, C. N. K.: Performance evaluation of the self- organizing map for feature extraction, Journal of Geophysical Research, 111, C05018, doi:10.1029/2005JC003117, ISSN 0148-0227, 2006

Liu, Y. and Weisberg R.H.: A review of self-organizing map applications in meteorology and oceanography. In: Self-Organizing Maps-Applications and Novel Algorithm Design, 253-272, 2011.

Llasat, M.C.: An objective classification of rainfall events on the basis of their convective features. Application to rainfall intensity in the north east of Spain. International Journal of climatology, 21, 1385-1400, 2001.

Marzuki, M., Hashiguchi, H., Yamamoto, M. K., Mori, S., and Yamanaka, M. D.: Regional variability of raindrop size distribution over Indonesia, Ann. Geophys., 31, 1941-1948, 2013.

Molini, L., Parodi, A., Rebora, N., Craig, G. C.: Classifying severe rainfall events over Italy by hydrometeorological and dynamical criteria. Quarterly Journal of the Royal Meteorological Society, 137(654), 148-154, 2011.

de Montera, L., Barthes, L., and Mallet, C.: The effect of rain-no rain intermittency on the estimation of the Universal Multifractal model parameters, J. of Hydrometeorology, 10, pp. 493–506, 2009.

Moussa, R. and Bocquillon, C.: Caractérisation fractale d'une série chronologique d'intensité de pluie. Rencontres hydrologiques Franco-Romaines, 363-370, 1991.

Suh, S.-H., You, C.-H., and Lee, D.-I.: Climatological characteristics of raindrop size distributions in Busan, Republic of Korea, Hydrol. Earth Syst. Sci., 20, 193-207, 2016.

Tapiador, F. J., Checa, R., and de Castro, M.: An experiment to measure the spatial variability of rain drop size distribution using sixteen laser disdrometers. Geophysical Research Letters, 37(16), ID L16803, 2010.

Testud, J., S., Oury, P., Amayenc, and Black, R. A.: The concept of ''normalized'' distributions to describe raindrop spectra: A tool for cloud physics and cloud remote sensing, J. Appl. Meteor., 40, 1118–1140, 2001.

Ulaby, F. T., Moore, R. K., and Fung, A. K.: Microwave Remote Sensing: Fundamentals and Radiometry. Vol. I.

Artech House, 321-327, 1981.
Uriarte, E. A. and Martín, F. D., Topology Preservation in SOM, World Academy of Science, Engineering and Technology. International Journal of Computer, Electrical, Automation, Control and Information Engineering 2, 9, 2008Verrier, S., Barthès L., Mallet C.: Theoretical and empirical scale dependency of Z-R relationships: Evidence, impacts, and correction, Journal of Geophysical Research: Atmospheres, 118 (14), 7435-7449, 2013.
Vesanto, J. and Alhoniemi, E.: Clustering of the self-organizing map. IEEE Transactions on Neural Networks, Vol. 11, 586–600, ISSN 1045-9227, 2000

2.3. Conclusion et synthèse

L'approche par caractéristiques montre qu'il est possible de définir un ensemble de variables macro physiques susceptibles de représenter au mieux les événements de pluie. Bien entendu nous ne prétendons pas que les cinq variables retenues constituent un sous ensemble unique. Par ailleurs, l'ensemble de départ considéré qui est constitué de 23 variables est très certainement incomplet. En effet nous avons considéré dans cette étude des points de vue plutôt centrés sur des problématiques météorologiques et hydrologiques, on aurait pu cependant considérer d'autres points de vue tels que l'agriculture ou la sécurité routière qui peuvent ajouter d'autres variables d'intérêt à notre ensemble de départ.

Cette étude a permis de mettre au point un algorithme basé sur une méthode neuronale d'analyse multidimensionnelle non supervisée qui réalise une typologie des évènements de pluie en les regroupant par similitude. Cet algorithme permet d'obtenir une caractérisation parcimonieuse des événements pluvieux. De manière plus générale, une façon de réduire la dimensionnalité des données sans perdre leur variabilité est de les représenter sous forme de classes en les regroupant par similitude. Les classes permettent de caractériser des situations types. Elles permettent aussi d'inférer des paramètres non observables (par l'absence d'instrument de mesure adéquat par exemple) en les associant à chacune des classes. Dans notre cas nous avons montré qu'on peut déduire des informations concernant les propriétés microphysiques des évènements en fonction de leur classe, elle-même déterminée à partir d'un jeu réduit (5) de caractéristiques macro-physiques.

Comme il a était dit dans l'introduction du chapitre, l'objectif a été de vérifier que l'analyse basée sur une approche statistique permettait de décrire la physique de la pluie. Retrouver les deux types d'évènements tels que classiquement défini par les experts avec une partition en deux classes était un moyen de valider la technique (non supervisée). Or, l'analyse du déploiement de la carte (figure 3) et des différentes projections sur l'espace des caractéristiques (figures 5,7 et 9) laisse à croire que l'on peut passer d'un événement et/ou classe à l'autre de façon relativement continue (un continuum)[7]. On peut en effet signaler des similitudes de caractéristiques entre des événements de classes voisines. Dans ce cas (continuum), on parle souvent d'ordination de l'espace, les techniques d'analyse tendent à proposer une relation d'ordre qui décrit l'espace continu. **Cependant, si une classification/partition est exigée (pour des besoins de séparation et/ou focalisation), le**

[7] La notion de continuum est détaillée dans la partie 5.3.1 du chapitre 5.

dendrogramme (figure 6) suggère une classification optimale en trois classes d'évènements plus homogènes qui agrège les deux sous-classes 1 et 2 dans une première classe regroupant 80% des événements, les deux sous-classes 3 et 4 dans une deuxième classe avec 8% des événements, et enfin isole la sous-classe 5 avec 12% des événements. Cette partition en trois classes obtenue sur les cinq caractéristiques est plus ou moins corroborée par la projection des caractéristiques microphysiques de la pluie (figure 9). Cette dernière (figure 9) propose une ordination intéressante de l'espace des événements décrite par les indices des classes. La question du nombre optimal de classes de pluie et la recherche d'une partition optimale est traitée dans les chapitres 5 et 6.

L'article a été cité dans plusieurs articles pour la méthodologie proposée (Alghamdi H. M. et Selamat A., 2019) ou pour des études relatives à l'analyse de la variabilité régionale des précipitations à des échelles infra-horaires (Pohle I. et al., 2018). L'article de Barbosa (Barbosa et al., 2018) utilise l'approche par événement pour une étude de la région Nord du Brésil, en effectuant la statistique descriptive de caractéristiques des événements de pluie pour différentes conditions climatiques. Dans leur article Parchure et Gedam (Parchure A. S. et Gedam S. K., 2018) analysent 23 000 évènements observés dans la métropole de Mumbay à l'aide d'une carte de Kohonen, ils distinguent six catégories liées à la topographie complexe de la région.

Chapitre 3 : Etudes des observations pluviométriques issues de pluviomètres à auget

3.1. Introduction

La caractérisation des événements de précipitations et la classification réalisées au chapitre précédant utilisent des données de disdromètres avec un temps d'agrégation d'une minute. Cette résolution permet une détermination fine des dates de début et de fin des évènements et permet de bien discriminer les évènements pluvieux des périodes de sécheresse. Cependant, les données de disdromètres à la résolution de la minute sont relativement rares alors qu'au contraire les observations à base de pluviomètres sont abondantes. Dans ce chapitre nous étudions la possibilité d'étendre aux données issues de pluviomètres à auget les résultats obtenus à partir de disdromètres. Plus précisément, il s'agit de répondre à la question :

Les résultats obtenus et présentés au chapitre précédent à l'aide d'observations de disdromètres agrégées à la minute sont-ils généralisables pour des observations de pluviomètres à auget basculant agrégées à une résolution temporelle plus grossière ?

En d'autres termes : la définition d'un événement et sa caractérisation à l'échelle de la macro physique à l'aide des cinq paramètres précédemment définis, est-elle dépendante du type d'instrument et du choix du temps d'agrégation ?

Comme cela a été décrit au chapitre 1, les pluviomètres à auget basculant et les disdromètres ont des principes de mesures radicalement différents. On rappelle que dans le cas de pluviomètres à auget, la quantité d'eau mesurée est discrétisée à la différence des disdromètres qui mesurent les gouttes individuellement ce qui facilite le calcul de l'intensité de pluie y compris pour des résolutions fines, donnant ainsi un effet de continuité.

Avant d'étudier la possibilité de généraliser la description à cinq paramètres, nous allons tout d'abord mesurer l'impact du changement d'instrument sur la qualité des séries mesurées. Autrement dit : Quelle conséquence la discrétisation induite par les augets a-t-elle sur la qualité des séries temporelles mesurées?

3.2. Simulation de séries temporelles de précipitations « pseudo-pluviomètre » à auget

Du fait de la forte variabilité spatiale et temporelle de la pluie et de son caractère discret (succession de gouttes), on peut observer de façon sporadique des différences notables entre deux instruments rigoureusement identiques (idéalement) situés côte à côte. Ces différences pourront être d'autant plus marquées que le temps d'agrégation est faible. En effet, la présence de turbulences peut localement perturber le flux de gouttes à la verticale d'un appareil de mesure, cela a un effet particulièrement important sur les petites gouttes. De même, la mesure des grosses gouttes qui sont les plus rares est sujette à une erreur statistique importante dès lors que le temps d'agrégation et/ou la surface de collecte est faible. Il est donc délicat pour étudier l'effet de la discrétisation introduite par les augets de comparer directement les observations co-localisées d'un disdromètre et d'un pluviomètre à auget. Pour dissocier l'effet dû à la discrétisation de celui des autres effets, des séries temporelles de « pseudo-pluviomètres » (notées PP) ont été simulées à partir de séries temporelles issues du disdromètre du SIRTA. Ces séries ont été réalisées pour différents volumes d'auget et plus spécialement avec des « volumes » correspondant à des hauteurs d'eau de 0,1 et 0,2 mm qui sont les volumes les plus courants ($v = 0{,}1\ ou\ 0{,}2\ mm$ pour les pluviomètres SIRTA et $v = 0.2\ mm$ pour les pluviomètres Météo-France). Ainsi, seul le phénomène de discrétisation dû aux augets est pris en compte à l'exclusion de tout autre phénomène (évaporation, vent …).

Pour simuler un pseudo-pluviomètre à auget basculant de « volume » v, on cumule dans le temps le volume d'eau apporté par chaque goutte. Dès que ce dernier atteint la hauteur équivalente au volume d'auget v on génère une impulsion. Dans un premier temps, les temps de basculement sont enregistrés en respectant le format de Costello et Williams (1991) pour garder toute information sur les basculements. Ensuite, ce format est converti au format Sadler et Busscher (1989) pour rendre la série comparable aux séries agrégées. Le même algorithme permet la simulation d'un pseudo-pluviomètre à pesée en prenant v égal à la précision du pluviomètre à pesée ($v = 0.01$mm pour une version commerciale[8]). Par la suite on appelle volume d'auget v le volume d'auget d'un pluviomètre à auget et/ou la précision d'un pluviomètre à pesée. L'ordre de grandeur informe sur le type d'appareil. Par convention, on notera $v = 0\ mm$ le volume « d'auget » d'un disdromètre.

Même si nous avons souligné précédemment qu'il était délicat de comparer directement deux instruments co-localisés, nous avons toutefois comparé les observations de nos pseudo-

[8] www.ott.com/fr-fr/produits/accessoires-180/ott-pluvio2-175/

pluviomètres avec celles obtenues à l'aide de deux pluviomètres à auget situés sur le même site (distance 20 et 200 mètres du disdromètre), l'idée étant non pas de réaliser une comparaison fine mais tout au moins de vérifier que les ordres de grandeurs obtenues étaient identiques. La comparaison est réalisée sur la période du 1 janvier 2012 au 31 décembre 2013.

Le tableau 3.1 présente des caractéristiques déduites des deux séries issues des deux pluviomètres du SIRTA et des séries pseudo-pluviomètres pour $v_1 = 0{,}1\ mm$ et $v_2 = 0{,}2\ mm$. Les caractéristiques ont été estimées pour un pas de temps $T = 1\ min$ qui correspond à la résolution native des observations des pluviomètres collectées par le SIRTA.

Pour un instrument donné, il est intéressant de comparer les caractéristiques obtenues pour les deux volumes d'auget. En effet certaines caractéristiques sont peu sensibles à une variation du volume, c'est le cas notamment des intensités maximales et dans une moindre mesure l'écart type des intensités. L'explication tient simplement au fait qu'en période de forte pluie le temps de remplissage d'un auget est rapide et largement inférieur au temps d'agrégation (1 minute) et cela aussi bien pour un auget de 0,1 mm que pour un auget de 0,2 mm. A contrario, l'occurrence de pluie est extrêmement sensible au volume des augets avec une diminution de près de 40% de celle-ci lorsqu'un passe d'un auget de 0,1 mm à 0,2 mm. Ici au contraire ce sont les pluies faibles qui vont gouverner cette différence et pour lesquelles le temps de remplissage des augets est grand comparé au temps d'agrégation. Dans ces conditions, la probabilité qu'il y ait zéro basculement d'auget en présence de pluie est nettement supérieure avec un « gros » volume d'auget. Cela était bien évidemment beaucoup moins vrai en présence de pluies fortes pour lesquelles il y a plusieurs basculements durant le temps d'agrégation. Concernant la caractéristique « valeurs moyennes sur les échantillons pluvieux », il existe pratiquement un rapport deux sur cette variable entre un auget de volume v_1 de 0,1 mm et v_2 de 0,2 mm. On peut remarquer que s'il n'y avait eu qu'un seul basculement par temps d'agrégation, nous aurions obtenu exactement un rapport deux. On déduit donc que sur un temps d'agrégation d'une minute il y a de façon très majoritaire un seul basculement.

Concernant les différences pluviomètres – pseudo-pluviomètres on remarque que les écarts dépendent de la caractéristique considérée. Ainsi la caractéristique « valeur moyenne sur les échantillons pluvieux» varie très peu (2,6%). Cela reste vrai dans une certaine mesure pour les écarts types avec des différences de l'ordre de 10 à 15%. On remarque également qu'il pleut plus souvent pour un pluviomètre que pour un pseudo-pluviomètre. On peut imaginer que d'une part certains types d'hydrométéores ne sont pas pris en compte par le disdromètre comme par exemple la bruine légère (constituée de gouttelettes de diamètres inférieurs à la sensibilité de l'appareil), le givre ou la condensation sur les parois du collecteur du pluviomètre qui finissent

tôt ou tard par ruisseler au fond du collecteur ainsi que la neige qui est systématiquement enlevée des observations du disdromètre.

	Pluviomètres		Pseudo − pluviomètres	
Volume d'auget v	0,1 mm	0,2 mm	0,1 mm	0.2mm
Occurrence de pluie [%]	0.91	0.56	0.72	0.39
Valeur max sur les échantillons pluvieux $[mm.h^{-1}]$	138	132	90	96
Valeur moyenne sur les échantillons pluvieux $[mm.h^{-1}]$	7.45	13.47	7.33	13.46
Ecart type sur la série $[mm.h^{-1}]$	0,94	1.15	0.81	0.96
Ecart type sur les échantillons $[mm.h^{-1}]$	6,55	7,32	5,87	6,56

Tableau 3.1. Statistiques des séries de pluviomètres et de Pseudo-pluviomètres à la résolution d'une minute et pour 2 volumes d'augets

Les courbes de densité de probabilité de hauteur d'eau des pluviomètres et des pseudo-pluviomètres sont présentées à la figure 3.1. Globalement les densités de probabilité des séries pseudo-pluviomètres sont très proches de celles issues des pluviomètres pour un volume d'auget donné. Les faibles écarts observés ne sont pas significatifs et nous concluons de ces diverses comparaisons que les séries de pseudo-pluviomètres ont un sens et qu'elles peuvent être utilisées pour la suite de l'étude.

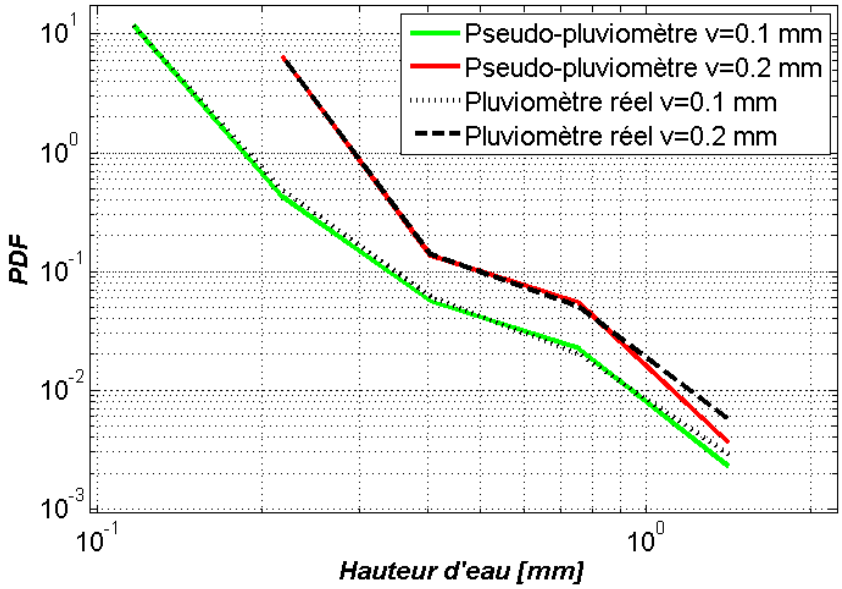


Figure 3.1.Densité de probabilité des hauteurs d'eau en millimètres pour les deux pluviomètres du SIRTA (traits pointillés) et les deux pseudo-pluviomètres (traits continus) à la résolution d'une minute. Les couleurs correspondent aux volumes d'auget

3.3. Relation volume d'auget / temps d'agrégation

Dans cette partie on étudie l'impact de l'agrégation temporelle et du volume d'auget sur les caractéristiques des séries chronologiques d'intensité de pluie. Pour cela, on compare les statistiques des séries des pseudo-pluviomètres pour différents volumes v (rappelons que le disdromètre correspond au volume d'auget nul $v = 0\ mm$) et différents temps d'agrégation T.

3.3.1. Occurrence de pluie

Un intervalle de temps est dit pluvieux pour une surface de collecte donnée si une quantité d'eau précipitante supérieure à un seuil donné a été mesurée durant le temps d'agrégation T considéré (Hubert et Carbonnel, 1989). En présence de précipitations faibles le remplissage d'un auget prend du temps et peut être supérieur au temps d'agrégation. Ce phénomène est plus ou moins gênant suivant le temps d'agrégation considéré et l'intensité de la pluie. On définit l'état sec lorsque la hauteur d'eau recueillie est inférieure au volume d'auget donné, et l'état pluvieux dans le cas contraire. La figure 3.2 présente pour le disdromètre et les deux pseudo-pluviomètres le pourcentage d'occurrence de la pluie en fonction du temps d'agréation T pour des durées d'agrégation comprises entre 1 minute et 6 jours. Au fur et à mesure de l'augmentation du temps d'agrégation le pourcentage d'occurrence augmente pour atteindre 100 % au-delà d'un temps d'agrégation de quelques dizaines de jours (il pleut au moins une fois sur une période de 10 à 20 jours en île de France). On peut d'ores et déjà noter qu'au-delà d'un temps d'agrégation d'une journée, le volume de l'auget joue assez peu. A l'inverse, lorsqu'on travaille avec des temps d'agrégation courts la taille des augets est prépondérante, ainsi à la résolution d'une minute on passe d'une occurrence d'environ 0,4 % pour un volume $v = 0{,}2\ mm$ à 4,28 % pour un volume nul $v = 0mm$ (disdromètre).

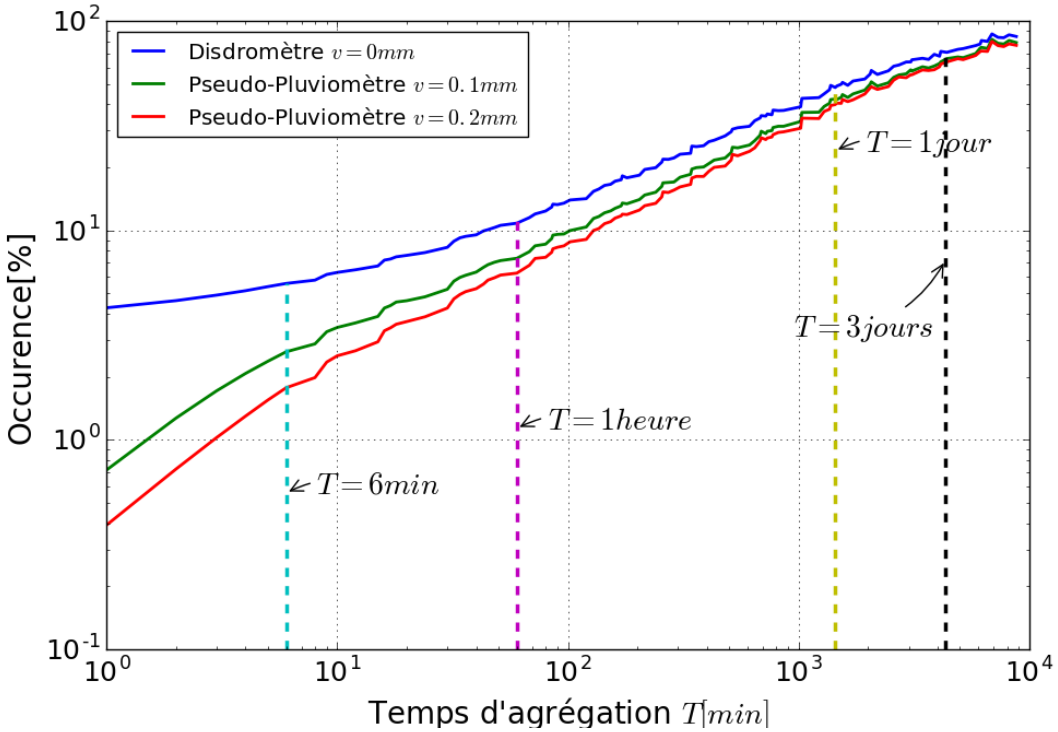


Figure.3.2 Occurrence de pluie observée entre le 1 janvier 2012 au 31 décembre 2013 sur le site du SIRTA *série observée par le disdromètre (v=0 mm) et les pseudo-pluviomètres (v=0.1 mm et v= 0.2 mm) pour différents temps d'agrégations de 1 minute à une semaine.*

volume d'agrégation v \temps d'agrégation T	1min	6min	1 heure	1 jour	3 jours	7 jours
Disdromètre $v = 0\ mm$	4.28%	5.59%	11.37%	48.02%	71.81%	90.97%
Pseudo-pluviomètre $v = 0.1\ mm$	0.72%	2.64%	7.84%	41.86%	66.07%	86.18%
Pseudo-pluviomètre $v = 0.2\ mm$	0.39%	1.77%	6.73%	39.81%	64.43%	85.23%

Tab.3.2 occurrence de pluie observée entre le 1 janvier 2012 au 31 décembre 2013 sur le site du SIRTA *observée par le disdromètre et séries simulées (v=0.1 mm et v= 0.2 mm) pour différents temps d'agrégations (T de 1 min à une semaine)*

Le tableau 3.2 résume les pourcentages d'occurrence entre le disdromètre et les deux pseudo-pluviomètres. On note que le pourcentage d'occurrence décroit pour des volumes d'auget importants et ce pour toutes les valeurs de T. Pour des observations horaires par exemple, le disdromètre indique un pourcentage d'occurrence de 11,37% alors que l'on note 6,73% pour un volume d'auget $v = 0{,}2\ mm$ et 7,84% pour un volume d'auget $v = 0{,}1\ mm$. Les écarts sont d'autant plus important que T est faible (pour $T = 1\ min$, le disdromètre observe 4,28% alors que l'on note 0,39% pour $v = 0{,}2\ mm$ et 0,72% pour $v = 0{,}1\ mm$). Pour un temps d'agrégation de l'ordre d'une à deux heures les écarts entre les pseudo-pluviomètres deviennent négligeables alors que les écarts relatifs entre pseudo-pluviomètres et disdromètre restent significatifs jusqu'à des durées d'un à deux jours, puis s'atténuent au-delà.

3.3.2. Maximum des intensités de pluie

La figure 3.3 représente l'intensité de pluie maximale en fonction du temps d'agrégation pour le disdromètre et les deux pseudo-pluviomètres. Comme on peut le constater, les trois courbes sont confondues, la valeur maximale est donc peu sensible au volume d'auget utilisé. Comme expliqué plus haut, on peut en effet supposer qu'en présence de pluie forte le remplissage des augets est rapide et, est largement inférieur au temps d'agrégation *T* quel que soit le volume de l'auget considéré (0,1 ou 0,2 mm). L'intensité maximale des précipitations décroit rapidement au fur et à mesure que le temps d'agrégation augmente traduisant ainsi la forte variabilité temporelle des précipitations. Enfin, on peut noter une queue de distribution en loi puissance caractéristique des processus de type multifractal.

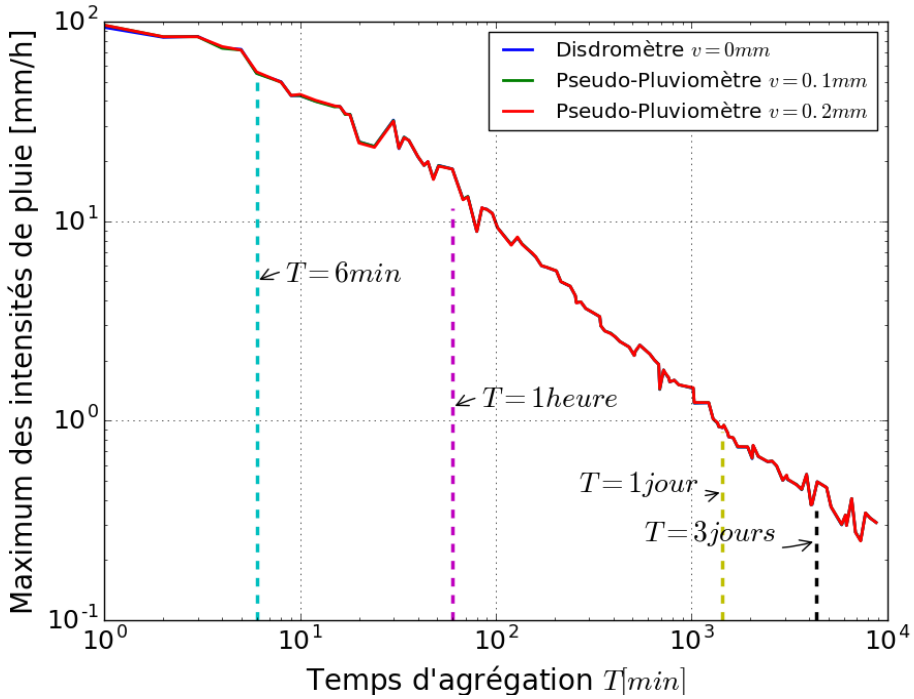


Figure 3.3 Valeurs maximales observées entre le 1 janvier 2012 au 31 décembre 2013 sur le site du SIRTA *par le disdromètre et les pseudo-pluviomètres en fonction du temps d'agrégation T*

3.3.3. Distribution des hauteurs d'eau

La figure 3.4 représente la densité de probabilité de hauteur d'eau cumulée *H* (conditionnée à $H > 0$) recueillie par les trois instruments (le disdromètre et les deux pseudo-pluviomètres) pour des valeurs de *T* d'une minute, une heure et un jour. Concernant les faibles cumuls, seul le disdromètre peut les observer. Pour les pluviomètres à augets les faibles cumuls ne font pas basculer l'auget, ces cumuls sont donc redistribués à des dates ultérieures pour lesquelles le cumul est suffisant pour faire basculer l'auget. En particulier, la fréquence relative du cumul égal au volume de l'auget est surreprésentée pour des temps d'agrégation d'une

minute. Dans tous les cas de figure, les courbes des pseudo-pluviomètres sont au-dessus de celles issues du disdromètre. Cependant l'écart se réduit notablement lorsque le temps d'agrégation augmente, celui-ci devenant négligeable pour une durée T d'un jour. Cet écart peut simplement s'expliquer en considérant le théorème de Bayes :

$$P(\,H > Seuil/H > 0) = \frac{P(H > Seuil)}{P(H > 0)} \quad (3.1)$$

A la résolution d'une minute, l'ensemble des densités de probabilité de hauteur d'eau supérieures au volume d'auget sont surestimées d'un facteur 16 ($v = 0{,}2\,mm$) et 8 ($v = 0{,}1\,mm$) qui s'explique en partie par les facteurs 11 et 6 sur l'occurrence de pluie à cette résolution. A la résolution équivalente à $T = 6min$ (non représentée) cette surestimation, causée artificiellement par le non basculement de l'auget aux pas de temps précédents, conduit à un facteur d'ordre 2 ou 3 suivant le volume d'auget pour l'ensemble des hauteurs. Ce facteur est de l'ordre de 1,6 ou 1,8 pour un pas de temps horaire et se réduit à 1,2 au pas de temps journalier. Cette augmentation des hauteurs d'eau observables par les pluviomètres est à mettre en regard de la diminution de l'occurrence de pluie (cf. section 3.3.1 de ce chapitre). Pour les résolutions supérieures ou égales à 6 min les facteurs observés sont pratiquement identiques à ceux observés sur l'occurrence de pluie (Tab 3.2).

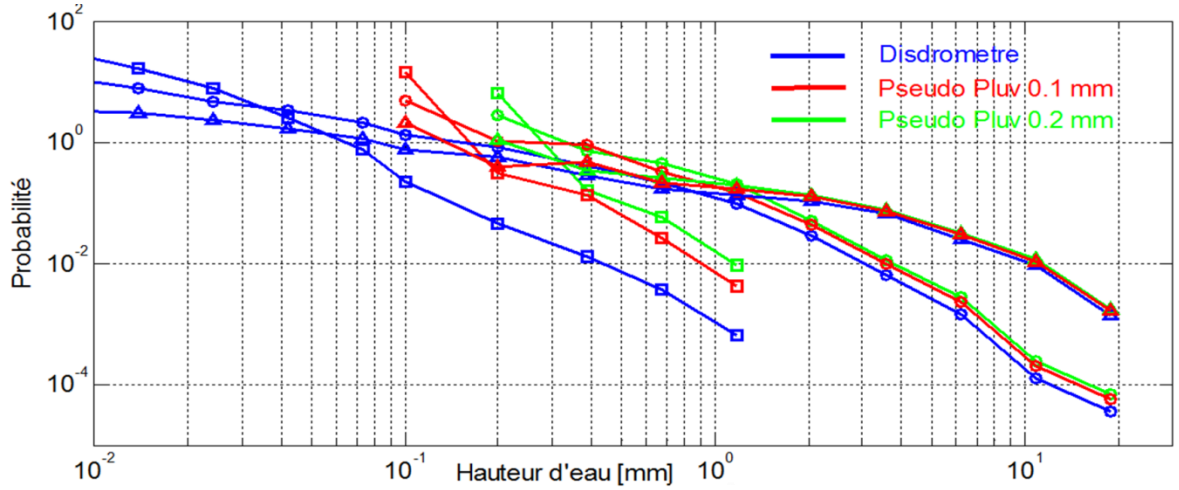


Figure 3.4 Densité de probabilité des hauteurs d'eau. Les couleurs correspondent aux différents volumes d'auget v, les symboles aux différents temps d'agrégation temporelle T : 1 min (□), 1 heure (○) et 1 jour (△)

3.3.4. Variabilité des séries temporelles

La figure 3.5 représente l'écart type de l'intensité de pluie en fonction de la durée d'agrégation T pour les 3 types d'appareil. On remarque que l'écart type est indépendant du

volume d'auget lorsque le temps d'agrégation est supérieur à quelques dizaines de minutes. Des études concernant les propriétés d'invariance d'échelle de la pluie (Verrier et al., 2011) ont montré qu'il existait plusieurs régimes qui se traduisent par des cassures visibles sur le spectre des séries. Une première cassure est située aux alentours de 30 minutes et définit un régime d'invariance d'échelle compris entre quelques secondes et 30 minutes. Ce régime caractérise les relations statistiques inter-échelles à l'intérieur même des événements de pluie alors qu'aux pas de temps plus importants le régime est régi en bonne partie par l'intermittence des précipitations (variabilité inter-évènement) (Verrier et al., 2011, Akrour et al., 2015). Au-delà de 30 minutes l'écart type reflète donc principalement la variation des hauteurs d'eau cumulées sur la totalité des évènements de pluie, il s'agit d'une variabilité inter-évènements qui ne dépend pas du volume d'auget utilisé. En deçà de 30 min l'écart type caractérise la variabilité intra-événement. Cependant, dans le cas des pluviomètres à augets, l'effet de quantification dû aux augets qui introduisent des valeurs nulles supplémentaires correspondant aux pas de temps au cours desquels l'auget n'a pas eu le temps de se remplir, augmente artificiellement la variabilité observée. Cela est bien visible sur la partie gauche de la figure 3.4.

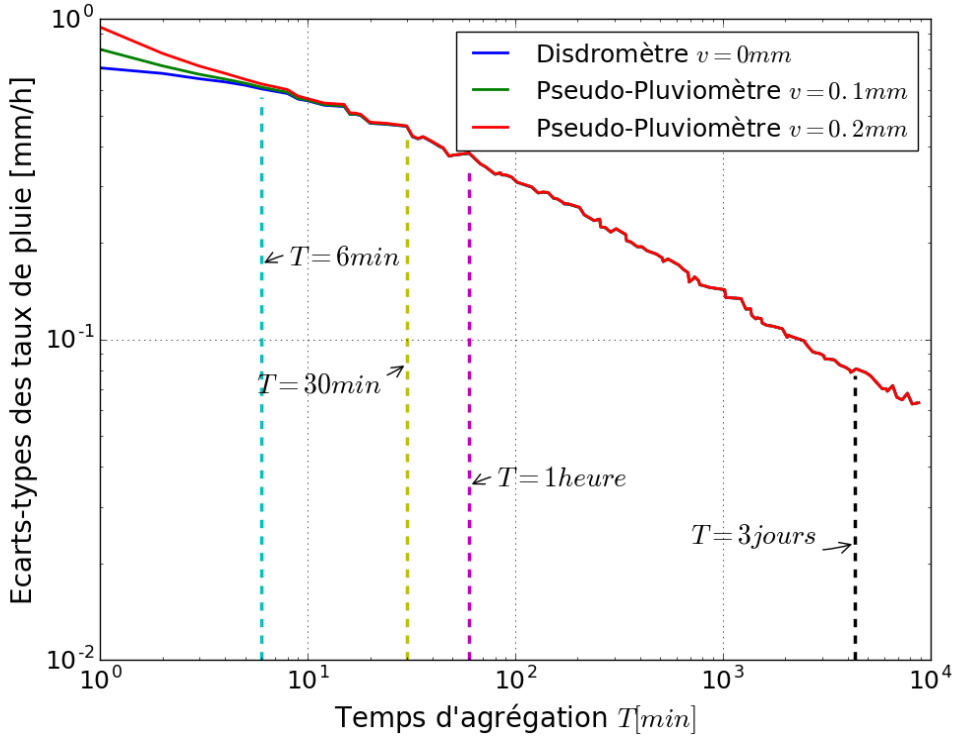


Figure.3.5 Ecart-types des intensités de pluies observées entre le 1 janvier 2012 au 31 décembre 2013 sur le site du SIRTA estimés par le disdromètre et les pseudo-pluviomètres pour différents temps d'agrégations (1 minute à une semaine)

3.4. Influence du volume d'auget: Quel temps d'agrégation pour un volume d'auget donné ?

Les résultats précédents montrent qu'en fonction de la caractéristique considérée (occurrence, distribution, valeur maximale, écart-type) celle-ci devient indépendante du volume de l'auget au-delà d'un certain temps d'agrégation. Ainsi quelle que soit la caractéristique, pour un temps d'agrégation T supérieur à un jour, l'utilisation d'un disdromètre ou d'un pluviomètre avec un volume d'auget $v = 0{,}1\ mm$ fournit des résultats similaires. Il se pose donc la question suivante :

Pour un volume d'auget v, peut-on définir quantitativement un temps d'agrégation $T(v)$ de sorte que les caractéristiques calculées à partir de la série agrégée non discrétisée soient « proches » de celles calculées sur la série non agrégée discrétisée ?

La recherche d'une réponse à cette question amène à se poser les questions suivantes :

- Le temps d'agrégation $T(v)$ est-il unique ? (injection)
- La relation entre le volume d'auget v et le temps d'agrégation $T(v)$ est-elle bijective ?
- Qu'elle est la robustesse de la relation $T(v)$

La partie suivante s'attache à rechercher cette relation entre T et v et les propriétés qui en découlent.

3.4.1. Construction de la fonction objective

Afin de répondre à la question : « **quel temps d'agrégation pour un volume d'auget donné ? »,** on se propose de comparer les hauteurs d'eau cumulées dans le temps $H_{cum}(t)$ suivant deux cas de figure :

1. On simule un pluviomètre à auget à partir du disdromètre : La série d'intensité de pluie est discrétisée à l'aide d'un volume d'auget fixé et on note les dates de basculement

2. On intègre la série du disdromètre de durée L suivant un pas de temps fixe (temps d'agrégation).

Soit un réel a vérifiant $0 < a < L$.

$$H_{cum}(a) = \int_0^a RR(t)\,dt = \int_0^k RR(t)\,dt + \int_k^a RR(t)\,dt \quad (3.2)$$

Plaçons-nous dans le premier cas, c'est à dire d'un instrument constitué d'un auget de volume v mais sans agrégation temporelle et prenons k l'instant correspondant au dernier basculement précédent a, et notons-le $k^v(a)$.

La quantité $\int_0^{k^v(a)} RR(t)\,dt$ représente la hauteur d'eau cumulée estimée par le pluviomètre et on peut écrire :

$$H_{cum}^v(a) = \int_0^{k^v(a)} RR(t)\,dt \quad (3.3)$$

$H_{cum}^v(a)$ est l'approximation de $H(a)$ compte tenu d'une discrétisation de volume d'auget v :

$$H_{cum}(a) \cong H_{cum}^v(a) \quad (3.4)$$

De plus cette discrétisation suppose que $H_{cum}^v(a)$ est un multiple de v. Donc :

$$\exists\, n \in \mathbb{N}, \quad H_{cum}^v(a) = n(a)v \quad (3.5)$$

$n(a)$ représente le nombre de basculements antérieurs à a.

On note $\varepsilon^v(a) = \int_{k^v(a)}^{a} RR(t)\,dt$ le cumul d'eau non détecté par le pluviomètre, il représente la quantité d'eau restante dans l'auget entre l'instant du dernier basculement $k^v(a)$ et l'instant a. On a :

$$\varepsilon^v(a) < v \quad (3.6)$$

Plaçons-nous maintenant dans le cas d'une agrégation temporelle avec un temps d'agrégation T et fixons la valeur de k égale au nombre de pas de temps de durée T enregistrés pendant la durée totale a:

$$k^\lambda(a) = a\ div\ T \quad (3.7)^9$$

Par conséquent,

$$\exists\, m \in \mathbb{N}, \quad k^\lambda(a) = mT \quad (3.8)$$

La quantité $\int_0^{k^\lambda(a)} RR(t)\,dt$ représente la hauteur d'eau cumulée estimée en considérant l'agrégation temporelle et on peut écrire :

$$H_{cum}^\lambda(a) = \int_0^{k^\lambda(a)} RR(t)\,dt \quad (3.9)$$

[9] div est l'opérateur de division entière

$H_{cum}^{\lambda}(a)$ est une approximation de $H(a)$ avec un temps d'agrégation T:

$$H_{cum}(a) \cong H_{cum}^{\lambda}(a) \quad (3.10)$$

La fonction $H_{cum}^{\lambda}(a)$ est par construction une fonction en escalier, chaque marche ayant une largeur T. Dans ces conditions, l'équation 3.9 s'écrit sous la forme :

$$H_{cum}^{\lambda}(a) = \sum_{i=1}^{k^{\lambda}(a)} RR_i^{\lambda} \times T \quad (3.11)$$

De façon similaire, on note $\varepsilon^{\lambda}(a)$ le cumul d'eau non considéré par l'agrégation temporelle. On a :

$$\varepsilon^{\lambda}(a) = \int_{k^{\lambda}(a)}^{a} RR(t)\, dt \quad (3.12)$$

A partir des équations 3.4 et 3.10 on déduit :

$$\forall a > 0, \quad H_{cum}^{v}(a) \cong H_{cum}^{\lambda}(a) \quad (3.13)$$

En injectant dans l'équation 3.13 les équations 3.5 et 3.11 :

$$\forall a > 0, \quad n(a)v \cong T \times \sum_{i=1}^{k^{\lambda}(a)} RR_i^{\lambda} \quad (3.14)$$

L'équation 3.14 est fonction des paramètres temps d'agrégation T et volume d'auget v. Elle constitue la contrainte principale pour l'étude de la relation entre le temps d'agrégation T et le volume d'auget v.

Une solution simple pour résoudre l'équation 3.14 consiste à rechercher les couples (T, v) qui minimisent la somme des erreurs quadratiques sur l'ensemble de la série, notée $f(T, v)$:

$$f(T, v) = \int_0^L \left(T \times \sum_{i=1}^{k^{\lambda}(a)} RR_i^{\lambda} - n(a)v \right)^2 da \quad (3.15)$$

En pratique, soit on cherche $T(v)$ pour un volume d'auget v donné qui minimise l'équation 3.15, soit on cherche $v(T)$ pour un temps d'agrégation T donné qui minimise l'équation 3.15.

3.4.2. Optimisation de la fonction objective

Une recherche séquentielle a été menée pour la minimisation de l'équation 3.15 consistant à faire varier le volume d'auget v de 0,001 à 0,5 mm avec un pas de 0,001 mm, puis de 0,5 à 3 mm avec un pas de 0,01 mm et le temps d'agrégation T de 1 à 10800 secondes (3 heures) avec un pas d'une seconde, puis de trois heures à un jour avec un pas de temps d'une minute. La figure 3.6 montre les couples temps d'agrégation / volumes d'auget optimaux obtenus à partir de notre série de référence de 2 ans issue du disdromètre. Il apparait clairement que pour des temps d'agréation inférieurs à une vingtaine de minutes, il est nécessaire de disposer d'augets de très faible volume inférieurs à 0,05 mm. Au-delà de 20 minutes d'agrégation, les choses ne sont plus aussi nettes et il apparait une dispersion importante des volumes en fonction du temps d'agrégation. L'analyse de l'allure de l'équation 3.15 a montré que le minimum de la fonction $f(T, v)$ est peu marqué et relativement bruité, expliquant ainsi cette dispersion. Toutefois il apparait clairement une borne supérieure du volume pour un temps d'agrégation donné. La droite verte de la figure 3.6 représente cette borne.

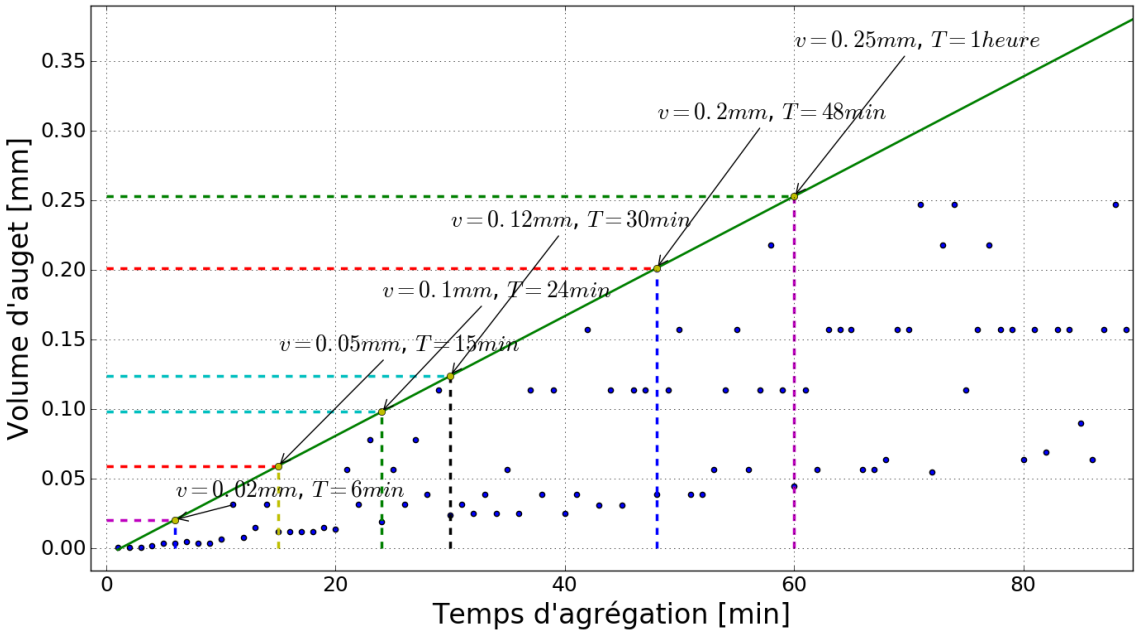


Figure 3.6 : Couples volume d'auget $v(T)[mm]$ / temps d'agrégation $T[min]$ optimum déduits de l'équation 15 par recherche séquentielle. En vert la droite délimitant la borne supérieure de $v(T)$

Une régression linéaire nous permet de déduire une relation de proportionnalité entre v et T:

$$v(T) = 6{,}95 \times 10^{-5}\, T \quad (3.16)$$

Avec T en secondes et v en mm.

Ou de façon équivalente :

$$T(v) = 14400\, v \quad (3.17)$$

A titre d'exemple, on peut voir que l'utilisation de pluviomètres dotés d'auget de 0,2 mm (cas de la majorité des pluviomètres) nécessite au moins un temps d'agrégation $T = 48\, min$ pour gommer l'effet de discrétisation de l'auget tandis qu'un temps d'agrégation $T = 24\, min$ est suffisant avec un pluviomètre de volume $v = 0{,}1\, mm$.

La figure 3.7-a montre l'enregistrement de l'événement pluvieux du 22 décembre 2012 caractérisé par une pluie faible et continue. Sur la figure 3.7-b sont présentées : la série de référence (disdromètre) en bleu, la série pseudo-pluviomètre avec un volume d'auget $v = 0{,}2\, mm$ (resp. $v = 0{,}1\, mm$) en vert (resp. en rouge) pour un temps d'agrégation T égal à une minute. La continuité de la série perçue par le disdromètre est perdue dans le cas des séries pseudo-pluviomètres pour lesquelles on enregistre des périodes non pluvieuses séparant les basculements. C'est le cas notamment à partir de l'instant $t = 650\, min$ dans lequel on n'observe plus de basculements jusqu'à l'instant $t = 720\, min$ laissant présager qu'il pourrait s'agir de deux évènements distincts.

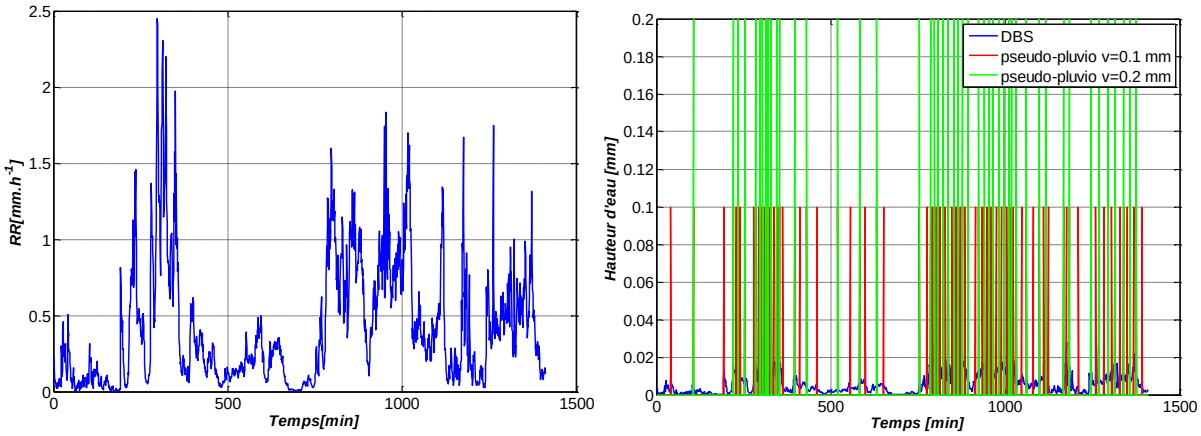


Figure 3.7 l'événement pluvieux du 22 décembre 2012 : (a) la série de référence issue du disdromètre agrégée à 1 min (b) pseudo-pluviomètres au pas d'agrégation T=1 min

La figure 3.8 présente l'effet de l'agrégation temporelle sur cette même série pour un temps d'agrégation de 6 minutes. Malgré une agrégation temporelle de 6 minutes, l'effet de la discrétisation dû aux augets est toujours bien visible. Cela est en accord avec la relation présentée précédemment qui indiquent que des durées d'agrégation de 24 et 48 minutes sont nécessaires pour des volumes d'auget de 0,1 et 0,2 mm respectivement.

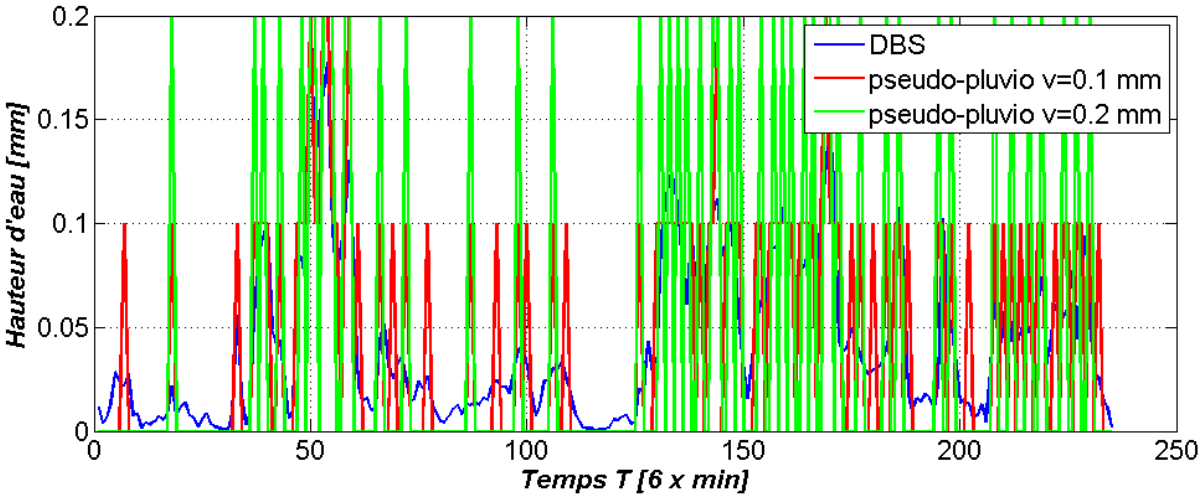


Figure 3.8 Evènement pluvieux du 22 décembre 2012 agrégé à $T = 6\ min$. En bleu disdromètre, en rouge et vert pseudo-pluviomètres

En augmentant le temps d'agrégation à 24 minutes puis à 48 minutes, l'effet des augets s'estompe peu à peu. Ainsi sur la figure 3.9-a, où le temps d'agrégation a été fixé à 26 minutes, la courbe déduite du pseudo-pluviomètre à $0{,}1\ mm$ se rapproche nettement de celle du disdromètre tandis que celle correspondant au pseudo-pluviomètre à $0{,}2\ mm$ présente encore quelques écarts notables en présence de pluie très faibles, par exemple entre les instants 20 et 30 où le pseudo-pluviomètre n'enregistre pas de traces de pluie aux $21^{ème}$, $23^{ème}$ et $25^{ème}$ pas de temps sachant pourtant qu'il pleut. L'effet induit par les augets de $0{,}2\ mm$ n'a pas été compensé à ces pas de temps-là, validant l'insuffisance du temps d'agrégation de $26\ min$ (<48 min) pour de tel volume d'auget. Avec un temps d'agrégation de 50 minutes (> 48 minutes) (figure 3.9-b), la courbe relative au pseudo-pluviomètre à $0{,}2\ mm$ tend à se rapprocher nettement de celle du disdromètre, même s'il existe encore quelques écarts. Ainsi, l'utilisation d'un temps d'agrégation de 50 minutes a pratiquement gommé les effets de la discrétisation.

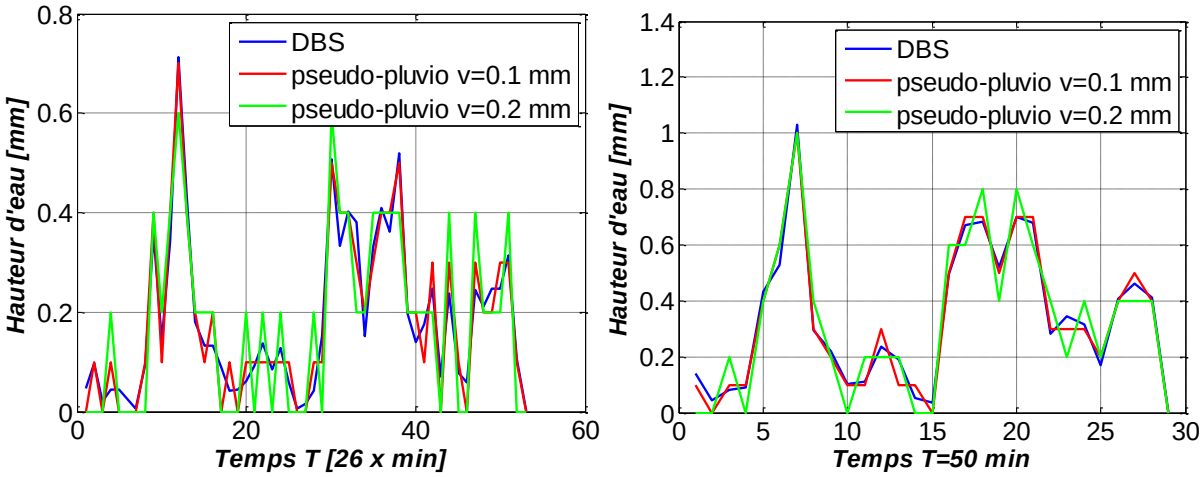


Figure 3.9 Idem figure 3.8 : à gauche $T = 26\ min$, à droite $T = 50\ min$

3.4.3. Discussion des résultats

Pour que les observations issues de pluviomètres à auget de 0,1 ou 0,2 mm soient parfaitement identiques aussi bien en termes d'occurrence que de distribution des hauteurs d'eau observées des temps d'agrégation de 24 ou 48 minutes sont nécessaires. D'après la relation trouvée précédemment, un temps d'agrégation de 6 minutes nécessiterait un auget de 0.02 mm, autrement dit l'utilisation d'un pluviomètre à pesée serait nécessaire. La continuité de l'observation de la pluie est alors assurée et l'information sur le support de pluie est donc exacte en chaque point de l'axe du temps.

