HAL
open science

Introduction d'outils de l'intelligence artificielle dans la prévision de pluie par radar

Andreas Neumann

To cite this version:

Andreas Neumann. Introduction d'outils de l'intelligence artificielle dans la prévision de pluie par radar. Hydrologie. Ecole Nationale des Ponts et Chaussées, 1991. Français. NNT : . tel-00520834

HAL Id: tel-00520834

https://pastel.hal.science/tel-00520834v1

Submitted on 24 Sep 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

79320

NS 15981(2)

Mémoire présenté pour l'obtention du titre du
Docteur de l'École Nationale des Ponts et Chaussées

"Sciences et Techniques de l'Environnement"

INTRODUCTION D'OUTILS DE L'INTELLIGENCE ARTIFICIELLE DANS LA PRÉVISION DE PLUIE PAR RADAR

ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES
CENTRE PÉDAGOGIQUE DE DOCUMENTATION ET DE COMMUNICATION

par

Andreas NEUMANN

Soutenue le 13 Décembre 1991

JURY

M. Guy JACQUET	Directeur de Thèse
M. Rémy POCHAT	Président
M. Jean-Gabriel GANASCIA	Rapporteur
M. Yves POINTIN	Rapporteur
M. Marc GILET	Examinateur
M. Isztar ZAWADZKI	Examinateur

E.N.P.C
DOC08404

06

Remerciements

Ma reconnaissance très chaleureuse va tout d'abord à Monsieur Guy JACQUET, mon directeur de recherche, avec qui j'avais le plaisir de travailler pendant les dernières trois années. Cette thèse doit infiniment à sa connaissance profonde du domaine et son soutien hors du commun.

Je tiens à remercier particulièrement Monsieur Rémy POCHAT, qui m'a très aimablement accueilli au CERGRENE et qui m'a fait l'honneur de présider mon jury.

J'adresse ma profonde gratitude à Monsieur le Professeur Jean-Gabriel GANASCIA et à Monsieur Yves POINTIN, qui ont accepter de prendre le temps pour revoir ce mémoire et être les rapporteurs de cette thèse.

Ma reconnaissance va également à Monsieur Marc GILET pour avoir agréé d'évaluer ce travail avec le regard du météorologiste.

Je souhaite exprimer ma gratitude toute particulière à Monsieur Istar ZAWADZKI, qui n'a pas hésité de participer à mon jury, et avec qui j'avais le plaisir de mener des discussions fructueuses lors de mes séjours à Montréal.

Je remercie vivement mon ami et "prédécesseur" au CERGRENE, Monsieur Thomas EINFALT, qui m'a initié aux problèmes de la prévision et qui était toujours présent pour me donner des conseils adéquats.

Les données radar utilisées dans cette étude ont été mises à disposition par la société RHEA, à laquelle j'adresse ma profonde gratitude.

Faute de pouvoir les citer tous, je tiens à remercier globalement tous les interlocuteurs, avec qui j'ai eu des discussions sur mon travail pendant ces dernières trois années, et qui m'ont souvent apportés des idées très fructueuses.

De la même façon j'exprime ma gratitude toute particulière à l'équipe du CERGRENE, qui m'a donné un cadre excellent pour la réalisation de cette thèse.

Enfin, je remercie tous ceux, qui m'ont aidé à passer à travers des bas-fonds de la langue française pendant la rédaction de ce mémoire, et qui ont souvent montré une énorme volonté et patience.

Résumé

L'objectif de l'étude présentée est le développement d'un système de prévision de pluie par radar, qui est adapté aux besoins de l'hydrologie urbaine. Un système automatisé structuré, baptisé PROPHETIA, est présenté, dont le fonctionnement est basé sur l'observation des cellules de pluie. L'algorithme de PROPHETIA de prévision de pluie à partir d'une série d'images $(I_1,..,I_n)$, mesurées aux instants $t_1,..,t_n$, comprend quatre étapes:

- identification et description des échos des cellules sur l'image actuelle I_n,
- appariement des cellules observées sur les images $I_1,..,I_{n-1}$ avec les échos sur l'image I_n,
- caractérisation des cellules dans l'intervalle (t_1,t_n),
- prévision de pluie par extrapolation des caractéristiques dans l'avenir.

Une technique de seuillage est appliquée pour l'identification des cellules. Pour leur appariement sur des images successives, une base de règles sous la forme d'un arbre de décision a été constituée par apprentissage automatique à partir d'exemples, qui ont été définis manuellement. La très bonne performance de la base de règles est mise en évidence par la comparaison avec les appariements manuels.

La prévision de PROPHETIA repose dans un premier temps sur la seule caractéristique de l'advection des cellules. Les résultats de cette prévision sont analysés selon un nouveau critère hydrologique, baptisé TMP. La qualité atteint par PROPHETIA est comparée à celle d'autres systèmes de prévision. PROPHETIA est surtout plus performante pour les pluies convectives.

L'examen détaillé des erreurs de la prévision par PROPHETIA a révélé que leur origine provient de l'hypothèse d'absence de développement des cellules à l'horizon de la prévision. L'étude des facteurs influant sur le développement des cellules a mené à la proposition d'un modèle des cellules reliant le développement aux masses d'air alimentant la cellule. La localisation du développement des cellules de pluie convective de la base de données est possible et apporterait un gain de prévision si le taux de ce développement pouvait être prédit, comme cela a été démontré pour un échantillon de 12 pluies convectives.

Or celui-ci dépend manifestement, comme l'étude des cycles de vie de quelques cellules l'a montré, de la possibilité de caractériser correctement les secteurs géographiques d'influence très différente sur la convection: une meilleure caractérisation de ces secteurs devrait être l'objectif qui suivrait celle-ci.

Mots-clés: Apprentissage automatique, arbres de décision, hydrologie urbaine, ID3, prévision de pluie par radar, reconnaissance de formes, traitement d'images.

Abstract

The work presented in this thesis aims at the development of a short-term radar rainfall forecasting system, that is adapted to use in urban hydrology. The system, named PROPHETIA, is based on rain cell tracking, in order to take individual advection into account. To forecast the rainfall at an instant t_n, based on the series of radar images $I_1,..,I_n$, the algorithm of PROPHETIA contains four main parts:

- identification and description of the rain cell echoes on image I_n,
- coupling of rain cells observed on the images $I_1,..,I_{n-1}$ with the echoes defined on image I_n,
- characterization of the rain cells in the interval (t_1,t_n),
- forecast of the rainfall by extrapolation of rain cell characteristics.

A threshold technique is used for echo identification. The coupling algorithm is based on the application of a decision tree, that has been generated by machine learning from human-defined examples. The high performance of this rule base is demonstrated by comparison with the human-defined couplings.

In the first part of the study, the forecast is based on rain cell advection alone. The performance of the system is analyzed with respect to a new hydrological evaluation criteria, named TMP. It is shown that the error rate of forecasts by PROPHETIA is lower than that of other techniques, especially for convective rainfall events.

The principal remaining forecast error is due to convective rain cell development during the forecast lead time. A model of this development is suggested, that allows for localizing the air masses alimenting the rain cells. Based on a statistical analysis of convective rain cells in terms of the model, a method is proposed, that takes rain cell growth/decay rates into account. It is shown, that a significant improvement of forecast quality can be achieved with this technique, if the development rates are known in advance.

In the last part of the study, the possibilities of forecasting the rain cell development are explored. An examination of a sample set demonstrates the necessity of taking local variations of the surface conditions into account. Possible solutions to this problem are discussed.

Key words: Decision tree induction, ID3, image processing, machine learning, pattern recognition, rain cell tracking, radar rainfall forecasting, urban hydrology.

Avant-Propos

"L'intelligence artificielle est tout, ce qui n'est pas encore réalisé"
Énoncé provocant de D.R. Hofstadter, dans "Gödel, Escher, Bach: an Eternel Golden Braid"

Est ce qu'une machine peut être intelligente? Un des premiers à avoir évoqué cette question était le fondateur de l'informatique moderne, A. Turing, lors d'un discours à Londres (1947, non publié, cité par Michie 1986). Il remarqua que sa conception d'une machine universelle inclut la possibilité, que la machine modifie elle-même son fonctionnement; et il conclut, qu'on est obligé de considérer ce comportement comme étant intelligent.

Aujourd'hui encore, l'utilisation du terme "intelligence" en association avec des machines provoque souvent des polémiques. Afin de clarifier cette discussion, il faut d'abord préciser ce qu'on comprend sous "intelligence". Le "Grand Larousse de la langue française" de 1971 propose la définition suivante:

> "(l'intelligence est) *au sens large (...) l'ensemble des fonctions de l'esprit ayant pour objet la connaissance. (...) Dans un sens restreint (...)* (elle est) *l'activité conceptuelle de l'esprit, par laquelle le sujet forme les idées et élabore les lois générales. (...)*"

Cette définition permet des interprétations bien différentes, jusqu'à celle prononcée par A. Binet, créateur de la psychométrie scientifique et des tests mentaux:

> *"l'intelligence, c'est ce que mesure mon test"*

Pour préciser la définition, il convient de regarder l'intelligence sous deux angles différentes: le point de vue philosophique, et le point de vue psychologique.

Dans la philosophie occidentale, l'intelligence est la fonction qui établit la supériorité de l'homme vis-à-vis d'autres êtres. Depuis l'antiquité, cette conception métaphysique a déterminé la façon, dont l'homme définit sa position dans l'univers. Dans ce siècle la création de nouveaux "êtres", des ordinateurs, nécessitait la rédéfinition du terme "intelligence", qui a été modifiée avec chaque progrès obtenu en informatique: des activités, qui ont été qualifiées comme intelligentes il y a peu de temps, sont ainsi souvent dévaluées à des simples algorithmes dont on nie le caractère intelligent. Ce point de vue, qui est très bien exprimé par la citation de Hofstadter, contraint alors à penser qu'une machine ne peut jamais être intelligente.

En psychologie, on ne peut pas se contenter d'une telle conception floue. Faute de pouvoir donner une définition générale et exacte, l'intelligence est accordée à des activités plus au moins indépendantes, dont l'"Encyclopaedia Universalis" de 1989 distingue six types:

- l'abstraction et la formation de concepts,
- le raisonnement inductif et la découverte de règles,
- le raisonnement déductif,
- la solution de problèmes,
- le raisonnement et la construction de systèmes formels,
- la gestion des activités cognitives.

Tous ces activités sont incontestablement dans un certain degré réalisables par des machines, qu'on doit par conséquent qualifier comme intelligentes. La compréhension de l'intelligence comme une liste de capacités précises amène Newell et Simon (1963) à l'énoncé que *"the free behaviour of a reasonable intelligent human can be understood as the product of a complex but finite and determinate set of laws"*. Dans les années 50 et 60, ce point de vue a mené a un esprit de l'"on peut tout faire" dans une grande partie de la communauté des chercheurs en intelligence artificielle, qui a souvent été critiqué. Weizenbaum (1976) par exemple remarque, que la liste des activités intelligentes doit être prolongée par la créativité, l'imagination, et l'intuition,

dont il est difficile d'imaginer la réalisation sur ordinateur. Il explique l'impasse, à laquelle ont abouti la quasi-totalité des recherches de systèmes modélisant le raisonnement humain d'une façon générale, par la négligence de ces aspects de l'esprit.

Depuis la fin des années 70, l'approche généraliste a cédé à l'intention plus prudente de développer des systèmes performants dans un domaine précis. L'étude présentée dans ce mémoire s'intègre dans cette axe, qui considère l'intelligence artificielle très pragmatiquement comme une science, qui cherche à formaliser et à modéliser des mécanismes intellectuels.

Table des Matières

Introduction générale 1

I La prévision de pluie par radar 4

I.1 Les systèmes de précipitation des zones tempérées 6

I.1.1 La précipitation provoquée par une ascendance frontale des masses d'air 7

I.1.2 La précipitation due à une ascendance convective 9

I.1.3 La précipitation de caractère mixte 10

I.2 L'observation de la pluie par radar 13

I.2.1 Les principes de la mesure de pluie par radar 13

I.2.2 Les sources d'erreur de la mesure de pluie par radar 14

I.2.2.1 Erreurs de mesure de la réflectivité 15

I.2.2.2 Erreurs introduites par la transformation de la réflectivité en intensité de pluie 16

I.2.2.3 Erreurs introduites par le calcul des lames d'eau 17

I.2.3 Caractéristiques des données utilisées dans cette étude 19

I.2.3.1 Caractéristiques techniques des données 19

I.2.3.2 Réduction de l'influence des erreurs de mesure 21

I.3 La prévision de pluie à des échéances courtes 24

I.3.1 Utilisation des prévisions de pluie en hydrologie urbaine 24

I.3.2 La prévision de pluie à l'aide des modèles numériques 24

I.3.3 La prévision de pluie par extrapolation des observations 25

I.3.4 Sources d'erreur de la prévision de pluie par extrapolation 28

I.4 Conclusion 30

II Méthodes de l'apprentissage automatique 31

II.1 Introduction à l'apprentissage automatique 33

II.1.1 Introduction d'une notion générale de l'apprentissage 33

II.1.2 Finalités de l'apprentissage automatique 35

II.1.3 La représentation symbolique de connaissances 36

II.1.4 Techniques d'apprentissage 42

II.2 Apprentissage inductif des arbres de décision 44

II.2.1 Génération et application des arbres de décision 44

II.2.1.1 Définitions 44

II.2.1.2 La génération des arbres de décision 46

II.2.1.3 Classification d'objets avec des arbres de décision 47

II.2.2 Sources des difficultés de l'emploi des arbres de décision pour résoudre des problèmes réels de classification 48

II.2.2.1 Difficultés liées à l'ensemble des attributs 48

II.2.2.2 Difficultés liées à l'ensemble des exemples d'apprentissage 51

II.3 Conclusion 56

III Proposition d'une méthodologie de l'apprentissage automatique d'une base de règles pour l'appariement des cellules de pluie sur l'image radar 57

III.1 L'identification des cellules de pluie sur l'image radar 59

III.1.1 L'ensemble des échos simples 59
III.1.2 L'ensemble d'échos imaginaires 60
III.1.3 Séquences strictes d'échos 61
III.2 Formulation d'un contexte de classification pour le problème de l'appariement d'échos 63
III.3 Définition des attributs pour le contexte de l'appariement 64
III.4 Définition de l'ensemble de l'apprentissage 65
III.5 Description de l'algorithme de l'apprentissage 68
III.5.1 Description de l'algorithme principal 68
III.5.2 Proposition d'une méthode de traitement de valeurs systématiquement inconnues 71
III.5.3 Proposition d'un critère de choix des attributs utilisant une plus grande profondeur de recherche dans l'arbre 73
III.5.4 Réflexion sur la complexité des algorithmes proposés 74
III.6 Algorithme de l'appariement par un arbre de décision 76
III.6.1 La classification de couples d'échos par un arbre de décision 76
III.6.2 L'appariement d'échos à l'aide d'un arbre de décision 76
III.7 Conclusion 79

IV Application de la méthodologie développée aux données radar de Trappes . . . 80
IV.1 Analyse des données radar utilisées 83
IV.1.1 Les événements de pluies choisis 83
IV.1.2 La définition des échos et l'appariement manuel 83
IV.1.3 La définition automatique des échos imaginaires 85
IV.2 Apprentissage automatique d'un arbre de décision pour l'appariement des échos . . 86
IV.2.1 Génération des ensembles d'exemples positifs et négatifs 86
IV.2.2 Induction d'un arbre de décision 87
IV.2.2.1 Comparaison des arbres générés par les algorithmes de IAD.O et IAD.S 87
IV.2.2.2 Choix d'un arbre de décision 89
IV.3 Application de l'arbre de décision sélectionné à l'appariement des échos 93
IV.3.1 Taux d'erreur de l'appariement 93
IV.3.2 Réflexion sur l'importance des résultats obtenus pour les objectifs de cette étude 97
IV.3.2.1 La prévision des lames d'eau 97
IV.3.2.2 L'observation du développement des cellules 97
IV.3.3 Comparaison des résultats obtenus avec ceux d'autres systèmes de prévision . . 98
IV.4 Conclusion 99

V La prévision de pluie basée sur la seule advection des cellules de pluie 100
V.1 Le système de prévision PROPHETIA 102
V.1.1 La caractérisation des séquences strictes d'échos 102
V.1.1.1 La définition de l'advection pour un seul pas de temps 102
V.1.1.2 L'extrapolation de l'advection des cellules de pluie 102
V.1.2 La prévision des lames d'eau 104
V.2 L'évaluation des prévisions 106
V.2.1 Critères de l'évaluation de la prévision par radar 106
V.2.2 Proposition du nouveau critère d'évaluation TMP 107
V.3 La prévision de pluie par le système PROPHETIA 110

V.3.1 Résultats des prévisions avec différentes règles de l'appariement 110
V.3.2 Analyse des sources d'erreurs de la prévision 112
V.3.2.1 Erreurs dues à un appariement incorrect des échos 119
V.3.2.2 Erreurs dues à une mauvaise description de l'advection des cellules 119
V.3.2.3 Erreurs dues à un développement local de la pluie 120
V.3.2.4 Erreurs dues à un développement global de la pluie 120
V.3.3 Comparaison de la performance de PROPHETIA avec celle d'autres méthodes de prévision 121
V.3.4 Examen des améliorations possibles par la modification du mode de définition des échos simples 126
V.3.5 Analyse de la complexité du système PROPHETIA 129
V.4 Conclusion 130

VI La prise en compte du développement des cellules de pluie convective pour la prévision de pluie 131
VI.1 Formulation du problème 133
VI.2 Examen des mécanismes du développement des cellules de pluie convective 135
VI.2.1 Facteurs influant sur le développement des cellules de pluie convective 135
VI.2.1.1 La convection comme source des précipitations 135
VI.2.1.2 L'influence des contrastes locaux sur la convection 137
VI.2.2 Proposition d'un modèle du développement des cellules de pluie convective . . . 138
VI.3 Étude de la répartition spatiale de la croissance/décroissance dans les cellules de pluie convective 140
VI.3.1 La croissance des cellules de pluie convective 140
VI.3.2 La décroissance des cellules de pluie convective 142
VI.4 Examen de l'amélioration possible de la prévision par prise en compte du développement des cellules de pluie convective 144
VI.5 La prévision du taux de croissance/ décroissance pour les cellules de pluie convective 149
VI.5.1 L'identification de régions favorables à l'intensification et à l'affaiblissement de la pluie 149
VI.5.2 L'extrapolation du développement observé des cellules 149
VI.5.3 Application de l'apprentissage automatique à la prévision du taux de croissance/décroissance des cellules 151
VI.5.4 La prise en compte des contrastes locaux pour la prévision du développement des cellules de pluie convective 151
VI.6 Conclusion 157

Conclusion générale 158

Références bibliographiques 161

Index terminologique, Glossaire 167

Annexes A-1
A.1 Définitions des attributs du contexte CT_{AP} A-2
A.2 Présentation du système de visualisation des images radar, d'appariement manuel et d'analyse des résultats de la prévision A-6
A.3 Présentation graphique des événements de pluie utilisés dans cette étude A-11

Figures

Figure I.1: Visualisation schématique des fronts d'une perturbation cyclonique extratropicale (d'après Triplet et Roche 1977) 7

Figure I.2: Influence des caractéristiques thermiques du secteur chaud sur le type de pluie déclenchée (d'après Triplet et Roche 1977) 8

Figure I.3: Modèle des flux de masses d'air autour d'une perturbation (d'après Harrold et Austin 1974) 9

Figure I.4: Durée de vie des cellules convectives (d'après Battan 1973) 10

Figure I.5: Mesure par radar d'une bande pluvieuse étroite 11

Figure I.6: Mesure par radar d'une bande pluvieuse large 11

Figure I.7: Mesure par radar d'une pluie convective 12

Figure I.8: Mesure par radar d'une pluie de caractère mixte 12

Figure I.9: Coupe verticale à travers une zone pluvieuse frontale, montrant six erreurs de mesure de l'intensité (d'après Browning et Collier 1982) 15

Figure I.10: Présentation schématique de l'erreur introduite par la conversion d'unités polaires de mesure de radar en unités quadratiques 16

Figure I.11: Hauteur du faisceau radar en fonction de la distance r et de l'angle d'élevation Θ (d'après Collier 1989) 18

Figure I.12: Schéma montrant la nécessité d'une prise en compte de l'advection dans l'intégration temporelle des images radar 18

Figure I.13: Fond de carte de la région couverte par les images du radar de Trappes 20

Figure I.14: Echos fixes sur l'image radar de Trappes (mesure du 10 juin 1989 à 10h04 TU) 23

Figure I.15: Qualité des prévisions en fonction des intervalles d'échéance, pour différentes techniques de prévision (d'après Browning 1980) 28

Figure I.16: Exemple d'une modélisation non-linéaire aggravant le taux d'erreur (a) et possibilité théorique d'amélioration d'erreur (b) (d'après Doswell 1986) 29

Figure II.1: Description des relations entre les objets du contexte CT_{EX} par un réseau de frames 37

Figure II.2: Neurone formel à n entrées et trois fonctions-types de neurones (d'après Lippmann 1987) 40

Figure II.3: Différentes architectures de réseaux de neurones: (a) réseau complètement connecté, (b) réseau en une couche, (c) réseau en multicouches (d'après Matheus et Hohensee 1987) 41

Figure II.4: Arbre de décision pour le problème de classification d'animaux en espèces non dangereuses (classe 0) et espèces dangereuses (classe 1) 45

Figure II.5: Arbre de décision du contexte CT_{EX}, dans lequel l'attribut $AGE(.)$ est transformé en attribut nominal 51

Figure II.6.a: Exemple d'un arbre de décision trop spécialisé, avec des feuilles déterministes 53

Figure II.6.b: Le même arbre réduit au noyau important, avec des feuilles probabilistes 53

Figure III.1: Définition d'échos pour les valeurs r_s=1 et d_{max}=1 (a) et r_s=2, d_{max}=1.5 (b) 60

Figure III.2: Représentation schématique du principe du suivi des cellules à l'aide des échos imaginaires 62

Figure III.3: Présentation schématique de l'influence des ensembles d'apprentissage sur la qualité de la règle induite 66

Figure III.4: Présentation schématique du flux des données dans l'algorithme incrémentelle de génération d'un ensemble d'exemples négatifs 67

Figure III.5: Sous-arbre construit par IAD.S dans le cas où une partie des valeurs de l'attribut est inconnue 71

Figure IV.1: Présentation schématique de la démarche poursuivi pour l'apprentissage d'un arbre de décision de l'appariement des échos et pour la vérification des algorithmes proposés 82
Figure IV.2: Nombre d'échos imaginaires générés par l'algorithme III.1 en fonction de la distance maximale (ensemble des 20 pluies) 85
Figure IV.3: Taux d'échos imaginaires retrouvés par l'algorithme III.1 en fonction du facteur c de la distance maximale (ensemble des 20 pluies) 85
Figure IV.4: Arbre de décision initial ADD_{INI} de l'algorithme de définition de l'ensemble des exemples négatifs (les seuils souples sont indiqués entre parenthèses) 86
Figure IV.5: Génération de l'ensemble des exemples négatifs par l'algorithme III.2 87
Figure IV.6: Connaissance des valeurs des attributs pour l'ensemble de l'apprentissage EX 87
Figure IV.7: Taux d'erreur maximal, minimal et moyen de la classification des ensembles d'exemples par les arbres générés 88
Figure IV.8: Distribution des probabilités estimés de l'appartenance à la classe positive des exemples de test (classification par l'arbre ADD_{APP}) 90
Figure IV.9: Le meilleur arbre de décision $ADD.O=ADD_{APP}$ généré par IAD.O 91
Figure IV.10: Le meilleur arbre de décision *ADD.S* généré par IAD.S 92
Figure V.1: Présentation schématique du flux des données entre les différents systèmes développés 101
Figure V.2.a: Différence de la vitesse entre différentes vecteurs antérieurs et le vecteur postérieur de 60 minutes $adv_{12}(.)$ 103
Figure V.2.b: Différence de la direction entre différentes vecteurs antérieurs et le vecteur postérieur de 60 minutes $adv_{12}(.)$ 103
Figure V.3: Durée de vie des séquences strictes de la base des données 104
Figure V.4.a: Influence de la taille des bassins de l'évaluation sur le taux d'erreur (TMP moyen de 20 pluies) 108
Figure V.4.b: Influence du seuil de l'intensité de l'évaluation sur le taux d'erreur (TMP moyen de 20 pluies) 108
Figure V.5: Courbes d'efficacité de la prévision avec différentes méthodes d'appariement (prévisions de 60 minutes) 111
Figure V.6: Taux d'erreur de la prévision avec les appariements manuels comparé aux taux de la prévision avec les deux arbres de décision (prévisions de 60 min) 111
Figure V.7.1-
Figure V.7.20: Taux d'erreur de la prévision par PROPHETIA 113
Figure V.8: Taux d'erreur de PROPHETIA comparé au taux d'erreur des autres techniques, et ligne de TMP égale ($\Delta_p t$ = 60 min) 121
Figure V.9.a: Courbes d'efficacité pour les pluies convectives ($\Delta_p t$ = 60 min) 123
Figure V.9.b: Courbes d'efficacité pour les pluies frontales ($\Delta_p t$ = 60 min) 123
Figure V.10: Différence entre TMP de CORRCROIS et TMP de PROPHETIA en fonction de la taille moyenne des échos ($\Delta_p t$ = 60 min) 124
Figure V.11: Explication de l'erreur de la prévision pour la pluie du 21.9.1990. La région parisienne est indiquée par la ligne en tirets. Présentés sont les isohyètes de 2,10,15 et 20 mm/h. 125
Figure V.12: Amélioration du taux d'erreur par définition multiple des échos pour cinq pluies frontales 128
Figure VI.1: Développement de la masse d'une cellule convective (cellule de la pluie du 6.6.1989) 133
Figure VI.2: Modèle schématique d'une cellule convective, présentant aussi les vents autour du système (d'après Chalon 1978) 136
Figure VI.3: Présentation schématique du modèle de développement des cellules de pluie convective 139
Figure VI.4: Les différentes zones autour d'une cellule croissante 140

Figure VI.5.a: Surface des quatre zones comme pourcentage de la surface totale (moyenne +/- écart-type) 141
Figure VI.5.b: Répartition de la masse augmentée sur les quatre zones (moyenne +/- écart-type) . . . 141
Figure VI.6: Les différentes zones autour d'une cellule décroissante 143
Figure VI.7.a: Surface des deux zones comme pourcentage de la surface totale (moyenne +/- écart-type) 143
Figure VI.7.b: Répartition de la masse diminuée sur les deux zones (moyenne +/- écart-type) 143
Figure VI.8.a-
Figure VI.8.l: Taux d'erreur de la prévision par PROPHETIA.II 145
Figure VI.9: Courbes d'efficacité de la prévision avec et sans prise en compte du développement (pluies convectives, ($\Delta_P t$ = 60 min) 148
Figure VI.10.a: *tcd* sur 5 minutes avant et après t_0 pour 492 cellules de pluie convective 150
Figure VI.10.b: *tcd* sur 10 minutes avant et après t_0 pour 492 cellules de pluie convective 150
Figure VI.10.c: *tcd* sur 15 minutes avant et après t_0 pour 492 cellules de pluie convective 150
Figure VI.11: Les régions de l'extension spatiale des mesures météorologiques au sol 152
Figure VI.12.a-
Figure VI.12.h Etude du cycle de vie des cellules de pluie convective 155

Tableaux

Tableau I.1: Echelle de réflectivité des données radar utilisées dans cet étude (échelle CALAMAR, d'après Agostini-Blanchet 1988) 19

Tableau I.2: Spécifications techniques du radar Rodin de Trappes 19

Tableau I.3: Prise en compte de principales sources d'erreur de la mesure par radar 22

Tableau I.4: Intervalle maximal d'extrapolation pour quelques types de pluie (d'après Zipser 1990) 29

Tableau II.1: Valeurs d'attributs des objets du contexte de classification d'animaux 34

Tableau II.2: Comparaison de quatre techniques de représentation de connaissances 41

Tableau IV.1: Pluviométrie en Ile de France par rapport à la normale des années 1951-1980 (source: La Météorologie, série VII, n 28-30, 33-35) 83

Tableau IV.2: Caractéristiques des données sélectionnées 84

Tableau IV.3: Performance des deux techniques de génération d'arbres de décision 88

Tableau IV.4: Taux d'erreur des deux arbres sélectionnés et taux d'erreur minimaux des 50 arbres générés par chaque algorithme 89

Tableau IV.5: Tableau d'évaluation des appariements effectués par l'algorithme III.7 93

Tableau IV.6: Performance de l'algorithme III.7 avec la règle initiale de l'apprentissage ADD_{INI} et avec la règle générée par l'apprentissage ADD_{APP} 95

Tableau V.1: Analyse des sources principales des erreurs de prévision, provoquant des surestimations (+) et des sous-estimations (−) des lames d'eau 118

Tableau V.2: Comparaison des taux d'erreur de quatre méthodes de prévision 122

Tableau V.3: Valeurs moyennes et maximales du temps de calcul du système PROPHETIA pour des prévisions de 60 minutes 129

Tableau VI.1: Comparaison des taux d'erreur de la prévision sans et avec prise en compte du développement avec la connaissance exacte et erronée du *TCD* 148

Algorithmes

Algorithme I.1: Prévision automatisée non structurée 26
Algorithme I.2: Prévision automatisée structurée 27
Algorithme II.1: Moteur d'inférences en chaînage avant 38
Algorithme II.2: Algorithme de ID3 de génération d'un arbre de décision 45
Algorithme II.3: Algorithme de classification d'objets par un arbre de décision 47
Algorithme II.4: Induction d'arbres de décision avec fenêtrage d'exemples 51
Algorithme III.1: Algorithme de définition hiérarchique d'échos imaginaires 61
Algorithme III.2: Algorithme incrémental de génération d'un ensemble d'exemples négatifs d'apprentissage 67
Algorithme III.3: Algorithme de IAD.O de génération d'un arbre de décision 70
Algorithme III.4: Le traitement des valeurs inconnues dans l'algorithme IAD.S, qui est basé sur IAD.O 72
Algorithme III.5: Algorithme du choix d'attribut de test de IAD.L avec une profondeur de k, basé sur IAD.O 73
Algorithme III.6: Algorithme de classification d'objets dont des valeurs d'attributs peuvent être inconnues 77
Algorithme III.7: Algorithme de définition et d'appariement d'échos de pluie sur l'imagerie radar 78
Algorithme V.1: Algorithme de prévision des lames d'eau de PROPHETIA 105

INTRODUCTION GÉNÉRALE

Dans les dernières deux décennies, les questions relatifs à la protection de l'environnement ont rencontré une sensibilité croissante dans les pays industrialisés. La dégradation de la qualité des eaux de surface a été reconnue comme un des problèmes le plus menaçantes pour les milieux naturels. Cette dégradation concerne directement la santé publique: par exemple proviennent en région parisienne 98% de l'eau potable des eaux superficiels (Herremans 1990).

La pollution des eaux est originaire de trois sources principales: des effluents industriels, des effluents domestiques, et de la pollution diffuse agricole. Les eaux pluviales ont un impact sur l'environnement surtout dans les zones urbanisées: par les ruissellements directs, qui amènent les substances toxiques des surfaces imperméabilisées, et par le débordement des réseaux unitaires provoqué par les fortes pluies. Afin de limiter la décharge des eaux usées des réseaux unitaires en périodes de pluie, des techniques automatisées de la gestion des réseaux d'assainissement en fonction de son état ont été développées, qui permettent l'exploit des éléments de stockage d'une façon optimisée.

Une amélioration considérable de la gestion peut être obtenue, si l'on dispose d'une prévision à court terme des débits, qui est rendu possible par la modélisation informatisée du réseau. Le paramètre déterminant d'entrée étant la pluie, sa prévision avec une résolution spatiale et temporelle assez fine apporte un gain supplémentaire important. Car la mesure de la pluie par radar offre un moyen unique d'obtenir une telle prévision, les méthodes de la prévision de pluie par radar ont été sujet de nombreuses recherches dans les années 70 et 80, qui continuent à ce jour.

L'étude présentée dans ce mémoire s'intègre dans cet axe de recherche. Depuis la disponibilité des mesures radar en France au début des années 80, le CERGRENE a accentuée la recherche de l'exploit des mesures, en collaboration avec la Météorologie Nationale de France et le département de Seine-St.Denis. Les travaux de Andrieu (1986), Agostini-Blanchet (1988), Einfalt (1988), Denoeux (1989), et beaucoup d'autres ont apportés une grande expertise dans ce domaine. L'intégration opérationnelle du système de prévision SCOUT (Einfalt et *al.* 1990), et du système de l'estimation à priori de la fiabilité de la prévision (Denoeux et *al.* 1990), dans le système de gestion en Seine-St.Denis témoignent du succès de ces efforts.

Néanmoins, les études menées ont révélé des problèmes importants, qui nous ont amenée au travail présentée ici. La qualité de la prévision, et ainsi son utilité, dépend de la capacité du système de suivre correctement les structures météorologiques sur l'image radar. L'approche heuristique proposée par Einfalt (1988) présente des insuffisances dans certaines situations, que nous essayerons a dépasser. L'intérêt de développer un approche plus systématique à ce problème a été souligné récemment par les modifications dans le traitement des données radar par la Météorologie Nationale en 1988 et 1991.

Le premier objectif de cette étude est l'observation correcte des phénomènes météorologiques sur l'image radar, à une échelle utile pour la prévision. Le problème central de cette observation est la reconnaissance des formes des cellules de pluie sur des images consécutives, afin de permettre leur suivi lorsqu'elles traversent la région observée. La reconnaissance des formes est un des axes principaux de la recherche en intelligence artificielle, dont la plus grande partie des travaux se concentre sur la vision automatique, avec des applications surtout en robotique. Comme l'image vidéo d'objet réels, l'image radar présente une modélisation simplifiée de phénomènes réels très complexes. L'identification des objets, et leur observation lorsque leur caractéristiques évoluent, est un problème commun aux deux domaines. Pour sa résolution, nous tenterons un approche de l'apprentissage automatique, qui permet la découverte de régularités qui ne sont pas évident à l'observateur humain.

La prévision de la pluie par radar consiste en l'extrapolation des observations faites. La validité de l'extrapolation dépend du rapport entre le temps d'échéance de la prévision, et la durée de vie des structures observées. Le développement de la pluie pendant l'intervalle d'échéance est reconnu comme la source principale d'erreur de cette méthode de prévision. Ce développement est à très petite échelle un processus chaotique, qui présente cependant à une échelle plus grande des

régularités, que l'observateur humain arrive à interpréter et, s'il dispose d'une certaine expérience, à prédire. Le deuxième objectif de cette étude est le développement d'une méthodologie, qui permet de révéler les régularités de ce processus et de formaliser leur application pour la prévision de la pluie.

