Typisation des situations météorologiques par classification automatique en vue de la prévision locale

Making a typology of meteorological conditions by automatic classification for purposes of local forecasting

G. Der Megreditchian, G. Legendre et M. Pasquier

Météorologie Nationale

L'approche classique de l'étude et de la prévision statistique de l'occurrence d'un phénomène, et plus précisément des précipitations, repose essentiellement sur des méthodes d'Analyse Discriminante [3]. La stabilité de la méthode, lors du passage du fichier d'apprentissage au fichier test. impose un nombre réduit de prédicteurs alors que d'autre part l'exhaustivité de la description des situations météorologiques en exige un grand nombre.

A cette première contradiction s'ajoute celle, plus fondamentale, de l'hypothèse simpliste qui préside à l'emploi de ces méthodes classiques, à savoir que la Nature effectuerait un tirage aléatoire choisissant en fonction de la situation météorologique l'occurrence ou la non-occurrence des précipitations.

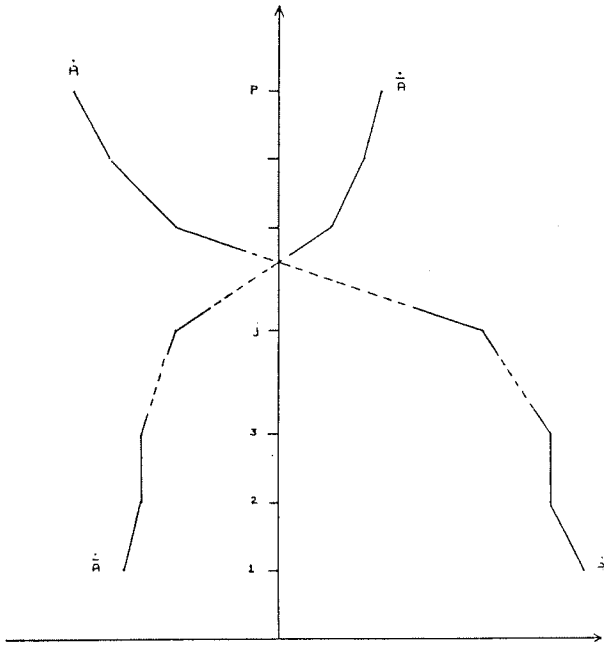
Cette dernière hypothèse n'est pas physiquement justifiée car, en réalité, les précipitations peuvent résulter de situations fort différentes et identiquement être absentes dans nombre de cas dissemblables. Il faut donc considérer qu'il y a au sein des situations météorologiques plusieurs populations distinctes que nous pouvons appeler "types de temps" favorables ou défavorables à l'occurrence de la pluie [1]. C'est à la recherche de ces situations-types qu'a été appliquée la méthode de classification automatique réalisée par l'algorithme "des Nuées Dynamiques". On a également élaboré une méthode de visualisation de ces situations dite Méthode des Profils.

Une façon de visualiser une situation paramétrée par les prédicteurs (x_1, x_j, x_p) consiste à tracer son profil. On appellera profil du vecteur x de coordonnées (x_1, x_j, x_p) la représentation graphique de ses coordonnées dans un système d'axes, obtenue en reliant dans l'ordre des indices j les points d'abscisse x et d'ordonnée entière j. De la même façon on définit et on représente le profil moyen d'un ensemble de situations. Plus précisément on peut distinguer a posteriori sur un fichier d'apprentissage le profil moyen des situations pour lesquelles on a observé l'occurrence de précipitations, et celui des cas pour lesquels cet événement n'a pas été constaté. On obtient ainsi deux profils dits "complémentaires". Si l'on centre chacune des variables x, ces deux profils sont homothétiques dans le rapport inverse des cardinaux de leur population au signe près. Le profil moyen de l'ensemble des situations est alors le profil nul (axe vertical). Il s'agit là d'une propriété utile à la représentation des profils et à la visualisation des situations. En effet, si l'on ordonne les prédicteurs de telle sorte que l'un des deux profils complémentaires soit croissant ou décroissant, l'autre profil sera au contraire décroissant ou croissant par le jeu de l'homothétie. Sur ces deux profils de base peuvent être superposés des profils de situations particulières ou des profils-types obtenus par une méthode de classification automatique, ce qui permet d'analyser la tendance d'un profil à se rapprocher, au sens d'une métrique qu'il convient de définir, de telle ou telle modalité, ainsi que les particularités de structure qu'il présente (Fig. 1).

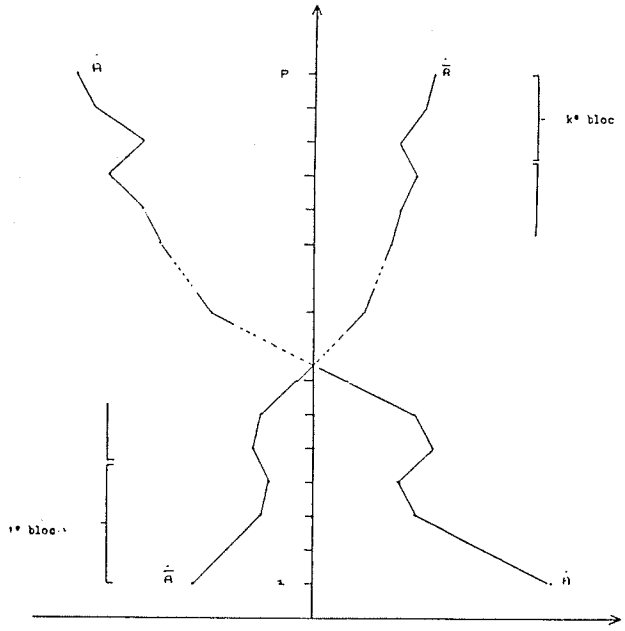
La Méthode des Nuées Dynamiques [2] est une des méthodes les plus répandues de Classification Automatique. Elle a pour but de fournir une partition (c'est-à-dire un partage en k classes disjointes) devant expliquer au mieux la structure interne d'un ensemble de n individus, au sens d'une métrique préalablement définie.