Il convient de noter que la relation volume d'auget / temps d'agrégation obtenue ici est caractéristique de la région Île de France. Bien évidemment, cette relation est sujette à la climatologie locale, la prédominance d'évènements stratiformes de faibles intensités induit d'autant plus d'erreurs de quantification alors que cela serait beaucoup moins vrai pour une région impactée majoritairement par des événements convectifs plus intenses.

La relation entre le temps d'agrégation T et le volume d'auget v qui a été obtenue assure une équivalence entre les observations des deux types de système de mesure (disdromètre/pluviomètre à auget) à chaque instant. Le plus souvent en pratique, le pas de temps et le volume d'auget sont fixés par des contraintes extérieures, il convient donc à l'utilisateur de vérifier si la relation temps d'agrégation / volume d'auget est vérifiée.

On s'intéresse dans la suite de ce chapitre aux conséquences de l'agrégation temporelle sur la quantité d'information de la série agrégée par rapport à la série non agrégée.

3.5. Influence du temps d'agrégation sur l'observation de la variabilité des précipitations

Dans cette partie on étudie l'impact de l'agrégation temporelle sur la variabilité des séries temporelles des précipitations, **la discrétisation due aux augets n'est pas prise en compte ici**. L'agrégation temporelle est une opération « destructrice » de l'information au regard de la variabilité des précipitations. Dans le cas d'une série d'intensité de pluie, l'agrégation temporelle remplace un ensemble de valeurs par leur moyenne, introduisant ainsi une perte de l'information sur la variabilité. La figure 3.10 présente en bleu un zoom de l'intensité de pluie sur l'intervalle de temps allant de la 225ème minute à la 405ème minute de l'événement du 22 décembre 2012 mesuré par le disdromètre pour des temps d'agrégation T d'une minute (en bleu) et T égal à 45 minutes (en rouge) ainsi que la moyenne globale de la

série (en bleu clair). Les lignes grises représentent l'écart entre les intensités de pluie à 1 minute et à 45 minutes. On peut voir que la variation des intensités de pluie à la résolution initiale ($T = 1min$) est très importante, elle diminue considérablement en agrégeant à $T = 45\ min$. On peut dire de façon qualitative que la perte d'information sur la variabilité introduite par l'opération d'agrégation est très importante. Nous nous proposons dans cette section de quantifier cette perte.

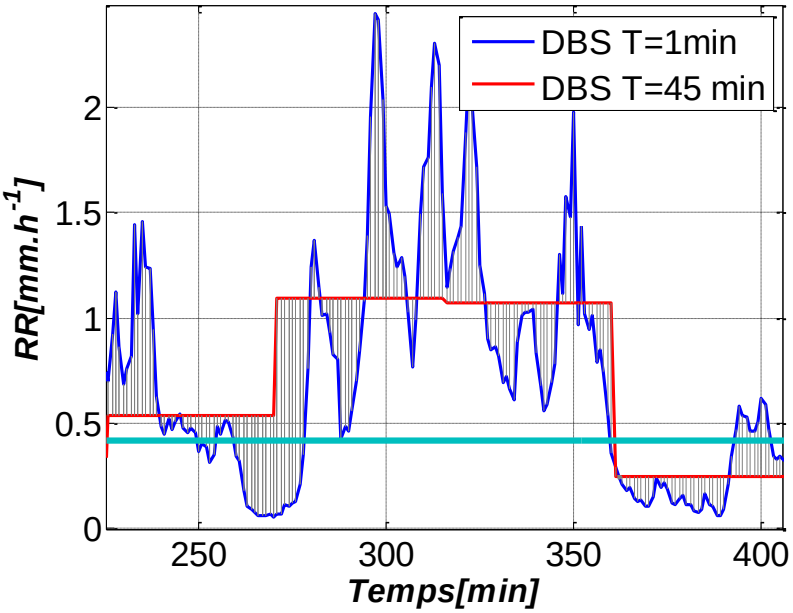


Figure 3.10 : Evénement pluvieux agrégé à $T = 1\ min$ (bleu) et à $T = 45\ min$ (rouge). Valeur moyenne en bleu clair.

3.5.1. Quantification de l'information conservée / perdue

Considérons la série temporelle d'intensité de pluie RR^{λ} de durée L à la résolution λ correspondant au temps d'agrégation T et RR^{λ_m} la série agrégée à la résolution λ_m correspondant au temps d'agrégation $T_m = mT$. On rappelle que l'i^{ème} terme de RR^{λ_m} vaut (cf. chapitre 1 eq. 1.6) :

$$RR_i^{\lambda_m} = \frac{1}{m} \sum_{j=m(i-1)+1}^{mi} RR_j^{\lambda} \quad i = 1 \dots \lambda_m \quad Chapitre\ 1\ eq.\ 1.6$$

La perte introduite notée $I_p\left(RR_i^{\lambda_m}\right)$ peut être mesurée par :

$$I_p\left(RR_i^{\lambda_m}\right) = \sum_{j=m(i-1)+1}^{mi} \left(RR_j^{\lambda} - RR_i^{\lambda_m}\right)^2 \quad (3.18)$$

La perte d'information sur la variabilité après l'agrégation temporelle à la résolution λ_m est donc

$$I_p^{\lambda_m} = \sum_{i=1}^{\lambda_m} I\left(RR_i^{\lambda_m}\right) \quad (3.19)$$

En remplaçant par (3.18), on trouve :

$$I_p^{\lambda_m} = \sum_{i=1}^{\lambda_m} \sum_{j=m(i-1)+1}^{mi} \left(RR_j^{\lambda} - RR_i^{\lambda_m}\right)^2 \quad (3.19)$$

- Pour $\lambda_m = \lambda$, correspondant à $T_m = T$, la série n'est pas modifiée $I_p^{\lambda} = 0 \rightarrow$ pas de perte d'information
- Pour $\lambda_m = 1$ correspondant à $T_m = L$, la série se résume à un seul échantillon égal à la valeur moyenne RR_m de la série : la perte d' information sur la variabilité est maximale :

$$I_p^1 = \sum_{j=1}^{\lambda} \left(RR_j^{\lambda} - RR_m\right)^2 \quad (3.20)$$

De la même façon, l'information conservée par l'agrégation temporelle notée $I_c^{\lambda_m}$ peut être mesurée par:

$$I_c^{\lambda_m} = \sum_{i=1}^{\lambda_m} m \times \left(RR_i^{\lambda_m} - RR_m\right)^2 \quad (3.21)$$

- Pour $\lambda_m = \lambda$, correspondant à $T_m = T$, la série est la série initiale $I_c^{\lambda} = I_p^1$ toute l'information est conservée
- Pour $\lambda_m = 1$ correspondant à $T_m = L$, on obtient $I_c^1 = 0$ l'information conservée est nulle

I_c^{λ} donne donc **une mesure de la variabilité totale de la série à la résolution fine λ et on la note I^t**. La relation fondamentale de Huygens permet, pour toutes les résolutions intermédiaires λ_m, de décomposer l'information totale en information perdue et information conservée :

$$I^t = I_p^{\lambda_m} + I_c^{\lambda_m} \quad (3.22)$$

Quelques conséquences :

1- Le pourcentage d'information conservé lors d'une agrégation temporelle à la résolution λ vers la résolution λ_m de la série RR^{λ} est :

$$\frac{I_c^{\lambda_m}}{I^t} \times 100 \quad (3.23)$$

Ce pourcentage d'information conservée diminue de 100 à 0% quand le temps d'agrégation T_m augmente de 1 à L.

2- La perte relative introduite par l'agrégation temporelle à la résolution λ vers la résolution λ_m de la série RR^{λ} est :

$$\frac{I_p^{\lambda_m}}{I^t} \times 100 = \left(1 - \frac{I_c^{\lambda_m}}{I^t}\right) \times 100 \quad (3.24)$$

Ce pourcentage d'information perdu augmente de 0 à 100% quand le temps d'agrégation T_m augmente de T à L.

3- A partir des définitions on a pour toutes les résolutions intermédiaires $I_c^{\lambda_m} = \lambda\, Var\left(RR^{\lambda_m}\right)$, les pourcentages d'information conservés et d'information perdus peuvent être exprimés par des rapports de variance:

$$\frac{I_c^{\lambda_m}}{I^t} \times 100 = \frac{Var(RR^{\lambda_m})}{Var(RR^{\lambda})} \times 100 \quad (3.25)$$

$$\frac{I_p^{\lambda_m}}{I^t} \times 100 = \left(1 - \frac{Var(RR^{\lambda_m})}{Var(RR^{\lambda})}\right) \times 100 \quad (3.26)$$

Lorsqu'on passe d'une résolution λ_n à une résolution λ_m, le pourcentage d'information conservée/gagnée (variance expliquée) est donné par la formule :

$$Ve(\lambda_n \to \lambda_m) = \frac{Var(RR^{\lambda_m})}{Var(RR^{\lambda_n})} \times 100 \quad (3.27)$$

1- Pour $\lambda_m < \lambda_n$: l'opération effectuée est une agrégation temporelle $Ve(\lambda_n \to \lambda_m)$ représente le pourcentage d'information conservée et varie entre 0 et 100%.

2- Pour $\lambda_m > \lambda_n$: l'opération effectuée est un raffinement temporel $Ve(\lambda_n \to \lambda_m)$ représente le pourcentage d'information gagné et varie entre 100 et l'infini.

On note $Ve(\lambda_m) = Ve(\Lambda \to \lambda_m)$ la variance expliquée (pourcentage d'information conservée) lorsqu'on agrège de la résolution la plus fine (Λ) à la résolution λ_m.

3.5.2. Propriétés multifractales des précipitations et relations d'échelle

Les propriétés multifractales (Schertzer et Lovejoy, 1987) des précipitations permettent l'estimation d'un moment d'ordre q quelconque à une échelle d'observation (résolution) λ_m quelconque à partir de la connaissance de ce moment à une autre échelle λ_n :

$$M^{\lambda_m}(q) = M^{\lambda_n}(q)\,(\frac{\lambda_m}{\lambda_n})^{K(q)} \quad (3.28)$$

$K(q)$ est appelée fonction d'échelle des moments (Schertzer et Lovejoy, 1987).

La variance statistique étant le moment d'ordre 2, l'équation 28 devient :

$$Var\left(RR^{\lambda_m}\right) = Var\left(RR^{\lambda_n}\right)(\frac{\lambda_m}{\lambda_n})^{K(2)} \quad (3.29)$$

Par conséquent,

$$\frac{Var(RR^{\lambda_m})}{Var(RR^{\lambda_n})} = (\frac{\lambda_m}{\lambda_n})^{K(2)} \quad (3.30)$$

On en déduit que :

$$Ve(\lambda_n \to \lambda_m) = 100 \times (\frac{\lambda_m}{\lambda_n})^{K(2)} = 100 \times (\frac{T_m}{T_n})^{-K(2)} \quad (3.31)$$

Remarques :

1. Le pourcentage d'information conservée par l'agrégation temporelle est une fonction puissance du rapport des temps d'agrégation considérés.
2. Cette fonction prend la valeur 100 pour $\lambda_n = \lambda_m$ qui correspond à une conservation totale de l'information de la série.
3. Quel que soit la valeur prise par $K(2)$ cette fonction est strictement monotone.

Dans notre cas ($T_n = 1\,min$), le pourcentage d'information conservée pour une agrégation temporelle T_m s'écrit:

$$Ve(\lambda_m) \propto 100 \times (T_m)^{-K(2)} \quad (3.32)$$

3.5.3. Application : Impact de l'agrégation temporelle sur les séries des taux précipitants en Île de France

On utilise la série des intensités de pluie mesurée par le disdromètre sur le site du SIRTA entre 2012 et 2013 comme série de référence à l'échelle la plus fine exploitable ($T = 1\,minute$) (De Montera et al., 2009), nous avons estimé empiriquement le pourcentage d'information conservé pour T_m allant de $1\,minute$ à $10080\,minutes$ avec un pas $\Delta T = 1\,min$. Les résultats sont présentés à la figure 3.11.

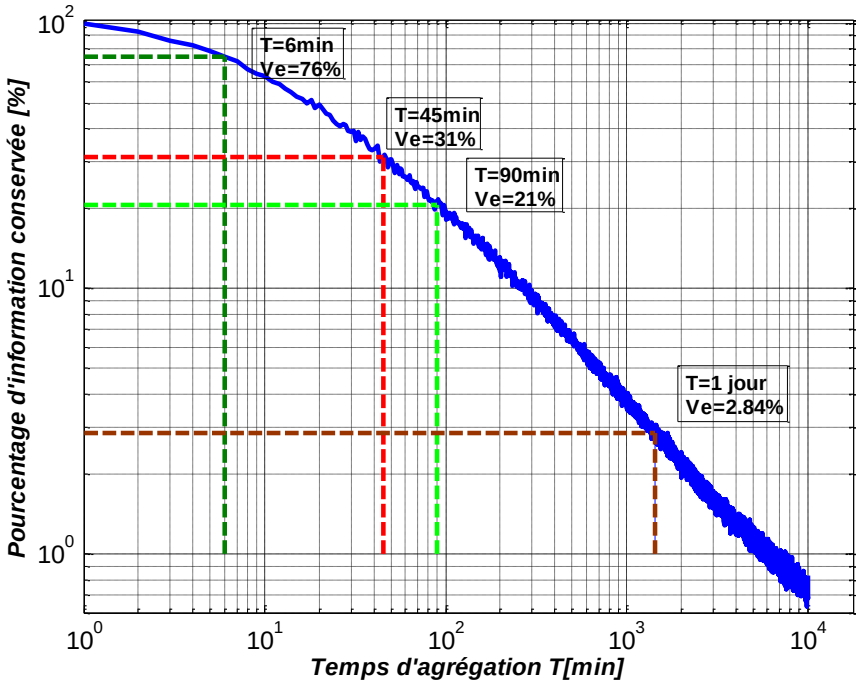


Figure 3.11 : Pourcentage d'information conservée en fonction du temps d'agrégation considéré par rapport à un temps d'agrégation de 1 minute

Le pourcentage d'information conservé a une allure strictement décroissante, on enregistre une décroissance rapide au début et un ralentissement avec l'augmentation du temps d'agrégation. On peut observer en échelle log-log une décroissance entre 1 et 30 minutes avec une pente relativement faible suivi d'un coude et d'une décroissance linéaire avec une pente plus importante. La queue de la distribution est de type loi puissance avec une cassure aux alentours de 30-40 minutes en accord avec les différents régimes d'invariance d'échelle mis en évidence par l'analyse multifractale des précipitations (Verrier et al., 2011).

Estimation de K(2) pour la résolution la plus fine Λ ($T_n = 1min$)

D'après l'équation 3.32 on peut écrire :

$$\log\left(\frac{Ve(\lambda_m)}{100}\right) = cste \; - K(2) \times \log(T_m) \quad (3.33)$$

L'estimation de $K(2)$ par régression linéaire de la queue de la distribution donne une valeur $K(2) = 0{,}71$. Cette valeur est en bon accord avec celle estimée dans les travaux de doctorat de S. Verrier (Verrier, 2011) avec une valeur de 0,65. Dans ces conditions, le pourcentage d'information conservé s'écrit:

$$Ve(\lambda_m) \propto 100 \times (T_m)^{-0.71} \quad (3.34)$$

3.5.4. Discussion des résultats

Le tableau 3.3 répertorie le pourcentage d'information conservé $Ve(T_n \rightarrow T_m)$ pour quelques valeurs de T couramment utilisées dans la littérature. Il montre que même une agrégation sur des temps relativement courts de 6 minutes on ne conserve que 76% de l'information par rapport à un temps d'agrégation d'une minute. Cette perte de 24% de l'information peut être un facteur limitant dans le cadre de certaines études, par exemple si on étudie la variabilité des précipitations à très fine échelle. A l'échelle de la journée, seule 2,84% de l'information est conservée par rapport à une observation à la minute, tandis qu'on passe à 11% d'information conservée par rapport à une observation horaire.

Cette perte d'information de la variabilité est à mettre en relation avec la variabilité artificielle induite par les augets. Nous avons montré dans les sections précédentes qu'au-delà de 24 ou 48 minutes la variabilité induite par ces derniers devenait négligeable alors que pour ces mêmes durées la perte d'information due à l'agrégation est importante. L'utilisation de pluviomètres à auget pose le dilemme suivant : utiliser un temps d'agrégation court permettant d'obtenir une bonne restitution de la variabilité avec toutefois un temps d'agrégation suffisamment long pour que l'effet des augets soit négligeable. Un compromis doit être fait suivant le type d'étude mené.

$T_n \backslash T_m$	1min	6min	30min	45min	1heure	90min	2heures	1jour
1 min	100	76	39	31	26	21	17	2.84
6 min		100	52	41	35	28	23	3.82
30min			100	80	68	53	45	7.30
45 min				100	85	66	56	9.11
1 heure					100	78	66	11
90 min						100	84	14
2 heures							100	16
1 jour								100

Tableau 3.3 : Pourcentage d'information conservée pour différents temps d'agrégation

3.6. Influence du volume des augets sur l'estimation du support des précipitations

Dans cette partie on étudie l'impact de la discrétisation induite par l'auget sur la qualité des séries temporelles des précipitations.

La discrétisation induite par les augets introduit une non linéarité forte, et tout comme l'agrégation c'est une opération destructrice de l'information concernant l'observation du support des précipitations. Dans le cas d'une série d'intensité de pluie, la discrétisation remplace

toutes les valeurs précédant le basculement de l'auget par des valeurs nulles introduisant ainsi une perte d'information sur le support. La figure 3.12 présente en bleu un zoom sur l'intervalle de temps entre la première minute et la 405ème minute de la série des intensités de pluie de l'événement du 22 décembre 2012 (figure 3.6-a) mesuré par le disdromètre avec un temps d'agrégation de $T = 1\ min$ et en vert sa discrétisation par le pseudo-pluviomètre à $v = 0{,}2\ mm$. On peut voir que le support des précipitations sans discrétisation est continu, cette information de continuité diminue considérablement en utilisant un volume d'auget $v = 0{,}2\ mm$, on compte 1410 échantillons non nuls pour la série du disdromètre contre 48 échantillons pour tout l'événement discrétisé. De plus, la discrétisation décale la première détection (le début de l'événement) de 106 minutes. La perte d'information sur le support peut donc dans certain cas être importante.

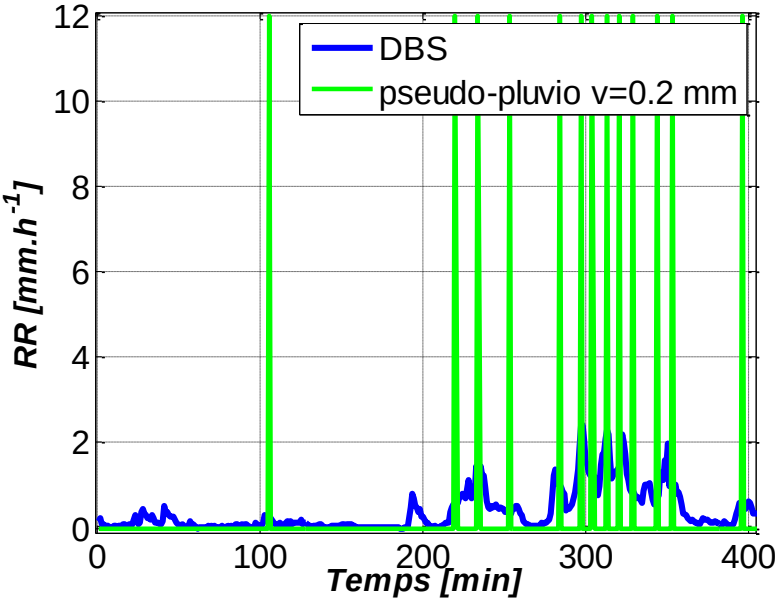


Figure 3.12 : Evénement du 22 décembre 2012 observé par le disdromètre et le pseudo-pluviomètre à 0,2 mm

3.6.1. Quantification de l'information sur l'occurrence conservée

D'une façon similaire à 5.1, si on considère la série temporelle des intensités de pluie RR^v correspondant au volume d'auget v, et RR^{v_m} sa discrétisation avec un volume d'auget $v_m = v + m\,\Delta v$. Une mesure (partielle) de l'information conservée sur le support après cette discrétisation notée $O_p^{v_m}$ peut être fournie par le nombre de valeurs non nulles de la série :

$$O_c^{v_m} = \sum_{i=1}^{\lambda} \chi(RR_i^{v_m}) \quad (3.33)$$

Avec $\chi(RR_i^{\lambda})$ la fonction indicatrice du support des précipitations :

$$\chi\left(RR_i^{v_m}\right) = \begin{cases} 1 & si \quad RR_i^{v_m} > 0 \\ 0 & si \quad RR_i^{v_m} = 0 \end{cases} \quad (3.34)$$

Lorsqu'on passe d'une discrétisation avec une volume d'auget v_n à une discrétisation avec un autre volume d'auget v_m, le pourcentage d'information conservée sur l'occurrence peut être obtenu en comparant l'information conservée aux deux volumes :

$$qe(v_n \rightarrow v_m) = \frac{O_p^{v_m}}{O_p^{v_n}} \times 100 \quad (3.35)$$

1- Pour $v_n < v_m$: l'opération effectuée est une accumulation d'eau $qe(v_n \rightarrow v_m)$ représente le pourcentage d'information conservée et varie entre 0 et 100%.
2- Pour $v_n > v_m$: l'opération effectuée est une répartition du cumul d'eau et $qe(v_n \rightarrow v_m)$ représente le pourcentage d'information gagné et varie entre 100 et l'infini.

On note $qe(v_m) = qe(0 \rightarrow v_m)$ le pourcentage d'information conservée après une discrétisation induite par le volume d'auget v_m.

3.6.2. Application : Impact du volume d'auget sur les séries d'intensité de pluie en Île de France

En prenant la même série des intensités de pluie qu'à la partie 5.3 (disdromètre, T=1 minute), nous avons estimé empiriquement le pourcentage d'information conservé pour v_m variant de $v = 0\ mm$ à $v = 50\ mm$ avec un pas $\Delta v = 0{,}001 mm$. Les résultats sont présentés à la figure 3.13.

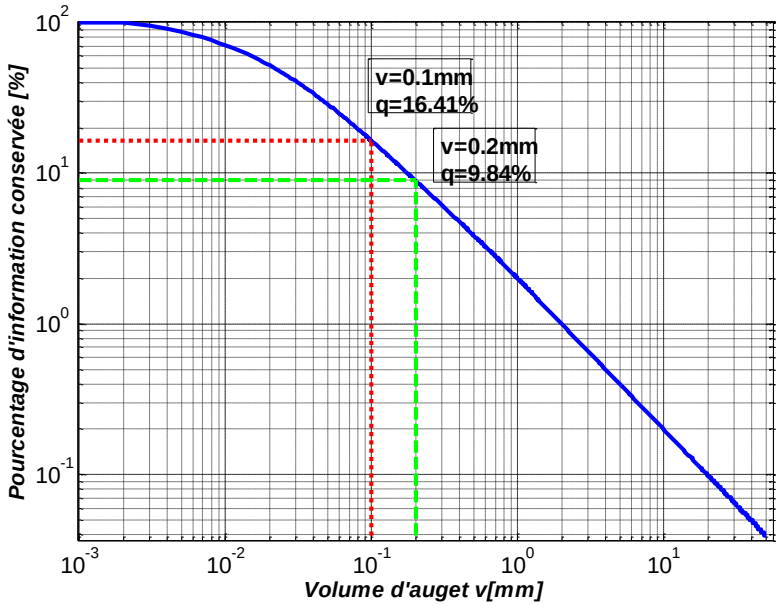


Figure 3.13 : Pourcentage d'information conservée sur l'occurrence en fonction du volume d'auget considéré par rapport à un volume d'auget nul

Le pourcentage d'information conservé a une allure strictement décroissante, on enregistre une décroissance rapide au début et un ralentissement avec l'augmentation du volume d'auget. On peut observer en échelle log-log une décroissance jusqu'à 0,02mm environ avec une pente relativement faible suivi d'un coude et d'une décroissance linéaire avec une pente plus importante. La décroissance linéaire de la queue de la distribution traduit une fois de plus une loi puissance. L'ajustement de cette dernière permet de déduire la relation suivante :

$$qe(v_m) \propto (v_m)^{-1.0016} \quad (3.36)$$

Cette relation montre que l'exposant (aux erreurs d'estimation prés) est égal à -1, on en déduit donc que le produit $qe\ v_m$ est constant : **la quantité d'information conservée diminue d'autant plus que le volume augmente**.

Le tableau 3.4 présente le pourcentage d'information conservée de l'occurrence $qe(v_n \rightarrow v_m)$ pour quelques volumes d'auget types. Ce tableau montre par exemple l'intérêt d'utiliser un pluviomètre à 0,1 mm plutôt que 0,2 mm pour lequel ce dernier ne conserve que 66% de l'information par rapport au premier. Au regard de ce tableau, l'utilisation d'un disdromètre pour l'étude du support de la pluie est fondamentale avec seulement 16% et 8,95% d'information conservée pour des pluviomètres respectivement de 0,1 et 0,2 mm.

$v_n \backslash v_m$	0 mm	0.02 mm	0.1 mm	0.15 mm	0.2 mm	0.5 mm	1 mm
0 mm	100	52.03	16.4	11	9.84	3.88	2.00
0.02 mm		100	38	33	25	14	7.95
0.1 mm			100	86	66	36	21
0.15 mm				100	77	41	24
0.2 mm					100	54	31
0.5 mm						100	57
1 mm							100

Tableau 3.4 : Pourcentage d'information conservée sur l'occurrence pour différents volumes d'auget

Après avoir étudié les impacts dus aux augets et de l'agrégation temporelle sur la qualité globale des séries temporelles des précipitations, la dernière partie de ce chapitre est consacrée à leur impact sur l'analyse par évènement effectuée au chapitre précédent.

3.7. Analyse par événement – sensibilité aux paramètres T et v

3.7.1 Impact de l'agrégation temporelle sur la définition des événements

La définition d'un point de vue temporel d'un événement de pluie dépend du temps inter-évènement-minimal MIT et du volume d'eau minimal requis MEV choisis (voir chap.1 7.1). Sous réserve que le volume d'auget v de l'instrument de mesure soit inférieur au volume d'eau minimal requis MEV (l'évènement génère au moins un basculement et, est de ce fait détecté), étudier l'effet de l'agrégation temporelle sur la définition de l'événement revient à étudier son effet sur le temps minimal inter-événements MIT.

On reprend la série temporelle du taux précipitants RR^{λ} de longueur L à la résolution fine λ correspondant au temps d'agrégation T, et RR^{λ_m} la série agrégée à la résolution λ_m correspondant au temps d'agrégation $T_m = mT$. Deux cas sont à envisager:

- 1er cas ($T_m \leq MIT$) : dans le cas d'un temps d'agrégation T_m inférieur au MIT, la période de non pluie qui sépare deux événements successifs est détectée. Ainsi le nombre d'événement reste inchangé.
- 2ème cas ($T_m > MIT$) : dans le cas d'un temps d'agrégation T_m supérieur au MIT, la période de non pluie peut être fusionnée avec une période pluvieuse ; ainsi deux événements indépendants risquent d'être regroupés en un seul événement. Par conséquent, le nombre d'évènements de précipitations définis décroit.

Exemple : effet de l'agrégation temporelle sur le nombre d'événements :

Au chapitre 2, pour la série temporelle d'intensité de pluie issue du disdromètre avec $T = 1\ min$, nous avons choisi un temps inter-évènements minimal $MIT = 30\ min$ et un volume d'eau minimal $MEV = 1mm$ pour définir un événement de précipitations. Le tableau 3.5 présente le nombre d'événements de pluie identifiés à des résolutions différentes.

T	1 min			6 min			1 $heure$			1 $jour$		
v [mm]	0	0.1	0.2	0	0.1	0.2	0	0.1	0.2	0	0.1	0.2
Nombre d'événements	1284	1054	948	1150	917	846	809	639	575	117	114	116
Nombre d'événements avec Cumul>1mm	234	276	291	227	275	287	219	274	295	77	84	87

Tableau 3.5 : nombre d'évènements de précipitations en fonction du temps d'agrégation

A quelques événements près, la définition de l'événement est stable par rapport aux résolutions temporelles $T = 1\ min$, $6\ min$ et $1\ heure$. En revanche pour la résolution d'un jour les événements identifiés précédemment fusionnent pour former des journées de pluie.

3.7.2. Effet sur la classification des évènements de précipitations

La classification réalisée au chapitre 2 sur les données du disdromètre agrégées à $T = 1\ min$, utilisait une définition de l'événement de précipitations qui se basait sur un $MIT = 30\ min$ et un $MEV = 1mm$. Cela nous a conduits à un ensemble de 234 événements de précipitations.

Afin de voir l'effet de l'agrégation temporelle et de la discrétisation de l'auget sur la caractérisation puis la classification des événements de précipitations, et pour des raisons de comparabilité quatre expériences de classification par cartes topologiques ont été menées sur le même ensemble d'événements en simulant une agrégation temporelle et/ou une discrétisation de l'auget. Pour l'étude de la variabilité des précipitations à fine échelle, la communauté scientifique utilisent des données de pluviomètres à auget ($v = 0{,}2\ mm$) agrégées à $T = 5\ min$ (Kann et al. 2015). Les données dont nous disposons pour la suite viennent de Météo-France qui dispose d'un large réseau de pluviomètres à auget ($v = 0.2\ mm$), et agrège ses données à $T = 6\ min$. Par conséquent, les paramètres choisis pour ces quatre expériences sont $T = 1\ et\ 6\ min$ et $v = 0\ et\ 0.2\ mm$:

1- **Expérience 0** : c'est la série de référence issue du disdromètre à *T*=1 minute.
2- **Expérience 1** : influence de l'agrégation temporelle. C'est la série du disdromètre à $T = 6\ min$. Cette valeur est inférieur au $MIT = 30min$ considéré, le risque de fusion des événements est écarté. De plus la durée minimale a été enregistrée pour le 34ème événement et vaut $D_e^{34} = 7\ min$.
3- **Expérience 2** : influence de la discrétisation dû aux augets. Une série pseudo-pluviomètre avec $v = 0.2\ mm$ et T = 1 minute est calculée à partir de la série du disdromètre. Le volume minimal requis étant $MEV = 1\ mm$, le risque de non-détection des événements est écarté. De plus, tous les événements enregistrent au moins 4 basculements, l'information sur la variabilité n'est donc pas totalement perdue.
4- **Expérience 3** : les deux effets agrégation temporelle et discrétisation de l'auget sont combinés $v = 0.2\ mm$ et T = 6 minutes.

Paramètres des cartes topologiques: les cartes ont toutes la même taille 8x8 = 64 neurones.

Initialisation : les neurones des différentes cartes ont été initialisés avec les valeurs de la carte obtenue au chapitre 2. Cette initialisation nous garantit que la variation des positions des neurones caractérise les différents effets.

Apprentissage : l'apprentissage pour les trois expériences utilise les cinq variables définies au chapitre précédent (durée de l'événement pluvieux, son écart-type, son maximum d'intensité, sa variation absolue du taux précipitant, et sa hauteur d'eau). Ces variables calculées sur des séries ayant des taux d'agrégation temporelle ou de discrétisation différents ne caractérisent pas les évènements de manière exactement identique

Analyse des résultats

Déploiements des cartes : Les différentes matrices des distances montrent pour chaque carte une répartition relativement équilibrées des événements sur l'ensemble des neurones (en moyenne : 4 événements par neurone) manifestant leur bon déploiement.

Le Tableau 3.7 montre les différentes projections des cartes pour sept des 23 variables définies au chapitre 2 et recalculées pour les différentes expériences. Les cartes relatives aux 5 variables utilisées pour l'apprentissage sont représentées sur les 5 premières lignes. Les différentes expériences sont représentées en colonnes. La dernière ligne de cartes représente le résultat de la classification en deux classes des neurones à l'aide d'une classification hiérarchique ascendante.

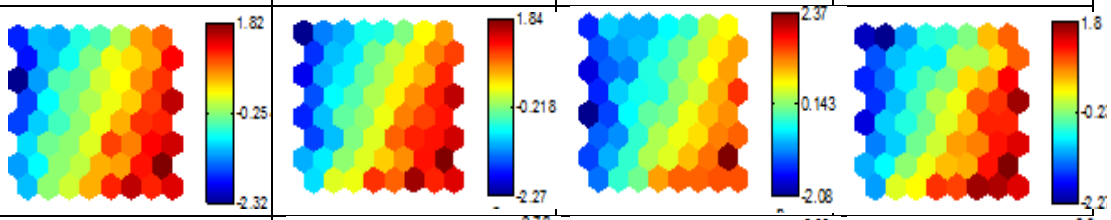
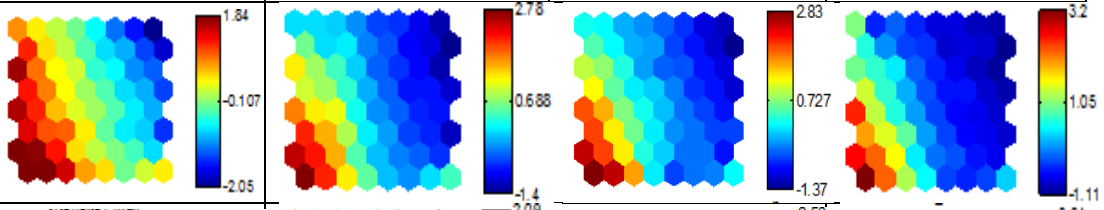
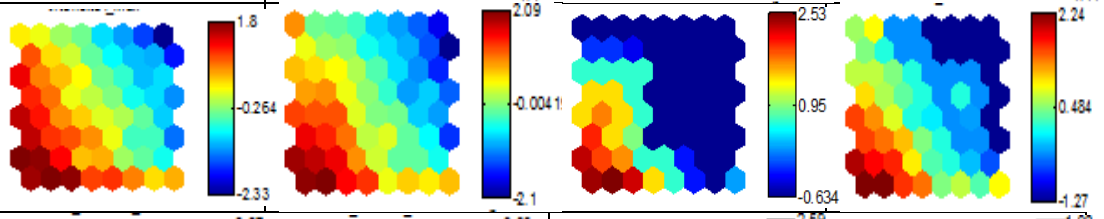
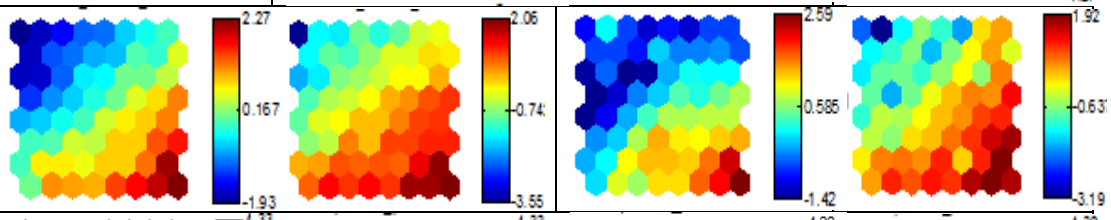
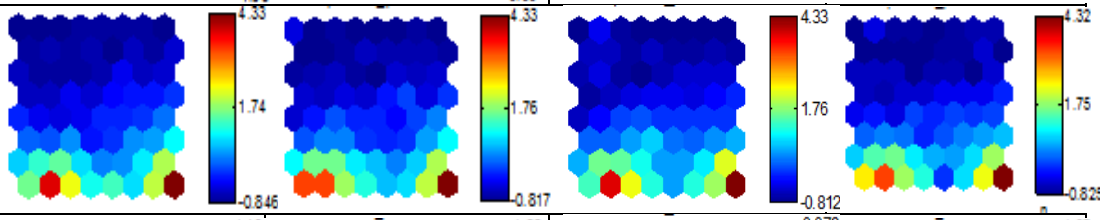
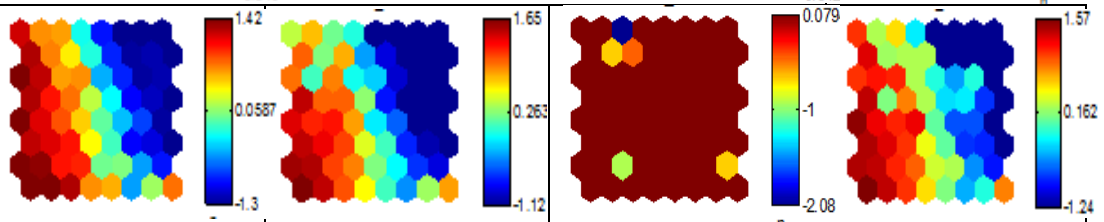
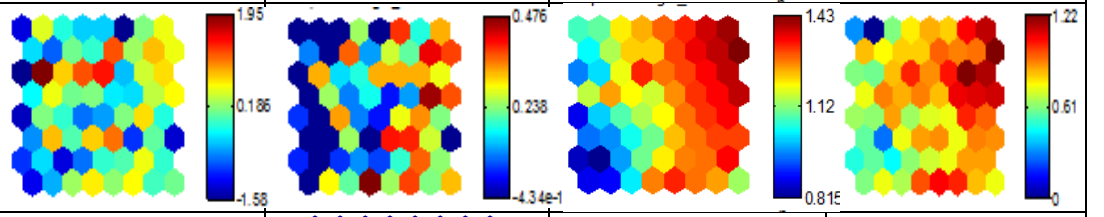
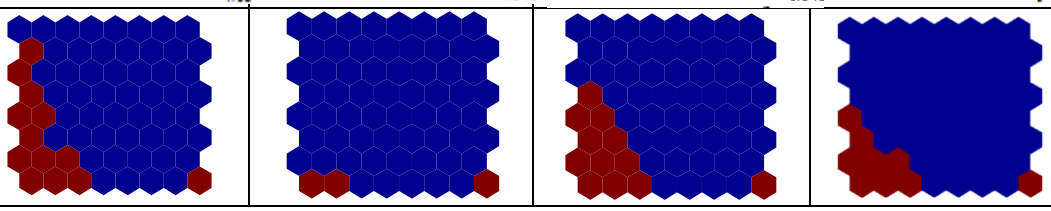
	Disdromètre ($v = 0\ mm$)		Pseudo-Pluviomètre ($v = 0.2\ mm$)	
Temps d'agrégation	$T = 1\ min$ Expérience 0	$T = 6min$ Expérience 1	$T = 1min$ Expérience 2	$T = 6min$ Expérience 3
D_e				
σ_R				
R_{max}				
P_{C1}				
R_d				
β_{L3}				
$D_{d\%e}$				
Classification en 2 classes				

Tableau 3.6 : Cartes topologiques obtenues pour diverses variables et pour les quatre expériences (échelles de couleur normalisée [-3, 3] pour faciliter la comparaison des expériences)

Dans un premier temps nous analyserons l'impact sur les différentes variables de l'agrégation temporelle de la discrétisation dû aux augets et de la combinaison des deux effets. Dans un second temps nous étudierons les conséquences sur la partition des évènements en 5 classes telle que réalisée au chapitre 2. Il s'agit de savoir si les impacts observés sur les cinq variables d'apprentissage sélectionnées au chapitre 2 modifient la topologie du nuage de points des évènements de pluie représenté dans l'espace de ces cinq variables.

Sensibilité des variables au changement d'instrument

Concernant les variables d'apprentissage, les projections des deux variables « durée de l'événement » (D_e) et « hauteur d'eau cumulée » (R_d) restent relativement inchangées pour les trois expériences. Ces deux variables sont peu sensibles au changement d'instruments et/ou à leur configuration (temps d'agrégation), les informations qu'elles expliquent sont relativement bien conservées. Au contraire, les différences observées pour les trois variables : écart-type σ_R, intensité maximale R_{max} et la variation absolue du taux précipitant P_{C1} illustrent une sensibilité importante aux changements des paramètres (T, v) :

Expérience 0 : C'est la série de référence, les cartes topologiques ont été présentées au chapitre 2 et copiées pour rappel dans le tableau 3.6.

Expérience 1 (influence de l'agrégation temporelle $T = 6\ min, v = 0\ mm$) : l'agrégation temporelle réduit globalement les valeurs des écart-types des intensités de pluie et les valeurs maximales caractérisant les événements. Cette réduction n'est pas homogène sur toute la carte, certaines zones de la carte sont plus affectées que d'autres. D'après la classification réalisée au chapitre 2 les neurones les plus affectés par cette réduction correspondent à la classe 5 (évènements convectifs très intenses et de courte durée). Afin d'étudier l'impact de l'agrégation temporelle en fonction du type d'évènements, la partition réalisée au chapitre 2 a été utilisée. Pour chacune des 5 classes d'évènements identifiés, l'évolution du pourcentage d'information conservée et des valeurs maximales ont été calculés pour des agrégations temporelles entre 1min et 1 heure. La figure 3.14 montre clairement que la classe 5 perd beaucoup d'information comparée aux quatre autres classes, corroborant ainsi l'observation faite précédemment à partir de la carte topologique.

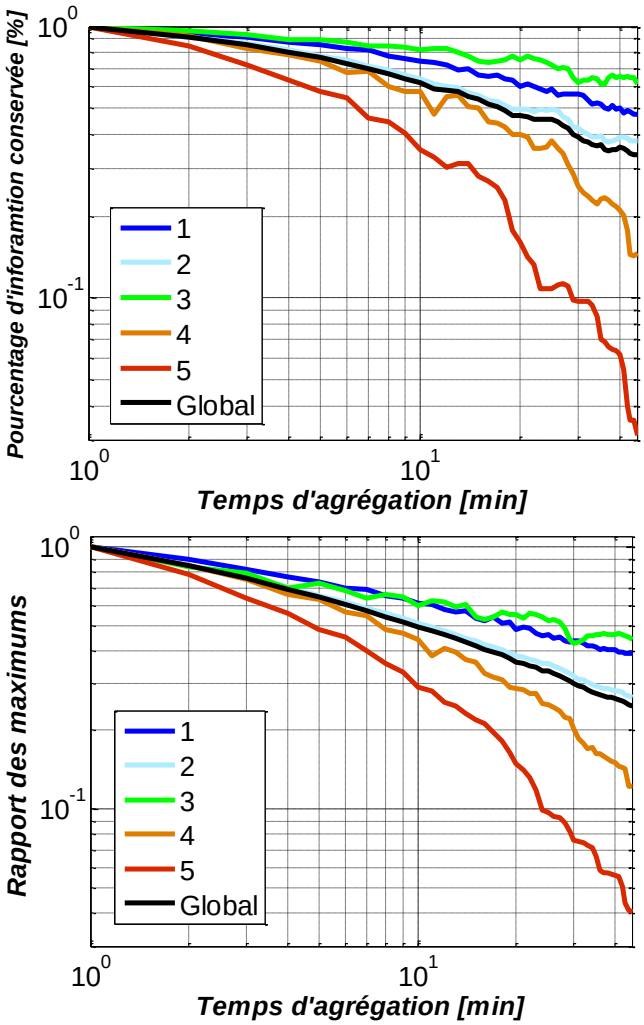


Figure 3.14. (haut) *Pourcentage d'information conservée en fonction du temps d'agrégation considéré par rapport à un temps d'agrégation de 1 minute pour chaque « valeur moyenne» de classe d'événements. La courbe en noir correspond au résultat global présenté dans la figure 3.11. (bas) rapport des valeurs maximales d'intensité de pluie en fonction du temps d'agrégation considéré par rapport à un temps d'agrégation de 1 minute pour chaque « valeur moyenne » de classe d'événements.*

Expérience 2 (influence de la discrétisation due aux augets $T = 1\,min,\ v = 0{,}2\,mm$): La discrétisation due aux augets semble avoir un effet similaire sur la variable écart type à celui dû à l'agrégation temporelle. Concernant la carte relative à la variable « intensité maximale » R_{max}, on observe que les neurones correspondant à des faibles valeurs sont très affectés par la discrétisation qui conduit à des valeurs nulles. En effet, pour les événements qui n'enregistrent pas un volume supérieur ou égal à v sur la durée d'agrégation T la discrétisation due aux augets a pour effet de mettre soit la valeur R= 0, soit la valeur $R = 0{,}2\frac{60}{1} = 12\,mm.h^{-1}$. D'après la classification réalisée au chapitre 2 les neurones les plus affectés correspondent à la classes 1 (bruine ou pluie très faible ; l'intensité maximale moyenne de la classe $R_{max}^{C1} = 2.08\,mm.h^{-1}$) et en partie à la classe 2 (évènements stratiformes; l'intensité maximale moyenne de la classe

$R_{max}^{C1} = 10\ mm.h^{-1}$). Pour les neurones correspondant à des évènements de faible intensité maximale, la discrétisation due aux augets conduit à une représentation binaire de l'intensité de pluie qui a tendance à gommer les différences qui existent entre les événements de la 1ère classe et une grande partie des événements de la 2ème classe. Pour les événements convectifs, les écarts de valeur s'expliquent simplement par le fait qu'un basculement même insignifiant en quantité d'eau (0,2 mm) est à T=1 minute équivalent à une différence d'intensité de 12 $mm.h^{-1}$.

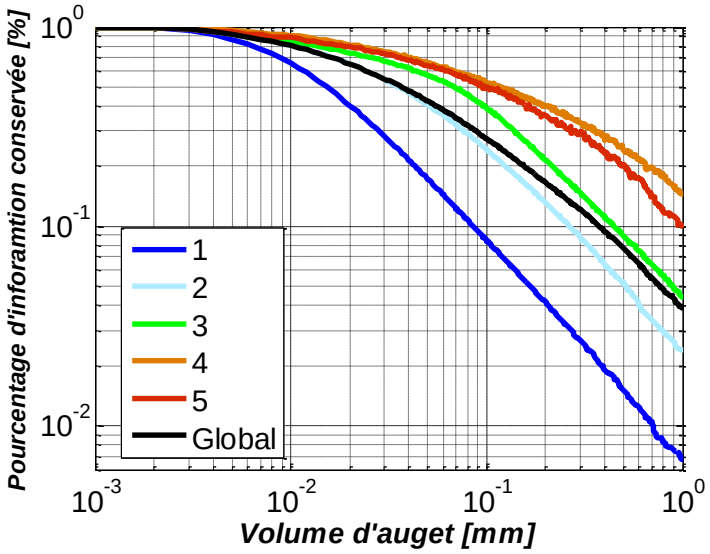


Figure 3.15. (a) *Pourcentage d'information conservée en fonction du volume d'auget pour chaque « valeur moyenne» de classe d'événements. La courbe en noir correspond au résultat global présenté dans la figure 3.13*

Expérience 3 : concernant la variable écart-type les deux opérations (agrégation temporelle et discrétisation de l'auget) semblent avoir le même effet sur les valeurs, **la combinaison des deux opérations reste « destructrice » de l'information contenue dans la variable écart-type.**

Par contre, dans le cas de la variable intensité maximale R_{max}, la carte topologique semble retrouver sa bonne structuration. Les conséquences de la discrétisation due aux augets (augmentation des intensités maximales) ont été gommées par l'agrégation temporelle (diminution des intensités maximales).

Le même effet est observé en comparant les projections des cartes sur la variable variation absolue du taux précipitant P_{C1} ; l'expérience 1 montre que l'agrégation fait augmenter les valeurs prises par cette variable alors que la discrétisation de l'auget a un effet inverse (diminuer les valeurs prise par cette variable). La combinaison des deux effets semble tendre vers un rééquilibrage. Cette hypothèse de rééquilibrage est soutenue par la mauvaise structuration de la carte topologique relative à la variable β_{L3} (indicateur du caractère convectif) pour la deuxième expérience alors qu'elle est bien structurée lors des deux autres expériences.

La variable β_{L3} est paramétrée par une valeur seuil d'intensité de pluie fixée à $b = 3mm.h^{-1}$ (cf. chapitre 2). β_{L3} est proportionnelle au pourcentage de points enregistrant des intensités supérieures à ce seuil, le passage par un effet d'auget relève toutes les valeurs d'intensité enregistrées à $12\ mm.h^{-1}$ ce qui donne un pourcentage de 100% pour tous les événements. Lors de l'agrégation temporelle à T=6min, le pic correspondant à un basculement est regroupé avec les valeurs adjacentes, souvent des zéros dans le cas des événements stratiformes. La valeur de $12\ mm.h^{-1}$ est donc réduite à $\frac{12}{6} = 2\ mm.h^{-1}$ en dessous de la valeur du seuil b. Dans le cas des événements convectifs continus et forts l'agrégation garde une valeur au-dessus du seuil, et par conséquent le caractère discriminant de la variable est vite rééquilibré par l'agrégation temporelle.

Pour une utilisation correcte de la variable β_L avec un pluviomètre à auget, il conviendrait donc de changer le seuil b afin de l'adapter au volume de l'auget et à la durée d'agrégation T. Cette solution bien qu'intéressante présente l'inconvénient de ne pas être stable, i.e. pour chaque configuration (T, v) utilisée il faudrait choisir une valeur adéquate du seuil. Cette remarque peut être généralisée à toutes les variables paramétrées (dans notre cas β_L, Psc).

Sensibilité de la partition en 2 classes au changement d'instrument

Les 5 variables d'apprentissage réagissent différemment à l'accroissement du temps d'agrégation et/ou au volume d'auget. Elles perdent une partie de l'information qui pourrait être utilisée pour la discrimination des événements. Cette perte d'information peut être définitive (cas de la variable écart-type) où « compensable » par une recherche d'équilibre entre les deux paramètres T et v (cas de la variable P_{c1}). Par conséquent, pour chaque configuration (T, v) les frontières des classes résultats de la classification automatique changent et ne sont pas stables pour un changement d'instruments et/ou un changement du temps d'agrégation. Le tableau 3.7 illustre l'effet sur une partition non supervisée en 2 classes. Si on suppose que dans les 2 cas la partition en 2 classes permet de distinguer les évènements stratiformes des évènements convectifs, on remarque que les évènements identifiés comme stratiformes avec un disdromètre le seront également avec le pluviomètre. Par contre les évènements classés comme convectifs, c'est à dire caractérisés par une intensité moyenne et maximale élevée, une variabilité importante et une durée relativement courte changent parfois de classe avec le pluviomètre et rejoignent la classe des stratiformes.

Matrice de confusion		Pseudo-pluviomètre ($v = 0.2mm$ et $T = 6min$)	
		stratiforme	convectif
Disdromètre ($v = 0\ mm$ et $T = 1\ min$)	stratiforme	188	0
	convectif	28	18

Tableau 3.7: matrice de confusion de la partition en 2 classes obtenues à partir de la série du disdromètre avec T = 1 min et de la série du pseudo-pluviomètre à auget de 0.2 mm avec T=6 min

D'autre part, alors que la perte de l'information sur la variabilité expliquée par l'écart-type semble inévitable lors de l'accroissement de T ou v, un « gain » d'information sur la variabilité « pourcentage de non pluie intra-événement $D_{d\%e}$ est observé ». En effet, les deux cartes topologiques associées à cette variable pour les expériences 0 et 1 (disdromètre T=1 et 6 minutes) montrent des cartes non structurées sans regroupement possible. L'introduction de la non linéarité due aux augets permet au contraire de structurer parfaitement cette variable (expérience 2 et 3, $v = 0{,}2\ mm$) alors qu'elle n'a pas été apprise, que ce soit pour un temps d'agrégation T d'une minute ou de 6 minutes. Les valeurs fortes de la variable (correspondant à une forte proportion de zéros) sont situées dans la partie droite des deux cartes topologiques (expériences 2 et 3) dans une zone labellisée stratiforme, en bon accord avec le fait que les événements stratiforme génèrent des basculements moins fréquents.

3.8. Conclusion

Nous avons montré dans ce chapitre que les effets du temps d'agrégation T et du volume des augets v sont intimement liés. L'étude sur les séries entières a montré que le type d'instrument utilisé a un effet important sur la mesure de l'occurrence de pluie qui est à mettre en regard de celui tout aussi important observé sur la distribution des hauteurs d'eau. La mesure de l'occurrence de pluie est particulièrement sensible pour les temps d'agrégation infra-horaire et reste liée à la configuration instrumentale. De même la distribution des hauteurs d'eau inférieures à quelques volumes d'auget est très sensible à la configuration instrumentale. Pour les temps d'agrégation T inférieur à 30 min, l'écart-type qui caractérise la variabilité intra-événement est aussi très sensible au type d'instrument utilisé. De plus, pour les temps d'agrégation plus importants, il (l'écart-type) caractérise la variabilité inter-événement et n'est plus sensible au type d'instrument. Les valeurs maximales des séries entières correspondent à des évènements très intenses dont l'observation n'est pas sensible au volume des augets.

En fonction du type d'instrument dont sont issues les séries entières analysées, le choix du temps d'agrégation devra être dans la mesure du possible adapté au volume d'auget. La minimisation de la somme des erreurs quadratiques engendrées par l'agrégation temporelle et

par le volume d'auget permet d'obtenir empiriquement l'équation 3.16 (resp. 3.17). Cette relation détermine le volume d'auget maximal pour un temps d'agrégation donné (resp. le temps d'agrégation minimal pour un volume d'auget donné). Ces équations empiriques sont caractéristiques de la région d'Île de France, dans laquelle la prédominance d'évènements stratiformes de faibles intensités peut induire des erreurs de quantification importantes. Appliquée à d'autres séries observées par des disdromètres installés dans d'autres régions, la méthodologie proposée pourra permettre de définir les équations de ce type pour les différentes régions.

Le plus souvent en pratique, le pas de temps et le volume d'auget sont fixés par des contraintes extérieures. Les résultats présentés dans le Tableau 3.3 permettent de quantifier la perte d'information sur la variance due à l'agrégation temporelle, on note par exemple qu'un pas de temps journalier ne permet de conserver que 11% de la variabilité contenue dans une série horaire et une série horaire ne contient elle-même que 26 % de la série à une minute. Le pas de temps journalier ne conserve donc que 2,8% de la variabilité à une minute. De manière analogue le Tableau 3.4 permet de quantifier la perte d'information sur l'occurrence due à la discrétisation par l'auget. Les pluviomètres les plus courants dont l'auget a un volume de 0.1 ou 0.2 mm ne permettent de conserver que 16 ou 9% de l'information relative à l'occurrence des précipitations.