Nous avons pris soin de formuler les problèmes et leurs résolutions proposées dans une forme compacte et précise. D'où la nécessité d'introduire une nouvelle notion pour certaines aspects de cette étude. Cette présentation a été choisie pour deux raisons: premièrement, elle permet la comparaison des différentes méthodes à un niveau abstraite; et deuxièmement, elle facilite le transfert de la méthodologie développée vers d'autres applications.

Ce mémoire est organisé de la façon suivante:

Dans le **premier chapitre** nous introduisons le lecteur dans le domaine de la prévision de pluie et nous concrétisons les deux objectifs de l'étude. Après avoir exposé des différentes phénomènes météorologiques qui nous intéressent, nous examinons la technique de la mesure par radar et les sources d'erreur de cette technique. Ensuite, nous étudions les différentes méthodes existantes de prévision sous l'angle de leur utilité pour atteindre les objectifs fixés.

Le **deuxième chapitre** présente une introduction dans l'apprentissage automatique. Une notion générale de l'apprentissage sera élaborée, qui permet la comparaison objective des techniques principales existantes. La représentation de connaissances sous forme d'arbres de décision sera étudiée de façon détaillé.

Dans le **troisième chapitre**, nous développons une méthodologie pour l'observation des structures sur l'image radar. Nous proposons des méthodes de l'identification et de la description des structures, et nous formalisons le problème de leur observation. Pour les problèmes spécifiques à ce contexte, nous proposons des résolutions originales.

La méthodologie proposée sera appliquée dans le **quatrième chapitre** aux données radar de Trappes. Les algorithmes préconisés seront vérifiés et les résultats seront critiqués. Un arbre de décision sera généré par apprentissage à partir d'exemples, qui servira comme la base de connaissances d'un système de prévision.

Dans le **cinquième chapitre** nous élaborons le système de prévision PROPHETIA, qui est basé sur la seule extrapolation de l'advection de la pluie. Après avoir proposé un critère d'évaluation, la performance du système sera critiquée et les sources d'erreurs seront examinées.

Des problèmes relatifs au deuxième objectif de cette étude seront examinés en **sixième chapitre**. Le développement de la pluie sera analysé, et les possibilités de son prise en compte pour la prévision seront examinés.

I

LA PRÉVISION DE PLUIE PAR RADAR

Ce premier chapitre a pour but de présenter les divers problèmes qui ont conduit à cette étude. Dans un premier temps sont exposés les différents phénomènes météorologiques provoquant de la précipitation. Les mécanismes atmosphériques à l'origine de ces phénomènes sont décrites, et l'importance des caractéristiques des phénomènes pour la prévision est examinée.

L'observation de la précipitation par radar fait l'objet de la deuxième partie. Nous exposerons le principe de cette technique ainsi que les erreurs qui influent sur la qualité des mesures. Ensuite, les caractéristiques des données utilisées dans cette étude seront décrites.

En troisième partie, les objectifs accessibles pour une prévision de pluie seront déterminés. Nous comparerons les techniques de prévision existantes sous l'angle des difficultés qu'elles présentent pour atteindre ces objectifs.

I.1 Les systèmes de précipitation des zones tempérées

Les phénomènes météorologiques provoquant des précipitations nous intéressent particulièrement dans les aspects ayant une influence sur la prévisibilité des observations, comme la dynamique et la mesurabilité des structures pluviales.

A l'origine de la pluie il y a des mouvements de l'atmosphère qui se manifestent à des échelles de différents ordres de grandeur. La connaissance des phénomènes varie selon l'échelle à laquelle leurs observations sont faites. En météorologie, on distingue principalement trois échelles:

- L'***échelle synoptique*** ou macroéchelle qui est l'échelle globale. Le maillage est donné par la description de l'état de l'atmosphère par des observations météorologiques régulières. Ces observations sont effectuées à travers le globe par radiosondages et par satellite. La grille horizontale est de quelques centaines de kilomètres.
- L'échelle moyenne ou ***mesoéchelle***, son maillage est compris entre un kilomètre et quelques centaines de kilomètres.
- La petite échelle ou ***microéchelle*** qui est une échelle assez fine pour décrire les processus physiques locaux dans l'atmosphère. Son maillage s'étend entre quelques mètres et quelques centaines de mètres.

Les mesures par radar météorologique ont généralement une résolution et un rayon d'observation qui correspond à la borne inférieure de la mesoéchelle.

A l'origine d'un déclenchement de la pluie il y a la condensation de la vapeur d'eau dans l'atmosphère. De tous les processus conduisant à la condensation, seule l'ascendance des masses d'air peut engendrer la pluie. Un nuage pluvieux est la manifestation visible d'un système dynamique, dans lequel une multitude de processus thermodynamiques a lieu de manière chaotique (Lovejoy 1984). Dans un tel système l'air humide des basses couches de l'atmosphère est amené en altitude, où la détente de la masse d'air engendre une croissance de l'humidité relative jusqu'à la saturation. A un certain niveau au-dessus de la saturation de l'air, la vapeur d'eau commence à se condenser. Des gouttelettes d'eau naissent et s'agrandissent selon les conditions locales conduisant à la précipitation.

Bien que ce mécanisme ne soit pas parfaitement connu, on peut constater que la quantité d'eau précipitée dépend, entre autres, de la hauteur des nuages et de la vitesse de l'ascendance. Ainsi des études ont montré que dans 97% des cas, le sommet des nuages pluvieux atteint une température négative (étude cité par Triplet et Roche 1977).

L'ascendance de l'air étant identifiée comme le processus principal contrôlant la pluie, on distingue trois types de pluie selon le mécanisme à l'origine de l'ascendance (cf. Viers 1968, Harrold et Austin 1974, Sauvageot 1982):

- la pluie provoquée par une ascendance frontale,
- la pluie provoquée par une ascendance convective,
- la pluie provoquée par des effets orographiques.

Vu la faible importance des effets orographiques pour les pluies étudiées dans cette étude, nous nous contentons d'examiner les deux premiers effets.

I.1.1 La précipitation provoquée par une ascendance frontale des masses d'air

Le modèle classique des perturbations de la zone tempérée, la théorie norvégienne, a été développé au début du 20e siècle. Bien que ce modèle présente une vue très simplificatriste de l'atmosphère, il permet néanmoins la compréhension des phénomènes climatologiques de grande échelle (échelle synoptique), qui se manifestent dans les latitudes moyennes sous la forme de perturbations cycloniques extratropicales. Dans ce modèle, les masses d'air sont séparées par des ***fronts***, surfaces séparant une masse d'air relativement chaud d'une masse d'air relativement froid. Si la masse d'air froid se trouve à l'avant par rapport au déplacement d'ensemble, on parle d'un ***front chaud***, dans le cas contraire d'un ***front froid***. Front chaud et front froid sont associés à des dépressions cycloniques, qui, dans la forme simple, comprennent un secteur chaud inséré entre deux secteurs froids. Dans le cas où les deux fronts s'unissent, on parle d'un ***front occlus*** (figure I.1).

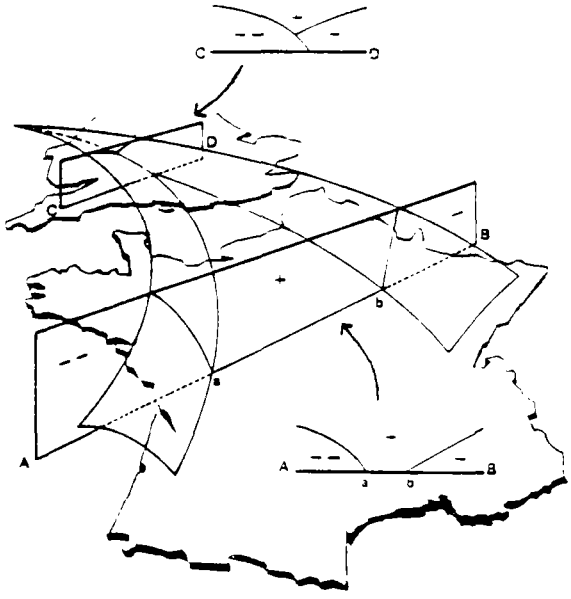


Figure I.1: Visualisation schématique des fronts d'une perturbation cyclonique extratropicale (d'après Triplet et Roche 1977)

Les surfaces séparant les différentes masses d'air ont une inclinaison d'environ 1‰ pour le front chaud et de 1% pour le front froid. Cette pente et la différence de densités des masses d'air provoquent un soulèvement de l'air chaud le long des surfaces frontales, ce qui amène à une formation de nuages frontaux. Le type de ces nuages et de la précipitation éventuellement déclenchée dépendent des caractéristiques thermodynamiques des masses d'air, notamment des caractéristiques du secteur chaud. Si l'air chaud est convectivement stable, la précipitation provoquée sera de type stratiforme et d'une intensité faible, tandis que l'air chaud instable donne lieu à des averses, souvent encastrées dans des régions de pluie stratiforme (figure I.2)

Les développements des techniques d'observation de l'atmosphère dans les deux dernières décennies, notamment le radar et le satellite, ont permis de créer de nouveaux modèles de perturbations à une échelle plus fine. Ils sont basés sur des flux d'air circulant autour d'une perturbation cyclonique (Harrold et Austin 1974, Browning 1985). D'après ces modèles il semblerait

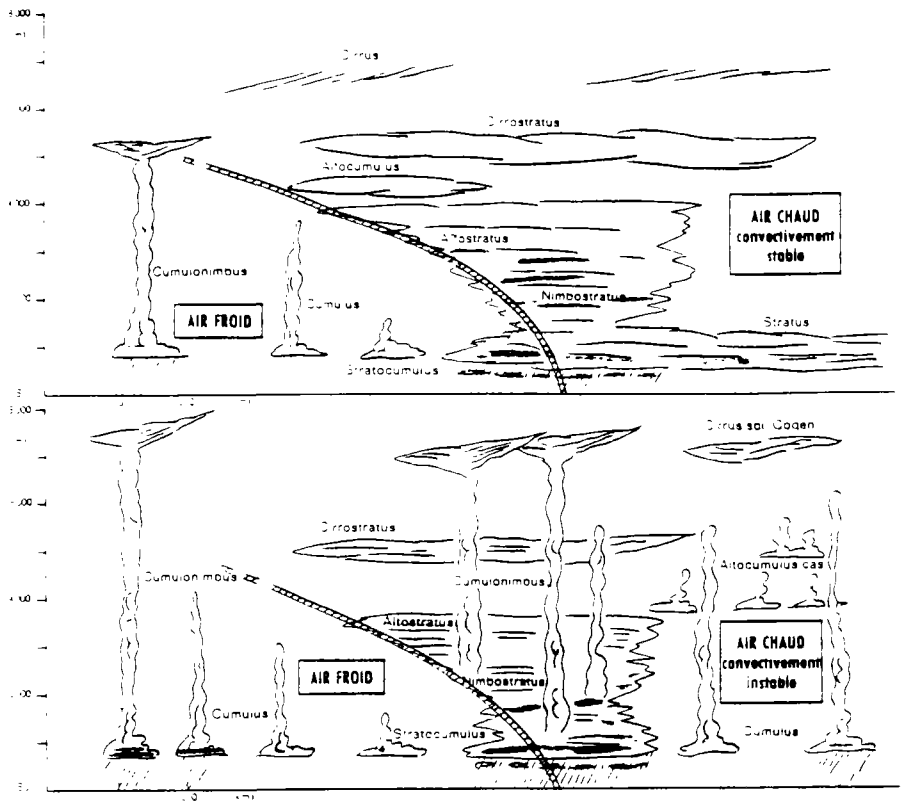


Figure I.2: Influence des caractéristiques thermiques du secteur chaud sur le type de pluie déclenchée (d'après Triplet et Roche 1977)

que la plus grande partie de la précipitation se forme dans le flux de la ceinture chaude ("conveyor belt"), flux d'une masse d'air chaud et humide qui monte en avant du front froid (figure I.3). Si ce flux tourne en altitude en arrière du front froid, on parle d'une situation ***anabatique***, s'il tourne en altitude vers l'avant du front, d'une situation ***katabatique***. Les flux qui déterminent la structure de la précipitation d'un tel système sont montrés dans figure I.3, dans laquelle la pluie stratiforme est désignée par les régions pointillées, tandis que les triangles indiquent des averses.

Ces systèmes frontaux sont à l'origine de précipitations de types très différents. D'après Browning (1985), la structure de la pluie dans le même système n'est guère uniforme. Néanmoins, on peut distinguer deux catégories principales.

Les ***bandes pluvieuses étroites***, associées aux fronts froids anabatiques, ont une extension de quelques kilomètres de largeur et de quelques centaines de kilomètres de longueur. Elles sont souvent imbriquées dans de vastes zones de précipitation faible. Le déplacement de la bande pluvieuse est identique à celui du front au sol. La durée de vie d'une telle structure s'étend sur plusieurs heures. L'intensité de la précipitation peut être très forte (figure I.5).

Les ***bandes pluvieuses larges*** sont associées à l'ascension lente de la ceinture chaude en avant d'un front froid katabatique et se déplacent avec la perturbation cyclonique. La dimension typique est de quelques dizaines de kilomètres de largeur et quelques centaines de kilomètres de longueur, avec une durée de vie d'une dizaine d'heures. La précipitation est généralement faible, mais des cellules intenses de courte durée de vie peuvent se développer à l'intérieur de ces bandes (figure I.6).

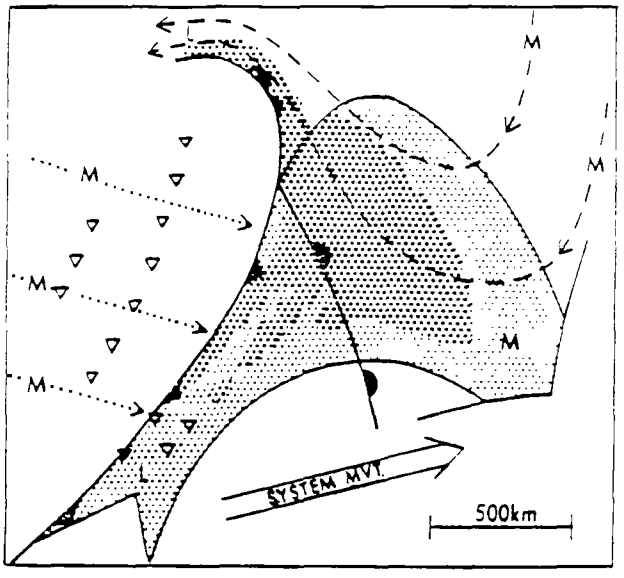


Figure I.3: Modèle des flux de masses d'air autour d'une perturbation (d'après Harrold et Austin 1974)

I.1.2 La précipitation due à une ascendance convective

Les ascendances convectives se produisent dans une masse d'air verticalement instable. L'air relativement chaud et humide du sol est ainsi rejeté en altitude, ce qui entraîne la saturation et la condensation d'une partie de la vapeur d'eau. La précipitation est d'autant plus importante, que l'instabilité de l'air est forte.

La pluie provoquée uniquement par des ascendances convectives possède un caractère cellulaire, les cellules ayant une surface de quelques kilomètres carrés (Harrold et Austin 1974, Austin et Houze 1972) (figure I.7). En conséquence, elles ne sont observables qu'à une échelle plus petite que l'échelle synoptique. Souvent les cellules sont organisées en structures plus grandes allant de quelques dizaines de kilomètres carrés à quelques milliers de kilomètres carrés (orages, lignes de grains).

En raison de la caractéristique locale du développement de la pluie convective, la durée de vie des cellules individuelles est souvent limitée à des intervalles inférieurs à une heure, comme le montre les analyses d'observations faites par radar (Battan 1973, Lopez et *al.* 1984) (figure I.4). Les structures, dans lesquelles les cellules sont imbriquées peuvent cependant souvent être observées pendant plusieurs heures.

Le déplacement des cellules de pluie convective est influencé et par le vent en altitude, et par leur développement. Il est donc parfois difficilement prévisible.

En raison des difficultés qui posent le fort développement et le déplacement hétérogène des cellules de pluie convective pour la prévision de la pluie, le dernier chapitre de cette étude sera consacré à une analyse de ces aspects.

Distribution of Duration of Radar Clouds for Different Sizes of Echoes

Duration (Min)	Maximum Horizontal Dimension of Echo (Km)								
	0–1.5	1.6–3.1	3.2–4.7	4.8–6.3	6.4–7.9	8.0–9.5	9.6–11.1	11.2–12.7	Total
0.0–4.9	....	....	....	....	....	....	....	....	....
5.0–9.9	1	2	....	....	....	....	....	....	3
10.0–14.9	1	3	6	....	....	....	....	....	10
15.0–19.9	1	9	5	....	....	....	....	....	15
20.0–24.9	...	7	6	4	1	....	....	....	18
25.0–29.9	....	2	5	1	....	....	....	....	8
30.0–34.9	....	....	5	....	....	....	....	....	5
35.0–39.9	....	....	1	1	2	....	....	....	4
40.0–44.9	....	....	1	....	....	....	....	1	2
45.0–49.9	....	....	....	....	....	1	....	....	1
50.0–54.9	....	....	....	....	....	1	....	....	1
Total	3	23	29	6	3	2	....	1	67

Figure I.4: Durée de vie des cellules convectives (d'après Battan 1973)

I.1.3 La précipitation de caractère mixte

On observe plus souvent des précipitations de caractère mixte, où les cellules convectives sont imbriquées dans des zones larges de pluie frontale, que des précipitations de caractère uniquement frontale ou convective. La figure I.8 montre la mesure d'une zone pluvieuse associée à un front chaud, à l'intérieur de laquelle se trouvent des cellules intenses, provoquées, dans le cas présent, par une convection locale déclenchée par la présence d'une forêt, source d'humidité au sol. La durée de vie de la grande structure étant déterminée par la perturbation, les cellules imbriquées ont une durée de vie limitée comme dans le cas de cellules purement convectives. Le déplacement des cellules intérieures est souvent différent du déplacement de la grande zone.

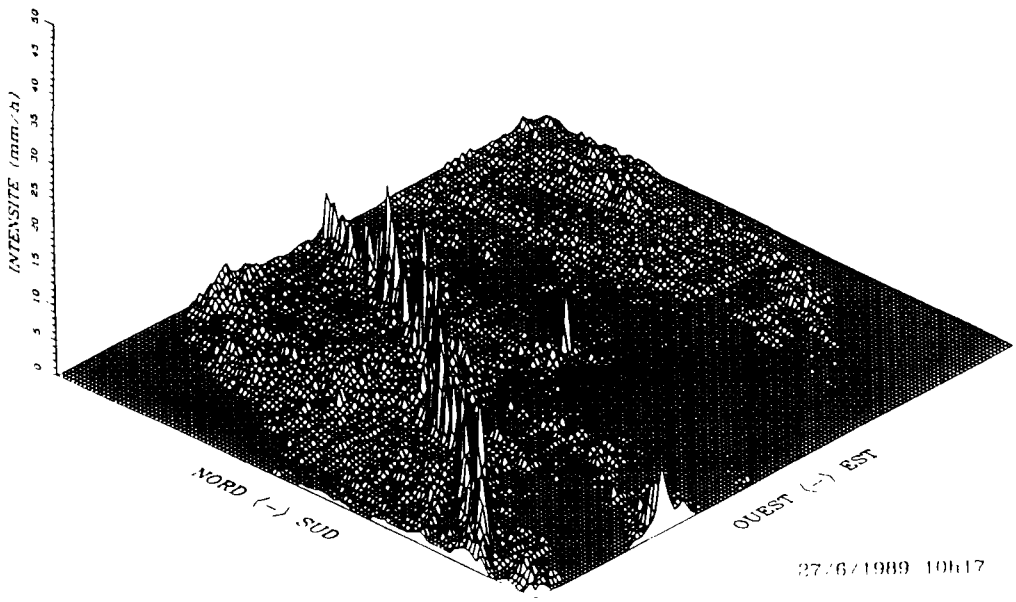


Figure I.5: Mesure par radar d'une bande pluvieuse étroite

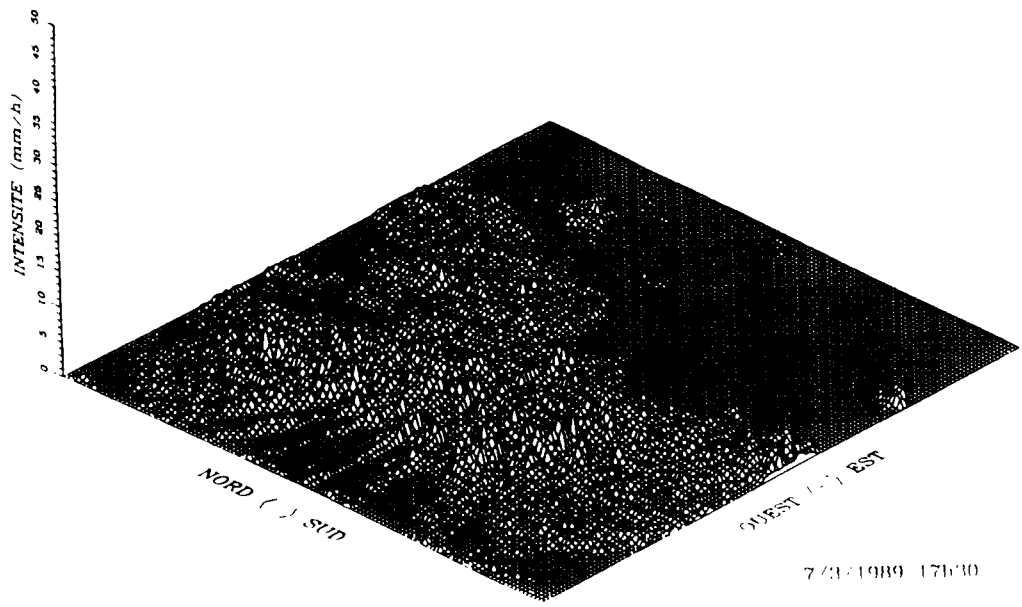


Figure I.6: Mesure par radar d'une bande pluvieuse large

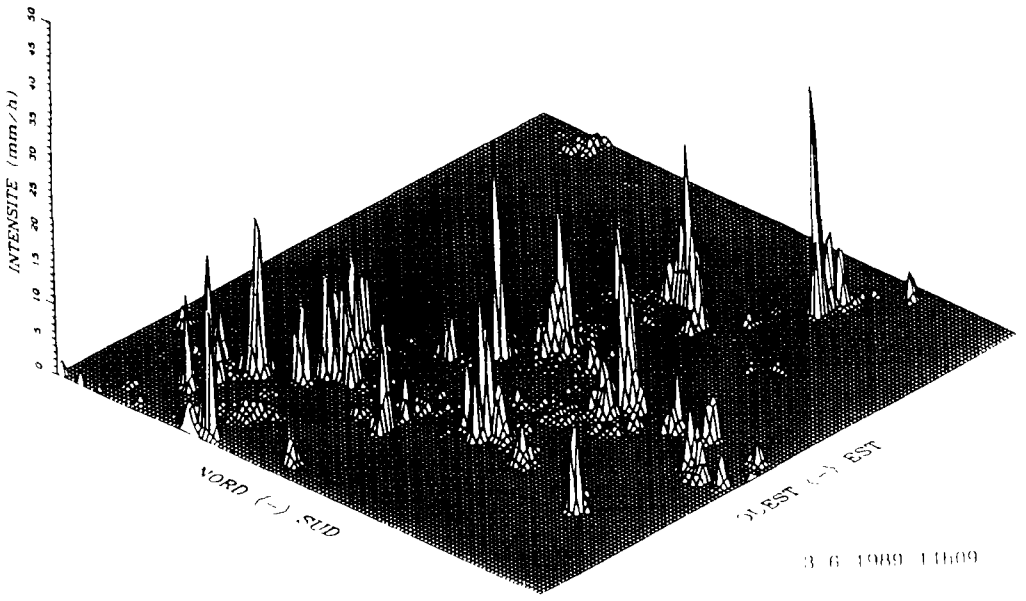


Figure I.7: Mesure par radar d'une pluie convective

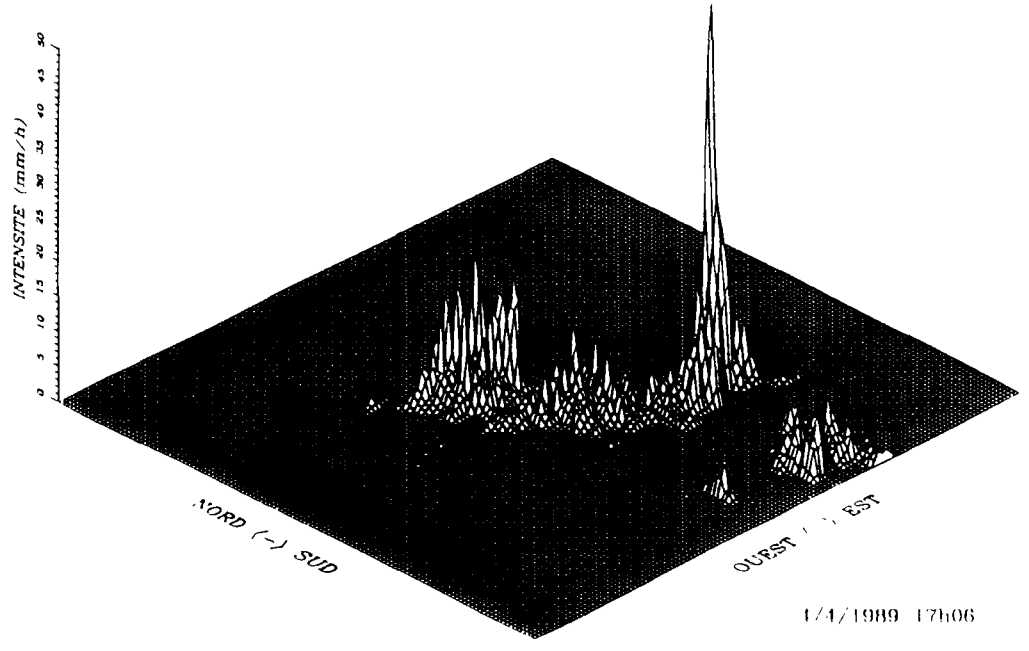


Figure I.8: Mesure par radar d'une pluie de caractère mixte

I.2 L'observation de la pluie par radar

Peu après la découverte du radar (**radio detecting and ranging**) dans les années 1930, la possibilité de son utilisation pour des observations hydrométriques a été connue (Hitschfeld 1986). Le gain que peuvent apporter des mesures radar par rapport aux mesures ponctuelles par pluviomètres est fondé sur la haute résolution spatiale. De nombreux programmes de recherche ont été menés afin de résoudre les problèmes techniques de la mesure des précipitations par radar, ayant entre autres pour but d'estimer l'intensité de la pluie, de déterminer les lames d'eau précipitées, et d'étudier la structure tridimensionnelle de la précipitation.

Par la suite nous rappelons les principes de la mesure de la pluie par radar, les notions de base et les principaux problèmes techniques ayant une importance pour la prévision (cf. par exemple Battan 1973, Sauvageot 1982, Collier 1989, pour une analyse plus complet du sujet).

I.2.1 Les principes de la mesure de pluie par radar

Un faisceau d'ondes électromagnétiques est diffusé quand il traverse un volume de gaz contenant des gouttes d'eau. Cette propriété permet en principe deux techniques différentes de mesure quantitative de la précipitation: La mesure de la réflectivité et la mesure du degré d'atténuation du faisceau (Battan 1973). Seule la première méthode est utilisée de manière opérationnelle pour la mesure de pluie par radar.

La valeur mesurée par cette méthode est la puissance P_r de l'onde radar réfléchie par un volume V à distance r de l'antenne. Si V est complètement rempli de gouttes d'eau de petits diamètres comparés à la longueur d'onde du radar, l'équation suivante est valable:

$$P_r = C \frac{Z}{r^2}$$

avec une constante C définie par les caractéristiques du radar utilisé, notamment par la longueur d'onde. La valeur Z est appelée le ***facteur de réflectivité*** de V. Z est défini comme

$$Z = \int_0^\infty D^6 n_a(D)\, dD \quad [\mathrm{mm}^6/\mathrm{m}^3]$$

où $n_a(D)dD$ représente le nombre de gouttes d'eau <u>en altitude</u> de diamètre compris entre D et $D+dD$ contenu dans V. Z est souvent exprimé en dBZ selon

$$Z_{dBZ} = 10 \log_{10}\left(\frac{Z\ \mathrm{mm}^6/\mathrm{m}^3}{1\ \mathrm{mm}^6/\mathrm{m}^3}\right)$$

La valeur qu'on cherche à mesurer est l'intensité R_s de la précipitation au sol, qui s'exprime par

$$R_s = \frac{\pi}{6} \int_0^\infty V_t(D)\, D^3 n_s(D)\, dD \quad [\mathrm{mm/h}]$$

$n_s(D)dD$ étant le nombre de gouttes d'eau <u>au sol</u> de diamètre compris entre D et $D+dD$, et $V_t(D)$ étant la vitesse terminale de chute d'une goutte de diamètre D.

On peut tenter d'établir une relation fonctionnelle entre ces deux paramètres, en estimant la vitesse terminale de chute d'une goutte par $V_t(D) = cD^\alpha$ avec c et α constants. La distribution des diamètres des gouttes d'eau a été estimée par l'expression exponentielle $n(D) = n_0 e^{-aD}$.

Si l'on suppose une répartition homogène de la précipitation dans V et une intensité R_s constante on obtient la relation

$$Z = AR_s^b$$

A et b étant constants. En conséquence, si tous les paramètres étaient connus, on pourrait estimer l'intensité de la pluie au sol à partir des mesures par radar en altitude. Or, dans le cas de l'utilisation du radar comme seul moyen de mesure, plusieurs paramètres sont inconnus:

- La distribution des diamètres des gouttes d'eau est difficilement estimable en altitude ($n_a(D)$), et elle est généralement inconnue au sol ($n_s(D)$).
- La vitesse de chute des gouttes ($V_t(D)$) dépend fortement de l'advection verticale de l'air, valeur qui, normalement, est inconnue.

Ainsi, les paramètres A et b de la relation Z-R établie ne peuvent pas être déterminés directement. Plusieurs approches ont été faites afin d'établir une relation empirique. De nombreuses valeurs ont été proposées dans la littérature pour les paramètres A et b (cf. Battan (1973) pour un résumé des propositions). Une relation souvent utilisée pour des pluies en latitude moyenne est celle proposée par Marshall et Palmer (Marshall et Palmer 1948) avec les valeurs $A = 200.$ et $b = 1.6$.

I.2.2 Les sources d'erreur de la mesure de pluie par radar

L'***image radar*** est une matrice $I(n,m)$; la valeur de chaque élément (nommé ***pixel***) de la matrice est une estimation de la réflectivité d'un volume d'air défini, à partir de laquelle on déduit l'intensité de la pluie pour une surface. Généralement ces surfaces sont de forme rectangulaire. Trois types d'images radar ont une importance en hydrologie:

- L'image radar type ***PPI*** (plan position indicator) est une mesure horizontale avec un angle d'élévation fixe. La précipitation d'un pixel p est alors mesurée à une altitude qui diffère en fonction de la distance du pixel par rapport au radar.
- L'image du type ***CAPPI*** (constant altitude plan position indicator) est une image composée de mesures dans la même couche atmosphérique de plusieurs PPI's d'élévations différentes.
- La valeur d'un pixel d'une image radar type ***VIL*** (vertically integrated liquid water) représente la quantité d'eau intégrée verticalement à partir de mesures sous des angles d'élévation différents.

L'image radar utilisée dans cette étude est de type PPI. Par la suite, nous étudierons brièvement les erreurs de mesure spécifiques à ce type d'image.

Les erreurs de mesure de la pluie par radar ont fait l'objet de nombreuses études (par exemple Harrold et *al.* 1975, Zawadzki 1984, Andrieu 1986). Denoeux (1989) présente de façon exhaustive les erreurs ayant une influence sur la prévision et la mesure des lames d'eau. Ici nous ne considérons que les principaux défauts de la technique relative à l'hydrologie de surface.

En hydrologie, on cherche généralement à mesurer ou à prévoir la lame d'eau précipitée sur un bassin donné dans un intervalle de temps précis. Or, le radar fournit une mesure instantanée de la réflectivité d'un volume d'air en altitude. Cette différence implique plusieurs types d'erreurs, dont certains sont montrés dans la figure I.9. On peut classer les erreurs en trois groupes:

- des erreurs de mesure de la réflectivité,
- des erreurs de transformation de la réflectivité en intensité de pluie en altitude,
- des erreurs de transformation de l'intensité de la pluie en altitude en lame d'eau au sol.

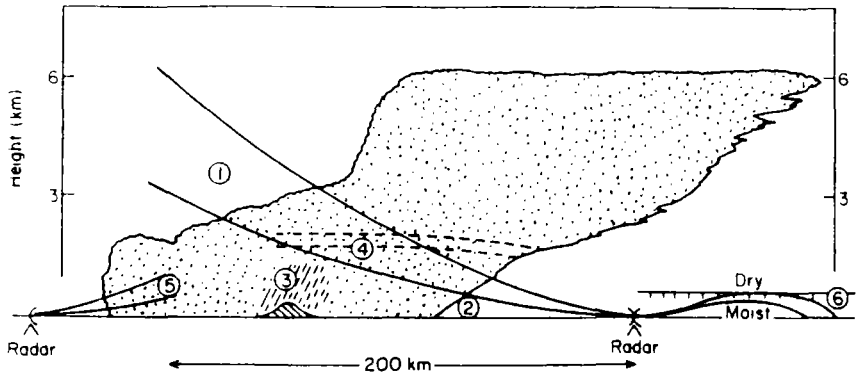


Fig. 6. Cross-section through an area of frontal precipitation illustrating six sources of error in the radar measurement of surface rainfall intensity, namely: (1) radar beam overshooting the shallow precipitation at long ranges; (2) low-level evaporation beneath the radar beam; (3) orographic enhancement above hills which goes undetected beneath the radar beam; (4) anomalously high radar signal from melting snow (the bright band); (5) underestimation of the intensity of drizzle because of the absence of large droplets; and (6) radar beam bent in the presence of a strong hydrolapse causing it to intercept land or sea.