Appliquée à un ensemble de situations météorologiques, elle devrait donc permettre de dégager des classes, c'est-à-dire des "types" de temps. Il s'agit là d'un problème complexe a priori puisque le nombre de variantes possibles des partitions d'un ensemble de n objets en k classes est donné par le nombre de Stirling de 2^e espèce :

$$S[n, k] = \frac{1}{k!} \sum_{j=1}^{k} (-1)^j C_k^j [k-j]^n.$$

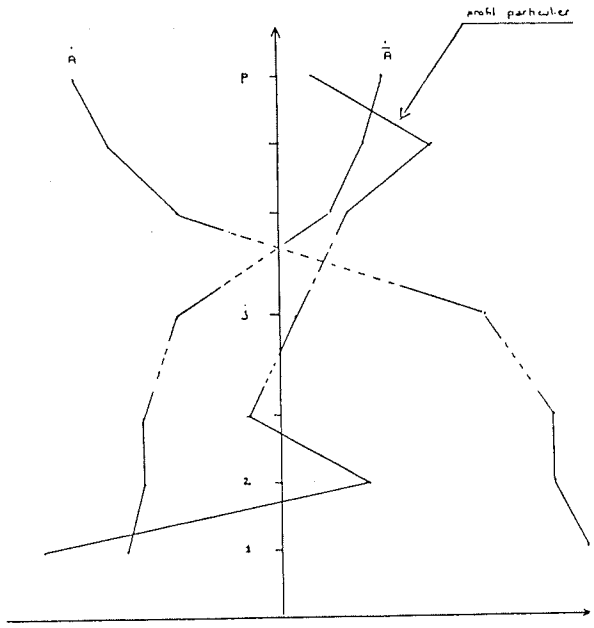


1 *a* – Ordre inital des variables.



1 *b* – Variables réordonnées.

Figure 1



1 *c* – Profils moyens pluie et non-pluie et profil d'une journée.

A titre indicatif rappelons que $S\ (15,4) = 42\ 335\ 950$.

La méthode à laquelle nous nous sommes plus particulièrement intéressés est celle, basée sur un algorithme original de E. Diday, qui s'inscrit dans la lignée des algorithmes utilisant une caractérisation des classes par des étalons. (MacQueen). Ayant muni un ensemble F d'une certaine métrique, ce type d'algorithme permet d'obtenir une partition suboptimale des éléments de cet ensemble. L'agglomération des classes s'effectue autour d'ensembles d'éléments appelés "noyaux", qui sont des parties de F. Une classe de F sera donc représentée par un ou plusieurs de ses éléments qui seront ses étalons.

En début d'algorithme doivent être précisés :

1) K. le nombre de classes exigé ;
2) Nk. le nombre d'étalons de la K-ième classe.

La structure des noyaux initiaux et la précision de cette structure peuvent présenter de nombreuses variantes. On définit la distance entre un élément et un noyau comme étant la somme des distances de cet élément à tous les éléments de l'ensemble du noyau divisée par le cardinal de ce noyau. Le processus de la méthode est ensuite itératif, le schéma d'une étape de ce processus étant globalement le suivant :

1) on procède à une agrégation autour des étalons représentant les classes : un individu sera affecté à la classe numéro k pour laquelle la distance entre cet individu et son novau sera minimale. Après avoir affecté tous les éléments à une classe, on calcule l'inertie de la partition qui est la somme des inerties internes de chaque classe par rapport à l'étalon le plus proche de son centre de gravité (que l'on appelle le "meilleur étalon"). On définit l'inertie interne d'une classe par rapport à un de ses éléments comme étant la somme des distances de chacun de ses éléments à cet élément particulier.

2) on recherche de nouveaux étalons pour chaque classe. Pour une classe donnée on détermine la plus petite distance de l'ensemble des éléments de cette classe avec l'un de ses étalons. Cela revient à déterminer "le meilleur" étalon.

3) on détermine ensuite les M éléments de la classe pour lesquels la distance à l'ensemble des autres éléments de la classe est inférieure à une certaine valeur d_0. Si $M < N(k)$, on revient en (2) en calculant un nouveau d_1, la plus petite distance de l'ensemble des éléments de la classe avec l'un des $N(k) - 1$ premiers étalons de cette classe et ceci jusqu'à obtenir M "candidats étalons" avec $M \geqslant N(k)$, ou bien une impossibilité par suite de $N(k)$ échecs.

4) parmi ces M éléments on choisit les $N(k)$ pour lesquels la distance est la plus petite. Ils deviennent les nouveaux étalons de la classe numéro k. On recommence le processus (2), (3), (4) pour chacune des classes. L'arrêt du processus a lieu après l'étape (1) lorsqu'on constate que l'inertie de la partition ne s'est pas améliorée par rapport à celle précédemment calculée. L'organigramme de la figure 2 présente schématiquement le processus logique de l'algorithme de Diday. On peut justifier l'amélioration au sens de la diminution de l'inertie et par conséquent la convergence de cet algorithme. Il faut surtout retenir que ce procédé est suboptimal dans la mesure où il n'effectue par une étude exhausive, mais atteint un optimum local. Le fait que les noyaux appartiennent au fichier conduit à une relative simplicité dans l'interprétation des résultats. D'autre part, de multiples choix de distances sont possibles. La partition finale dépend largement de l'initialisation pouvant être réalisée, soit par la donnée d'une partition initiale, soit par tirage au hasard, soit par la procédure suivante :

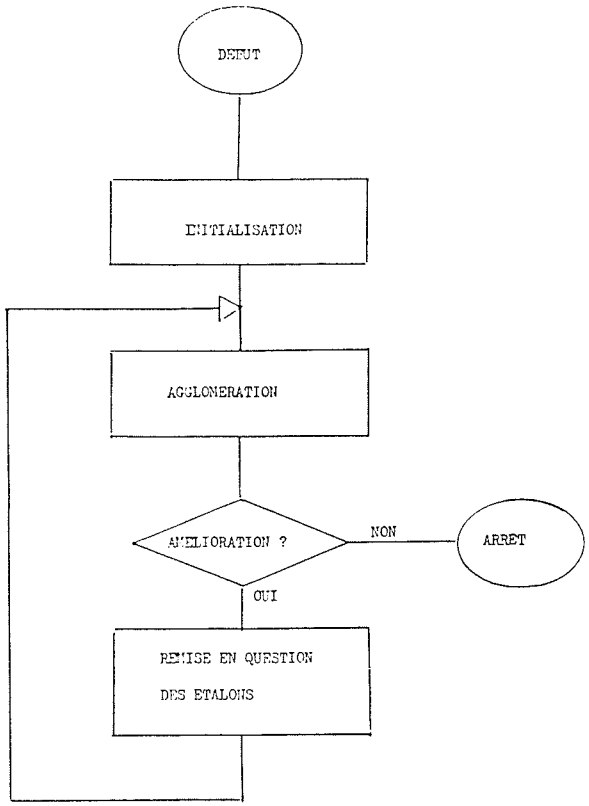


Figure 2 – Organigramme de l'algorithme des Nuées Dynamiques.