Dans la seconde partie de ce chapitre dédiée à l'effet du type d'instrument (notamment le volume d'auget) sur l'analyse par événement. On vérifie que pour des temps d'agrégation inférieur à une heure la définition des événements reste stable. Trois des variables caractéristiques des évènements identifiés au chapitre 2 (écart-type σ_R, intensité maximale R_{max} et variation absolue du taux précipitant P_{C1}) sont très sensibles aux changements des paramètres (T, v). L'étude de la sensibilité par type d'événement montre clairement que l'agrégation temporelle a un effet particulièrement significatif sur la caractérisation des évènements convectifs, seul un temps d'agrégation très faible permet de conserver l'information sur la variabilité et la valeur maximale des intensités de ce type d'évènements. Ces évènements sont par contre peu sensibles à la discrétisation par le volume d'auget. Pour les évènements stratiformes de faible intensités et les bruines, au contraire, on observe peu de sensibilité au temps d'agrégation mais une perte d'information considérable liée à la discrétisation par le volume d'auget. L'étude par événement à l'aide d'un pluviomètre est rendue délicate car trop dépendante du choix de *T* et *v.* Les 5 variables d'apprentissage sélectionnées au chapitre 2 ont une sensibilité dépendante de la nature des évènements considérés. Cela conduit à une modification importante de la topologie du nuage de point des évènements représenté dans l'espace de ces 5 variables. Il semble illusoire de déterminer des caractéristiques à la fois stables

par rapport aux changements des paramètres (T, v) et discriminantes par rapport aux processus physiques sous-jacents aux évènements de précipitations.

La méthode de sélection des variables présentée au chapitre précédent (2) pourrait permettre de vérifier si les variables adéquates sont indépendantes du couple de paramètres (T, v). Une première tentative avait montré que les caractéristiques sélectionnées (discriminantes) ne sont pas les mêmes pour des données du pluviomètre à auget ($v = 0.2mm$) et des temps d'agrégations importants. De plus, le jeu de caractéristiques obtenu par cette approche risque d'être dépendant de la région d'étude ce qui peut poser des problèmes pratiques.

Dans la suite de cette étude nous chercherons à nous affranchir de la description par caractéristique des séries temporelles en proposant une mesure de similarité entre séries.

Chapitre 4 : comparaison des séries temporelles de pluie par mesure de similarité

4.1. Introduction

Au chapitre 2, nous avons présenté une description des événements de pluie à partir d'un nombre réduit de variables caractéristiques. Nous avons en effet montré, qu'à partir des cinq variables : durée de l'événement D_e(#1), écart-type des taux précipitants σ_R(#9), maximum des taux précipitants R_{max}(#11), coefficient de variation du taux précipitants d'ordre 0.5 P_{C1}(#14) et cumul d'eau R_d (#13), il est possible de décrire relativement bien un événement de pluie. Cette approche a été développée à partir d'intensités de pluie collectées par un disdromètre à une résolution temporelle équivalente au pas de temps $T = 1min$.

Nous avons montré par la suite au chapitre 3 que l'utilisation de ces mêmes cinq variables estimées à partir de pluviomètres à auget à une résolution plus grossière ne permet pas de discriminer correctement les différents type d'événements de pluie. La cause principale, étant la discrétisation apportée par les augets qui nécessitent l'utilisation de temps d'intégration relativement long pour l'obtention de volume d'eau représentatifs. Cet accroissement du temps d'intégration entraine une perte importante de l'information sur la variabilité de l'intensité de pluie réduisant ainsi la capacité des variables caractéristiques à représenter un événement de pluie.

Dans le présent chapitre nous nous intéressons à une approche complétement différente qui vise à comparer deux à deux des séries temporelles d'intensité de pluie. Contrairement à l'étude présentée au chapitre 2, l'avantage de l'approche considérée est qu'elle ne nécessite pas l'extraction de variables caractéristiques, il s'agit de définir une mesure de dissimilarité entre deux séries de pluie globales ou partielles (un ou plusieurs événements). On peut donc s'attendre à ce qu'une classification basée sur une telle mesure soit moins sujette au problème de discrétisation dû aux augets. Le principe retenu est celui dit de la déformation temporelle dynamique (Dynamic Time Warping, DTW). A partir du concept de base de la temporisation dynamique (DTW) développé initialement pour le traitement de la parole qui consiste à associer les échantillons d'une série temporelle à ceux d'une autre en déformant le temps afin que la distance entre les deux séries chronologiques soit minimale, nous avons adapté une mesure multi-échelles (IMs-DTW) de dissimilarité aux spécificités des séries temporelles de précipitation (intermittence et multifractalité). La structure de l'alignement obtenue (que nous appellerons « path » dans la suite) n'est pas le chemin optimal global mais plutôt le chemin optimal sous des contraintes multi-échelles, qui respecte l'indépendance entre des événements de pluie distincts qui ne doivent pas être associés.

L'article ci-après est publié dans le « Journal of Data Science and Analytics ». Il montre l'intérêt de la mise en œuvre de la DTW dans un contexte multi-échelles, cadre qui permet de bien prendre en compte la présence de zéros et ainsi la faculté de distinguer correctement les événements de pluie.

La section 1 expose la problématique et la spécificité des séries chronologiques de précipitation dont les structures ne peuvent être analysées qu'à fine échelle mais qui sont alors constituées d'évènements de pluie séparés par des séries souvent très longues de valeurs nulles. La section 2 présente l'algorithme DTW (Dynamic Time Warping) et des variantes qui ont été proposées pour accélérer les calculs et mieux contrôler les chemins possibles. L'algorithme multi-échelle retenu permet un appariement entre évènements vus sur les deux séries à grande échelle et un appariement des pics d'intensité à l'intérieur des évènements appariés. Les contraintes permettent de réduire le temps de calcul et évitent l'association irréaliste de pic d'intensité d'un événement à un autre événement sans rapport avec le premier. Au final l'application de l'algorithme fournit une fonction de déformation de l'axe du temps appelée 'path' ou structure d'alignement et une mesure des différences entre les intensités de pluie des deux séries considérées après appariement appelée dissimilarité. La section 3 concerne l'application aux séries de précipitation. La robustesse de l'algorithme est vérifiée. Une étude de cas où on applique l'algorithme à un ensemble de séries (comparaison des paires), montre que l'analyse du « path » constitue un bon outil de suivi des cellules de pluie alors que l'analyse de dissimilarité fournit des informations sur la déformation de la cellule de pluie lorsqu'elle se déplace.

Iterative Multiscale Dynamic time warping (IMs-DTW): a tool for rainfall time series comparison

Dilmi M. D., Mallet C., Barthès L., Chazottes A. International Journal of Data Science and Analytics, 2019, DOI : 10.1007/s41060-019-00193-1.

4.2. Article: Iterative Multiscale Dynamic Time Warping (IMs-DTW): A tool for rainfall time series comparison

Mohamed Djallel Dilmi[1], Laurent Barthes[1], Cécile Mallet[1], Aymeric Chazottes[1]

Abstract
In many domains, such as weather forecasting, hydrology or civil protection, it is an important issue to characterize rainfall variability and intermittency in, either or both, a given time period or area. A variety of sensors, for instance, rain gauges, weather radars, and satellites are widely used for this purpose. Techniques to establish the similarity between rainfall time series are commonly based on the comparison of some extracted characteristic parameters (cumulative rainfall height, extreme values, rain occurrence, mean rain rate, etc.). The present study focuses on the development of a tool allowing to compare directly rainfall time series at a fine temporal scale. It allows quantifying the dissimilarity between the time series and determining a non-linear relationship between their time axes. This study presents an algorithm based on a Multiscale Dynamic Time Warping (MsDTW) approach, it is based on the DTW algorithm applied on an iterative multiscale framework we called IMs-DTW. This proposed algorithm is well suited for rain time series allowing point-to-point pairing between pairs of rainfall time. It takes the intermittency and the non-stationarity of the precipitation process into account. An application to measurements observed by four pluviometers located in the Paris area makes it possible to interpret the obtained results and to compare the IMs-DTW with more usual statistical features.

Keywords Multiscale dynamic time warping, rain gauges network, time series comparison, precipitations, warping path, measure of dissimilarity, spatio-temporal variability of the rain.

✉ Mohamed Djallel Dilmi
djallel.dilmi@latmos.ipsl.fr

This work was supported by the CNES/TOSCA ATMEAU_GPM project. The authors gratefully acknowledge Météo France for providing rain gauges data from their Radome network.

Laurent Barthes
laurent.barthes@latmos.ipsl.fr

Cécile Mallet
cecile.mallet@latmos.ipsl.fr

Aymeric Chazottes
aymeric.chazottes@latmos.ipsl.fr

[1] LATMOS/CNRS/UVSQ/Université Paris-Saclay
11 boulevard d'Alembert, 78280 Guyancourt, France

1 Introduction

In recent years, a wide range of data mining techniques has been developed and applied in various fields. Many of these techniques are far more flexible than more classical modeling approaches and could be usefully applied to environmental problems [35]. In the field of the water cycle, rainfall observation is essential in many areas such as climate study, weather forecast, urban hydrology [1] or extreme-event study. The way these observations are used differs depending on the application. Therefore, rain observations are generally used in combination with numerical models representing the physical processes occurring in the atmosphere or in the soil. The basic principle of these numerical models is to represent the state of the atmosphere on a regular 2D or 3D grid composed of pixels of a given size (spatial resolution). Then applying the equations of the model, the temporal evolution is computed at regular time intervals (temporal resolution). In the case of weather forecasts, observational data are meshed to the grid model and averaged at the timescale of the model and then incorporated into the numerical model to improve the prediction (data assimilation). In general, whatever the model considered, the choice of the spatial and temporal scales of the model is a critical point. A too fine spatio-temporal resolution leads to unrealistic computational costs while a too coarse resolution does not allow to represent the local variability of the geophysical fields. In practice, a compromise has to be found, depending on the study. In the case of global studies, given the coarse temporal and spatial resolutions, no compromise is needed. On the contrary, it is much more complicated for regional studies involving finer scales. This is especially true for precipitation, as rainfall has high spatio-temporal variability and is an inherently intermittent process. Hence, as soon as the grid size of the model is larger than a few hundred meters, rainy and dry regions can share the same pixel, which leads to a poor representation of the field. From a temporal point of view, the same problem exists as soon as the temporal resolution is greater than a few minutes. In addition, especially for convective rain events, there is such a

spatial (temporal) variability in rainfall rate that a grid size (temporal resolution) greater than a few hundred of meters (a few minutes) cannot accurately represent the natural variability of precipitation [2, 3].

Some common techniques for analyzing time series like spectral analysis, autoregressive model, principal component analysis or logistic regression are widely used. They allow modeling some features which summarize the time series. However, these techniques are not necessarily well suited for the modeling of rainfall (or rainfall properties) at a fine scale (see for example Cristiano et al. [1] in the context of urban hydrology). Considering rainfall properties, one can notice that a rainfall-rate time series contains dry periods which are composed of a succession of zero values. The percentage of pure zero values in rainfall-rate time series is generally quite large (about 95% in the Paris area for an integration period T of a few minutes). This property is not always in agreement with the assumptions (explicit or implicit) made by commonly used models which often assume, for example, the presence of an additive (Gaussian) noise. Moreover, most of these statistical models are only able to model stationary processes, which is not always the case for rainfall processes.

One method to obtain information regarding the characteristics of precipitation at a particular location and for a specific application is the use of the concept called “rain event” [4]. Such a concept is a convenient way to summarize precipitation time series in a small number of macrophysical features so that they make sense for particular applications. However, different points of view exist concerning the concept of "rain event". For weather studies, a "rain event" is associated with a localized atmospheric disturbance in time and space. For hydrological studies, a "rain event" is rather defined from its ground track, materialized by a measured amount of water. A number of definitions of the concept of rain event [5, 6] have been investigated in the literature [7]. There is a wide variety of criteria for dividing precipitation records into rain events. Dunkerley [8, 9] carried out an analysis of the Inter-Event Time (IET) in order to check the influence of this parameter on the definition of rainfall events and its influence on the average rainfall rate. As highlighted in their study, when determining a value for the IET the compromise between independence of rain events and intra-event variability of rain rates is crucial. The selection of the IET directly impacts the estimated macrophysical features.

In the present study, to overcome the difficulties commonly encountered in the analysis of rainfall data at a given time scale (resolution) or by rain event, we will focus on a tool using a multiscale approach. The main advantage of this approach is that the entire temporal structure is taken into account without any loss of information (unlike averaging), without the arbitrariness of a time scale choice and without the definition of any inter-event time characteristics or macrophysical rain features. The proposed tool, based on the Dynamic Time Warping (DTW) algorithm [13, 14], provides a measure of the dissimilarity between time series. The use of a similarity/dissimilarity measure, looser than a (metric) distance, leads to more appropriate comparisons. Indeed, a metric distance ensures the well-known condition: $d(x, y) = 0 \Leftrightarrow x = y$. This is not appropriate for comparing the time series. Indeed two shifted time series are very similar even though their distance can be very large. Although various similarity/dissimilarity measures have been formulated in the past (see [13, 33, 36] for a complete review), the DTW is particularly interesting for rain studies. Actually, in addition to a measure of dissimilarity between two time series, it provides the temporal lags between them. Such information can be very useful for example to analyze the trajectories of rain cells through a network of rain gauges. In a more general way, time warping techniques could be useful to compare time series but also to deal with pattern recognition, extreme event detection or event clustering. In this context, clustering based on k-medoids or hierarchical clustering seems to be a good approach when combined with a dynamic temporal distortion framework [10, 11]. When applied, for example, on a rain gauge network (urban area for example), this tool could help to deduce some rain spatial features. Finally, it can be noted that, unlike most classical approaches, the DTW does not require a second-order stationarity hypothesis. In the present study, we will apply this tool for the comparison of rain gauge time series located in the same zone of interest. As said previously, a wide variety of similarity measures between time series exists. Among them, the simplest is the Euclidean distance (which is also a measure of dissimilarity). However, this approach is not appropriate when the time series have different time lags or if they have different lengths. In both cases, it leads to a bad estimate of the dissimilarity. Another common technique is the cross-correlation function which is a measure of similarity of the temporal displacement of one time series relative to the other [12]. However, it is only able to catch linear relationships between two stochastic processes and does not allow to take into account some other nonlinear effects like local contraction or dilatation of the time axis (time scaling) of one time series versus the other one. In the framework of precipitation study, these effects can frequently occur. It can be caused by the variation of a rain cell advection velocity (due to horizontal wind) from a rain gauge to another located further. Convection or evaporation have the same effect. Subject to the processes of convection (vertical transport), advection (horizontal winds) or evaporation, the precipitation moves at variable speeds while deforming. Rain is a non-stationary phenomenon. Contrarily to classical approaches the proposed algorithm makes it possible to better take into account precipitation intermittency and non-stationarity.

In this paper we will focus on a method derived from the original Dynamic Time Warping method (DTW)

[14]. This choice as preconized by Sung et al. [15] is conditioned by the good behavior of this method in presence of nonlinear signal transformations like time scaling and/or time shifting which are present in rain time series. They compared four distances namely the DTW, the Earth Mover's Distance (EMD) [16], the Fréchet Distance [17] and Hausdorff Distance [18]. They concluded that the DTW shows the best performances in all of the experiments they conducted.
In this work we consider rainfall rate time series RR measured by a rain gauges network. Each rain gauge allows measuring the cumulated rain height over an integration time period T. This parameter is usually converted to rainfall rate $RR\ [mm.h^{-1}]$ by dividing the rain height by the time period T (generally expressed in hour). Rainfall rate is well suited to describe rain variability and thus to classify rain events into different categories like stratiform or convective events. The information provided by rainfall rates is the main input of all hydrological models. A rain gauge time series, of duration D, is a sequence of N samples where $N = \frac{D}{T}$. From this native rainfall rate time series, we can compute a subsequent rainfall time series RR^c which is averaged over consecutive time periods equal to cT. The parameter c called the resolution factor or the compression rate [14], is an integer value ranging between 1 and N. c values greater than 1 correspond to time series with a coarser temporal resolution. Hence, for a given compression rate c, we denote N^c the number of samples of RR^c :

$$N^c = \frac{D}{cT} \qquad (1)$$

According to these notations, the time series RR^c can be expressed as a sequence of rainfall rates (eq. 2).

$$RR^c = (RR_1^c, RR_2^c, \dots, RR_{N^c}^c) \qquad (2)$$

RR_i^c, the i^{th} element of RR^c is calculated from the native time series RR^1, thanks to the following equation:

$$RR_i^c = \frac{1}{c}\sum_{j=c(i-1)+1}^{ci} RR_j^1 \quad with\ i \in [\![1, N^c]\!] \qquad (3)$$

The time series RR^c is also known as the temporal aggregation of the native time series. By definition, it represents the piecewise aggregate approximation (PAA) of the finer precipitation time series RR^1. Eq. 3 simply states that the precipitation time series is composed of N^c equal-sized "frames" RR_i^c. In the following, we will use equivalently RR^1 and RR as well as N^1 and N. Figure 1 illustrates a precipitation time series at native resolution ($c = 1$) and its PAA for $c = 2$ and 4.

2 The Multiscale Dynamic Time Warping algorithms (MsDTW)

Let's first introduce the Dynamic Time Warping (DTW) algorithm and some of its variants. The DTW was introduced in 1978 by Sakaoe and Chiba [19] as the Dynamic Programming Algorithm (DP-Algorithm).

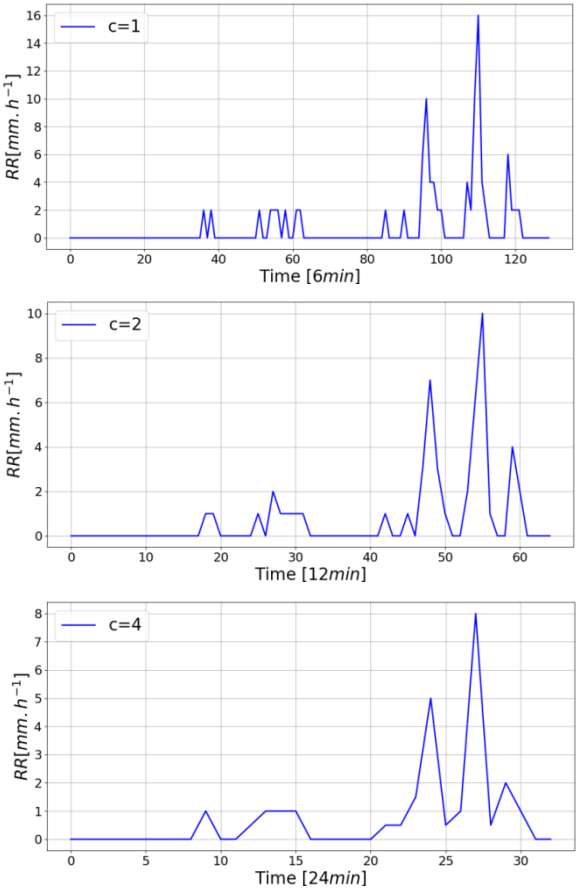


Fig. 1 Example of PAA of rainfall time series for c=1, 2 and 4 obtained from a rain gauge with T= 6 min

An extensive literature exists on either improving the method (especially to make it faster) or applying it in various fields such as speech recognition. The DTW algorithm computes the time distortion needed to align the two time series. This alignment is calculated in order to minimize a distance between two series. Several variants exist which introduce constraints in order to yield to an algorithm less computationally expensive. For a complete description of the DTW algorithm under constraints see for example Zhanga et al. [20]. Here, we only give a brief description of the algorithm in the framework of precipitation time series alignment.
Let's denote A^1 and B^1 two rainfall rate time series observed with the same integration time T (Fig. 2a). The two time series do not necessarily share the same number of samples. Hence we denote respectively N_A^1 and N_B^1 the lengths of the times series A^1 and B^1. As done previously the two time series can be expressed as sequences:

$$\begin{cases} A^1 = (A_1^1, A_2^1, \dots A_i^1, \dots, A_{N_A^1}^1) \\ B^1 = (B_1^1, B_2^1, \dots B_j^1, \dots, B_{N_B^1}^1) \end{cases} \quad (4)$$

Sakoe and Chiba [19, 21, 22] proposed to consider an $i-j$ *plane* (Fig. 2b) where time series A^1 and B^1 are developed respectively along the $i-axis$ and the $j-axis$. This plane is reported in the literature as the distance matrix D [23]. The timing differences between time series A^1 and B^1 can be depicted by a sequence P^1 of K points $p_k^1 = (i_k^1, j_k^1)$ belonging to the distance matrix D:

$$P^1 = (p_1^1, p_2^1, \dots, p_k^1, \dots, p_K^1) \quad (5)$$

It is worth noticing that the length of the sequence K is not known at this point.
This sequence represents a mapping from the time axis of time series A^1 onto that of time series B^1. Sakoe and Chiba [21] called this mapping a warping function and it is currently known as a warping path [23, 24]. When there is no timing difference between two time series (with the same number of samples), the path coincides with the diagonal line $j_k^1 = i_k^1$ on the distance matrix D as shown in Fig.2a and 2b. Otherwise, it deviates further from the diagonal line as the timing difference grows (Fig. 2c and 2d).

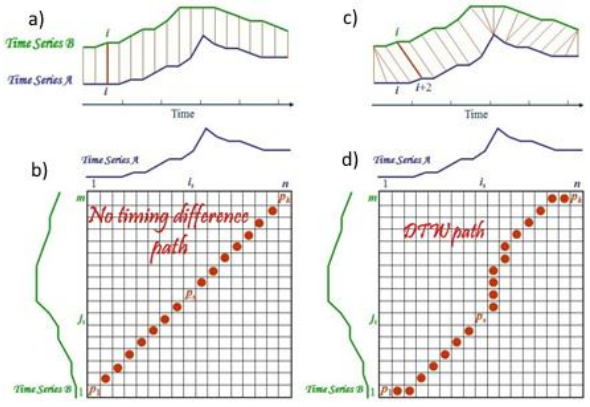


Fig.2 **a** Comparison between two time series with no timing difference. The grey lines represent the alignments between the two time series. **b** Warping path associated to **a.** **c** Comparison between two time series with some timing differences. **d** The warping path associated to **c** (from [23]).

In order to compute an optimal mapping between the two time series, a tool is needed to evaluate how close they are. Let's denote $d(p_k^1)$ the distance corresponding to the k^{th} point of the warping path which corresponds to the distance between two samples $A_{i_k}^1$ and $B_{j_k}^1$. In our case, we simply chose a quadratic distance:

$$d(p_k^1) = \left(A_{i_k}^1 - B_{j_k}^1\right)^2 \quad (6)$$

For the DTW, we denote d_{DTW} the time-normalized dissimilarity between two sequences, here, the two rainfall time series A^1 and B^1. This dissimilarity relies on a corresponding optimal warping path P_{DTW}^1. Both the dissimilarity and the warping path were given by Sakoe and Chiba [19] as the solution of an optimization problem:

$$\begin{cases} d_{DTW}(A^1, B^1) = min_P \left[\sqrt{\frac{\sum_{k=1}^K d(p_k^1) \times w_k}{\sum_{k=1}^K w_k}}\right] \\ P_{DTW}^1(A^1, B^1) = Arg\ min_P \left[\frac{\sum_{k=1}^K d(p_k^1) \times w_k}{\sum_{k=1}^K w_k}\right] \end{cases} \quad (7)$$

with w_k the non-negative weighting coefficients introduced intentionally to allow the dissimilarity flexible characteristics. The denominator $\sum_{k=1}^K w_k$ is employed to make the dissimilarity score independent of K, the length of the warping path P_{DTW}^1. Hence, the dissimilarity d_{DTW} is the minimum weighted average on all the possible warping paths.
All warping paths are not necessarily appropriate and, since initially the DTW algorithm was developed for speech recognition, some constraints were added by the authors in accordance with speech features:

1. Boundary conditions: $p_1^1 = (1,1)$ and $p_K^1 = (N_A^1, N_B^1)$. The path starts at the bottom left and ends at the top right.
2. Monotonicity: $i_k^1 - i_{k-1}^1 \geq 0$ and $j_k^1 - j_{k-1}^1 \geq 0$. The path will not turn back on itself, both the i and j indexes either stay the same or increase, they never decrease.
3. Continuity: $i_k^1 - i_{k-1}^1 \leq 1$ and $j_k^1 - j_{k-1}^1 \leq 1$. The path advances one step at a time. Both i and j can only increase by at most 1 on each step along the path.

In many cases, searching for the optimal path may result in undesired effects because the global optimal path may not necessarily be the one desired and may even be unrealistic. As an example, in the presence of a succession of constant values (series of null values corresponding to dry periods in our case), a large number of points of one time series is mapped to a single point of the other one. A common way to overcome this problem is to restrict the warping path. This is done in such a way that it has to follow a direction in the neighborhood of the diagonal. To do so two additional constraints are commonly used:

4. Warping window condition: $|i_k^1 - j_k^1| \leq \delta$ where δ is a threshold restricting the path. It enforces the recursion to stop at a certain depth. This constraint is known as Sakoe-Chiba band [19] (Fig. 3a). Besides limiting extreme or degenerate mappings, it allows to speed-up the DTW distance calculation.
5. Slope constraint condition: $\frac{j_{k+t}^1 - j_k^1}{i_{k+t}^1 - i_k^1} \leq t$ and $\frac{i_{k+s}^1 - i_k^1}{j_{k+s}^1 - j_k^1} \leq s$. For a warping path, the slope should be neither too steep nor too gentle. Here, in a sequence of k consecutive points of the warping path, for one step in the i-direction, we are allowed t steps in the j-direction. Likewise, for one step in the j-direction, we are allowed s steps in the i-direction [19].

The constraints proposed above were used in the first versions of the dynamic time warping algorithm. Since then various modifications have been proposed to speed up the DTW computations as well as to better control the possible routes of the warping paths. Itakura [25] proposed another constraint known as the Itakura parallelogram (Fig. 3b).
With such constraints, the quality of an alignment depends heavily on the choice of one or more parameters, which is quite subjective. Moreover, as pointed by Cassisi and al. [13] the use of such constraints does not guarantee a good alignment between the two time series since it does not allow the optimal warping path to leave the region as defined in steps 4 and 5.

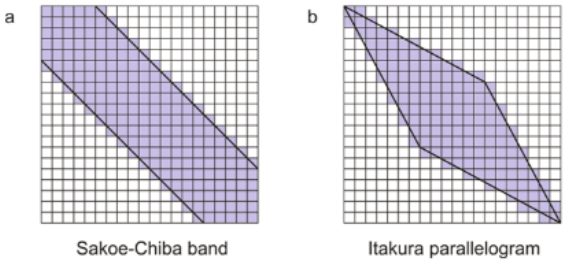


Fig. 3 Examples of global constraints. **a** Sakoe-Chiba band. **b** Itakura parallelogram (from Cassisi et al [13])

Since the objective function in eq. 7 is a rational expression, its minimization is an unwieldy problem. However, if the denominator $\sum_{k=1}^{K} w_k$ in eq. 7 (called normalization coefficient) is independent of the warping path P^1. It can be considered constant and equal to L:

$$\forall P^1, \sum_{k=1}^{K} w_k = L \qquad (8)$$

It can be put out of the brackets, resulting in the following equation:

$$d_{DTW}(A^1, B^1) = \frac{1}{\sqrt{L}} min_P\left[\sqrt{\sum_{k=1}^{K} d(p_k^1) \times w_k}\right] \qquad (9)$$

With this simplification [19], the warping path P_{DTW}^1 can be (and is traditionally) computed using dynamic programming [26] with an $O(N_A^1 N_B^1)$ "quadratic" complexity. To enable this simplification several weighting coefficient definitions were proposed: symmetric, asymmetric (see a complete discussion in [19, 27]). All these weighting coefficient definitions share the same disadvantage: they favor a type of path to the others.
In the other hand, without any a priori information, we are looking for a weighting coefficient definition that does not suffer from such a disadvantage and evaluates equally the paths. In this case, the obvious definition for the weighting coefficients is:

$$\forall\, k, w_k = 1 \quad so \quad \sum_{k=1}^{K} w_k = K \qquad (10)$$

This brings the minimization of the objective function back to the problem encountered previously in eq. 7 (i.e. the dependence of the dissimilarity score on *K*, the unknown number of points on the warping path).
One way to overcome this problem is to reformulate the objective function keeping the same concept. To simplify the minimization computation we replace *K* by a known value of the same magnitude. We chose a normalized quasi-symmetric form of the DTW [23] reformulated below:

$$\begin{cases} d_{DTW}(A^1, B^1) = \sqrt{\frac{2}{(N_A^1 + N_B^1)}}\, min_P\left[\sqrt{\sum_{k=1}^{K} d(p_k^1)}\right] \\ P_{DTW}^1(A^1, B^1) = Arg\ min_P[\sum_{k=1}^{K} d(p_k^1)] \end{cases} \qquad (11)$$

This optimization problem can be computed using dynamic programming. It has the following properties:

1. If $N_A^1 = N_B^1 = N$ then $d_{DTW}(A^1, B^1) = \sqrt{\frac{1}{N}\sum_{k=1}^{K} d(p_k^1)}$
2. The d_{DTW} is less sensitive to the difference between the lengths of the two series N_A^1 and N_B^1
3. Consequence of eq. 11 : If $N_A^1 = N_B^1 = 1$ then $d_{DTW}(A^1, B^1) = \sqrt{(A_1^1 - B_1^1)^2} = |A_1^1 - B_1^1|$

To speed-up the DTW Keogh and Pazzani [14] took advantage of the fact that we can efficiently approximate most time series by a piecewise aggregate approximation (PAA), so they proposed to apply the DTW on the time series at a coarser resolution $c > 1$ rather than at the native resolution ($c = 1$). They advanced that:

$$d_{DTW}(A^1, B^1) \cong d_{DTW}(A^c, B^c) \qquad (12)$$

They called their algorithm Piecewise Dynamic Time Warping (PDTW). One important limitation of this approach, however, is that the user must carefully choose the compression rate parameter c. Indeed, a too coarse resolution can lead to an inaccurate or even completely useless warping path [24, 28]. Although this approach may be interesting for processes characterized by a low, temporal or spatial, variability, this is not valid in the case of rain. Indeed rain is well known to exhibit a multifractal behavior [2, 38, 39] characterized by variability that can greatly increase with the resolution.
This limitation motivates other approaches such as the Multiscale DTW (hereafter MsDTW) approach which is based on a multilevel mapping achieved through the use of several resolutions. In this regard, Chu et al. [29] presented an algorithm named Iterative Deepening Dynamic Time Warping (IDDTW). The basic principle is to iteratively apply the Piecewise Dynamic Time Warping (PDTW) for different resolutions starting at a very coarse resolution. At each iteration, this algorithm decides whether to apply the PDTW to a higher resolution or to keep the current PDTW approximation.

Unfortunately, IDDTW like PDTW is not acceptable for precise alignment of the time series. Later, Salvador et al. [30] proposed the FastDTW algorithm. The authors state that : "A multilevel approach works well if a large problem is difficult to solve all at once, but partial solutions can effectively be refined at different levels of resolution. The dynamic time warping problem can also be solved with a multilevel approach". This approach avoids applying the brute force of the standard DTW algorithm by using the multiscale framework.
The FastDTW algorithm can be divided into four steps:

1- Initialization: The two time series are initially averaged to a low resolution ($c >> 1$) using a piecewise aggregate approximation (PAA). The DTW algorithm is run (using constraints 1 to 3) and a warping path is found for the current resolution (solid line in Fig. 4a).
2- A projection operator projects this warping path to the next higher resolution giving a list of cells defining a new constraint that only these cells will be evaluated by the DTW algorithm (dark gray cells in Fig. 4b).
3- Aware that the entire optimal warp path may not be contained within the projected path, Salvador et al. [30] have released a little bit this constraint by allowing an additional number of cells to be evaluated (light grey cells in Fig. 4b). For that, they introduced a radius parameter r which controls the additional number of cells on each side of the projected path that will also be evaluated.
4- The DTW algorithm is run with this released constraint (and constraints 1 to 3) and an optimal warp path is found. This optimal warp path is then used to find a new released constraint for the next higher resolution. The procedure is repeated (go to step 2) until the full resolution is reached.

The Fig. 4 (from [30]) provides an illustration of the iterative process for a resolution factor c equal, respectively, to 8, 4, 2 and 1. In this figure, the solid black line represents the optimal warping path for the current resolution factor c. In this example, the optimal path does not move too far from the local diagonal path. However, with a radius r greater than 1, the algorithm allows going far from the local diagonal. This case is illustrated in Fig. 5a where the FastDTW has been applied to two rain gauges time series recorded in two cities (Trappes and Villacoublay) located near Paris. The FastDTW provides good alignments except for spurious peaks. This is the case for example at the 205 time index where a peak in the Trappes time series is associated with another peak in the Villacoublay time series at the time index 420. The global optimal path P^1_{DTW} leads to a shift of 215 lags between the two peaks, which correspond to a delay of 215x6 min = 1290 min, i.e. more than 21 hours. This alignment is unrealistic considering the distance between the two cities (15 km). This kind of situation will occur whenever the rain is observed by a rain gauge but not by the other. In other words, the trajectory of the rain cell meets only one of the two rain gauges.
We want to avoid the unrealistic association of a spurious parasitic peak of one time series to an unrelated rain event from the other. Hence, instead of the global optimal path, we would better choose a local minimum path that respects the independence between distant rain events that should not be linked. Therefore, we only want to consider the intra-rain event deformation mapping a rain event to another. Purposely we set the radius r to 0. This corresponds to removing the third step of the FastDTW algorithm and therefore not to release the constraint defined in step 2. In contrast with Fig. 5a, Fig 5b shows that the peak at time index 205 in the Trappes time series is now associated with a zero value in the Villacoublay time series. It would be much more realistic if it was associated with the peak located at time index 420 instead. We are aware that the found path is not necessarily the global optimal path but rather an optimal path under multiscale constraints. In the following, this particular configuration will be called IMs-DTW for Iterative-Multiscale DTW. For the sake of simplicity, the optimal warping path will be denoted by P_{DTW} instead of P^1_{IMSDTW}. Finally, it can be noted that even if it was not the main objective, a radius defined to zero considerably reduces the calculation time.

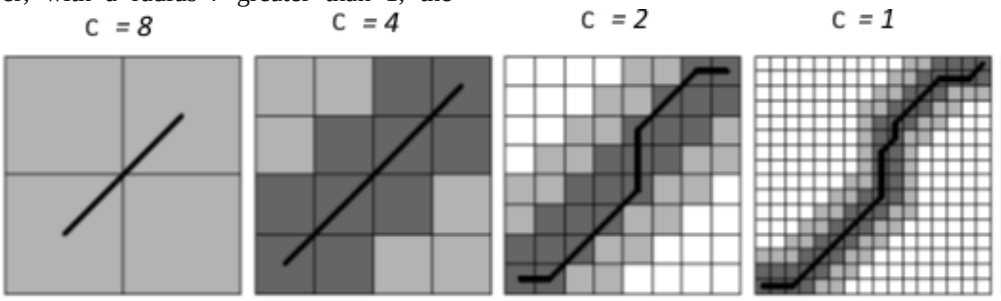


Fig. 4 Illustration of the Salvador et al. algorithm [30]. Dark gray squares are cells that will be evaluated by the constrained DTW. They are derived from the previous step by a "projection" operator. Light grey cells correspond to the released constraint for a radius equal to 1 square. They are also evaluated by the constraint DTW algorithm. White cells are not taken into account by the DTW algorithm. The solid line corresponds to the derived optimal path.

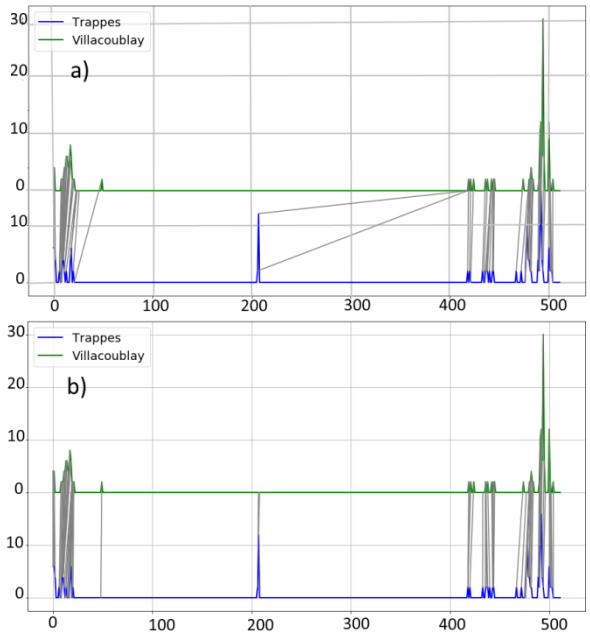


Fig. 5 **a** FastDTW alignments for two rain gauges time series measured in the cities of Trappes and Villacoublay with a time resolution of T=6 min. The grey lines represent the alignments between the two time series. **b** Same figure but for IMs-DTW alignments

3 The Iterative Multiscale DTW algorithm

3.1 Experiment outline

In order to evaluate the efficiency of the IMs-DTW algorithm on precipitation time series, several experiments were conducted. Here we present an experiment for a precipitation time series resulting from 10-days of tipping bucket rain gauge measurements, recorded close to the city of Trappes (27 km southwest of Paris) between June 7, 2009, and June 17, 2009. This time series called reference time series in the following is characterized by an integration time T equals to 6 minutes and by a tipping-bucket volume corresponding to an equivalent rain height of h = 0.2 mm.

This time series of precipitation (Fig. 6a) is characterized by four early rain events between June 7 and June 11. Then a dry period of several days is followed by a last rainy episode. Hereafter, we call this precipitation time series A which corresponds to A^1 of section 2 (for sake of clarity the resolution factor exponent $c = 1$ will be omitted in the following).

This 10-day period is particularly interesting since it regroups both several short dry periods (few hours or less) and a long dry period (a few days). It will allow us to show the behavior of the IMs-DTW algorithm in different meteorological situations. In the next subsection, this time series is used to illustrate the ability of the algorithm to map rainfall time series through the estimated warping path.

3.2 Robustness of the IMs-DTW on precipitation time series

To demonstrate the robustness of the IMs-DTW algorithm we simulate a set of transformed time series. They were built from the reference time series A by applying three kinds of transformations namely: adding lags, varying rain rates by multiplying non-zero rainfall rates values by a random value and finally inserting some spurious rain events. In the following, the IMs-DTW algorithm is applied between the time series A and the transformed time series.

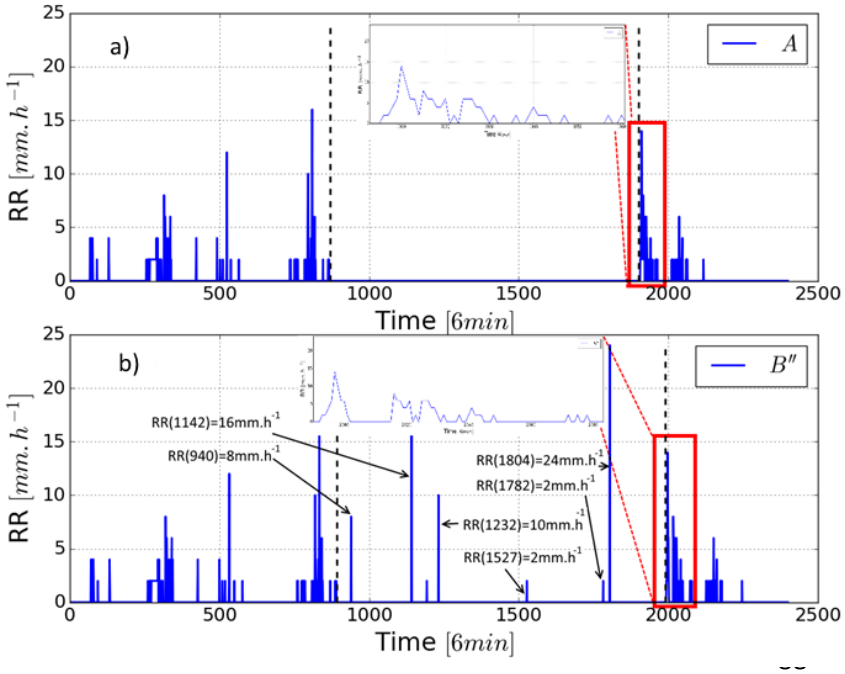


Fig. 6 **a** The reference precipitations time series A recorded in Trappes between June 7, 2009 and June 17, 2009 with an integration time T = 6 min.
b A realization of the transformed time series B'' built from the reference precipitation time series A. Black arrows indicate inserted rain events. The event at time index 2000 is zoomed.

To ensure that we cover a wide range of situations, we have simulated 1000 transformed time series with the following 3-step procedure:

1. The first transformation is a time shifting. It can be seen has the time lags introduced by a fluctuating wind (advection velocity) transporting a frozen rain event from a place to another. At each time step the time series B is delayed from A by an offset $h(i)$ corresponding to an accumulation of random time lags α_j. This shifted rainfall time series B is computed by adding zero rain rate values. It can be expressed by Eq. 13:

$$\begin{cases} B_i = A_{i-h(i)} \\ B_{i-h(i)-y(i)} = 0 \quad \text{when } \alpha_i \neq 0 \end{cases} \qquad (13)$$

With the sequence $y(i) = (1, \dots, \alpha_i)$ and the sum $h(i) = \sum_{j=1}^{i} \alpha_j(\lambda)$ where $\alpha_j(\lambda)$ is drawn according to a poisson law of parameter λ (λ=0.1 in our case).

2. The second transformation is an amplitude modulation. Since a rain event is not frozen when it is transported/advected from a place to another, we have to modulate the values of the rain rates to take the variability of rain into account. Hence, the B' transformed time series is computed by a correction to the non-zero rainfall rates of an already shifted B time series. To exhibit significant rain rate variations from A, the non-null rainfall rate values are multiplied by a random value β_i drawn uniformly between 0.5 and 2. This new type of time series is denoted B' :

$$B'_i = \beta_i B_i \quad when\ B_i \neq 0 \qquad (14)$$

This transformation does not disrupt dry periods since only non-zero rainfall rates are considered.

3. The third transformation adds spurious rain events. Finally, as explained in the previous section, a rain event can be seen by a rain gauge in a specific area but not by another one located a few kilometers away. For this reason, the transformed time series B'' is derived from B' by inserting spurious rain events. The number of added rain events is drawn according to a uniform law between 2 and 30. Figure 6b shows a realization of a transformed time series B''.

To evaluate the IMs-DTW performance, we first analyzed the warping paths $P_{DTW}(A,B)$ linking the time series A to the different realizations of the shifted time series B. As expected, thanks to the multiscale constraints, the IMs-DTW was able to find the right warping paths whatever the added time delays. The calculated $d_{DTW}(A,B)$ dissimilarity, based on the $P_{DTW}(A,B)$ warping path, manifests two behaviors depending on where the time delays were added:

- Case 1: a time series B is generated as mentioned above but time shifts only occur inside a dry event (consecutive zero sequences). In other words, the rain event patterns, i.e. non zero sequences, are preserved. In this situation, as expected the dissimilarity between A and B is always null: $d(A,B) = 0$.

- Case 2: a time series B is generated but this time one or more zero rain rate values are added inside the rain events. (A zero is added in between consecutive non-zero rain rate values.) It creates intra-event dry periods, therefore, changing event durations, and generally modifying the rain event properties. These intra-event dry periods are expected to be associated to non-zero rain rates. Consequently, the $d_{DTW}(A,B)$ dissimilarity is different from 0. This case is illustrated by the event at the time index 2000 in Fig. 6b. In Fig. 7a the blue curve shows an example of the estimated $P_{DTW}(A,B)$ warping path. One can see that during inter-event dry periods, the path remains on a local diagonal (see, for example, the time period between 880 and 1900). The associated lags are, consequently, constant (Fig. 7b). Similarly, when the second type of transformed time series B' was tested, the IMs-DTW was also able to find the right paths (not shown).

Now let's consider the third type of transformation (transformed time series B'') in which some spurious rain events were inserted. Again two situations have to be considered:

- Case 1: In time series B'', the inserted rain event is "far" (few hours) from the rain events in A. In this situation, the inserted rain event is considered by the algorithm to exist only in time series B''. Therefore it is associated with zero values in time series A. This case is illustrated for example in Fig. 6a and b for index time ranging from 1142 to 1232. (In this case the rain events are present only in time series B''.) The figure7b shows that during this period the warping path remains on the local diagonal and consequently the rain event is associated with zero values in A.

- Case 2: In time series B'', the inserted rain event is "close" to a rain event in A. In this case, the IMS-DTW considers the two rain events as a single event and the time separating them as an intra-event dry period. The inserted rain event is then associated with the closest rain event in A. This case is visible in Fig. 7 in which the event beginning at time index 940 in B'' has been grouped with an event of time series A at time index 863.

Finally, for the 1000 simulations, the IMS-DTW proposed acceptable warping paths P_{DTW} that were consistent with what might be expected from the observation of rainfall. Indeed, two rainy events occurring with a time lag of more than a few hours can be considered independent. The presence of a rainy

episode in a time series does not necessarily imply its presence/absence in the other one.

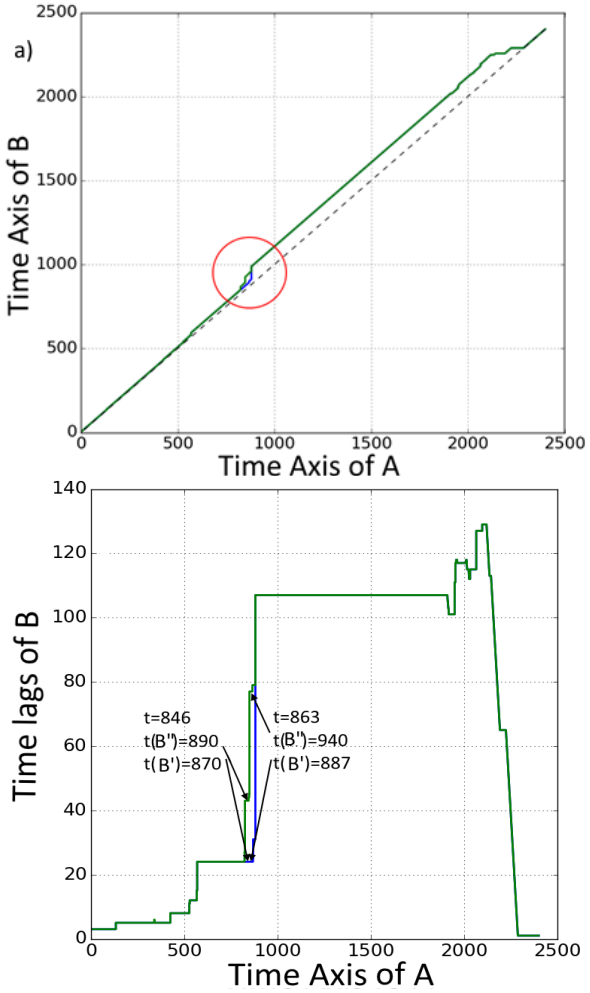


Fig. 7 **a** In blue, the warping path $P_{DTW}(A,B)$ between the time series A time series B. In green $P_{DTW}(A,B'')$ the warping path between time series A and time series B'' shown in Fig. 6b. The dashed line represents the diagonal line. **b** The corresponding time lags.

In this subsection, we have shown the good behavior of the IMS-DTW algorithm in various simulated situations based on a real time series (time series A) and the main transformations encountered in the context of rainfall observation. In the following, the study focuses on how the method can be used to compare two real time series.

3.3 Case study

In this section, we wish to assess the performance of the IMS-DTW on real rainfall time series observed by rain gauges located in the same urban area. We chose to present an analysis that is performed on four relatively short precipitation time series (13 hours) simultaneously recorded by four meteorological stations (Trappes, Le Bourget, Roissy, and Nangis) located near Paris. The measurements occurred on June 10, 2009 between 06:00 and 19:00 with an integration time $T = 6$ min. The figure 8 shows the four time series.

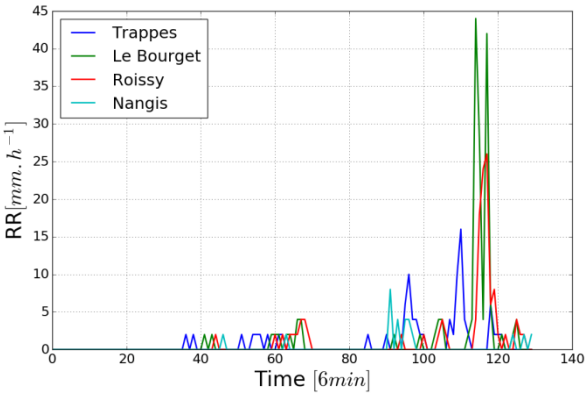


Fig. 8 Precipitations time series recorded in Trappes, Le Bourget, Roissy and Nangis on June 10, 2009 between 06:00 and 19:00 with an integration time T=6 min.

The similarities between the rain patterns on the four series suggest that it could be the same rain event recorded by the four stations. This suggestion is partially confirmed by radar images (not shown) which also indicate that the rain cells at the origin of the four recordings move from Southwest to Northeast. Radar images also allow the following observations:

1- The same rain cell was at the origin of the time series recorded at Trappes first and then at Le Bourget, and Roissy. The meteorological station of Trappes is located to the Southwest of Le Bourget and Roissy. On the radar image, the rain cell is thus observed there a little earlier than for the other two. One notes moreover that it seems less intense.

2- A different rain cell was at the origin of the time series recorded at Nangis station which is located in the extreme southeast of the area (60 km from Roissy).

As a result, since Roissy and Le Bourget are close (around 10 km between the two) we expect that their time series should just be slightly delayed. Moreover the lags obtained between Nangis station and the three others have no physical meaning.

To illustrate the benefit of considering a time warping using the IMS-DTW for real precipitation time series comparison we first computed the Euclidian distance widely used in the context of precipitation time series comparison [31]. For A and B representing precipitation time series at the native resolution (i.e. for a resolution factor $c=1$) recorded at the stations S_i and S_j, the Euclidian distance is defined by:

$$d(S_i,S_j) = \sqrt{\sum_{k=1}^{N}(A_k - B_k)^2} \qquad (15)$$

This distance is sensitive to N the length of the time series. To eliminate this undesired effect, we computed a normalized Euclidian distance defined by:

$$d_{NTD}(S_i,S_j) = d(P_{NTD},S_i,S_j) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(A_k - B_k)^2} \quad (16)$$

In eq. 16 we add the term P_{NTD}, for the "No Timing Difference Path", which corresponds to a diagonal warping path, i.e. $p_k^1 = (i_k^1, i_k^1)$. In fact, the normalized Euclidean distance is a special case of the dissimilarity defined by eq. 11 and therefore it is possible to compare these two distances. Similarly, we will denote the correlation coefficient between the two time series $r_{NTD}(S_i,S_j) = r(P_{NTD},S_i,S_j)$, and the maximum of cross-correlation $r_{\tau,Max}(S_i,S_j) = r_{Max}(P_\tau,S_i,S_j)$. P_τ is the warping path associated to the time lag τ_{delay} ($\tau_{delay} = nT$ with T the integration time), this latter corresponds to the time lag maximizing the cross-correlation function. This means that the warping path P_τ is located on the *n*-super/sub diagonal of the distance matrix *D*, i.e:

$$p_k^1 = (i_k^1, j_k^1) = (i_k^1, i_k^1 \mp n) \quad with\ n = \frac{\tau_{delay}}{T}$$

The Normalized Euclidian distance $d_\tau(S_i,S_j) = d(P_\tau,S_i,S_j)$ considering the time delay τ_{delay} was also performed. Table 1 provides the corresponding values for six (S_i,S_j) station pairs while Fig. 9 shows the warping paths P_{NTD} (Fig. 9a) and P_τ (Fig. 9b) for the pairs of stations Trappes/ Le Bourget and Trappes/ Nangis.

The without sliding approach (i.e. based on the P_{NTD} warping path) compares rainy periods to non-rainy periods (see Fig. 9a). The smallest normalized Euclidian distance $d_{NTD}(S_i,S_j)$ (Table 1 column 3) is obtained for the Trappes / Nangis pair while these two stations are the furthest apart. As mentioned before they did not record the same rain cell. In fact, the obtained distances seem to be only sensitive to the rain amount of the two considered time series when the stations are sufficiently distant. This example shows that the Euclidian distance is meaningless in the context of rainfall time series comparison. Concerning the $r_{NTD}(S_i,S_j)$ correlation coefficient (Table 1 column 4), it is close to zero except for the pair Le Bourget / Roissy. There is no linear correlation between the pairs except when the stations are close enough like for the pair previously mentioned. Thus, like the normalized Euclidean distance, the correlation coefficient is ineffective in this context and classical approaches without sliding (P_{NTD}) do not generally allow to identify the similarities/dissimilarities between precipitation time series.

In the second approach, we compare the time series pairs taking into account a shift by a constant time lag τ_{delay} which is equivalent to the use of the warping path P_τ. Very low cross-correlation values (Table 1 column 7) are obtained when Nangis station belongs to a pair. As stated in the without sliding approach, the two recorded rain time series do not come from the same rain cell and are therefore are not correlated. For the three first pairs of stations Trappes/ Le Bourget, Trappes/ Roissy and Le Bourget/ Roissy the time delays τ_{delay} are consistent with both the geographic distances (Table 1 column 2) and the advection velocity observed on the radar images (not shown). In these cases, the maximum of cross-correlation $r_{\tau,Max}(S_i,S_j)$ is thus representative. In addition, for these pairs, the normalized Euclidian distance $d_\tau(S_i,S_j)$ decreased compared to $d_{NTD}(S_i,S_j)$ by a factor ranging from 16 to 50%. Trappes and Roissy are not as dissimilar as suggested by the normalized Euclidian distance. On the other hand, for the stations of Le Bourget and Roissy we expected, given their proximity (10 km), a smaller dissimilarity (5.05 $mm.h^{-1}$). When the Nangis station belongs to a pair the $d_\tau(S_i,S_j)$ distances remain almost unchanged compared to the $d_{NTD}(S_i,S_j)$.