Figure I.9: Coupe verticale à travers une zone pluvieuse frontale, montrant six erreurs de mesure de l'intensité (d'après Browning et Collier 1982)

I.2.2.1 Erreurs de mesure de la réflectivité

Par la mesure radar on cherche à mesurer la réflectivité $Z(V)$ d'une masse d'air V, dont la taille et la localisation sont bien définies. Des erreurs de mesure sont introduites par les caractéristiques du matériel utilisé (stabilité du récepteur, signal minimal détecté, angle d'ouverture, longueur d'onde, calage du système) et par le traitement des signaux reçus. Parmi les erreurs due au matériel on mentionne notamment le problème d'atténuation du faisceau radar par la pluie, effet bien connu pour les longueurs d'ondes courtes (Zawadzki 1984, Collier 1989).

Si V est de forme cubique, le traitement des données introduit des erreurs dues au problème de transformation des mesures de coordonnées polaires en coordonnées cartésiennes. Les unités de mesure ayant la forme d'un segment d'anneau, une conversion est nécessaire afin d'obtenir les mesures pour des unités quadratiques (figure I.10). Comme la résolution spatiale de la mesure est décroissante avec la distance r du volume V du radar, l'exactitude de cette conversion diminue en proportion avec la distance.

Un autre problème de mesure est celui de la ***propagation anormale*** du faisceau radar. Le faisceau peut être réfracté d'une manière forte dans le cas d'un très fort gradient de température avec l'altitude. La réflectivité mesurée ne correspond alors pas à celle du volume ciblé V. Il peut arriver que le faisceau intercepte le sol et engendre ainsi des ***échos de sol***. Généralement ces échos peuvent facilement être filtrés, car le signal réfléchi par un écho de sol est de moindre variabilité que le signal réfléchi par la pluie.

Finalement, la quantité d'informations fournies par le radar nécessite souvent une simplification des données, pour faciliter l'archivage et le transfert. Souvent les mesures continues sont discrétisées (de 8 à 256 niveaux), chaque niveau correspondant à un intervalle de réflectivité. L'influence de l'échelle de digitalisation sur l'utilité des données varie en fonction de l'application des données (Agostini-Blanchet 1988, Cluckie et *al.* 1989).

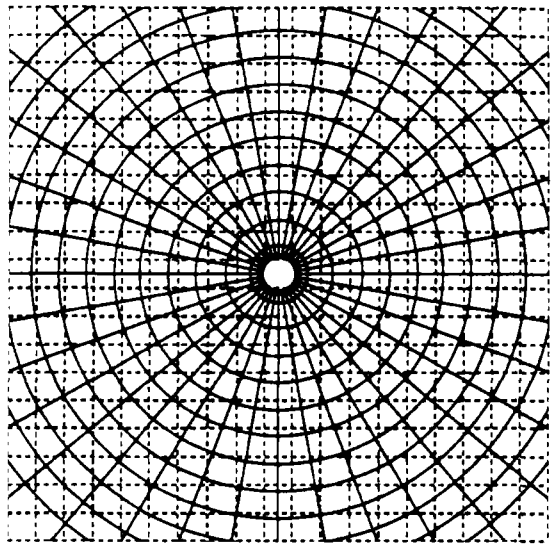


Figure I.10: Présentation schématique de l'erreur introduite par la conversion d'unités polaires de mesure de radar en unités quadratiques

I.2.2.2 Erreurs introduites par la transformation de la réflectivité en intensité de pluie

Les problèmes à l'origine de la méconnaissance des paramètres de la relation *Z-R* ont déjà été mentionnés. Plusieurs techniques ont été proposées afin d'améliorer les estimations de l'intensité de la pluie mesurée par radar. A cause de la multitude des paramètres inconnus, les approches empiriques sont les plus employées. Zawadzki et Calheiros (1987) proposent une démarche d'établissement d'une relation statistique entre les réflectivités mesurées par radar et les intensités mesurées au sol. Ils soulignent cependant que cette méthode n'optimise pas la mesure pour chaque pluie individuellement.

D'autres méthodes reposent sur une estimation des paramètres de la rélation *Z-R* après mesure. La distribution des diamètres des gouttes d'eau ayant une importance prépondérante dans cette relation, la plus grande partie de techniques cherche à déterminer la distribution des diamètres pour la pluie mesurée. Russchenberg et Baptista (1990) montrent, que la variabilité de cette distribution lors d'une même pluie peut être très importante. Zawadzki (1984) constate cependant une influence mineure de cette variabilité comparé aux autres sources d'erreur.

Au sol, la distribution des gouttes $n_s(D)$ d'une pluie peut être mesurée par un disdromètre, méthode de mesure ponctuelle qui suppose une homogénéité verticale et horizontale de la distribution (Breuer 1976). Selon Russchenberg et Baptista (1990), cette condition n'est généralement pas remplie.

Une méthode de la mesure de la distribution en altitude $n_a(D)$ consiste à utiliser la technique du radar à multiple polarisation pour la détermination de la forme moyenne des gouttes. Les grandes gouttes étant plus aplaties que les petites, la relation entre la réflectivité horizontale et la réflectivité verticale permet d'estimer le volume d'eau par goutte d'eau moyen. Un résumé de la mesure par radar à multiple polarisation se trouve dans Rogers (1984).

Une autre technique de mesure en altitude consiste en une combinaison des mesures des réflectivités du faisceau radar et de l'atténuation d'un signal radio émis par satellite. Les deux grandeurs étant dépendantes de la taille des gouttes d'eau, on peut en déduire la distribution des tailles des gouttes (Russchenberg et Baptista 1990).

Ces trois techniques sont basées sur un calage des équations établissant la rélation *Z-R*. Cette approche reste insuffisante, car la distribution des tailles des gouttes d'eau n'est qu'un paramètre parmi d'autres inconnus. Par conséquent, les méthodes empiriques de calage direct des paramètres *A* et *b* de la rélation *Z-R* sont aujourd'hui les plus utilisées. Une méthode simple consiste en une adaptation des paramètres *A* et *b* au type de pluie mesurée, mais elle ne permet pas de prendre en compte la variabilité dans l'espace et dans le temps de ces paramètres.

Une méthode utilisée de manière opérationnelle est la calibration des mesures effectuées par radar à l'aide des mesures de l'intensité de la pluie au sol par pluviomètre (Browning 1979, Andrieu 1986). Les valeurs des intensités estimées sont calées par des facteurs de correction obtenus sur les points de mesure par pluviomètre. Cette méthode provoque des difficultés liées au fait que les mesures présentent des caractéristiques différentes. En effet, le radar mesure de façon discontinue dans le temps et de façon continue dans l'espace, et le pluviomètre mesure de façon continue dans le temps et discontinue dans l'espace. De surcroît, la sensibilité au vent des mesures par pluviomètre peut être source des erreurs importantes. Wilson et Brandes (1979) estiment la sous-estimation de l'intensité de pluie par le pluviomètre à 20-40%, pour une vitesse du vent de 10-35 m/s.

D'autres erreurs sont induites par l'hypothèse de base de cette méthode: l'objet mesuré est la pluie liquide. Or, il peut s'agir d'autres formes de précipitation, notamment de la neige, de la grêle et du grésil, avec des valeurs de réflectivité très différentes de celle de la pluie. A l'intérieur des nuages, la couche de la ***bande brillante***, dans laquelle on trouve des particules de neige fondante, possède ainsi un facteur de réflectivité multiple de celui d'une pluie de la même intensité. Il est généralement difficile de supprimer ces erreurs sur une mesure de radar classique. Seul un radar à polarisation multiple permet de distinguer la pluie des autres types de précipitation (Lipschutz et *al.* 1986).

Les ***échos fixes***, provoqués par des immeubles ou collines qui interrompent le faisceau du radar, sont une autre cause d'erreur. Ces échos peuvent être détectés et éliminés par la méthode utilisée pour les échos de sol.

I.2.2.3 Erreurs introduites par le calcul des lames d'eau

Il existe deux sources d'erreurs dans le calcul des lames d'eau: Les écarts entre l'intensité de la pluie mesurée en altitude et l'intensité au sol, et l'erreur qui est introduite par le fait que la mesure radar est instantanée, tandis que la lame d'eau est une grandeur continue.

L'estimation de l'intensité de la pluie au sol à partir de la mesure radar en altitude est basée sur l'hypothèse de l'homogénéité de cette intensité pendant la chute. Or, plusieurs effets peuvent invalider cette hypothèse; notamment l'évaporation, l'intensification et l'advection horizontale et verticale de la pluie pendant la chute. Zawadzki (1984) montre que le taux d'erreur ponctuel introduit par l'advection verticale est faible, alors que celui provoqué par l'advection horizontale peut aller jusqu'à 100% pour les mesures ponctuelles. Ce taux est cependant moins élevé pour les mesures des surfaces.

Le taux d'erreur introduit par le changement de la pluie pendant la chute dépend de la hauteur de la mesure. La hauteur *H*, distance entre le faisceau radar et la terre, est fonction de la distance. Elle est approximativement donnée par la formule

$$H = r \tan\Theta + \frac{r^2}{2R}$$

où r représente la distance entre le point de mesure et le radar, Θ l'angle d'élévation, et R le rayon de la terre. La hauteur croît alors rapidement avec la distance (figure I.11), rendant ainsi les erreurs de mesure plus importantes lorsque les distances sont plus élevées. Zawadzki (1984) insiste sur le caractère aléatoire de la différence constatée entre pluie en altitude et pluie au sol, et il conclut que les mesures faites au-delà de 100 km du radar sont difficilement utilisables.

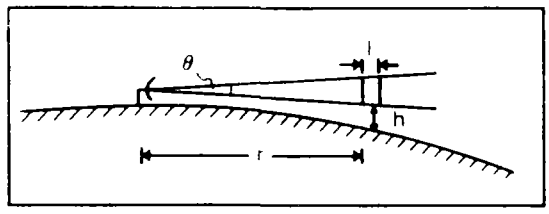


Figure I.11: Hauteur du faisceau radar en fonction de la distance r et de l'angle d'élevation Θ (d'après Collier 1989)

Finalement, l'erreur introduite par une intégration temporelle trop simplifiée des mesures instantanées peut être importante. Blanchet et *al.* (1989) montrent que cette erreur peut atteindre 30% pour les mesures ponctuelles sur des intervalles de mesure de 2 min. Pour diminuer cette erreur il faut tenir compte de l'advection des champs de pluie (figure I.12).

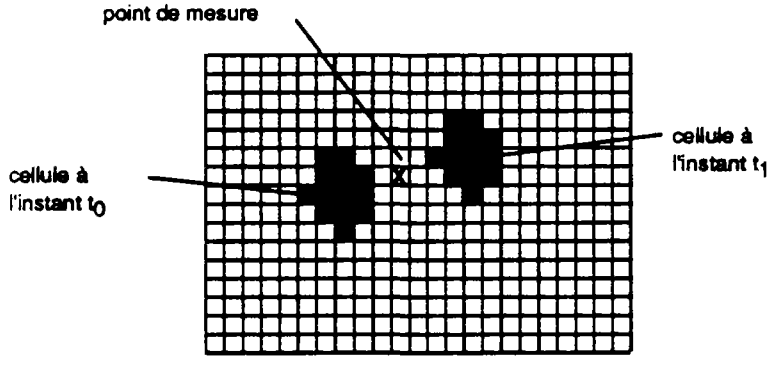


Figure I.12: Schéma montrant la nécessité d'une prise en compte de l'advection dans l'intégration temporelle des images radar

Niveau	Réflectivité dBZ	Intensité (rélation de Marshall et Palmer) mm/h
0	<16	<0.36
1	16-25	0.36-1.33
2	25-30	1.33-2.73
3	30-34	2.73-4.86
4	34-38	4.86-8.70
5	38-41	8.70-13.3
6	41-43	13.3-17.8
7	43-45	17.8-23.7
8	45-47	23.7-31.6
9	47-48	31.6-36.5
10	48-50	36.5-48.6
11	50-52	48.6-64.8
12	52-53	64.8-74.9
13	53-55	74.9-100.
14	55-58	100.-154.
15	>58	>154.

Tableau I.1: Echelle de réflectivité des données radar utilisées dans cet étude (échelle CALAMAR, d'après Agostini-Blanchet 1988)

I.2.3 Caractéristiques des données utilisées dans cette étude

I.2.3.1 Caractéristiques techniques des données

Pour cette étude, nous avons utilisé des données du radar météorologique de Trappes. La région couverte par la mesure du radar est montrée par la figure I.13. Ce radar fait partie du réseau ARAMIS de la Météorologie Nationale de France (Cheze et *al.* 1991). Il est du type RODIN et utilise une longueur d'onde de la bande C. Le tableau I.2 résume les principales caractéristiques du radar.

Producteur	Thomson
Diamètre de l'antenne	3 m
Rotation	72 sec
Ouverture	1.3°
Longueur d'onde	5.23 cm
Puissance du signal	250 kW
Durée d'une impulsion	2 µs
Fréquence	330 Hz
Seuil de détection	8 dBZ à 100 km

Tableau I.2: Spécifications techniques du radar Rodin de Trappes

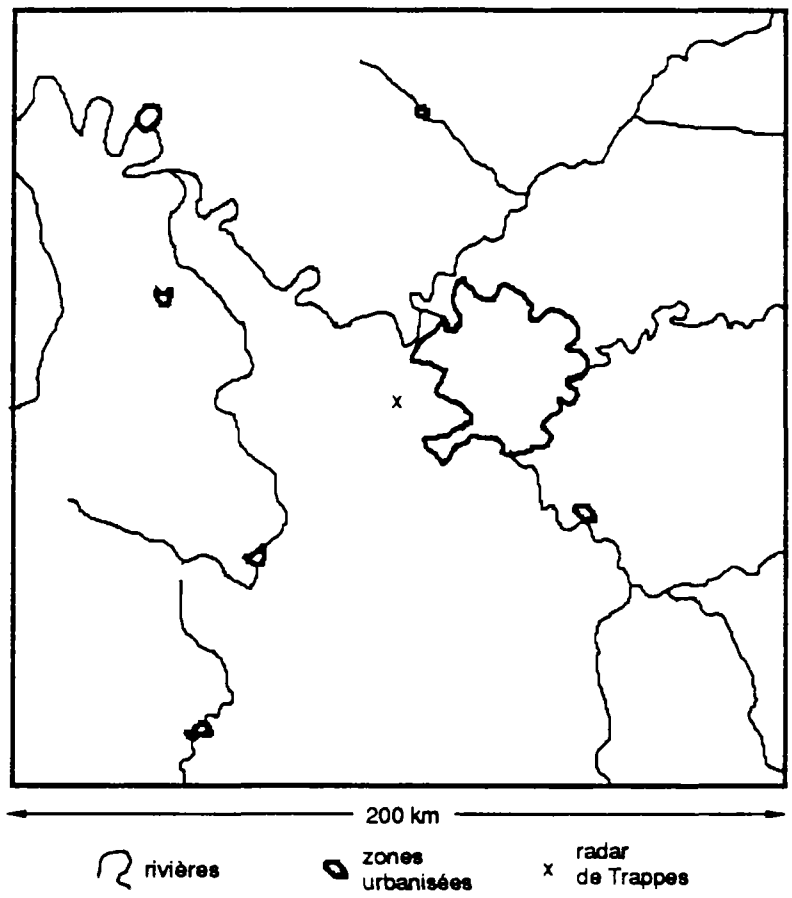


Figure I.13: Fond de carte de la région couverte par les images du radar de Trappes

Les images radar utilisées sont des images spécifiquement conçues pour l'hydrologie par la société RHEA en France. Elles ont les caractéristiques suivants:

- les images sont du type PPI,
- elles couvrent un carré de 200 km de coté avec une résolution spatiale de 800 m,
- l'intervalle d'archivage des images est de 4 à 6 minutes,
- les mesures sont discrétisées en 16 niveaux selon une échelle optimisée pour l'hydrologie urbaine (tableau I.1).

I.2.3.2 Réduction de l'influence des erreurs de mesure

Ci-dessus nous avons examiné les erreurs possibles de mesure de la pluie par radar. Plusieurs précautions ont été prises afin de réduire leur importance pour les résultats obtenus dans cette étude:

- Toutes les images choisies ont été vérifiées à l'aide d'une visualisation (cf. la description du système d'analyse et visualisation en annexe). Ainsi les erreurs de mesure dues à la propagation anormale et à la bande brillante ont été écartées. Quelques rares périodes de forte atténuation du faisceau radar ont été repérées et exclues.
- La discrétisation des mesures en 16 niveaux selon une échelle optimisée (Agostini-Blanchet 1988) réduit la perte d'information due à cette transformation.
- Grâce à la lente rotation et la courte durée de l'impulsion du radar de Trappes, la transformation des mesures polaires en coordonnées cartésiennes, effectuée par la Météorologie Nationale, provoque relativement peu d'inexactitudes.
- Outre le filtrage des données par la Météorologie Nationale, nous avons supprimé manuellement les échos fixes proche du radar à l'aide du système de visualisation. Grâce à la topographie de la Région Parisienne, les échos fixes ne posent pas de problème majeur au delà d'une distance d'environ 5 km du radar (figure I.14).
- Toutes les données utilisées ont été enregistrées dans la période mars-octobre. Les précipitations sous forme de neige ou de grêle ne sont pas traitées.
- On s'intéresse aux mesures et aux prévisions pour la région de l'agglomération Parisienne, qui se trouve à une distance entre 20 kilomètres et 50 kilomètres au nord-est du site du radar de Trappes. A une telle distance le faisceau du radar est à une altitude inférieure à 650 mètres, limitant ainsi les effets d'intensification et d'évaporation de la pluie au-dessous du faisceau.
- Pour tout calcul de lames d'eau nous avons appliqué systématiquement une intégration des images radar avec l'advection des champs de pluie, afin d'éviter des erreurs dues à l'échantillonnage temporel des mesures.

Pour la transformation de la réflectivité en intensité de pluie, nous avons utilisée la relation proposée par Marshall et Palmer (1948). Des comparaisons entre les valeurs estimées et les intensités mesurées ponctuellement par pluviomètres ont montré que, pour la plus grande partie des données utilisées dans cette étude, la pluie mesurée au sol est d'une intensité supérieure à celle estimée à partir du radar. La sous-estimation est de l'ordre de 10% à 50%, selon le type de pluie. Il a été montré que ce facteur ne change pas de façon importante pendant le passage d'un même champs de pluie. Faute d'un réseau pluviométrique dense s'étendant sur toute la région observée (d'une densité équivalent à celle existant dans le département de Seine-St.Denis par exemple), un ajustement s'adoptant à chaque champs de pluie et évoluant dans le temps semble difficile.

Mais la relative stabilité de la rélation *Z-R* au sein d'un champ de pluie nous permet cependant la comparaison relative des prévisions des lames d'eau avec les estimations des lames d'eau à partir des images radar. L'évaluation des prévisions sera faite uniquement sur la base de cette comparaison. Tout au long de cette étude, la dénomination "intensité de pluie" est utilisée pour désigner des estimations d'intensités réelles, et la dénomination "lames d'eau" est utilisée pour désigner des estimations des lames d'eau réelles.

Le tableau I.3 résume les méthodes de prise en compte des erreurs de mesure par radar, qui ont été appliquées dans cette étude.

ECOLE NATIONALE DES PONTS ET CHAUSSÉES
CENTRE PÉDAGOGIQUE DE DOCUMENTATION ET DE COMMUNICATION

Source d'erreur	Mode de prise en compte
Atténuation du faisceau radar	Sélection des données à l'aide de la visualisation
Transformation en données cartésiennes	Erreur de moindre importance à cause des caractéristiques du radar
Echos de sol	Sélection des données à l'aide de la visualisation
Discrétisation des données	Erreur de moindre importance à cause de l'échelle optimisée
Méconnaissance de la rélation Z-R	Travail avec des valeurs relatives pour les intensités de pluie et les lames d'eau
Neige, grêle, bande brillante	Sélection des données à l'aide de la visualisation
Echos fixes	Suppression manuelle dans un rayon de 5 km du radar, au-delà erreur de moindre importance
Evaporation et intensification de la pluie sous le faisceau radar	Limitation des prévisions à une région située à une distance du radar inférieure à 50 km
Echantillonnage temporel	Prise en compte de l'advection pour le calcul des lames d'eau

Tableau I.3: Prise en compte de principales sources d'erreur de la mesure par radar

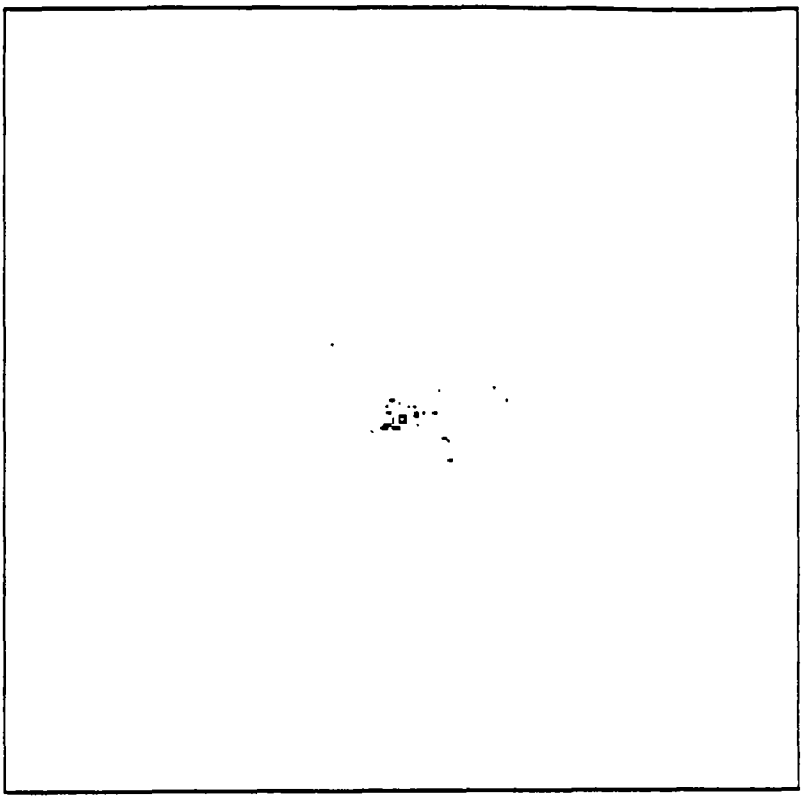


Figure I.14: Echos fixes sur l'image radar de Trappes (mesure du 10 juin 1989 à 10h04 TU)

I.3 La prévision de pluie à des échéances courtes

I.3.1 Utilisation des prévisions de pluie en hydrologie urbaine

Outre la construction, la gestion optimisée des réseaux urbains d'assainissement prend une importance croissante en hydrologie urbaine. Les objectifs de la gestion des réseaux sont: la sécurité des habitants de la région et des travailleurs dans le réseau, la protection de l'environnement et l'optimisation du coût de la gestion. Pour atteindre ces buts, on a développé des systèmes automatisés de gestion en temps réel, qui, basés sur des mesures d'état du système, proposent des stratégies de gestion. Delattre et *al.* (1986), Einfalt et Denoeux (1987), Frérot (1987), KNMI (1990), et Khelil (1990) décrivent quelques systèmes existants. Il a été montré que l'intégration d'une prévision du ruissellement peut améliorer la gestion de façon importante (Damant et *al.* 1983, Schilling et Petersen 1987, Denoeux 1989). Cela nécessite une prévision des lames d'eau à une échelle spatiale et temporelle adaptée au système à gérer. L'amélioration de la gestion est d'autant plus forte, que la pluie est diversifiée. Pendant un orage d'été, les possibilités d'optimiser la gestion sont ainsi souvent plus grandes que lors d'une pluie stratiforme; en effet cette dernière remplit le réseau de manière uniforme et minimise l'intérêt des améliorations apportées par une gestion optimisée; tandis que l'orage crée des problèmes locaux dont la résolution nécessite plus une gestion disposant d'une bonne connaissance de la situation globale.

Schilling et Petersen (1987) ont montré que, pour le réseau de Brème, la plus grande amélioration obtenue de la gestion est due à l'intégration de la connaissance des pluies et débits de 30 minutes à 2 heures d'avance, alors que l'amélioration supplémentaire due à l'élargissement de l'horizon à 3 à 4 heures est faible. Damant et. *al.* (1983) parlent, quant à eux, d'un horizon nécessaire de 30 minutes à 1 heure pour le système de Montréal.

En ce qui concerne l'échelle spatiale utile à la prévision, la résolution minimale nécessaire doit correspondre à la taille des surfaces des bassins introduits dans le module de gestion, qui est en général de quelques km^2.

Il est difficile de fixer un taux d'erreur acceptable de la prévision. Denoeux (1989) remarque, qu'une prévision peut être considérée comme utile, si elle améliore les résultats de la gestion. En conséquence, on ne peut pas évaluer une prévision, si on ne définit pas au préalable son intégration dans le système de gestion existant. Le critère d'évaluation des prévisions, qui a été utilisé dans cette étude, sera défini ultérieurement. Pour le moment, nous nous contentons de formuler comme objectif que le taux d'erreur des prévisions soit "assez bas", pour améliorer la gestion d'un réseau d'assainissement par rapport à la gestion sans prévision.

I.3.2 La prévision de pluie à l'aide des modèles numériques

La méthode "classique" de la prévision est la modélisation numérique des processus physiques de l'atmosphère. Ces modèles sont initialisés par des mesures décrivant l'état de l'atmosphère (la pression, le vent, la température et l'humidité), effectuées, à heures régulières, soit à partir de stations météorologiques, soit à partir de satellites. Après cette initialisation, le développement de l'atmosphère est modélisé pour une durée de quelques jours.

Avec l'augmentation de la puissance des super-ordinateurs, la qualité de la modélisation et de la prévision a été considérablement améliorée durant les dernières décennies, surtout grâce à un raffinement de la résolution spatiale et temporelle des modèles.

Néanmoins, il y a des limites à une telle progression (Torterotot 1988, Bougeault et *al.* 1989, Sutton et Conway 1989):

- Une réduction du maillage de modélisation d'un facteur de 2 augmente le temps de calcul d'un facteur de 8.
- Le maillage des mesures, actuellement compris entre 100 km et 1000 km, doit être réduit au même degré que le maillage du modèle.

Pour être le plus exact possible, on utilise souvent deux modèles différents: l'un, avec une maille large à l'échelle synoptique, couvre tout le globe, et l'autre, avec une maille plus fine à l'échelle moyenne, couvre la région pour laquelle il faut établir les prévisions. Les résultats obtenus à partir du modèle synoptique complètent les mesures; ensemble ils servent à l'initialisation du modèle fin. La Météorologie Nationale française emploie ainsi le modèle EMERAUDE, pour la modélisation de l'atmosphère globale, et le modèle PERIDOT, qui ne couvre que la partie ouest de l'Europe. Ce dernier emploie une résolution d'une grille horizontale quasi-quadratique de 35 km environ, et de 15 niveaux verticaux. Le pas de temps de calcul est de 4 minutes (Imbard et *al.* 1986, Torterotot 1988). Un principe similaire est aussi employé par le Meteorological Office en Grande Bretagne, le modèle fin utilisant un maillage de 15 km (Sutton et Conway 1989, Golding 1987).

Or, l'échelle des phénomènes que peut décrire un modèle est de quatre fois la taille de la maille, soit 140 km pour le modèle PERIDOT (Torterotot 1988), échelle qui est trop grossière pour la plus grande partie des phénomènes importants en hydrologie urbaine. Autre problème majeur pour obtenir les prévisions souhaitées à l'aide des modèles numériques est celui du "spin up", le temps de simulation nécessaire au modèle afin de développer des solutions stables. Ce temps est généralement de l'ordre de deux heures, temps qui s'ajoute à l'intervalle d'échéance et qui augmente le taux d'erreur.

Ces modèles ont alors une importance faible pour les prévisions à courtes échéances et à haute résolution spatiale, qui sont nécessaires en hydrologie urbaine.

I.3.3 La prévision de pluie par extrapolation des observations

Les techniques de mesure du radar et du satellite offrent la possibilité d'observation de la pluie avec une haute résolution spatiale. Cette bonne connaissance de l'état de l'atmosphère peut conduire à une prévision au moyen d'une extrapolation des observations. A partir de ces mesures, l'advection et le développement des phénomènes (nuages, zones de pluie) sont déterminés et extrapolés dans le futur immédiat. La validité d'une telle prévision dépend de la qualité de l'observation et de la stabilité du phénomène observé.

Les techniques existantes d'extrapolation à partir des mesures par radar diffèrent selon leur degré d'automatisation et selon la complexité qui caractérise les observations.

Les méthodes ***semi-automatiques*** travaillent en interaction avec le prévisionniste. Une application d'une telle méthode aux États-Unis est décrite par Huff et *al.* (1980). Un autre exemple de ce type est le système FRONTIERS du Meteorological Office de la Grande Bretagne (Sutton et Conway 1989, Browning 1979). Il s'agit d'un système d'aide à l'interprétation des images radar et satellite. Après avoir visualisé l'image radar ou satellite, le prévisionniste définit manuellement les champs de précipitation. Il assigne à chaque champ un vecteur d'advection et un facteur de croissance/décroissance, qu'il dérive subjectivement des images. A partir de ces informations, le système fournit les prévisions. Sutton et Conway (1989) signalent, que ce système laisse une part trop importante des décisions au prévisionniste, qui, souvent, n'arrive pas à utiliser toutes les opportunités offertes. En particulier l'option de prévoir le développement des champs de pluie ne peut pas être exploitée suffisamment.

D'où l'intérêt des méthodes automatisées, qui offrent un champ d'application plus vaste, par exemple comme système d'alerte. Les techniques ***automatisées non structurées*** sont des méthodes automatisées de prévision, qui emploient une approche basée sur l'image entière, et qui ne tient pas compte des structures météorologiques. Elles comprennent deux étapes:

La première est la ***caractérisation*** de la pluie sur l'image courante par la détermination d'un vecteur d'advection moyen de déplacement.

La deuxième est la ***prévision***, qui extrapole ce vecteur pour la prédiction des déplacements de toutes les structures se trouvant sur cette image.

(1) Caractérisation de la pluie sur l'image actuelle I_t
(2) Prévision par extrapolation des caractéristiques dans l'avenir

Algorithme I.1: Prévision automatisée non structurée

Dans la plus grande partie des cas d'application d'une méthode non structurée, le vecteur $\vec{v}$ de l'advection est déterminé par la méthode du coefficient de la corrélation croisée:

Soit deux mesures, images $I_{t-\Delta t}$ et I_t, représentées sous forme de deux matrices de réflectivité $M_{t-\Delta t}$ et M_t. On définit

$$\hat{M}_{t-\Delta t} = M_{t-\Delta t} - m_{t-\Delta t}$$

$$\hat{M}_t = M_t - m_t$$

avec $m_{t-\Delta t}$ et m_t les réflectivités moyennes des deux images. Le coefficient $\rho(\vec{v})$ de la corrélation croisée des matrices $M_{t-\Delta t}$ et M_t est définie pour l'advection $\vec{v}=(v_x,v_y)$ comme

$$\rho(\vec{v}) = \frac{\sum_x \sum_y (\hat{M}_{t-\Delta t}(x,y) \cdot \hat{M}_t(x+v_x,y+v_y))}{\sqrt{\sum_x \sum_y \hat{M}_{t-\Delta t}(x,y)^2 \cdot \sum_x \sum_y \hat{M}_t(x,y)^2}}$$

et on cherche $\vec{v}_{max}$ avec

$$\rho(\vec{v}_{max}) = \max_{\vec{v}} (\rho(\vec{v}))$$

Un exemple d'un système opérationnel de ce type est le système SHARP, développé à l'université de McGill, Montréal (Austin et Bellon 1974, Bellon et Austin 1978).

Une autre méthode pour déterminer le vecteur d'advection $\vec{v}$ est l'utilisation du champ du vent en altitude pour l'estimation du déplacement des champs de pluie. Ces vecteurs peuvent être obtenus par des mesures atmosphériques ou par des modèles numériques. Tatehira et *al.* (1981) décrivent une telle application.

Les méthodes non-structurées ne définissent pas sur l'image radar de structures ayant une signification physique. Le déplacement individuel et le développement des champs de pluie ne peuvent pas être pris en considération. Avec de telles approches, la prévision dans des situations non homogènes est alors difficile.

Les techniques ***automatisées structurées*** de prévision permettent de prendre en compte des caractéristiques différentes grâce à l'analyse des mesures au niveau des structures

météorologiques. Ces structures sont les champs de pluie ou, à une échelle plus petite, les cellules de pluie.

(1) Identification et description des structures sur l'image radar actuelle I_t

(2) Appariement des structures observées sur les images radar de l'intervalle (t-$n\Delta t$, t-Δt) avec les structures définies sur l'image I_t

(3) Caractérisation des structures observées dans l'intervalle (t-$n\Delta t$, t)

(4) Prévision par extrapolation des caractéristiques dans l'avenir

Algorithme I.2: Prévision automatisée structurée

A un instant donné, une telle technique comprend quatre étapes (algorithme I.2). A l'inverse des méthodes dites non-structurées, la caractérisation de la pluie est basée sur une description des structures qui sont appariés d'une image à l'autre. Une prévision est effectuée individuellement pour chaque structure.

Parmi les premiers à avoir développé des systèmes de prévision structurés on peut citer Barklay et Wilk (1970), Blackmer et *al.* (1973), et Ostlund (1974). Crane (1979) décrit une approche particulièrement adaptée aux situations convectives. Bjerkaas et Forsyth (1980) ont développé le système WEATRK, qui emploie une technique structurée s'appuyant sur des mesures tridimensionnelles par radar. Forfoulu-Georgiou et *al.* (1990) ont tenté une approche de reconnaissance basée sur la description des formes des cellules par des coefficients de Fourrier.

Un exemple opérationnel d'un système automatisé structuré est le système SCOUT, décrit par Einfalt (1988). Ce système est intégré comme système automatique d'alerte dans le cadre de la gestion automatisée du réseau d'assainissement du département de la Seine-St.Denis (Einfalt et *al.* 1990).

Plusieurs auteurs ont proposé une combinaison des modèles numériques avec des méthodes d'extrapolation. (Browning 1980, Austin et Bellon 1982). Ces approches sont basées sur une extrapolation des observations pour la prévision immédiate et une utilisation de modèles numériques pour les grands intervalles d'échéance (figure I.15). Une question n'est pas résolue: à partir de quel intervalle d'échéance les modèles numériques donnent ils des prévisions plus fiables que les techniques d'extrapolation? Cet intervalle maximal dépend fortement du type de situation météorologique.

Un autre problème, qui subsiste, est le développement d'une méthode de transition entre les différentes techniques. Austin et *al.* (1990) demandent l'établissement de liens entre les systèmes de prévision à des échéances courtes et moyennes, en particulier le développement de techniques d'initialisation des modèles numériques à l'aide des mesures par radar et par satellite.