1) on tire au hasard le premier étalon
2) on prend pour deuxième le plus éloigné du premier ;
3) on prend pour troisième le plus éloigné des deux premiers et ainsi de suite jusqu'à finir par la remise en question du premier.

Cette forme d'initialisation, bien qu'inadaptée à certains cas précis de fichier, a l'avantage de ne pas présenter d'aberrations (étalons voisins). Le problème du choix du nombre de classes demeure assez ouvert, et est en fait étroitement lié au fichier étudié. Tout au plus peut-on dans le cas général fournir un ordre de grandeur de k en fonction de n, étant entendu que la fluctuation permise autour de ces valeurs indicatrices est très large.

On peut prendre, en l'absence d'autres informations complémentaires, par exemple les valeurs "raisonnables" suivantes :

$$k = 1 + 3{,}22 \log(n), \quad \text{ou} \quad k = 5 \log_{10}(n).$$

Le fichier étudié comportait 616 observations quotidiennes pour la période du 01/04/75 au 31/03/77 pour la station de Rennes. Le but de l'étude était la prévision des précipitations à courte échéance.

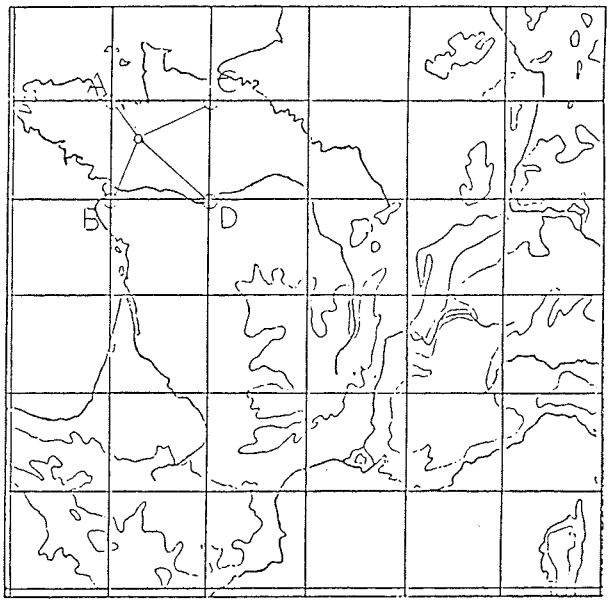


Figure 3 – Position relative des points de grille et de la station Rennes.

Le prédictand était donc une variable binaire, la variable indicatrice de l'occurrence des précipitations.

Sur les 610 observations, 410 se rapportaient à la non occurrence et 206 à l'occurrence des précipitaions.

Le vecteur des prédicteurs comportait 93 variables fournies par le modèle dynamique AMETHYSTE [4] pour les quatres points de grille entourant Rennes (Fig. 3).

Le fichier test contenait 221 observations, (du 01/04/77 au 31/03/78), dont 127 jours secs et 94 jours pluvieux.

Une procédure de Sélection Progressive Ascendante et de Sélection Aléatoire Adaptative avait permis de retenir les 8 prédicteurs suivants :

x_1 : humidité au point A à 12 h TU.
x_2 : la composante $W.E$ du vent au point B à 6 h TU.
x_3 : la composante $W.E$ du vent au point C à 6 h TU.
x_4 : la stabilité de la couche 700/500 mbar au point B à 12 h TU.
x_5 : l'occurrence des précipitations au point D entre 6 et 12 h.
x_6 : la stabilité de la couche 900/700 mbar au point B à 18 h TU.
x_7 : la température à 900 mbar au point A à 18 h TU.
x_8 : $\sqrt{R}$.

Sur ce nouveau fichier des prédicteurs sélectionnés on a calculé les profils moyens des classes "pluies pour les variables centrées réduites".

On dispose ainsi, en introduisant une distance adéquate entre profils, d'une méthode simple de prévision basée sur la proximité relative des profils moyens types (Fig. 4).

La stratégie triviale de la persistance fournit alors une deuxième méthode dont la qualité servira de référence.

La méthode de Classification Automatique a été appliquée au fichier des huit meilleurs prédicteurs en

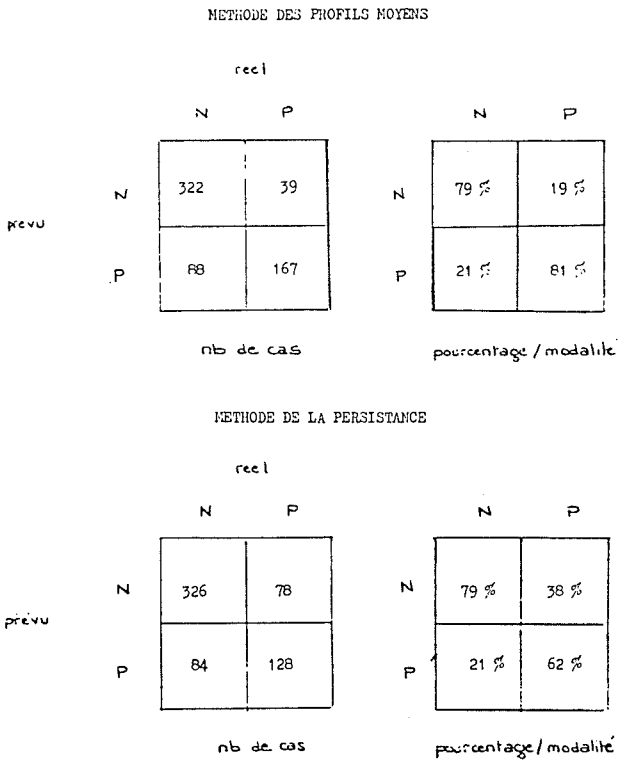


Figure 4 – Qualité de la prévision pour les stratégies triviales.