Table 1 Indicators of dissimilarity between the six (S_i,S_j) pairs for P_{NTD}, P_τ and P_{DTW} warping paths.

	Distance [km]	$\mathbf{P_{NTD}}$ warping path		$\mathbf{P_\tau}$ warping path			$\mathbf{P_{DTW}}$ warping path	
Pair of stations $\mathbf{S_i}$, $\mathbf{S_j}$		$d_{NTD}(S_i,S_j)$ [$mm.h^{-1}$]	$r_{NTD}(S_i,S_j)$	τ delay [min]	$d_\tau(S_i,S_j)$ [mm /h]	$r_{\tau,Max}(S_i,S_j)$	$d_{DTW}(S_i,S_j)$ [mm/h]	$r_{DTW}(S_i,S_j)$
Trappes/ Le Bourget	34.78	6.35	-0.05	30	5.05	0.61	3.82	0.81
Trappes / Roissy	44.84	4.24	-0.03	42	2.58	0.75	1.53	0.91
Le Bourget/ Roissy	10.41	4.42	0.65	12	3.84	0.79	3.27	0.83
Trappes/ Nangis	77.96	2.24	0.17	-30	2.14	0.35	1.68	0.64
Le Bourget/ Nangis	60.44	6.08	-0.04	60	6.09	0.21	5.38	0.51
Roissy/ Nangis	61.36	3.86	-0.06	60	3.83	0.15	3.07	0.62

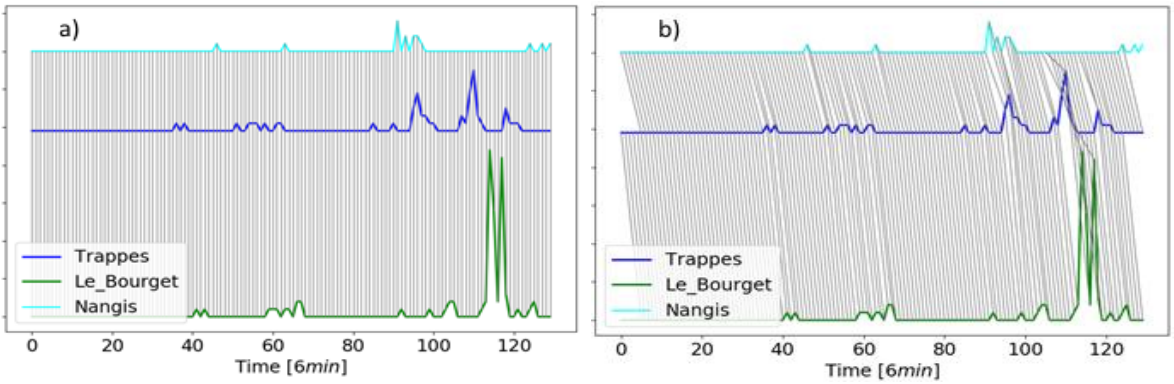


Fig. 9 **a** Sequences P^1 of points $p_k^1 = (i_k^1, j_k^1)$ defined by eq. 6 for the P_{NTD} warping path for the pairs of stations Trappes/ Le Bourget and Trappes/Nangis. The grey lines represent the alignments between the two time series. **b** Same figure than **a** but with the P_τ warping path.

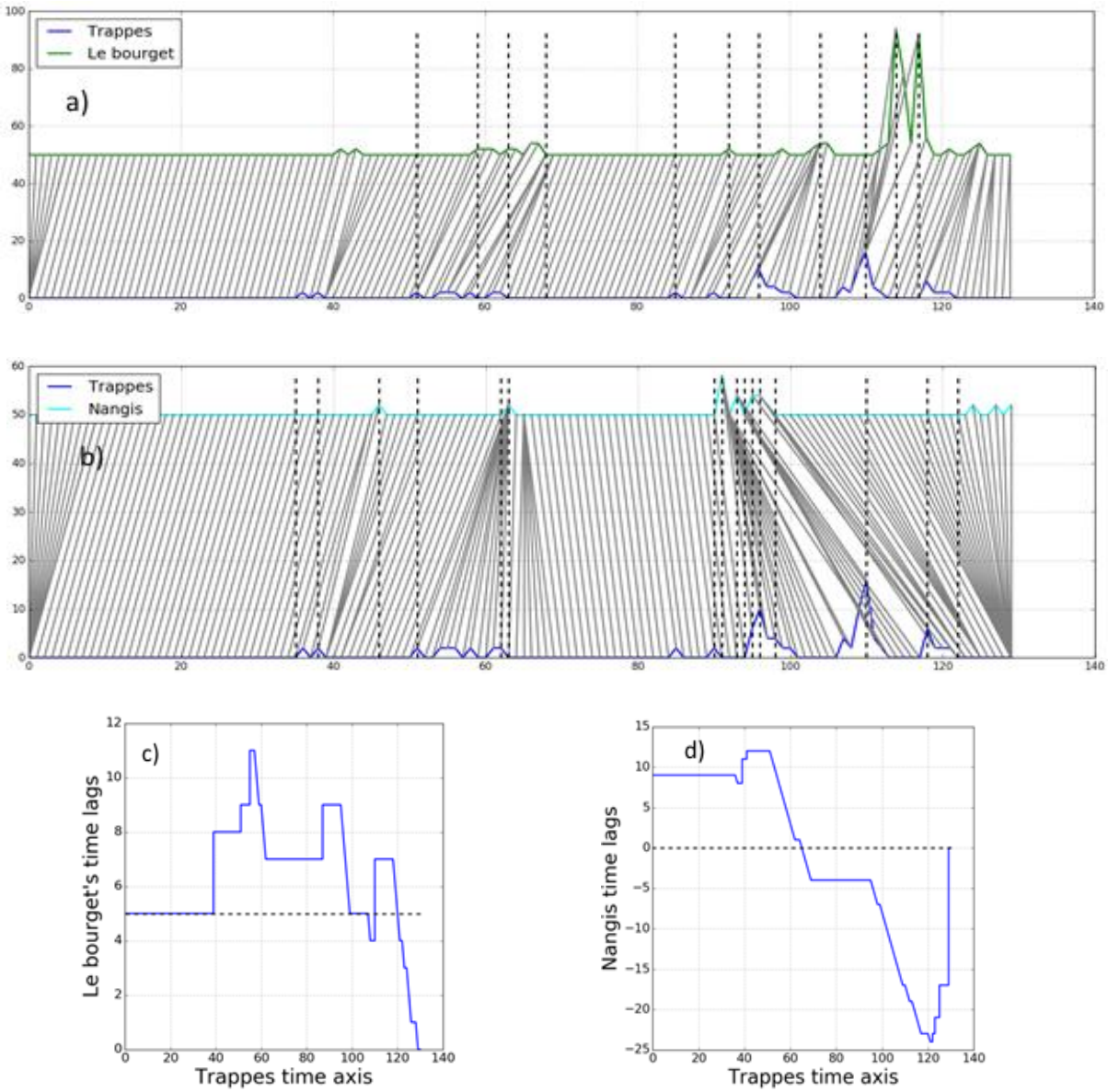


Fig. 10 **a** warping path for the pair of stations Trappes/ Le Bourget. **b** warping path for the pair of stations Trappes/Nangis. **c** the obtained time lags for the pair of station Trappes and Le Bourget. **d** the obtained time lags for the pair of station Trappes and Nangis

Their corresponding $r_{\tau,Max}(S_i,S_j)$ cross-correlations remain meaningless. Finally, we can conclude that the use of a constant time lag (warping path P_τ) does not make it possible to estimate good indicators representing the dissimilarities between the stations.

After outlining these findings, we performed the IMs-DTW on the six different pairs and we analyzed the warping paths P_{DTW}. Figures 10 a and b show the sequences P^1 of points p_k^1 of P_{DTW} for the pairs Trappes/ Le Bourget and Trappes / Nangis.

In the case of the pair Trappes/ Le Bourget, the assumption of a dynamic warp improves the associations of the points p_k^1 of the sequence P^1 unlike with P_τ warping path where the matching respects the general behavior even if there are some association mismatches. For example, the second Le Bourget peak recorded at the 118 time index is associated to a zero rain rate value in the time series of Trappes while the P_{DTW} made a better association by associating the peak in Le Bourget Time series with a peak in the Trappes time series. In a general way, the IMs-DTW was able to find a better matching of all the couples of points (see Fig. 10 c and d). The algorithm realizes good matching, associating together the beginning and the ending times of rainy periods, their peaks and intra/inter-event dry periods independently of the considered situation. Table 1 column 8 provides the distances $d_{DTW}(S_i,S_j)$ for the six (S_i,S_j) pairs while column 9 provides the correlation coefficients $r_{DTW}(S_i,S_j)$.

When considering a dynamic warping path, the compared time series are not as dissimilar as suggested by the classic approaches. For the six pairs, the dissimilarities $d_{DTW}(S_i,S_j)$ decrease and become smaller than the ones calculated with the two other warping paths (P_{NTD} or P_τ). In the same way, the correlation coefficient increase and become much more important especially for the 3 pairs Trappes/ Le Bourget, Trappes/ Roissy, and Le Bourget/ Roissy.

For the first three pairs, where the same rain cell is observed, the IMs-DTW was able to spot the similarities between the time series as depicted by the dissimilarities d_{DTW} or the correlation coefficients r_{DTW}. Indeed whatever the considered pair the dissimilarities are smaller than with P_{NTD} or P_τ warping paths. The same is true for the correlation coefficients, which are higher. For the last three pairs, where the time series represented two different rain cells, the IMs-DTW allows an alignment of the rain patterns. Thus the dissimilarities decreased and the correlation coefficients increased but to a lesser extent.

In a general way, the IMs-DTW will find an optimal path minimizing the distance between two time series, however, the question is: is there a physical sense behind the P_{DTW} warping path? In other words, is it possible to use P_{DTW} to deduce if two rain time series are produced by the same rain cell? In the affirmative, can the deduced delays be linked to the time for the rain cell to move from one station to another? In this case, the dissimilarity $d_{DTW}(S_i,S_j)$ could be interpreted as a measurement of the spatio-temporal evolution of the same rain cell between the different locations. Indeed, if the assumption of a rain cell displacement is verified, the warping path P_{DTW} should have a mean behavior that is more or less comparable to P_τ. Let us introduce the average time difference delay τ_{ATD} which is the average value of the lags estimated from P_{DTW} during a rain event and σ_{ATD} the associated standard deviation. Table 2 allows comparing τ_{ATD} and τ_{delay} (obtained previously in Table 1) for the six (S_i,S_j) pairs. In addition, the estimated advection velocity is estimated by dividing the geographical distance by τ_{ATD}.

Table 2 Time delay τ_{delay}reported from Table 1 , the average time difference τ_{ATD}, the corresponding standard deviations σ_{ATD} and the estimated advection velocity for the 6 pairs (S_i,S_j).

Pair of stations (S_i,S_j)	τ_{delay} [min]	τ_{ATD} [min]	σ_{ATD} [min]	the estimated advection velocity $[m.s^{-1}]$
Trappes/ Le Bourget	30	37	12.66	15.60
Trappes / Roissy	42	45	12.78	16.29
Le Bourget / Roissy	12	8	4.14	21.90
Trappes / Nangis	-30	11	66.53	118.98
Le Bourget / Nangis	60	106	51.78	9.49
Roissy / Nangis	60	97	40.08	10.58

For the three first pairs, as expected the average time difference τ_{ATD} is quite close to the time delays τ_{delay} and the standard deviations σ_{ATD} remain low. In such a way the coefficients of variations $(\sigma_{ATD}/\tau_{ATD})$ are much smaller than unity (respectively 0.34, 0.28, 0.51). In addition, the estimated advection velocities (Table 2 last column) are quite homogeneous and in agreement with the velocities obtained by radar measurements (not shown). For the last three pairs, τ_{ATD} and τ_{delay} are very different and the standard deviations σ_{ATD} are much higher than previously. This example illustrates that the standard deviation σ_{ATD} and the difference τ_{delay}-τ_{ATD} are good indicators to decide whether or not the warping path P_{DTW} is due to the same rain cell and thus can be interpreted as time lags.

In order to characterize and compare the precipitation time series, a large number of features derived from rain rates are commonly used. In Dilmi et al. [32], the authors show that among the many existing parameters a rain event can be fairly well described using only five parameters. Among these parameters, four parameters are commonly used in meteorology or hydrology namely the event duration, the rain rate peak, the rain

event depth, and the standard deviation while the last one called the absolute rain rate variation is less used. The values of these five characteristics are shown in Table 3 for the June 10, 2009, rain event.

Table 3 Main features of the four considered time series

	Trappes	Le Bourget	Roissy	Nangis
Event Duration [6 min]	86	85	84	85
Rain amount [mm]	10	17.4	13.2	3.2
Standard deviation of rain rates [mm.h^{-1}]	0.2536	0.7174	0.44	0.01
Maximum of rain rate [mm.h^{-1}]	16	44	26	8
Absolute rain rate variation of order 0.5	53.20	66.12	52.46	27.21

Except for the event duration, we observed an important discrepancy of the obtained features. The time series with the most dissimilar characteristics being that of Nangis. The pair Trappes /Roissy presents the most similar characteristics, and it is interesting to note that the IMs-DTW provides for this pair the lower dissimilarity (1.53 mm.h^{-1}) and the higher correlation coefficient $r_{DTW}(Trappes, Roissy)$ (0.91). Trappes and Roissy are the most similar rainfall time series. Concerning the pair Roissy / Le Bourget whose stations are located close to each other (10 km) the maximum rain rates values given by Table 3 suggest that the most intense part of the rain cell has passed over the Bourget station but not above Roissy, giving higher rain rates to the former. Nevertheless, their features remain close as suggested by the IMs-DTW dissimilarity (3.27 mm.h^{-1}) and its correlation coefficient $r(P_{DTW}, Bourget, Roissy)$(0.83).
Contrarily to the previous indicators, commonly used, the IMs-DTW dissimilarity takes the temporality of the rain event into account. Even if, not made explicit, it is well suited for physical phenomena such as advection (time shift between series) or diffusion (warping of the time series).

Since the rain is a multiscale phenomenon subject to non-stationarity the time scale are conditioned on the context. The hierarchical fitting of the algorithm enables it to find the similarity where it lies. The algorithm is well balanced since it forbids the released constraint of the FastDTW. Indeed, since rain at a given time scale is conditioned on larger scales, this highlights the coherence through a range of scales. Two time series are similar if they behave coherently both in time scales and in time. This aspect which is not commonly used is well fitted for rain time series which are known to display multifractal properties.

4 Conclusion

The present study focused on the comparison of rainfall time series through the use of the concept of dissimilarity. Unlike conventional approaches that define a number of features to describe a physical process, the time warping approach provides a measure of dissimilarity between two time series without going through this step. This kind of approach was first developed for signal processing and in particular for speech recognition. The basic concept of dynamic time warping (DTW) is to associate the samples of a time series with those of another by warping the time so that the distance between the two time series is minimal.
Precipitation is characterized by a multifractal behavior leading to strong inhomogeneities and strong variability in rainfall rate. Therefore, at a fine resolution, time series obtained from devices located close enough to each other can differ significantly while maintaining common features such as the overall shape of the event. We showed that the DTW in a multiscale framework (IMs-DTW) by comparing rainfall time series at different time scales allows taking into account (at least partially) these features. Indeed, the multiscale approach of the algorithm that constrains fine-scale associations by the association of sequences on a larger scale is well suited to time series composed of subsequences in which scale relations exist, as in multifractal objects, thus allowing distributions of local statistics not to be identically distributed. As another specificity rain time series encompass a great number of zeros among which a set of non-null rainfall rates belonging to a specific time interval defined as a rainfall event. An important issue for precipitation studies is to ensure that non-zero rainfall rates from a rainfall event are not associated with another event. Again, the multiscale approach helps to distinguish rain events from each other and thus allows to correctly associate the samples in each time series.
When considering intra-event rain samples the analysis of the warping path provides useful information, especially for rain cells monitoring. Indeed, we have shown that if the same rain cell is at the origin of the two time series the associated warping path has a regular behavior which can be more or less considered as a time shift corresponding to the travel time of the cell rain from one meteorological station the other. The analysis of the warping path regularity is thus a good indicator to detect a rain cell passing through. When a rain gauge network is available, the calculation of the warping path between each pair of stations allows the identification of the rain cell through the network and thus, analyzing the warping paths offers a good tool for rain cell tracking. Moreover, the dissimilarity analysis

will provide information on the spatio-temporal distortion of the rain cell as it moves.

Finally, since the proposed algorithm is derived from the FastDTW algorithm but with more restrictive constraints, it also significantly reduces the computation time compared to the standard DTW algorithms but also to a certain extent compared to the FastDTW algorithm (radius equal to 1).

This work was primarily dedicated to the ability to use DTW algorithms to provide a pertinent measure of dissimilarity for rainfall time series. In future works, we will focus on applications using the IMs-DTW in the framework of precipitation such as rain cell tracking or rain events clustering [34, 37].

The code source (in Python) and rain gauges data set are available on the GitHub deposit at the following address: https://github.com/djallelDILMI/IMs-DTW.

References

1. Cristiano, E., Veldhuis, M.,Giesen, N.: Spatial and temporal variability of rainfall and their effects on hydrological response in urban areas – a review In Hydrol. Earth Syst. Sci., **21**, 3859–3878 (2017)
2. Verrier, S., de Montera, L., Barthès, L., Mallet, C.: Multifractal analysis of African monsoon rain fields, taking into account the zero rain rates problem, J. of Hydrology, pp. **389**(1),111-120 (2010).
3. Verrier, S., Mallet, C., Barthès, L.: Multiscaling properties of rain in the time domain, taking into account rain support biases Journal of Geophysical Research- Atmospheres, J. Geophys. Res., **116**, doi:10.1029/2011JD015719 (2011).
4. Llasat, M.C.: An objective classification of rainfall events on the basis of their convective features. Application to rainfall intensity in the north east of Spain. In International Journal of climatology, **21**, 1385-1400 (2001).
5. Eagleson, P. S.: Dynamic Hydrology, McGraw-Hill, New York (1970)
6. Brown, B.G., Katz, R. W., and Murphy, A.H.: Statistical analysis of climatological data to characterize erosion potential: 4. Freezing events in eastern Oregon/Washington. Oregon Agricultural Experiment Station Spec. Rep. No. 689, Oregon State University (1984).
7. Larsen, M. L. and Teves, J. B.: Identifying Individual Rain Events with a Dense Disdrometer Network, Advances in Meteorology, ID582782 (2015).
8. Dunkerley, D.: Rain event properties in nature and in rainfall simulation experiments: a comparative review with recommendations for increasingly systematic study and reporting, Hydrological Processes, **22** (22), pp. 4415–4435 (2008a).
9. Dunkerley, D.: Identifying individual rain events from pluviography records: a review with analysis of data from an Australian dryland site, Hydrological Processes, **22**(26), pp. 5024–5036, (2008b).
10. Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.H.: Time-series clustering – A decade review, Information Systems, 53, pp16-38, (2015)
11. Sarda-Espinosa, A.: Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package, R package, https://cran.rproject.Org/web/packages/dtwclust/vignettes/dtwclust.pdf (2017)
12. Pearson, K.: Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. 187: 253–318. ISSN 1364-503X. doi:10.1098/rsta.1896.0007 (1896).
13. Cassisi, C., Montalto, P., Aliotta, M., Cannata, A., Pulvirenti, A.: Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining, Advances in Data Mining Knowledge Discovery and Applications. Adem Karahoca (Ed.), InTech, DOI: 10.5772/49941 (2012).
14. Keogh, E., Pazzani, M.: Scaling up Dynamic Time Warping for Datamining Applications. In Proc. Of the Sixth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, pp.285-289. Boston, Massachuseetts (2000).
15. Sung, P., Syed, Z., Guttag, J.: Quantifying Morphology Changes in Time Series Data with Skew. Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. International Conference on Acoustics, Speech and Signal Processing. 477-480 (2009).
16. Rubner, Y. , Tomasi, C., Guibas, L.J.: A Metric for Distributions with Applications to Image Databases, in Proc. IEEE ICCV, pp.59–66, 1998.
17. Aronov, B., Har-Peled, S., Knauer, C., Wang, Y., Wenk, C.: "Fréchet distance for curves, revisited," in ESA'06, London, UK, pp. 52–63, Springer-Verlag (2006).
18. Huttenlocher, D. P., Klanderman, G.A., Rucklidge, W.J. "Comparing images using the Hausdorff distance," IEEE Trans. PAMI, **15**, no. 9, pp. 850–863 (1993).
19. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoustics, Speech, and Signal Proc., Vol. ASSP-26 (1978).
20. Zhanga, Z. , Tavenardb, R., Baillyb, A., Tangc, X., Tanga, P., Corpetti, T.: Dynamic Time Warping Under Limited Warping Path Length. Information Sciences, **393**, pp 91-107, (2017).
21. Sakoe, H., Chiba, S.: A similarity evaluation of speech patterns by dynamic programming, Dig. Nat. Meeting, Inst. Electron. Comm. Eng. Japan, p. 136 (1970).

22. Sakoe, H.,Chiba, S.: A dynamic programming approach to continuous speech recognition, Proc. 7th ICA, Paper 20 CI3 (1971).
23. Tsiporkova E.: Dynamic Time Warping Algorithm for. PPT presentation available at: http://www.psb.ugent.be/cbd/papers/gentxwarper/DTWAlgorithm.ppt, date of the last visit (2013)
24. Zinke, A., Mayer, D.: Iterative Multi Scale Dynamic Time Warping. Computer Graphics technical reports, CG-2006/1 (2006).
25. Itakura, F.: Minimum Prediction Residual Principle Applied to Speech Recognition. In IEEE Trans. Acoustics, Speech, and Signal Proc. Vol. ASSP-23, pp 52-72 (1975).
26. Bellman, R., Dreyfus, S.: Applied Dynamic Programming. New Jersey: Princeton Univ. Press (1962).
27. Sakoe, H., Chiba, S.: Comparative study of DP-pattern matching techniques for speech recognition, Tech. Group Meeting Speech, Acoust.SOC. Japan,Preprints (S73-22), (1973).
28. Muller, M., Mattes, H., Kurth, F.: An Efficient Multiscale Approach to Audio Synchronization. In Proc. ISMIR, Victoria, Canada, pp. 192-197 (2006)
29. Chu, S., Keogh, E., Hart D., Pazzani, M.: Iterative Deepening Dynamic Time Warping for Time Series. In Proc. Of the Second SIAM Intl. Conf. on Data Mining. Arlington, Virginia (2002).
30. Salvador S., Chan., P.: FastDTW: Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis, **11**(5), 561-580 (2007).
31. Tokay, A., Öztürk, K.: An Experimental Study of the Small-Scale Variability of Rainfall, Journal of Hydrometeorology, 13 (1), pp 351-365 (2012)
32. Dilmi, M. D., Mallet, C., Barthes, L., Chazottes, A.: Data-driven clustering of rain events: microphysics information derived from macro-scale observations. Atmos. Meas. Tech., **10**, 1–18 (2017).
33. Goshtasby A.A.: Similarity and Dissimilarity Measures. In: Image Registration. Advances in Computer Vision and Pattern Recognition. Springer, London (2012)
34. Truong, C.D. & Anh, D.T. A novel clustering based method for time series motif discovery under time warping measure. Int J Data Sci Anal, **4** (2), 113-126 (2017), doi.org/10.1007/s41060-017-0060-3
35. Wang, S. and Eick, C.F.: A data mining framework for environmental and geo-spatial data analysis. Int J Data Sci Anal **5**(2-3), 83-98 (2018). doi.org/10.1007/s41060-017-0075-9
36. van Gennip, Y., Hunter, B., Ma, A. et al. Unsupervised record matching with noisy and incomplete data Int J Data Sci Anal, **6**(2), 109-129 (2018). doi.org/10.1007/s41060-018-0129-7
37. Endo, Y., Toda, H., Nishida, K. et al. Classifying spatial trajectories using representation learning. Int J Data Sci Anal **2**(3-4) 107-117 (2016), doi.org/10.1007/s41060-016-0014-1
38. De Montera, L, Barthes L., Mallet C., Golé P: The Effect of Rain-No Rain Intermittency on the Estimation of the Universal Multifractals Model Parameters. Journal of Hydrometeorology, American Meteorological Society, **10** (2), 493-506 (2009)
39. Akrour N., Chazottes A., Verrier S., Mallet C., Barthès L: Simulation of yearly rainfall time series at microscale resolution with actual properties: Intermittency, scale invariance, and rainfall distribution. Water Resources Research, American Geophysical Union, **51**(9), 7417-7435 (2015).

4.3. Conclusion et synthèse

A partir du concept de base de la déformation temporelle dynamique (DTW) nous avons adapté une mesure multi-échelles (IMs-DTW) de dissimilarité aux spécificités des séries temporelles de précipitations (intermittence et multifractalité). La structure de l'alignement obtenue (ie. path) respecte l'indépendance entre des événements de pluie distincts qui ne doivent pas être associés. L'algorithme obtenu présente un intérêt opérationnel important avec un temps de calcul réduit qui permet son application ultérieure à des séries de précipitations longues observées par un réseau de pluviomètres. Pour une machine équipée d'un processeur i7 2.4GH et d'une mémoire 16Go, la comparaison de deux séries temporelles longues ($L = 1\ année$) à fine échelles ($T = 6\ min\ \ i.e. \Lambda = 87600$), la version classique provoque une surcharge mémoire sans convergence alors que la version optimisée converge en 7 secondes. Pour la comparaison des séries courtes, le gain sur le temps de calcul est considérable, on passe d'un temps de 2 heures à une exécution en temps réel.

Dans l'étude bibliographique réalisée, qui concerne des séries temporelles de nature différentes (principalement en traitement de la parole, de la musique, de la biométrie, …), la structure d'alignement a un rôle purement technique d'appariement des séries. Nous avons montré que dans le cas des séries de précipitations non seulement la dissimilarité elle-même mais également la structure d'alignement apporte une information pertinente sur l'écart entre les deux séries observées. Nous avons montré en particulier que si une même cellule de pluie est à l'origine des deux séries temporelles observées par deux pluviomètres distants, la structure d'alignement associée à un comportement régulier qui peut être interprété comme un décalage temporel correspondant au temps de parcours par la cellule de pluie entre les stations pluviométriques. La structure d'alignement permet donc de déterminer si deux évènements observés par deux pluviomètres distants correspondent ou non au même processus et s'il s'agit de la même cellule, la vitesse d'advection peut alors être estimée.

Dans le cadre de cette thèse, l'IMs-DTW est utilisée au chapitre 6 pour analyser les séries observées par un réseau pluviométrique en île de France : la structure d'alignement permet l'identification des cellules de pluie traversant le réseau, l'analyse des dissimilarités associées fournit des informations sur la distorsion spatio-temporelle de la cellule de pluie lorsqu'elle se déplace (cf. 6.4).

Au-delà de ce travail de thèse, cette mesure peut être utilisée pour diverses applications qui reposent sur les séries temporelles de précipitations. Elle peut par exemple participer à l'amélioration de l'étalonnage des radars météorologiques à partir un réseau de pluviomètres en

contribuant à faciliter l'appariement entre série pluviométriques et observation radar. Des recherches peuvent également être menées pour améliorer les méthodes de krigeage actuellement utilisées pour réaliser l'interpolation spatiale de séries pluviométrique et qui repose actuellement sur la modélisation des variogrammes. Pour faciliter l'accès de la communauté scientifique concernée à cet algorithme les codes seront disponibles en ligne dès que l'article ci-dessus sera publié.

Chapitre 5 : analyse et classification des séries temporelles de précipitations à l'aide de l'IMs-DTW

5.1. Introduction

Au chapitre 3, nous avons mis en évidence un effet important de l'agrégation temporelle et de la discrétisation de l'auget sur les cinq variables qui ont permis au chapitre 2 de discriminer les différents types d'évènements. Pour pallier à la difficulté de déterminer des caractéristiques à la fois stables par rapport aux changements des paramètres (T, v) et discriminantes par rapport aux processus physiques sous-jacents aux évènements de pluie, nous avons cherché à nous affranchir de la description par caractéristique des séries chronologiques en proposant au chapitre 4 une mesure de similarité entre séries.

L'adaptation de l'IMs-DTW a été motivée par la trop grande sensibilité des méthodes de classification basées sur l'extraction de caractéristiques, au type d'instrument et au temps d'agrégation. Cette forte sensibilité impacte la stabilité des classifications et l'information tirée des mesures réalisées. Dans ce chapitre nous discutons de la sensibilité de l'IMs-DTW aux changements d'instruments et aux changements des temps d'agrégation.

Dans la majorité des travaux utilisant la comparaison par déformation temporelle, seule la dissimilarité d_{DTW} est considérée. On rappelle qu'en comparant deux séries temporelles de précipitations, l'IMs-DTW fournit un nombre réel quantifiant la dissimilarité noté d_{DTW} mais également un alignement (path) sous contraintes multi-échelles ou « localement optimal », noté P_{DTW} . Nous avons vu au le chapitre précédent que deux couples de séries temporelles ayant la même valeur de d_{DTW} peuvent avoir été alignées de façon complètement différente, l'analyse du path P_{DTW} peut permettre de distinguer les deux situations suivantes : le déplacement du même événement observé par deux stations ou bien deux évènements distincts.

Nous définirons dans un premier temps la matrice de dissimilarités d'un nombre quelconque N de séries temporelles. Nous analyserons la stabilité de la dissimilarité d_{DTW} et de l'alignement P_{DTW} au type d'instrument et au temps d'agrégation. Nous présenterons ensuite une méthode de classification basée sur la matrice de dissimilarités.

Matrice de dissimilarité

Les mesures de dissimilarités de N séries temporelles de précipitations comparées deux à deux sont stockées dans une matrice carrée symétrique définie positive appelée matrice de dissimilarités et notée D :

$$D = \left[d_{ij}\right]_{i,j=1\ldots N}$$

Avec d_{ij} =d_{DTW}(S_i,S_j) mesure de dissimilarité entre la série Si et la série Sj. Puisque la dissimilarité d'une série avec elle-même est nulle cette matrice possède des zéros sur la diagonale. De même, la mesure de dissimilarité étant symétrique, la matrice est symétrique.

Données utilisées

L'IMs-DTW se veut une approche alternative et/ou complémentaire à l'approche basée sur l'extraction des caractéristiques. Tous les tests menés dans cette partie utilisent l'ensemble des 234 événements présentés et utilisés dans les chapitres deux et trois.

La comparaison des 234 événements issus du seul DBS agrégés à $T = 1min$ produit $\frac{234\times233}{2} = 27261$ alignements et la matrice des dissimilarités D est donc de dimension 234×234. Nous noterons cette matrice $D_{234\times234}$, elle est représentée graphiquement sur la partie de droite de la figure 1.

Sachant qu'il s'agit d'événements identifiés par un même instrument mais à des périodes de temps différents, les alignements trouvés ne traduisent pas une dynamique (on ne compare pas le même événement) mais une ressemblance de type motifs/forme. Toutefois, nous avons procédé à une inspection visuelle sur un échantillon de 200 alignements choisis aléatoirement : les alignements assemblent bien les pics avec les pics, les périodes non pluvieuses avec les périodes non pluvieuses conservant la cohérence recherchée. La figure 5.1 de gauche représente l'alignement trouvé entre le 62ème événement qui a eu lieu le 4 juillet 2012 entre 04:38 et 04:58 et le 92ème événement qui a lieu le 11 octobre 2012 entre 08:00 et 08:40. On voit que les sept premières minutes de l'événement 62 marquées par des intensités faibles sont associées aux dix premières minutes du 92ème événement de même nature, les deux averses sont associées elles aussi et les minutes suivantes caractérisées par des intensités faibles et annonçant la fin des événements sont également bien associées.

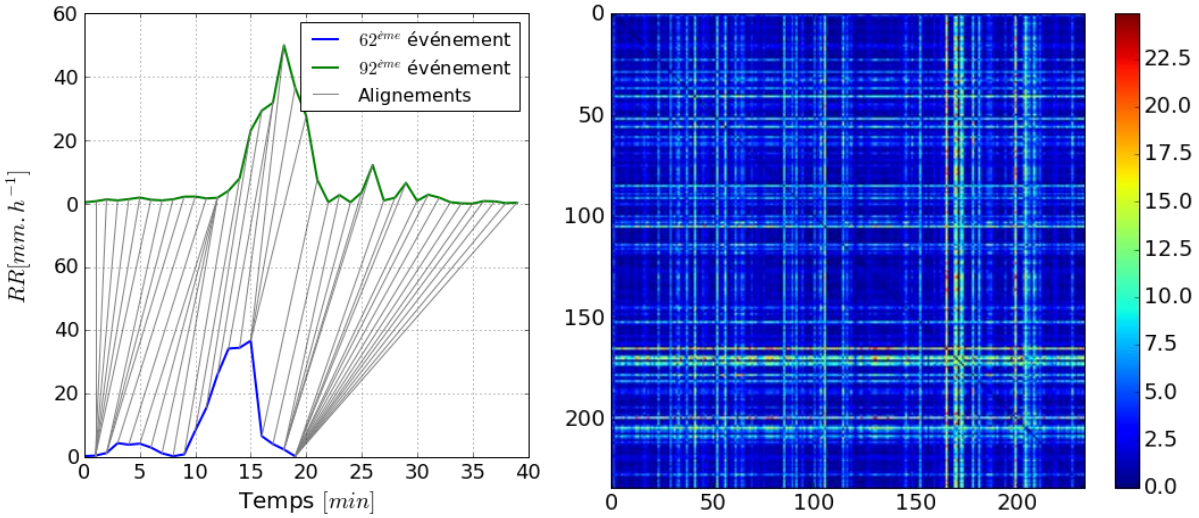


Figure 5.1. A gauche alignement de deux séries DBS à T=1min entre le $62^{ème}$ événement (en bleu) et le $92^{ème}$ événement (en vert), les alignements sont représentés en gris, à droite la représentation graphique de la matrice de dissimilarités $D_{234\times234}$

Sur la figure 5.1 de droite, qui représente la matrice de dissimilarité D associé aux 234 événements, l'ensemble des événements de précipitations n'ayant pas été réarrangé, le numéro de l'événement représente l'ordre d'arrivée de ce dernier (axe du temps). On remarque des groupes d'événements étalés sur 10 à 20 jours pouvant plus ou moins correspondre à l'échelle synoptique. On peut voir par exemple, que le groupe d'événement de violents orages ayant traversé l'île de France entre le 12 juin et le 19 juin 2013 (représentés par la série d'événements du 166ème au 177ème) manifestent de grandes dissimilarités avec le reste des événements.

5.2. Stabilité par rapport au type d'instrument et au temps d'agrégation

L'impact du changement d'instrument et de l'agrégation temporelle sur les performances de l'IMs-DTW doit donc être discuté sur les deux « sorties » de la méthode : d_{DTW} et P_{DTW}. Concernant la discrétisation par volume d'auget, si deux séries mesurées par le DBS se ressemblent, leurs discrétisations à volume d'auget v sont sensés se ressembler aussi (accumulation d'une quantité d'eau constante), mais les basculements peuvent être plus ou moins décalés dans le temps. Par conception, l'IMs-DTW repère ces décalages temporels et aligne par conséquent les basculements générés, l'alignement retrouvé entre les deux séries après discrétisation (pseudo-pluviomètre) est donc sensé résumer l'alignement avant la discrétisation. On peut s'attendre à ce qu'une ressemblance entre deux séries soit conservée lors

d'une discrétisation par volume d'auget v sous contrainte d'utilisation de valeurs raisonnables de v.

L'agrégation temporelle étudiée au chapitre 3, présente un effet « destructeur » de l'information sur la variabilité. Malgré les travaux de Keogh et Pazzani (2000), mentionnés au chapitre précédent, concernant la conservation des mesures de dissimilarités, la conception de l'IMs-DTW garantit seulement la comparabilité des alignements mais pas forcément la conservation des mesures des dissimilarités. Cependant si le temps d'agrégation est « raisonnable » on peut s'attendre à une conservation relative de la mesure de dissimilarité, dans le cas contraire une perte d'information importante peut être observée.

Les valeurs des mesures de dissimilarités peuvent changer après différentes transformations (discrétisation et/ou agrégation) sans pour autant forcément modifier l'information expliquée par cette mesure. Le test d'égalité n'est pas toujours une stratégie judicieuse. Il se peut que l'égalité ne soit pas vérifiée alors qu'on a une conservation de l'information expliquée par cette mesure « pouvoir discriminatif ». Dans ce qui suit, en absence d'une relation affine ou linéaire entre les mesures, on utilisera le coefficient de corrélation de rang de Spearman ρ comme mesure de conservation de l'information expliquée par la mesure de dissimilaité IMs-DTW.

5.2.1. Stabilité de l'IMs-DTW au volume d'auget au pas de temps d'1 min

Avec un raisonnement similaire à celui du Chapitre 3, nous avons simulé des séries de pseudo- pluviomètres avec des volumes d'auget $v = 0.1\ mm$ et $= 0.2\ mm$. L'application l'IMs-DTW sur les 234 événements transformés donne comme précédemment $\frac{234\times233}{2} = 27261$ alignements et des matrices de dissimilarités notée $D_{0.1}$et $D_{0.2}$.

La figure 5.2-gauche (resp. droite) présente l'alignement trouvé entre le 62ème et 92ème événements discrétisés à $v = 0.1mm$ (resp. $v = 0.2mm$). Dans le cas de la discrétisation à $v = 0.1mm$ on voit que les trois premiers basculements correspondant à chaque début des deux événements sont bien liés conservant la liaison observée sur la figure 5.1 (sans discrétisation), les deux averses conservent leurs alignements et leurs fins sont bien associées malgré l'absence de basculements sur l'événement 62. La même remarque est faite pour la discrétisation à $v = 0.2\ mm$.

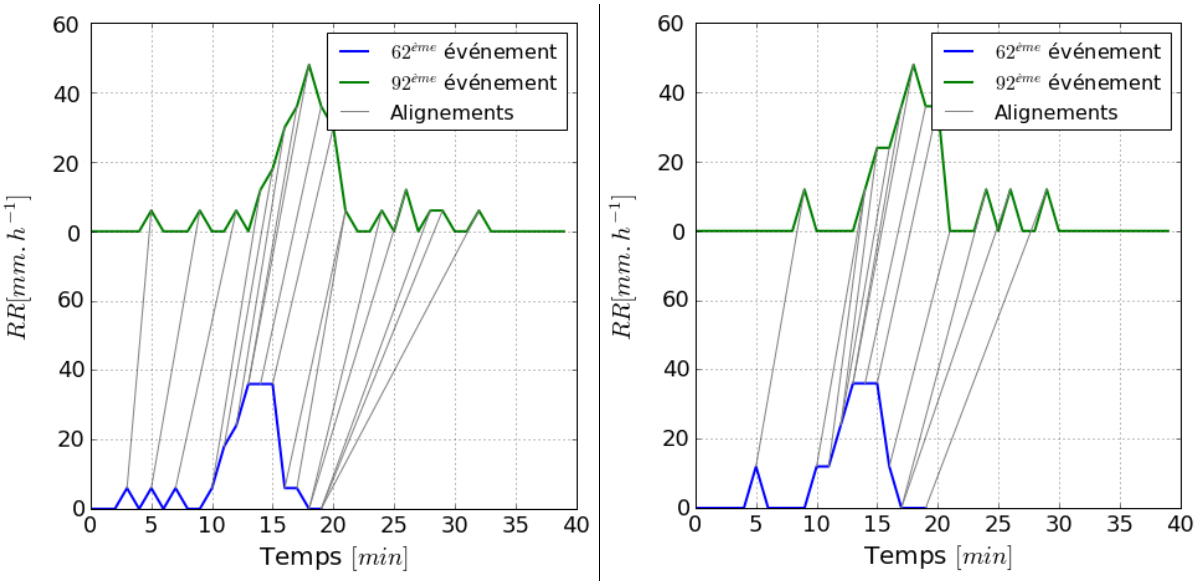


Figure 5.2. Les alignements trouvés entre le 62ème événement en bleu et le 92ème événement en vert, les alignements sont représentés en gris (à gauche les séries au format pseudo-pluviomètre v=0.1mm et à droite les séries au format pseudo-pluviomètre v=0.2mm)

Discussion concernant la conservation des formes des alignements : bien que d'autres approches puissent être utilisées pour l'analyse des appariements, nous avons choisi une comparaison par descriptions. Petitjean (2011) propose une liste d' *"indices de descriptions permettant de retranscrire de façon condensée l'information complexe dans l'alignement"*. Nous avons retenu parmi ceux proposés, et comme au chapitre précédent, la moyenne et l'écart-type des différences temporelles pour discuter de la qualité des alignements. Toutefois, nous avons aussi procédé à une comparaison visuelle qualitative d'un échantillon aléatoire de 100 alignements pour chaque discrétisation. La figure 5.3 présente sous forme d'histogrammes 2D les moyennes τ_{ATD} et les écart-types σ_{ATD} des différences temporelles avant discrétisation et post-discrétisations ($v = 0.1mm$ et $v = 0.2\ mm$).

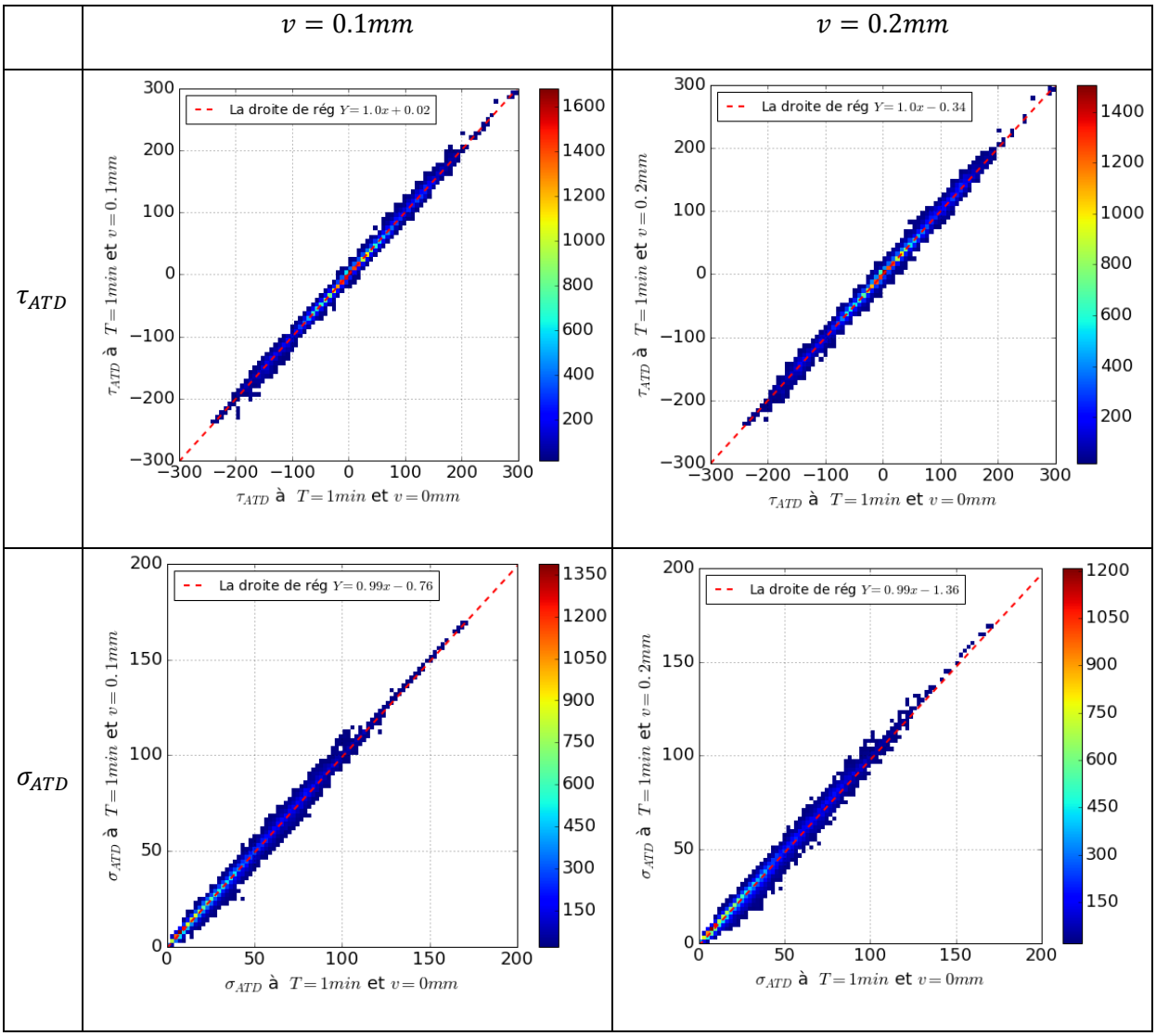


Figure 5.3.histogrammes 2D des moyennes (resp. écart-types) des différences temporelles entre les 234 événements deux à deux représentant les grandeurs post-discrétisation en fonction des grandeurs initiales - en rouge les droites des régressions linéaires estimées (seuil d'histogramme : 20 points/pixel)

A première vue, l'effet de la discrétisation à $v = 0.1mm$ ressemble à celui à $v = 0.2mm$. La discrétisation semble sans effet significatif sur les descripteurs ; on note la forte densité des points sur la première bissectrice pour les quatre histogrammes justifiant des ajustements de type régression linéaire. Les coefficients de corrélation de Pearson et de Spearman entre les moyennes des différences temporelles τ_{ATD} lors de l'utilisation du DBS et celles obtenues après discrétisations sont égaux à 0.99 pour les deux discrétisations $v = 0.1mm$ et $v = 0.2mm$. De plus, les deux coefficients directeurs des droites de régression entre les moyennes des différences temporelles sont égaux à 1 avec des coefficients de détermination égaux à 0.99 soutenant l'hypothèse d'égalité des moyennes. Une conclusion similaire sur l'égalité des écart-types des différences moyennes peut être faite à partir des deux figures 5.3-c et 5.3-d.

On peut noter également que les moyennes des différences temporelles peuvent dans certains cas prendre des valeurs absolues fortes, elles décrivent en moyenne les distorsions nécessaires pour faire "matcher" deux événements. Deux situations se présentent : la première est la comparaison d'un événement court avec un événement long, l'alignement dans ce cas compare implicitement les durées des événements, on cite comme exemple le cas de la comparaison du 34ème événement "le plus court avec une durée de 7 minutes" avec le 121ème évènement "le plus long avec une durée de 1410 minutes" qui implique forcément une distorsion importante. La deuxième situation "intéressante" est celle où cette grande distorsion est obtenue pour deux évènements de durée similaire et causée par des positions différentes des pics d'intensités maximales sur chaque événement (l'alignement favorise l'association des pics d'intensités). Cette situation est illustrée sur la figure 5.4 présentant l'alignement trouvé à partir des séries du DBS entre le 42ème et le 53ème évènement et soutient un résultat du chapitre précédent concernant l'importance de l'alignement P_{DTW} de deux séries temporelles. En effet, l'information sur la différence de positions des pics d'intensités est très intéressante et aide à différencier deux situations hydrologiquement différentes a affirmé D. Dunkerley dans son rapport de révision (Dunkerley, 2016) de l'article présenté au chapitre 2 :

'' *far more of the rain tends to become overland flow if the largest intensity peak occurs late in the rainfall event, and much less becomes overland flow if the intensity peak is early*''

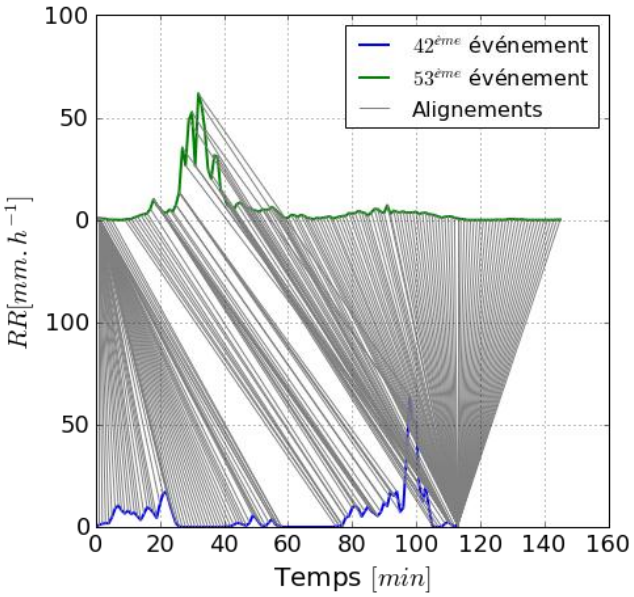


Figure 5.4. L'*alignement trouvé entre le 42ème événement en bleu et le 53ème événement en vert, les alignements sont représentés en gris, les séries DBS à T=1min*

Discussion concernant la conservation des dissimilarités : La comparabilité des mesures de dissimilarité est contrainte par la ressemblance des alignements. Cette dernière étant vérifiée, la

comparaison des matrices de dissimilarités reflèterait l'effet de la discrétisation de l'auget. La figure 5.5 présente les matrices de dissimilarités $D_{0.1}$ et $D_{0.2}$ associées aux discrétisations de volumes d'augets $v = 0.1mm$ et $v = 0.2\ mm$ respectivement et *T=1 min*.

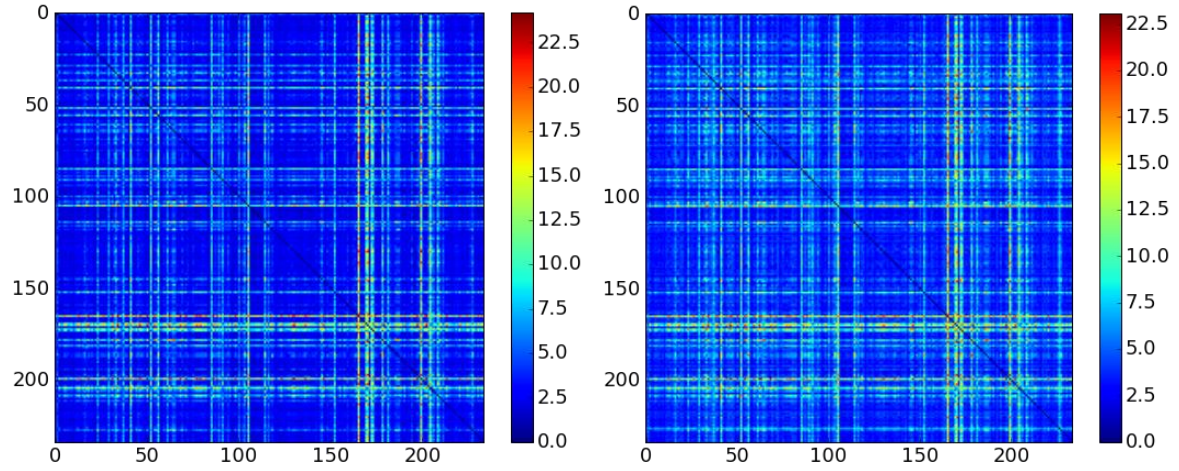


Figure 5.5. Matrices de dissimilarités des 234 événements transformés : à gauche les événements sont au format pseudo-pluviomètre à v=0.1mm et à droite l les événements sont au format pseudo-pluviomètre à v=0.2mm, T=1min

En comparant la matrice D de la figure 5.1 relative au DBS et $D_{0.1}$ de la figure 5.5 puis les matrices $D_{0.1}$ et $D_{0.2}$ on remarque que les dissimilarités à valeurs faibles et modérées s'accentuent de plus en plus au fur et à mesure que le volume d'auget augmente et qu'au contraire apparait une atténuation des dissimilarités à valeurs fortes. Cependant, pour ces volumes d'auget (0.1mm et 0.2mm) les dissimilarités semblent rester discriminantes. L'ordre des ressemblances/dissemblances entre les événements est relativement conservé ; les groupes d'événements successifs qui se ressemblent sur la matrice D se ressemblent aussi sur les deux autres matrices $D_{0.1}$ et $D_{0.2}$ et les patterns d'événements à l'échelle synoptique restent perceptibles. De même, les violents orages de juin 2013 conservent leurs démarcations après les deux discrétisations. Le tableau 5.1 résume les statistiques standards des distributions des mesures et confirme ces observations.

Paramètre	DBS (v=0mm)	Pseudo-P (v=0.1mm)	Pseudo-P (v=0.2mm)
Mode	1.08	2.17	3.22
Moyenne	3.48	4.30	5.00
Ecart-type	3.50	3.07	2.78
Médiane	2.11	3.14	4.17
1er quartile	1.08	2.41	3.26
3ème quartile	4.68	5.01	5.76
Ecart-interquartiles	3.6	2.6	2.5
Valeur minimale	0.11	0.73	0.79
Valeur maximale	24.89	24.18	23.06
Etendu	24.78	23.45	22.27

Tableau 5.1. Statistiques des mesures des dissimilarités pour les trois configurations (v=0mm, v=0.1 mm et v=0.2mm) à T=1min

La valeur minimale des mesures de dissimilarité observée par le DBS augmente après la discrétisation par les augets alors que la valeur maximale diminue. Les séries temporelles de précipitation mesurées par un pluviomètre (de volume d'auget v) enregistrent des taux précipitants multiples de v et la somme des différences des taux précipitants (notée S_d) est donc un multiple de v. Par conséquent, une mesure de dissimilarité non nulle entre deux séries vaudrait au moins la valeur $pas = \frac{2v}{len(s_1)+len(s_2)}$ au lieu d'une valeur proche de 0 et prend pour valeur maximale la mesure arrondie à un multiple de pas. Ce constat pourrait expliquer l'augmentation (resp. diminution) des valeurs minimales (resp. maximales) après discrétisation par l'auget.

En moyenne les dissimilarités s'accentuent après discrétisation, on note la présence de biais systématiques 0.8 (resp. 1.5). D'autre part, la comparaison d'un événement i observé par le DBS avec ce même événement observé par un pseudo-pluviomètre fournit une mesure de dissimilarité qui pourrait décrire l'effet de la discrétisation de l'auget sur ce même événement i. La moyenne de ces mesures sur l'ensemble des 234 événements vaut 1.7 (resp. 2.9) pour une comparaison DBS pseudo-pluviomètre à $v = 0.1mm$ (resp. $v = 0.2\ mm$).

De même, la comparaison des valeurs des 3èmes quartiles montre que le biais (l'écart) diminue mais reste positif. Ceci est en adéquation avec les remarques générales d'accentuation

observée des valeurs des dissimilarités et est vérifiée pour d'autres volumes d'augets simulés (0.3, 0.4 et 0.5 non présentées ici).