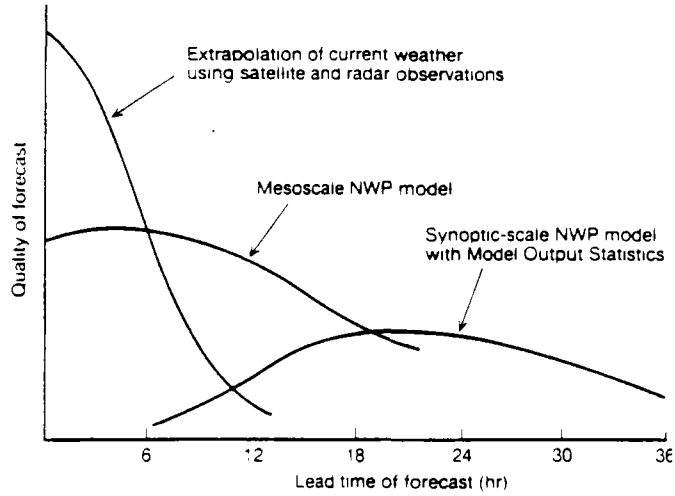


Figure I.15: Qualité des prévisions en fonction des intervalles d'échéance, pour différentes techniques de prévision (d'après Browning 1980)

I.3.4 Sources d'erreur de la prévision de pluie par extrapolation

Les erreurs de la prévision de pluie basée sur des mesures radar ont deux sources principales:

- les erreurs dues à l'estimation de la pluie à partir de l'image radar,
- les erreurs dues à la seule prévision de pluie.

Les erreurs du premier type comprennent les erreurs de mesure, mentionnées précédemment. Une autre erreur de ce type est due au rayon de mesure limité du radar. Naturellement, des extrapolations ne sont possibles que pour les pluies observées sur l'image radar. Les possibilités de prévision pour des bassins proches du bord de l'image radar sont ainsi limitées. Les bassins, auxquels on s'intéresse dans cette étude, sont situés entre 20 km et 50 km au nord-est du site du radar de Trappes. Avec une advection moyenne des pluies de 50 km/h en direction de l'est, l'intervalle d'échéance maximale de la prévision est de 2 heures environ. Cet intervalle correspond aux demandes formulées au chapitre I.3.1. Néanmoins, pour une advection forte de direction différente cet intervalle maximal peut être ramené à moins d'une heure.

Les erreurs du deuxième type sont dues à la seule méthode de prévision. Bellon et Austin (1984) estiment que, pour un système de prévision non-structuré, les deux types d'erreurs ont une influence équivalente sur les écarts entre prévisions et pluies réelles. Browning et *al.* (cité par Collier 1981) confirment ces résultats.

Dans cette étude nous étudions plus particulièrement les erreurs du deuxième type, dont Collier (1989) distingue trois groupes:

- erreurs dues à une caractérisation insuffisante de la pluie observée,
- erreurs dues à un changement de l'advection de la pluie pendant l'intervalle d'échéance de la prévision,
- erreurs dues au développement de la pluie pendant l'intervalle d'échéance de la prévision.

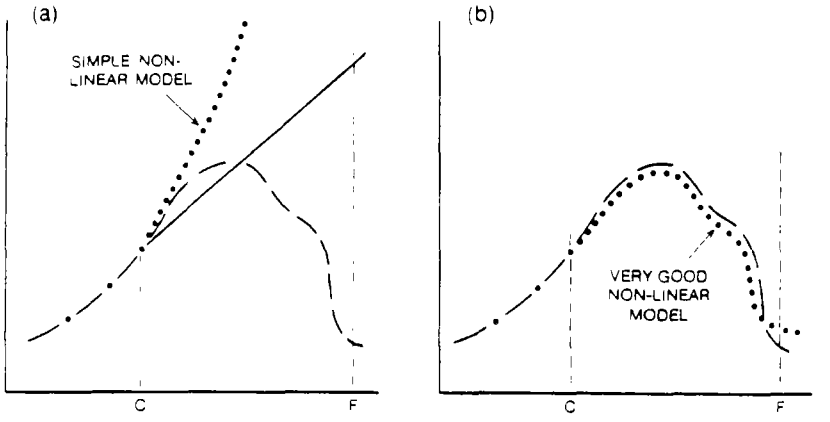


Figure I.16: Exemple d'une modélisation non-linéaire aggravant le taux d'erreur (a) et possibilité théorique d'amélioration d'erreur (b) (d'après Doswell 1986)

L'importance des erreurs dues à la caractérisation varie selon la complexité du paramètrage des caractéristiques de la pluie. Une méthode non-structurée est moins complexe en ce qui concerne l'advection et elle introduit des erreurs dans l'estimation des déplacements prévus. La qualité d'une méthode structurée dépend cependant de la capacité de suivre les structures définies (l'étape de l'appariement). Les erreurs dues à un développement de la pluie peuvent théoriquement être réduites par les méthodes structurées, car une séparation des structures à caractéristiques différentes est essentielle pour une telle approche.

Les erreurs dues au changement de l'advection et des intensités pendant l'intervalle d'échéance de la prévision sont fortement liées, car l'advection de la pluie est influencée par son développement. Elles marquent les limites principales des méthodes d'extrapolation. Ces erreurs de prévision sont causées par la non-connaissance du cycle de vie des phénomènes prévus. Zipser (1990) propose une validité des extrapolations en fonction de la durée de vie du phénomène, limitant l'intervalle maximal de prévision à un quart de la durée de vie. Dans le tableau I.4 sont résumés les intervalles maximales résultants pour quelques phénomènes météorologiques. Des listes plus complètes se trouvent dans Collier (1989) et Doswell (1986).

Phénomène météorologique	Horizon de prévision valide
Cellule orageuse	5 - 20 min
Orage sévère	10 min - 1 h
Orage associée à une perturbation	1 - 2 h
Précipitation frontale	plusieurs heures

Tableau I.4: Intervalle maximal d'extrapolation pour quelques types de pluie (d'après Zipser 1990)

Par rapport aux techniques linéaires d'extrapolation, l'emploi de techniques non-linéaires peut légèrement améliorer les prévisions. Plusieurs auteurs ont cependant souligné que les techniques non-linéaires comportent de forts risques d'aggravation des taux d'erreur (figure I.16). Peu de travaux comparant les taux d'erreur des différentes techniques de prévision ont été publiés. Les études accessibles, comme celle de Elvander (1976) et celle de Einfalt (1988), montrent une forte influence du type de la pluie sur la qualité relative des systèmes de prévision.

I.4 Conclusion

Nous avons exposé le principe de la mesure de pluie par radar et les phénomènes météorologiques, qui sont observables par ce moyen. La technique de mesure et le traitement des données introduisent plusieurs erreurs, dont la plus grande partie a pu être écartée pour cette étude.

Le but de l'étude présentée ici est le développement d'une méthode de prévision de pluie adaptée aux besoins de l'hydrologie urbaine. Nous avons fait ressortir les caractéristiques suivantes de la prévision utile dans ce domaine:

On cherche des prévisions

- **des lames d'eau précipitées,**
- **à court terme de 30 minutes à 2 heures,**
- **avec une résolution de quelques km²,**
- **avec un taux d'erreur acceptable.**

Les modèles numériques ne permettent pas la prévision à une échelle spatiale et temporelle assez fine. Les méthodes d'extrapolation des observations faites par le radar sont à ce jour le seul moyen d'obtenir les prévisions recherchées. Parmi ces méthodes, nous distinguons entre les méthodes non structurées, qui travaillent au niveau de l'image radar entière, et les méthodes structurées, qui travaillent au niveau des structures météorologiques identifiées sur l'image radar.

La plupart des pluies importantes en hydrologie urbaine est caractérisée par une structure diversifiée. Afin d'observer les pluies de façon détaillée et de prendre en compte le cycle de vie des structures météorologiques, une technique automatisée structurée de prévision a été choisie pour cette étude.

Notre objectif est la diminution des deux principales sources d'erreurs de la prévision: l'observation insuffisante de la pluie et la prévision de son développement futur. Le premier objectif concerne les étapes de l'identification et de l'appariement des structures (étapes (1) et (2) de l'algorithme I.2), tandis que le deuxième objectif se rapporte aux étapes de la caractérisation et de la prévision (étapes (3) et (4)).

Les deux objectifs principaux de cette étude sont:

- **L'amélioration de l'observation de la pluie sur l'imagerie radar.**
- **La prise en compte du développement de la pluie pendant l'intervalle d'échéance de la prévision.**

II

MÉTHODES DE L'APPRENTISSAGE AUTOMATIQUE

L'intelligence artificielle peut être définie comme la science de la simulation des processus cognitifs. Ces processus comprennent:

- l'acquisition des connaissances,
- l'archivage des connaissances,
- l'application des connaissances.

On distingue entre ***connaissances procédurales***, qui sont indépendantes du domaine d'application, et ***connaissances spécifiques*** d'un domaine d'application. On cherche à séparer explicitement ces deux types de connaissances dans les systèmes intelligents, afin de permettre une transmission de la connaissance procédurale d'un domaine à un autre.

Les premiers systèmes développés, basés sur une approche généraliste, ont cependant montré une limite de l'abstraction de la connaissance procédurale du problème traité (cf. par exemple Newell et Simon 1963). Au contraire, les approches spécifiques à un domaine ont été plus performant. Un exemple d'une technique très spécifique est le système du jeux de dames, qui a été développé par Samuel (1963). Ce logiciel, dans sa version finale, a gagné la plupart des jeux contre l'homme. Les techniques d'apprentissage et de représentation des connaissances appliquées dans ce système sont cependant trop spécifiques et ne permettent pas une application dans d'autres domaines.

Un autre problème est posé par le fait, que les systèmes d'une représentation peu explicite des connaissances n'en permettent pas la vérification et l'explication, surtout si la connaissance est comprise implictement dans l'algorithme du système.

Aujourd'hui une partie de la recherche en intelligence artificielle se consacre à l'élaboration de ses fondements théoriques. Dans cette axe on cherche à définir d'une façon abstraite les termes "intelligence", "apprentissage" et "connaissance". Malgré les efforts entreprises, aucune théorie satisfaisante n'a été développée à ce jour. McDermott (1987) conclut dans un résumé de l'état de la recherche fondamentale récente, qu'une justification théorique des approches vers les systèmes à base de connaissances et vers l'apprentissage automatique n'a pas encore été formulée.

Le plus grand axe de recherche s'oriente vers l'application, en négligeant ces problèmes fondamentaux. On cherche à développer des systèmes, qui maîtrisent une connaissance correcte, complète et explicite d'un domaine et qui sont performants dans leur application. L'acquisition des connaissances spécifiques du domaine d'application reste un obstacle important pour le développement de tels systèmes. Souvent ces connaissances ne peuvent pas être formulées explicitement par les experts humains et nécessitent des mécanismes d'acquisition automatique des connaissances.

Par conséquent, le travail en intelligence artificielle se concentre en grande partie sur des méthodes d'apprentissage automatique, dont les trois objectifs principaux sont:

- l'analyse théorique de l'espace des méthodes d'apprentissage possibles,
- l'étude et la modélisation des processus humains d'apprentissage,
- l'étude de systèmes d'apprentissage pour la résolution d'un problème donné.

Dans ce chapitre nous examinerons la problématique du développement de systèmes d'apprentissage pour un problème donné, et les solutions proposées à cet objectif.

II.1 Introduction à l'apprentissage automatique

L'apprentissage automatique est un sujet de recherche relativement récent, qui a connu un intérêt croissant dans les dernières deux décennies. Dû à ce fort développement, une théorie générale de cette matière n'est pas encore établie. Les notions utilisées dans la littérature sont en grande partie orientées vers des applications. Par la suite nous introduisons une notion, qui permet la comparaison des différentes techniques de l'apprentissage à un niveau abstrait. Ensuite, nous approfondirons les trois principaux caractéristiques des méthodes d'apprentissage automatique:

- la finalité de l'apprentissage,
- le type de représentation des connaissances acquises,
- la technique d'apprentissage.

II.1.1 Introduction d'une notion générale de l'apprentissage

Définition II.1:

Un ***contexte de connaissances*** est un quadruplet $CT=(O,A,P,S)$, qui est défini par

O : L'ensemble des ***objets*** (ou ***individus***).

A : L'ensemble des ***attributs***, chaque $a \in A: O \rightarrow V$ étant une fonction de l'ensemble des objets dans un espace des valeurs.
Si $V \subset \mathbf{R}$, a est appelé ***attribut linéaire***; si $V \subset \mathbf{N}$ est un ensemble fini, a est appelé ***attribut linéaire qualitatif***, sinon ***attribut linéaire quantitatif***.
Si $V=\{v_1,...,v_k\}$ et V n'est pas hiérarchisé, a est appelé ***attribut nominal***.
Si V est un ensemble hiérarchisé dans une structure autre que linéaire, a est appelé ***attribut structuré***.

P : L'ensemble des ***problèmes***.

S : L'ensemble des ***solutions***.

Pour illustrer cette notion, nous introduisons un exemple très simple d'un problème de classification d'animaux en deux classes:

- *espèces non dangereuses (classe 0)*
- *espèces dangereuses (classe 1)*

Nous définissons le contexte $CT_{EX}=(O_{EX},A_{EX},P_{EX},S_{EX})$ *par*

O_{EX} *={ A1, A2, A3, A4, A5, A6 }*

A_{EX} *={ ESPÈCE(.), TAILLE(.), AGE(.), COULEUR(.), ANIMAL_DE_PROIE(.) } avec*

ESPÈCE(o) $\in$ *{"tigre", "rhinocéros", "lion", "chat", "vache", "poisson rouge"}*

TAILLE(o) $\in$ *{"petit","moyen","grand"}*

AGE(o) $\in \mathbf{R}$

COULEUR(o) $\in$ *{"rouge","jaune","gris","blanc"}*

ANIMAL_DE_PROIE(o) $\in$ *{"faux","vrai"}*

P_{EX} *={ DANGEREUX(o), o*$\in O$ *}*

S_{EX} *={ "faux", "vrai" } = { 0, 1 }*

L'attribut ESPÈCE(.) est un attributs structuré, car une taxonomie existe implicitement pour l'ensemble de ses valeurs (par exemple sont les tigres et les lions des félines, les rhinocéros non), tandis que TAILLE(.) est un attribut linéaire qualitatif, AGE(.) est un attribut linéaire quantitatif, et COULEUR(.) et ANIMAL_DE_PROIE(.) sont des attributs nominaux.

Les valeurs d'attributs des objets sont rassemblées dans le tableau II.1. Cet exemple sera utilisé ultérieurement dans ce chapitre.

objet	*ESPECE*	*TAILLE*	*AGE*	*COULEUR*	*ANIMAL DE_PROIE*	*DANGEREUX*
A1	*tigre*	*moyen*	*5*	*rouge*	*vrai*	*vrai*
A2	*lion*	*moyen*	*10*	*jaune*	*vrai*	*vrai*
A3	*rhinocéros*	*grand*	*50*	*gris*	*faux*	*vrai*
A4	*chat*	*petit*	*1*	*gris*	*vrai*	*faux*
A5	*vache*	*moyen*	*7*	*blanc*	*faux*	*faux*
A6	*poisson rouge*	*petit*	*0.5*	*rouge*	*faux*	*faux*

Tableau II.1: Valeurs d'attributs des objets du contexte de classification d'animaux

Soit $CT=(O,A,P,S)$ un contexte de connaissances. Dans les définitions suivantes, nous considérons des systèmes, qui permettent de résoudre une partie des problèmes de ce contexte et dans lequel les connaissances procédurales sont séparées des connaissances spécifiques du contexte. On appelle l'ensemble des connaissances spécifiques d'un tel système ***base de connaissances***, et l'algorithme de leur application l'***interprétation*** des connaissances. Un problème résoluble par le système sera appelé ***problème admissible*** du système.

Définition II.2:

Un ***système à base de connaissances*** de CT est un quadruplet $K=(CT,\hat{P},BC,IC)$, qui est défini par

$\hat{P}\subset P$: L'ensemble des problèmes admissibles de K.

BC : La base de connaissances de K sur la solution de problèmes $p\in\hat{P}$.

IC : Le système de l'interprétation de connaissances, défini comme fonction $IC:\hat{P}\times BC\rightarrow S$.

La construction d'un système à base de connaissances nécessite la définition d'un ***langage de description***, qui permet la représentation des objets et des connaissances, et l'expression des problèmes et des solutions du contexte CT. Afin d'alléger le texte, nous supposons que ce langage soit implicitement défini par BC et IC.

Si l'on suppose qu'il existe une méthode pour juger, si la solution d'un problème est correcte ou non, nous pouvons définir

Définition II.3:

Un système à base de connaissances $K=(CT,\hat{P},BC,IC)$ est ***consistant***, si

$\forall p\in\hat{P}$: $IC(p,BC)$ est une solution correcte du problème p

Définition II.4:

Un système à base de connaissances $K=(CT,\hat{P},BC,IC)$ est ***complet***, si $\hat{P}=P$.

L'apprentissage automatique est une modification permanente d'un système à base de connaissances, qui utilise un ensemble d'informations afin d'augmenter la performance du système:

Définition II.5:

Soit $CT=(O,A,P,S)$ un contexte de connaissance, $K=(CT,\hat{P},BC,IC)$ un système à base de connaissances et I un ensemble d'informations sur le contexte CT. Un ***apprentissage automatique de connaissances*** AC est une transformation

$$AC : (I, K) \rightarrow K'=(CT,\hat{P}',BC',IC')$$

qui augmente la performance du système relatif à la solution des problèmes dans le contexte donné.

Remarquons que la définition II.5 inclut notamment le cas de l'apprentissage sans connaissance préalable ($BC=\emptyset$ et $\hat{P}=\emptyset$). Par la suite nous supposons que le système d'interprétation IC reste inchangé pendant l'apprentissage AC ($IC' \equiv IC$).

Une grande partie des systèmes d'apprentissage a pour but de générer des descriptions de classes d'objets $c \subset O$ dans un contexte CT. Souvent la notion ***concept*** est utilisée pour une telle classe.

II.1.2 Finalités de l'apprentissage automatique

Le but de l'apprentissage automatique est une augmentation de la performance d'un système à base de connaissances. Selon Fohmann (1985), cette augmentation peut concerner:

- un élargissement quantitatif de la capacité de résoudre des problèmes,
- une amélioration qualitative de la capacité de résoudre des problèmes,
- une augmentation de l'efficacité avec laquelle les problèmes sont résolus.

Soit $AC : (I,K) \rightarrow K'=(CT,\hat{P}',BC',IC)$ un apprentissage de connaissances dans un contexte $CT=(O,A,P,S)$. Un ***élargissement quantitatif*** de la performance de K est une ***généralisation*** de la base de connaissances qui conserve les solutions de K aux problèmes $p \in \hat{P}$. Cette généralisation est effectuée par une modification de la base de connaissances BC tel que

$$\hat{P}' \supseteq \hat{P}$$

et

$$\forall p \in \hat{P} : IC(p,BC) = IC(p,BC')$$

Une ***amélioration qualitative*** de la performance du système K nécessite une fonction de mesure de qualité $Q:(P,S) \rightarrow \mathbb{R}$. Si cette fonction est définie, une amélioration de la qualité des solutions consiste en une modification de la base de connaissances, telle que

$$\hat{P}'=\hat{P}$$

et

$$\forall p \in \hat{P} : Q(p, s=\mathrm{IC}(p,BC)) \leq Q(p, s'=IC(p,BC'))$$

et

$$\exists p \in \hat{P} : Q(p, s=\mathrm{IC}(p,BC)) < Q(p, s'=IC(p,BC'))$$

Une ***augmentation de l'efficacité*** de K concerne des aspects de la complexité du processus de solution d'un problème et de la compréhensibilité de la solution trouvée. Ce but peut être atteint par une modification ou par une réorganisation de la base de connaissances, à condition que les solutions des problèmes restent inchangées. Si $EFF(K)$ est une mesure de l'efficacité du système K, une augmentation de l'efficacité s'exprime comme suit:

$$\hat{P}'=\hat{P}$$

et

$$\forall p \in \hat{P} : IC(p,BC) = IC(p,BC')$$

et

$$EFF(K') > EFF(K)$$

II.1.3 La représentation symbolique de connaissances

Le fonctionnement et la performance d'un système à base de connaissances et de l'apprentissage automatique sont en grande partie déterminés par le type de représentation de la connaissance. Dans ce paragraphe nous rappelons les principales techniques de représentation qui ont été développées en intelligence artificielle. Le classement choisi est inspiré par la présentation de Barr et Feigenbaum (1981).

a) Logique

Dans certains contextes les connaissances s'expriment naturellement comme ensemble de formules logiques. Le domaine de la logique étant très vaste, nous nous referons à la littérature pour une description de son application en intelligence artificielle (par exemple Winston 1977).

Le type de problèmes résolubles par un système logique est la preuve d'hypothèses, étant donné que les formules de la base de connaissances sont vraies. Dans un tel système, l'apprentissage consiste en une déduction de formules et leur intégration dans la base de connaissances.

La représentation de connaissances en termes logiques est précise et modulaire. Autre avantage, les mécanismes d'inférence sont bien définis, ce qui facilite la construction de la base de connaissances et la vérification des solutions.

De nombreux auteurs ont mis en évidence l'importance de la logique pour l'intelligence artificielle (par exemple Nilsson 1980). Le langage de programmation PROLOG à été développé pour faciliter des telles applications. Néanmoins, l'inefficacité de la représentation, les limites en ce qui concerne le traitement des contradictions, les difficultés d'un raisonnement probabiliste, et les limites de l'expression de meta-connaissances (connaissance sur la connaissance) rendent la représentation de connaissances en termes de logique pure souvent impraticable.

b) Frames

Le concept de ***frames*** (cadres) a été proposé par Minsky (1975) comme une manière de représenter des connaissances sous forme de graphes dans lesquels les noeuds représentent des objets ou des concepts et dont les arcs indiquent des relations entre objets.

Un frame est une structure, qui possède des caractéristiques invariables (par exemple un nom) et variables, nommées "slots" (fentes). Ces dernières sont de deux types:

- les slots d'attributs contiennent des valeurs d'attributs pour l'objet représenté par le frame,
- les slots relationnels indiquent des relations entre différents objets.

Par les slots relationnels la base de connaissances forme un graphe. La figure II.1 montre un exemple d'une base de connaissances du contexte CT_{EX} sous forme de frames.

Un système de frames peut donner des solutions aux problèmes concernant des relations entre les objets. L'apprentissage dans un tel système consiste en une modification des valeurs de slots ou en une génération et intégration de nouveaux frames dans le réseau.

La flexibilité structurelle des réseaux de frames permet une grande variété d'interconnexions et ainsi une expression simple et évidente de structures complexes. Pour des grandes quantités de données, la mise à jour de la base de connaissances et la vérification de son contenu posent cependant souvent des problèmes.

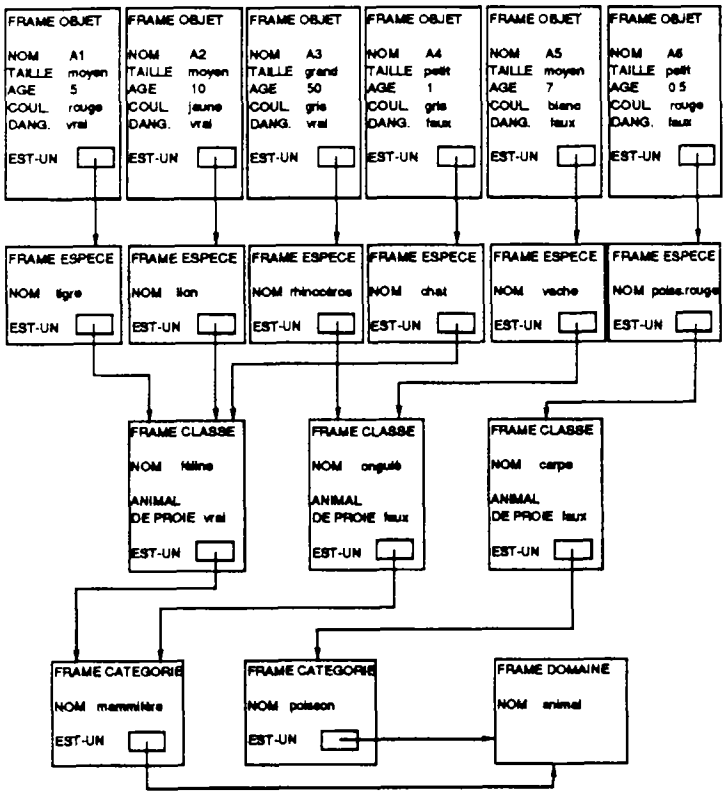


Figure II.1: Description des relations entre les objets du contexte CT_{EX} par un réseau de frames

c) Productions

Dans un système de productions, la base de connaissances est divisée en deux parties (Waterman et Hayes-Roth 1978):

- la mémoire des règles (MR),
- la mémoire de travail (MT).

Cette division est inspirée par la séparation entre la mémoire courte (MT) et la mémoire longue (MR) du cerveau humain. La mémoire de travail contient des faits et des hypothèses sur un problème que le système cherche à résoudre. Un fait est une expression, dont la valeur de vérité est déjà établie, tandis qu'une hypothèse est une expression, dont on cherche à déterminer cette valeur.

La mémoire des règles contient des règles permettant la déduction de nouveau faits à partir de faits déjà connus. Les règles sont représentées sous forme de ***productions***, notion connue de la théorie des langages formels. Une production est composée de deux parties, la condition et la conclusion, dont la signification est: si la condition est vraie, alors la conclusion l'est aussi. Si α est la condition, et β la conclusion de la production p, on écrit $p : \alpha \rightarrow \beta$.

Le système de l'interprétation de la base de connaissances d'un système de productions est appelé ***moteur d'inférences***. Il peut fonctionner selon deux mécanismes:

- appliquer des productions, dont la vérité de la condition est démontrée par les faits contenu dans MT, et ajouter la conclusion comme nouveau fait dans MT (***chaînage avant*** ou "forward chaining"),
- appliquer des productions, dont la conclusion correspond à une hypothèse contenue dans MT, et ajouter la condition comme nouvelle hypothèse dans MT (***chaînage arrière*** ou "backward chaining").

Généralement les productions contiennent des variables, et leur application nécessite une ***unification*** (ou ***instanciation***) des variables et des constantes contenu dans les règles et les faits ou hypothèses. Ce processus peut être défini comme suit:

- une constante s'unifie avec la même constante,
- une variable s'unifie avec une constante,
- une variable s'unifie à une autre variable, qui n'est pas encore unifiée avec une constante; si une d'elles est après unifiée avec une constante, l'autre l'est immédiatement aussi.

Si $\alpha \rightarrow \beta$ est une production contenant les variables $(x_1,\ldots,x_n)$, et $\Im : (x_1 \rightarrow c_1,\ldots,x_n \rightarrow c_n)$ est une instanciation des variables, nous utilisons la notion $\alpha^{\Im} \rightarrow \beta^{\Im}$ pour désigner la production instanciée. L'algorithme II.1 montre le fonctionnement d'un moteur d'inférences en chaînage avant.

Donné : Une mémoire de règles MR.
Une mémoire de travail MT.
Un ensemble F de faits $F=\{f_1,\ldots f_n\}$.

Cherché : La conséquence des faits F

Algorithme :

(0) MT := F

(1) $\tilde{R}$:= { $\alpha_k \rightarrow \beta_k \in$ MR avec: il existe une instanciation $\Im_k$ de $\alpha_k \rightarrow \beta_k$ telle que $\alpha_k^{\Im_k}$ est vérifié par MT et $\beta_k^{\Im_k} \notin$ MT }

(2) Si $\tilde{R} = \varnothing$ STOP

Sinon

(3) Choisir un $(\alpha_k^{\Im_k} \rightarrow \beta_k^{\Im_k}) \in \tilde{R}$

(4) MT := MT $\cup\ \beta_k^{\Im_k}$

(5) Continuer avec (1)

Algorithme II.1: Moteur d'inférences en chaînage avant

Dans le contexte CT_{EX}, un système de productions $K=(CT_{EX},P,BC,IC)$ peut être défini comme suit:

$BC = \{$ p_1: *(TAILLE(o)="moyen"* $\wedge$ *ANIMAL_DE_PROIE(o)="vrai")* $\rightarrow$ *(CATÉGORIE(o)="féline")*

p_2: *(CATÉGORIE(o)="féline"* $\wedge$ *AGE(o)>3)* $\rightarrow$ *(DANGEREUX(o)="vrai")*

p_3: *(TAILLE(o)="petit")* $\rightarrow$ *(DANGEREUX(o)="faux")* $\}$

Il est alors

$\hat{P} = \{ DANGEREUX(o), o \in O \wedge$
$(TAILLE(o)="petit" \vee (TAILLE(o)="moyen" \wedge ANIMAL_DE_PROIE(o)="vrai")) \}$

K est un système non complet, car par exemple une réponse au problème DANGEREUX(A3) est impossible. La résolution du problème DANGEREUX(A2) est, sous application de l'algorithme II.1, déterminé de la façon suivante:

(0) $MT = \{ ESPECE(A2)="lion", TAILLE(A2)="moyen", AGE(A2)=10,$
$COULEUR(A2)="jaune", ANIMAL_DE_PROIE(A2)="vrai" \}$

(1) $\tilde{R} := \{ p_1$ *avec l'instanciation* $(o \rightarrow A2) \}$

(2) $\tilde{R} \neq \emptyset$

(3) *choisir* p_1

(4) $MT := MT \cup \{ CATÉGORIE(A2)="féline" \}$

(1) $\tilde{R} := \{ p_2$ *avec l'instanciation* $(o \rightarrow A2) \}$

(2) $\tilde{R} \neq \emptyset$

(3) *choisir* p_2

(4) $MT := MT \cup \{ DANGEREUX(A2)="vrai" \}$

(1) $\tilde{R} = \emptyset$

(2) *STOP*

Un résumé des systèmes de productions se trouve dans Davis et King (1977). Ils caractérisent les contextes bien appropriés pour l'emploi des systèmes de productions comme suit:

- la connaissance existante du contexte est diffuse,
- les processus du contexte peuvent être représentés comme un ensemble d'actions indépendantes,
- la connaissance du contexte est facilement séparable en règles et faits.

L'apprentissage dans les systèmes de productions consiste en une modification de la mémoire des règles, en ajoutant, supprimant ou modifiant des règles. Les avantages de la représentation des connaissances sous forme de règles sont multiples:

- la connaissance est organisée de façon très modulaire, facilitant l'ajout et la suppression des règles,
- les solutions données s'expliquent par la liste des règles appliquées, les solutions sont alors vérifiables et des règles fausses peuvent être détectées,
- un raisonnement probabiliste est possible par l'ajout d'un coefficient de probabilité à chaque règle.

Cependant, de grandes quantités de connaissances sont difficiles à gérer, car le choix des productions applicables (étape 1 de l'algorithme II.1) nécessite l'évaluation de toutes les conditions avec toutes instanciations. Un autre inconvénient est que l'expression des connaissances algorithmiques par des productions est difficile.

Les systèmes de productions ont trouvé des applications sous forme de systèmes experts (cf. Hayes-Roth et *al.* (1983) pour un résumé des techniques et applications des systèmes experts).

Un ensemble de productions peut être représenté dans la forme d'un arbre, ce qui permet leur apprentissage itératif d'une manière très efficace. Bien que cette technique réduise l'enchaînement possible des productions, elle est bien adaptée à des contextes moins complexes. Nous approfondissons cette technique dans le paragraphe II.2.

d) Réseaux de neurones

La technique de représentation des connaissances sous forme de réseaux de neurones est inspirée par le modèle biologique du cerveau humain. Le cerveau est essentiellement composé d'éléments de base appelés ***neurones***, qui possèdent chacun plusieurs entrées (synapses) et une sortie (axone) qui peut former l'entrée d'autres neurones. Les neurones, dont le nombre est estimé a 10^{11}, sont ainsi interconnectés de manière très complexe. Cette complexité rend une modélisation directe de neurones biologiques sur ordinateur impossible.

Les réseaux de neurones artificiels se constituent de neurones formels, qui sont définis comme suit.

Définition II.6:
Un ***neurone formel*** est une fonction

$$nf : \mathbb{R}^n \rightarrow (-1,+1)$$
$$(e_1,\ldots,e_n) \rightarrow f_{nf}\left(\sum_{j=1}^{n} c_j e_j - \theta\right)$$

Les arguments de nf sont appelés les ***entrées***, l'image des entrées la ***sortie***, et θ le ***seuil*** du neurone nf.

Des exemples de différentes types de la fonction f_{nf} se trouvent dans figure II.2. Un réseau de neurones est formé par un ensemble de neurones formels, les entrées et sorties des neurones étant interconnectés. La connexion complète (figure II.3a) n'est souvent pas réalisable pour les réseaux comportant un nombre élevé de neurones. Les structures les moins complexes sont les réseaux en une seule couche (figure II.3b) ou en plusieurs couches (figure II.3c). Dans toutes les structures une partie des neurones possède des entrées connectées à l'environnement (neurones d'entrée), et une autre partie possède des sorties accessibles à l'environnement (neurones de sortie).

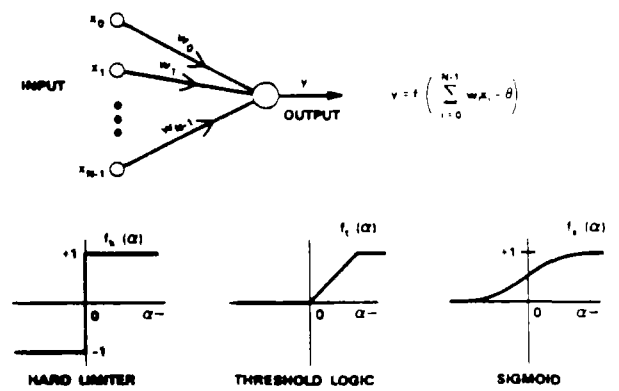


Figure II.2: Neurone formel à n entrées et trois fonctions-types de neurones (d'après Lippmann 1987)

Le type de problèmes résolubles par un réseau de neurones, $\{nf_1,\ldots,nf_k\}$, est la classification d'objets décrits par des attributs linéaires quantitatifs, chacun en rélation biunivoque avec une entrée du réseau. Souvent les valeurs de sortie sont limitées aux valeurs 0 et 1, et le nombre de neurones de sortie est égal au nombre de classes d'un concept. On cherche à déterminer les paramètres c_{ij} de neurones nf_i tels que la sortie 1 à un neurone de sortie indique l'appartenance de l'objet à une classe précise.