CLASSIFICATIONS

CLASSES	3	4	5	6	7	8	9	10
1	26,7	14,2	34,2	4,8	68,4	22,0	46,3	4,5
2	4,1	2,6	3,5	0,0	0,0	0,0	0,0	0,0
3	72,3	73,2	76,9	76,9	85,8	80,3	80,8	80,8
4		32,8	32,0	32,0	28,8	28,2	17,8	17,6
5			6,1	5,6	4,9	4,0	2,0	2,9
6				54,2	44,3	70,0	70,0	70,0
7					4,8	5,5	5,6	8,3
8						68,4	75,3	75,3
9							18,9	19,2
10								46,3

a)

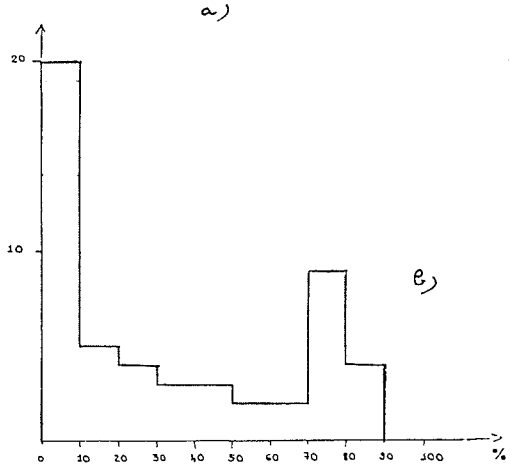


Figure 5 – Pourcentage de pluie (*PP*) dans les différentes classes (5 *a*) et histogramme du nombre de classes pour *PP* fixée (5 *b*).

essayant différentes distances (euclidienne, Mahalanobis, profils).

Le nombre de classes, estimé entre 9 et 10 par les formules susmentionnées, fut précisé par une exploration systématique entre 3 et 10.

La méthode fut utilisée avec une phase d'initialisation conduisant à des étalons suffisamment éloignés, et un certain nombre de paramètres permettant de caractériser chaque classe dont :

ET effectif total de la classe, *EN* effectif non-pluie de la classe, *EP* effectif pluie de la classe, *PN* pourcentage non-pluie de la classe, *PP* pourcentage de pluie de la classe, *RN* représentativité de la classe par rapport aux non-pluies, *RP* représentativité de la classe par rapport aux pluies.

La représentativité d'une classe par rapport à la modalité pluie est ici le pourcentage des éléments de la population Pluie (Non-Pluie) qui appartiennent à cette classe.

La figure 5 indique le pourcentage de pluie (*PP*) pour les différentes classes des *P* classification et l'histogramme du nombre de classes ayant une valeur donnée de *PP*. On découvre ainsi une structure tranchée des classes obtenues.

Les classes favorisant la pluie ($PP > 50$ %), les classes favorisant la non-pluie ($PP < 18$ %), et les classes indécises (18 % $< PP < 50$ %).

La visualisation des paramètres respectifs pour ces 3 sortes de classes nous montre l'évolution de la structure définie par la typisation en fonction du nombre de classes (Fig. 6). On remarque que les classes augmentent rarement leur effectif, que les classes de pluie s'épurent de leurs non-pluies et inversement au cours de leur "vieillissement", que la seule anomalie de structure concerne la classe numéro 6, qui perd son caractère Pluie lors de la classification en 7 classes pour le retrouver ensuite.

Soulignons ici l'étonnante stabilité entre les classifications en 5 et 6 classes.

Par ailleurs les classifications en 7 classes ou plus perdent de leur intérêt à cause de la trop grande importance du risque dans la structure qu'elles proposent.

L'aspect dynamique des classes qui permet de cerner la meilleure structure à adopter se retrouve aussi par examen des résultats graphiques grâce à l'édition des profils moyens de toutes les classes de chacune des 8 classifications (Fig. 7).

De nombreuses remarques particulières pourraient être faites quant à cette évolution dynamique et graphique des classes. Bornons-nous à indiquer que c'est une première manière d'étudier la structure optimale du fichier disponible.

Pour concrétiser la faisabilité de cette approche, il faut étudier son comportement en tant qu'outil prévisionnel. La qualité de cet outil sera mesurée et comparée en introduisant une matrice des coûts, qui fixe le taux des pénalités infligées aux erreurs de prévision.

Celle que nous avons adoptée (Fig. 8) pénalise lourdement la non-prévision de pluie et n'est pas indulgente envers une prévision systématique de risque. Pour mieux appréhender cette fonction de coût, précisons les coûts global et journalier correspondant à 3 stratégies triviales (calculés sur le fichier d'apprentissage) (Fig. 9).

On peut distinguer deux sortes de coûts :