L'analyse des étendus, des écarts-type puis des écart-interquartiles, fait apparaitre un effet de compression des valeurs des dissimilarités : l'intervalle des valeurs est de plus en plus restreint lorsque le volume d'auget augmente.

Pour visualiser la relation entre les mesures et les déformations causées par la discrétisations, la figure 5.6 présente sous forme d'histogramme 2D la distribution des $\frac{234\times234}{2} = 27378$ points représentant les dissimilarités entre les couples d'événements obtenus à partir du DBS (en abscisse) et par le pseudo-pluviomètres (en ordonnée) pour les deux volumes d'auget ($v = 0.1mm$ et $v = 0.2mm$) et pour un temps d'agrégation $T = 1\ min$.

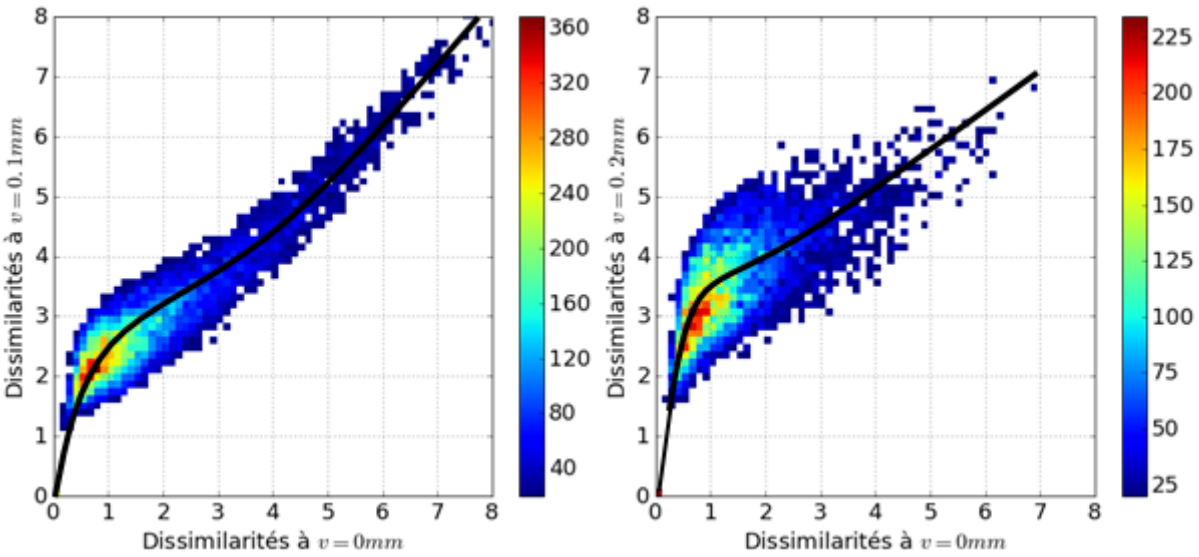


Figure 5.6.Histogrammes 2D des dissimilarités des événements après discrétisation (axe vertical) en fonction des dissimilarités des mêmes événements avant discrétisations à T=1min (axe horizontal) : à gauche pseudo-pluviomètre v=0.1mm en fonction du DBS v=0mm, à droite pseudo-pluviomètre v=0.2mm en fonction du DBS v=0mm (seuil d'histogramme : 20 points/pixel)

En plus de confirmer les constations faites à partir des statistiques unidimensionnelles (présence d'un biais, compression,…), ces deux histogrammes explicitent la relation entre les mesures initiales (DBS) et post-discrétisations. Ils présentent la même forme à une différence de paramètres près, les commentaires faits sur le premier peuvent être transposés sur le deuxième.

Concernant la figure 5.6-gauche qui traite de l'effet de la discrétisation à $v = 0.1mm$, la distribution des points, bien que faisant apparaitre une certaine dispersion, dessine une courbe. Pour les faibles valeurs de dissimilarité, la pente de celle-ci est constante et est relativement

importante, ensuite elle décroit et tend vers 1 lorsque la courbe se rapproche de la bissectrice. Cette courbe fait apparaitre une relation bijective et montre donc que deux couples d'événement observés par le DBS et ayant des similarités différentes ont également des similarités différentes lorsqu'ils sont observés par le pseudo-pluviomètre. Cette constatation est corroborée par les coefficients de corrélation de rang de Spearman qui valent 0.93 pour $v = 0.1mm$ et 0.85 pour $v = 0.2mm$.

Concernant les cas les plus fréquents (les points situés sur la courbe en noir) nous avons constaté que les couples d'événements possèdent un nombre moyen de basculement d'augets et un temps inter-basculement sensiblement identique. Par contre, si on considère les points situés au voisinage de la courbe, il apparait que les évènements peuvent réagir de façon différente à la discrétisation de l'auget. Nous avons constaté les deux situations typiques suivantes :

1- **Points au-dessous de la courbe** : considérons par exemple la dissimilarité obtenue par un DBS entre une bruine (support long et continu avec une intensité très faible <1mm/h) et un événement convectif (intensité forte). La dissimilarité obtenue sera relativement forte car la mesure de la dissimilarité permet de détecter toute cette différence. Considérons maintenant la même situation mais observée par un pluviomètre à auget. L'évènement convectif sera peu affecté par la discrétisation, par contre, l'évènement de bruine sensé durer longtemps avec une faible intensité sera observé par le pluviomètre avec seulement un seul basculement (ou tout au moins un nombre très réduit de basculements) mais d'intensité relativement forte (on rappelle qu'un basculement à v=0,2 mm et T=1 min correspond à une intensité de pluie de 12 mm/h !). Ainsi l'effet de la discrétisation sur un événement de bruine sera de réduire les durées de pluie de l'événement et d'augmenter leur intensité de sorte que l'évènement de bruine tend à ressembler un peu à un événement convectif. Il en résulte qu'au lieu de comparer un évènement stratiforme avec un événement convectif, on compare un évènement « un peu convectif » avec un évènement convectif, d'où la diminution de la dissimilarité.

2- **Points au-dessus de la courbe** : ce cas est fréquent lorsqu'on considère des événements stratiformes. Par exemple soit deux événements qui enregistrent des intensités constantes mais proches (5.9 mm/h et $6mm/h$) sur toute la durée de l'événement. L'analyse par le DBS détecterait une faible dissimilarité entre ces deux événements. Le passage au format pseudo-pluviomètre à $v = 0.1mm$ génèrerait une suite continue sans 0 dans le cas de la série à $6mm/h$ alors qu'il introduirait périodiquement des périodes de non-pluie dans la série à $5.9mm/h$. Cette introduction de valeurs nulles a pour effet de rendre les deux séries encore plus différentes que ce qu'elles étaient, et par conséquent une mesure de dissimilarité plus forte.

5.2.2. Stabilité de l'IMs-DTW au volume d'auget à la résolution 6 min

La figure 5.7 reprend l'étude précédente (figure 5.6) mais pour un temps d'agrégation $T = 6\,min$. Les deux distributions manifestent moins de dispersions comparées aux nuages à $T = 1min$. Encore une fois les effets dus à la discrétisation se ressemblent pour $v = 0.1mm$ et $v = 0.2mm$ sont du même type. La distribution à $v = 0.1mm$ (figure de gauche) fait apparaitre une relation forte, toujours positive, avec une forme affine cette fois, le biais systématique diminue et est proche de 0. Un ajustement linéaire donne l'équation $d_{0.1}^{6} = 0.98\, d_{0.1}^{1} + 0.13$ avec un coefficient de détermination $r_2 = 0.97$ qui suggère une conservation des valeurs des dissimilarités entre les séries DBS et pseudo-pluviomètre. De plus, les coefficients de corrélation de Spearman grimpent à 0.98 (resp. 0.94) pour v=0.1 mm (resp. v=0.2mm) au lieu de 0.93 (resp. 0.84) et les coefficients de corrélation linéaire de Pearson sont à 0.99 (resp. 0.98) pour v=0.1mm (resp. v=0.2mm) et suggère une conservation importante du pouvoir discriminant de la mesure de dissimilarités des séries discrétisées.

Ces constatations corroborent l'hypothèse de relation de « compensation » d'effets entre le volume d'auget et le temps d'agrégation vue au troisième chapitre où nous avions montré que les effets de l'auget sont masqués à condition de prendre un temps d'agrégation suffisant.

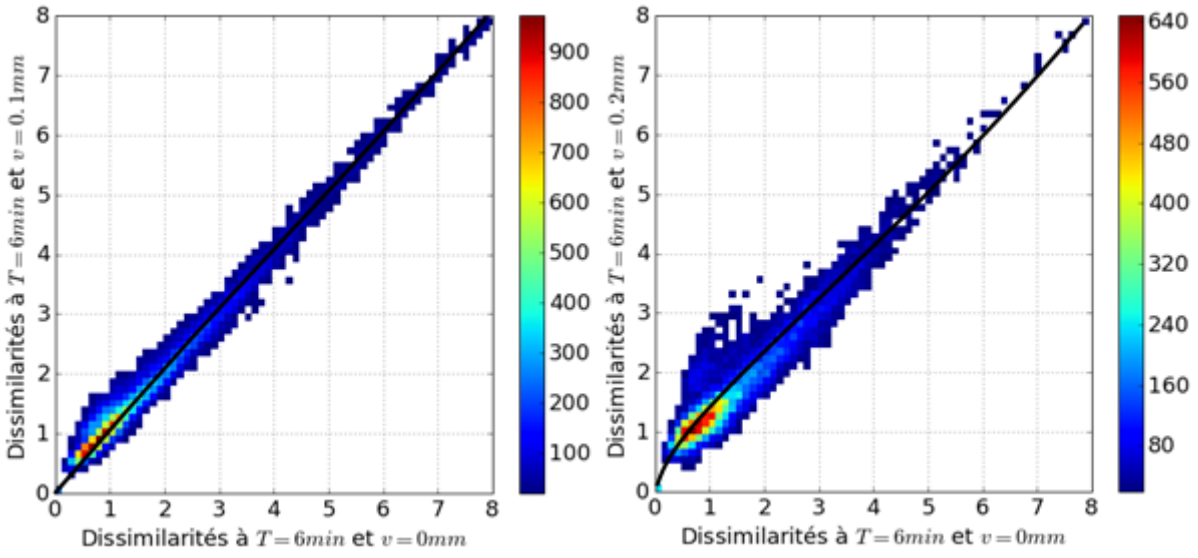


Figure 5.7. Histogrammes 2D des dissimilarités des événements après discrétisation en fonction des dissimilarités des événements post-discrétisations : à gauche données pseudo-pluviomètre v=0.1mm agrégées à T=6min en fonction des données DBS v=0mm agrégées à T=6min à droite données pseudo-pluviomètre v=0.2mm agrégées à T=6min en fonction des données DBS v=0mm agrégées à T=6min (seuil d'histogramme : 20 points/pixel)

L'augmentation du temps d'agrégation d'un pluviomètre à auget ($T = 6min$ dans notre étude) fait tendre la dissimilarité vers celle qui serait obtenue à l'aide d'un disdromètre. On peut donc

par conséquent supposer que l'utilisation d'un pluviomètre est relativement équivalente à l'utilisation d'un disdromètre à la résolution $T = 6min$ pour la classification des séries chronologiques de pluie à partir de l'IMs-DTW.

La figure 5.8 présente les matrices des dissimilarités $(D^6, D^6_{0.1}, D^6_{0.2})$ à $T = 6min$ des 234 événements pour les mesures DBS et les deux transformations pseudo-pluviomètres.

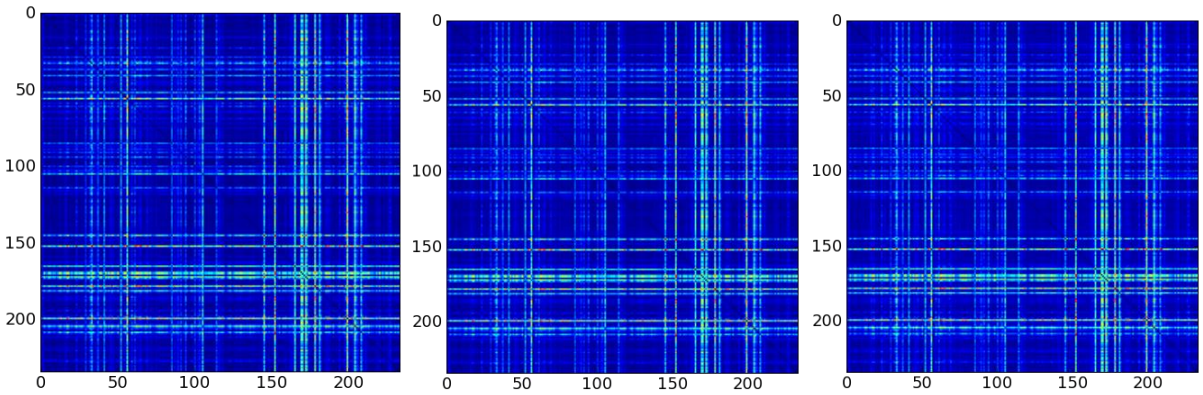


Figure 5.8. Matrices de dissimilarités $D^6, D^6_{0.1}$ et $D^6_{0.2}$ comparant les 234 événements transformés à T=6min pour les différents volumes d'augets: de gauche à droite les événements sont au format DBS (v=0mm), pseudo-pluviomètre à v=0.1mm et pseudo-pluviomètre à v=0.2mm

Comme on peut s'y attendre, les groupes d'événements successifs qui se ressemblent sur la matrice D (figure 1.A) se ressemblent sur les matrices $D^6, D^6_{0.1}$ et $D^6_{0.2}$et les patterns d'événement à l'échelle synoptique restent perceptibles. De même, les violents orages de juin 2013 conservent leurs démarcations.

La comparaison deux à deux des matrices de dissimilarité $D^6, D^6_{0.1}$ et $D^6_{0.2}$ est complémentaire à la comparaison de la figure 5.7 et la ressemblance entre les trois matrices rejoint la relation affine (identitaire) observée entre les mesures de dissimilarités dans la partie précédente. Le tableau 5.2 résume les statistiques standards des distributions des trois mesures et confirme l'équivalence des types instruments à la résolution de 6 min.

Paramètre	DBS (v=0mm)	Pseudo-P (v=0.1mm)	Pseudo-P (v=0.2mm)
Mode	0.76	0.78	0.79
Moyenne	3.59	3.67	3.78
Ecart-type	4.28	4.24	4.21
Médiane	1.86	1.97	2.25
1er quartile	0.97	1.14	1.29

3ème quartile	4.20	4.25	4.40
Ecart-interquartiles	3.23	3.11	3.11
Valeur minimale	0.11	0.26	0.32
Valeur maximale	32.75	32.90	32.92
Etendu	32.64	32.64	32.6

Tableau 5.2. Statistiques des mesures des dissimilarités pour les trois configurations (v=0mm, v=0.1 mm et v=0.2mm) à T=6min

En effet, à cette résolution ($T = 6min$) les statistiques des mesures sont presque égales pour les trois configurations (v=0mm, v=0.1mm et v=0.2mm).

5.2.3. Stabilité de l'IMs-DTW au temps d'agrégation

Dans la partie précédente, nous nous sommes principalement focalisés sur l'influence du volume des augets (et donc du type d'instruments) pour deux résolutions temporelles. Dans cette partie, nous allons nous intéresser, pour chacun des instruments séparément, à la sensibilité de la méthode au temps d'agrégation T. Comme nous l'avons vu au chapitre 4, la recherche de l'alignement optimal par la méthode IMs-DTW entre deux séries se fait par descente d'échelle, l'alignement à une échelle fine est un descendant de l'alignement à l'échelle grossière. Cette propriété assure la comparabilité des mesures de dissimilarités, leur comparaison reflète donc l'effet de l'agrégation temporelle.

L'impact de l'agrégation temporelle sur la mesure des dissimilarités est visible en comparant les valeurs et les statistiques des dissimilarités à $T = 1min$ (figure 5.1.A et 5.5 et tableau 5.1) avec celles obtenues à $T = 6min$ (figure 5.8 et tableau 5.2) pour chaque instrument. Globalement, pour le DBS les statistiques obtenues à $T = 1min$ et à $T = 6min$ sont assez proches à l'exception de la valeur maximale. Pour les deux séries discrétisées, l'utilisation d'un temps d'agrégation de 6 minutes tend à faire converger leurs statistiques vers celles du DBS. Cela est conforme à la remarque concernant l'effet de compensation de la discrétisation par l'agrégation temporelle.

La figure 5.9 compare les distributions des dissimilarités obtenues pour deux temps d'agrégation différents ($T = 1min$ en abscisse et $T = 6min$ en ordonnée) et pour chacun des 3 types d'instrument (DBS, pseudo-pluviomètre 1 et pseudo-pluviomètre 2).

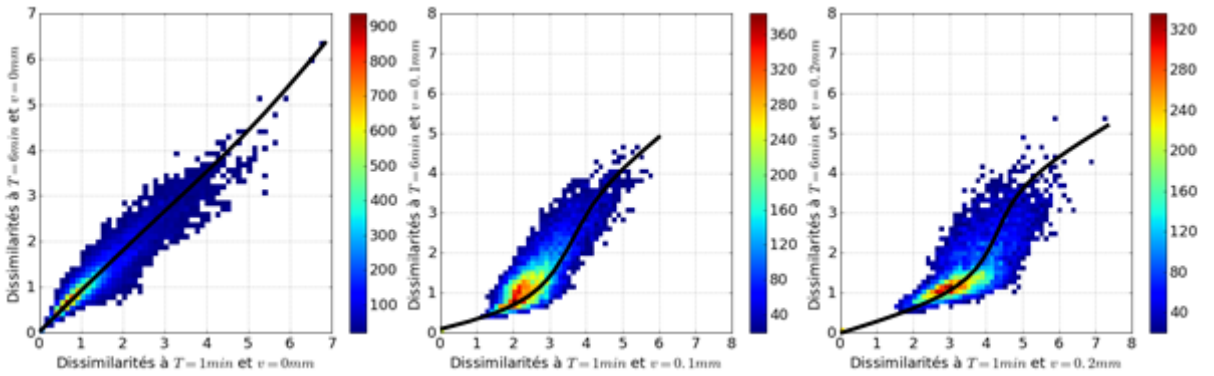


Figure 5.9. Histogrammes 2D des Dissimilarités des événements après agrégation temporelle à T= 6min en fonction des dissimilarités des événements à T=1min pour les trois volumes d'auget: de gauche à droite le DBS (v=0mm) , le pseudo-pluviomètre v=0.1mm et le pseudo-pluviomètre v=0.2mm (seuil d'histogramme : 20 points/pixel)

La distribution des points sur la figure de gauche (représentant l'effet de l'agrégation temporelle sur les mesures du format DBS) manifeste une légère dispersion mais fait apparaitre une relation affine positive entre les mesures initiales (DBS) à $T = 1min$ et celles après agrégation temporelle à $T = 6\ min$ pour le même instrument (DBS). L'ajustement affine entre les deux grandeurs donne l'équation $d_0^6 = 1.08\, d_0^1 - 0.19$ avec un coefficient de détermination r_2 =0.74 qui suggère une conservation des valeurs des dissimilarités entre les mesures des dissimilarités à $T = 1min$ et les mesures après agrégation temporelle à $T = 6min$. De plus, le coefficient de corrélation de Spearman entre les deux grandeurs vaut $r_s^{DBS} = 0.94$ et corrobore la conservation du "pouvoir discriminatif ". Concernant, les deux distributions (figure 5.9-centre et figure 5.9-droite) qui présentent les effets de l'agrégation temporelle sur des données pseudo-pluviomètres, on note la symétrie des courbes par rapport à la première bissectrice comparée aux courbes obtenues dans le cas de l'étude de l'effet de la discrétisation due aux augets (figure 5.6). Cette observation en adéquation avec les observations précédentes confirme l'hypothèse que l'effet de l'agrégation temporelle provoque un effet opposé à celui de la discrétisation par les augets. Par analogie, on peut dire que l'information expliquée par les mesures des dissimilarités est relativement conservée. Ces constations de conservation relative de l'information expliquée sont corroborées par les coefficients de corrélation de rang de Spearman entre les mesures à $T = 1min$ et $T = 6min$ qui valent 0.90 pour la discrétisation à $v = 0.1mm$ et 0.85 pour $v = 0.2mm$.

Une conséquence de ces deux remarques est que les mesures de dissimilarité "faibles" à la résolution fine ($T = 1min$) diminuent en agrégeant les séries mais que les dissimilarités "fortes" augmentent après agrégation – une explication de cet effet a été détaillée dans la partie 3.4 du chapitre 3. Inverse à celui de la discrétisation par volume d'auget, la composition des deux opérations (combinaison des deux effets) donnerait en principe des mesures des dissimilarités

stables pour des valeurs de v et T bien choisies. La partie suivante se propose d'étudier cette combinaison des opérations.

5.2.4. Combinaison des effets de discrétisation et d'agrégation temporelle

Actuellement, la majeure partie des observations in situ des précipitations est réalisée à l'aide de pluviomètres à augets. Ainsi, les données issues des pluviomètres de Météo France sont prétraitées puis partagées après agrégation à $T = 5{\sim}6min$. Par conséquent, les effets agrégation temporelle et discrétisation dus aux augets sont en pratique toujours combinés. Il est donc nécessaire d'analyser plus en détail l'effet combiné de ces deux opérations sur la mesure de dissimilarité par IMs-DTW.

La figure 5.10 présente les histogrammes 2D des couples d'événements selon leurs dissimilarités à $T = 1min$ et $v = 0mm$ en abscisse et à $T = 6min$ en ordonnée ($v = 0.1mm$ figure de gauche et $v = 0.2mm$ figure de droite).

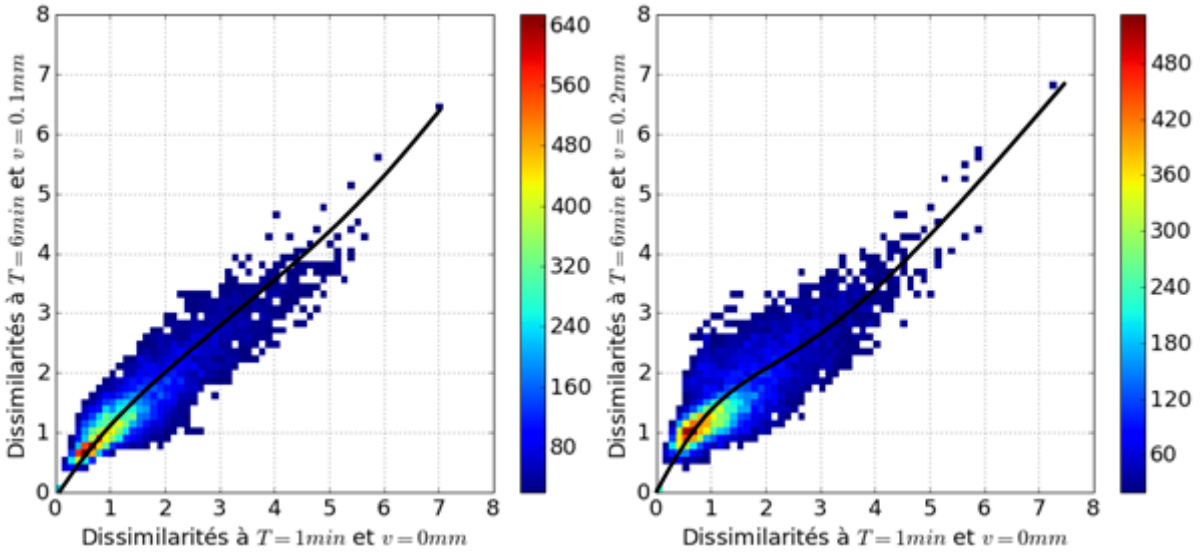


Figure 5.10. Dissimilarités des événements après discrétisation puis agrégation temporelle à T= 6min en fonction des dissimilarités des événements à T=1min pour les deux volumes d'auget: à gauche pseudo-pluviomètre v=0.1mm et à droite pseudo-pluviomètre v=0.2mm (seuil d'histogramme : 20 points/pixel)

A première vue, les deux distributions (figure 5.10) présentent les mêmes étendues de dispersion que sur les deux distributions de la figure 5.9. Toutefois, on note l'absence des deux concavités et du point d'inflexion. Malgré la dispersion, les nuages sont positionnés autour de la première bissectrice suggérant une relation linéaire entre les deux mesures.

On peut conclure que l'alignement est par construction totalement stable par rapport à l'agrégation temporelle et très stable par rapport à la discrétisation due aux augets. Cette discrétisation augmente en moyenne la valeur de la dissimilarité entre événements mais cet effet est compensé par l'agrégation temporelle pour un temps d'agrégation $= 6min$. Nous allons donc dans la suite de ce chapitre utiliser l'IMs-DTW pour la classification de séries temporelles de précipitation.

5.3. Exploitation de l'IMs-DTW pour l'analyse des séries temporelles de précipitations

Les méthodes de clustering ou classification non supervisée (recherche de groupes homogènes dans un jeu de données) regroupent les individus qui partagent des caractéristiques communes, ce qui se traduit par des critères de similarité. Les groupes constitués doivent être le plus distincts possibles, en ce sens que deux individus appartenant à deux groupes différents doivent être le plus dissimilaires possible. La constitution de groupe homogène permet d'identifier un nombre réduit de prototypes, représentatifs de chaque groupe, et qui peuvent être analysés en détail.

5.3.1. Quelle méthode de classification choisir ?

Plusieurs méthodes de classification, basées sur les dissimilarités, existent dans la littérature (Aghabozorgi et al., 2015), parmi lesquelles :

- Les méthodes de regroupement hiérarchique (Everitt, 1974) ;
- les méthodes connexionnistes telles que les cartes auto adaptatives (SOFM) (Rahmel, 1995)(Kohonen, 1982)
- les méthodes basées sur les centroïdes telles que les algorithmes des *k*-moyennes ou k-médoïdes (Kaufman & Rousseeuw, 1987)

Chaque méthode a ses propres propriétés, ses paramètres, ses points forts et ses points faibles et est adaptée à une problématique spécifique. **Le choix de l'algorithme de classification utilisé dépend donc de l'objectif visé. Ce dernier cerne le choix des paramètres de l'algorithme.** Par exemple, les méthodes de regroupement hiérarchique sont contraintes par la définition du bon critère d'agrégation des sous-classes. Selon le critère choisi, le résultat du regroupement peut changer de façon significative.

D'autre part, l'analyse par la classification bien qu'étant une approche intéressante, suppose l'existence d'une classification et/ou son unicité qui promet une convergence des algorithmes de classification. Dans le jeu de données des 234 événements l'hypothèse d'existence et unicité d'une classification en deux classes est soutenue par une connaissance à priori : la typologie des précipitations, i.e. l'existence de deux types de pluie: convectives et stratiformes (Molini et al. 2011) (cette vision binaire n'est d'ailleurs vraie que pour des séquences relativement courtes et homogènes). Dans le cadre de l'exploration des données sans connaissances à priori, cette contrainte d'existence et d'unicité n'est pas toujours vérifiée notamment dans le cas d'un jeu de données tel que l'on peut passer d'un point à l'autre de façon relativement continue (un continuum). Dans ce cas au lieu de parler de classification au sens groupement, on parle d'analyser une tendance ou un gradient, et la classification sert à réduire le nombre d'échantillons pour une perspective d'ordination. Ceci peut correspondre à des séries de précipitations relativement longues qui contiendront différentes séquences stationnaires et convectives en proportion variable. La figure 5.11-a illustre le cas où le jeu de données admet une classification optimale tandis que la figure 5.11-b illustre le cas où la notion de classes n'existe pas.

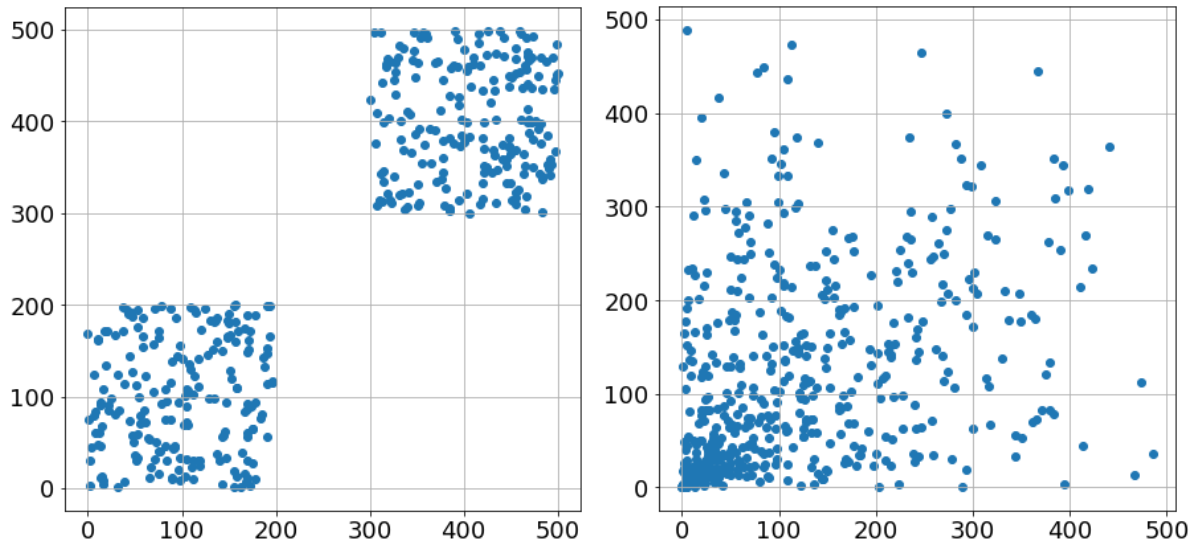


Figure 5.11 (a) l'ensemble des points accepte une classification en deux classes, (b) l'ensemble des points n'accepte pas une classification en k classes

Le jeu de données présenté à la figure 5.11-a est réparti en deux composantes facilement identifiables (une frontière est visible) et les algorithmes de classification connus peuvent converger vers cette classification en deux composantes. Le deuxième jeu de données présenté sur la figure 5.11-b est sous forme d'un continuum et la notion de classes n'existe pas. Cette nature continue du jeu de données donnerait une « infinité » de classifications différentes en k

classes en fonction des paramètres choisis (initialisations et critères). L'algorithme de classification sert à réduire le nombre d'éléments à analyser tout en conservant la densité de probabilité (échantillonnage) offrant ainsi une bonne description du jeu de données (quantification vectorielle).

Pour l'analyse des données à notre disposition (jeu de données SIRTA + MétéoFrance), le besoin de choisir un algorithme "compromis" qui sert pour la classification lorsque les conditions d'existence et/ou d'unicité sont vérifiées ou de méthode d'échantillonnage dans le cas d'une structure du type continuum, nous a conduit à éliminer les méthodes de regroupement hiérarchique. Parmi les méthodes existantes, nous avons choisi les « *k*-médoïdes » (le choix sera justifié dans le paragraphe suivant).

5.3.2. Algorithme des *k*-médoïdes (Kaufman & Rousseeuw, 1987)

Présentation des k-médoïdes basée sur les dissimilarités

En statistiques, un **médoïde** est le représentant le plus central d'une classe. La méthode dite **k-médoïdes** est une méthode de partitionnement qui repose sur la recherche de *k* médoïdes (représentants), un pour chaque classe. Dans sa version Partition Around Medoids (PAM) qui s'inspire beaucoup des k-moyennes, le médoïde de la classe est l'objet pour lequel la dissimilarité moyenne (ou de manière équivalente la similarité moyenne) par rapport à tous les objets de la classe est minimale (resp. maximale).

La méthode des k-médoïdes est une méthode de classification plus robuste vis-à-vis des données aberrantes (*outliers*) que celle des k-means (k-moyennes). De plus, traiter un médoïde (dans le cas des k-médoïdes) permet une analyse poussée des caractéristiques d'un élément représentatif alors que le traitement d'un centre dépend de la définition de l'opérateur moyenne et des caractéristiques conservées.

Algorithme des k-médoïdes :

Entrée : X (n ind.), d #mesure de dissimilarité, k # classes
Initialiser k médoïdes M_k REPETER **1-Affectation** : Affecter chaque individu à la classe dont le médoïde est le plus proche (au sens de la mesure de dissimilarité d). **2-Représentation** : Recalculer les médoïdes des classes à partir des individus rattachés. JUSQU'À Convergence
Sortie : Une partition des individus C caractérisée par les k médoïdes de classes M_k

Algorithme k-medoids -version PAM- (Kaufman & Rousseeuw, 1987)

D'un point de vue pratique, les mesures des dissimilarités entre individus sont calculées une fois pour toutes et sont stockées dans la matrice des dissimilarités. L'algorithme utilise donc en entrée cette matrice, ce qui conduit à une converge plus rapide et réduit considérablement la complexité de l'algorithme.

Entrée[10] : D matrice de dissimilarité, k # classes

Sur le plan méthodologique, deux grandes approches sont souvent utilisées: la première, de type analyse statistique s'intéresse à une vision globale du jeu de données à disposition, la deuxième dite « étude de cas » consiste à rendre compte du caractère évolutif et complexe des phénomènes par l'étude des processus sous-jacents.

Si les partisans de l'approche statistique reprochent à « l'étude de cas » de tirer des conclusions spécifiques relatifs à un cas particulier sans pouvoir de généralisation, les partisans de l'approche « étude de cas » reprochent à l'approche statistique son caractère généraliste. L'algorithme des k-médoïdes présente un bon compromis entre ces deux approches, il permet une analyse statistique globale via les clusters et une analyse approfondie des processus via les représentants qu'il propose. Ces derniers étant des individus choisis (séries temporelles), ils conservent toute l'information et surtout tout leur sens contrairement à l'algorithme des K-moyennes où les représentants dépendent de l'opération de calcul des centres (calculer un représentant à partir de moyennes de séries de pluie n'aurait aucun sens).

[10] Cette ligne remplace la ligne d'entrée dans la version initiale de l'algorithme

5.4. Application à la classification des événements de précipitations

L'analyse basée sur la mesure des dissimilarités IMs-DTW se veut une approche alternative à l'approche basée sur les caractéristiques avec pour objectif de minimiser les problèmes dus à la discrétisation liée aux augets.

Afin de pouvoir comparer avec les classifications du chapitre 2, qu**atre expériences de classification par k-médoïdes ont été menées sur le même ensemble des 234 événements en reprenant des conditions identiques à celles du chapitre 3 : agrégation temporelle et/ou discrétisation de l'auget.**

5.4.1. Représentant (série temporelle moyenne) des 234 événements au sens de la mesure de dissimilarité IMs-DTW

On s'intéresse d'abord au représentant de tout l'ensemble des 234 événements. L'algorithme des k-médoïdes est utilisé en prenant la matrice des dissimilarités D (calculée sur l'ensemble des 234 événements) comme matrice des dissimilarités, et un nombre de représentants $k = 1$. L'algorithme retourne le représentant optimal de cet ensemble au sens de l'IMs-DTW. Ce dernier étant proche du centre du nuage des événements, il fournit une information sur le sens que l'on peut attribuer à la mesure de dissimilarité. Ainsi, la classification avec **k-médoïdes avec k=1** peut aider à l'interprétation de la mesure de l'IMs-DTW. Pour les expériences menées, quel que soit l'initialisation de départ, le représentant retourné est toujours 121ème événement correspondant au 22 décembre 2012. La figure 5.12 présente cet évènement « vu » par les trois instruments.

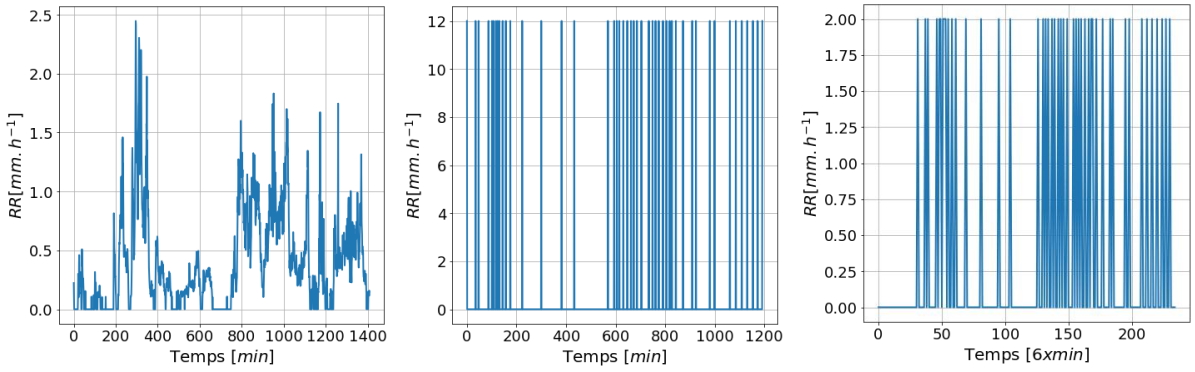


Figure 5.12 séries temporelles représentant le 121ème événement du 22 décembre 2012: à gauche série DBS pour T=1min, au centre pseudo-pluviomètre v=0.2mm à T=1min et à droite le pseudo-pluviomètre v=0.2mm à T=6min

Le fait que l'événement du 22 décembre 2012 soit le représentant des événements quel que soit le type de données (DBS, pseudo-pluviomètre, temps d'agrégation différent) confirme la stabilité de l'IMs-DTW. Cependant, cette approche présente l'inconvénient de ne pas fournir un résultat interprétable, autrement dit, dans notre cas on sait que l'événement est le représentant optimal au sens de l'IMs-DTW et que ce résultat est insensible au changement d'instrument et au temps d'agrégation (dans une certaine limite) mais on ne sait pas pourquoi cet évènement a été choisi. D'un autre côté, l'approche basée sur les caractéristiques (telles qu'elles ont été définies par exemple au chapitre 2), certes plus sensible au type d'instrument et au temps d'agrégation, garde l'avantage de fournir un résultat interprétable d'un point de vue physique. Ce constat motive l'interprétation des résultats des k-médoïdes à l'aide de caractéristiques physiques.

L'analyse par caractéristiques basée sur des cartes de Kohonen et présentée au chapitre 2 a montré que la majorité des caractéristiques sont structurées selon la première ou la deuxième diagonale donnant suite à deux comportements qui caractérisent deux sous-groupes de caractéristiques (voir fig. 2.4 chap. 2) :

- le premier groupe suit la première diagonale et est représenté par l'intensité maximale R_{max} et l'écart-type des intensités σ_R et les caractéristiques corrélées avec. Ce groupe décrit la variabilité des intensités des événements de pluie.
- Le deuxième groupe suit la deuxième diagonale et est représenté par la durée D_e et les caractéristiques corrélées avec. Ce groupe décrit les variations du support de pluie (de l'intermittence).

En se basant sur l'analyse par caractéristiques, l'événement du 22 décembre 2012 (affecté au neurone 64 –bas à droite dans la carte) présente les valeurs moyennes selon le premier groupe de caractéristiques (variations des intensités) (ex. intensité maximale, écart-type des intensités,…) mais des valeurs extrêmes pour le deuxième groupe.

La mesure de dissimilarité $d_{IMs-DTW}$ décrit les variations des intensités de pluie (la variabilité temporelle des précipitations) sans considérer les variations du support (ex. la durée) ; un résultat plus ou moins attendu par rapport à la normalisation considérée qui visait une indépendance des durées (# longueurs) des événements.

Toutefois l'information associée au deuxième groupe de caractéristiques (dont la durée,…) en plus d'autres caractéristiques telle que la position des pics est conservée par le biais de l'alignement. **Une classification qui prendrait en compte l'information des alignements donnerait un représentant dont des valeurs du second groupe seraient proches des valeurs moyennes obtenues sur l'ensemble des individus.**

Nous verrons dans la partie 3.4 l'exploitation de l'information de l'alignement et sa corrélation avec le deuxième groupe de caractéristiques.

Analyse de l'événement n° 121 du 22 décembre 2012

L'événement 121 est affecté au 64ème neurone qui compose la 3ème sous-classe (cf. chapitre 2), cette sous classes regroupe au total 3 événements. Il est caractérisé par une faible variation d'intensité de pluie : une intensité maximale R_{max} et un écart-type des intensités σ_R faibles en totale adéquation avec la caractéristique moyenne de la région d'Île-de-France. Le caractère de durée de cet événement ne représente pas la moyenne des durées des événements car la mesure de dissimilarité ne considère pas cette information. De plus la figure 5.12 montre un résultat très intéressant, du fait de sa faible intensité de pluie, cet événement génère des basculements orphelins lorsqu'on considère les pseudo-pluviomètres au temps d'agrégation d'une minute (si un basculement survient alors on observerait un et un seul basculement par pas de temps). Cette constatation reste valable avec un temps d'agrégation de $T = 6min$. De même, l'analyse sur les 234 événements montre que les événements précédemment classés stratiformes (cf. chapitres 2 et 3) présentent aussi cette particularité. Ce résultat est en adéquation avec le climat en Île-de-France dominé par les précipitations de faibles intensités.

Puisque la mesure de dissimilarité semble liée aux variations de l'intensité, les classifications résultantes par les k-médoïdes pour différentes valeurs de k présenteraient des variations d'intensités (variabilités) similaires entre les évènements d'une même classe et des variations dissimilaires entre des classes différentes. L'existence d'une classification classique (basée sur la physique des processus) en deux classes suggère dans un premier temps de choisir $k = 2$, et de considérer l'algorithme à des fins de classifications et non d'échantillonnage. A partir de $k > 3$, on considérera l'algorithme à des fins d'échantillonnage. L'étude pour k égal à 2 puis à 3 est présentée à la partie suivante, puis les résultats de l'échantillonnage pour $k > 3$ font l'objet de la partie d'après.

5.4.2. Classification pour $k = 2$ et $k = 3$

Initialisation : notre jeu de données comprend 234 événements, on dispose donc de $C_{234}^2 = 27261$ initialisations différentes des 2-représentants. La convergence rapide de l'algorithme a motivé l'exploration de toutes les combinaisons possibles.

Traitement : on garde la liste des représentants (médoïdes) la plus fréquente pour chaque expérience.

Analyse des résultats : deux aspects sont discutés, le premier : les représentants retournés par l'algorithme, le deuxième : la stabilité des clusters (classes).

Analyse des médoïdes :

	Disdromètre ($v = 0\ mm$)		Pseudo-Pluviomètre ($v = 0.2\ mm$)	
Temps d'agrégation	$T = 1\ min$ Expérience 0	$T = 6min$ Expérience 1	$T = 1min$ Expérience 2	$T = 6min$ Expérience 3
$k = 2$ Représentants des deux classes	[2 121]	[2 121]	[2 41]	[2 121]
$k = 3$ Représentants des trois classes	[2 121 109]	[2 121 109]	[2 121 41]	[2 121 41]

Tableau 5.3. Les représentants des classes retournés (combinaison la plus fréquente) par l'algorithme des k-médoïdes pour les quatre expériences menées pour k=2 et k=3 classes.

Pour $k = 2$, le représentant numéro 2 est présent pour les 4 expériences, il s'agit de l'événement du 03 janvier 2012 entre 14:39 et 19:47 (figure 5.13). Il représente une classe d'événements caractérisée par de fortes variabilités de l'intensité. Concernant le représentant de l'autre classe, l'évènement 121 a été sélectionné dans trois cas sur quatre. Cet évènement est caractérisé par au plus un basculement par pas de temps, une caractéristique commune aux événements de cette classe (stratiformes). Seule, l'expérience numéro 2 a retenu comme représentant l'événement 41 (Figure 5.14), qui apporte une quantité d'eau moindre, à la place de l'évènement 121 (Figure 5.12).

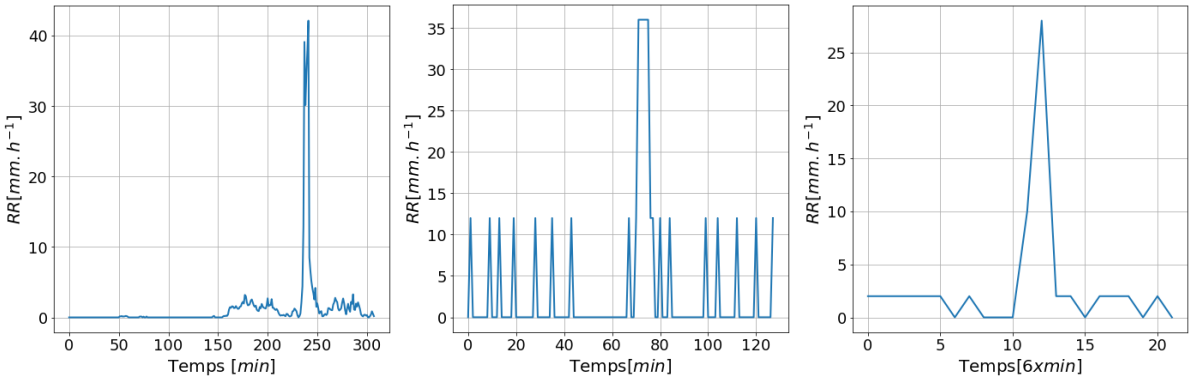


Figure 5.13 Evénement n°2 du 03 janvier 2012 (14:39 à 19:47): à gauche série DBS pour T=1min, au centre pseud-pluviomètre v=0.2mm à T=1min et à droite le pseudo-pluviomètre v=0.2mm à T=6min

Analyse de l'événement n° 2 du 3 janvier 2012 :

L'événement 2 est caractérisé par une variation importante de l'intensité de pluie : une intensité maximale R_{max} élevée et un écart-type des intensités σ_R et un β_L relativement forts en adéquation avec la description d'un événement convectif. La figure *5.*13 met l'accent sur un effet très intéressant, du fait de sa forte intensité maximale R_{max}, cet événement garde relativement la même forme au format pseudo-pluviomètre. Cette constatation reste valable lors d'une agrégation à $T = 6min$. De même, l'analyse sur les 234 événements montre que les événements précédemment classés convectifs (cf. chapitres 2 et 3) présentent aussi cette particularité.

Discussion de la stabilité de la classification

		Disdromètre ($v = 0\ mm$)				Pseudo-Pluviomètre ($v = 0.2\ mm$)			
Appareil	Temps d'agrégation	$T = 1\ min$ Expérience 0		$T = 6min$ Expérience 1		$T = 1min$ Expérience 2		$T = 6min$ Expérience 3	
Disdromètre ($v = 0\ mm$)	$T = 6min$ Expérience 1	9	6						
		12	207						
Pseudo-Pluviomètre ($v = 0.2\ mm$)	$T = 1min$ Expérience 2	16	4	8	12				
		5	209	7	207				
	$T = 6min$ Expérience 3	7	6	10	3	6	7		
		14	207	5	216	14	207		

Tableau 5.4. Les différentes matrices de confusion par paires de classifications associées aux quatre expériences pour k=2.

Remarques générales : concernant l'effet de l'agrégation : à chaque comparaison et quels que soient les instruments de mesure utilisés on trouve ~20 événements mal classés (21 dans le cas

du pluviomètre et 18 dans le cas du disdromètre) alors que dans le cas d'une discrétisation par l'auget sans changement d'échelle temporelle, le nombre d'événements mal classés est au plus de 8~9.

Pour k=3-médoides, la classification se raffine et la classe des événements à intensités faibles se divise donnant lieu à deux classes.

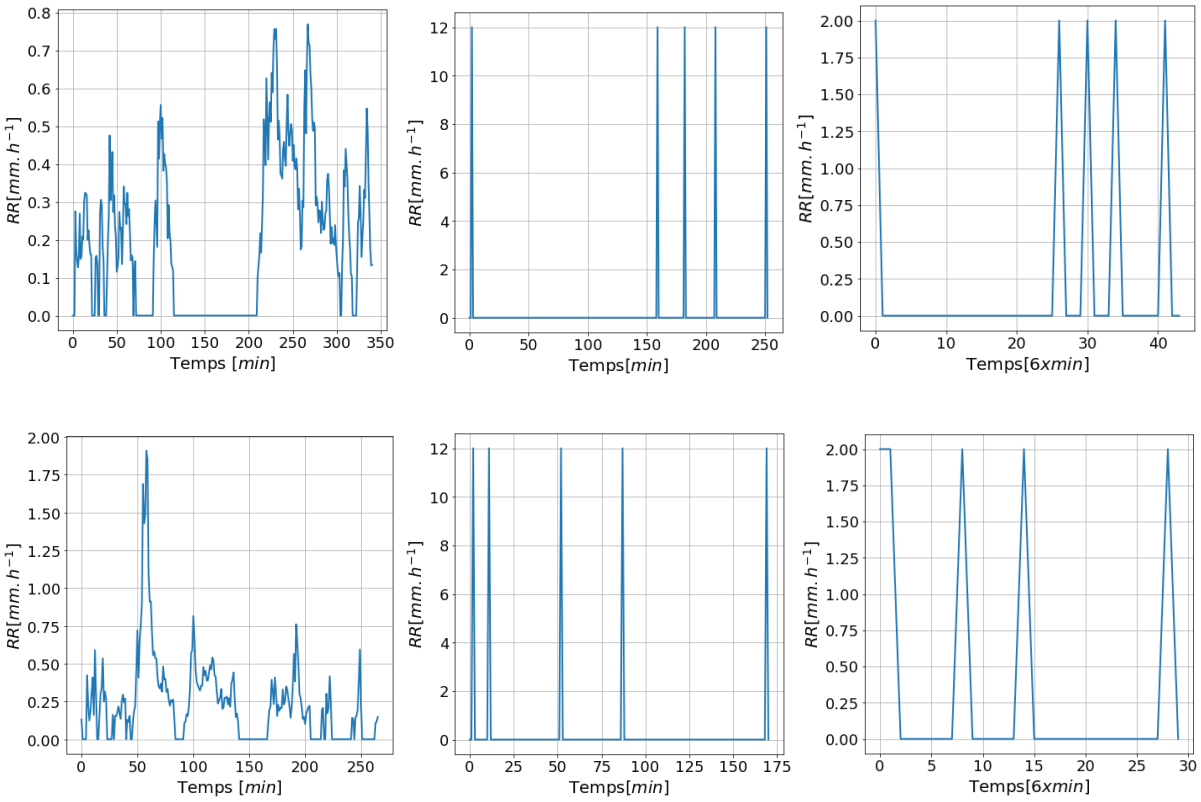


Figure 5.14 séries temporelles de l'événement 41 (figures de haut) et de l'évènement 109 (figure en bas): à gauche DBS pour T=1min, au centre pseudo-pluviomètre v=0.2mm à T=1min et à droite le pseudo-pluviomètre v=0.2mm à T=6min

On a vu au chapitre 3 que le temps qui sépare deux basculements successifs du même événement est corrélé positivement à l'intensité enregistré : plus l'intensité est faible, plus la durée qui sépare deux basculements successifs est longue. Les deux événements représentants 41 et 109 représentent qui représentent des pluies faibles ne manifestent pas une différence notable, ce qui pourrait expliquer le fait que l'algorithme propose le 41ème événement à la place du 109ème comme représentant après discrétisation par l'auget.

De plus, à quelques différences près, la classification en trois classes propose des pourcentages comparables et un ordonnancement de l'espace des événements similaire à celui proposé au chapitre 2 (page 42 et 43). On peut voir en effet que la classe des événements de

faibles intensités représentée par le 41ème événement regroupe en moyenne ~80% du nombre total des événements (comparable au 80%), la classe des événements longs représentée par le 121ème événement regroupe ~11% (comparable au 8%) alors que la classe des événements intenses représentée par le 1er événement recense ~9% (comparable au 12%). En outre, l'ordre des dissimilarités entre les événements propose d'ordonner les trois classes d'une façon similaire à celui proposé par la carte topologique du chapitre 2 :

$$(classe\ des\ événements\ de\ faibles\ intensités\ (\#41)\ \rightarrow$$
$$classe\ des\ événements\ longs\ (\#121) \rightarrow classe\ des\ événements\ intenses\ (\#1)).$$

Ce résultat nous a encouragé à étudier la question suivante : si on s'intéresse à un problème d'échantillonnage, effectuer une classification avec k>3 donnerait-il les mêmes représentants pour les différentes expériences (volumes d'augets et temps d'agrégation)?

5.4.3. Classification en plusieurs classes (k>3)

Un autre point intéressant est l'utilisation des k-médoïdes pour l'échantillonnage (réduire le nombre d'événements à traiter). La stabilité de la liste des représentants pour différentes valeurs de k vis-à-vis du changement d'instrument et de l'agrégation temporelle est un résultat appréciable.

Pour toutes les valeurs de k comprises entre 4 et 100, nous avons testé empiriquement si pour la même initialisation, on trouverait les mêmes représentants pour les quatre expériences précédemment décrites ($v = 0.2\ mm$ et/ou $T = 6min$). Le nombre de cas possibles étant très grand, on a testé 10000 initialisations pour chaque valeur de k testée. Le résultat était très encourageant, pour toutes les valeurs de k testées, nous avons trouvé exactement les mêmes représentants pour les différentes expériences. Ce résultat tend à montrer la robustesse de l'utilisation des k-médoïdes exploitant l'IMs-DTW vis-à-vis d'un changement d'instrument et/ou du temps d'agrégation.

5.4.4. Exploitation de l'alignement (information expliquée par l'alignement)

A la fin de la partie 3.1, nous avons avancé que l'alignement retourné par l'IMs-DTW contient une information importante qui peut être associée au 2ème groupe des caractéristiques du support (représentées par la durée D_e , et le coefficient P_{C1}). Dans cette partie nous allons fournir quelques éléments pour étayer cette hypothèse.

L'alignement retourné par l'IMs-DTW est représenté sous la forme d'un graphe. En conséquence, l'exploitation de l'information expliquée par cet alignement est quelque peu

compliquée. Une solution simple consiste à résumer cette information à l'aide d'un unique descripteur. Pour cela, plusieurs travaux proposent la définition de descripteurs comme par exemple Sorlin et al. (2003) dans le cas d'un alignement en général, et Petitjean (2011) pour décrire un type d'alignement DTW.