L'apprentissage dans un réseau de neurones consiste en une modification des paramètres c_{ij} des neurones nf_i selon une règle d'apprentissage, afin d'obtenir le comportement souhaité du réseau. La règle d'apprentissage est basée sur la mesure de différence entre la valeur de sortie souhaitée et la valeur obtenue par des exemples.

Un avantage de cette méthode de représentation des connaissances est, que dans un réseau en couches l'interprétation et l'apprentissage des connaissances peuvent être effectués d'une façon parallèle, ce qui permet un modélisation efficace des neurones sur des ordinateurs multiprocesseurs.

Les inconvénients de cette technique sont liés à la pauvreté du langage de description des connaissances. La connaissance acquise par apprentissage automatique est difficilement compréhensible pour l'homme, une vérification de la connaissance est alors souvent impossible.

Des applications de réseaux de neurones se trouvent dans des contextes qui nécessitent une assimilation de grandes quantités d'informations, par exemple la reconnaissance de la parole et la vision automatique.

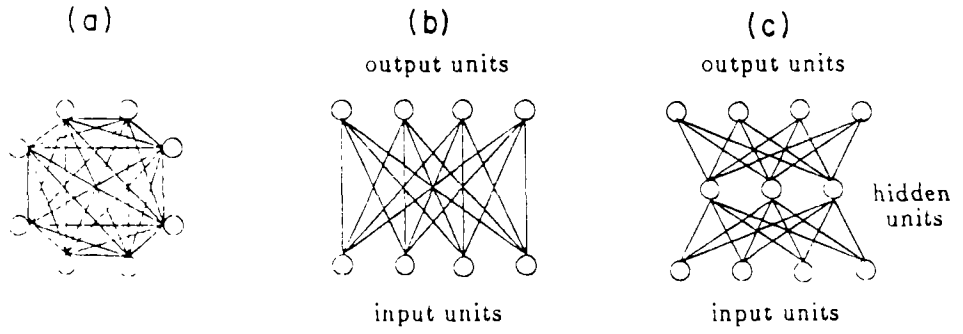


Figure II.3: Différentes architectures de réseaux de neurones: (a) réseau complètement connecté, (b) réseau en une couche, (c) réseau en multicouches (d'après Matheus et Hohensee 1987)

e) Comparaison des techniques de représentation de connaissances

La technique de représentation des connaissances soulève des contraintes à la recherche des solutions dans un contexte donné. Ces contraintes font partie de ce que Mitchell (1982) a appelé le "learning bias", qui restreint à priori l'apprentissage de connaissances dans le contexte. Il est alors primordial de choisir la technique de représentation (et le langage de description) en fonction des caractéristiques du contexte: le type du problème, les types d'attributs, et la quantité d'informations. Un autre critère important est la compréhensibilité des solutions proposées par le système. Le tableau II.2 résume le degré d'adaptation des quatre techniques examinées à des exigences différentes.

Technique	Type de problème	Type d'attribut			Contextes incertains	Grandes quantités de données	Compréhens. de la solution
		nominal	linéaire	structuré			
logique	déduction	+	0	—	—	0	0
frames	relations	+	+	+	—	—	+
productions	déduction	+	+	0	+	0	+
neurones	classification	—	+	—	0	+	—

Tableau II.2: Comparaison de quatre techniques de représentation de connaissances
(+ : bien adapté 0 : moins adapté — : non adapté)

II.1.4 Techniques d'apprentissage

Souvent on utilise un classement des techniques d'apprentissage selon la complexité de la tâche effectuée par un système d'apprentissage de connaissances *AC* (cf. par exemple Carbonell et *al.* 1983). Cette complexité est déterminée par l'inférence effectuée par le système afin de transformer l'information *I*, qui est accessible au système, en connaissances dans la représentation interne. Dans la suite, $K=(CT,\hat{P},BC,IC)$ représente un système à base de connaissances qui est transformé en $K'=(CT,\hat{P}',BC',IC)$ par *AC*.

a) Implantation de la connaissance

L'apprentissage par implantation est l'apprentissage "par coeur" de connaissances. L'information *I* est sous une forme directement utilisable dans la base de connaissances du système. L'unique tâche de l'apprentissage est la mémorisation de connaissances.

$$AC_{implan} : (I,K) \rightarrow K'=(CT,\hat{P}', \mathbf{BC'=BC \cup I}, IC)$$

Un exemple de ce type d'apprentissage est la programmation d'un ordinateur ou la modification directe d'une base de données.

b) Apprentissage par instruction

Comme dans le cas d'implantation, toute la connaissance est fournie explicitement au système, mais sous une forme qui nécessite une transformation *t* de l'information fournie en termes de représentation dans la base de connaissances.

$$AC_{instruc} : (I,K) \rightarrow K'=(CT,\hat{P}', \mathbf{BC'=BC \cup t(I)}, IC)$$

La formulation explicite de la connaissance reste la tâche d'un instructeur. Un exemple de l'apprentissage par instruction est la définition de règles d'un système expert.

c) Apprentissage à partir d'exemples

Souvent la définition explicite et complète de la connaissance, nécessaire pour les deux types d'apprentissage mentionnés ci-dessus, n'est pas praticable. Un système d'apprentissage à partir d'exemples génère la connaissance de façon inductive à partir de données, qui représentent des exemples positifs ou ***instances*** et des exemples négatifs ou ***non-instances*** d'un concept. La tâche de l'instructeur est le choix des exemples présentés au système et l'ordre de leur présentation.

Le but de l'apprentissage est la génération d'une description d'un concept assez général pour couvrir tous les exemples positifs, et assez spécifique pour exclure les exemples négatifs. Si les exemples ne sont pas tous accessibles à la fois, le système *AC* doit alors disposer de deux mécanismes:

- Généralisation de *BC* pour le cas où un exemple positif est trouvé, qui n'est pas couvert par le concept représenté par *BC*.
- Spécification de *BC* pour le cas où un exemple négatif est trouvé, qui est couvert par le concept représenté par *BC*.

Le fonctionnement d'un système d'apprentissage à partir d'exemples est défini par

$$AC_{exemp} : (I,K) \rightarrow K'=(CT,\hat{P}', \mathbf{BC'=BC \cup f(BC,I)}, IC)$$

Des exemples d'application de cette technique sont la génération des règles par induction d'arbres de décision (Quinlan 1986), l'apprentissage dans les réseaux de neurones (Lippmann 1987) et la génération de réseaux sémantiques à partir d'exemples donnés (Winston 1975).

d) Apprentissage par observation

L'apprentissage par observation est un apprentissage à partir d'exemples qui ne sont ni choisis ni structurés par un instructeur. La tâche d'apprentissage est le développement de concepts et la génération de la description de concepts à partir d'observations aléatoires ou influencées par le système (apprentissage par expérimentation). La génération de concepts (***groupement conceptuel***) nécessite une connaissance globale du contexte *CT*, pour permettre la définition de concepts ayant une signification dans *CT*.

$$AC_{observ} : (I,K) \rightarrow K'=(CT,\hat{P}', \mathbf{BC'=BC \cup g(BC,I,CT)}, IC)$$

La plus grande partie des systèmes de groupement conceptuel développés génèrent une structure hiérarchique de concepts qui inclue des degrés différents d'abstraction (par exemple Fisher 1987, Ganascia 1987, Lebowitz 1987, Michalski et Stepp 1986).

II.2 Apprentissage inductif des arbres de décision

Les problèmes de classification forment une partie importante des contextes de connaissances en intelligence artificielle. Pour ces problèmes la représentation de connaissances sous forme d'un arbre de décision est particulièrement efficace. Par la suite nous examinons cette technique sous les aspects de l'apprentissage et de l'application de connaissances à des problèmes réels.

II.2.1 Génération et application des arbres de décision

Le concept de la représentation de connaissances sous forme d'arbres de décision a été développé par Hunt et *al.* (1966). Des systèmes de génération d'arbres de décision ont été les sujets de recherche dans une multitude d'applications. Les systèmes développés sont entre autres ID3 (Quinlan 1984,1986), CART (Breiman et *al.* 1984), et ASSISTANT (Cestnik et *al.* 1987).

II.2.1.1 Définitions

Définition II.7:
Un contexte $CT=(O,A,P,S)$ est appelé ***contexte de classification***, si les problèmes $p \in P$ sont des tâches de classification des objets $o \in O$ en partition $S=\{c_1,\ldots,c_k\}$, avec

$$\bigcup_{i=1}^{k} \{o \in O, c(o)=c_i\} = O \quad et \quad \forall i \neq j \quad \{o \in O, c(o)=c_i\} \cap \{o \in O, c(o)=c_j\} = \emptyset$$

où $c(o)=c_i$ signifie que l'objet $o \in O$ appartient à la ***classe*** c_i.

Définition II.8:
Un ***arbre*** est un graphe directionnel non bouclé dont l'ensemble N de noeuds est fini, et dans lequel un chemin unique existe entre un noeud étiqueté n_0 et chaque noeud $n \in N$. Soit deux noeuds n_1 et n_2 de l'arbre:

- S'il existe un chemin de n_1 à n_2, n_1 est ***prédécesseur*** de n_2 et n_2 est ***successeur*** de n_1.
- S'il existe une jonction entre n_1 et n_2 en direction de n_2, n_1 est ***parent*** de n_2 et n_2 est ***enfant*** de n_1.
- Si n_1 n'a pas d'enfants, n_1 est appelé ***noeud terminal*** ou ***feuille***.
- Si n_1 n'a pas de parents, n_1 est égal à n_0 et est appelé ***racine***.

Définition II.9:
Soit $CT=(O,A,P,S)$ un contexte de classification. Un ***arbre de décision*** ADD de CT est un arbre possédant les propriétés suivantes:

- à chaque noeud non terminal n est associée une fonction de ***test***

  $f^n : (O,A) \rightarrow \{ n_j, n_j \text{ est enfant de } n \}$
- à chaque noeud terminal n est associé un élément $c^n \in S$

Dans la plus grande partie des applications, les fonctions de test sont restreintes à l'évaluation d'un seul attribut, dont les valeurs possibles sont coordonnées aux enfants du noeud. Cet attribut est appelé ***attribut de test*** du noeud.

La figure II.4 montre un exemple d'un arbre de décision du contexte CT_{EX} de classification d'animaux (cf. page 33).

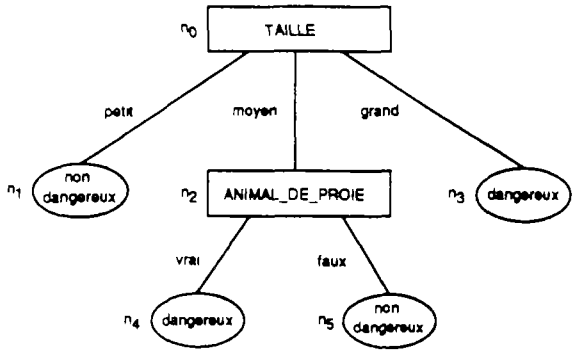


Figure II.4: Arbre de décision pour le problème de classification d'animaux en espèces non dangereuses (classe 0) et espèces dangereuses (classe 1)

Donné : Un contexte de classification $CT=(O,A,P,S)$ avec $A=\vec{a}=\{a_1,\ldots,a_k\}$ un ensemble d'attributs nominaux.
Un ensemble X_A d'objets, leurs valeurs d'attributs et leur classe
$X_A = \{ \{o_1, \vec{a}(o_1), c(o_1)\} ,\ldots, \{o_l, \vec{a}(o_l), c(o_l)\} \}$
appelé l'***ensemble d'exemples d'apprentissage***.

Cherché : Un arbre de décision *ADD* qui classe l'ensemble X_A correctement:
$\forall o \in X_A : F_{ADD}(o,\vec{a}(o)) = c(o)$

Algorithme :

(0) Générer la racine $n:=n_0$ et associer X_A à la racine, $X_A^n:=X_A$

(1) Si tous les exemples associés à n sont de la même classe c_i, transformer n en noeud terminal avec $c^n:=c_i$

Sinon

(2) Choisir un attribut de test $a^n \in A$ du noeud n

(3) Générer un noeud enfant de n pour chaque valeur de a^n

(4) Diviser X_A^n selon leurs valeurs de a^n

(5) Associer les sous-ensembles aux noeuds enfants correspondants

(6) Continuer avec (1) pour chaque noeud enfant de n

Algorithme II.2: Algorithme de ID3 de génération d'un arbre de décision

Un arbre de décision *ADD* définit une fonction $F_{ADD} : (O,A) \rightarrow S$, qui associe à chaque objet $o \in O$ une classe $c \in S$ et représente ainsi la base de connaissances d'un système $K=(CT,\hat{P},BC,IC)$ du contexte de classification.

Un arbre de décision est une forme spécifique de représentation d'un ensemble de productions, car chaque chemin de la racine n_0 à un noeud terminal correspond à une production

$$p : (f^{n_0}(o,\vec{a}(o)) = n_{i_1}) \wedge (f^{n_{i_1}}(o,\vec{a}(o)) = n_{i_2}) \wedge \ldots \wedge (f^{n_{i_{k-1}}}(o,\vec{a}(o)) = n_{i_k}) \rightarrow c(o) = c^{n_{i_k}}$$

II.2.1.2 La génération des arbres de décision

L'utilité des arbres de décision comme forme de représentation de connaissances est liée à la possibilité de générer l'arbre de manière inductive par apprentissage à partir d'exemples. L'algorithme II.2 rappelle le principe de cette technique, qui est utilisée dans ID3 (Quinlan 1986).

Le choix de l'attribut de test a_n associé à un noeud n (étape (2) de l'algorithme II.2) est effectué selon une mesure de l'impureté des noeuds.

Définition II.10:
Soit $CT=(O,A,P,S)$ un problème de classification de k classes, ADD un arbre de décision pour CT, n un noeud d'ADD, X_A^n l'ensemble d'exemples d'apprentissage associé à n, et p_i la fréquence de la classe i dans X_A^n:

$$p_i = \frac{|\{o \in X_A^n \wedge c(o)=c_i\}|}{|X_A^n|} \quad (i=1,\ldots,k)$$

où $|.|$ signifie le cardinal d'un ensemble. Une ***mesure d'impureté*** de n est une fonction

$$\Phi : \{(p_1,\ldots,p_k) \in [0,1]^k, \Sigma p_i = 1\} \rightarrow [0,1]$$

avec:

- $\Phi(p_1,\ldots,p_k)$ maximal $\Leftrightarrow \forall i : p_i = 1/k$
- $\Phi(p_1,\ldots,p_k) = 0 \Leftrightarrow \exists j : p_j = 1$ et $\forall i \neq j : p_i = 0$
- $\forall s,t : \Phi(p_1,\ldots,p_t,\ldots,p_s,\ldots,p_k) = \Phi(p_1,\ldots,p_s,\ldots,p_t,\ldots,p_k)$

La mesure d'impureté le plus souvent utilisé (Quinlan 1986, Cestnik et *al.* 1987) est la mesure d'***entropie***

$$\Phi_{entrop}(p_1,\ldots,p_k) = -\sum_{j=1}^{k} p_j \log_2 p_j \quad (\text{avec } 0 \log_2 0 := 0)$$

qui est basé sur la théorie de l'information (cf. par exemple Volle 1985). La mesure de ***gini*** est proposé par Breiman et *al.* (1984)

$$\Phi_{gini}(p_1,\ldots,p_k) = \sum_{j=1}^{k} \sum_{\substack{i=1 \\ i \neq j}}^{k} p_i p_j$$

Étant donné une mesure d'impureté de Φ, on exprime la réduction de l'impureté par le test a_i comme

$$\Delta\Phi(n,a_i) = \Phi(n) - \sum_{j=1}^{k_i} \frac{|X_A^{n_j^i}|}{|X_A^n|} \Phi(n_j^i)$$

où n_j^i dénote le $j^{ième}$ enfant de n généré par la division de l'ensemble d'apprentissage X_A^n en k_i sous-ensembles selon a_i comme attribut de test. Dans l'étape (2) de l'algorithme II.2, l'attribut a_i maximisant la réduction de l'impureté est choisi comme attribut de test a^n du noeud n.

Dans cette étude, nous utiliserons la mesure d'entropie comme mesure d'impureté. Par la suite, nous noterons simplement Φ au lieu de Φ_{entrop}.

Le fonctionnement de l'algorithme II.2 sera illustré à l'aide de l'exemple de classification d'animaux. A la racine de l'arbre, la réduction de l'impureté (mesurée par l'entropie) se calcule pour l'attribut TAILLE selon

$$\Delta\Phi(n_0, TAILLE) = 1 - (\frac{2}{6}(-\frac{2}{2}\log\frac{2}{2} - \frac{0}{2}\log\frac{0}{2}) + \frac{3}{6}(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}) + \frac{1}{6}(-\frac{0}{1}\log\frac{0}{1} - \frac{1}{1}\log\frac{1}{1}))$$
$$= 0.54$$

Cette réduction est supérieure à celle obtenue par l'attribut de test ANIMAL_DE_PROIE, qui se calcule selon

$$\Delta\Phi(n_0, ANIMAL_DE_PROIE) = 1 - (\frac{1}{2}(-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}) + \frac{1}{2}(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3})) = 0.08$$

L'attribut TAILLE sera alors préféré à ANIMAL_DE_PROIE comme attribut de test pour la racine de l'arbre.

II.2.1.3 Classification d'objets avec des arbres de décision

La classification d'un objet o par un arbre de décision est effectué par la recherche du chemin de la racine à une feuille d'arbre, sur lequel tous les tests correspondent aux valeurs des attributs de o. Cette recherche est réalisée selon l'algorithme II.3.

Donné : Un contexte de classification $CT=(O,A,P,S)$.
Un arbre de décision ADD pour CT.
Un objet $o \in O$ et le vecteur de valeurs d'attributs $\vec{a}(o)$.

Cherché : La classe $c(o)$ de l'objet o.

Algorithme :

(0) Noeud n := racine de ADD

(1) Si n est un noeud terminal $c(o) := c^n$, c^n étant la classe associée à n
STOP

Sinon

(2) $n := a^n(o)^{\text{ième}}$ enfant de n, a^n étant l'attribut de test de n

(3) Continuer avec (1)

Algorithme II.3: Algorithme de classification d'objets par un arbre de décision

La classification par l'algorithme II.3 de l'objet A3 du contexte CT_{EX}, en utilisant l'arbre montré dans la figure II.4, est effectuée comme suit:

(0) $n := n_0$

(1) n n'est pas terminal

(2) TAILLE(A3)="moyen" $\Rightarrow$ $n := n_2$

(1) n n'est pas terminal

(2) ANIMAL_DE_PROIE(A3)="vrai" $\Rightarrow$ $n := n_4$

(1) n est terminal $\Rightarrow$ $c(A3)=c^n$*=classe 1 (dangereux)*
STOP

Les arbres de décision sont employés pour la résolution de deux types de problèmes (Breiman et *al.* 1984):

- comme classificateur de l'ensemble d'apprentissage X_A,
- comme prédateur de la classification d'objets $o \in O$ dont la classe $c(o)$ est inconnue.

Dans le premier cas, la finalité de l'apprentissage est une grande efficacité de la base de connaissances, qui s'exprime surtout dans la taille de l'arbre généré. Dans le deuxième cas la qualité de la classification en termes de taux d'erreur est plus important que l'efficacité. Le taux d'erreur peut être estimé à l'aide d'un ensemble d'exemples test Y, avec $X_A \cap Y = \emptyset$.

II.2.2 Sources des difficultés de l'emploi des arbres de décision pour résoudre des problèmes réels de classification

La qualité descriptive ou prédictive d'un arbre de décision généré à partir d'exemples dépend de l'ensemble des exemples d'apprentissage et de l'ensemble des attributs. Il est donc primordial de choisir des exemples représentatifs de la tâche de classification et de définir les attributs importants pour la description des objets et significatifs par rapport à la classification.

L'application à des problèmes réels des algorithmes décrits rencontre souvent des difficultés, qui sont dues aux contraintes de la méthode, non seulement pendant la génération de l'arbre, mais aussi pendant la classification des objets. Par la suite nous examinons les principaux obstacles et nous étudions les moyens d'améliorer la méthode.

II.2.2.1 Difficultés liées à l'ensemble des attributs

a) Attributs pouvant prendre un grand nombre de valeurs

Quinlan (1986) remarque que la mesure d'impureté de l'entropie favorise les attributs pouvant prendre un grand nombre de valeurs par rapport aux attributs prenant un nombre plus petit de valeurs. Ce comportement n'est pas souhaitable, car les attributs prenant un grand nombre de valeurs divisent l'ensemble d'exemples d'apprentissage en petits sous-ensembles, dont la fréquence des classes peut être non représentative pour l'ensemble des objets.

Dans le contexte CT_{EX}, la réduction de l'impureté à la racine de l'arbre par l'attribut TAILLE est de 0.54 (cf. page 47). L'impureté du noeud enfant correspondant à la valeur "moyen" est de

$$x = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.92$$

Si un attribut TAILLE' est défini avec quatre valeurs ("petit", "moyen$_1$", "moyen$_2$", "grand"), et p_i exemples ont la valeur "moyen$_i$", dont p_i^d sont dangereux et p_i^n ne le sont pas (i=1,2), l'impureté des deux noeuds enfants de n_0 correspondant aux valeurs "moyen$_1$" et "moyen$_2$" est

$$x_i = -\frac{p_i^d}{p_i}\log_2\frac{p_i^d}{p_i} - \frac{p_i^n}{p_i}\log_2\frac{p_i^n}{p_i} \qquad (i=1,2)$$

Si l'on exclue le cas trivial $p_1=0$, il n'y a que deux cas différents:

(1) $p_1^d=1, p_1^n=0, p_2^d=1, p_2^n=1 \Rightarrow x_1=0, x_2=1 \Rightarrow$

$$\Delta\Phi(n_0, TAILLE') = 1 - (\frac{2}{6}\cdot 0 + \frac{1}{6}\cdot 0 + \frac{2}{6}\cdot 1 + \frac{1}{6}\cdot 0) = 0.66$$

(2) $p_1^d=0, p_1^n=1, p_2^d=2, p_1^n=0 \Rightarrow x_1=0, x_2=0 \Rightarrow$

$$\Delta\Phi(n_0, TAILLE') = 1 - (\frac{2}{6}\cdot 0 + \frac{1}{6}\cdot 0 + \frac{2}{6}\cdot 0 + \frac{1}{6}\cdot 0) = 1.$$

L'attribut TAILLE' sera alors dans tous cas préféré à l'attribut TAILLE.

Une solution de ce problème est la limitation du nombre de valeurs à une valeur k fixe. Pour une valeur de k égale à 2 cette limitation implique un arbre de décision binaire, chaque noeud non terminal ayant deux enfants. Cestnik et *al.* (1987) appliquent cette technique. Quinlan (1986) constate que cette restriction diminue l'efficacité de l'arbre construit, car un arbre binaire est d'une hauteur plus grande qu'un arbre qui n'est pas contraint à deux enfants par noeud. Il propose une équilibrage des attributs prenant différents nombres de valeurs par une modification du critère de choix d'attributs de test, en introduisant une mesure IV de la valeur d'attributs pour la réduction de l'impureté, qui est dans le cas d'un attribut a avec k valeurs $\{v_1,...,v_k\}$ défini comme

$$IV(n,a) = -\sum_{i=1}^{k} \frac{|\{o \in X_A^n, a(o)=v_i\}|}{|X_A^n|} \log_2 \frac{|\{o \in X_A^n, a(o)=v_i\}|}{|X_A^n|}$$

La valeur d'un test sur l'attribut a associé au noeud n est défini comme

$$Q(n,a) = \frac{\Delta\Phi(n,a)}{IV(n,a)}$$

Dans l'exemple donnée ci-dessus, il est

$IV(n_0, TAILLE) = 1.46$ *et* $Q(n_0, TAILLE) = 0.37$

tandis qu'il est

$IV(n_0, TAILLE') = 1.92$ *et* $Q(n_0, TAILLE') = 0.34$ *(cas 1)*

$Q(n_0, TAILLE') = 0.52$ *(cas 2)*

L'attribut TAILLE' sera alors préféré à l'attribut TAILLE uniquement dans le deuxième cas, dans lequel les deux classes sont parfaitement séparées.

Quinlan constate une performance améliorée par rapport à l'emploi du critère d'impureté basé sur la réduction de l'entropie seule. Un choix préalable d'attributs est cependant nécessaire, car la valeur $Q(n,a)$ n'est pas définie pour tous attributs ($IV(.)$ peut être égale à 0).

Les attributs prenant un grand nombre de valeurs posent un autre problème non résolu par cette modification: le fait qu'une partie des valeurs d'attributs peut être insignifiante pour la classification. En particulier dans le cas d'un petit nombre d'exemples d'apprentissage, l'arbre de décision peut être divisé de manière inutile. Cheng et *al.* (1988) proposent un test vérifiant si les valeurs d'attributs sont significatives. Toutes les valeurs ne dépassant pas un seuil de signification sont rassemblées dans une valeur par défaut. Néanmoins, la difficulté de définir ce seuil d'acceptation subsiste.

b) Attributs d'une signification limitée

Les tests permis dans l'algorithme de ID3 sont limités aux tests sur un seul attribut, et une introduction de tests plus généraux comme "$2a_i(o)+a_j(o)=v$" ou "$a_i(o)=v_1 \wedge a_j(o)<s$" est exclue dans cet algorithme. La découverte de règles efficaces de discrimination n'est possible que si certaines (au moins un) des attributs définis ont une influence sur l'appartenance d'un individu à une classe donnée.

Afin de permettre la découverte de structures linéaires dans les données, Breiman et *al.* (1984) ont admis des tests de la forme "$\Sigma f_i a_i < s$" dans l'arbre, qui combinent tout attributs linéaires. Ils ont proposé un algorithme de recherche de la meilleur combinaison, et reportent des résultats satisfaisants de cette technique. Il n'est cependant pas exclu, que cet algorithme génère des combinaisons sans signification. L'intégration dans le système d'une description très sophistiquée du contexte (relations entre les attributs, les valeurs, les règles, *etc.*), comme par exemple réalisée dans le système CHARADE (Ganascia 1987), est nécessaire pour permettre la découverte automatique des combinaisons significatives entre attributs.

c) Attributs linéaires et attributs structurés

L'algorithme décrit d'induction des arbres de décision permet uniquement l'utilisation d'attributs nominaux comme attributs de test. Dans des problèmes réels de classification, les objets sont cependant souvent décrits par des attributs linéaires ou structurés (valeurs de mesure, nombre d'observations, etc.).

Les attributs linéaires qualitatifs et les attributs structurés peuvent être traités directement comme attributs nominaux si le nombre de leurs valeurs est fini. Ce traitement provoque cependant une perte importante d'informations, car la description des relations entre les objets implicitement comprises dans la taxonomie des valeurs n'est pas conservée.

Une transformation d'attributs linéaires quantitatifs en attributs nominaux nécessite une discrétisation de l'espace des valeurs des attributs par une partition. Si $a \in A$ est un attribut qualitatif dont les valeurs se trouvent dans l'intervalle réel (r_1,r_2), $r_1,r_2 \in \mathbb{R} \cup \{\infty\}$, une partition $r = \{(r_{11},r_{12}),\ldots,(r_{k1},r_{k2})\}$ de (r_1,r_2) définit un attribut linéaire qualitatif $a'(.) = r(a(.))$, qui peut être traité comme attribut nominal. A part des pertes d'informations déjà mentionnées, cette transformation donne lieu à une altération de la signification de l'attribut qui est déterminée par le choix du nombre et des valeurs des seuils r_{ij}. Un mauvais choix des seuils peut ainsi fortement diminuer l'importance de l'attribut pour la classification.

Pour éviter ce problème on peut procéder à une optimisation du choix de la partition. Si n est un noeud terminal d'un arbre de décision, X_A^n l'ensemble d'apprentissage associé à n, $a_j \in A$ un attribut qualitatif, et R_{aj} l'ensemble fini de partitions de $a_j(X_A^n) = \{a_j(o_{i1}),\ldots,a_j(o_{in})\}$, on choisit la partition $r \in R_{aj}$ qui définit l'attribut a_j' de façon optimale par rapport au critère de choix d'attributs dans l'algorithme d'induction de l'arbre.

Pour certains mesures d'impureté, il faut définir un nombre maximal de classes de la partition pour éviter des partitions trop fines. Un nombre maximal de 2 génère des partitions selon une valeur de seuil et implique des arbres de forme binaire.

La figure II.5 montre un arbre de décision du contexte CT_{EX}, dans lequel l'attribut linéaire quantitatif AGE(.) est transformé en attribut nominal. Dans cet exemple, un seuil a été introduit, qui divise l'ensemble des valeurs de l'attribut en deux sous-ensembles.

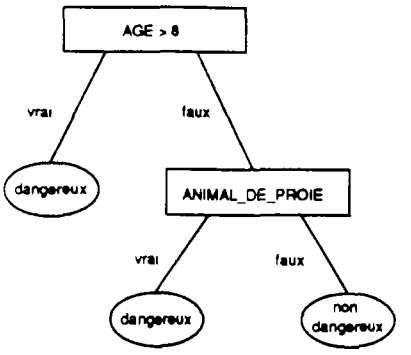


Figure II.5: Arbre de décision du contexte CT_{EX}, dans lequel l'attribut $AGE(.)$ est transformé en attribut nominal

II.2.2.2 Difficultés liées à l'ensemble des exemples d'apprentissage

a) Induction avec un grand nombre d'exemples

En raison des limites de capacité des ordinateurs, la génération d'un arbre de décision à partir d'un nombre d'exemples très élevé peut poser des problèmes techniques, car la méthode d'induction d'arbres par algorithme II.2 nécessite l'accessibilité de l'ensemble X_A complet.

Donné : Un problème de classification $CT=(O,A,P,S)$.
Un ensemble d'exemples d'apprentissage X_A.

Cherché : Un arbre de décision ADD pour X_A.

Algorithme :

(0) Choisir aléatoirement un fenêtre $W_0 \subset X_A$, $i:=0$

(1) Construire un arbre ADD à partir de W_i

(2) $W_{i+1} := W_i \cup \{(o,\vec{a}(o)) \in X_A \wedge F_{ADD}(o,\vec{a}(o)) \neq c(o)\}$

(3) Si $W_{i+1} = W_i$ STOP

Sinon

(4) $i := i+1$ et continuer avec (1)

Algorithme II.4: Induction d'arbres de décision avec fenêtrage d'exemples

Quinlan (1984) propose un algorithme de fenêtrage pour la construction d'arbres à partir de grands ensembles d'apprentissage (algorithme II.4), qui est basée sur l'hypothèse stipulant, qu'une grande partie de l'information nécessaire pour l'induction de la règle est comprise dans une partie relativement petite de l'ensemble.

Wirth et Catlett (1988) ont comparé la performance de deux versions de ID3, avec et sans application de l'algorithme de fenêtrage, dans plusieurs champs d'application. Ils concluent que les arbres de décision construits par ID3 avec fenêtrage sont généralement plus grands et moins corrects que les arbres construits sans fenêtrage. Ils ont aussi trouvé une nette augmentation du temps de calcul due au fenêtrage.

Une autre possibilité de traiter de grands nombres d'exemples est l'application d'une version incrémentale de l'algorithme de génération d'arbres de décision (algorithme II.2). Des versions incrémentales de ID3 ont été développé par Schlimmer et Fisher (1986) et par Utgoff (1988, 1989).

Une génération incrémentale d'arbres de décision est basée sur la présentation des exemples "un par un". Une mémorisation de la distribution des valeurs d'attributs à chaque noeud permet une détermination du meilleure attribut de test pour les noeuds après chaque nouveau exemple présenté. Cette technique n'est pas applicable si les attributs sont du type linéaire quantitatif, car une mémorisation des valeurs d'attributs correspond dans ce cas à une mémorisation de l'ensemble des exemples.

b) Traitement du bruit dans les valeurs des attributs des exemples

Un problème courant dans les applications réelles est l'existence de bruit dans les données, dû soit aux erreurs de mesure, soit à une faible exactitude de mesure soit à un traitement incorrect des données. Deux types d'erreur sont provoqués par le bruit: la qualité prédictive d'un arbre de décision est diminuée si il est généré à partir d'exemples erronés, et la classification d'objets est incorrecte si les valeurs d'attributs sont faussées. Dans les deux cas un raisonnement probabiliste est plus approprié que le raisonnement exact de la technique décrite au chapitre précédent.

L'induction d'arbres de décision par l'algorithme II.2 construit un arbre dont toutes les feuilles ne contiennent que des exemples d'une seule classe. Si des noeuds terminaux contenants des exemples de plusieurs classes sont admis, une attribution probabiliste d'une classe à un objet peut être estimée à partir de la fréquence des exemples d'apprentissage. L'ensemble de solutions S d'un contexte de classification de k classes est alors changé en

$$S = \{(p_1,\ldots,p_k) \in (0,1)^k \,,\ \sum_{i=1}^{k} p_i = 1\}$$

p_i indiquant la probabilité conditionnelle d'appartenance d'un objet à la classe i. Souvent on utilise des coefficients de vraisemblance $s_i = 2p_i - 1$ au lieu des probabilités.

Dans l'arbre montré dans la figure II.6.a, les feuilles contiennent des objets d'une seule classe. Par cet arbre, un animal de proie est classé comme étant dangereux, si il a plus de 3 ans. Ce choix diminue évidemment la qualité prédictive de l'arbre, car un animal de proie d'un age de trois ans peut très bien être dangereux. Une réduction de l'arbre au noyau important, montré dans la figure II.6.b, avec une interprétation des feuilles comme prédateurs probabilistes permet des meilleur résultats. Par cet arbre, chaque animal de proie est classé comme dangereux avec une probabilité de 0.66.

Breiman et *al.* (1984) et Quinlan (1987b) proposent des techniques de taillage d'arbres de décision selon des critères de complexité et de l'importance des branches de l'arbre. Quinlan (1987b) compare ces techniques à une méthode de transformation de l'arbre de décision en un ensemble optimisé de productions. Il trouve que la dernière méthode est plus performant en ce qui concerne la qualité prédictive et l'efficacité de la base de connaissances.