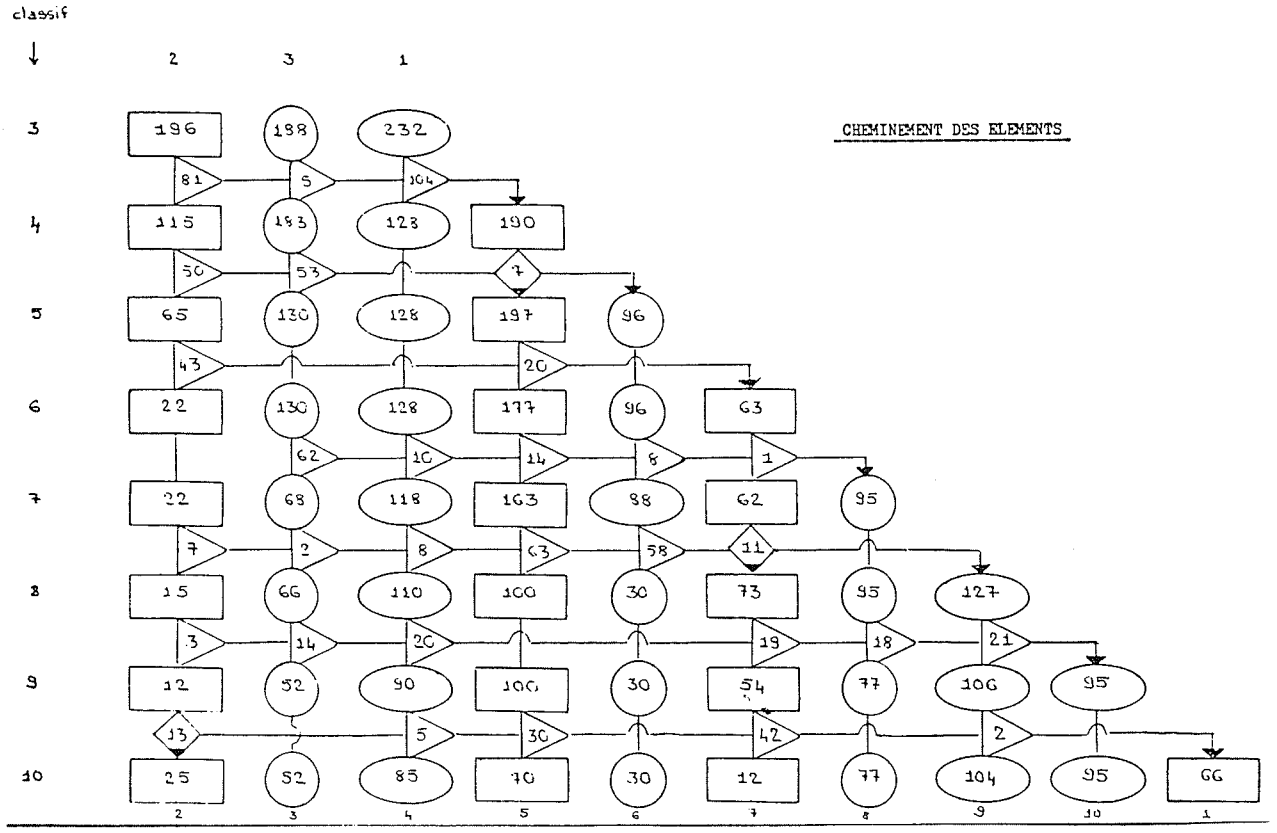


Figure 6 – Evolution dynamique des éléments des classes.

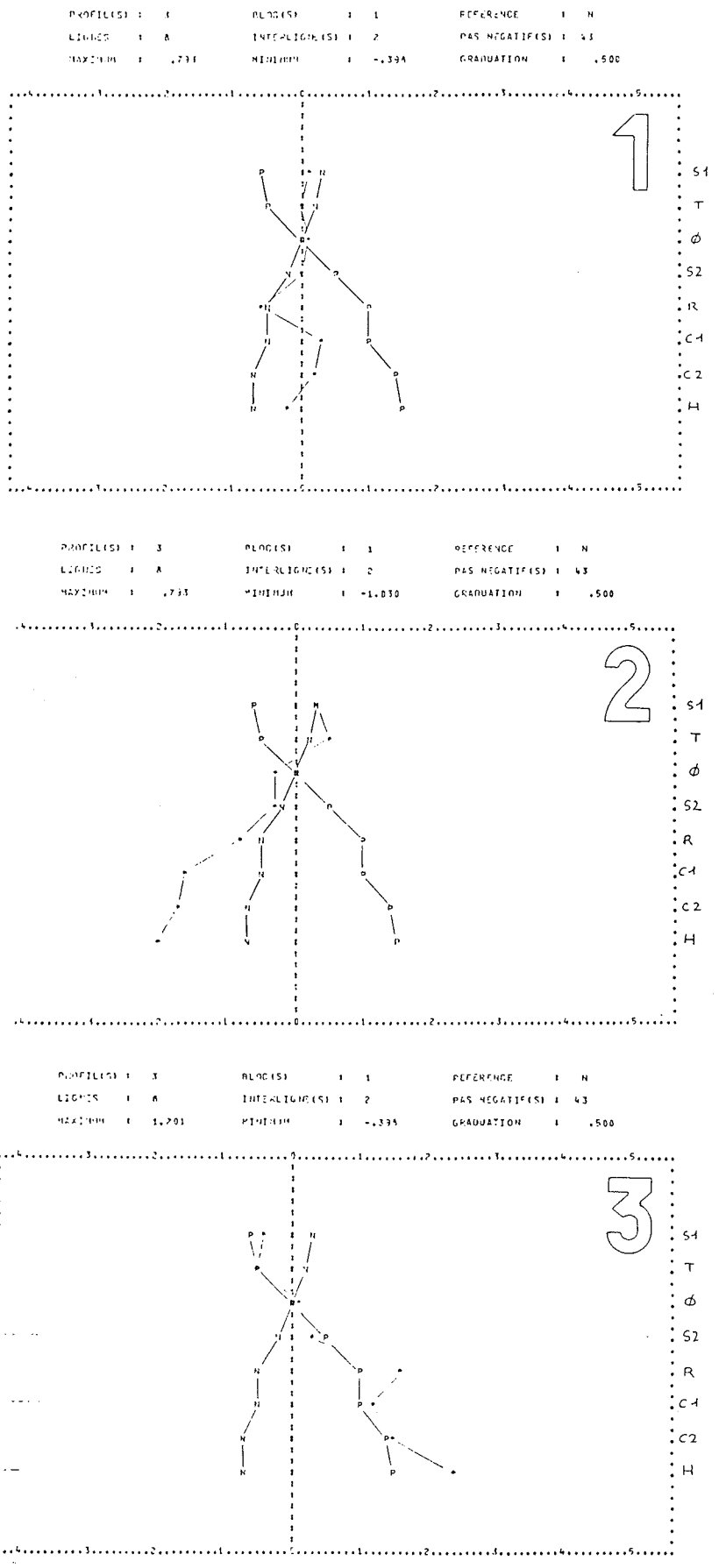


Figure 7 – Profils moyens global, pluie et non-pluie pour le cas de 3 classes.

	N	P
N	0	8
R	2	1
P	4	0

Figure 8 – Matrice des coûts utilisés.

	Coût global	Coût journalier
- prévision systématique de Non-pluie	1648	2,68
- prévision systématique de Risque	1026	1,67
- prévision systématique de Pluie	1640	2,66

Figure 9 – Coûts affectés aux stratégies triviales.

Coût "sans risque" et "avec risque", la suppression de la notion de risque fait alors entrer les classes de ce type dans la catégorie "Non Pluie".

La méthodologie que nous avons décrite a été ensuite appliquée au fichier test. La méthode de décision a été la suivante. Pour chacune des classifications nous avons repris les coordonnées des étalons de chaque classe et nous avons affecté chaque élément du fichier à la classe dont il est le plus proche au sens de sa distance à l'étalon correspondant. Puis nous avons prévu *N*, *R* ou *P* pour chaque année selon la nature antérieure de la classe où elle se trouvait affectée.

Il est alors intéressant de représenter les 4 courbes suivantes (Fig. 13) de variation du coût journalier :

"avec risque" sur fichier d'apprentissage ;
"sans risque" sur fichier d'apprentissage ;
"avec risque" sur fichier test ;
"sans risque" sur fichier test ;

La superposition de ces 4 courbes montre, que le coût est en général minimum pour 5, 6 et 7 classes.