Pour illustrer le lien nous avons choisi le descripteur d'alignement proposé dans Sorlin et al. (2003) pour la comparaison des graphes[11], et adapté pour notre problématique :

$$desc_{alignement}(i,j) = \frac{\max(N_i, N_j) + split(aligement)}{longueur(alignement)} \quad (5.1)$$

Avec :

N_i $(resp.\ N_j)$ **:** la longueur de la série temporelle i (resp. j)

$split(aligement)$ **:** le nombre d'éclatements de l'alignement (distorsion temporelle)

$longueur(alignement)$ **:** La longueur de l'alignement trouvé entre la série i et la série j

Le descripteur d'alignement $desc_{alignement}(i,j)$ est symétrique, et définit une nouvelle mesure de dissimilarité entre les deux séries temporelles i et j. La comparaison des 234 événements donne lieu à une nouvelle matrice des dissimilarités notée $D_{alignement}$ difficile à interpréter.

Le positionnement multidimensionnel non métrique (NMDS) (Kruskal, 1964)[12] appliqué à la matrice $D_{alignement}(234 \times 234)$ propose un espace de représentation de petite dimension (2 ou 3 dans notre cas) qui conserve au mieux l'information de l'espace initial (dissimilarités contenues dans la matrice $D_{alignement}(234 \times 234)$). Autrement dit : les points (# événements) de l'espace de représentation seront positionnés de telle sorte que les distances des uns aux autres dans cet espace reflètent aux mieux les dissimilarités contenues dans la matrice $D_{alignement}$. Par conséquent, deux événements ayant un descripteur d'alignement faible seront voisins dans l'espace de représentation et vice versa. Le tableau 5.5 présente les différentes configurations (NMDS) des 234 points représentants les événements associées aux différentes matrices de dissimilarités (Exp. 0, 1, 2 et 3) en $deux\ dimensions$, puis en $trois\ dimensions$. La couleur des points (#événements) est en fonction de la variable durée de l'événement $(De)_v^T$ pour T et v variant en fonction de l'expérience.

[11] Ce descripteur est initialement inspiré de l'index de Tversky (1977)

[12] Le positionnement multidimensionnel non métrique est une technique utilisée pour la visualisation d'information pour explorer les dissimilarités des données, Elle peut etre définie comme une procédure de construction d'une configuration de points qui donne la meilleure approximation de la matrice des dissimilarité

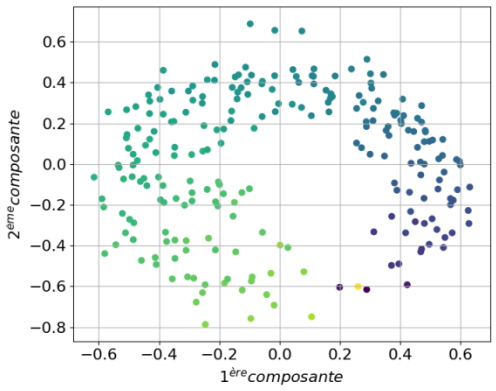
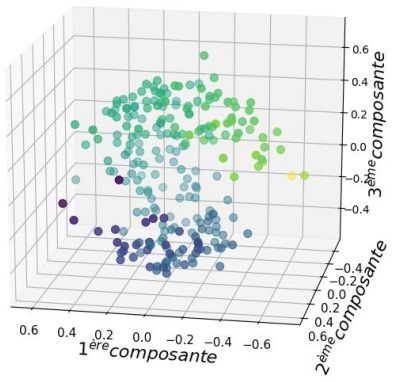
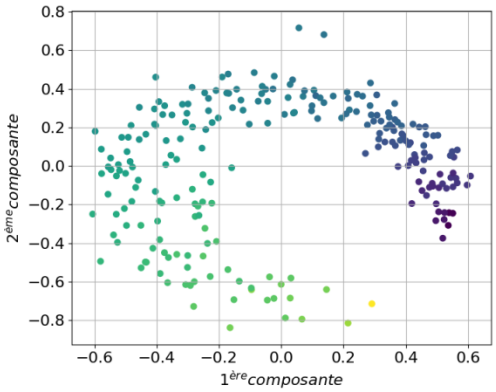
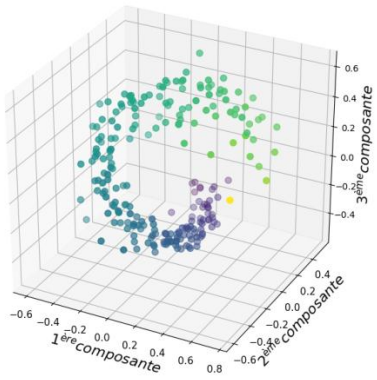
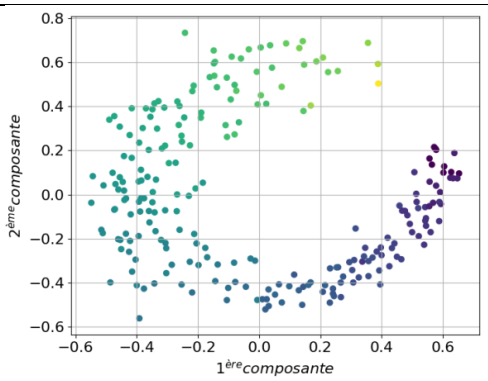
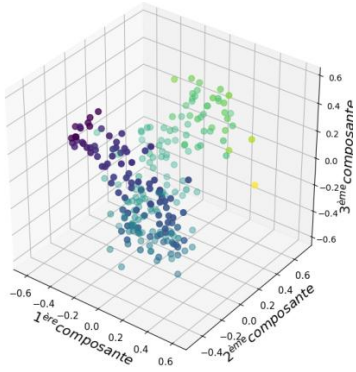
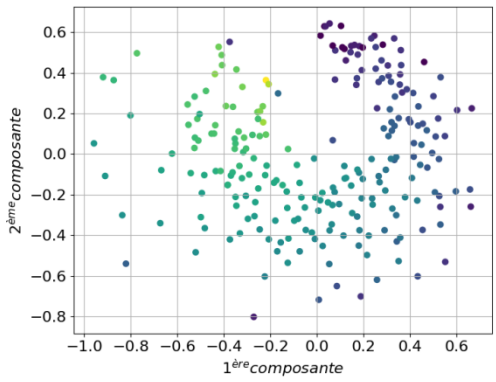
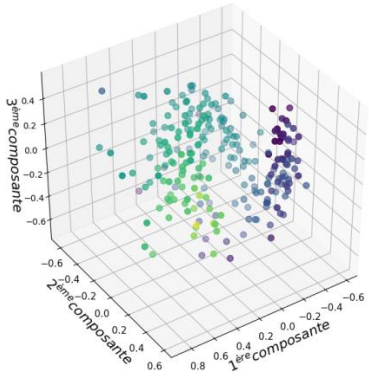
Variable de coloration : Durée de l'événement $(De)_v^T$		
	Espace NMDS	
Dim de l'espace	2	3
Exp. 0 ($v = 0\ mm$) $T = 1\ min$		
Exp. 1 $T = 6min$ ($v = 0\ mm$)		
Exp. 2 ($v = 0.2\ mm$) $T = 1\ min$		
Exp. 3 ($v = 0.2\ mm$) $T = 6min$		

Tableau 5.5. Visualisation en 2D (colonne 2) et en 3D (colonne 3) par NMDS de la matrice des alignements avec une échelle de couleur relative à la variable Durée de l'événement

Concernant l'expérience 0 et pour le positionnement en deux dimensions, le nuage de points est structuré sous forme d'anneau, les couleurs des points qui représentent la durée de l'événement vont du bleu (7 minutes) au jaune (1 jour) et donnent lieu à un gradient homogène. Cependant cette représentation laisse à penser que les événements longs sont positionnés aux cotés des événements courts. Le positionnement en trois dimensions conserve l'homogénéité du gradient durée et fait apparaitre une forme d'hélice et une différence entre les événements longs et courts apparait sur une dimension de l'espace de représentation (la première dimension pour l'exp. 0).

Les différentes structurations (anneau 2D / hélice 3D) ainsi que les colorations homogènes corroborent l'hypothèse qu'il existe un lien entre l'alignement issu de l'IMs-DTW et le deuxième groupe de caractéristiques (variabilité due au support de pluie) représenté par la variable «Durée de l'événement $(De)_v^T$ », et cela quel que soit l'expérience considérée 0, 1, 2 ou 3.

L'analyse présentée ci-dessus, bien que brève, suscite la question suivante : une combinaison du descripteur d'alignement et de la mesure de dissimilarité $d_{IMs-DTW}$ permettrait-elle de définir une nouvelle mesure de dissimilarité qui prendrait en considération l'ensemble des caractéristiques (les deux groupes de caractéristiques) ?

5.5. Définition d'une méthode d'analyse mixte

L'algorithme *k*-médoïdes basé sur les dissimilarités fournit $\boldsymbol{k}$ représentants optimaux en sortie. Dans notre cas on dispose donc de k séries temporelles de précipitations (# événements) réduisant l'analyse à cet ensemble uniquement. Or, comme nous l'avons vu dans la partie 5.3, la dissimilarité ne fournit aucune information permettant d'interpréter les résultats en termes de caractéristiques physiques. Autrement dit, on ne sait pas pourquoi certaines séries sont regroupées dans une même classe. Un tel point faible de l'approche basée sur les dissimilarités constitue le point fort de l'approche basée sur les caractéristiques où l'interprétation des résultats/caractérisation des classes est faite à l'aide de ces dernières (caractéristiques).

L'idée de cumuler les avantages des deux approches nous a poussés à considérer l'approche mixte pour l'analyse :

1- Effectuer une classification à l'aide de la méthode des k-médoïdes basée sur les dissimilarités IMs-DTW (mesure directe et multi-échelle des différences entre séries).
2- Choisir un ensemble de variables/caractéristiques descriptives pour analyser les k-classes : le choix peut être subjectif « on choisit la caractéristique qu'on sait interpréter »

sans que ce choix soit couteux, car le choix des variables n'impacte pas le résultat de la classification mais donne une information sur les caractéristiques partagées entre les membres de la même classe.

3- Analyser les k-médoïdes en utilisant les techniques d'études de cas (situationnelle) : « étude de processus », caractérisation et analyse du caractère évolutif… avec une relative possibilité de généralisation sur l'ensemble des séries temporelles de la même classe de chaque médoïde.

5.6. Conclusion

La mesure de dissimilarité a été développée pour s'affranchir de la description par caractéristiques des séries chronologiques de précipitation. Après avoir vérifié la stabilité de cette mesure par rapport au type d'instrument et au temps d'agrégation (paramètres T, v), nous avons utilisé cette mesure pour réaliser une classification des évènements avec un algorithme basé sur les k-médoïdes. Nous avons ainsi pu vérifier la grande stabilité des clusters obtenus relativement aux changements des paramètres (T, v) et le pouvoir discriminant de la mesure de dissimilarité. La partition obtenue regroupe les évènements ayant des valeurs similaires pour les variables caractéristiques du cumul, des intensités maximales ou moyennes. A partir d'un descripteur de l'alignement nous avons pu réaliser une matrice de dissimilarité des alignements entre séries et mettre en évidence les liens entre les alignements et les variables caractéristiques du support. L'alignement contient également des informations relative à la position des pics dans la série (qui n'ont pas été étudiées ici) ; Cette constatation ouvre à terme la perspective de pouvoir réaliser une classification des séries, robuste par rapport au type d'instrument et au temps d'agrégation et qui regroupe les séries ayant à la fois des variations des intensités similaires et des propriétés du support similaires. La mesure de dissimilarité des intensités comme celle des alignements permet de réaliser des classes mais ne permet pas directement d'interpréter les résultats en termes de caractéristiques physiques. Une approche mixte est ainsi proposée : les k-medoïdes basées sur les mesures de dissimilarité, permettent de réaliser des partitions robustes aux changements des paramètres (T, v) , les caractéristiques telles que celles définies au chapitre deux peuvent être utilisées pour l'interprétation des clusters obtenus.

Chapitre 6 : Analyse des séries temporelles de précipitations d'Île-de-France

6.1. Présentation des données

Dans le cadre de cette étude, Météo-France a mis à notre disposition des données d'observations du réseau de pluviomètres à auget de la région parisienne. La figure 6.1 montre la répartition des vingt-six stations. Des relevés pluviométriques sont disponibles à fine échelle ($T = 6min$) couvrant la période 2006-2015 (10 ans) pour vingt stations strictement incluses dans la région « Île-de-France » et six stations dans les régions voisines. Techniquement, les données représentent des séries temporelles de cumul de précipitations issues des pluviomètres à augets de volume $v = 0.2\,mm$, les données sont ensuite agrégées à $T = 6min$ puis stockées dans la base de données au début de chaque heure (par paquet de 10 relevés), avec une précision au dixième de mm[13].

[13] Documentations météo-France disponible sur le site

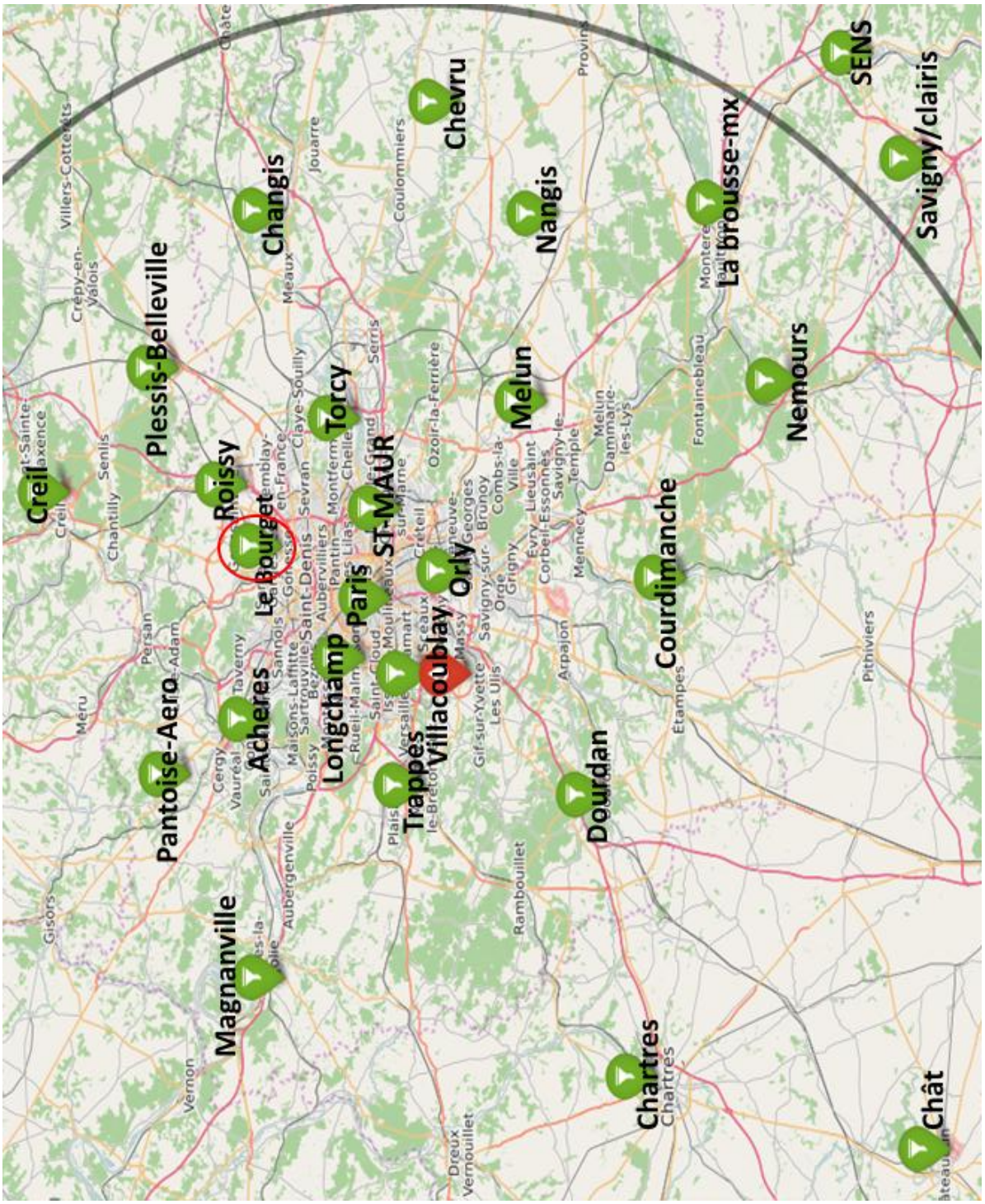


Figure 6.1 répartition des 26 stations Météo-France (en vert) sur la région d'Île-de-France et les régions voisines, le symbole rouge montre l'emplacement de la plateforme SIRTA ayant servie pour les chapitres précédent, le cercle rouge entoure la station du Bourget (représentante des 26 stations au sens des K-médoïdes avec k=1)

On note la densité du réseau autour de Paris et la diminution de cette densité en s'éloignant du centre.

Des données historiques des deux stations « Paris-Montsouris » et « Saint-Maur » au pas de temps journalier (T=1jour) pour des périodes relativement longues (142 et 136 années respectivement) sont également mises à disposition et font l'objet de l'étude climatique dans la partie 6.5.

6.1.1. Prétraitement des données

Les données mises à disposition par météo-France étant sous format brut, un prétraitement a été appliqué. Ce prétraitement a consisté à supprimer les données aberrantes (valeurs négatives de cumul d'eau), identifier les périodes des données manquantes pour différentes causes, …

Après nettoyage et plusieurs réunions avec des experts météo-France[14], nous avons conclu que seule la période 2009-2015 permettait de disposer simultanément des observations sur les 26 stations.

6.1.2. Objectifs de l'étude

Les données des 26 stations couvrant les sept années de la période 2009-2015 peuvent être exploitées de plusieurs façons:

- Une classification des stations où les données d'une station sont considérées comme une seule série temporelle, la classification des 26 séries temporelles permet dans ce cas d'étudier la variabilité spatiale.
- Une classification des événements de précipitations où la série d'une station est décomposée en plusieurs sous-séries. Dans ce cas, nous verrons que la classification des événements de précipitations peut généraliser les résultats des chapitres précédents (2 et 5).
- Une classification (étendue spatialement) des événements de précipitations où un événement est défini par 26 séries temporelles marquant le passage de ce dernier sur chaque station. La classification étendue des événements pourra permettre d'associer aux types de précipitations des comportements particuliers.

[14] Olivier Laurentin était membre du premier comité de thèse

Par ailleurs, les données des deux séries historiques « Paris-Montsouris » et « Saint-Maur » seront utilisées pour investiguer l'impact du changement climatique sur l'évolution des précipitations.

La méthode de classification utilisée est celle décrite au chapitre précédent qui repose sur l'algorithme des k-médoïdes appliqué à la matrice de dissimilarité entre les individus à classer.

6.2. Étude de la variabilité spatiale des précipitations (classification des stations)

6.2.1. Analyse des dissimilarités globales entre stations

La classification des stations peut nous informer sur la variabilité spatiale des précipitations. La matrice de dissimilarité est constituée des 'dissimilarités entre stations'. La 'dissimilarité entre deux stations' est la dissimilarité obtenue en appliquant l'IMsDTW aux séries chronologiques complètes (7ans à la résolution de 6 min) de deux stations. La figure 6.2 présente la matrice des dissimilarités $D_s(26 \times 26)$ associée aux *vingt-six* stations, où d_{ij} est la mesure de dissimilarité obtenue par IMsDTW entre les deux stations i et j.

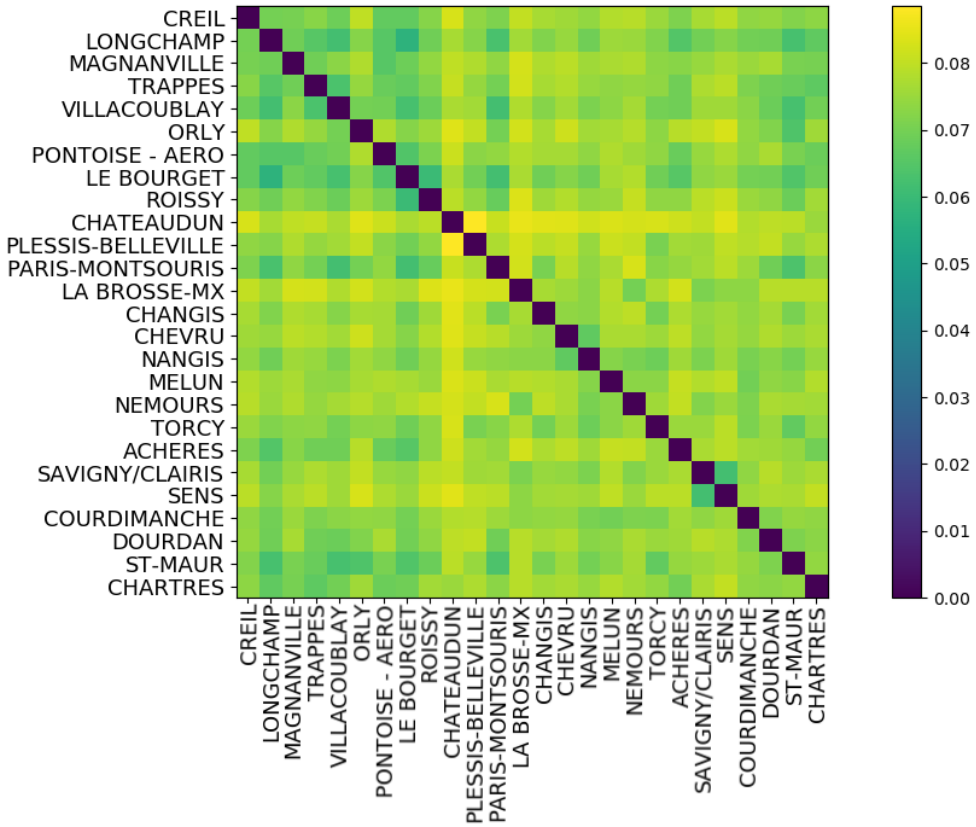


Figure 6.2 matrice des dissimilarités D_S (26×26) entre les vingt-six stations

On remarque que les valeurs des dissimilarités sont faibles par comparaison aux valeurs du chapitre cinq partie 5.1.1. Cette différence notable est causée par la normalisation (formule 11 du chapitre 4) par les longueurs des séries. Dans le cas de séries complètes, contrairement au cas des évènements, les périodes inter-événements (séries identiquement nulles pour toutes les stations) sont globalement beaucoup plus longues que les évènements.

De plus, on constate que les valeurs faibles des dissimilarités caractérisent la comparaison des villes autour de Paris (ilot parisien). Ceci résulte de la densité spatiale du réseau des pluviomètres sur cette zone "parisienne". En effet, la dissimilarité explique bien la distance géographique entre les stations, par exemple, la station de Châteaudun située à une distance relativement importante des autres stations enregistre des valeurs de dissimilarités relativement importantes aussi sur la matrice des dissimilarités. La figure 6.3 affiche les dissimilarités entre les stations en fonction des distances géographiques.

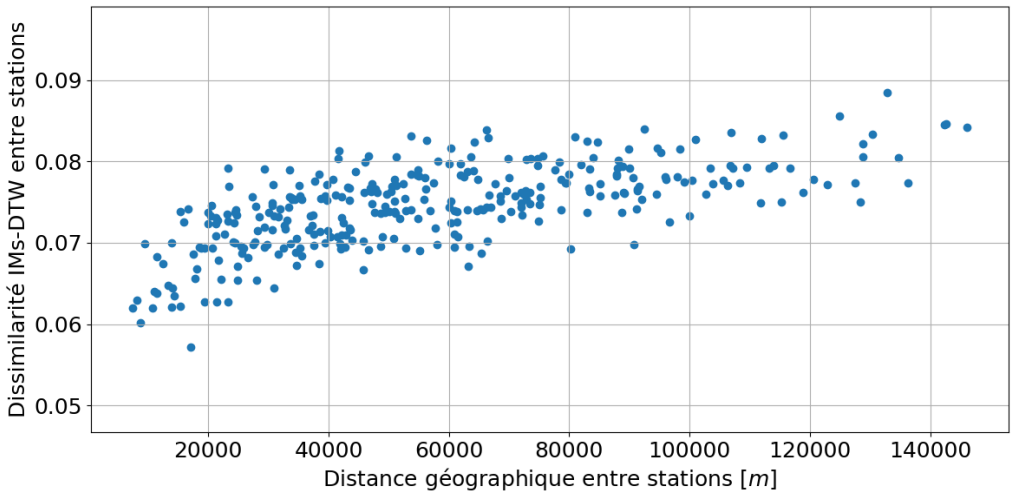


Figure 6.3 la dissimilarité entre les stations en fonction de la distance géographique (valeurs des diagonales ignorées)

La valeur de dissimilarité en fonction de la distance géographique enregistre une tendance positive. Le coefficient de corrélation de Pearson vaut $r = 0.66$ (0.71 *en échelle* $log - log$) entre la dissimilarité mesurée entre deux stations et la distance à vol d'oiseau qui les sépare, confirme l'existence de cette relation. La station de Châteaudun, est visiblement la station la moins similaire aux autres, ce qui peut s'expliquer par sa position géographique relativement éloignée.

Comme dans le chapitre précédent, on obtient le représentant de tout l'ensemble des 26 stations de mesures en appliquant l'algorithme des k-médoïdes appliqué à la matrice des dissimilarités D_{26} (calculée sur l'ensemble des 26 stations) avec k=1. La station du **"Bourget"**

(#7) est le représentant de l'ensemble des stations. Le cercle rouge sur la figure 6.1 montre la position de cette station. L'analyse détaillée de cette station fait l'objet de la section 6.3.1.

La classification des stations en plusieurs classes pose le problème du choix du nombre de classes k. Analyser l'évolution de la distorsion du nuage après classification peut présenter un moyen pour choisir ce nombre de classes. La distorsion représente la perte d'information qui résulte d'une classification en k classes (i.e. erreur de quantification dans le cas d'une dissimilarité). Par définition, décroissante en fonction de k, une distorsion relative de 100% est associée à $k = 1$, ce qui veut dire qu'avec une classification en une seule classe (un seul représentant) on perd toute l'information sur la variabilité du nuage de points. D'autre part, une classification en $k = 26$ (avec N le nombre d'individus #séries temporelles) ne produit aucune perte d'information sur la variabilité du nuage et de ce fait la distorsion est égale à 0. La figure 6.4 présente la distorsion relative pour la classification des stations en fonction du nombre de classes utilisé k en variant $k = 1 \dots 26$.

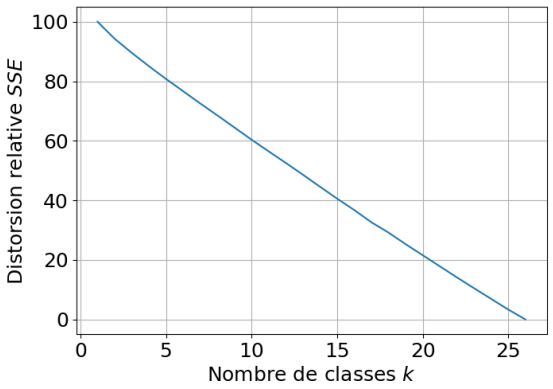


Figure 6.4 distorsion relative du nuage des stations en fonction du nombre de classes k

La courbe des distorsions relative sur l'ensemble des 26 stations a une allure décroissante monotone régulière. En l'absence de toute cassure (coude), cette allure laisse à penser à une structure de type continuum ne faisant intervenir aucun partitionnement marqué. De plus, pour toute valeur de k et pour chaque initialisation différente, l'algorithme converge vers une partition (classification) différente. Cependant, un point commun à toutes ces partitions est la conservation de la topologie spatiale (i.e. les classes présentent une continuité spatiale). Cette observation soutient fortement l'hypothèse d'un continuum. Une partition en deux classes divise l'ensemble des stations en deux groupes relativement équilibrés séparé par un axe variable selon l'initialisation de la méthode.

Pour conclure sur la variabilité spatiale des précipitations, nous pouvons dire qu'à l'échelle de la série complète, il semble difficile d'obtenir des informations sur la variabilité

spatiale des précipitations en exploitant seulement la mesure de dissimilarités. La zone d'étude est peu étendue, avec peu de reliefs et qui pourrait justifier un comportement particulier de quelques stations. De plus, la qualité des données est homogène sur l'ensemble des stations. Ces deux raisons justifient que le pluviomètre est globalement homogène sur toutes les stations ou et qu'aucune station ou groupe de stations ne se distingue des centres sur la période de 7 ans considérée globalement.

6.2.2. Analyse de l'information de l'alignement global des stations

Nous avons vu précédemment (chapitre 4) que dans le cas où l'alignement est interprétable, les deux descripteurs (moyenne et écart-type) permettent de fournir des renseignements sur la signification physique des alignements. Nous avons donc exploité cette propriété dans la partie suivante. La figure 6.5.a (resp. b) présente la matrice des moyennes (resp. écart-types) des alignements $\tau_{ATD}(26 \times 26)$ (resp. $\sigma_{ATD}(26 \times 26)$) associée aux couples des *vingt-six* stations, où $\tau_{ATD}[i,j]$ (resp. $\sigma_{ATD}[i,j]$)représente la moyenne des décalages entre les deux stations i et j.

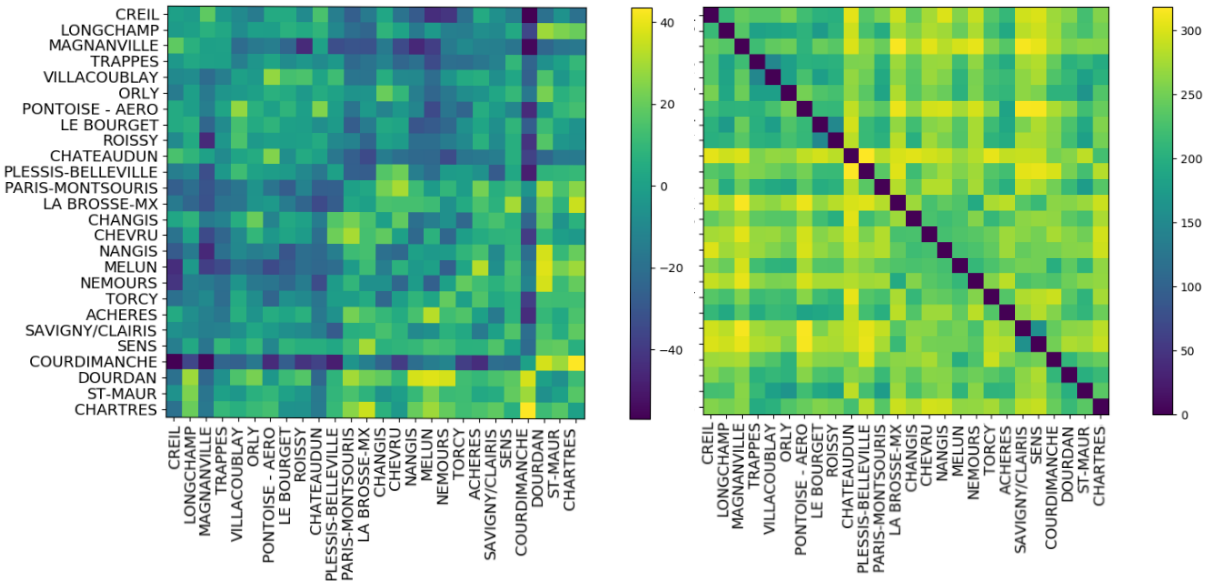


Figure 6.5 à gauche matrice des moyennes des décalages $\tau_{ATD}(26 \times 26)$ entre les vingt-six stations pas [6 x min], à droite matrice des écart-types des décalages $\sigma_{ATD}(26 \times 26)$

D'après les interprétations de la structure de l'alignement réalisées au chapitre 4, les écart-types faibles (en vert) sur la figure de droite correspondent à des couples de stations qui observent globalement les mêmes évènements. Pour ces couples, la moyenne des décalages fournie par la figure de gauche peut être interprétée comme un décalage temporel moyen. Les écart-types élevés (en jaune) correspondent à des couples de stations pour lesquels l'algorithme

a apparié ensemble des cellules de pluie distinctes, les décalages entre ses structures ne peuvent alors être interprétés comme des décalages temporels.

Pour les couples de stations proches telles que (Roissy ; Le Bourget), (Trappes ; Villacoublay), (Savigny; Sens), (St Maur; Orly) par exemple les écarts-types sont faibles, la plupart des événements appariés par l'algorithme correspondent bien à la même cellule pluvieuse observée successivement par les deux stations. La lecture de la carte de gauche nous indique dans ce cas dans quel ordre les deux stations enregistrent les évènements et avec quel écart de temps ce qui peut nous permettre d'estimer les vitesses d'advection moyenne.

A l'inverse, les couples de stations (Magnanville/Châteaudun), (Magnanville/Labrosse), (Chateaudun/ Plessi-Belleville) n'observent généralement pas les mêmes cellules de pluie et les valeurs des moyennes des décalages ne peuvent être interprétées.

Les valeurs relativement fortes observées sur la figure de gauche pour la station de Courdimanche nous indiquent des décalages temporels importants entre cette station et la plupart des autres stations. Cependant les valeurs de la figure de droite nous indiquent que les évènements précipitant observés par la station de Courdimanche sont généralement observés par Melun, Orly et Dourdan, souvent par Villacoublay Nemours ou Saint Maur (écart type faibles) et plus rarement par les stations au Nord de la capitale telles que Magnaville, Creil, Pontoise-Aero, Plessis-Belleville (écarts-type forts).

Inversement, la station du Bourget, qui est la station centrale du point de vue des dissimilarités (section précédente) présente des valeurs d'écart-type faible avec la plupart des stations excepté les stations au sud et à l'ouest de la capitale (Châteaudun, Chartres, Courdimanche, Nangis, La Brosse, Savigny, Sens et Chevreu).

6.3. Étude de la variabilité temporelle des précipitations (Classification des événements de précipitations)

La pluviométrie est globalement homogène sur la zone d'étude considérée, l'analyse par événement a pour objectif de compléter l'étude de cas présentée au chapitre 4 et qui ne concernait que quelques évènements et de vérifier si les résultats obtenus au chapitre 2 lors de la classification des évènements observés par le DBS à la résolution de *T=1 min* sont confirmés par la classification des évènements observés par les pluviomètres à la résolution de *T=6 min*.

6.3.1. Définition des événements de précipitations

Au chapitre 2, nous avions également choisi un temps inter-événement minimal (absence de précipitations) $MIT_{DBS} = 30\ min$. Cela nous a menés à l'identification de 120 événements par année en moyenne. Au chapitre 3, la transposition des résultats obtenus à l'aide du DBS sur les données pluviomètres, a montré que l'utilisation brute des paramètres et/ou seuils décrits au chapitre 2 n'est pas toujours pertinente, et avons suggéré une révision de ces paramètres (qui en effet dépendent de l'instrument et du temps d'agrégation utilisés).

Les données du réseau pluviométrique de Météo-France ajoutent une dimension spatiale à la définition des événements. Nous avons donc étendu cette définition (temporelle) du MIT à un aspect spatial en ajoutant la contrainte à la définition de ce dernier : absence de précipitation sur un certain laps de temps sur l'ensemble de la zone étudiée (150x150 km² environ). Le choix du temps inter-événement minimal n'est pas anodin, ainsi un $MIT = 30\ min$ semble inapproprié compte tenu des dimensions de la zone étudiée. En effet, une cellule de pluie peut fort bien se trouver au-dessus de la zone d'étude mais ne pas être détectée durant 30 minutes compte tenu de la distance inter stations pluviométriques. Dans ce cas de figure, le nombre d'évènement de pluie estimés est aux alentours de 2000, soit autour de 285 évènements par an en moyenne, ce qui est en contradiction avec le nombre moyen d'événement moyen trouvé au chapitre 2. Nous avons donc choisi empiriquement un $MIT = 3.2\ heures$ afin d'obtenir un nombre d'évènements sensiblement identique à celui trouvé au chapitre 2. Au final, nous obtenons **873 événements** sur les sept années de données, équivalent à une moyenne de 124 événements par année.

Finalement, un événement de précipitations est représenté par 26 séries temporelles de taux précipitants. Le tableau 6.1 présente un exemple de séries temporelles de précipitations, chaque colonne représentant une station et chaque ligne un évènement parmi les 873 événements identifiés sur les sept années d'étude. Le premier est un événement stratiforme avec un seul basculement de l'auget par pas de temps mais pour une durée totale relativement longue, l'événement 191 est de nature convective avec des intensités fortes et une durée faible, l'événement 873 est faible en intensité et de courte de durée.

La visualisation multi-site du même évènement permet, en particulier pour les évènements intense une analyse spatiale de l'évènement, l'événement 191 est observé par la station de Chartres, puis Le Bourget puis Creil ce qui correspond à une trajectoire Sud/Nord.

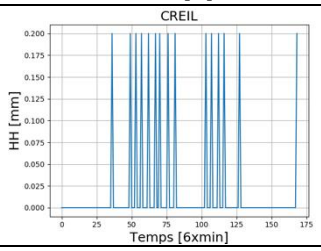
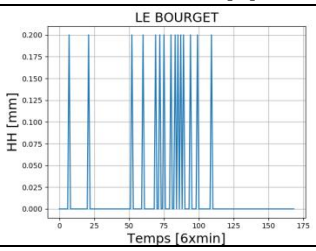
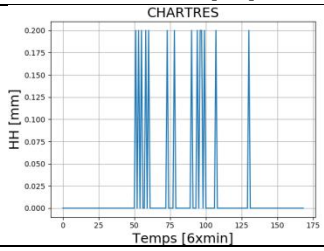
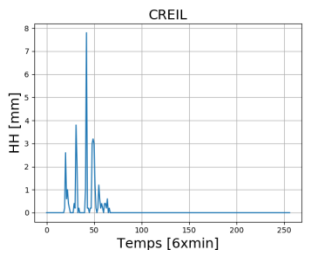
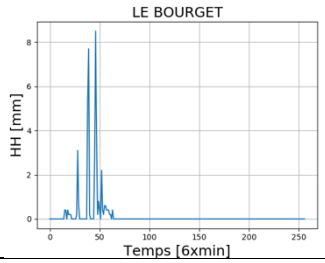
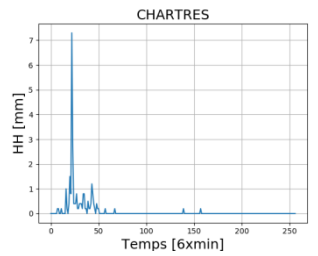
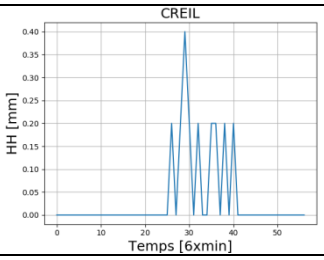
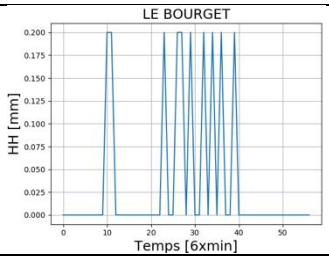
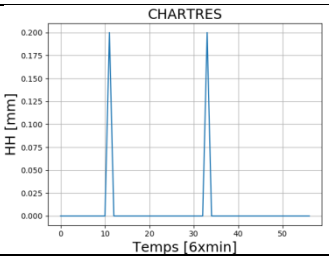
	CREIL [0]	...	Le BOURGET [7]	...	CHARTRES [25]
1		...		...	
⋮	⋮	⋱	⋮	⋱	⋮
191		...		...	
⋮	⋮	⋱	⋮	⋱	⋮
873		...		...	

Tableau 6.1 : Le tableau "événements X stations" représenté sous la forme classique en analyse des données

6.3.2. Classification des événements de précipitations pour la station du Bourget

Avant de nous intéresser à l'ensemble des stations, nous commençons par étudier la station du Bourget définie au paragraphe 6.2 comme étant la représentante la plus proche de l'ensemble des autres stations du point de vue de l'IMsDTW. La figure 6.6 représente la série temporelle des hauteurs d'eau à la résolution de six minutes. L'événement 191 enregistre le pic le plus important avec une valeur proche de 8,5 *mm*.

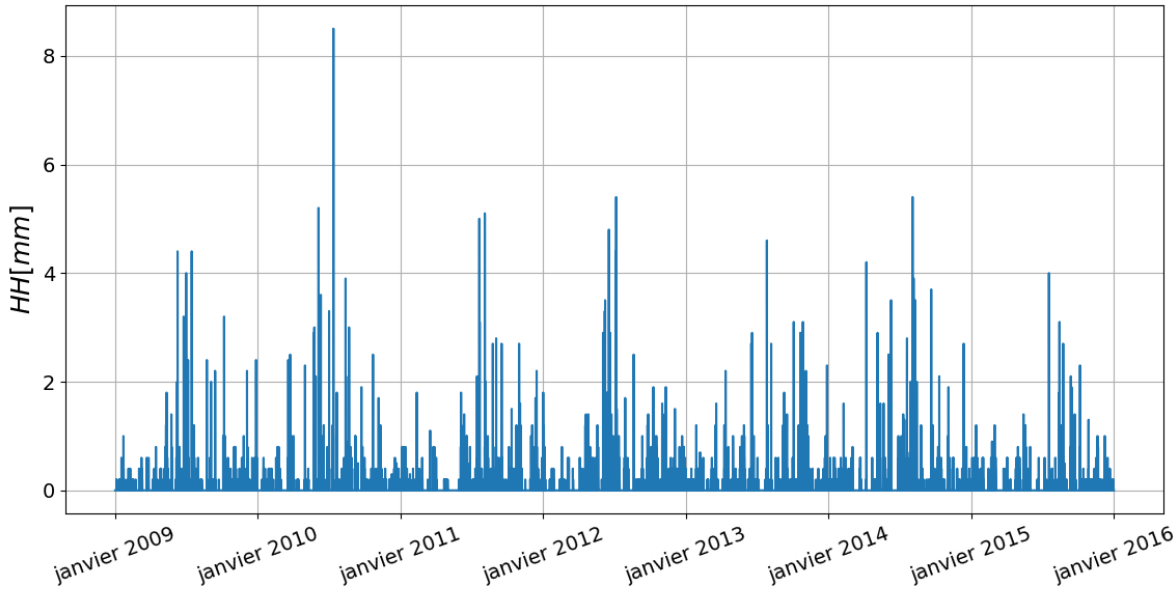


Figure 6.6 série temporelle mesurée par un pluviomètre (v=0.2mm) sur la station Le Bourget entre 2009 et 2015 à pas de temps T=6min

La figure 6.7 présente à gauche un exemple d'alignement proposé entre le premier et le deuxième événement, à droite la matrice des dissimilarités $D_{Ev}^{Bourget}(873 \times 873)$ où d_{ij} représente la mesure de dissimilarité entre deux événements i et j observés par la station du Bourget. Cette matrice de dissimilarités entre les 873 évènements observés par la station Le Bourget est utilisée pour la classification des évènements.

Rappelons que dans cette section, contrairement à la section précédente, l'IMDTW est appliquée à des événements différents observés par la station Le Bourget à des dates différentes, la structure d'alignement dans ce cas ne traduit pas une dynamique mais une ressemblance de type motifs/forme qui ne peut être interprétée.

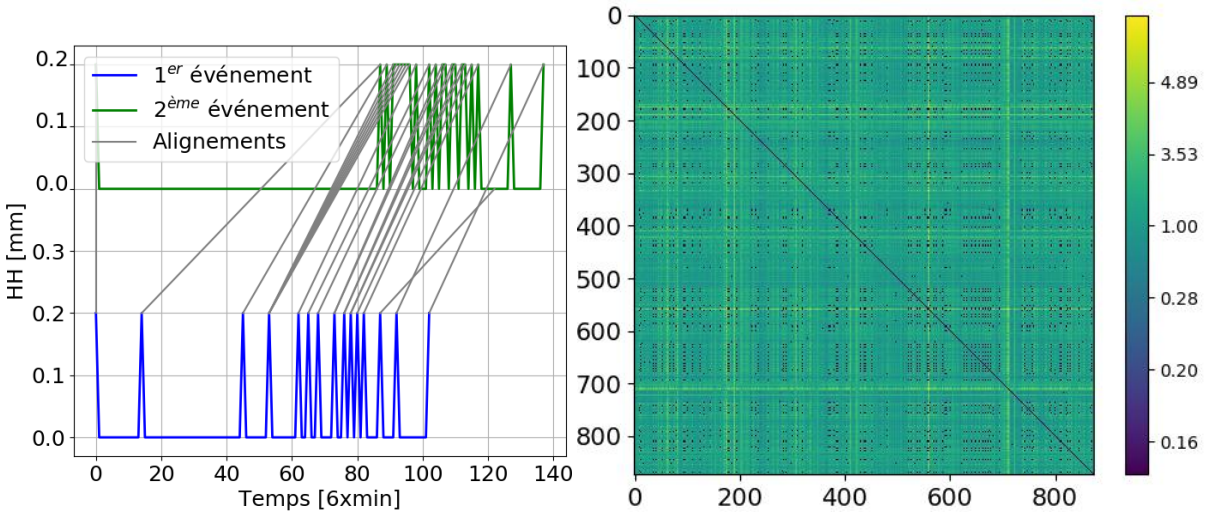


Figure 6.7 Pour la station le Bourget A gauche l'alignement trouvé entre le 1er événement en bleu et le 2ème événement en vert, les alignements sont représentés en gris, à droite la matrice des dissimilarités $D_{Ev}^{Bourget}(873 \times 873)$ entre les séries temporelles des 873 événements mesurées sur la station le Bourget

Sur l'échantillon des alignements visualisés, les appariements semblent en adéquation avec l'objectif (à savoir bien associer le support des précipitations en associant les débuts, les fins des évènements et en associant les pics à l'intérieur des évènements apparié). En analysant la matrice des dissimilarités $D_{Ev}^{Bourget}$ on note une occurrence importante de dissimilarité nulles (en bleue). Cela se produit chaque fois qu'il n'a pas plu sur la station du Bourget durant les deux événements considérés (par contre il a plu sur d'autres stations durant ces événements).

D'une façon générale, les résultats et les observations du chapitre 5 se vérifient, on remarque notamment que l'ordre de grandeur des dissimilarités est comparable à celui estimé sur les séries simulées du chapitre 5. La démarcation des événements intenses est également visible sur cette matrice (ex. L'événement 191 présenté dans le tableau 6.1 se démarque sur la matrice des dissimilarités et enregistre de fortes dissimilarités par rapport aux autres événements).

6.3.2.1. Représentant (série temporelle) des 873 événements au sens de la mesure de dissimilarité IMs-DTW

On s'intéresse au représentant de l'ensemble des 873 événements observés par la station du Bourget. L'algorithme des k-médoïdes est utilisé en prenant la matrice des dissimilarités $D_{873\times873}$ (calculée sur l'ensemble des 873 séries temporelles représentants les événements sur la station du Bourget) comme matrice des dissimilarités, et un nombre de représentants $k = 1$.

Quel que soit l'initialisation de départ, le représentant retourné est toujours le 262ème événement ayant traversé l'Ile de France entre 22 janvier 2011 à 08:00 et le 24 janvier 2011 à 19:12. La figure 6.9-a présente la série temporelle correspondante.

À propos de l'évènement 262 sur la station le Bourget:

Si on se limite à la station du Bourget, la caractéristique majeure de l'événement 262 est la présence d'au plus un basculement par pas de temps. Un événement de ce type a également été obtenu comme représentant de l'ensemble des évènements au chapitre 5.

6.3.2.2. Classification en plusieurs classes ($\mathbf{k = 2, 3}$ et 4)

La définition spatiotemporelle d'un événement tel que nous l'avons redéfini au début de la partie 6.3 n'interdit pas que lors du passage d'une petite cellule de pluie localisée, il est fort probable que celle-ci ne soit « vue » que par un nombre restreint de stations. La classification des 873 événements en deux classes sépare les évènements pluvieux (représentant événement 262) des évènements sans pluie (représentant événement 7). On vérifie en effet que la classe représentée par l'événement 7 ayant traversé l'Île-de-France entre le 19 janvier 2009 à 20:00 et le 20 janvier 2009 à 20:00 regroupe tous les événements "à taux précipitants nuls" sur la station Le Bourget.

Si on considère cet ensemble des événements "nuls" sur la station du Bourget comme une classe à part, la classification avec un nombre de classe $k = k_p + 1$ avec k_p le nombre de classes des événements pluvieux reproduit parfaitement les résultats obtenus au chapitre 5 partie 5.3.2, à savoir une division de la classe stratiforme en deux sous-classes.

Ainsi, la classification en trois classes ($k = 3$ / « $k_p = 2$ ») sépare la classe des événements pluvieux en deux classes « stratiformes et convectifs » représentées respectivement par les événements 262 et 389 ayant traversé l'ile de France entre le 20 avril 2012 à 01:36 et le 22 avril 2012 à 21:36. De même, la figure 6.8 présente l'évolution de la distorsion relative SSE en fonction du nombre de classes k choisi pour la classification des événements enregistrés sur la station le Bourget et suggère que le nombre de classes optimal pour la classification est de $k = 4$ ($k_p = 3$) corroborant ainsi les résultats des chapitres 2 et 5 sur **l'existence de trois classes.**

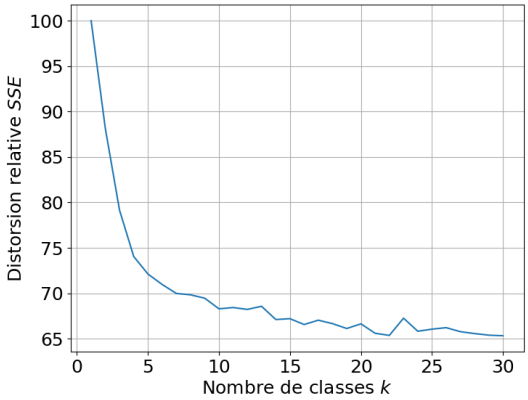


Figure 6.8 la distorsion relative SSE en fonction du nombre de classes k choisi pour la classification des événements enregistrés sur la station le Bourget, Un coude apparait à k=4

La classification en quatre classes ($k = 4$ et « $k_p = 3$ ») raffine la classification en séparant l'ensemble des événements stratiformes en deux classes à intensité moyenne d'une part et faible d'autre part, et sont représentées respectivement par les deux événements 262 et 272 ayant *traversé l'ile de France entre le 12 mars 2011 à 14:24 et le 13 mars 2011 à 20:48*.

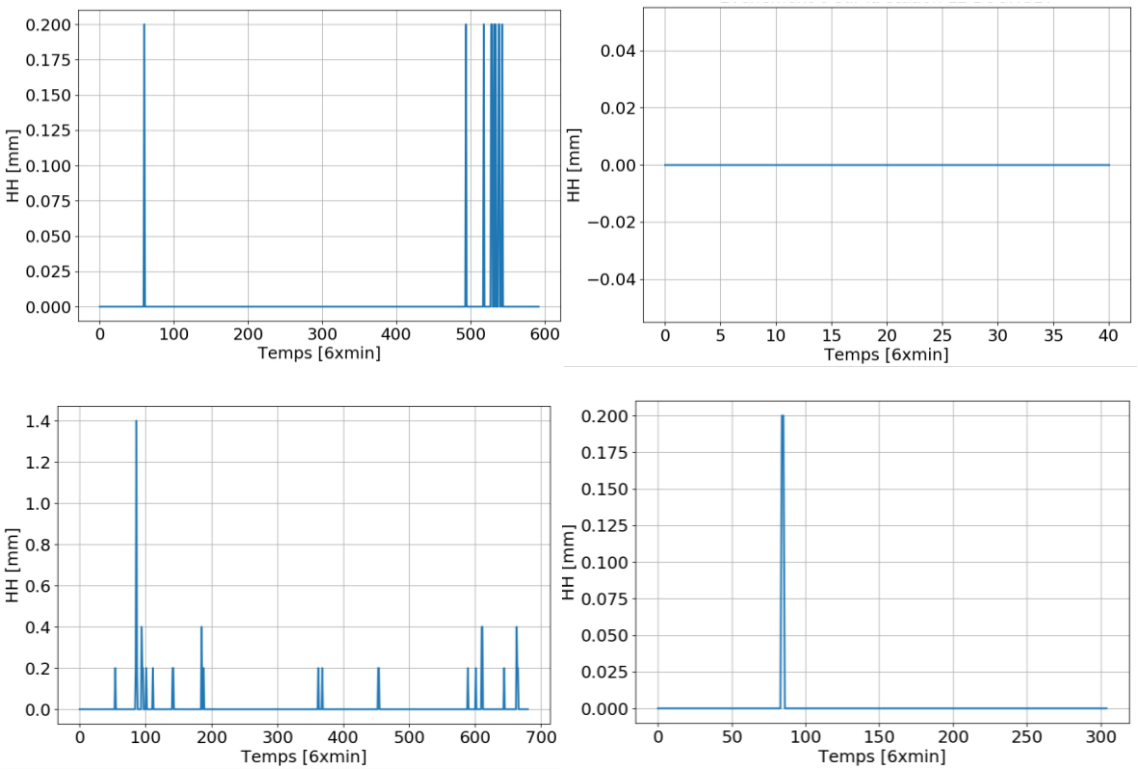


Figure 6.9 Séries temporelles représentant le passage des évènements représentant les classes à la station Le Bourget : En haut, à droite : l'événement 262 (traversant l'ile de France entre 22 janvier 2011 à 08:00 et le 24 janvier 2011 à 19:12). En haut, à gauche : l'événement 7 (traversant l'ile de France entre le 19 janvier 2009 à 20:00 et le 20 janvier 2009 à 20:00). En bas, à droite : l'événement 389 (traversant l'ile de France entre le 20 avril 2012 à 01:36 et le 22 avril 2012 à 21:36). En bas, à gauche l'événement 272 (traversant l'ile de France entre le 12 mars 2011 à 14:24 et le 13 mars 2011 à 20:48).

L'analyse des partitions et des représentants obtenus permet globalement de retrouver les classes d'évènements obtenues aux chapitres 2 et 5. Les évènements analysés n'étant pas les mêmes (pas la même station, pas la même période) qu'au chapitre 2 cette conclusion repose sur une analyse visuelle des référents et une analyse qualitative de la nature des évènements regroupés.

6.4. Étude de la variabilité spatio-temporelle des précipitations (classification des événements sur toutes les stations)

Considérer l'information spatiale n'est pas supposé changer le résultat mais juste donner plus d'informations sur la variabilité spatiale de chaque type d'évènements. Au final la question à laquelle on tente de répondre dans cette partie est: les différents types d'événements de précipitations présentent-ils des différences d'un point de vue spatial ?