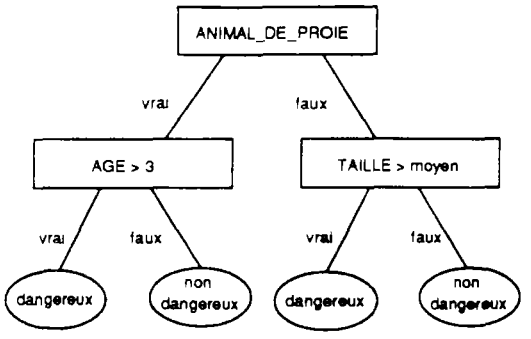


Figure II.6.a: Exemple d'un arbre de décision trop spécialisé, avec des feuilles déterministes

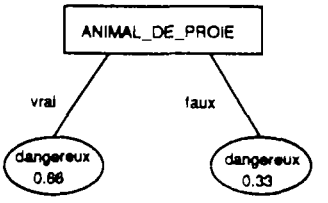


Figure II.6.b: Le même arbre réduit au noyau important, avec des feuilles probabilistes

Une autre possibilité de réduction de l'arbre est l'introduction de conditions d'arrêt dans l'algorithme de génération d'arbres (étape (1) d'algorithme II.2), en estimant la valeur prédictive d'un attribut de test. Quinlan (1986) propose une méthode basée sur le test de χ^2, qui détermine la confiance avec laquelle on peut rejeter l'hypothèse stipulant que l'attribut a_i, choisi comme attribut de test pour noeud n, est indépendant de la distribution des classes d'exemples associés à n. Si cette confiance n'est pas très faible (par exemple 5%), le noeud est déclaré comme noeud terminal. La signification d'un tel test est plus grande s'il est appliqué à un ensemble d'exemples test X_T différent de l'ensemble d'apprentissage X_A.

Dans le cas d'attributs linéaires dans un contexte de valeurs faussées, leur transformation en attribut nominal pose le problème supplémentaire de la définition des seuils de classes qualitatives. Les seuils fixes ne permettent pas de tenir compte du fait qu'un petit changement d'une valeur peut être dû au bruit, et devrait en conséquence ne pas provoquer le changement de la classification.

Soit n un noeud non terminal d'un arbre de décision binaire, et "$a^n(o)<s$" le test du noeud n. La classe 1 soit attribuée à l'objet o avec la probabilité $p_{<s}$ si $a^n(o)<s$, et avec la probabilité $p_{\geq s}$ dans l'autre cas. Si $p_{<s} \neq p_{\geq s}$ et $a^n(o)=s-\delta$, un bruit $b>\delta$ peut alors changer la classe attribuée à o.

Quinlan (1987a) propose la définition de ***seuils souples*** par l'introduction de seuils supplémentaires $s^-<s$ et $s^+>s$. La décision souple est prise comme suit:

$$
\begin{array}{lll}
\text{si } a^n(o) < s^- & : & p(c(o)=1) = p_{<s} \\
\text{si } s^- \leq a^n(o) < s & : & p(c(o)=1) = p_{<s} + (p_{\geq s} - p_{<s}) \cdot \dfrac{a^n(o) - s^-}{2(s - s^-)} \\
\text{si } s \leq a^n(o) \leq s^+ & : & p(c(o)=1) = p_{\geq s} + (p_{\geq s} - p_{<s}) \cdot \dfrac{s^+ - a^n(o)}{2(s^+ - s)} \\
\text{si } s^+ < a^n(o) & : & p(c(o)=1) = p_{\geq s}
\end{array}
$$

Ainsi un intervalle de transition permettant, à proximité du seuil s, des petites variations des valeurs d'attributs sans changement important de la classification effectuée est introduit. Quinlan propose une définition des seuils s^- et s^+ comme suit:

Si le meilleur test d'un noeud n est ($a^n<s$), une partie $T \subset X_A^n$ d'exemples est mal classée par le sous-arbre dont n est la racine. Si le test est changé à ($a^n<s'$) avec $s' \neq s$, la partie

d'exemples qui sont mal classés est augmentée, car le seuil s était choisi comme seuil optimal. L'écart-type ET de l'erreur peut être estimé par

$$ET = \sqrt{\frac{(|T|+0.5)\cdot(|X_A^n|-|T|-0.5)}{|X_A^n|}}$$

s^- et s^+ sont choisis de façon que la différence entre la taille de l'ensemble T et la taille des ensembles T^+ et T^- d'exemples, qui sont mal classés si le test était ($a^n<s^+$) ou ($a^n<s^-$), soit égale à ET: $|T^+|=|T^-|=|T|+ET$.

Cette technique utilise le caractère linéaire des attributs quantitatifs, qui est négligé par la transformation en attributs nominales. Elle exploite alors une partie de l'information perdue pour la classification, si des seuils "durs" sont utilisés.

c) Traitement de valeurs inconnues des attributs

Dans des applications réelles on rencontre souvent le cas, que la valeur d'un attribut est inconnue. Un tel attribut sera appelé ***attribut incomplètement valué***. Cette source d'erreur demande des traitements spéciaux dans les étapes

(1) du choix d'attribut (étape (2) d'algorithme II.2),

(2) de la division de l'ensemble des exemples (étape (4) d'algorithme II.2),

(3) de la classification (étape (2) d'algorithme II.3).

Des différentes manières de résoudre ces problèmes sont résumées dans Quinlan (1989). Il s'agit de:

- négliger les exemples de valeurs inconnues ((1) et (2)),
- ajouter une nouvelle valeur "inconnue" à la liste de valeurs de l'attribut ((2) et (3)),
- estimer la valeur d'un attribut à partir de la distribution des valeurs d'autres exemples ((1), (2) et (3)),
- évaluer le résultat pour toutes valeurs possibles et estimer le résultat pour l'objet en considération par une pondération des résultats ((3)).

Quinlan a trouvé, que l'arbre de décision est moins prédictif, lorsque les valeurs inconnues sont négligés pendant la génération. La meilleure performance a été obtenue avec les méthodes, qui estiment la distribution des valeurs inconnues à l'aide des exemples de valeur connue pendant la génération d'arbre.

En cas de méconnaissance de la valeur d'un attribut de test pendant la classification d'un objet, la meilleure méthode est l'évaluation des résultats pour toutes valeurs possibles et la pondération de ces résultats selon la distribution observée des valeurs.

Supposons que dans le contexte CT_{EX} la valeur de l'attribut TAILLE soit inconnue pour l'objet A2. La distribution connue des valeurs de cet attribut est alors (2,2,1). Pendant la génération de l'arbre, la valeur TAILLE(A2) est choisi aléatoirement avec les probabilités

p(TAILLE(A2)="petit") = 0.4

p(TAILLE(A2)="moyen") = 0.4

p(TAILLE(A2)="grand") = 0.2

Pour la classification de l'objet A2 par l'arbre montre par la figure II.4, la méthode proposée consiste en une évaluation du résultat pour chaque valeur possible de l'attribut TAILLE:

si TAILLE(A2)="petit" alors c(A2)=0 (non dangereux)

si TAILLE(A2)="moyen" alors c(A2)=1 (dangereux)

si TAILLE(A2)="grand" alors c(A2)=1 (dangereux)

Le résultat de la classification est la moyenne pondérée selon la distribution des valeurs connues de l'attribut:

$$c(A2) = 0.4 \cdot 0 + 0.4 \cdot 1 + 0.2 \cdot 1 = 0.6$$

Ce résultat peut être interprété comme probabilité 0.6 d'appartenance de l'objet A2 à la classe 1 (dangereux).

Breiman et *al.* (1984) ont proposé de générer pour chaque test dans l'arbre une liste de tests sur des attributs différents, ordonnée dans l'ordre décroissant de similarité des résultats avec le test original ("*surrogate splits*"). Si la valeur du premier attribut de test est inconnue pour un objet, il est classé selon le deuxième test, si cette valeur est aussi inconnue, selon le troisième test, et ainsi de suite. Pour le choix des tests, les objets de valeur inconnue sont négligés, en supposant que leur nombre est petit comparé au nombre d'objets de valeur connue. Cet algorithme exige un nombre fixe d'enfants pour chaque noeud, comme c'est le cas par exemple pour l'arbre binaire.

d) Représentativité de l'ensemble de l'apprentissage

Il est évident qu'un arbre de décision prédictif ne peut être développé qu'à partir d'un ensemble d'exemples représentatifs pour le concept faisant l'objet de l'apprentissage. Un problème se pose si ce concept est l'objet d'un changement, c'est à dire la représentativité de l'ensemble d'apprentissage change après la construction de l'arbre de décision. Ce phénomène est appelé "concept drift" dans la littérature anglaise. Dans ce cas l'arbre de décision doit changer avec le concept. Une méthode incrémentale de construction, comme celle décrite par Schlimmer et Fisher (1986) et par Utgoff (1988, 1989), est alors indispensable.

II.3 Conclusion

Nous avons exposé les méthodes les plus importantes, qui ont été développées en intelligence artificielle pour l'apprentissage automatique de bases de connaissances. La notion générale, introduite en première partie de ce chapitre, nous a permis de comparer les méthodes d'une façon abstraite sous les angles de la finalité de l'apprentissage, de la représentation symbolique de connaissances, et de la technique de l'apprentissage. L'idée centrale de cette notion est la formalisation du problème donné comme un *contexte de connaissances*. L'adaptation des méthodes examinées aux contextes de différentes caractéristiques est montrée dans le tableau II.2.

Les systèmes de productions, et plus concrètement les arbres de décision, présentent une technique de la représentation de connaissances, qui est particulièrement adaptée aux contextes incertaines demandant une bonne compréhensibilité des solutions proposées. Grâce à ces avantages, la plus grande partie des systèmes commercialisés est basée sur ces techniques.

Nous avons présenté l'algorithme principal de la génération d'arbres de décision à partir d'exemples donnés, ainsi que l'algorithme de classification d'objets par d'arbres de décision. Les problèmes, que peut rencontrer l'application de ces algorithmes aux problèmes réels, ont été exposés, et des diverses résolutions à ces problèmes ont été examinées.

Notre but est la génération d'une base de règles pour la reconnaissance des formes sur l'image radar, afin de pouvoir suivre automatiquement les structures météorologiques. Parmi les techniques examinées, l'induction d'arbres de décision semble être celle, qui est le mieux adaptée à cet objectif. Son application nécessite cependant en premier lieu la formalisation du problème dans un contexte de connaissances, et notamment la spécification de l'ensemble des objets de classification. Le chapitre prochain sera consacré à cette précision et à l'élaboration de l'algorithme de l'apprentissage.

III

PROPOSITION D'UNE MÉTHODOLOGIE DE L'APPRENTISSAGE AUTOMATIQUE D'UNE BASE DE RÈGLES POUR L'APPARIEMENT DES CELLULES DE PLUIE SUR L'IMAGE RADAR

Comme indiqué au premier chapitre, nous cherchons un progrès dans l'observation des structures météorologiques. Une prévision fiable par extrapolation des observations nécessite l'identification des structures, qui ont une persistance temporelle et des caractéristiques uniformes. Dans des conditions orageuses ces structures sont des cellules convectives, tandis que dans des conditions frontales l'observation à une échelle plus grande, par exemple au niveau des bandes de pluie, peut être suffisante. Nous utiliserons la notion ***cellule*** dans un sens large pour toutes ces structures, et préciserons sa signification où ça sera nécessaire.

Ce troisième chapitre a pour sujet la proposition d'une méthode permettant le suivi des cellules sur l'image radar. Dans la première partie nous présenterons la technique de leur identification. Ensuite, nous formulerons le problème de l'appariement dans un contexte de connaissances, dans lequel les techniques de l'apprentissage à partir d'exemples, examinées au chapitre précédent, peuvent être appliquées. En troisième partie nous développerons les attributs utilisés pour la description des exemples de l'apprentissage. Dans un quatrième temps, la méthode de définition de l'ensemble d'exemples sera exposée. La technique de l'apprentissage d'une base de règles sera présentée en cinquième partie. En dernière partie, la méthode de l'application des règles pour l'appariement automatique sera exposée.

III.1 L'identification des cellules de pluie sur l'image radar

III.1.1 L'ensemble des échos simples

Nous recherchons à identifier les cellules dans l'intervalle temporel de leur présence dans la région couverte par l'image radar, ce qui nécessite l'identification les ensembles de pixels correspondant aux cellules sur des images successives.

Nous utilisons la notion ***écho*** pour un ensemble de pixels, dont la réflectivité dépasse un seuil donné r_s, et nous faisons distinction entre

- ***échos de pluie***, qui sont des ensembles de pixels dont la réflectivité est reconnue comme provenant de la pluie, et
- ***échos de sol***, qui sont des ensembles de pixels dont la réflectivité provient des échos fixes ou des propagations anormales.

Une ***séquence d'échos*** est une suite $(e_1,..,e_n)$ d'échos sur des images successives $I_1,..,I_n$, chaque couple (e_i,e_{i+1}) étant apparié.

Comme déjà mentionné, les échos de sol ne seront pas traités dans cette étude, ils ont été écartés à l'aide de la visualisation et de la validation de toutes images utilisées. Par la suite, la notion "écho" signifiera toujours "écho de pluie". Pour leur définition sur l'image radar, deux méthodes différentes sont applicables.

La première méthode est un groupement de pixels basé sur une définition de leur connectivité, qui est donnée par une valeur d_{max} de distance maximale entre deux pixels connectés. Ainsi, avec une valeur de d_{max} égale à 1, deux pixels sont connectés, si ils ont une cote en commun, tandis qu'avec une valeur de d_{max} égale à 1.5, deux pixels sont connectés, si ils ont un point en commun (cf. figure III.1).

Cette définition est employée dans la majorité des systèmes de prévision (cf. par exemple Blackmer et *al.* 1973, Ostlund 1974, Tsonis et Austin 1981, Collier 1981, Einfalt 1990). Les valeurs r_s et d_{max} sont déterminées selon le type de l'application de la prévision. Plusieurs auteurs utilisent des valeurs fixes, tandis que Einfalt applique une méthode qui détermine la valeur du seuil en fonction de la distribution de l'intensité de pluie sur l'image radar.

Crane (1979) propose une autre méthode de définition des échos, qui repère les échos à partir des maxima locaux de l'intensité. Cette méthode est basée sur l'hypothèse, que chaque cellule de pluie possède un seul maximum. Les bornes d'un écho e à réflectivité maximale $r_{max}(e)$ sont définies par un niveau de réflectivité $r_{inf}(e) := r_{max}(e) - r_d$, où r_d est constant. Cette technique est particulièrement adaptée à l'observation des cellules convectives (Rosenfeld 1987, Brémaud 1991). Toutefois, cette deuxième méthode présente un algorithme beaucoup plus complexe que la première.

Pour cette étude, nous avons adopté la première méthode de définition, afin de faciliter la comparaison des caractéristiques d'échos telles que la masse, la taille et la forme. Pour les mêmes raisons, nous avons choisi le seuil de réflectivité et la distance définissant la connectivité constants pour toutes les images traités. Un inconvénient de cette fixation est la difficulté d'observer des cellules d'une forte intensité de pluie, qui sont imbriquées dans des champs de pluie stratiforme. Dans de telles situations, les écho définis n'ont pas nécessairement des caractéristiques uniformes.

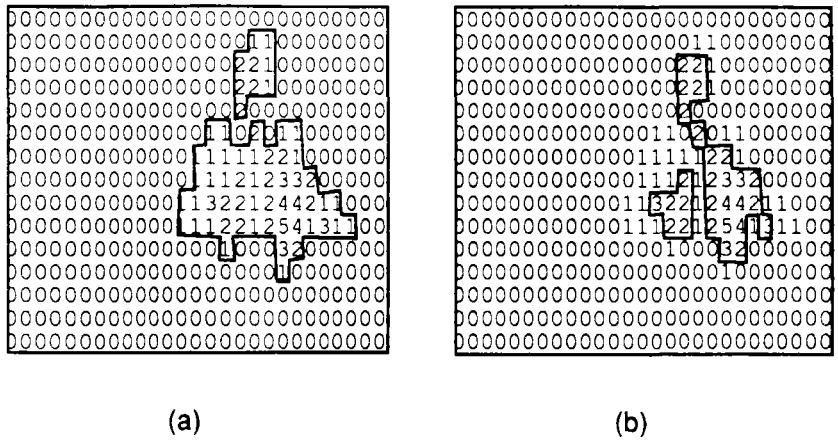


Figure III.1: Définition d'échos pour les valeurs r_s=1 et d_{max}=1 (a) et r_s=2, d_{max}=1.5 (b)

Subjectivement les valeurs r_s=25 dBZ et d_{max}=1.5 pixels ont été considérées comme optimales par rapport à l'objectif de définition des échos correspondant à des structures qui ont une certaine durée dans le temps et des caractéristiques uniformes. Afin de limiter le nombre d'échos, deux seuils inférieurs ont été introduits: La surface des échos doit être égale ou supérieure à 3.2 km^2 (5 pixels), et la masse pluvieuse des échos, obtenue comme produit de l'intensité moyenne et de la surface, doit dépasser $5 \cdot 10^4$ m^3/h.

L'ensemble des échos ainsi définis sur une image I est appelé l'***ensemble d'échos simples*** de I et est noté $E_s(I)$.

III.1.2 L'ensemble d'échos imaginaires

Le phénomène observable sur les images radar du type PPI est la pluie à une certaine altitude. La structure météorologique des cellules peut cependant être telle que, à un instant donné, il existent des zones de faible pluie à son intérieur, ce qui provoque que ses zones de pluie intense ne sont pas connectés sur l'image radar. Dans ce cas, la cellule est représentée par une agglomération d'échos simples, qui peuvent faire l'objet de ***scissions*** et ***fusions*** sur des images successives. Dans ce cas, la seule définition des échos simples n'est pas suffisant pour le suivi de la cellule.

Einfalt et *al.* (1990) proposent comme solution la définition d'échos dits "imaginaires", qui sont formés par une agglomération d'échos simples. Nous adoptons cette notion et définissons l'***ensemble d'échos imaginaires*** $E_i(I)$ d'une image I comme ensemble de sous-ensembles de $E_s(I)$, en excluant les ensembles élémentaires:

$$E_i(I) = \wp(E_s(I)) \setminus \{\{e\}, e \in E_s(I)\}$$

Le suivi d'une cellule, qui a été observé sur les images $I_1,\ldots,I_n$ comme une séquence d'échos *es* et dont l'écho est séparée en deux échos simples e_1 et e_2 sur l'image I_{n+1}, est rendue possible par l'appariement de *es* avec l'écho imaginaire $\{e_1,e_2\}$ (figure III.2). De la même façon, la fusion de deux échos simples peut être traitée.

Nous définissons l'***ensemble d'échos*** $E(I)$ d'une image radar I comme union de l'ensemble d'échos simples et de l'ensemble d'échos imaginaires

$$E(I) = E_i(I) \cup E_s(I) \equiv \wp(E_s(I))$$

La prise en compte de tous les échos imaginaires n'est cependant pas efficace, car seule une très petite partie des échos imaginaires est nécessaire pour le suivi correct des cellules. Nous appelons cette partie l'***ensemble d'échos imaginaires utiles***.

Einfalt et *al.* (1990) appliquent une méthode hiérarchique de définition d'une chaîne binaire d'agrégations d'échos, qui génère le sous-ensemble $E_i'(I) \subset E_i(I)$. La méthode est basée sur une fonction δ de distance entre deux échos e_1 et e_2:

$$\delta(e_1,e_2) = \frac{\text{masse}(e_1) + \text{masse}(e_2)}{\text{masse}(e_1) \cdot \text{masse}(e_2)} \cdot \| cg(e_1) - cg(e_2) \|^2 \qquad [\text{h/mm}]$$

où $cg(e_i)$ représente le centre de gravité de l'écho e_i (algorithme III.1). Nous adoptons cette méthode, qui génère n-1 échos imaginaires pour une image avec n échos simples.

Donné : Une image radar I, l'ensemble d'échos simples $E_s(I)$ et une fonction de distance entre échos $\delta : E(I) \times E(I) \rightarrow \mathbb{R}$

Cherché : Un ensemble $E_i'(I) \subset E_i(I)$ d'échos imaginaires utiles.

Algorithme :

(0) $E_i'(I) := \emptyset$; $E := \{\{e\}, e \in E_s(I)\}$

(1) Choisir $e_1, e_2 \in E \cup E_i'(I)$ tel que $\delta(e_1,e_2)$ est minimal et $e_1 \cap e_2 = \emptyset$.

(2) Si un tel couple d'échos n'existe pas STOP.

Sinon

(3) $E_i'(I) := E_i'(I) \cup \{e_1 \cup e_2\} \setminus \{\{e_1\},\{e_2\}\}$

(4) $E := E \setminus \{\{e_1\},\{e_2\}\}$

(5) continuer avec (1)

Algorithme III.1: Algorithme de définition hiérarchique d'échos imaginaires

III.1.3 Séquences strictes d'échos

En utilisant les définitions données ci-dessus, une cellule est représentée dans un intervalle de temps (t_1,t_n) par une séquence d'échos $(e_1 \in E(I_1),..,e_n \in E(I_n))$. Toutefois, dû aux scissions et fusions des échos simples, il existent des séquences, qui ne représentent pas des cellules ayant des caractéristiques uniformes. La considération de toute séquence comme cellule n'est alors pas utile, même si les appariements sont corrects. Par la suite, seules les séquences correspondant à la définition suivante seront considérés comme représentation des cellules.

Définition III.1:

Soit $(I_1,...,I_n)$ une suite d'images radar. Une ***séquence stricte d'échos*** de $I_1,...,I_n$ est définie comme suit:

(1) Un couple $(e_1 \in E(I_i), e_2 \in E(I_{i+1}))$ est une séquence stricte d'échos, si une des conditions suivantes est remplie:

(i) e_1 et e_2 sont appariés,

(ii) e_1 est un écho imaginaire, dont un élément représentant 90% de la masse est apparié à e_2,

(iii) e_2 est un écho imaginaire, dont un élément représentant 90% de la masse est apparié à e_1,

(iv) e_1 et e_2 sont des échos imaginaires, et les éléments de e_1 sont appariés biunivoquement aux éléments de e_2.

(2) Une séquence $(e_1 \in E(I_i),...,e_k \in E(I_{i+k}))$ est une séquence stricte d'échos, si $(e_1,...,e_{k-1})$ est une séquence stricte d'échos et (e_{k-1},e_k) est une séquence stricte d'échos.

L'ensemble de séquences strictes d'échos d'une suite d'images sera notée $SE(I_1,...,I_n)$.

La définition III.1 implique, qu'une séquence d'échos est une séquence stricte, si aucune scission ou fusion n'a lieu, ou si les échos, qui sont crées par une scission, se fusionnent après. Une exception est faite pour les scissions et fusions, qui ne concernent qu'une petite partie d'un écho représentant moins de 10% de sa masse totale.

La définition III.1 est appliquée dans cette étude pour tous les algorithmes concernant l'observation des cellules dans le temps. Afin de réduire l'influence du bord de l'image sur le calcul des caractéristiques des cellules, on ne tient compte que des échos, dont le centre de gravité est situé à une distance minimale du bord de l'image. Nous définissons cette distance minimale en fonction de la taille de l'écho:

$$distance_{min}(\, cg(écho)\, ,\, bord\,) = \sqrt{\frac{taille(écho)}{\pi}}$$

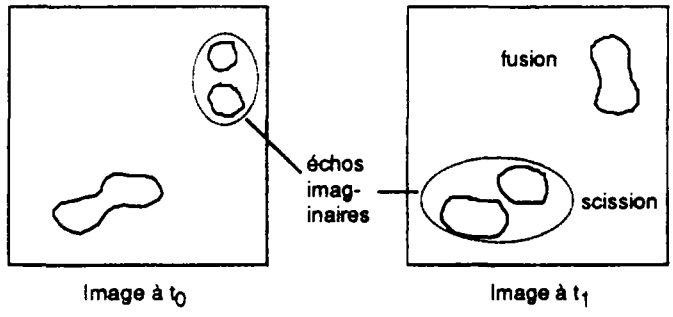


Figure III.2: Représentation schématique du principe du suivi des cellules à l'aide des échos imaginaires

III.2 Formulation d'un contexte de classification pour le problème de l'appariement d'échos

La définition des échos étant un problème plutôt algorithmique, leur appariement est cependant un problème de classification, pour lequel nous définissons par la suite un contexte de classification de deux classes $CT_{AP}=(O_{AP},A_{AP},P_{AP},S_{AP})$.

Nous considérons à un instant t_{n+1} l'image radar I_{n+1}. Soient $I_1,...,I_n$ les images précédentes, mesurées aux instants $t_1,...,t_n$. Nous cherchons à apparier correctement les échos sur l'image I_{n+1} avec les cellules observées dans l'intervalle (t_1,t_n). Si *es* est une séquence d'échos sur les images radar $I_1,...,I_n$, et *e* est un écho sur l'image I_{n+1}, il s'agit d'associer le couple (*es*,*e*) à une de deux classes:

- la classe "+" des bons appariements: l'écho *e* est provoqué par la même cellule que les échos de la séquence *es*;
- la classe "-" des mauvaises appariements: l'écho *e* est provoqué par une autre cellule que les échos de la séquence *es*.

Nous définissons l'ensemble des objets O_{AP} comme l'ensemble des couples (*es*,*e*), où *es* est une séquence d'échos sur les images radar $I_1,...,I_n$, et *e* est un écho sur l'image I_{n+1}:

$$O_{AP} = \{(es=(e_{n-k}\in E(I_{n-k}),...,e_n\in E(I_n)),\ e\in E(I_{n+1}))\}$$

Les attributs A_{AP} décrivent les caractéristiques des couples (*es*,*e*). Les caractéristiques correspondent aux paramètres de l'écho *e*, de la séquence *es* et des paramètres de développement basés sur des différences entre les structures de *es* et la structure *e*:

$$A_{AP} = \{a_1(.),...,a_k(.)\}$$

Le problème de la classification P_{AP} est d'estimer la probabilité, qu'un couple (*es*,*e*) appartient à la classe des bons appariements:

$$P_{AP} = \{p(\text{classe}(o)=+),\ o\in O_{AP}\}$$

L'espace des solutions S_{AP} est l'intervalle [0,1] des probabilités que le couple (*es*,*e*) est un bon appariement:

$$S_{AP} = \{p\in[0,1]\}$$

Une solution est considérée comme étant correcte, si la probabilité estimée de l'appartenance d'un couple (*es*,*e*) à la classe "+" est plus grande que 0.5 pour un bon appariement, ou si cette probabilité est égale ou plus petite que 0.5 pour un mauvais appariement.

Nous cherchons un système à base de connaissances $K_{AP}=(CT_{AP},P_{AP},BC_{AP},IC_{AP})$ qui donne la solution correcte pour tous couples (*es*,*e*).

Car une séquence $es\in SE(I_1,...,I_n)$ des images $I_1,...,I_n$ est définie univoquement par l'écho $e_n\in E(I_n)$, qui fait partie de la séquence, nous ne faisons dans la suite pas distinction entre l'appariement (es,e_{n+1}) et l'appariement (e_n,e_{n+1}).

III.3 Définition des attributs pour le contexte de l'appariement

Dans le contexte CT_{AP}, on cherche à caractériser les objets par des attributs ayant une signification pour la tâche de classification en bons et mauvais appariements. Un objet est un couple $(es=(e_1 \in E(I_1),...,e_n \in E(I_n)), e_{n+1} \in E(I_{n+1}))$ d'une suite d'images $(I_1,...,I_n,I_{n+1})$. La suite d'images correspond à une période d'un événement pluvieux EV, où deux images I_k,I_{k+1} sont écartées temporellement d'une intervalle $\Delta_{k,k+1}t$.

Les paramètres utiles à la classification peuvent être divisés en cinq catégories:

(A1) caractéristiques de l'événement de pluie,

(A2) caractéristiques de la pluie sur image I_{n+1},

(A3) caractéristiques de l'écho e_{n+1},

(A4) caractéristiques de l'écho e_{n+1} par rapport aux caractéristiques des échos $e_1,...,e_n$,

(A5) changement de caractéristiques entre e_n et e_{n+1} par rapport aux changements dans la séquence es

Les paramètres des catégories A4 et A5 sont évidemment ceux sur lesquels doit reposer la classification finale, car ils sont basés sur une comparaison des échos, tandis que ceux des catégories A1, A2, et A3 peuvent être utiles pour un préclassement des couples (es,e), pouvant par exemple servir à séparer les orages d'été des pluies stratiformes d'hiver.

Néanmoins, le préclassement n'est possible que lorsque la méthode de l'apprentissage permet d'évaluer sa valeur pour la tâche de classification. Dans l'algorithme de ID3 cette évaluation n'est pas possible, car l'importance d'un attribut est évalué uniquement en fonction du degré de discrimination des classes qu'il effectue (cf. algorithme II.2). Or, l'utilité des paramètres qui servent à un préclassement est limitée par cette méthode de l'apprentissage. Les attributs, qui seront utilisées dans cette étude, sont alors ceux des catégories A4 et A5. La liste complète des paramètres utilisés dans cette étude se trouve en annexe.

Les attributs de la cinquième catégorie semblent à priori plus utiles que ceux de la quatrième, car ils traitent les différences des caractéristiques des échos e_n et e_{n+1} d'une manière relative à la variation des caractéristiques dans la séquence es: par exemple l'attribut, qui exprime la distance entre écho e_n et écho e_{n+1}, parait moins utile que l'attribut, qui exprime la différence entre les distances observées dans la séquence es et la distance entre écho e_n et écho e_{n+1}, car la dernière permet une plus grande généralité. Ces attributs seront appelés ***attributs historiques***.

Or, les valeurs des attributs historiques sont inconnues lorsque la longueur de la séquence est trop petite. Ceci est par exemple le cas pour une cellule naissante observée sur deux images seulement. Aussi, pour des raisons de stabilité, plusieurs paramètres ne sont fiables qu'après une certaine durée d'observation. L'advection des cellules, en effet définie par le déplacement des centres de gravité des échos, en est un exemple. Car cette valeur est influencée par la résolution spatiale de la mesure et par le développement des intensités à l'intérieur des cellules, le calcul d'une valeur moyenne sur une période comprenant plusieurs images radar est indispensable pour disposer d'une valeur fiable.

Remarquons que toutes les valeurs d'attributs sont entachées d'incertitude, dû aux inexactitudes de mesure et du traitement des données (par exemple la discrétisation des valeurs).

III.4 Définition de l'ensemble de l'apprentissage

Dans le contexte CT_{AP} nous développerons une méthode d'induction d'une base de connaissances à partir d'exemples donnés. Deux ensembles de l'apprentissage sont nécessaires pour cette induction: un ensemble d'exemples positifs

$$X^{+} = \{(es_i, e_i), es_i \in SE(I_{k_i}, \ldots, I_{k_i+d}), e_i \in E(I_{k_i+d+1})\}$$

dont les éléments sont des bons appariements, et un ensemble d'exemples négatifs

$$X^{-} = \{(es_j, e_j), es_j \in SE(I_{k_j}, \ldots, I_{k_j+d}), e_j \in E(I_{k_j+d+1})\}$$

dont les éléments sont des mauvais appariements.

Le choix des séquences d'images radar, qui seront utilisées pour la définition des ensembles de l'apprentissage, est guidé par la nécessité de représenter tous les types de pluie. Les différences entre les types de pluie se manifestent dans les caractéristiques suivantes:

- les conditions météorologiques (convectif, front froid/chaud, ...),
- l'intensité de la pluie,
- la taille d'échos simples,
- le nombre d'échos simples sur l'image,
- la vitesse et la direction de l'advection des cellules.

Les caractéristiques des ensembles d'exemples d'apprentissage déterminent le degré de spécification de la règle induite. Un ensemble d'exemples positifs trop petit ou mal choisi, implique une règle trop spécifique, qui a pour résultat une fausse classification des bons appariements. Un mauvais choix de l'ensemble des exemples négatifs implique au contraire une règle trop générale, ce qui peut provoquer une fausse classification de cas négatifs. Comme le montre schématiquement la figure III.3, une règle consistante peut être générée à partir d'un ensemble d'exemples positifs comprenant un échantillon représentatif de l'espace des cas positifs, et d'un ensemble d'exemples négatifs proches des cas positifs. Winston (1975) a introduit la notion de la ***presque-instance*** pour un exemple négatif proche des instances positives dans l'espace défini par les valeurs d'attributs.

L'utilisation comme instances positives de l'ensemble d'apprentissage de tous les bons appariements sur les images choisies permet d'éviter de générer une règle trop spécifique. Toutefois, il faudra s'assurer que tous les types de pluie sont représentés dans cet ensemble.

Une définition de l'ensemble d'exemples négatifs est possible comme le complément de X^{+} dans l'ensemble des appariements sur les images choisies. Étant donné la taille de cet ensemble, cette solution n'est pas praticable: sur deux images successives avec 10 échos simples (et 9 échos imaginaires) définis sur chacune, le nombre d'appariements possibles est $(10+9)^2=361$, dont un maximum de 10 sont de bons appariements. Le choix d'un nombre limité de presque-instances comme exemples négatifs est nécessaire pour obtenir une base d'apprentissage efficace.

Mais ce choix n'est pas facile à faire; car la définition appropriée de distance entre exemples peut être totalement subjectif (exemple: une métrique euclidienne, qui accorderait le même poids à chaque attribut du contexte). Nous proposons une méthode incrémentale de recherche de presque-instances, qui emploie l'algorithme de l'apprentissage pour la découverte d'exemples susceptibles d'être mal classés par une règle trop générale (algorithme III.2 et figure III.4). La méthode est basée sur la technique de fenêtrage proposé par Quinlan (1984) pour le traitement d'ensembles d'exemples trop larges pour être utilisés en totalité par l'algorithme de l'apprentissage (algorithme II.4). Elle est toutefois différente, car toutes instances positives sont inclues à priori dans l'ensemble d'apprentissage, et les exemples négatifs sont uniquement ceux mal classées par une règle intermédiaire.

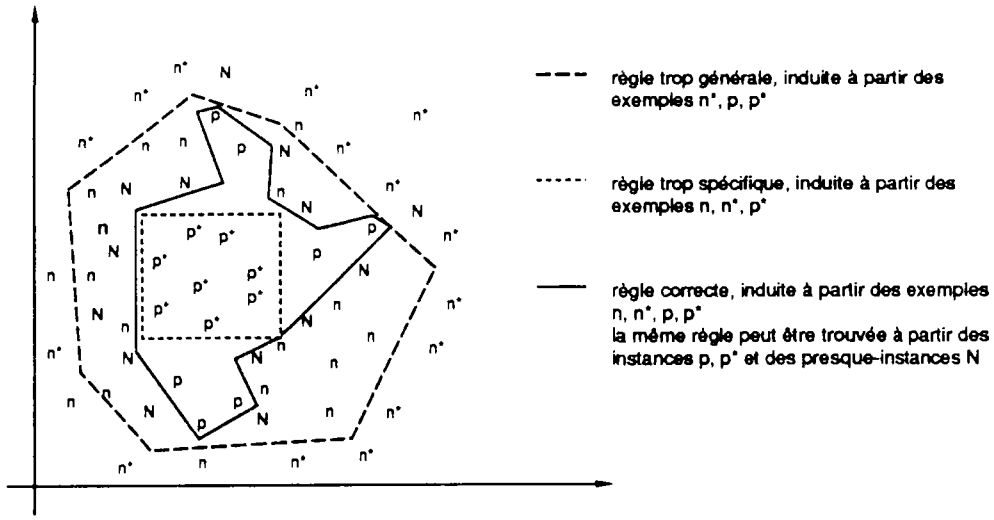


Figure III.3: Présentation schématique de l'influence des ensembles d'apprentissage sur la qualité de la règle induite

L'algorithme proposé est applicable dans les conditions suivantes:

- le nombre des instances du concept est beaucoup plus petit que le nombre des non-instances,
- une règle initiale peut être définie à priori.