On notera que la classification en 7 classes tient son succès de ses 2 classes de risque, puisque "sans risque" elle présente des coûts élevés. D'autre part, ce sont les partitions en 5 et 6 classes qui chutent le moins lors du passage du fichier d'apprentissage au fichier test. Cela nous a incité, compte tenu également, de l'inconfort résultant de la prolifération des classes de risque, à choisir la partition en 6 classes comme structure prévisionnelle optimale (Fig. 10, 12, 13).

C'est donc une partition qui fournit une prévision stricte (*N* ou *P*) dans environ 80 % des cas avec un taux d'erreur inférieur en moyenne à 18 % et un risque justifié à 32 %.

La faible chute de la qualité observée lors du passage au fichier test s'explique par le fait, que la classification a pour but de cerner une réalité météorologique et non une structure particulière du fichier (Fig. 11).

Des études complémentaires ont été réalisées en adoptant plusieurs étalons par classes, puis en définissant une stratégie complexe pour laquelle chaque classification particulière est assimilée à un expert et la décision finale s'obtient par amalgame des décisions particulières concordantes. Dans ce cas le coût journalier devient 0,93 sur le fichier d'apprentissage.

Illustrons maintenant la question évoquée sur la multivocité des situations favorables ou défavorables à l'occurrence d'un phénomène météorologique. Le profil de la situation du 21/12/66 est plus proche de *P* que de *NP*, pourtant il n'a pas plu ce jour là. La méthode des profils moyens donne une prévision erronée. Or la classification en 4 classes l'affecte judicieusement à la classe numéro 1. (86 % de *N.P.*) (Fig. 14).

Pour une étude plus fine de la structure obtenue nous l'avons complétée par un renseignement supplémentaire portant sur les profils des jours moyens types et le type de situations météorologique concernée, décrite par les cartes de surface et d'altitude. Cela nous a conduit à définir les types suivants :

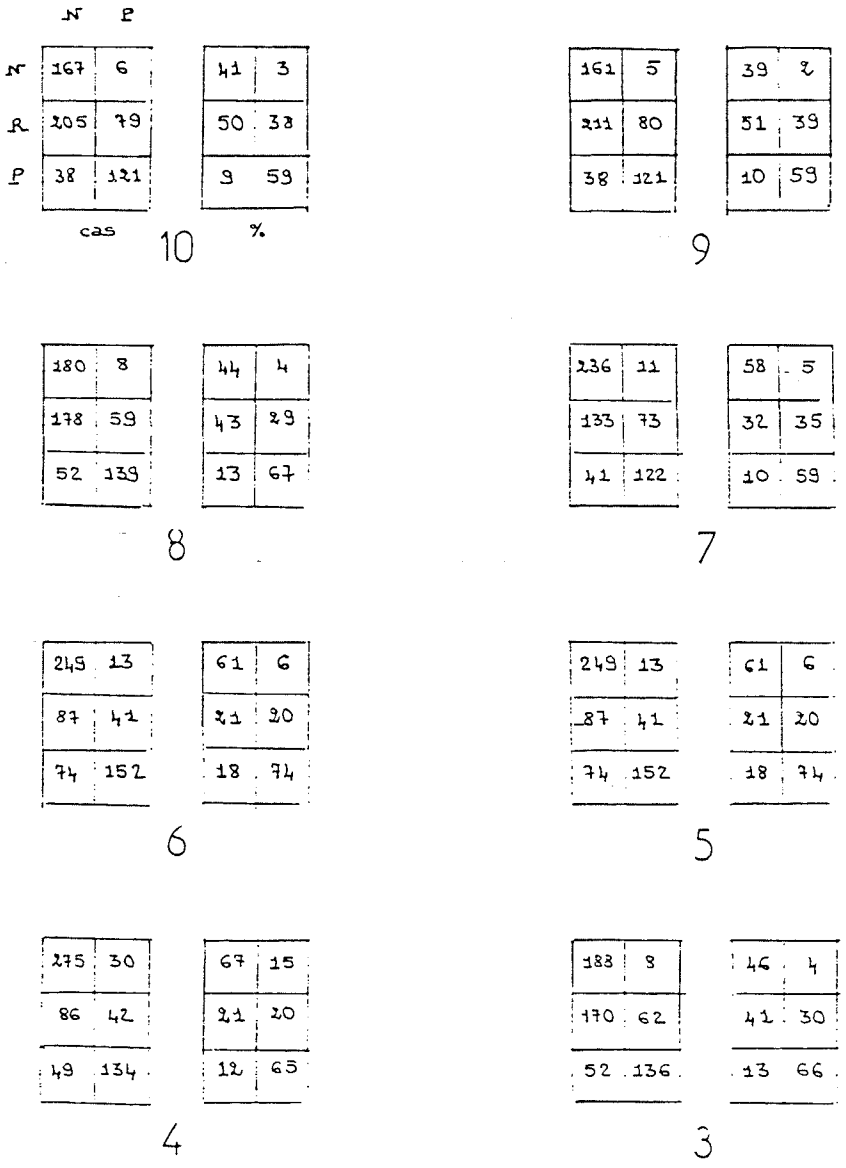
1) Situation anticyclonique stable d'hiver ;
2) Invasion d'air polaire continental ;
3) Entrée d'air polaire maritime ;
4) Instabilité convective ;
5) Situation anticyclonique stable d'été ;
6) Régime d'ouest perturbé de printemps et d'automne.

Il est naturellement gênant pour la météorologie synoptique d'avoir limité le nombre de classes à 6, mais cette restriction nous était imposée par le faible nombre de situations disponibles.

Toutefois la méthodologie proposée présente l'avantage d'améliorer l'applicabilité de l'approche statistique ; les décisions sont prises à l'intérieur de fichiers homogènes, alors que généralement on mélange toutes les situations, ce qui conduit à un fichier hétérogène et peu informatif.

De nombreux développements sont encore envisageables : augmentation de la taille du fichier et du nombre de classes, étude des sous-fichiers saisonniers, optimisation de la distance utilisée et des stratégies décisionnelles retenues.