Pour chaque couple d'évènement (i, j) , les dissimilarités « spatiales » sont calculées comme suit , à partir des dissimilarités du couple d'événement (i, j) sur chacune des 26 stations :

$$d_{Ev}^{S}(i,j) = \sqrt{\frac{1}{26}\sum_{k=1}^{26}\left(d_{Ev}^{k}(i,j)\right)^{2}} \quad (6.1)$$

6.4.1. Matrice des dissimilarités :

Les valeurs des dissimilarités "spatiales" entre les 873 événements sont stockées dans la matrice des dissimilarités $D_{Ev}^{S}(873 \times 873)$. La figure 6.10 de gauche présente cette matrice.

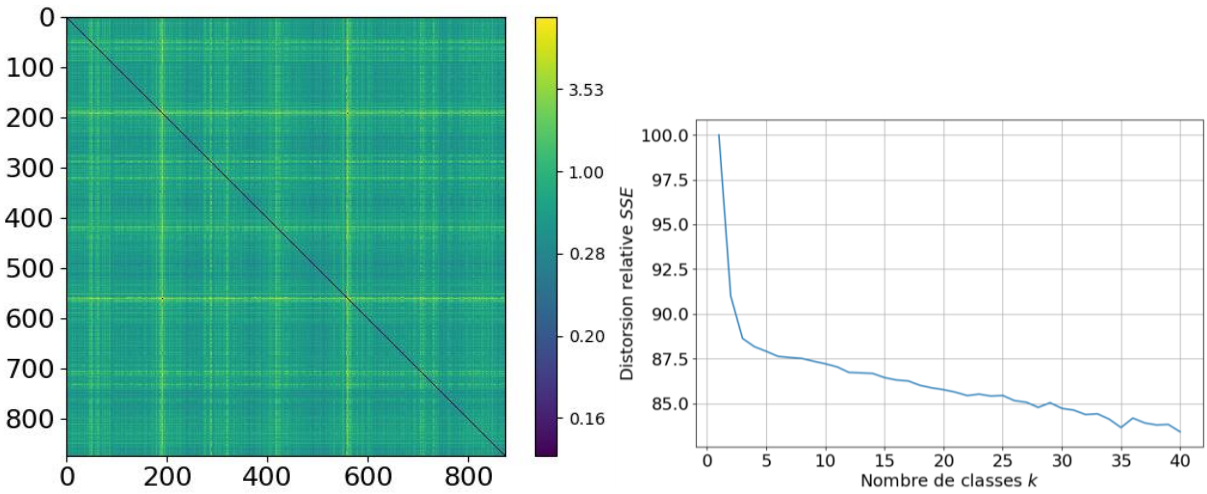


Figure 6.10 à gauche, la matrice des dissimilarités $D_{Ev}^{S}(873 \times 873)$ entre les séries temporelles des événements mesurées sur les 26 stations, à droite, la distorsion relative SSE en fonction du nombre de classes k choisi pour la classification des événements, Un coude apparait à k=3

Considérer l'information de toutes les stations fait que les événements de la classe «séries nulles» (représentée par l'événement 7) se sont différentiés les uns des autres. En effet, la présence des valeurs nulles présentes sur la matrice D_{Ev}^{S} se limite à la diagonale. Par conséquent le nombre optimal de classes est $k_{opt} = 3$ au lieu de $k = 4$ dans la partie précédente. De plus, cette considération accentue la démarcation de l'événement 191 sur la matrice de dissimilarités. Ceci est dû aux propriétés spatiales de ce type d'événements.

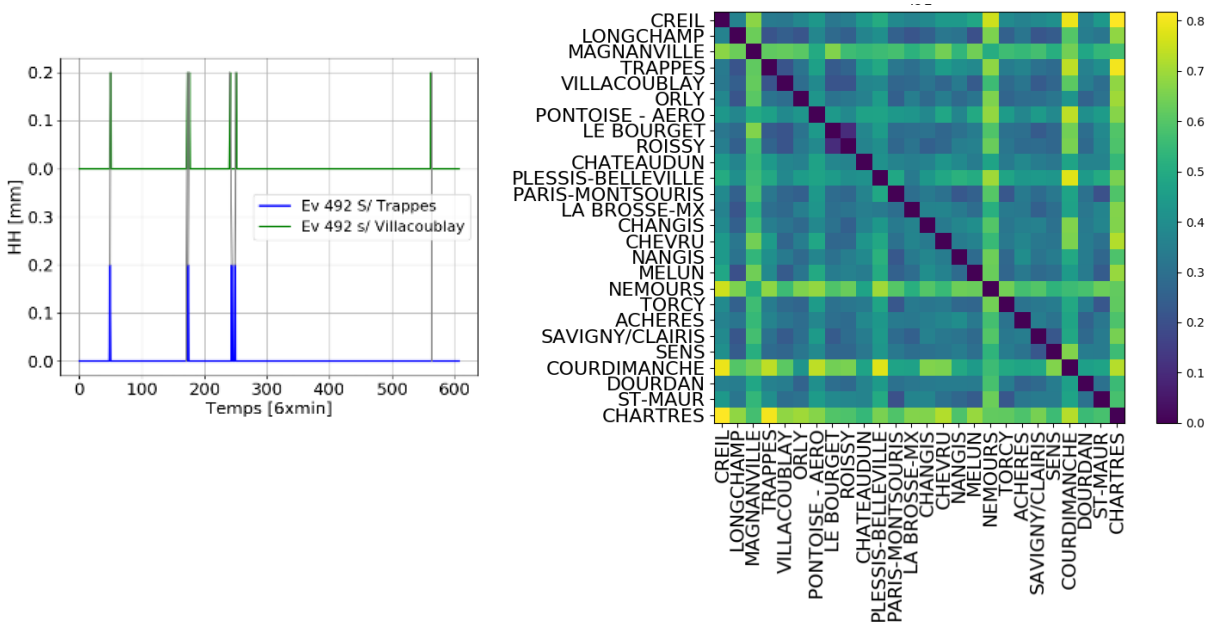
L'algorithme des k-médoïdes est utilisé en prenant la matrice des dissimilarités D_{Ev}^{S} . De la même façon que précédemment, on s'intéresse d'abord au représentant des 873 événements.

6.4.2. Événement représentant (26 séries temporelles) des 873 événements au sens de la mesure de dissimilarité IMs-DTW

Quel que soit l'initialisation de départ, le représentant retourné est toujours le 492ème événement ayant traversé l'ile de France entre 11 janvier 2013 à 09:36 et le 13 janvier 2013 à 22:24 et enregistré sur les 26 stations. Cet événement pour lequel nous allons analyser la matrice des dissimilarités et des alignements entre les 26 stations, correspond à la structure de précipitations qui domine en île de France. Il (événement 492) enregistre une trace sur toutes les stations et des variabilités relativement faibles ainsi que des cumuls d'eau modérés (entre 1mm et 8mm). De plus, à l'exception des stations (Magnanville, Chevreu, Melun, Nemours, Courdimanche, Sait-Maur et Chartres), les 19 stations restantes enregistrent au plus un basculement par pas de temps associé à un faible nombre de valeurs non nulles dans l'événement (entre 5 et 35). Cette description correspond à la définition d'un événement de

précipitations stratiforme en adéquation avec les résultats précédents et le caractère dominant de l'Ile-de-France.

Afin d'analyser la variabilité spatiale de l'événement 492, nous avons calculé la matrice des dissimilarités $D_{492}^{S}(26 \times 26)$ entre les 26 séries temporelles représentant l'événement 492 sur les différentes stations. La figure 6.11 de gauche présente deux exemples d'alignements proposés par l'IMsDTW pour deux couples de station tandis que la figure de droite présente la matrice des dissimilarités obtenu pour cet événement calculée sur l'ensemble des stations D_{492}^{S}.



La figure 6.11-a à gauche les alignements trouvés entre les séries temporelles représentant l'évènement 492 sur les stations Trappes (en bleu), Villacoublay(en vert sur la figure de haut) les alignements sont représentés en gris, à droite la matrice des dissimilarités $D_{492}^{S}(26 \times 26)$ entre les 26 séries temporelles représentant le passage de l'événement 492 sur les stations

Les alignements proposés par l'IMs-DTW semblent décrire une trajectoire de la cellule pluvieuse ; on voit sur l'alignement proposé sur le couple Trappes / Villacoublay (figure 6.11-a de haut), un décalage moyen de deux pas de temps ($\tau_{ATD} = 12\,min$). L'analyse des 325 alignements propose la direction sud-ouest/nord-est comme trajectoire, et la visualisation de la lame d'eau issue des radars de Météo France du 12 janvier 2013 corroborent cette trajectoire pour l'événement 492 (voir annexe 2.1). De plus, la trajectoire proposée pour ce dernier (représentant des événements) est en adéquation avec la direction du vent dominant dans la région d'Île-de-France.

On peut remarquer également que les valeurs des dissimilarités sur la matrice D_{492}^{S} varient entre 0 et 0.8 avec une moyenne de 0.4. Cette valeur correspond grosso modo à la variation de la dissimilarité lorsqu'on passe d'une station à une autre pour un événement donné. Toutefois, les stations de Magnanville, Courdimanche, Nemours et Chartres manifestent des dissimilarités importantes par rapport aux autres stations. Mise à part la démarcation de ces quatre stations les dissimilarités restent faibles et permettent d'avoir une idée de la valeur de référence de la variabilité en présence d'événements stratiformes. (cf. section 6.4.3.2).

6.4.3. Classification optimale ($k_{opt} = 3$)

La classification en trois classes ($k = 3$) fournit en plus de l'événement 492 décrit ci-dessus, les deux événements représentants 773 et 744.

	R_{max}	$mean(R_{\max})$	$std(R_{\max})$	$mean(R_{\mathrm{d}})$	$std(R_{\mathrm{d}})$	$mean(D_e)$	$std(D_e)$
492	8	3	1.77	2.81	2.07	450	112.05
773	8	4.84	1.37	6.62	2.62	498	98.59
744	12	6.96	2.34	22.11	7.33	1058	125.82

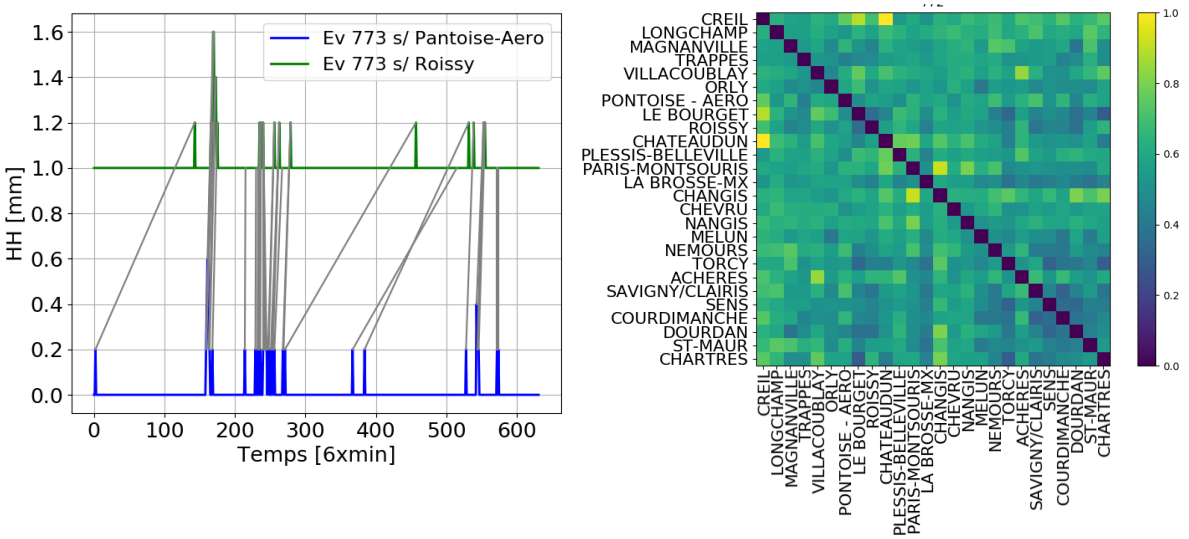
Tableau 6.2. caractéristiques des trois événements représentants

6.4.3.1. Description des événements

Analyse de l'événement 773 ayant traversé l'ile de France entre 30 janvier 2015 à 22:24 et le 02 février 2015 à 13:36

Cet événement enregistre des traces sur toutes les stations et des variabilités relativement fortes par rapport à l'évènement 492. Ainsi sur la station de Creil, l'événement enregistre une intensité maximale de 8 mm/h. Les cumuls d'eau enregistrés sur les stations sont modérés (entre 2mm et 12mm). De plus, cette fois toutes les stations enregistrent des intensités maximales équivalentes à plusieurs basculements par pas temps, mais des durées de support de pluie relativement plus grands (entre 10 et 60 pas de temps).

Afin d'analyser la variabilité spatiale de cet événement, nous avons calculé la matrice des dissimilarités $D_{773}^{S}(26 \times 26)$ entre les 26 séries temporelles représentant l'événement 773 sur les différentes stations. La figure 6.12 présente un exemple des alignements proposés entre les séries des stations de Roissy et de *Pontoise* ainsi que la matrice des dissimilarités D_{773}^{S}.



La figure 6.12-a à gauche l'alignement trouvé entre les séries temporelles représentant l'évènement 773 sur les stations Pontoise-Aero (en bleu), et Roissy (en vert), les alignements sont représentés en gris, à droite la matrice des dissimilarités $D^{S}_{773}(26 \times 26)$ entre les 26 séries temporelles représentant le passage de l'événement 773 sur les stations

Une fois de plus les alignements proposés par l'IMs-DTW semblent décrire une trajectoire de la cellule pluvieuse ; on voit sur l'alignement qui lie les stations de Pontoise-Aéro et Roissy (figure 6.12-a), un décalage moyen de huit pas de temps ($\tau_{ATD} = 48\,min$) et l'analyse des 325 alignements propose la direction nord-ouest/sud-est comme trajectoire. Ici aussi, la visualisation des cartes radars du 31 janvier 2015 corrobore l'hypothèse de trajectoire pour l'événement 773 (voir annexe 2.2).

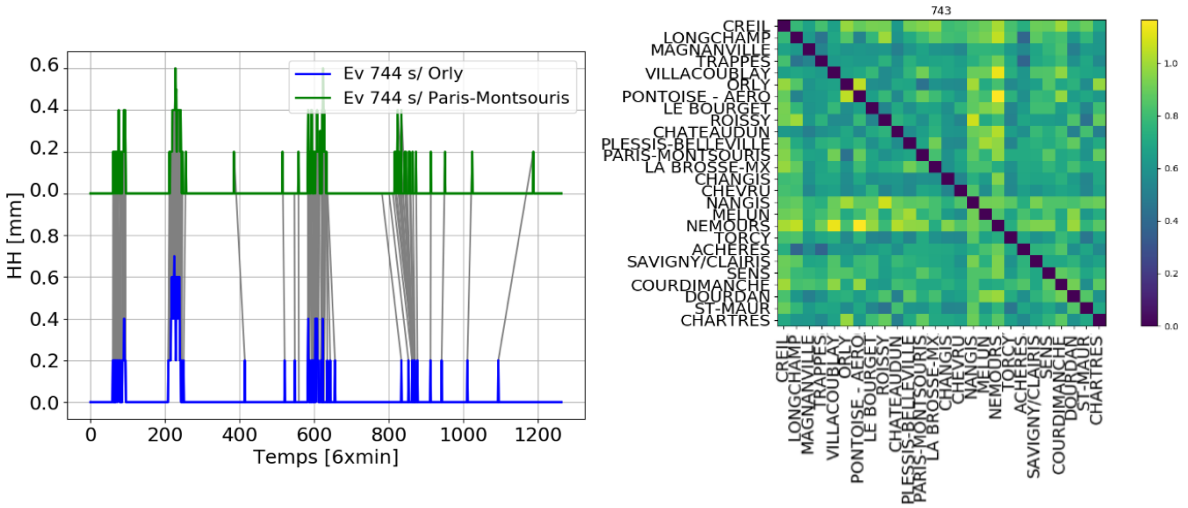
D'autre part, les valeurs des dissimilarités sur la matrice D^{S}_{773} varient entre 0 et 1 avec une moyenne de 0.55 plus forte que celle enregistrée pour l'événement 492. Cette augmentation s'explique par la forte variation de cet événement d'une station à l'autre par comparaison au 492ème (à variation lente). Cette description (trajectoire + variabilité relative) est en adéquation avec le type de cet évènement caractérisé par sa structure frontale. De plus, sur le coin bas à droite de la matrice des dissimilarités les valeurs semblent plus faibles que la moyenne et identifient une région spatiale homogène (cf. 6.4.3.2).

Analyse de l'événement 744 ayant traversé l'ile de France entre 14 novembre 2014 à 03:12 et le 19 novembre 2014 à 09:36

Cet événement enregistre lui aussi des intensités non nulles sur toutes les stations mais des variabilités très fortes par rapport aux deux évènements précédents. Cet événement se

différencie aussi des autres par ses cumuls d'eau importants (entre 5 mm et 40 mm) et ses durées de support de précipitations importants (entre et pas de temps).

De plus, cette fois deux groupes de stations enregistrent des intensités et des variations différentes. Ainsi, sur les stations de Nemours et Creil (resp. Nangis), l'événement enregistre des intensités maximales de 12 mm/h (resp. 10 mm/h) et des variations fortes. Cette description correspond à la définition d'un événement de précipitations convectif en adéquation avec une répartition spatiale localisée. Afin d'analyser la variabilité spatiale de l'événement 744 plus en détails, nous avons calculé la matrice des dissimilarités $D^S_{744}(26 \times 26)$ entre les 26 séries temporelles représentant l'événement 744 sur les différentes stations. La figure 6.13 présente un exemple des alignements proposés entre les séries ainsi que la matrice des dissimilarités D^S_{744}.



La figure 6.13-a à gauche l'alignement trouvé entre les séries temporelles représentant l'évènement 744 sur les stations Orly (en bleu) et Paris-Montsouris (en vert), les alignements sont représentés en gris, à droite la matrice des dissimilarités $D^S_{744}(26 \times 26)$ entre les 26 séries temporelles représentant le passage de l'événement 744 sur les stations

Les alignements proposés par l'IMs-DTW semblent adopter deux comportements différents ; un premier groupe présentant des retards réguliers sous forme de trajectoires et un autre groupe non réguliers. Cette différence de comportements pourrait décrire implicitement la localisation de l'événement en plus de la trajectoire de la cellule pluvieuse sur les stations principales (voir annexe 2.3). D'autre part, les valeurs des dissimilarités sur la matrice D^S_{744} varient entre 0 et 1.20 avec une moyenne de 0.70. Cet ordre de grandeur décrit le cycle de vie (la variation temporelle) d'un événement de précipitations de type « convectif ». De plus, les stations Creil, Nemours, Melun et Nangis manifestent de fortes dissimilarités avec le reste des

stations en adéquation avec l'observation des valeurs exceptionnelles des intensités (cf. partie suivante).

6.4.3.2. Variabilité spatiale des trois événements

L'analyse de la variabilité spatiale des événements sur l'ensemble des stations visait une caractérisation des différences entre les types d'événements. Les distorsions relatives calculées sur les matrices des dissimilarités quantifient les variabilités spatiales de chaque événement.

La figure 6.14 présente les distorsions relatives pour la classification des stations en fonction du nombre de classes utilisé k ($k = 1 \dots 26$) pour les trois événements 492, 773 et 744.

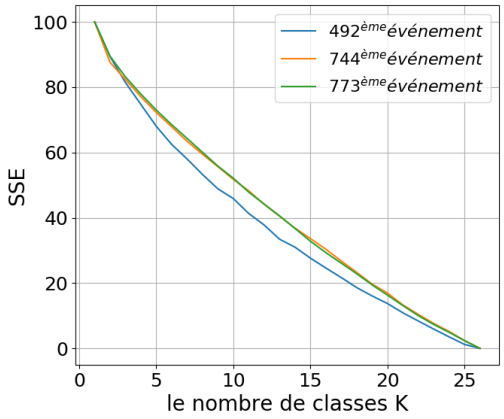


Figure 6.14 la distorsion relative SSE en fonction du nombre de classes k choisi pour la classification des stations pour les trois événements

Analyse comparative des distorsions

L'analyse des distorsions relatives montre que les deux événements 744 et 773 présentent les mêmes évolutions. Cette analyse suggère que ces derniers présentent les mêmes variabilités mais deux répartitions spatiales différentes.

D'autre part, l'événement 492 présente moins de variabilité « spatiale » comparé aux deux autres événements. D'une part les valeurs des dissimilarités (figure 6.11) sont plus faibles que celles des deux autres événements (figure 6.12 et 6.13), d'autre part considérer 10 classes donc 10 stations engendre une perte de 45 % de l'information sur la variabilité spatiale de cet événement (492) contre une perte de 54% "plus importante" pour les deux autres événements. Pour limiter la perte à un taux de 45 % pour le 744 et 773 il faut garder au moins 12 stations. Cette constations est en totale adéquation avec les descriptions des deux situations météorologiques stratiformes (précipitations à large échelle avec moins de variabilité spatiale) et convectives (précipitations localisées à fortes variabilité spatiale).

6.4.3.3. La classification optimale des événements de précipitations

Les différentes approches de classification employées dans les différents chapitres de ce manuscrit avaient pour but d'explorer/étudier la variabilité des événements de précipitations. Le tableau 6.3 reprend les statistiques de la première classification des événements (voir chapitre 2 page 42-43) et les compares aux statistiques de la dernière expérience présentée au-dessus (partie 6.4).

Configuration	Instrument : disdromètre Échelle : T=1min Approche : par caractéristiques Méthode : SOM Période : 2013-2014	Instruments : pluviomètres (26) Échelle : T=6min Approche : par formes Méthode : (K-médoïdes) Période : 2009-2015
Nombre optimal de classes (sans informations apriori)	3	3
Classification		
Classe 1 (stratiformes)	80%	71%
Classe 2 (intermédiaire)	8%	8%
Classe 3 (convectifs forts)	12%	20%

Tableau 6.3. Statistiques des classifications des événements selon différents points de vue.

On remarque que malgré les différences apparentes des deux jeux de données : instruments différents, dimensions (ponctuelle vs réseaux), échelles, approches, méthodes et périodes tous différents, le nombre optimal des classes proposé est toujours $k = 3$, avec une répartition stable sur les classes en adéquation avec les caractéristiques de la région d'Ile de France dominée par des précipitations stratiformes. Cette stabilité est vraie pour les différentes classifications évaluées dans ce manuscrit : les différentes expériences menées présentaient des statistiques similaires (voir partie 5.4.2) et de même pour la classification des 873 événements sur la station du Bourget qui présentait les mêmes statistiques (voir partie 6.3.2.2).

Hormis le fait de valider la méthodologie suivi dans ce travail, avec des configurations (instrument, échelle, approche, méthode, période étudiée et la dimension temporelle et/ou spatio-temporelle,…) qui changent à ce point, **on est amené à s'interroger sur la source/origine de cette stabilité des résultats (i.e. le nombre optimal de classes $= 3$, la répartition des événements ainsi que l'ordination des espaces).**

Si la répartition des événements trouve sa justification dans les caractéristiques de la région d'Ile de France dominée par des précipitations stratiformes, le nombre optimal de classes (k=3) remettrait en question la typologie rigide des événements (classification en stratiformes et

convectifs). Dans leurs deux rapports, D. Dunkerley[15] et B. Boudevillain[16] citent le fait que certains événements puissent présenter un mélange des deux processus physique et qu'une classification rigide en deux classes n'est pas très bien adaptée.

L'ordination des classes proposée par les différentes classifications (i.e. une classe intermédiaire positionnée entre les deux classes classiques stratiformes et convectives) corrobore cette hypothèse. En effet, cette classe pourrait bien regrouper les événements "mélanges des deux processus", les durées longues qui caractérisent les événements de cette classe offrent de bons supports pour l'évolution des événements au cours de leurs cycles de vie.

Une autre théorie peut être étudiée et est : l'impact de l'origine (point de naissance de la convection) et de la direction du vent sur le développement et le cycle de vie d'un événement de précipitations. Toutefois, la question reste ouverte et mérite de faire l'objet d'une étude en perspective.

[15] Rapport de révision (Dunkerley, 2016) de l'article présenté au chapitre 2
[16] Rapport sur la première version de ce document de thèse

6.5. Impact du changement climatique sur la variabilité/ évolution des précipitations

Le changement climatique est une réalité largement reconnue aujourd'hui aussi bien par la communauté scientifique (L'Agence Parisienne du Climat et Météo-France, 2016) ; que pour le grand public (Pech, 2019 ; Tabeaud, 2010). Il correspond à une modification durable (de la décennie au million d'années) des paramètres statistiques (paramètres moyens, variabilité) du climat global de la Terre ou de ses multiples climats régionaux. Ces changements peuvent être dus à des processus intrinsèques à la Terre, à des influences extérieures **ou, plus récemment, aux activités humaines**. Le changement climatique anthropique responsable du réchauffement climatique est le fait des émissions de gaz à effet de serre **engendrées par les activités humaines**, et qui modifient la composition de l'atmosphère de la planète. À cette évolution viennent s'ajouter les variations naturelles du climat. Du fait de la corrélation directe des émissions de gaz à effet de serre (notamment CO2) et de la température, l'impact du changement climatique sur cette dernière grandeur en est même devenu un indicateur important. **Néanmoins, l'impact du changement climatique sur les précipitations reste délicat à observer et à prévoir de par la nature intermittente et irrégulière de cette dernière[17].**

Dans ce contexte, l'objectif de cette partie est d'apporter des éléments de réponse aux questions suivantes: Les outils développés dans la présente étude peuvent–ils être utilisés pour analyser les changements dans les séries pluviométriques ? Permettent-ils de mettre en évidence des évolutions des précipitations extrêmes en Île-de-France?

6.5.1. Présentation des données

Les stations qui effectuent des mesures depuis au moins 100 ans sont rares. Elles apportent pourtant des informations inestimables sur l'histoire du climat régional récent. C'est pour cette raison que l'Organisation météorologique mondiale (OMM) vient de reconnaître officiellement trois stations centenaires en France, récompensées pour la qualité et la fiabilité de leurs données sur plus d'un siècle. La station de Paris-Montsouris en fait partie. Elle peut s'enorgueillir d'une série ininterrompue de mesures de températures, de pression atmosphérique, d'humidité et de précipitations depuis 1873.

[17] http://wikhydro.developpement-durable.gouv.fr/index.php/Changement_climatique_-_%C3%A9volution_des_pr%C3%A9cipitations

Pour répondre aux questions précédentes, on utilise les données de cette station historique « Paris-Montsouris » pour l'analyse. La figure 6.13 présente la série temporelle de précipitations couvrant les 142 années (1873-2015) au pas de temps journalier ($T = 1jour$).

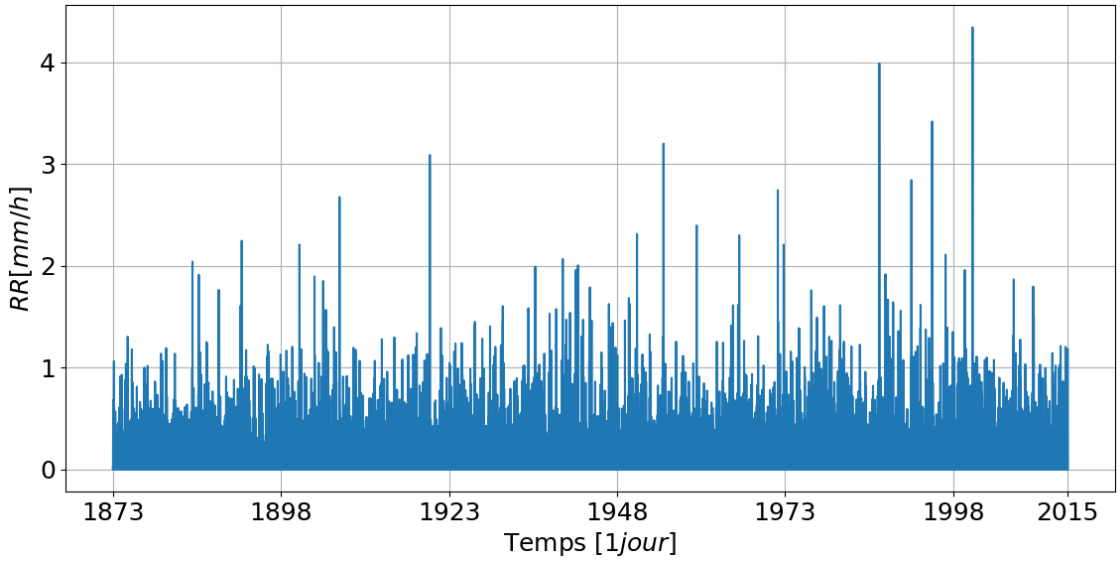


Figure 6.13 série temporelle mesurée par un pluviomètre sur la station Paris-Montsouris entre 1873 et 2015 à pas de temps T=1 jour

La visualisation de la figure 6.13 permet les constatations suivantes :

- Le record des cumuls journaliers est enregistré le 6 juillet 2001 ;
- D'autres valeurs exceptionnelles ($RR > 3mm.h^{-1}$) sont enregistrées le 15 juillet 1907, le 17 octobre 1920, le 19 juillet 1955, le 19 juillet 1972, le 19 juillet 1955, le 31 mai 1992 et le 02 juillet 1995…
- une diminution/absence des valeurs exceptionnelles ($RR < mm.h^{-1}$) sur l'intervalle des années 1923 et 1950.

6.5.2. Ce que l'on constate actuellement dans la littérature

La série « Paris-Montsouris » a fait l'objet de plusieurs études et publications (Mestre, 2000). Les climatologues de Météo-France ayant analysé les données ne voient *"… aucune évolution significative sur les inondations, les orages, les épisodes de grêle, les tornades ou les tempêtes…"* (l'Agence Parisienne du Climat et Météo-France, 2016). Moisselin et al. (2002) note la forte variabilité d'une année à l'autre, et la difficulté d'évaluer /décider « à l'œil nu » d'une tendance sur la série Paris-Montsouris. Cependant, la plupart des études réalisées à partir de la modélisation numérique du climat s'accordent sur une intensification des précipitations comme marqueur de l'évolution.

6.5.3. Analyse climatologique et comparaison des années

En climatologie, il est d'usage de comparer des années entre elles pour essayer de détecter une évolution. La définition civile de l'année rend difficile l'évaluation du contraste saisonnier à cause de la coupure de la saison hivernale. Ainsi des observations telles que *" Sur les 50 dernières années, les hivers semblent plus pluvieux alors que les étés sont un peu plus secs..."*[18] sont difficiles à évaluer. De ce fait, une définition qui couvre la comparaison des saisons semble plus appropriée. Dans ce travail, nous avons utilisé une définition hydrologique de l'année qui va du 1er septembre au 31 aout de l'année suivante, ainsi dans ce qui suit l'année X renvoi à l'année commençant le 1 septembre X. Cette définition donne lieu à 142 années que nous pourrons comparer entre elles avec l'IMS-DTW. La figure 6.14 présente un exemple des alignements proposés par la méthode entre les années 1890 et 1900.

La classification des années repose la matrice de dissimilarité constituée des 'dissimilarités entre deux années'. La 'dissimilarités entre deux années' est la dissimilarité obtenue en appliquant l' IMsDTW aux séries chronologiques d'un an à la résolution journalière observées par la station Paris-Montsouris deux années distinctes. Les résultats présentés dans cette section doivent être considérés comme préliminaires et doivent encore être consolidés par de nombreuses études complémentaires.

6.5.3.1. Evaluation des alignements

La structure d'alignement qui résulte de l'algorithme traduit une ressemblance de type motifs/forme qui ne peut être interprétée. La figure 6.14 illustre à titre d'exemple les alignements obtenus entre les années 1890 et 1900. La structure d'alignement assure l'association des pics maximaux et des périodes pluvieuses des deux années, on remarque que les associations de l'alignement présentent des décalages relativement faibles avec un maximum de deux semaines. Cette constatation se vérifie sur l'ensemble des alignements proposés par l'IMS-DTW entre les 142 années.

[18] http://wikhydro.developpement-durable.gouv.fr/index.php/Changement_climatique_-_%C3%A9volution_des_pr%C3%A9cipitations

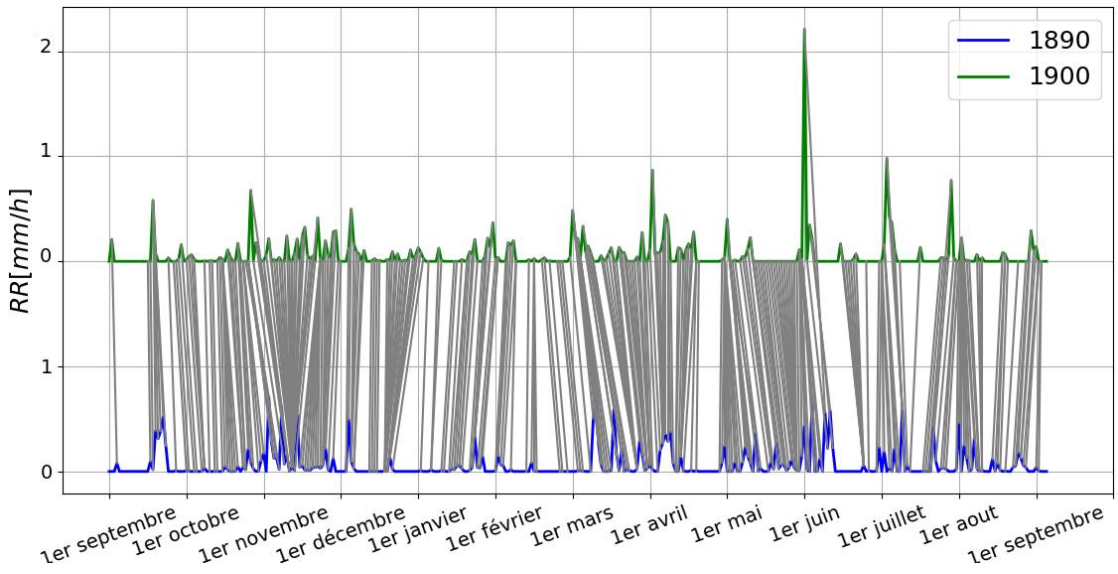


Figure 6.14 l'alignement qui lie les deux séries temporelles de précipitations représentants les deux années 1890 et 1900 mesurées à la station de paris Montsouris

Les décalages faibles enregistrés entre les années garantissent implicitement la comparaison saisonnière des années différentes (l'été comparé à l'été et l'hiver à l'hiver). Parfois, une saison peut être décalée de quelques semaines, et l'IMS-DTW permet donc non seulement de bien les recaler mais aussi d'évaluer précisément ce décalage.

6.5.3.2. Matrice des dissimilarités

Les résultats des dissimilarités entre deux années' sont stockés dans la matrice des dissimilarités $D_{Clim}(142 \times 142)$. La figure 6.15 la présente où d_{ij} représente la mesure de dissimilarité entre deux années i et j.

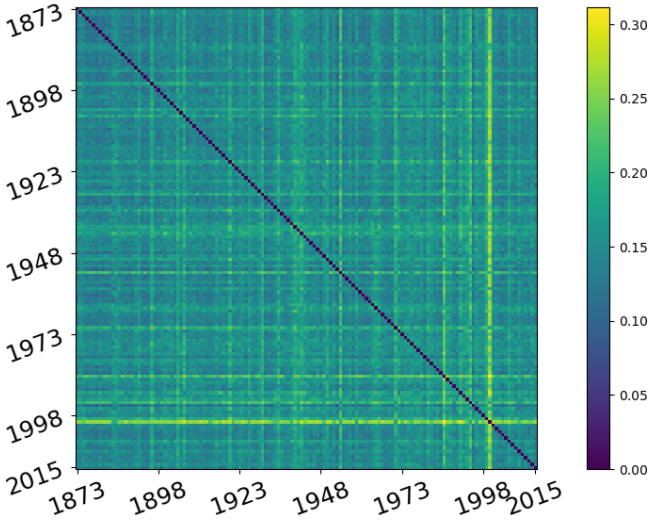


Figure 6.15 la matrice des dissimilarités D_{clim} (142×142) entre les 142 années observées par la station Paris-Montsouris

Comme cela était le cas pour la comparaison des stations, la présence des périodes non-pluvieuses fait diminuer l'ordre de grandeur des dissimilarités. On peut citer parmi les années qui se distinguent:

- L'année 1906 (#34) connue dans l'histoire parisienne pour la crue de la seine et répertoriée année exceptionnelle, présente des dissimilarités importantes avec les autres des années.
- On note aussi une bande qui marque les quatre années correspondant à la période du 1er septembre 1940 au 31 aout 1944 (#68 à 71) où les stations étaient sous l'occupation allemande. Cette démarcation est soulignée dans la présentation des données et pourrait s'expliquer par un changement de méthodologie et ou de technique de mesure.
- L'année 1954 (#82), répertoriée aussi année exceptionnelle pour la crue de la seine, les record enregistrés à l'époque présentent des dissimilarités avec les autres années.
- L'année 2000 (#128) présente de fortes dissimilarités (les maximums) par rapport aux autres années de l'étude.

En résumé, toutes les observations/remarques de la page 4 du rapport "le changement climatique à Paris" publié par Météo-France (Moisselin et al. 2002) sont vérifiées et visibles sur la matrice des dissimilarités D_{clim}.

De plus, les valeurs des dissimilarités semblent s'accentuer au fil des années, (un gradient visible de gauche à droite / de haut en bas). La classification des années permet d'approfondir l'analyse.

6.5.4. Classification des années

L'algorithme des k-médoïdes est utilisé en utilisant la matrice des dissimilarités D_{clim}. De la même façon que précédemment, on s'intéresse d'abord au représentant de ces 142 années.

6.5.4.1. Année représentante de toutes les années ($k = 1$)

Le représentant retourné est le 123ème élément correspondant à l'année du 1er septembre 1995 au 31 aout 1996 discutée dans la partie suivante (cf. 6.5.4.4). La figure 6.17.a présente la série temporelle du taux précipitant mesuré à Paris-Montsouris cette année-là.

6.5.4.2. Classification en plusieurs classes

De même, la figure 6.16 présente l'évolution de la distorsion relative SSE en fonction du nombre de classes k choisi pour la classification des années.

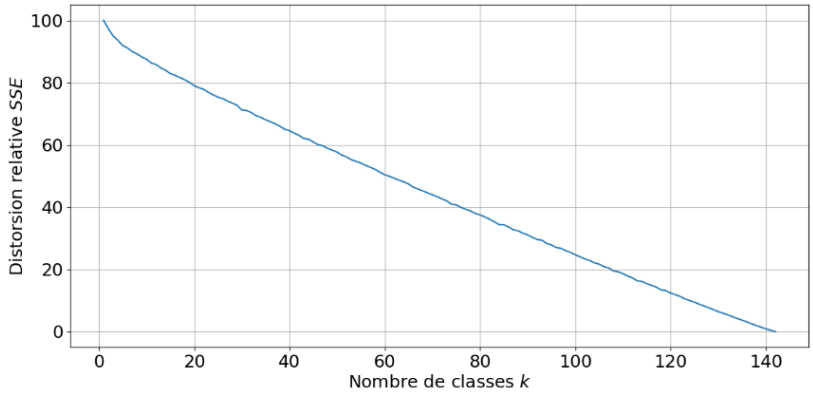


Figure 6.16 la distorsion relative SSE en fonction du nombre de classes k choisi pour la classification de des années enregistrées sur la station Montsouris.

En absence d'une cassure (un coude) significative sur la courbe des distorsions relatives, cette forme laisse à penser à **une structure de continuum**. Toutefois, la présence de d'une légère déformation à k=4 nous a encouragé à prendre ce nombre de classes à des fins d'échantillonnage (bien qu'on ne s'attende pas vraiment à une séparation claire des classes).

6.5.4.3. Classification en quatre classes (k=4)

La classification en quatre classes ($k = 4$) fournit les années 1884 (#12), 1931(#59), 1953(#81) et 1995(#123) comme représentants des types distincts de pluviométrie annuelle en île de France. L'année 1995 est également la représentante de toutes les années.

6.5.4.4. Caractérisation des représentants

Afin de caractériser les quatre représentants et les classes associées, nous avons choisi les cinq caractéristiques[19] : intensité maximale ($\boldsymbol{R_{max}}$), l'écart-type des intensités ($\boldsymbol{\sigma_R}$), le cumul d'eau ($\boldsymbol{R_d}$), le supports de précipitations ($\boldsymbol{S_R}$) , le Nombre d'épisode ainsi que la position de l'intensité maximale. A l'échelle journalière on considère comme un « épisode de pluie » une série de jours de pluie consécutifs,

	$R_{max}\ [mm.h^{-1}]$	$\sigma_R[mm.h^{-1}]$	$R_d[mm]$	$S_R[jour]$	$pic\ max$	Nb Event
1884	2.04	0.22	459.90	142	28 juin	56
1931	1.04	0.16	457.40	149	23 oct.	60
1953	0.53	0.13	382.90	146	24 aout	53
1995	0.66	0.14	378.10	134	04 juil.	62

Tableau 6.4. Variables descriptives des 4 années représentant des types de pluviométrie annuelle

[19] A noter que le choix des caractéristiques descriptives n'affecte pas le résultat et que d'autres caractérisations peuvent aussi être utilisées.

L'année 1884 (représentant de la classe 1): l'année qui commence au 1 septembre 1884 et finit le 31 aout 1985 année exceptionnelle intensité maximale fortes, cumul annuel fort, un écart-type journalier particulièrement élevée, pour un nombre total d'épisodes moyen. C'est une année fortement pluvieuse composée d'épisodes pluvieux en nombre et de durée standard mais avec un cumul moyen par épisode particulièrement important. Cette année se caractérise également par son hétérogénéité avec un épisode particulièrement intense en été et un épisode de sécheresse de 23 jours consécutifs.

L'année 1931 (représentant de la classe 2): l'année qui commence au 1 septembre 1931 et finit le 31 aout 1932 présente un cumul annuel pratiquement identique à l'année précédente pour une valeur maximale annuelle et un écart-type journalier moindres mais un nombre de jours de pluie et d'épisodes de pluie plus important. C'est une année fortement pluvieuse mais avec une répartition temporelle plus homogène en un nombre plus important d'épisodes. Les épisodes sont de même durée en moyenne que pour l'année précédente mais leur cumul moyen de pluie est assez nettement plus faible que pour l'année précédente car le cumul d'eau annuel est réparti sur un nombre plus important d'épisodes.

L'année 1953 (représentant de la classe 3): l'année qui commence au 1 septembre 1953 et finit le 31 aout 1954 **est une année sèche,** « … le régime des précipitations est resté déficitaire pour toutes les régions françaises ; … l'écart pluviométrique avec la normale 1900-1930 est de 197 mm à Paris… » (Chartier, 1954). L'année 1953 présente donc comme on peut s'y attendre un cumul, une valeur maximale annuelle, un écart-type journalier et un nombre d'épisodes nettement plus faible que les deux années précédentes. Le cumul moyen par épisode n'est cependant pas beaucoup plus faible que celui de l'année précédente car l'année 1953 se caractérise par un nombre relativement faible d'épisodes particulièrement longs.

L'année 1995 (représentant de la classe 4): Comme l'année 1953 , l'année 1995 est une année sèche, avec un cumul annuel et un nombre de jour de pluie particulièrement faible. Le nombre d'épisodes de pluie est par contre plus élevé que pour les 3 années précédentes. L'année 1995 est constituée d'un nombre important d'épisodes particulièrement courts et dont le cumul d'eau moyen par épisode est particulièrement faible. Elle présente également un épisode de sécheresse particulièrement long (28 jours)

Les deux années pluvieuses (resp. sèches) se distinguent entre elles par leurs structures temporelles, avec une intermittence très différente. Après avoir décrit les représentants, on s'intéresse aux classes représentées par ces derniers.

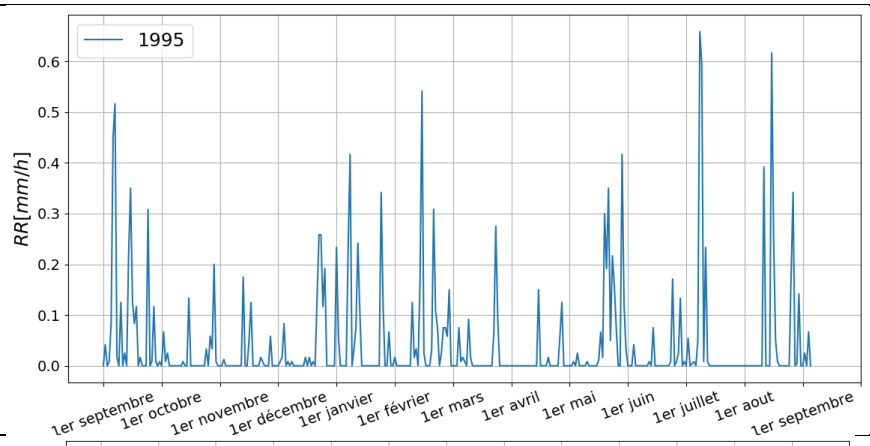
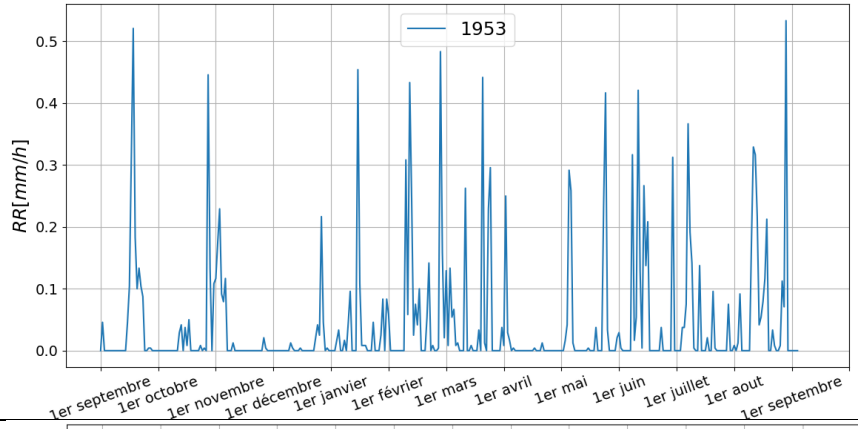
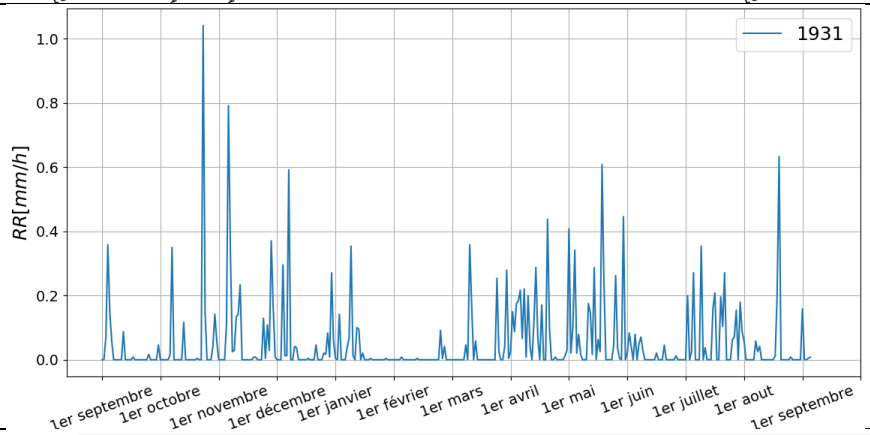
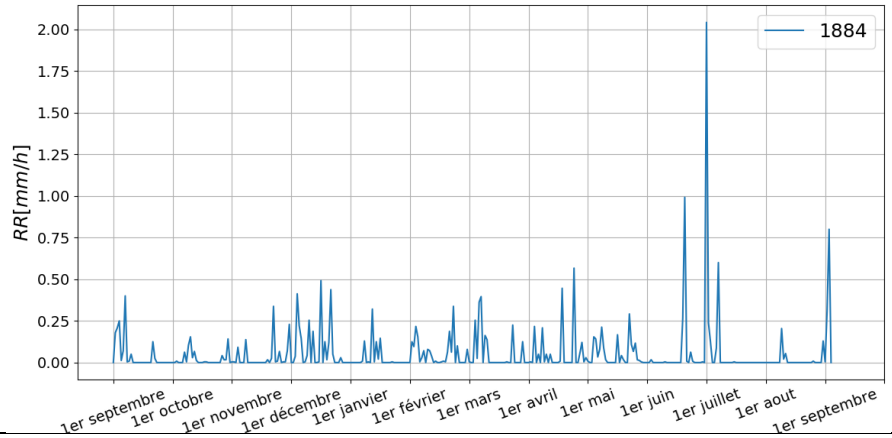
(a) *la 123ème année allant du 1er septembre 1995 au 31 aout 1996* représentante Classe 2	
(b) *La 81ème année allant du 1er septembre 1953 au 31 aout 1954* représentante Classe 1	
(c) *la 59ème année allant du 1er septembre 1931 au 31 aout 1932* représentante Classe 3	
(d) *la 12ème année allant du 1er septembre 1884 au 31 aout 1885* représentante Classe 4	

Figure 6.17 Séries temporelles des différents représentants des quatre classes mesurées à l'aide d'un pluviomètre v=0.2mm agrégée à T=1 jour

6.5.4.5. Caractérisation et labélisation des classes

La figure 6.18 présente par paires les caractéristiques des 142 années colorées en fonction de leurs classes d'appartenance. Les numéros des classes sont justifiés ultérieurement

- **La Classe 4 (rouge) :** représentée par l'année 1884 (#12). Les années appartenant à cette classe sont caractérisées par des intensités maximales $\boldsymbol{R_{max}}$ fortes, enregistrées dans 100% des cas durant l'été comme le montrent les figures c et d. Elles sont aussi caractérisées par les plus importantes variances $(\boldsymbol{\sigma_R})^2$ de l'ensemble avec des supports de précipitations $\boldsymbol{S_R}$ et des cumuls d'eau $\boldsymbol{R_d}$ dans la moyenne. Cette classe dénombre les 24 années suivantes: 1884, 1888, 1900, 1902, **1906,** 1914, 1934, 1935, 1939, 1941, 1950, 1952, **1954,** 1963, **1971,** 1972, 1976, 1978, **1991, 1994, 1996, 1999, 2000** et 2009 où les années en rouge sont répertoriées exceptionnellement intenses dans les bulletins d'histoire en adéquation avec la labélisation proposée. Il s'agit d'années pluvieuses avec un événement particulièrement extrême en été

- **La classe 3 (verte) :** représentée par l'année 1931 (#59). Les années appartenant à cette classe sont caractérisées par des cumuls d'eau $\boldsymbol{R_d}$ relativement forts, des supports de précipitations $\boldsymbol{S_R}$ relativement longs **(voir figure g),** des intensités maximales $\boldsymbol{R_{max}}$ et des variances $(\boldsymbol{\sigma_R})^2$ dans la moyenne **(voir figure b).** Cette caractérisation correspond au label « années pluvieuses ». Cette classe comprend les 37 années suivantes: 1874, 1877, 1878, 1880, **1882**, 1892, 1903, 1907, **1909,** 1911, 1912, 1913, 1919, **1923,** 1926, 1927, 1929, 1930, 1931, 1932, 1938, 1940, 1942, 1944, 1947, 1949, 1957, 1959, 1977, 1979, **1981,** 1992, 1998, 2011, 2012, 2013 et 2014 où les années en vert sont répertoriées dans les bulletins d'histoire comme des années pluvieuses en adéquation avec la labélisation proposée. Il s'agit d'années pluvieuses avec une répartition annuelle plus homogène que la classe 1.

- **La classe 1 (bleue) :** représentée par 1953 l'année (#81). Les années appartenant à cette classe sont caractérisées par des intensités maximales $\boldsymbol{R_{max}}$, des variances $(\boldsymbol{\sigma_R})^2$ et des cumuls d'eau $\boldsymbol{R_d}$ parmi les plus faibles de l'ensemble des années ainsi que des supports de précipitations $\boldsymbol{S_R}$ relativement faibles. L'ensemble de ces caractéristiques correspondent au label « années sèches ». Cette classe est composée des 34 années suivantes: 1873, 1881, **1883,** 1885, 1887, 1890, 1891, 1896, 1897, 1898, 1901, 1904, 1905, 1917, **1920, 1921**, 1924, 1933, 1936, 1945, **1946,** 1948, 1953, 1962, 1967, 1969, 1970, 1974, **1975**, 1980, 1983, **2003,** 2005 et 2008. Les années colorées en bleu sont

répertoriées sèches dans les bulletins d'histoire et sont en adéquation avec cette labélisation.

- **La classe 2 (orange) :** représentée par l'année 1995 (#123). Les années de cette classe sont caractérisées par des cumuls d'eau $\boldsymbol{R_d}$ et des supports de précipitations $\boldsymbol{S_R}$ relativement faibles mais des intensités maximales $\boldsymbol{R_{max}}$ et des variances $(\boldsymbol{\sigma_R})^2$ dans la moyenne **(voir figure b).** Cette caractérisation correspond au label « années modérées »**.** Cette classe dénombre les 47 années suivantes: 1875, 1876, 1879, 1886, 1889, 1893, 1894, 1895, 1899, 1908, 1910, 1915, 1916, 1918, 1922, 1925, 1928, 1937, 1943, 1951, 1955, 1956, 1958, 1960, 1961, 1964, 1965, 1966, 1968, 1973, 1982, 1984, **1985**, 1986, 1987, 1988, **1989**, **1990**, 1993, **1995**, 1997, 2001, 2002, 2004, 2006, 2007 et **2010.** Les années colorées en bleu sont répertoriées relativement sèches dans les bulletins météorologiques.

  L'analyse de la structure en épisode et en saison de toutes les années doit être réalisée pour déterminer quelles sont les caractéristiques des représentantes qui sont communes à l'ensemble des années d'une même classe.

6.5.4.6. Ordonnancement des classes

D'après les premières analyses présentées figure 6.18, les années ne sont pas réparties en classes distinctes mais il s'agit d'un continum. On peut en effet signaler des similitudes de structure temporelle entre l'année 1931 pluvieuse et l'année 1995 sèche .Ces deux années ont en effet un cumul moyen par jour de pluie et un nombre d'épisode de pluie similaire. On peut établir l'ordre suivant :

$$1953\ (classe\ 1 : sèches\ et\ hétérogène) \rightarrow 1995(classe\ 2 : sèches\ et\ homogène) \rightarrow$$
$$1931(classe\ 3 : pluvieuses\ et\ homogène) \rightarrow 1884(classe\ 4 : pluvieuses\ et\ hétérogène).$$

Les valeurs des dissimilarités entre les représentants des classes corroborent cet ordre proposé.