Ces deux conditions sont remplies dans le contexte de l'appariement: une règle initiale peut être définie en utilisant des contraintes physiques concernant par exemple la vitesse maximale de l'advection des cellules.

Donné : Un contexte $CT=(O,A,P,S)$ de classification de deux classes (+,—).
Un ensemble fini d'exemples $X \subset O$, dont la classe est connue.
L'ensemble d'exemples positifs $X^+=\{x \in X,\ c(x)=+\}$.
Une règle initiale BC_0 qui est correcte pour X^+.
Un système à base de connaissance $K=(CT,\mathcal{P},BC,IC)$.
Un système de génération de règles $G : \mathcal{P}(O) \rightarrow \Re$, avec $\Re$ l'espace de règles BC.

Cherché : Un ensemble $X^- \subset O$ de presque-instances de CT.

Algorithme :

(0) $BC := BC_0 ; X^- := \varnothing$;

(1) $X^+_{BC} := \{x \in X \wedge IC(BC,x)=+\}$

(2) $X^{-\prime} := \{x \in X^+_{BC} \wedge x \notin X^+\}$

(3) Si $X^{-\prime} = \varnothing$ STOP

Sinon

(4) $X^- := X^- \cup X^{-\prime}$

(5) $BC := G(X^+ \cup X^-)$

(6) continuer avec (1)

Algorithme III.2: Algorithme incrémental de génération d'un ensemble d'exemples négatifs d'apprentissage

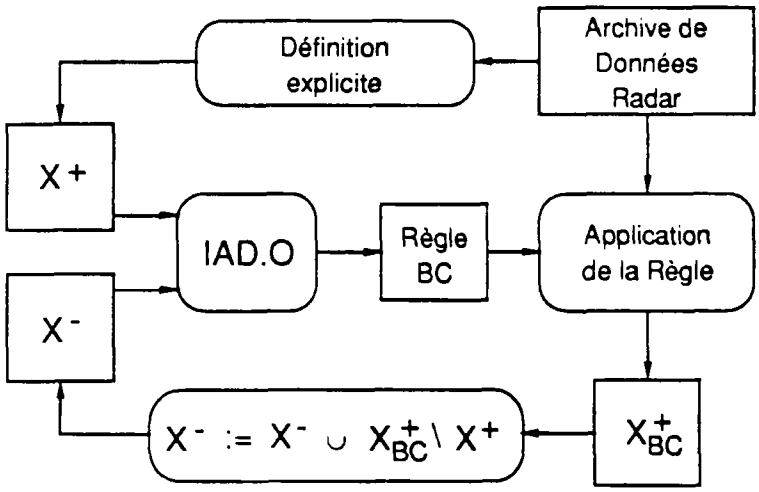


Figure III.4: Présentation schématique du flux des données dans l'algorithme incrémentelle de génération d'un ensemble d'exemples négatifs

III.5 Description de l'algorithme de l'apprentissage

La méthode de la génération de règles pour l'appariement des échos doit être adaptée aux caractéristiques du problème, notamment:

- des incertitudes dans les valeurs d'attributs,
- des valeurs d'attributs inconnues,
- une grande quantité de données,
- la méconnaissance d'une partie de l'espace d'objets.

La dernière contrainte s'explique par le fait qu'une base de données radar fini ne couvre pas nécessairement tous les types d'événement pluvieux.

Les méthodes statistiques classiques ne sont pas applicables dans ces conditions, notamment à cause de l'occurrence de valeurs inconnues. Parmi les techniques de l'intelligence artificielle, l'application d'un système de productions est particulièrement adaptée à ces problèmes (cf. tableau II.2). La technique des arbres de décision est une forme de ces systèmes, permettant la génération automatique de la base de connaissances. Dans la suite nous exposerons son application dans cette étude.

III.5.1 Description de l'algorithme principal

Nous avons développé un système d'apprentissage, nommé ***IAD.O***, permettant l'induction d'un arbre de classification pour l'appariement d'échos basé sur l'algorithme de ID3. Plusieurs mesures ont été prises afin de prendre en compte des particularités de ce problème.

a) Traitement des attributs linéaires

Les attributs du contexte CT_{AP} sont du type linéaire quantitatif. Pour leur transformation en attributs nominaux, on procède comme suit: si n est un noeud, pour lequel le meilleur attribut de test est recherché, et $a_k \in A_{AP}$ est un attribut linéaire, un attribut nominal a_k^s est défini par un seuil s selon

$$a_k^s(o) = \begin{matrix} 0 \text{ si } a_k(o) < s \\ 1 \text{ si } a_k(o) \geq s \end{matrix}$$

L'attribut $a_k^{s_{ak}}$ maximisant la réduction de l'impureté de n est sélectionné:

$$\Delta\Phi(n, a_k^{s_{ak}}) = \max_s(\Delta\Phi(n, a_k^s))$$

Pour des raisons de la simplicité, nous noterons par la suite l'attribut $a_k^{s_{ak}}$ également comme a_k et spécifierons la signification, où c'est nécessaire.

Les arbres générés sont alors de forme binaire. Ainsi, les problèmes relatifs aux nombres différents de valeurs des attributs et à la difficulté de valeurs non significatives ont été éliminés. Un inconvénient de cette méthode est le risque de générer des arbres plus grands que ceux générés par les méthodes autorisant plusieurs valeurs d'attributs.

Dans l'algorithme de ID3, chaque attribut ne peut être plus d'une fois l'attribut de test sur chaque chemin de la racine à un noeud terminal. La transformation des attributs linéaires en attributs nominaux génère en fait des attributs nominaux différents à partir d'un même attribut

linéaire. Le même attribut linéaire peut alors servir plusieurs fois comme attribut de test. La hauteur de l'arbre, limitée dans ID3 au nombre d'attributs, n'est alors limitée que par le nombre d'exemples. On peut cependant supposer, que dans le cas où le même attribut est souvent choisi comme attribut de test, ceci indique que les autres attributs n'ont pas une importance assez significative pour la classification. Dans IAD.O, nous limitons la hauteur de l'arbre à une valeur égale à deux fois le nombre d'attributs. Si cette limite est dépassée, nous considérons l'ensemble des attributs comme non significatif et arrêtons la procédure de génération.

b) Traitement du bruit dans les données

Les incertitudes dans les valeurs des attributs posent un problème particulier pour l'apprentissage d'une règle de l'appariement.

Pour IAD.O, nous avons pris deux mesures pour traiter ce problème:

- la génération d'un arbre de décision comme classificateur probabiliste,
- la définition de seuils souples, qui a été décrit au chapitre précédent.

Deux critères d'arrêt ont été introduits dans l'algorithme de la construction de l'arbre. Un noeud est déclaré comme feuille si:

- le nombre des exemples associés à un noeud est inférieur à 5 dans la classe la plus petite,
- l'hypothèse, que l'attribut choisi comme attribut de test est non-significatif pour la classification, ne peut pas être réfuté avec une certitude de 0.95.

Pour le deuxième critère l'ensemble d'exemples est divisé aléatoirement en un ensemble d'apprentissage X_A et un ensemble de test X_T. Le meilleur test est cherché à l'aide de l'ensemble X_A; sa signification pour la classification est ensuite testée par le test de χ^2 de l'indépendance de deux variables statistiques, appliqué à l'ensemble X_T.

c) Traitement de valeurs inconnues

Comme Quinlan (1989) le propose, on suppose la même fréquence des valeurs d'attributs pour les exemples dont les valeurs sont inconnues et pour les exemples dont la valeur est connue. Pour l'induction de l'arbre, les valeurs inconnues sont choisies aléatoirement avec la fréquence définie par les valeurs connues. Rappelons, que nous avons introduit la notion "attribut incomplètement valué" pour un attribut, dont une partie des valeurs est inconnue.

La technique appliquée est résumée dans l'algorithme III.3. Une probabilité p^n est associée à un noeud terminal n selon la répartition des classes dans l'ensemble de test associés à n:

$$p^n := \frac{|\{x \in X_T^n ; classe(x) = +\}|}{|X_T^n|}$$

Donné : Un contexte $CT=(O,A,P,S)$ de classification de deux classes (+,-) avec $A=\vec{a}=\{a_1,\ldots,a_k\}$ un ensemble d'attributs linéaires et $S=[0,1]$ la probabilité d'appartenance d'un objet o à la classe "+".

Un ensemble X^+ d'exemples positifs (classe +) et un ensemble X^- d'exemples négatifs (classe -).

Cherché : Un arbre de décision probabiliste *ADD*.

Algorithme :

(0) (a) Diviser aléatoirement l'ensemble $X=X^+\cup X^-$ en un ensemble de test X_T et un ensemble d'apprentissage X_A.

(b) Générer la racine $n:=n_0$ et associer X_T et X_A à la racine;

$X_T^n:=X_T$ et $X_A^n:=X_A$.

(1) Si la plus petite classe de X_A^n contient moins de 5 éléments, ou n est un noeud au niveau $2\cdot|A|$, transformer n en noeud terminal.

Sinon

(2) (a) Choisir un attribut $a^n\in A$ et une valeur de seuil s_{an} selon le critère de la plus grande réduction de l'entropie:

$$\Delta\Phi(n,a^n) = \max\{\Delta\Phi(n,a_k),a_k\in A\}$$

S'il existent des $x\in X_A^n$ dont la valeur de a^n est inconnue, choisir avant aléatoirement ces valeurs avec la fréquence définie par les $x\in X_A^n$ de valeur connue.

(b) Définir le test "$a^n<s_{an}$" pour le noeud n.

(c) Partitionner X_T^n selon le test de n en deux ensembles X_T^{ng} et X_T^{nd}. S'il existent des $x\in X_T^n$ dont la réponse du test est inconnue, choisir aléatoirement avec la fréquence définie par les $x\in X_T^n$ de réponse connue.

(d) Tester la signification du test par rapport à la fréquence des classes dans X_T^{ng} et X_T^{nd}.

(e) Si le test est négatif, transformer n en noeud terminal.

Sinon

(3) Générer deux noeuds ng, nd comme enfants de n.

(4) Partitionner X_A^n selon le test de n en deux ensembles X_A^{ng} et X_A^{nd}. S'il existent des $x\in X_A^n$ dont la réponse du test est inconnue, choisir aléatoirement avec la fréquence définie par les $x\in X_A^n$ de réponse connue.

(5) Associer X_A^{ng} et X_T^{ng} à ng, X_A^{nd} et X_T^{nd} à nd.

(6) Continuer avec (1) avec $n:=ng$, puis avec $n:=nd$.

Algorithme III.3: Algorithme de IAD.O de génération d'un arbre de décision

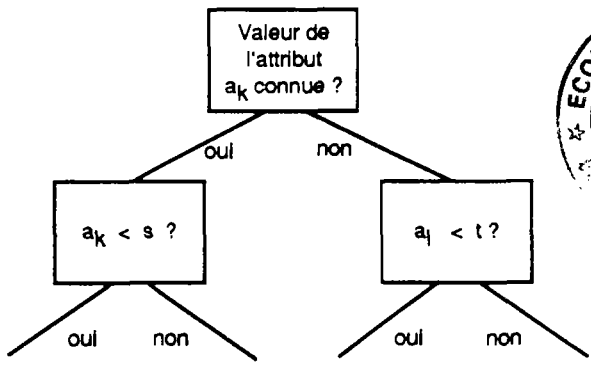
III.5.2 Proposition d'une méthode de traitement de valeurs systématiquement inconnues

L'hypothèse d'une fréquence des valeurs d'un attribut identique pour les exemples, dont la valeur est connue, et pour les exemples, dont la valeur est inconnue, n'est valable que si les valeurs manquent aléatoirement. Le traitement des valeurs inconnues dans l'algorithme de IAD.O est basé sur cette supposition.

Or, cette hypothèse n'est pas valable lorsque les valeurs sont systématiquement inconnues, ce qui est le cas dans le contexte de l'appariement d'échos. La méconnaissance des valeurs d'attributs historiques pour une cellule est la conséquence d'une des trois circonstances suivantes:

- la cellule pénètre dans la zone observée sur l'image radar,
- la cellule est en croissance et son écho vient de dépasser les seuils de définition des échos,
- la cellule vient de se séparer d'une autre cellule.

Les cas de valeur inconnue ne sont alors pas raccordables à ceux de valeur connue, et l'hypothèse, que la fréquence des valeurs connues est représentative de la fréquence des valeurs inconnues n'est pas nécessairement valable dans ce contexte. Afin d'en tenir compte, il est indispensable de séparer les exemples de l'apprentissage en exemples, dont la valeur d'un attribut donné est connue, et en exemples, dont la valeur est inconnue.



ECOLE NATIONALE CENTRE PÉDAGOGIQUE DE

Figure III.5: Sous-arbre construit par IAD.S dans le cas où une partie des valeurs de l'attribut est inconnue

Cette séparation elle-même ne diminue cependant pas nécessairement l'impureté dans l'arbre généré. Son utilité ne peut alors pas être mesurée par le critère de la réduction de l'impureté $\Delta\Phi(.,.)$, comme il a été défini.

Nous proposons la démarche suivante pour le choix d'un attribut de test pour un noeud n: pas de changement pour les attributs complètement valués. Pour les attributs incomplètement valués, une séparation de l'ensemble d'exemples associé à n en deux sous-ensembles est effectuée suivant que la valeur de l'attribut est connue ou non. Sur celui, dont la valeur de l'attribut est connue, on recherche la valeur de seuil permettant d'obtenir la meilleure réduction de l'impureté. Sur l'autre, on estime la réduction de l'impureté obtenable par la recherche d'un autre attribut, on appliquant un choix aléatoire des valeurs au cas où il s'agit d'un attribut incomplètement valué.

Cette procédure génère un sous-arbre de deux niveaux avec n comme racine (figure III.5). La valeur du test "a_k connu ?" est évalué en fonction de la réduction de l'impureté par ce sous-arbre. Le critère de choix d'un attribut de test a été modifié afin de rendre cette évaluation possible. La réduction de l'impureté d'un noeud n par un attribut a_k, dont la valeur est inconnue pour une partie X_A^{ni} de l'ensemble d'exemples X_A^n associé à n, est déterminée comme suit

$$\Delta\Phi'(n, a_k) = \Phi(n) - \frac{|X_A^{ni}|}{|X_A^n|}\Delta\Phi(ni, a_l) - \frac{|X_A^{nc}|}{|X_A^n|}\Delta\Phi(nc, a_k)$$

où ni est le noeud à lequel est associé l'ensemble d'exemples X_A^{ni} et nc est le noeud qui comprend les exemples X_A^{nc} dont la valeur de a_k est connue; $a_l \neq a_k$ étant le meilleur test de ni. Remarquons que pour les attributs complètement valués a_k, il est $\Delta\Phi'(n,a_k) = \Delta\Phi(n,a_k)$.

Les modifications entreprises concernent l'étape (2) de l'algorithme de IAD.O. La nouvelle version est nommée ***IAD.S*** (algorithme III.4).

(2) (a) Choisir un attribut $a^n \in A$ selon le critère de la plus grande réduction de l'entropie:

(i) Pour les attributs a_k, dont la valeur est connue pour tous $x \in X_A^n$, choisir le meilleur seuil s_{ak} et déterminer $\Delta\Phi'(n,a_k) = \Delta\Phi(n,a_k)$.

(ii) Pour les attributs a_k, dont la valeur est inconnue pour une partie d'exemples de X_A^n, diviser X_A^n en deux ensembles $X_A^{nc} = \{x \in X_A^n,\ a_k(x) \text{ connu}\}$ et $X_A^{ni} = \{x \in X_A^n,\ a_k(x) \text{ inconnu}\}$.

Pour X_A^{nc} déterminer $\Delta\Phi^c := \Delta\Phi(nc,a_k)$ comme sous (i).

Pour X_A^{ni} déterminer $\Delta\Phi^i := \max(\Delta\Phi(ni,a_l),\ a_l \neq a_k)$.

$\Delta\Phi'(n,a_k) = |X_T^{ni}|/|X_A^n| \cdot \Delta\Phi^i + |X_T^{nc}|/|X_A^n| \cdot \Delta\Phi^c$.

Diviser X_T^{nc} et X_T^{ni} selon les attributs choisis et tester la signification. Si un test est négatif, ne pas tenir compte de a_k.

(iii) Choisir l'attribut a^n avec $\Delta\Phi'(n,a^n) = \max\{\Delta\Phi'(n,a_k), a_k \in A\}$

Si a^n est un attribut dont toutes valeurs sont connues:

(b) Définir le test "$a^n < s_{an}$" pour le noeud n.

(c) Partitionner X_T^n selon le test de n en deux ensembles X_T^{ng} et X_T^{nd}.

(d) Tester la signification du test par rapport à la fréquence des classes dans X_T^{ng} et X_T^{nd}.

(e) Si le test est négatif, transformer n en noeud terminal.

Sinon

(f) Définir le test "a^n connu ?" pour le noeud n.

(g) Exiger l'attribut a^n pour le noeud enfant ng.

Algorithme III.4: Le traitement des valeurs inconnues dans l'algorithme IAD.S, qui est basé sur IAD.O

III.5.3 Proposition d'un critère de choix des attributs utilisant une plus grande profondeur de recherche dans l'arbre

La faiblesse de l'algorithme de ID3, qui consiste en l'estimation "myope" de la valeur d'un attribut pour la classification, a déjà été mentionnée. L'évaluation de la réduction de l'impureté directement après la division d'un noeud ne permet pas la découverte de relations complexes. L'attribut, qui décrit le type de l'événement de pluie comme convectif où frontal, en est un exemple. L'emploi de cet attribut n'augmente pas immédiatement la pureté dans un arbre de décision, car des exemples positifs et négatifs existent pour les deux cas. L'attribut est cependant susceptible de permettre l'induction de sous-arbres efficaces.

Dans l'algorithme de IAD.S, la profondeur de l'évaluation de l'importance d'un test sur un attribut a été augmentée, afin de rendre compte du fait que la séparation des exemples en exemples de valeur connue et exemples de valeur inconnue ne réduit pas l'impureté. Cette technique peut être généralisée de la façon suivante: A l'étape (2) de l'algorithme III.3, le meilleur sous-arbre de k niveaux au plus, minimisant l'impureté, est cherché pour le noeud n. L'attribut de test de la racine du sous-arbre est ensuite défini comme attribut de test de n.

Cette technique permet de déterminer la réduction de l'impureté par un enchaînement de conditions de tests comme $(a_k<s_{ak}\wedge\ldots\wedge a_l<s_{al})$. Elle offre une plus grande souplesse dans l'évaluation de la valeur d'attributs pour la tâche de classification. L'algorithme III.5 décrit les modifications de IAD.O nécessaires pour cette méthode, appelé ***IAD.L***.

(2) Pour tout attribut $a\in A$ et tout seuils s_a, construire un sous-arbre d'une profondeur maximale k avec la racine n selon l'algorithme de IAD.O. Evaluer la réduction de l'impureté par ce sous-arbre, et choisir le meilleur attribut comme attribut de test de n.

Algorithme III.5: Algorithme du choix d'attribut de test de IAD.L avec une profondeur de k, basé sur IAD.O

III.5.4 Réflexion sur la complexité des algorithmes proposés

La complexité *comp(A)* d'un algorithme *A* est une mesure de l'ordre de grandeur du nombre des opérations nécessaire pour son exécution, ce qui est en même temps une mesure relative du temps de calcul de sa réalisation sur ordinateur.

L'objectif de l'emploi de techniques de l'apprentissage automatique pour le problème de l'appariement est la découverte de règles efficaces. La méthode appliquée doit alors être choisie en fonction du taux d'erreur et de la complexité de la règle; la complexité de l'algorithme de l'apprentissage étant d'une moindre importance. Or, vu la grande quantité des données et les limites de capacité des ordinateurs, une réflexion sur ce sujet semble utile.

Afin de pouvoir comparer la complexité relative des différents algorithmes, la notion de l'ordre de complexité sera utilisée.

Définition III.2:
Une fonction $F(\vec{a}=(a_1,\ldots,a_n))$ est de l'***ordre de complexité*** $O(f(\vec{a}))$, si

$$\exists c \in \mathbb{R}\ \exists \vec{a}_0 \in \mathbb{R}^n : F(\vec{a}) \leq c \cdot f(\vec{a})\ \forall \vec{a}\ avec\ |\vec{a}| > |\vec{a}_0|$$

Les algorithmes proposés sont basés sur l'algorithme de ID3 (algorithme II.2). La complexité de ID3 dépend du nombre d'exemples d'apprentissage et du nombre d'attributs. Si *A* est l'ensemble d'attributs, X_A l'ensemble d'exemples d'apprentissage, et *D* la hauteur de l'arbre de décision construit, la complexité de ID3 est donné par

$$\text{comp}(ID3) = \sum_{k=|A|}^{|A|-D} |X_A| \cdot k = O(|X_A| \cdot |A|^2)$$

car *D* ne peut pas dépasser le nombre d'attributs.

L'algorithme de ID3 est uniquement applicable si tous les attributs sont nominaux. Les attributs du contexte de l'appariement sont cependant linéaires. L'estimation suivante de la complexité des algorithmes est basée sur ce cas.

Pour l'algorithme IAD.O, la complexité est augmentée par la recherche du meilleur seuil de définition des attributs nominaux. La complexité de cette recherche dépend du nombre de valeurs d'attributs, qui est limité par le nombre d'exemples d'apprentissage. Comme la hauteur de l'arbre est limitée à deux fois le nombre des attributs, on obtient comme complexité de IAD.O

$$\text{comp}(IAD.O) = \sum_{k=1}^{D} |X_A| \cdot \sum_{a \in A} |a(X_A)| = O(|X_A|^2 \cdot |A|^2)$$

Le traitement des valeurs inconnues dans l'algorithme IAD.S résulte en une complexité plus grande, provoquée par la profondeur augmentée de l'évaluation de la réduction de l'impureté pour les attributs incomplètement valués. Dans l'équation suivante, $A^i \subset A$ dénote l'ensemble des attributs incomplètement valués, et A^c l'ensemble des attributs complètement valués; pour un attribut a, X_A^{ac} dénote l'ensemble des exemples dont la valeur de *a* est connue, et X_A^{ai} est son complément dans X_A. La complexité de IAD.S se calcule comme suit

$$\begin{aligned}
\text{comp}(IAD.S) &= \sum_{k=1}^{D} |X_A| \cdot \Big(\sum_{a \in A^c} |a(X_A)| + \sum_{a \in A^i} \big(|X_A^{ac}| \cdot |a(X_A^{ac})| + |X_A^{ai}| \cdot \sum_{\substack{a' \in A \\ a' \neq a}} |a'(X_A^{ai})| \big) \Big) \\
&\leq \sum_{k=1}^{D} |X_A| \cdot \Big(\sum_{a \in A^c} |a(X_A)| + \sum_{a \in A^i} \big(|X_A| \cdot \sum_{\substack{a' \in A \\ a' \neq a}} |a'(X_A)| \big) \Big) \\
&= O(|A| \cdot |X_A|^2 \cdot (|A^c| + |A^i| \cdot |A| \cdot |X_A|))
\end{aligned}$$

Si tous les attributs sont complètement valués, IAD.S est de la même complexité que IAD.O, tandis que si tous les attributs sont incomplètement valués, la complexité est de $O(|A|^3 \cdot |X_A|^3)$.

Dans l'algorithme de IAD.L, le coût de la construction d'un sous-arbre de la profondeur k pour l'évaluation de la réduction de l'impureté entre dans le calcul de la complexité, qui s'exprime comme suit

$$\begin{aligned} comp(IAD.L) &= \sum_{k=1}^{D} (|X_A| \cdot \sum_{a \in A} |a(X_A)|)^k \\ &\leq \sum_{k=1}^{D} (|X_A|^2 \cdot |A|)^k \\ &= O(|A| \cdot (|X_A|^2 \cdot |A|)^k) \end{aligned}$$

Si la complexité de ID3 de la génération d'un arbre de décision à partir d'un ensemble de 1000 exemples, décrits par 10 attributs, est de 1, la complexité de IAD.O est alors de 10, celle de IAD.S prend une valeur entre 10 (si toutes valeurs sont connues) et 10^5 (si des valeurs inconnues existent pour chaque attribut), et celle de IAD.L avec une profondeur de 2 est de 10^6. Autrement dit, si la construction d'une règle par IAD.O prend 10 secondes, l'application de IAD.S prend entre 10 secondes et 1 jour, tandis que IAD.L nécessite 10 jours avant d'être terminée. Bien que ces estimations soient approximatives, elles montrent que l'application de l'algorithme IAD.L dans cette étude est exclue.

Dans des contextes comportant uniquement des attributs nominaux, l'algorithme de IAD.L peut cependant être applicable, car, alors, sa complexité est réduite à $O(|A| \cdot (|X_A| \cdot |A|)^k)$.

III.6 Algorithme de l'appariement par un arbre de décision

III.6.1 La classification de couples d'échos par un arbre de décision

L'algorithme de la classification d'objets par un arbre de décision a été défini au chapitre précédent (algorithme II.3). Afin de pouvoir classer des objets, dont une partie des valeurs est inconnue, il faut procéder à des modifications de cet algorithme.

Si n est un noeud d'un arbre ADD dont le test est "$a^n<s$", la classification d'un objet $o=(es,e)$, dont la valeur $a^n(o)$ est inconnue, est effectuée comme suit:

- la classification est achevée pour les deux possibilités $a^n(o)<s$ et $a^n(o)\geq s$,
- le résultat est déterminé comme la moyenne des deux classifications, pondérée selon la fréquence des valeurs de a^n dans l'ensemble d'exemples de test X_T

L'utilisation optionnelle des seuils souples a été introduite dans l'algorithme, dont le fonctionnement est résumé dans l'algorithme III.6. Le résultat de la classification d'un couple (es,e) est la probabilité estimée que (es,e) est un bon appariement.

III.6.2 L'appariement d'échos à l'aide d'un arbre de décision

L'appariement d'échos est la deuxième étape de l'algorithme des méthodes structurées de prévision (algorithme I.2). Nous appliquons un algorithme de l'appariement comme suit: A un instant donné t_{k+1}, tous les appariements possibles entre les séquences $es=(e_1^{es}\in E(I_1),\ldots,e_k^{es}\in E(I_k))$ d'échos sur les images $I_1,\ldots,I_k$ et les échos $e\in E(I_{k+1})$ sur l'image I_{k+1} sont examinés. Les couples, dont la probabilité de représenter un bon appariement est plus grande que 0.5, sont retenus. L'appariement des couples ainsi sélectionnés est effectué comme suit:

(1) dans l'ordre décroissant du coefficient de probabilité,

(2) dans l'ordre croissant de nombre d'échos simples faisant partie des deux échos e_k^{es} et e pour des couples de coefficients égaux,

(3) par choix aléatoire pour les couples de coefficients égaux et de nombre d'échos simples égaux.

Les couples, dont un écho déjà apparié fait partie, sont éliminées de la liste. Les échos imaginaires ne sont utilisés que si aucun de leurs parties a été appariée. Ce fonctionnement est résumé dans l'algorithme III.7.

Grâce au traitement des valeurs inconnues, la classification est possible pour chaque objet (es,e). Le système à base de connaissances $K_{AP}=(CT_{AP},P_{AP},BC_{AP},IC_{AP})$, dont BC_{AP} est défini par l'arbre de décision ADD, et IC_{AP} est une réalisation de l'algorithme III.7, est alors complet dans le contexte CT_{AP}.

Donné : Un contexte de classification $CT=(O,A,P,S)$ de deux classes (+,-), avec $S=\{p\in[0,1]\}$ la probabilité d'appartenance d'un objet à la classe +.

Un arbre de décision probabiliste ADD pour un système à base de connaissances de CT.

Un objet $o\in O$ et le vecteur de valeurs d'attributs $\vec{a}(o)$, dont des valeurs peuvent être inconnues.

Cherché : La probabilité $p(o)$ de l'appartenance de l'objet o à la classe +.

Algorithme :

(0) Noeud n := racine n_0 de ADD.

(1) Si n est un noeud terminal $p_n(o):=p^n$.

Sinon

(2) (a) Si le test de n est de la forme "$a^n<s$":

Soient s^- et s^+ les seuils souples du test, soient ng et nd les enfants de n.

Déterminer $p_{ng}(o)$: $n:=ng$ et continuer avec (1).

Déterminer $p_{nd}(o)$: $n:=nd$ et continuer avec (1).

(i) Si $a^n(o)$ est connu:

- Classification avec seuils durs:

si $a^n(o)<s$, $p_n(o):=p_{ng}(o)$.

si $a^n(o)\geq s$, $p_n(o):=p_{nd}(o)$.

- Classification avec seuils souples:

si $a^n(o)<s^-$, $p_n(o):=p_{ng}$

si $s^-\leq a^n(o)<s$, $p_n(o):=p_{ng}+(p_{nd}-p_{ng})\cdot(a^n(o)-s^-)/(2(s-s^-))$

si $s\leq a^n(o)\leq s^+$, $p_n(o):=p_{nd}+(p_{nd}-p_{ng})\cdot(s^+-a^n(o))/(2(s^+-s))$

si $s^+<a^n(o)$, $p_n(o):=p_{nd}$

(ii) Si $a^n(o)$ est inconnu:

$p_n(o):=p_{ng}(o)\cdot|X_T^{ng}|/|X_T^n|+p_{nd}(o)\cdot|X_T^{nd}|/|X_T^n|$

(b) Si le test de n est de la forme "a^n connu":

Soient ng et nd les enfants de n.

(i) Si $a^n(o)$ est connu: $n:=ng$ et continuer avec (1).

(ii) Si $a^n(o)$ est inconnu: $n:=nd$ et continuer avec (1).

(3) $p(o):=p_{n0}(o)$

Algorithme III.6: Algorithme de classification d'objets dont des valeurs d'attributs peuvent être inconnues

Donné : Une suite d'images $(I_1,...,I_k)$ $(k>0)$, dont les échos sont définis et appariés, et l'image radar I_{k+1}.
Un arbre de décision *ADD* du contexte de l'appariement.

Cherché : L'ensemble $AP(I_k,I_{k+1})$ d'appariements corrects.

Algorithme :

(0) $AP(I_k,I_{k+1}) := \emptyset$.

(1) Définir les échos $E_s(I_{k+1})$ et les échos imaginaires $E_i(I_{k+1})$.
$E(I_{k+1}) := E_s(I_{k+1}) \cup E_i(I_{k+1})$.

(3) Pour tous couples $(es = (e_1^{es} \in E(I_1),...,e_k^{es} \in E(I_k)), e \in E(I_{k+1}))$, déterminer $p=p(\text{classe}(es,e)=+)$ par classification avec *ADD* selon l'algorithme III.6. Créer une liste L des pairs dont $p > 0.5$.

(4) Ordonner la liste L en ordre:
- décroissant du coefficient de probabilité,
- croissant du nombre d'échos simples pour des couples de coefficients égaux,
- aléatoire pour les couples de coefficients égaux et de nombre d'échos simples égaux.

$L = \{(es_1,e_1),...,(es_n,e_n)\}$.

(5) Pour $l=1,...,n$ répéter:

Si es_l n'est pas apparié, et $e_k^{es_l}$ n'est pas un écho imaginaire ou aucun élément de $e_k^{es_l}$ est apparié, et

si e_l n'est pas apparié, et e_l n'est pas un écho imaginaire ou aucun élément de e_l est apparié:

Apparier es_l et e_l.

Algorithme III.7: Algorithme de définition et d'appariement d'échos de pluie sur l'imagerie radar

III.7 Conclusion

Dans ce chapitre nous avons développé une nouvelle méthode du suivi des cellules de pluie sur l'image radar, qui est basée sur l'appariement des échos d'une image à l'autre. Un problème particulier nous est posé par les scissions et fusions des cellules, qui nécessite la concrétisation de la poursuite des cellules. Cette concrétisation a été mis au point par la définition des *séquences strictes d'échos*, qui seront dorénavant interprétées comme représentations des cellules. La limitation de l'observation aux séquences strictes implique, que la fusion de deux ou plusieurs cellules donne lieu à la naissance d'une nouvelle cellule, si aucune des cellules fusionnant n'est très dominante. Cette convention n'est pas toujours conforme au processus météorologique, car, dans un sens physique, une ou plusieurs des cellules fusionnées peuvent continuer leur vie après la fusion. Mais, dû au seuillage fixe appliqué pour l'identification des échos, une observation plus détaillée ne nous est pas possible dans ces situations.

Le problème de l'appariement automatique des échos a été formalisé dans un contexte de classification, qui permet l'apprentissage automatique d'une base de connaissances à partir d'exemples. Le degré de généralité de la base est déterminé par le choix de l'ensemble d'exemples négatifs de l'apprentissage. Un algorithme a été proposée, qui permet la sélection automatique de *presque-instances* comme exemples négatifs, et réduit ainsi le nombre nécessaire d'exemples sans affecter la qualité de la base de connaissances générée. Le bon fonctionnement de cet algorithme sera vérifié dans le chapitre prochain.

Un algorithme de génération d'arbres de décision, nomme IAD.O, a été développé, qui est adapté aux problèmes spécifiques du contexte de l'appariement d'échos: le traitement des attributs linéaires, le traitement du bruit, et le traitement d'attributs incomplètement valués. Cet algorithme ne permet cependant pas de tenir compte du fait, que certaines valeurs d'attributs sont systématiquement inconnues. L'algorithme IAD.S a été proposé pour l'amélioration du traitement de ces attributs. Il utilise une plus grande profondeur de recherche dans l'arbre pour les attributs incomplètement valués. La généralisation de cette procédure, l'algorithme IAD.L, est très intéressant, car il permet l'évaluation de la combinaison de plusieurs attributs. A cause de sa complexité importante provoquée par le traitement des attributs linéaires, il n'est pas applicable au problème de l'appariement. Dans des contextes avec des attributs nominaux, il pourrait cependant présenter une amélioration importante par rapport aux méthodes existantes.