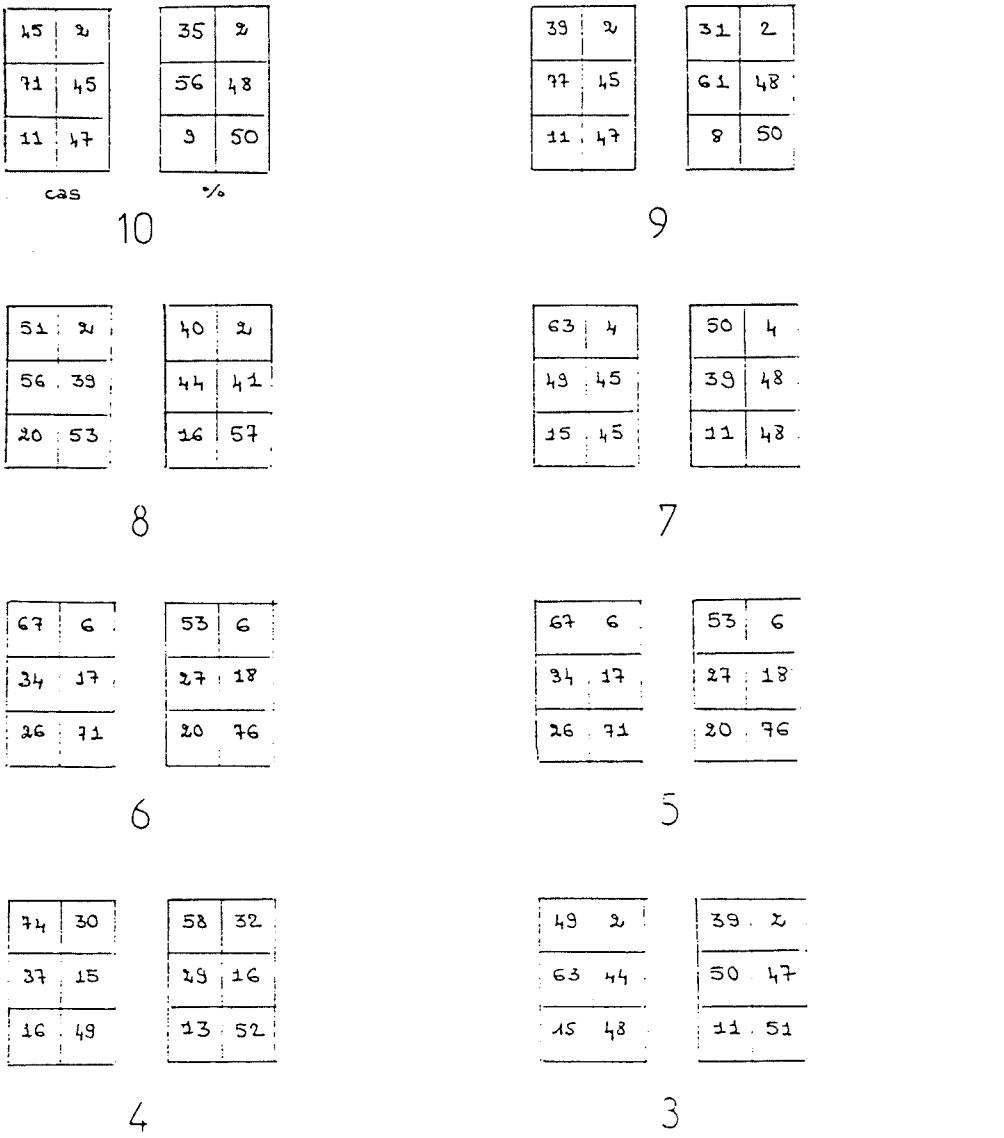
Les résultats encourageants que nous avons obtenus grâce à une méthodologie statistiquement mieux fondée et plus physique, plus proche de la décision synoptique, permettent d'envisager avec optimisme un débouché opérationnel pour un nombre de stations suffisamment important.



classifications

		3	4	5	6	7	8	9	10
coût global	sans R	768	772	728	728	836	744	832	832
	avec R	674	650	615	615	591	687	694	689
coût journalier	sans R	1,25	1,25	1,18	1,18	1,36	1,21	1,35	1,35
	avec R	1,09	1,06	1,00	1,00	0,96	1,12	1,13	1,12

Figure 10 – Qualité de la prévision sur fichier d'apprentissage.
10 *a* – Scores.
10 *b* – Coût global et journalier.



classifications

		3	4	5	6	7	8	9	10
coût global	sans R	428	424	288	288	452	408	420	420
	avec R	246	393	237	237	235	247	259	247
coût journalier	sans R	1,94	1,92	1,30	1,30	2,05	1,85	1,90	1,90
	avec R	1,11	1,78	1,07	1,07	1,06	1,12	1,17	1,12

Figure 11 – Qualité de la prévision sur fichier test.
11 *a* – Scores.
11 *b* – Coût global et journalier.

	ET	EN	EP	PN	PP	RN	RP
CLASSE 1	21	20	1	95.2%	4.8%	15.7%	1.1%
CLASSE 2	3	3	0	100.0%	0.0%	2.4%	0.0%
CLASSE 3	45	7	38	15.6%	84.4%	5.5%	40.4%
CLASSE 4	51	34	17	66.7%	33.3%	26.8%	18.1%
CLASSE 5	49	44	5	89.8%	10.2%	34.6%	5.3%
CLASSE 6	52	19	33	35.5%	63.5%	15.0%	35.1%
		127	94				

Figure 12 – Paramètres de la classification en 6 classes.