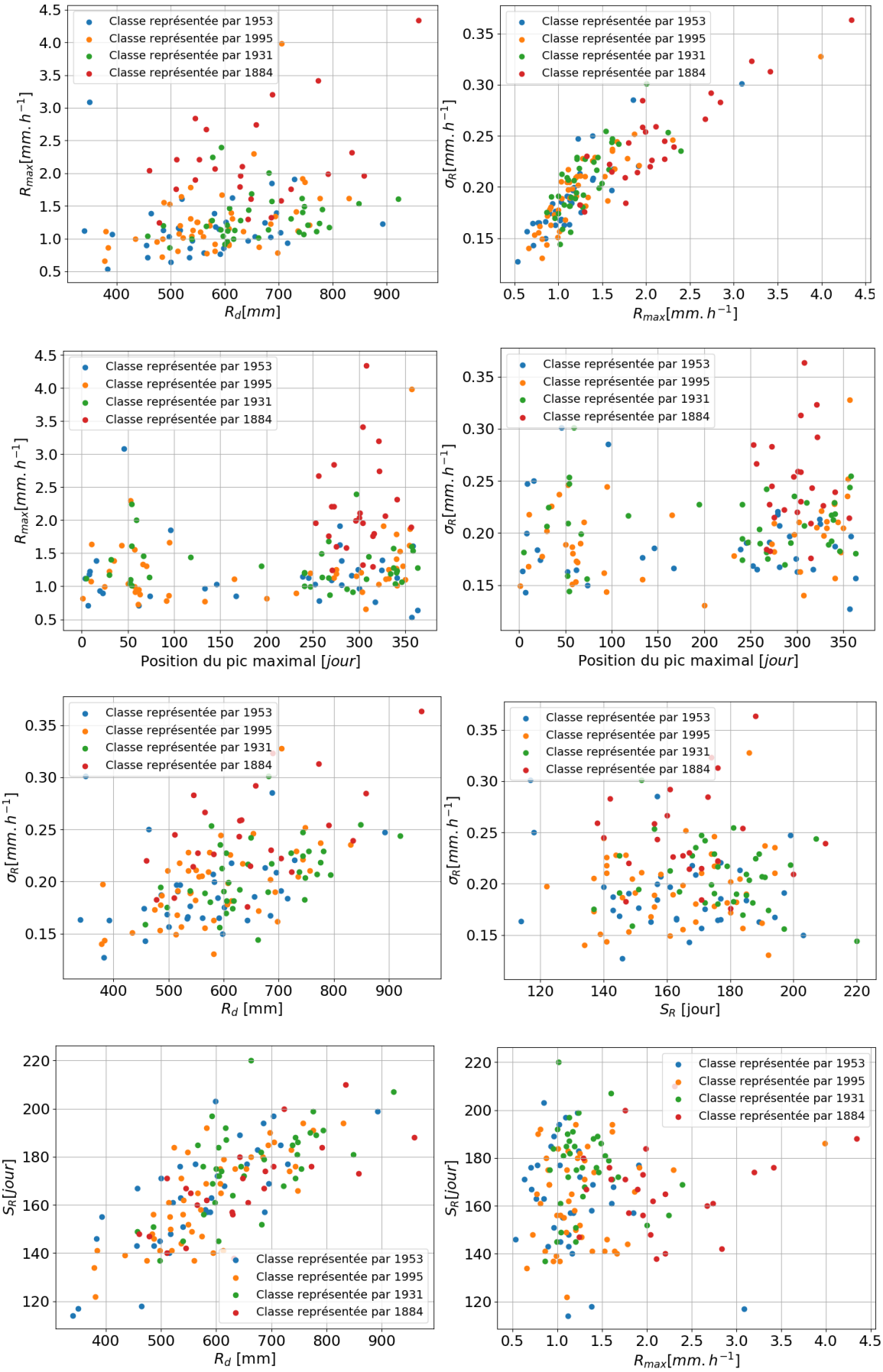


Figure 6.18 représentation par paires des caractéristiques des années, les années sont colorées en fonction de la classe d'appartenance

6.5.5. Évolution des précipitations et changement climatique

La série longue à notre disposition nous permet d'étudier l'évolution éventuelle des précipitations sur toute la période en considérant la fréquence de chaque classe dans 7 intervalles temporels successifs. La fenêtre temporelle est de 27 ans avec un chevauchement de 7 ans. Les Tableaux ci-dessous 6.5, 6.6 et 6.7 résument les résultats obtenus.

		Classe			
Fenêtre temporelle		C1 Sèche hétérogène	C2 Sèche homogène	C3 Pluvieuse homogène	C4 Pluvieuse Hétérogène
1873	1899	10	9	6	2
1892	1918	7	9	7	4
1911	1937	6	7	11	3
1930	1956	6	5	9	7
1949	1975	7	11	3	6
1968	1994	6	11	4	6
1987	2014	3	13	6	6

Tableau 6.5. effectifs par classe et par fenêtre

		Classe			
Fenêtre temporelle		C1	C2	C3	C4
1873	1899	0.38	0.35	0.23	0.07
1892	1918	0.27	0.35	0.27	0.15
1911	1937	0.23	0.27	0.42	0.11
1930	1956	0.23	0.19	0.35	0.27
1949	1975	0.27	0.42	0.11	0.23
1968	1994	0.23	0.42	0.15	0.23
1987	2014	0.11	0.48	0.22	0.22

Tableau 6.6. fréquence relative en % de chaque classe par fenêtre temporelle (somme par ligne)

		Classe			
Fenêtre temporelle		C1	C2	C3	C4
1873	1899	0.29	0.19	0.16	0.08
1892	1918	0.21	0.19	0.19	0.17
1911	1937	0.18	0.15	0.29	0.12
1930	1956	0.18	0.11	0.24	0.29
1949	1975	0.21	0.23	0.81	0.25
1968	1994	0.18	0.23	0.11	0.25
1987	2014	0.09	0.28	0.16	0.25

Tableau 6.7. fréquence relative en % de chaque fenêtre par classe (somme sur colonne))

Analyse de la tendance : avec une fenêtre temporelle de taille grossière on voit que la présence de la classe 1 (classe des années sèches hétérogènes) décroit au fils du temps alors que celle de

la classe 4 (classe des années pluvieuses hétérogènes) augmente ou stagne dans le même temps. Les classes intermédiaires (2 et 3) ont également une évolution inverse. Pour la classe 2 une diminution jusque vers 1950/1960 puis une augmentation alors que l'effectif de la classe 3 a l'évolution inverse.

6.6. Conclusion

L'IMs-DTW présentée au chapitre 4 associée à une méthode de classification présentée au chapitre 5 a été développée pour l'analyse et la classification de séries pluviométrique. Le travail informatique réalisé a rendu l'algorithme opérationnel et permet son application à un jeu de données réelles relativement important constitué des observations réalisées par 26 stations pendant une période de 7ans à fine résolution temporelle ainsi qu'une série journalière de 140 ans. Plusieurs expériences sont présentées pour déterminer les possibilités et les limites de la méthode proposée.

La classification des stations à partir des séries chronologiques complète ne permet de distinguer aucune station ou groupe de station. Globalement, les stations proches font des observations similaires et les stations éloignées ont des dissimilarités plus importantes. Ce résultat laisse penser qu'une normalisation adéquate par la distance inter-station peut permettre d'utiliser la matrice de dissimilarité entre stations pour détecter de manière automatique une station ayant des observations 'anormales' par rapport au reste du réseau.

La classification des évènements observés par la station Le Bourget permet de retrouver la classification similaire à celle d'un DBS à une résolution plus fine. Cette étude confirme les résultats du chapitre 5 sur l'intérêt de l'IMsDTW qui évite le calcul de caractéristques sensibles à la discrétisation par l'auget. La méthode a pour objectif de rendre (dans une certaine mesure) le résultat de classification indépendant du volume d'auget et du temps d'intégration.

L'efficacité opérationnelle de l'algorithme développé permet le calcul rapide des dissimilarités « spatiales » entre les 873 évènements, chaque évènement étant décrit par 26 séries chronologiques. La classification permet de détecter les structures dominantes pour les précipitations en île de France. Les représentants des classes ont été analysés et les conclusions confrontées à celles des observations radar présentées en annexe. Un travail complémentaire reste à réaliser pour déterminer dans quelle mesure les structures spatiales des représentants décrites sont communes à l'ensemble des évènements d'une même classe.

L'IMsDTW permet de déterminer les dissimilarités interannuelles à partir des observations de la série journalière de 142 ans de la station « Paris-Monsouris ». L'approche permet une classification des années sans a priori sur les caractéristiques retenues et en prenant ainsi en considération l'ensemble de la structure annuelle de précipitation. Une nette évolution temporelle des fréquences de certaines classes est observée. Pour confirmer et comprendre ce résultat, un travail complémentaire reste à réaliser pour déterminer dans quelle mesure les structures temporelles des représentants décrites sont communes à l'ensemble des années d'une même classe.

Conclusion

L'intérêt porté à la variabilité des précipitations à des échelles infra-journalières voire infra-horaires conduit à s'interroger sur l'utilisation des réseaux pluviométriques existants. Toutes les régions du globe ne disposant pas de radar météorologiques, les séries chronologiques mesurées par des pluviomètres à auget sont les seules qui sont suffisamment longues pour étudier la variabilité temporelle et suffisamment nombreuses pour étudier la variabilité spatiale à l'échelle d'une région.

En raison du caractère intermittent des précipitations, l'approche retenue dans cette thèse repose sur une analyse par événement des précipitations. Une première étude réalisée à partir des observations d'un disdromètre à la résolution d'une minute a permis de réaliser une première typologie des évènements de pluie en Île-de-France. La mise en évidence de relations entre les propriétés microphysiques des évènements et la classification réalisée à partir de caractéristiques macro-physiques confirme **la pertinence de l'approche par événement**. La description par caractéristiques, utilisée dans la partie deux, pose des questions lors de son application à des observations réalisées par des pluviomètres à auget.

La mesure pluviométrique du réseau de pluviomètres de Météo France est basée sur une technique qui induit une erreur de discrétisation plus ou moins importante suivant le volume d'auget des pluviomètres utilisés et qui doit être prise en compte pour des analyses à fine résolution. Nous avons montré que les effets du temps d'agrégation T et du volume des augets v sont intimement liés.

A partir de données pseudo-pluviométriques qui reproduisent les observations d'un pluviomètre pour différents temps d'agrégation et différents volumes d'auget nous avons étudié la **quantification des effets du temps d'agrégation T et du volume des augets v** pour la région d'Île de France. Cette étude a permis :

- de définir une relation empirique entre le volume d'auget maximal et un temps d'agrégation donné (resp. le temps d'agrégation minimal pour un volume d'auget donné)
- de quantifier la perte d'information sur la variance due à l'agrégation temporelle (le pas de temps journalier ne conserve donc que 2,8% de la variabilité à une minute)
- de quantifier la perte d'information sur l'occurrence due à la discrétisation par l'auget (un volume 0.2 mm ne permettent de conserver que 9% de l'information relative à l'occurrence des précipitations).

Le temps d'agrégation et le volume d'auget ont un impact sur la caractérisation des séries chronologiques et des évènements et ont permis de mettre en évidence :

- une grande disparité de sensibilité suivant la caractéristique considérée : l'écart type de l'intensité de pluie est par exemple particulièrement sensible au volume d'auget pour les temps d'agrégation infra-horaire.
- une grande disparité suivant le type d'évènements considéré : la caractérisation des évènements intenses ou convectifs (resp de faible intensité ou stratiformes) est très sensible au temps d'agrégation (resp. à la discrétisation par le volume d'auget) et peu sensible à la discrétisation par le volume d'auget (resp. au temps d'agrégation)

Nous avons donc cherché à nous affranchir de la description par caractéristiques en proposant une mesure de dissimilarité entre séries, basée sur la déformation temporelle dynamique (DTW). L'objectif est d'une part, de palier à la sensibilité des caractéristiques à l'instrument de mesure et au temps d'agrégation, et d'autre part, de proposer une méthode d'analyse des évènements de pluie sans a priori sur le choix de caractéristiques pertinentes.

Nous avons adapté une **mesure multi-échelles (IMs-DTW) de dissimilarité** aux spécificités des séries temporelles de précipitations (intermittence et multifractalité). Nous avons dans un premier temps réalisé un travail informatique pour rendre l'algorithme opérationnel, en réduisant considérablement le temps de calcul. Nous avons montré que, dans le cas d'un même épisode pluvieux observé par plusieurs pluviomètres, non seulement la dissimilarité entre les séries chronologiques mesurées nous informe sur l'écart entre les observations mais également **la structure d'alignement obtenue par l' IMs-DTW apporte une information pertinente** sur la dynamique de l'épisode considéré La structure d'alignement permet de déterminer si deux évènements observés par deux pluviomètres distants correspondent ou non au même processus pluvieux.

A partir des données pseudo-pluviométriques déjà utilisées nous avons vérifié la stabilité de IMs-DTW. Un algorithme de k-médoïdes basé sur la matrice de dissimilarité, calculée par l'IMs-DTW, a permis de réaliser une partition des évènements et de vérifier la **stabilité de la partition obtenue par rapport au temps d'agrégation T et du volume des augets v**. Un travail préliminaire sur l'alignement nous a permis de mettre en évidence des liens entre les alignements et la structure temporelle des évènements. L'approche sans caractéristique pose cependant une difficulté d'interprétation des représentants des classes et des clusters obtenus.

La méthode de classification développée a été appliquée au jeu de données réelles pour l'analyse de la pluviométrie en île de France. Plusieurs applications ont été envisagées, des

travaux complémentaires, qui nécessitent de disposer de données supplémentaires (données radar en île de France, classification de la circulation atmosphérique à l'échelle synoptique, séries chronologiques sur une région plus grande, …), sont nécessaires pour valider et interpréter les résultats obtenus concernant:

- La classification des stations à partir des 26 séries chronologiques complètes de 7 ans ;
- la classification de 873 évènements observés par une station pluviométrique (Le Bourget) ;
- la classification de 873 évènements observés par le réseau de 26 pluviomètres ;
- la classification des 142 séries annuelles de précipitation observées à la station de Paris-Montsouris entre 1873 et 2015.

Les perspectives des travaux menés au court de cette thèse sont nombreuses :

- la mise en ligne des algorithmes présentés permettra la mise à disposition de la communauté scientifique ;
- la consolidation des résultats de classification obtenus sur les données île de France pourra donner lieu à une publication ;
- une étude sur la normalisation de la dissimilarité par la distance géographique entre stations pluviométriques pourrait permettre d'utiliser l'algorithme pour détecter automatiquement des anomalies dans les observations de réseaux de pluviomètres ;
- une étude pour définir un descripteur de la structure d'alignement pertinent permettra de réaliser des analyses spatiotemporelles des évènements qui prennent mieux en compte la structure spatiale du réseau de pluviomètres.

Une perspective à plus long terme est la généralisation de la mesure de dissimilarité des séries chronologiques de précipitations (structure 1D) aux images de télédétection radar ou spatiale des précipitations (structure 2D).

Bibliographie

Aghabozorgi, S., Shirkhorshidi, A.S. and Wah, T.H. (2015). Time-series clustering – A decade review, Information Systems, 53, pp16-38,

Akrour, N., Chazottes, A., Verrier, S., Mallet, C., and Barthes, L. (2015). Simulation of yearly rainfall time series at microscale resolution with actual properties: Intermittency, scale invariance, and rainfall distribution. Water Resources Research, 51(9), 7417–7435.

Alghamdi H. M. et Selamat A., (2019). Arabic Web page clustering: A review. Journal of King Saud University - Computer and Information Sciences. 31 (1), 1-14

Aronov, B., Har-Peled, S., Knauer, C., Wang, Y. and Wenk, C. (2006). "Fréchet distance for curves, revisited," in ESA'06, London, UK, pp. 52–63, Springer-Verlag.

Atlas, D., Ulbrich, C.W., Marks, F. D., Amitai, E., and Williams, C. R. (1999). Systematic variation of drop size and radar-rainfall relations, J. Geophys. Res.-Atmos., 104, 6155–6169.

Balme, M., Vischel, T., Lebel, T., Peugeot, C., and Galle, S. (2006). Assessing the water balance in the Sahel: impact of small scale rainfall variability on runoff Part 1: rainfall variability analysis. Journal of Hydrology 331:336–348.

Barbosa, L. R., Almeida, C. D. N., Coelho, V. H. R., Freitas, E. D. S., Galvão, C. D. O. and Araüjo, J. C. D. (2018). Sub-hourly rainfall patterns by hyetograph type under distinct climate conditions in Northeast of Brazil: a comparative inference of their key properties. RBRH, 23, e46. pp 345-350.

Bellman, R. and Dreyfus, S. (1962). Applied Dynamic Programming. New Jersey: Princeton Univ. Press.

Berne, A., Delrieu, G., Creutin, J.D. and Obled, C. (2004). Temporal and spatial resolution of rainfall mesurements required for urban hydrology.2004- J. Hydrol. 299, 166-179

Bringi, V. N., Chandrasekar, V., Hubbert, J., Gorgucci, E., Randeu, W. L., and Schoenhuber, M. (2003). Raindrop size distribution in different climatic regimes from disdrometer and dual-polarized radar analysis, J. Atmos. Sci., 60, 354–365,

Brown, B.G., Katz, R. W., and Murphy, A.H. (1983). Statistical analysis of climatological data to characterize erosion potential: 1. Precipitation Events in Western Oregon. Oregon Agricultural Experiment Station Spec. Rep. No. 689, Oregon State University.

Brown, B.G., Katz, R. W., and Murphy, A.H. (1984). Statistical analysis of climatological data to characterize erosion potential: 4. Freezing events in eastern Oregon/Washington. Oregon Agricultural Experiment Station Spec. Rep. No. 689, Oregon State University.

Brown, B.G., Katz, R. W., and Murphy, A.H. (1985). Exploratory Analysis of Precipitation events with Implications for Stochastic Modeling. Journal of Climate and Applied meteorology(57-67).

Cassisi, C., Montalto, P., Aliotta, M., Cannata, A. and Pulvirenti, A. (2012). Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining, Advances in Data Mining Knowledge Discovery and Applications. Adem Karahoca (Ed.), InTech, DOI: 10.5772/49941.

Chartier, M., M., (1954). Climatologie. Quelques caractères de l'année 1953 en France. L'information Géogrpahique, 18-3, pp. 115

Chu, S., Keogh, E., Hart D. and Pazzani, M. (2002). Iterative Deepening Dynamic Time Warping for Time Series. In Proc. of the Second SIAM Intl. Conf. on Data Mining. Arlington, Virginia.

Ciach, G. J. (2003). Local random errors in tipping-bucket rain gauge measurements, J. Atmos. Oceanic Technol., 20, 752– 759

Coutinho, J.V., Almeida, C. Das, N., Leal. A.M. F. and Barbarosa, L. R. (2014). Characterization of sub-daily rainfall properties in three rain gauges located in northeast Brazil. Evolving Water Resources Systems: Understanding, Predicting and Managing Water–Society Interactions Proceedings of ICAR 2014, Bologna, Italy, 345-350.

Corbin, A. (dir.), (2013). La pluie, le soleil et le vent. Une histoire de la sensibilité au temps qu'il fait. Paris, Aubier, coll. « Historique », 246p.

Cosgrove, C.M. and Garstang, M. (1995). Simulation of rain events from rain-gauge measurements. International Journal of Climatology 15, 1021–1029.

Costello, T. A. and Williams Jr., H. J., (1991). Short duration rainfall intensity measured using calibrated time-of-tip data from a tipping bucket raingage. Agricultural and Forest Meteorology, 57, 1, pp.147-155.

Cristiano, E., Veldhuis, M. and Giesen, N. (2017). Spatial and temporal variability of rainfall and their effects on hydrological response in urban areas – a review In Hydrol. Earth Syst. Sci., 21, 3859–3878.

Daumas, F. (1982). Méthodes de normalisation de données, Revue de statistique appliquée, 30(4), 23-38.

Delahaye, J.-Y., Barthès, L., Golé, P., Lavergnat J., and Vinson, J.P. (2006). a dual beam spectropluviometer concept, Journal of Hydrology, 328(1-2), 110-120.

Dilmi, M. D., Mallet, C., Barthes, L. and Chazottes, A. (2017). Data-driven clustering of rain events: microphysics information derived from macro-scale observations. Atmos. Meas. Tech., 10, 1–18.

Driscoll, E. D., Palhegyi, G. E., Strecker, E. W., and Shelley, P. E. (1989). Analysis of storm events characteristics for selected rainfall gauges throughout the United States. *US Environmental Protection Agency, Washington, DC*.

Drogue, G., Jeannée, N., Adzijian-Gérard, J. and François D. (2010). La répartition spatiale des précipitations à différentes échelles dans le Nord-Est français : observations et cartographie. In : Bulletin de l'Association de géographes français, 8ème année, 2010-2. Approches spatiales multiscalaires en climatologie. pp. 245-260

Dunkerley, D. (2008). Rain event properties in nature and in rainfall simulation experiments: a comparative review with recommendations for increasingly systematic study and reporting, Hydrological Processes, 22(22), 4415–4435, a.

Dunkerley, D. (2008). Identifying individual rain events from pluviograph records: a review with analysis of data from an Australian dryland site, Hydrological Processes, 22(26), 5024–5036.

Dunkerley, D. (2016). Interactive comment on "Data driven clustering of rain events: microphysics information derived from macro scale observations" byM. D. Dilmi et al. Atmos. Meas. Tech., Discuss.

Eagleson, P. S. (1970). Dynamic Hydrology, McGraw-Hill.

Everitt, B. (1974). Cluster Analysis. London: Heinemann Educ. Books.

Galmarini, S., Steyn, D. G., and Ainslie, B. (2004). The scaling law relating world point-precipitation records to duration. Int. J. Climatol., 24, 533–546.

Gargouri, E. and Chebchoub, A. (2010). Modélisation de la structure de dépendance hauteur-durée d'événements pluvieux par la copule de Gumbel. Hydrological Sciences–Journal–des Sciences Hydrologiques, 53(4), 802-817.

Grazioli, J., Tuia, D., and Berne, A. (2015). Hydrometeor classification from polarimetric radar measurements: a clustering approach, Atmos. Meas. Tech., 8, 149-170.

Guyon, I. and Elisseeff, A. (2003). An Introduction to Variable and Feature Selection (Kernel Machines Section), 3, 1157--1182, 2003.

Haile, A. T., Rientjes, T. H. M., Habib, E., Jetten, V., and Gebremichael, M. (2011) Rain event properties at the source of the Blue Nile River, Hydrol. Earth Syst. Sci., 15, 1023-1034.

Habib, E., Witold, F. K. and Kruger, A. (2001). Sampling Errors of Tipping-Bucket Rain Gauge Measurements. Journal of Hydrologic Engineering. 6. 10.1061/(ASCE)1084-0699(2001)6:2(159).

Hill, D. J. (2013). Automated Bayesian quality control of streaming rain gauge data, Environ. Model. Software, 40, 289–301.

Holland, J. H. (1975). Adaptation In Natural And Artificial Systems, University of Michigan Press.

Huttenlocher, D. P., Klanderman, G.A. and Rucklidge, W.J. (1993). "Comparing images using the Hausdorff distance," IEEE Trans. PAMI, vol. 15, no. 9, pp. 850–863.

Hubert P. and Carbonnel J.P.(1989). Dimension fractales de l'occurrence de pluie en climat soudano-sahélien. *Hydrologie Continentale*, 4(1), 3-10. ISSN 0246-1528

Iguchi, T., Kozu, T., Kwiatkowski, J., Meneghini, R., Awaka, J., and Okamoto, K. (2009). Uncertainties in the rain profiling algorithm for the TRMM precipitation radar. J. Meteor. Soc. Japan, 87A, 1-30.

Itakura, F. (1975). Minimum Prediction Residual Principle Applied to Speech Recognition. In IEEE Trans. Acoustics, Speech, and Signal Proc. vol. ASSP-23, pp 52-72.

Kann, A., Meirold-Mautner, I., Schmid, F., Kirchengast, G., Fuchsberger, J., Meyer, V., Tüchler, L. and Bica, B. (2015). Evaluation of high-resolution precipitation analyses using a dense station network. Hydrology and Earth System Sciences. 19. 10.5194/hess-19-1547-2015.

Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids, *in In : Dodge Y & editor, ed., , North Holland/ Elsevier,* p 405-416, 1987.

Keogh, E. and Pazzani, M. (2000). Scaling up Dynamic Time Warping for Datamining Applications. In Proc. of the Sixth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, pp.285-289. Boston, Massachuseetts.

Kohonen, T. (1982). Self-organizing formation of topologically correct feature maps. Biological Cybernetics, 46, 59-69.

Kohonen, T. (2001). Self-Organizing Maps. Springer-Verlag, ISBN 3-540-67921-9, New York, Berlin, Heidelberg.

Kruskal, J. B., (1964). Nonmetric multidimensional scaling. A numerical method. Psychometrika, 29. pp.115-129.

L'Agence Parisienne du Climat et Météo-France, (2016). Le changement climatique en France au XXeme siècle, juillet 2015 - ISBN : 978-2-9548167-3-9.

Larsen, M. L. and Teves, J. B. (2015). Identifying Individual Rain Events with a Dense Disdrometer Network, Advances in Meteorology, 2015, ID582782.

Lavergnat, J. and Golé, P. (1998). A Stochastic Raindrop Time Distribution Model. Journal of Applied Meteorology, 37, 805-818.

Lavergnat, J. and Golé, P. (2006). A stochastic model of raindrop release: Application to the simulation of point rain observations, Journal of Hydrology, 328(1), 8-19.

Leroy, M. (2000). Estimation de l'incertitude de mesure des précipitations, documentation DSO/DOS, n°42, M001693, ccrom.meteo.fr/ccrom/IMG/pdf/note42-3.pdf

Liu, Y., Weisberg, R. H. and Mooers, C. N. K. (2006). Performance evaluation of the self-organizing map for feature extraction, Journal of Geophysical Research, 111, C05018, doi:10.1029/2005JC003117, ISSN 0148-0227.

Liu, Y. and Weisberg R.H. (2011). A review of self-organizing map applications in meteorology and oceanography. In: Self-Organizing Maps-Applications and Novel Algorithm Design, 253-272.

Llasat, M.C. (2001). An objective classification of rainfall events on the basis of their convective features. Application to rainfall intensity in the north east of Spain. International Journal of climatology, 21, 1385-1400.

Lloyd, J. W. (1974). The hydrogeology and Utilization of Brines in El Salado, Chile. Groundwater, 12: 72-77.

Marzuki, M., Hashiguchi, H., Yamamoto, M. K., Mori, S., and Yamanaka, M. D. (2013). Regional variability of raindrop size distribution over Indonesia, Ann. Geophys., 31, 1941-1948.

Mestre O., 2000 : Méthodes statistiques pour l'homogénéisation de longues séries climatiques. Thèse de doctorat de l'université Paul-Sabatier (Toulouse III).

Molini, L., Parodi, A., Rebora, N. and Craig, G. C. (2011). Classifying severe rainfall events over Italy by hydrometeorological and dynamical criteria. Quarterly Journal of the Royal Meteorological Society, 137(654), 148-154.

Moisselin, J.M., Schneider, M., Canellas, C. and Mestre, O., (2002). Le changement climatique en France au XXeme siècle : Étude des longues séries homogénéisées de données de température et de précipitations, La Météorologie, 38, pp. 45-56.

De Montera, L., Barthes, L., and Mallet, C. (2009). The effect of rain-no rain intermittency on the estimation of the Universal Multifractal model parameters, J. of Hydrometeorology, 10, pp. 493–506.

Moussa, R. and Bocquillon, C. (1991). Caractérisation fractale d'une série chronologique d'intensité de pluie. Rencontres hydrologiques Franco-Romaines, 363-370.

Muller, M., Mattes, H. and Kurth, F. (2006). An Efficient Multiscale Approach to Audio Synchronization. In Proc. ISMIR, Victoria, Canada, pp. 192-197.

Parchure A. S. and Gedam S. K., (2019). Self-organising maps for rain event classification in Mumbai City, India, ISH Journal of Hydraulic Engineering, DOI: 10.1080/09715010.2019.1581099

Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. 187: 253–318. ISSN 1364-503X. doi:10.1098/rsta.1896.0007.

Pech M. E. (2019). Les jeunes dans la rue pour le climat. (Article presse dans Le figaro) (14/03/2019).

Petitjean, F. (2011). Description des alignements formés par DTW. <hal-00647522>

Pohle, I., Niebisch, M., Müller, H., Schümberg, S., Zha, T.,Maurer, T. and Hinz, C. (2018). Coupling Poisson rectangularpulse and multiplicative microcanonical random cascade models to generate sub-daily precipitation timeseries, Journal of Hydrology, 562, 50-70.

Rahmel, J. U. (1995). Similarity-based Self-organized Clustring. Proceedings 95: Workshop Fuzzy Logic and Neural Networks

Rubner, Y. , Tomasi, C. and Guibas, L.J. (1998). A Metric for Distributions with Applications to Image Databases, in Proc. IEEE ICCV, pp.59–66.

Sadler, E. J., and Busscher, W. J. (1989). ‘‘High-intensity rainfall ratedetermination from tipping-bucket rain gauge data.’’Agronomy J., 68,126–129.

Salvador S. and Chan., P. (2007). Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis, 11, 5, 561-580.

Sakoe, H. and Chiba, S. (1971). A dynamic programming approach to continuous speech recognition, Proc. 7th ICA, Paper 20 CI3.

Sakoe, H. and Chiba, S. (1970). A similarity evaluation of speech patterns by dynamic programming, Dig. Nat. Meeting, Inst. Electron. Comm. Eng. Japan, p. 136.

Sakoe, H. and Chiba, S. (1973). Comparative study of DP-pattern matching techniques for speech recognition, Tech. Group Meeting Speech, Acoust.SOC. Japan,Preprints (S73-22).

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoustics, Speech, and Signal Proc., Vol. ASSP-26.

Sarda-Espinosa, A. (2017). Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package, R package, https://cran.rproject. org/web/packages/dtwclust/vignettes/dtwclust.pdf.

Schertzer, D. and S. Lovejoy (1987). *Physically based rain and cloud modeling by anisotropic, multiplicative turbulent cascades*. J. Geophys. Res. 92, 9692-9714.

Sorlin, S., Champin, P. A., Solnon, C., (2003). Mesurer la similarité de graphes étiquetés. 9èmes Journées Nationales sur la résolution pratique de problèmes NP-complets (JNPC), France, pp.325-339.

Suh, S.-H., You, C.-H. and Lee, D.-I. (2016). Climatological characteristics of raindrop size distributions in Busan, Republic of Korea, Hydrol. Earth Syst. Sci., 20, 193-207.

Sung, P., Syed, Z. and Guttag, J. (2009). Quantifying Morphology Changes in Time Series Data with Skew. Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. International Conference on Acoustics, Speech and Signal Processing. 477-480.

Susana Ochoa-Rodriguez et al. (2015). Impact of spatial and temporal resolution of rainfall inputs on urban hydrodynamic modelling outputs: A multi-catchment investigation Journal of Hydrology 531 389–407.

Tabeaud, M., (2010). Climats urbains. Savoirs experts et pratiques sociales , Ethnologie française, 4 (Vol. 40), pp. 685-694. doi : 10.3917/ethn.104.0685.

Tapiador, F. J., Checa, R., and de Castro, M. (2010). An experiment to measure the spatial variability of rain drop size distribution using sixteen laser disdrometers. Geophysical Research Letters, 37(16), ID L16803.

Testud, J., Oury, S., Amayenc, P. and Black, R. A. (2001). The concept of ‘‘normalized’’ distributions to describe raindrop spectra: A tool for cloud physics and cloud remote sensing, J.

Appl. Meteor., 40, 1118–1140.

Tokay, A. and Öztürk, K. (2012). An Experimental Study of the Small-Scale Variability of Rainfall, Journal of Hydrometeorology, 13 (1), pp 351-365.

Tsiporkova E. (2013). Dynamic Time Warping Algorithm for. PPT presentation available at: http://www.psb.ugent.be/cbd/papers/gentxwarper/DTWAlgorithm.ppt, date of the last visit.

Ulaby, F. T., Moore, R. K., and Fung, A. K. (1981). Microwave Remote Sensing: Fundamentals and Radiometry. Vol. I. Artech House, 321-327.

Uriarte, E. A. and Martín, F. D. (2008). Topology Preservation in SOM, World Academy of Science, Engineering and Technology. International Journal of Computer, Electrical, Automation, Control and Information Engineering 2, 9.

Verrier, S. (2011). Modélisation de la variabilité spatiale et temporelle des précipitations à la sub-mésoéchelle par une approche multifractale. PhD thesis, Université Versailles-St-Quentin-en-Yvelines / Université Paris-Saclay

Verrier, S., de Montera, L., Barthès, L. and Mallet, C. (2010). Multifractal analysis of African monsoon rain fields, taking into account the zero rain rates problem, J. of Hydrology, pp. 389,111-120.

Verrier, S., Mallet, C. and Barthès, L. (2011). Multiscaling properties of rain in the time domain, taking into account rain support biases Journal of Geophysical Research- Atmospheres, J. Geophys. Res., 116, doi:10.1029/2011JD015719.

Verrier, S., Barthès L., Mallet C. (2013). Theoretical and empirical scale dependency of Z-R relationships: Evidence, impacts, and correction, Journal of Geophysical Research: Atmospheres, 118 (14), 7435-7449.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map. IEEE Transactions on Neural Networks, Vol. 11, 586–600, ISSN 1045-9227.

Villarini, G., Mandapaka, P. V. ,Krajewski, W. F. and Moore, R. J. (2008), Rainfall and sampling uncertainties: A rain gauge perspective, J. Geophys. Res., 113, D11102, doi:10.1029/2007JD009214.

Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., Kendon, E. J., Lenderink, G., and Roberts N. M. (2014), Future changes to the intensity and frequency of short-duration extreme rainfall, *Rev. Geophys.*, 52, 522–555, doi: 10.1002/2014RG000464.

Zhanga, Z. , Tavenardb, R., Baillyb, A., Tangc, X., Tanga, P. and Corpetti, T. (2017). Dynamic Time Warping Under Limited Warping Path Length. Information Sciences, 393, pp 91-107.

Zinke, A. and Mayer, D. (2006). Iterative Multi Scale Dynamic Time Warping. Computer Graphics technical reports, CG-2006/1.

Annexes

Annexe 1 : stations Météo-France

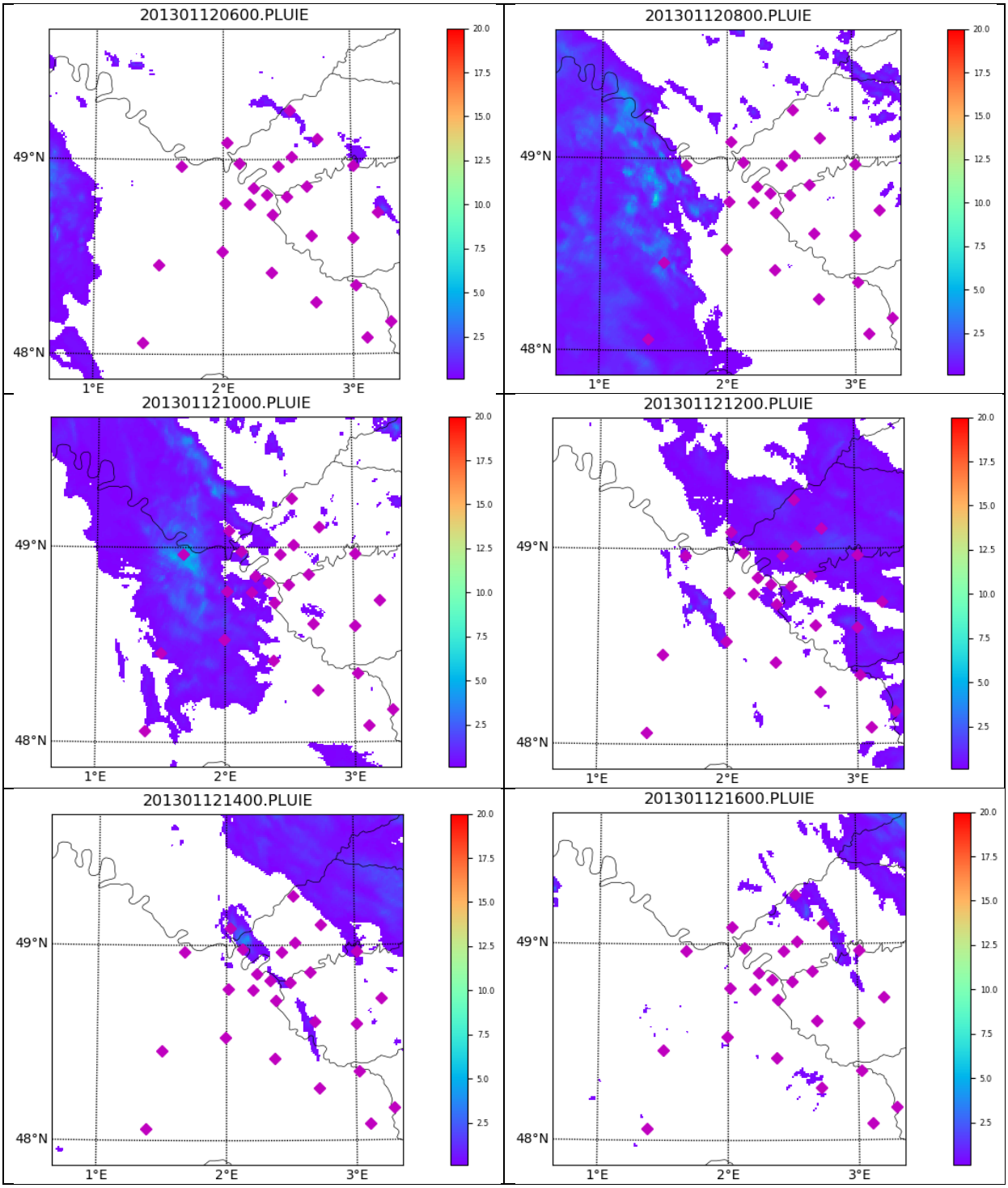
Nom	Numéro	Début	fin	disponibilités des données	occurrence de pluie		% de non pluie
disdromètre		2008	2015	86	$T = 1min$	4.28	95.72
Pluvio_Z1(0.1mm)		18/04/2005	02/09/2015	97.56		0,91	99,09
Pluvio_Z2(0.2mm)		19/07/2002	01/06/2016	98.52		0,56	99,44
Chartres	28070001	01/01/2006 00:00	31/12/2015 14:54	97,23	2,86		97,13
Châteaudun	28198001	01/06/2007 00:00	31/12/2015 23:54	97,24	2,3		97,7
Creil	60175001	28/06/2007 10:06	31/12/2015 23:54	96,3	2,66		97,33
Plessis-Belleville	60500004	01/01/2006 00:00	31/12/2015 23:54	97,67	2,51		97,48
Paris-Montsouris	75114001	01/01/2006 00:00	31/12/2015 23:54	96,96	2,41		97,58
Longchamp	75116008	14/03/2007 10:06	31/12/2015 23:54	99,33	2,15		97,84
La brousse-mx	77054001	01/01/2006 00:00	31/12/2015 23:54	97,56	2,41		97,58
Changis	77084001	01/01/2006 00:00	31/12/2015 23:54	96,9	2,75		97,24
Chevru	77113002	01/01/2006 00:00	31/12/2015 23:54	96,27	2,52		97,48
Nangis	77211001	01/01/2006 00:00	31/12/2015 23:54	96,88	2,73		97,27
Melun	77306001	01/01/2006 00:00	31/12/2015 23:54	95,35	4,51		95,49
Nemours	77333003	01/01/2006 00:00	31/12/2015 23:54	97,48	2,71		97,28
Torcy	77468001	01/01/2006 00:00	31/12/2015 23:54	97,6	2,5		97,5
Acheres	78005002	01/01/2006 00:00	31/12/2015 23:54	97,61	2,2		97,8

Magnanville	78354001	01/12/2006 00:00	31/12/2015 23:54	99,4	2,34	97,65
Trappes	78621001	01/12/2006 00:00	31/12/2015 23:54	97,29	3,46	96,53
Villacoublay	78640001	14/05/2007 11:06	31/12/2015 23:54	97,49	2,78	97,22
Savigny/clairis	89380001	01/01/2006 00:00	31/12/2015 23:54	97,77	2,59	97,41
Sens	89387002	01/01/2006 00:00	31/12/2015 23:54	96,59	2,36	97,64
Orly	91027002	01/12/2006 00:00	31/12/2015 23:54	98,91	2,89	97,1
Courdimanche	91184001	01/01/2006 00:00	31/12/2015 23:54	97,69	2,11	97,88
Dourdan	91200002	01/01/2006 00:00	31/12/2015 23:54	97,28	2,61	97,38
St-Maur	94068001	01/01/2006 00:00	31/12/2015 23:54	97,52	2,54	97,46
Pantoise Aero	95078001	01/12/2006 00:00	31/12/2015 23:54	97,59	3,59	96,41
Lebourget	95088001	01/12/2006 00:00	31/12/2015 23:54	98,44	2,78	97,2
Roissy	95527001	01/12/2006 00:00	31/12/2015 23:54	99,25	2,97	97,03

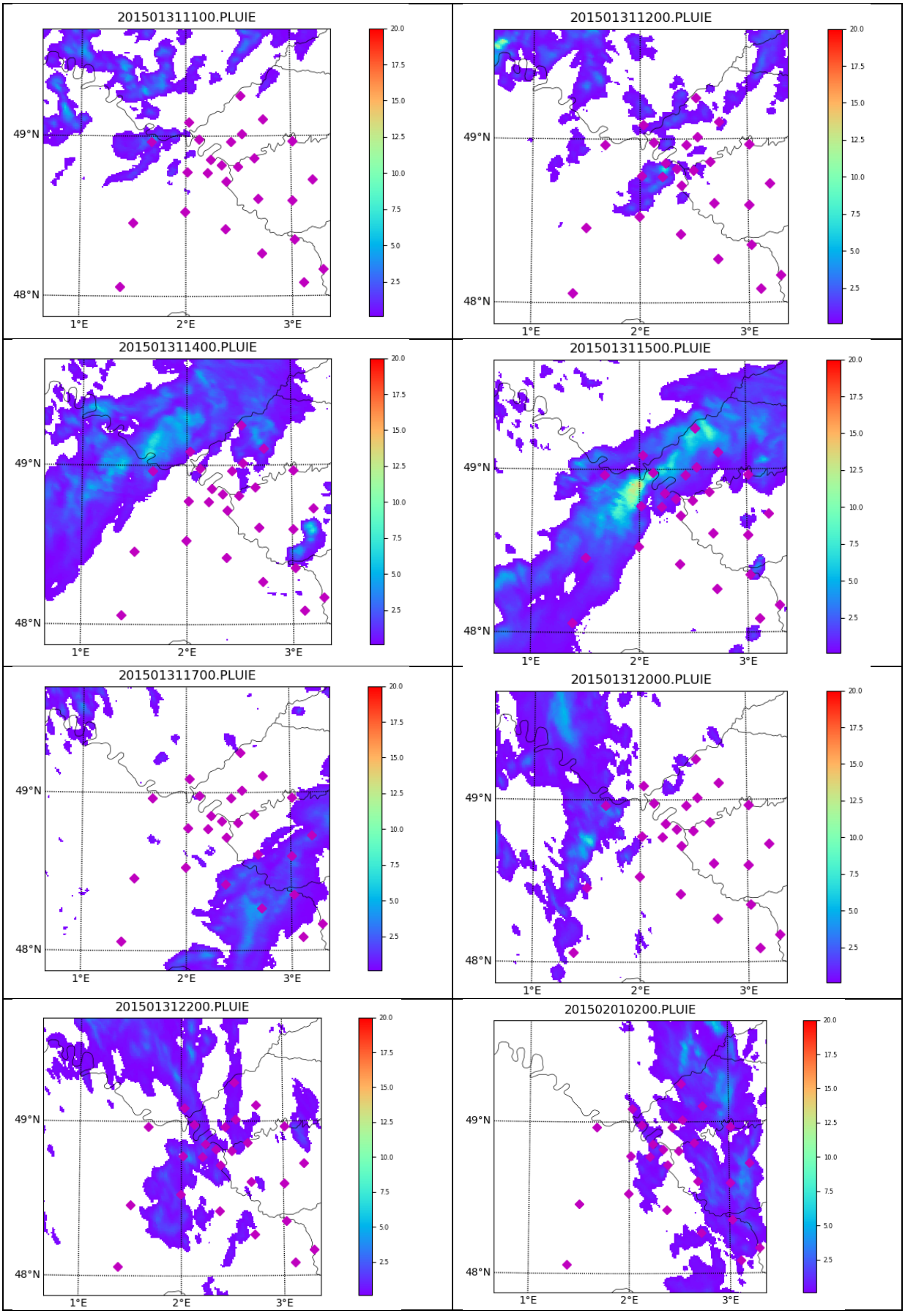
Annexe 2 : profils (cartes de radar) des événements représentants des trois classes

La présentation des cartes suit la convention des tableau en informatique : on parcourt la colonne et en suite les lignes

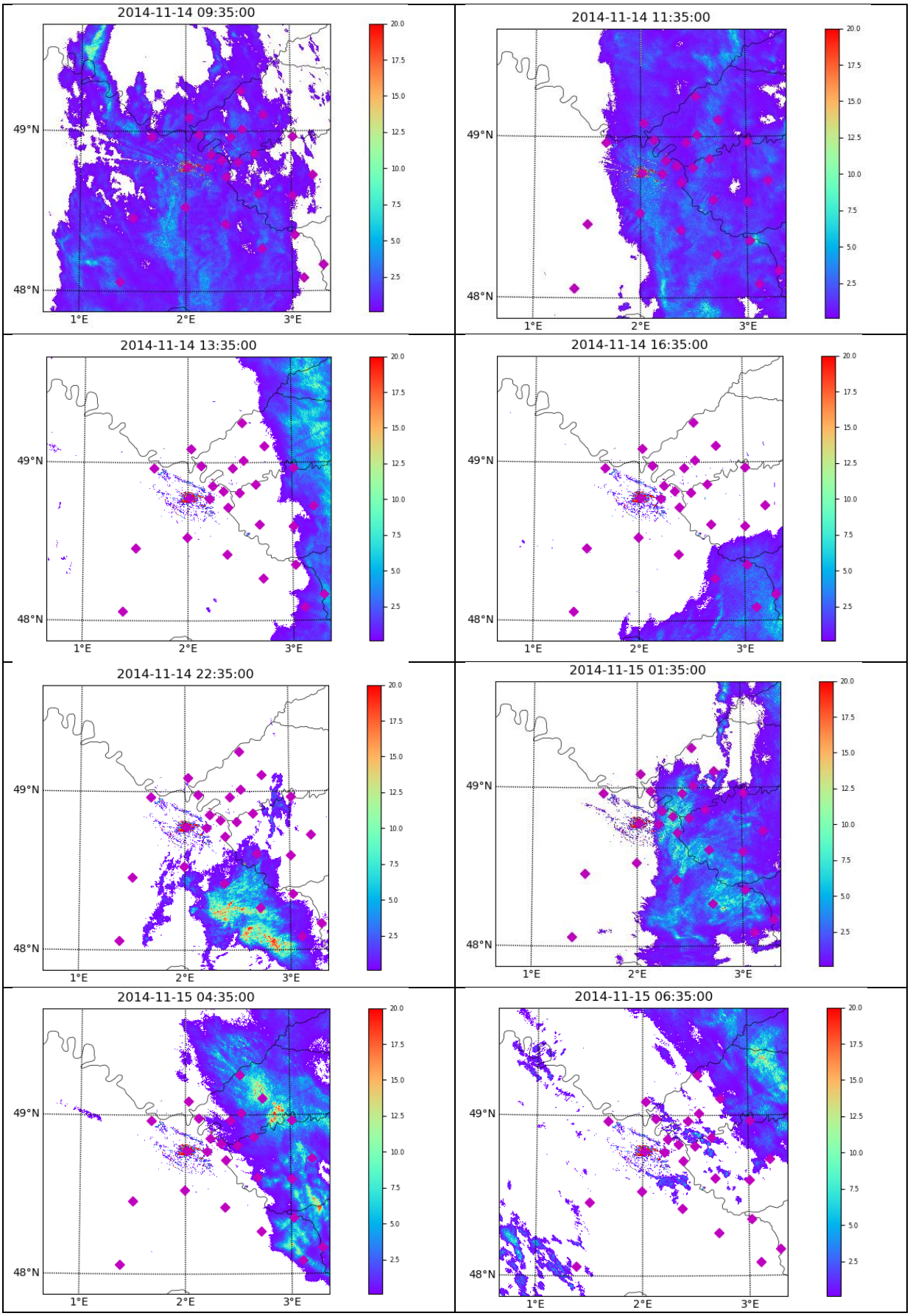
Annexe 2.1 : cartes radar du 12 janvier 2013 (événement 492)



Annexe 2.2 : cartes radar du 31 janvier 2015 (événement 773)



Annexe 2.3 : cartes radar du 14 novembre 2014 (événement 744)



Annexe 3 : histogramme des retards

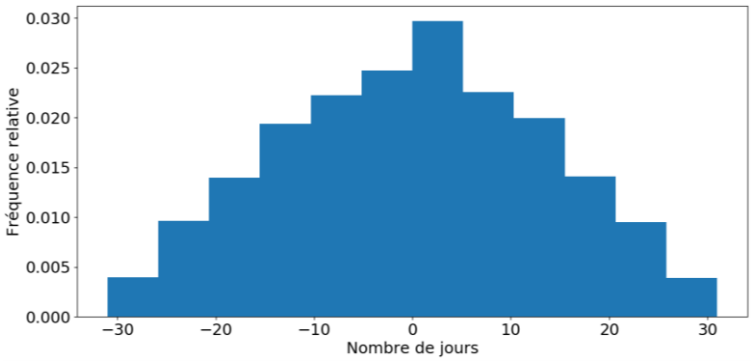


Figure histogramme des retards

Table des illustrations

Table du chapitre 1

Table du chapitre 2

Table du chapitre 3

Table du chapitre 4

Table du chapitre 5

Table du chapitre 6

Table des tableaux

Table du chapitre 1

Table du chapitre 2

Table du chapitre 3

Table du chapitre 4

Table du chapitre 5

Table du chapitre 6

Résumé

La question de l'impact du changement climatique sur l'évolution temporelle des **précipitations** ainsi que l'impact de l'ilot de chaleur parisien sur la répartition spatiale des précipitations motivent l'étude la **variabilité du cycle de l'eau** à fine échelle en Île-de-France. Le réseau pluviométrique de météo-France compte 26 stations en région parisienne et mesure les précipitations avec une résolution temporelle de 6 minutes depuis une dizaine d'années années, ce qui présente un important volume de données sous formes de **séries temporelles de taux précipitant** à analyser. **La classification non supervisée des séries temporelles** peut être utilisée pour explorer ce jeu de données.

La pluie est un processus intermittent, non stationnaire, et présente une variabilité extrême. Ses caractéristiques font que la plupart des méthodes de classification existantes ne peuvent pas être directement utilisées. L'objectif de la thèse est une recherche méthodologique en classification des séries temporelles afin de rendre possible l'interprétation des observations réalisées à différentes échelles tant spatiales que temporelles. Dans le cadre de cette thèse, nous proposons deux approches pour la classification et la comparaison de séries temporelles des précipitations à différentes échelles de temps et nous discutons de chacune d'elles.

La première approche « classique » est basée sur la description des séries par des **caractéristiques** définies par des experts du domaine. Ces dernières présentent une redondance d'information. Nous avons développé un algorithme de **sélection des caractéristiques** basé sur **les algorithmes génétiques (GA)** et les **cartes topologiques (SOM).** Une liste de caractéristiques obtenue à partir des données pluviométriques mesurées par un disdromètre (DBS) à fine échelle (1min) a permis de mettre en évidence une relation entre la microphysique des précipitations et l'observation macro-physique. Cependant, la liste optimale des caractéristiques est dépendante du moyen d'observation et de l'échelle de temps étudiée (i.e. la liste précédente n'est pas utilisable sur les données des pluviomètres à l'échelle de 6 min).

La deuxième approche basée sur la notion de dissimilarité entre les séries temporelles fait abstraction d'une description préalable par caractéristiques afin d'améliorer la généricité de l'approche. Nous avons développé une mesure de dissimilarité baptisée I**terative Multiscale Dynamic Time Warping (IMSDTW)** adaptée aux séries temporelles de précipitations, cette dernière qui s'inspire de la méthode Dynamic Time Warping (DTW) renvoie une mesure et un **alignement (path)** liant les deux séries comparées. Nous avons montré l'intérêt d'analyser l'alignement estimé et comment on peut l'utiliser pour détecter la trajectoire et étudier

l'évolution d'une cellule pluvieuse. La sensibilité de la dissimilarité et de l'alignement au changement d'instruments et à l'échelle de temps a été quantifiée.

L'approche basée sur l'IMSDTW a d'abord été appliquée à l'évaluation de la **variabilité spatiale** des précipitations en île de France. Une classification des 26 stations exploitant la mesure de dissimilarité conserve la topologie géographique sans apporter de nouvelles informations alors que l'exploitation de l'information des alignements prend en compte des propriétés météorologiques des évènements précipitants.

Pour l'évaluation de **la variabilité temporelle des précipitations**, une classification des événements de précipitation observés par une station a été réalisée. La classification obtenue correspond à la typologie des précipitations. L'approche proposée permet de réaliser une classification des évènements observés par l'ensemble du réseau pluviométrique en prenant ainsi en compte la variabilité spatiale.

L'application sur la série historique de Paris-Montsouris (1873-2015) permet de discriminer automatiquement les années « exceptionnelles » d'un point de vue météorologique. L'analyse de la fréquence d'apparition des années exceptionnelles au cours du temps peut être un indicateur de l'impact de l'évolution du climat sur l'évolution des précipitations.

Mots clés : Précipitations, Séries temporelles de taux précipitant, classification non supervisée des séries temporelles, sélection des caractéristiques, algorithmes génétiques (GA), cartes topologiques auto organisatrice (SOM), dynamic time warping (DTW), iterative multiscale dynamic time warping (IMSDTW), alignement (path).