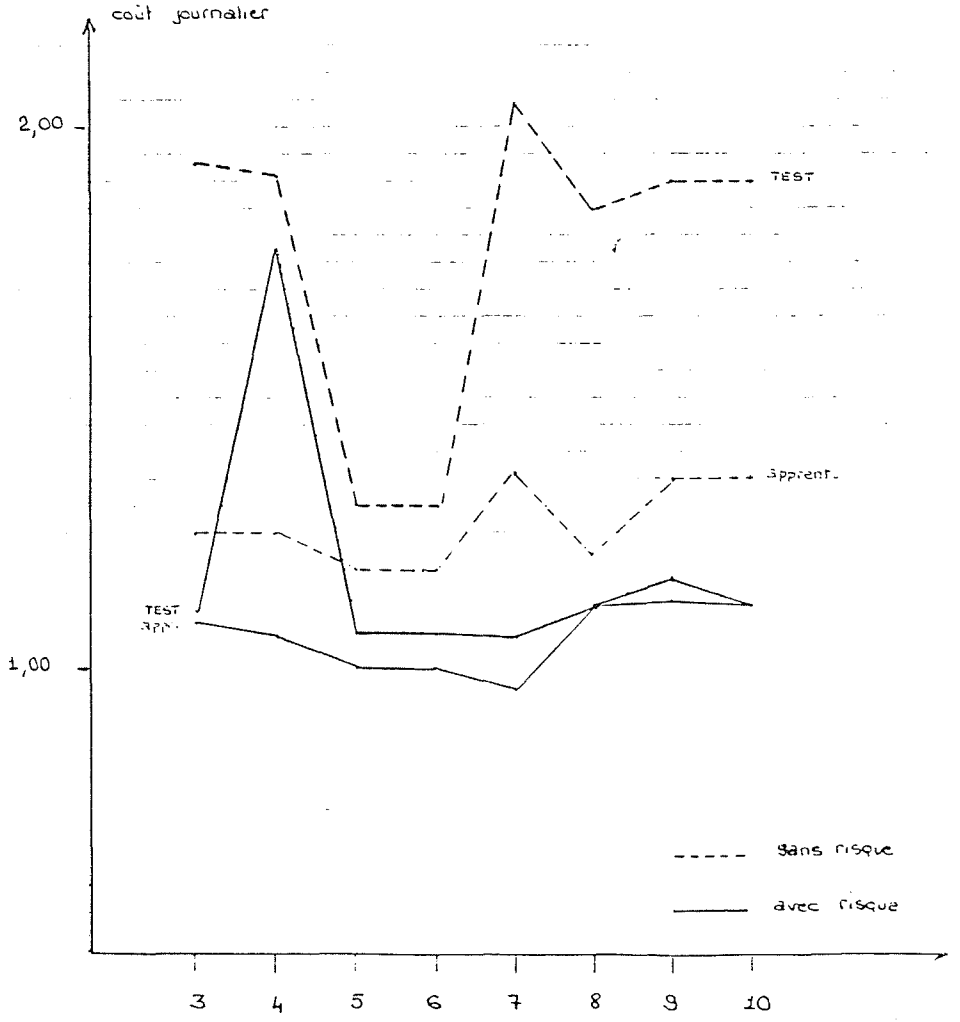


Figure 13 – Evolution des coûts en fonction du nombre de classes.

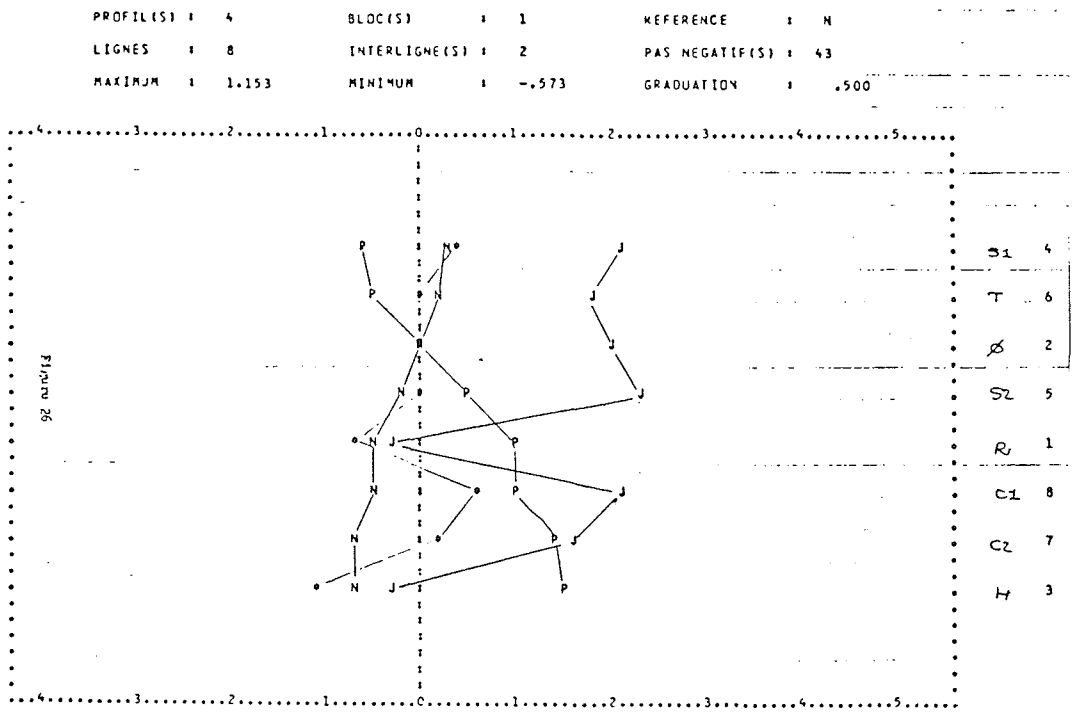


Figure 14 – Profils moyens pluie et non-pluie et profil de la journée du 26/12/66.

Bibliographie

[1] DER MEGREDITCHIAN G., RULLIERE M.C. – Classification automatique de la pluviométrie en France. *La Météorologie VI, Série 1*, juin 1975.

[2] DIDAY E. – Optimisation en classification automatique. *IRIA*, 1979.

[3] JAVELLE, DER MEGREDITCHIAN, CALVAYRAC, DUVERNET, VEYSSEIRE. – Operational forecasting of local weather parameters through statistical interpretation of *NWP* using "Perfect Prog" method. *WMO Symposium on probabilistic and statistical methods in weather forecasting.* Nice 1980.

[4] ROUSSEAU D. – Projet Améthyste. *Cahier n° 1, 2 et 7. Météorologie Nationale Note Technique n°* 81.

Discussion

Président : M. J. JACQUET

Le Président. – Merci M. LEGENDRE de nous avoir présenté cette approche systématique qui permet d'affiner la prévision locale du phénomène pluie. J'aimerais maintenant avoir des réactions d'utilisateurs opérationnels de prévisions, M. GUILLOT par exemple.

M. GUILLOT. – je n'ai pas pu suivre dans le détail la définition des classes. Avez-vous comparé avec des méthodes beaucoup plus frustes comme la persistance ?

M. LEGENDRE. – La comparaison a été faite avec la persistance. Le coût journalier de la persistance était de 1,25 contre un côut de 1 avec le modèle de classification à 6 classes. C'est une amélioration difficile à quantifier et le critère nous est personnel. Mais tout de même la méthode est meilleure que la persistance.

M. THIRRIOT. – Voilà plusieurs fois qu'on parle de la distance de Mahalanobis. Quelle est la définition de cette distance ?

M. DER MEGREDITCHIAN. – Mahalanobis est un Indien qui a utilisé une distance qui diffère essentiellement de la distance habituelle euclidienne par une pondération au moyen de la matrice de covariance inversée.

Le Président. – S'il n'y a pas d'autres questions, il me reste à remercier encore MM. DER MEGREDITCHIAN et LEGENDRE.