IV

APPLICATION DE LA MÉTHODOLOGIE DÉVELOPPÉE AUX DONNÉES RADAR DE TRAPPES

Ce chapitre a pour objet d'exposer les résultats obtenus par l'application de la méthodologie d'observation de la pluie, qui a été proposée au chapitre précédent. La figure IV.1 présente schématiquement les différentes étapes de la démarche poursuivi, à savoir:

(1) Sélection des données radar de 20 événements de pluie parmi les données disponibles des années 1989 et 1990. Identification et description des échos simples; définition manuelle des échos imaginaires utiles et appariement manuel des échos sur les images.

(2) Partition manuelle des pluies sélectionnées en deux ensembles P_1 et P_2.

(3) Création d'un ensemble EX^+ d'exemples positifs du contexte CT_{AP} à partir de l'ensemble P_1 et des appariements manuels. Génération d'un ensemble EX^- d'exemples négatifs par l'algorithme III.2.

(4) Partition aléatoire de l'ensemble $EX = EX^+ \cup EX^-$ en deux ensembles: l'ensemble X, comprenant deux tiers des instances, et l'ensemble Y, comprenant l'autre tiers des instances.

(5) Application des algorithmes IAD.O et IAD.S pour la génération d'arbres de décision à partir de l'ensemble X (50 applications de chaque technique).

(6) Classification des ensembles X et Y par les 100 arbres générés sous application de l'algorithme III.6 et détermination des taux d'erreur. Analyse et critique des résultats obtenus.

(7) Sélection du meilleur arbre ADD_{APP}.

(8) Application de l'algorithme III.7 sous utilisation de ADD_{APP} aux ensembles P_1 et P_2 et comparaison des résultats de l'appariement automatique avec les appariements manuels.

La vérification des résultats à l'aide d'ensembles de test est alors effectué à des étapes différentes:

- L'ensemble d'exemples X_T est sélectionné aléatoirement dans X par les algorithmes IAD.O et IAD.S avant l'induction des arbres de décision. Il est utilisé pour le test de la signification des tests définis dans l'arbre, et pour la définition des distributions des valeurs d'attributs utilisées pour la classification d'objets par algorithme III.6.
- L'ensemble d'exemples Y, non utilisé pour la génération des arbres, sert à l'estimation du taux d'erreur de la classification par l'algorithme III.6 utilisant les arbres générés.
- L'ensemble de pluies P_2, non utilisé ni pour la génération des arbres, ni pour la sélection du meilleur arbre, sert à l'estimation du taux d'erreur de l'appariement automatique des échos par l'algorithme III.7.

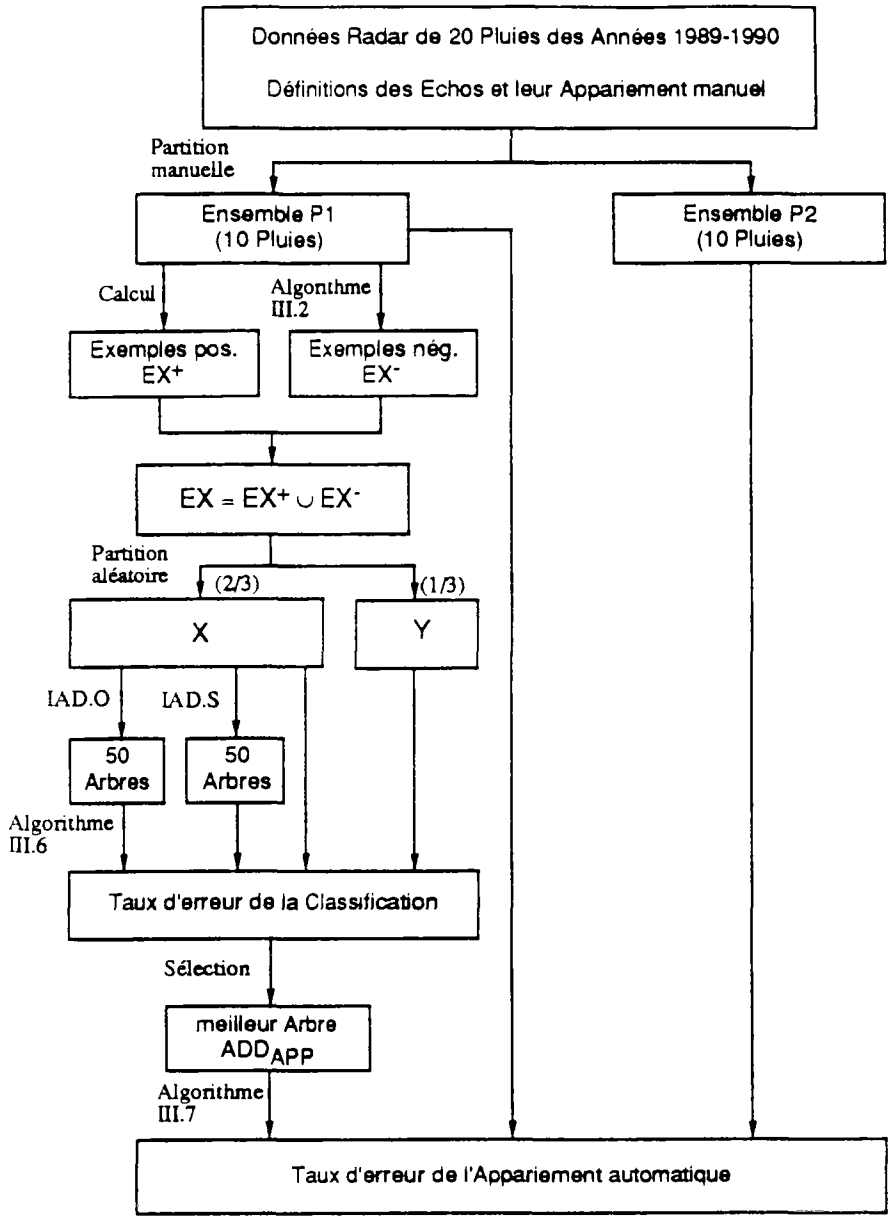


Figure IV.1: Présentation schématique de la démarche poursuivi pour l'apprentissage d'un arbre de décision de l'appariement des échos et pour la vérification des algorithmes proposés

IV.1 Analyse des données radar utilisées

IV.1.1 Les événements de pluies choisis

Les données utilisées dans cette étude sont des mesures du radar de Trappes des années 1989 et 1990. Vingt événements pluvieux de cette période ont été sélectionnés. Les critères de choix étaient les suivants:

- absence d'erreurs de la mesure et de l'archivage sur les images radar (par exemple échos de sol, échos fixes, atténuation,...),
- passage de la zone pluvieuse dans la région centrale de l'image radar (50 km autour de Trappes),
- diversité des types de pluies.

Les caractéristiques des pluies sélectionnées sont résumées dans le tableau IV.2. Des figures montrant des situations caractéristiques de chaque événement se trouvent en annexe. Le lecteur s'y référera opportunément à chaque analyse d'un cas particulier. La base de données comprend 990 images radar, qui couvrent environ 77 heures de mesure. La plus grande partie des pluies importantes des périodes avril-septembre des années 1989 et 1990 est comprise dans cet ensemble.

Le climat estival en Ile de France de ces périodes était marqué par une pluviométrie exceptionnellement déficitaire (tableau IV.1) par rapport à la normale. Néanmoins, les pluies sélectionnées constituent un échantillon, qui représente tous types de pluie habituel en région Ile de France.

année	avril	mai	juin	juillet	août	septembre
1989	220%	30%	100%	70%	60%	80%
1990	130%	35%	env. 150%	40%	65%	65%

Tableau IV.1: Pluviométrie en Ile de France par rapport à la normale des années 1951-1980 (source: La Météorologie, série VII, n° 28-30, 33-35)

IV.1.2 La définition des échos et l'appariement manuel

Par la méthode décrite au chapitre précédent, un ensemble d'échos simples a été défini pour chaque image radar des événements choisis. Au total 10824 échos simples ont été générés, le nombre d'échos par image étant fonction du type de la pluie et de la taille des cellules. Dans le tableau IV.2 sont rappelés les nombres moyens d'échos trouvés.

Tous les échos ont été appariés manuellement à l'aide d'un logiciel de visualisation des images et des échos, dont une description se trouve en annexe. Dans le cas de fusions ou scissions de cellules, des agglomérations d'échos simples ont été définies manuellement comme échos imaginaires, afin d'assurer le suivi correct des cellules. 1129 échos imaginaires ont ainsi été

définis. La base de données formée par les échos et leurs appariements manuels sera utilisée à diverses fins:

- la vérification de l'algorithme de définition automatique des échos imaginaires,
- la génération de l'ensemble des exemples de l'apprentissage,
- la méthode de référence de la prévision, destinée à vérifier le maximum d'efficacité accessible par la prévision basée sur la seule advection,
- l'observation correcte du développement des cellules de pluie.

Date de la pluie	Type de pluie	Nombre d'images	Durée	Advection moyenne		Nombre moyen d'échos	Taille moyenne d'échos	Intensité moyenne d'échos
			(h)	Vit. (km/h)	Dir. (deg.)	(/image)	(km²)	(mm/h)
7.3.89	bande large	43	3.00	29.3	59	10.5	269.7	2.59
4.4.89*	mixte	63	5.00	36.8	350	3.4	286.9	4.36
24.4.89*	convectif	51	4.00	51.2	43	7.8	140.8	3.83
27.4.89*	bande large	80	6.25	36.7	86	4.8	488.7	3.89
10.5.89*	convectif	39	3.00	22.9	56	11.9	68.4	4.39
2.6.89*	bande large	25	2.00	14.4	260	5.2	480.0	3.65
3.6.89	convectif	51	4.00	20.8	119	7.8	62.0	3.51
6.6.89	convectif	52	4.00	38.3	83	21.4	86.6	4.20
27.6.89*	bande étroite	64	5.00	47.6	94	12.2	228.4	3.22
10.7.89*	convectif	51	4.00	10.8	115	4.4	52.6	3.16
7.8.89*	convectif	64	5.00	34.1	61	8.4	123.8	3.98
12.9.89*	convectif	50	4.00	14.9	46	15.3	57.2	5.14
19.9.89*	bande étroite	38	3.00	69.4	29	9.6	86.2	3.32
23.4.90	convectif	46	3.50	32.7	220	23.2	46.1	5.10
14.5.90	bande large	76	6.00	29.1	59	6.8	215.1	3.79
9.6.90	convectif	44	3.50	32.1	163	19.0	60.2	5.55
26.6.90	convectif	39	3.00	34.1	63	9.8	136.7	6.21
21.9.90	mixte	38	3.00	80.8	122	15.2	275.7	4.43
24.9.90	convectif	33	2.50	57.0	86	12.9	124.7	7.72
30.9.90	convectif	43	3.50	40.9	65	18.5	242.4	11.53

Tableau IV.2: Caractéristiques des données sélectionnées (les pluies de l'ensemble P_l sont marquées *) (cf. la visualisation en annexe)

IV.1.3 La définition automatique des échos imaginaires

Au chapitre précédent, un algorithme pour la génération d'un ensemble d'échos imaginaires a été proposé pour définir les échos imaginaires utiles pour le suivi des cellules de pluie en cas de fusion et de scission des échos simples (algorithme III.1). Une vérification du fonctionnement de cet algorithme est possible par la comparaison de l'ensemble des échos imaginaires définis automatiquement, et de l'ensemble d'échos imaginaires définis manuellement. Ce dernier est considéré comme l'ensemble d'échos imaginaires utiles.

Sur une image radar avec k échos simples, l'algorithme III.1 génère k-1 échos imaginaires. Une fonction $\delta(.,.)$ de distance entre échos est utilisée pour la sélection des agglomérations d'échos qui définissent les échos imaginaires. Afin de limiter le nombre d'échos imaginaires générés, il est utile d'introduire dans cet algorithme une distance maximale δ_{max}. Nous définissons cette distance maximale comme fonction de l'écart temporel $\Delta t(I_1,I_2)$ entre deux images I_1 et I_2, car la probabilité de l'occurrence d'une fusion ou d'une scission d'échos simples est d'autant plus grande, que cet écart est plus grand:

$$\delta_{max}(I_1,I_2) = c \cdot \Delta t(I_1,I_2) \qquad [\text{h/mm}]$$

La figure IV.2 montre le nombre total des échos imaginaires et le nombre des échos imaginaires utiles générés par l'algorithme en fonction du facteur c. Le nombre total converge vers le nombre maximal d'échos imaginaires, qui est d'environ 9000; tandis que le nombre d'échos imaginaires utiles retrouvés atteint son maximum pour un facteur c de 3.2. 752 échos utiles sont trouvés par l'algorithme, dont 95% (714) avec un facteur de distance maximale égal à 1. Avec ce facteur, 4070 échos imaginaires sont générés au total.

La figure IV.3 montre le pourcentage d'échos utiles trouvés en fonction du facteur c. Pour un facteur de 1, 63% des échos imaginaires définis manuellement sont retrouvés. Ce taux peut être augmenté à 66%, si le facteur de la distance maximale est plus grand; néanmoins, cette légère amélioration est dévalorisée par le nombre d'échos inutiles très élevé, qui sont générés au même temps. Par conséquent, le facteur de c=1.0 a été retenu pour cette étude.

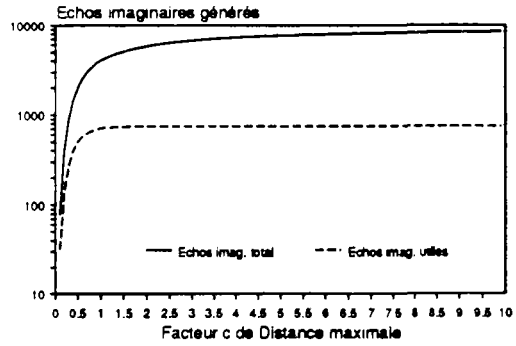


Figure IV.2: Nombre d'échos imaginaires générés par l'algorithme III.1 en fonction de la distance maximale (ensemble des 20 pluies)

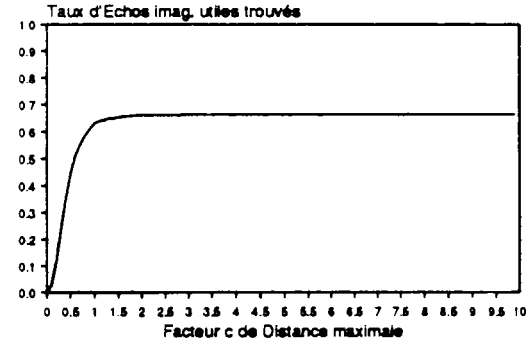


Figure IV.3: Taux d'échos imaginaires retrouvés par l'algorithme III.1 en fonction du facteur c de la distance maximale (ensemble des 20 pluies)

IV.2 Apprentissage automatique d'un arbre de décision pour l'appariement des échos

IV.2.1 Génération des ensembles d'exemples positifs et négatifs

Dans la base des données formée par les échos définis sur les images des 20 événements pluvieux choisis, la classe correcte (bon ou mauvais) est connue pour tous les appariements possibles: si l'appariement est défini manuellement, il s'agit d'un bon appariement, sinon il est mauvais. Une ambiguïté subsiste lorsque les échos imaginaires en tant que tels ne sont pas appariés mais seulement leurs parties. L'appariement de ces échos imaginaires n'est pas incorrecte, mais il engendre une diminution de l'exactitude de l'observation. Nous considérons ces cas comme mauvais appariements, car en général un appariement des échos simples est préférable à l'appariement des échos imaginaires.

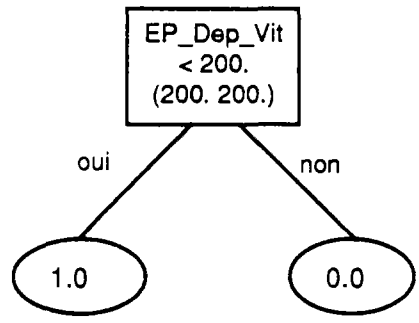


Figure IV.4: Arbre de décision initial ADD_{INI} de l'algorithme de définition de l'ensemble des exemples négatifs (les seuils souples sont indiqués entre parenthèses)

La base de données a été divisée en deux ensembles: l'ensemble P_1 des pluies, qui serviront comme données de l'apprentissage (pluies marquées "*" dans le tableau IV.2), et l'ensemble P_2 de pluies, qui serviront pour la vérification de la règle apprise. Chaque ensemble comprend 10 pluies.

L'ensemble EX^+ d'exemples positifs est généré à partir des appariements manuels des événements de l'apprentissage (2856 instances). Pour sélectionner un ensemble EX^- d'exemples négatifs, l'algorithme III.2 a été employé. La règle initiale utilisée par cet algorithme a été définie par la contrainte, que l'advection des cellules de pluie reste toujours inférieure à 200 km/h. Cette règle a été exprimée sous la forme d'un arbre de décision, qui sera noté ADD_{INI} (figure IV.4). L'algorithme IAD.S a été employé comme méthode G de génération des règles.

Afin de limiter le temps de calcul, l'algorithme III.2 a été arrêté au cycle où moins d'un exemple négatif supplémentaire par événement a été généré. Ceci était le cas après 10 cycles (figure IV.5). 1806 exemples négatifs ont été générés, qui forment l'ensemble EX^-.

L'algorithme de génération d'exemples a été proposé afin de sélectionner de préférence des presque-instances comme exemples négatifs. Une vérification du bon fonctionnement de cet algorithme a été réalisée à l'aide de l'outil de visualisation présenté en annexe. L'auteur a pu constater que, dans la plupart des cas, les mauvais appariements inclus dans l'ensemble EX^- sont des cas où l'écho apparié est proche de l'écho de l'appariement correct.

Les exemples, qui font partie de l'ensemble $EX = EX^+ \cup EX^-$, ont été caractérisés par les 21 attributs du contexte de l'appariement (cf. chapitre III et annexe). Pour 12 attributs, les valeurs sont inconnues pour une partie des exemples. La figure IV.6 montre la distribution des valeurs connues dans l'ensemble d'éléments de EX. Tous les attributs sont connus dans 25% des cas seulement.

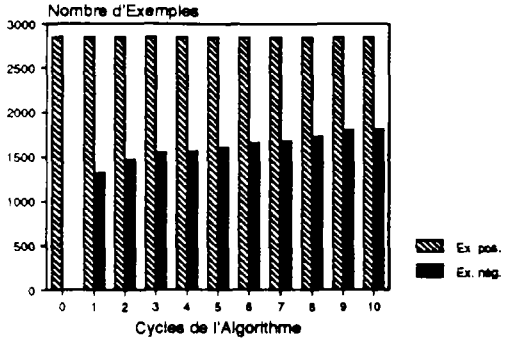


Figure IV.5: Génération de l'ensemble des exemples négatifs par l'algorithme III.2

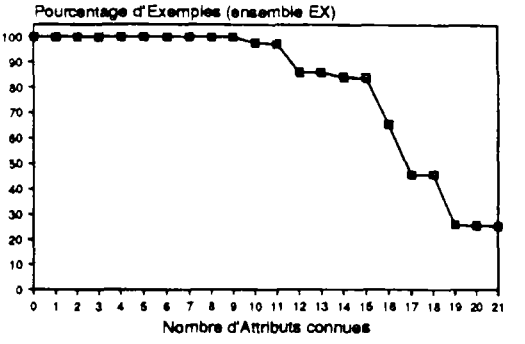


Figure IV.6: Connaissance des valeurs des attributs pour l'ensemble de l'apprentissage EX

IV.2.2 Induction d'un arbre de décision

Plusieurs techniques de l'induction des arbres de décision ont été proposées au chapitre précédent. L'algorithme IAD.L n'étant pas applicable à cause de sa complexité élevée, nous comparons dans la suite uniquement les performances des algorithmes IAD.O et IAD.S (algorithmes III.3 et III.4).

Pour établir cette comparaison, l'ensemble des exemples *EX* a été divisé aléatoirement en un ensemble *X*, comprenant environ deux tiers des cas, et un ensemble *Y* comprenant les autres cas. L'apprentissage des arbres a été effectué à partir de l'ensemble *X*, tandis que la performance des arbres a été testée par la classification de l'ensemble *Y*.

Chacun des deux algorithmes a été appliqué 50 fois. Les arbres générés par chaque algorithme ne sont pas identiques, car les algorithmes de génération ne sont pas déterministes:

- Au début des algorithmes (étape (0)), l'ensemble d'exemples *X* est partitionné aléatoirement en un ensemble X_A, à partir de lequel les tests des noeuds sont sélectionnés, et un ensemble X_T, qui sert pour le test de la signification des tests sélectionnés.
- Dans l'algorithme de IAD.O, les valeurs inconnues sont choisies aléatoirement.

IV.2.2.1 Comparaison des arbres générés par les algorithmes de IAD.O et IAD.S

Les critères d'évaluation des performances des algorithmes sont d'une part ses complexités, et de l'autre part la performance des règles générées, qui s'exprime par la taille des arbres de décision, et par les taux d'erreur de la classification de l'ensemble de *X* utilisé pour l'apprentissage et de l'ensemble *Y* de test.

La complexité des algorithmes se manifeste en besoins de mémoire vive et de temps CPU pour leurs exécutions. Dans la réalisation des algorithmes pour cette étude, celui de IAD.O fait usage de 420 kO de mémoire vive pour l'induction d'un arbre de décision à partir de 3000 exemples décrits par 21 attributs, contre 460 kO pour l'algorithme IAD.S. Le temps de l'exécution sur micro-ordinateur du type PC 386/20 est en moyenne d'environ 2.5 fois plus élevé pour IAD.S que pour

IAD.O (tableau IV.3). Néanmoins, ces différences n'ont que peu d'importance: la complexité des deux algorithmes est suffisamment basse pour permettre leur application.

Technique	Nombre de niveaux des arbres générés	Nombre de noeuds des arbres générés	Temps CPU de l'exécution de l'algorithme (sec)	Nombre d'attributs incomplètement valués utilisés dans l'arbre	Niveau moyen d'attributs incomplètement valués dans l'arbre
IAD.O	7.3	31.3	268	3.5	3.9
IAD.S	8.3	46.9	658	5.5	4.1

Tableau IV.3: Performance des deux techniques de génération d'arbres de décision (moyennes de 50 applications)

Les tailles moyennes des arbres générés par les deux algorithmes sont indiquées dans le tableau IV.3. Les arbres générés par IAD.S sont légèrement plus grands, à cause des noeuds supplémentaires qui sont introduits pour les attributs incomplètement valués (attributs, dont une partie des valeurs est inconnue). Néanmoins, la différence entre les arbres générés par les deux techniques n'est pas très importante.

Le critère le plus important pour l'application, qui nous intéresse dans cette étude, est la qualité de la classification des exemples par les arbres générés. La figure IV.7 montre les taux d'erreur maximaux, minimaux et moyens pour la classification des ensembles X, Y et l'ensemble $EX=X \cup Y$ par les arbres générés par IAD.O et IAD.S. Une distinction est faite entre la classification avec et sans utilisation des seuils souples. Un exemple est considérée comme mal classé, si:

- pour un exemple négatif, la probabilité d'appartenance à la classe positive estimée par la règle est supérieure à 0.5,
- pour un exemple positif, cette probabilité est inférieure ou égale à 0.5.

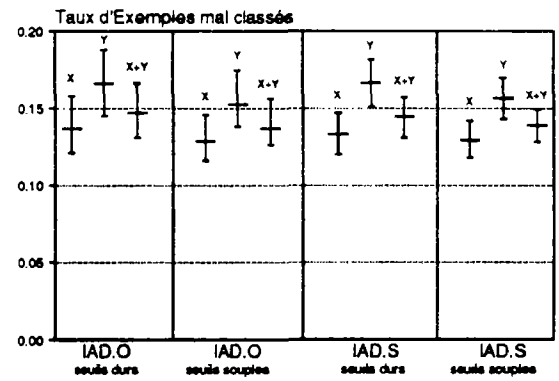


Figure IV.7: Taux d'erreur maximal, minimal et moyen de la classification des ensembles d'exemples par les arbres générés

L'application des seuils souples réduit le taux d'erreur en moyenne de 7% pour les arbres générés par IAD.O et de 4% pour les arbres générés par IAD.S. La réduction est plus élevée pour l'ensemble Y de test que pour l'ensemble X de l'apprentissage (8% *vs.* 6% pour IAD.O, 5% *vs.* 3% pour IAD.S). Ce résultat montre, que l'influence du bruit sur la classification peut être sensiblement réduit par les seuils souples, et que la classification devient plus fiable avec cette technique.

Peu de différences peuvent être constatées entre les taux d'erreur de la classification par les arbres générés par IAD.O de l'une part et IAD.S de l'autre part. Bien que la variabilité du taux soit

moins accentuée pour les 50 arbre générés par IAD.S, les taux d'erreur minimaux sont pratiquement équivalents pour les deux techniques. Ceci peut s'expliquer par deux raisons:

- La différence des distributions des valeurs d'attributs entre les exemples de valeur connue et les exemples de valeur inconnue est moins importante qu'estimée. Dans ce cas, la méthode de traitement des valeurs inconnues appliquée par IAD.O serait justifiée.
- L'importance des attributs incomplètement valués est moins accentuée qu'attendu. Dans ce cas, ces attributs peuvent être remplacés par ceux, dont toutes les valeurs sont connues.

En fait, comme le montre le tableau IV.3, les attributs incomplètement valués sont plus souvent utilisés dans les arbres construits par IAD.S. Leur niveau dans les arbres est cependant en moyenne égale au niveau des autres attributs utilisés; ce qui signifie que leur importance n'est pas supérieure. Nous avons donc surestimé leur importance relative dans la phase du développement de l'algorithme IAD.S.

IV.2.2.2 Choix d'un arbre de décision

De chaque ensemble de 50 arbres, celui résultant en taux d'erreur minimal de la classification de l'ensemble *EX* avec seuils souples a été analysé. Celui de l'ensemble généré par IAD.O (nommé *ADD.O*), est montre dans la figure IV.9, et celui de l'ensemble généré par IAD.S (nommé *ADD.S*) dans la figure IV.10. Pour les noeuds terminaux, les facteurs de probabilité p^n sont indiqués, et pour les noeuds non terminaux, les seuils souples sont présentés entre parenthèses. Une description des attributs se trouve en annexe.

Pour ces deux arbres, les taux d'erreur de la classification des ensembles *X*, *Y*, et $EX=X \cup Y$ sont montrés dans le tableau IV.4, qui indique aussi les taux d'erreur minimaux pour chaque ensemble d'arbres. Bien que le taux d'erreur de la classification de l'ensemble *EX* soit minimal pour *ADD.O* et *ADD.S*, ceci n'est pas le cas pour les deux sous-ensembles *X* et *Y*. Néanmoins, les taux sont très proche des taux minimaux. En conséquence, nous considérons les deux arbres comme les meilleurs générés par les deux algorithmes IAD.O et IAD.S.

Arbre	Taux d'exemples mal classés		
	X	*Y*	$EX=X \cup Y$
ADD.S	11.9	14.5	12.8
Taux minimaux des arbres générés par IAD.S	11.8	14.3	12.8
ADD.O	11.8	14.1	12.6
Taux minimaux des arbres générés par IAD.0	11.6	13.8	12.6

Tableau IV.4: Taux d'erreur des deux arbres sélectionnés et taux d'erreur minimaux des 50 arbres générés par chaque algorithme

Le taux d'erreur de la classification par *ADD.O* est légèrement inférieur à celui de la classification par *ADD.S*. Cette petite différence n'est cependant pas significative; la performance des deux arbres doit être considéré comme équivalente.

Dans tous les deux arbres, l'attribut "Co_Dep_Vit", qui exprime la vitesse de déplacement résultant de l'appariement de deux échos, se trouve à la racine, ce qui indique l'importance prépondérante de cet attribut. Le nombre d'attributs utilisés dans *ADD.O* est 8, dont 4 sont incomplètement valués. Dans *ADD.S*, le nombre d'attributs est 12, dont 6 sont incomplètement valués, et dont 7 sont aussi utilisés dans *ADD.O*. Un test de la forme "a_k connu" se trouve 4 fois dans *ADD.S*.

Chacun des deux arbres possède 8 niveaux, mais l'arbre *ADD.O* comprend seulement 37 noeuds, contre 49 de *ADD.S*. En conséquence, la classification par *ADD.S* est plus fine, le nombre moyen d'exemples associés à chaque noeud terminal étant plus petit que pour *ADD.O*.

Si on ne tient pas compte des seuils souples, chaque chemin de la racine à une feuille correspond à une règle individuelle de classification. Ainsi, une vérification de la plausibilité du raisonnement des arbres est possible. Pour tous les deux arbres, cette analyse n'a pas relevé de règles, qui sont à priori fausses.

On peut alors conclure, que la qualité des deux meilleur arbres générés par IAD.O et IAD.S est tout à fait équivalente, à part de la taille plus élevée de celui généré par IAD.S. La supériorité attendue de l'algorithme IAD.S n'est donc pas démontrée. L'arbre *ADD.O*, compte rendu de sa taille inférieure, sera utilisée ultérieurement dans cette étude. Nous le notons dorénavant comme arbre ADD_{APP}.

Le pourcentage d'exemples mal classés par cet arbre est de 11.8% pour l'ensemble *X*, et de 14.1% pour l'ensemble *Y*. Ce taux d'erreur ne prend cependant pas en considération le coefficient de probabilité de la classification. Dans l'algorithme III.7, ce coefficient détermine l'ordre, dans lequel les échos sont appariés. Ainsi, si un mauvais appariement d'un couple d'échos (es,e_1) est classé comme positif avec un coefficient de probabilité p_1, tandis que l'appariement correct (es,e_2) est classé positif avec un coefficient $p_2>p_1$, le deuxième appariement sera préféré au premier. Comme le montre la figure IV.8, 85% des exemples correctement classés le sont avec un coefficient de probabilité supérieur à 0.7, tandis qu'il est inférieur à 0.7 pour 55% des exemples incorrectement classés. Le taux des cas mal classés par l'algorithme III.7 sera alors inférieur aux taux d'erreur mentionnés ci-dessus pour la classification excluant le coefficient de probabilité.

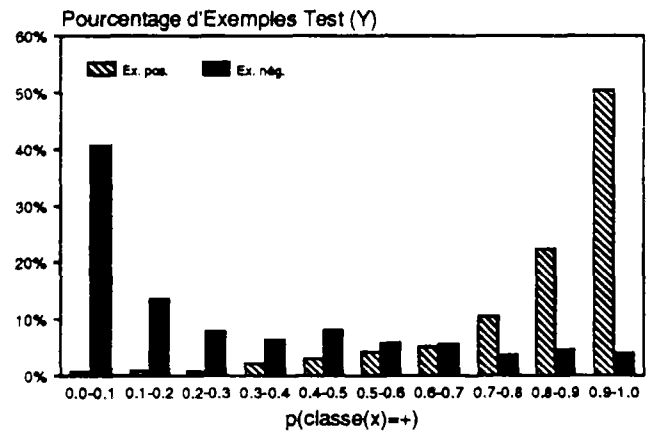


Figure IV.8: Distribution des probabilités estimés de l'appartenance à la classe positive des exemples de test (classification par l'arbre ADD_{APP})

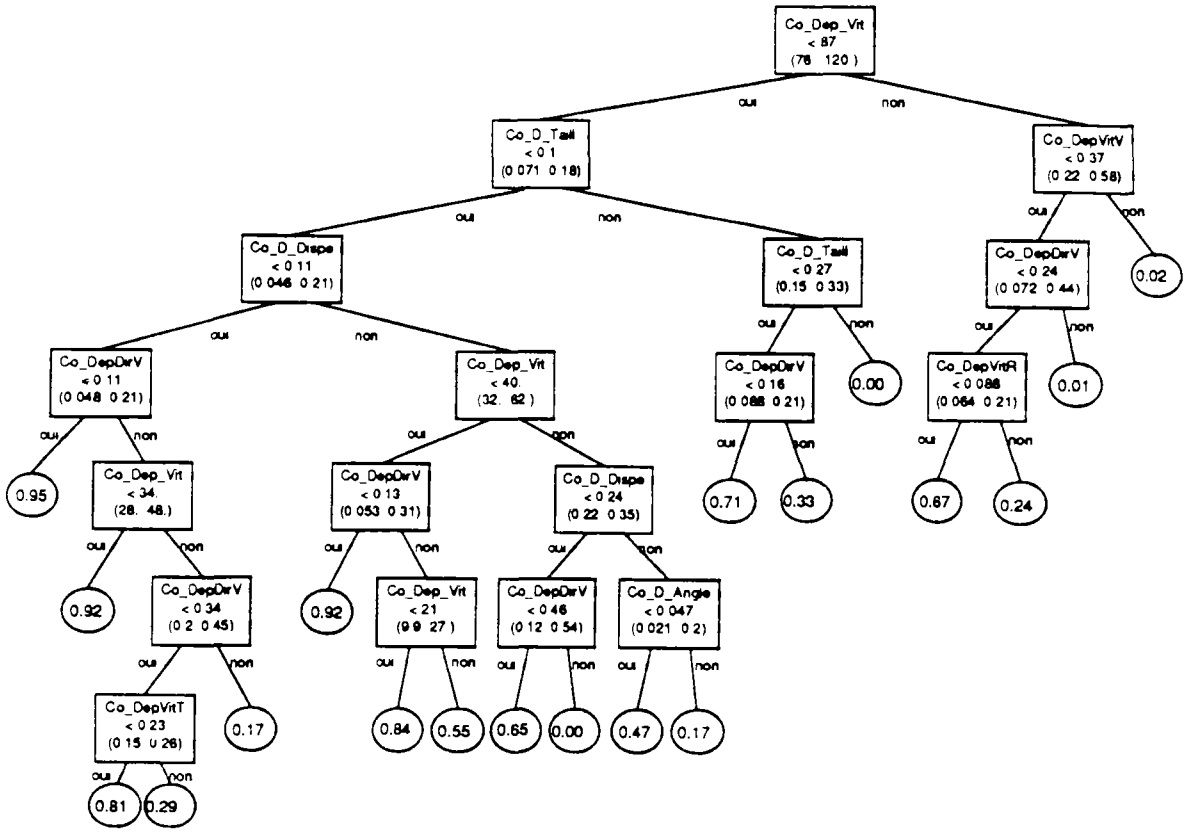


Figure IV.9: Le meilleur arbre de décision $ADD.O{=}ADD_{APP}$ généré par IAD.O

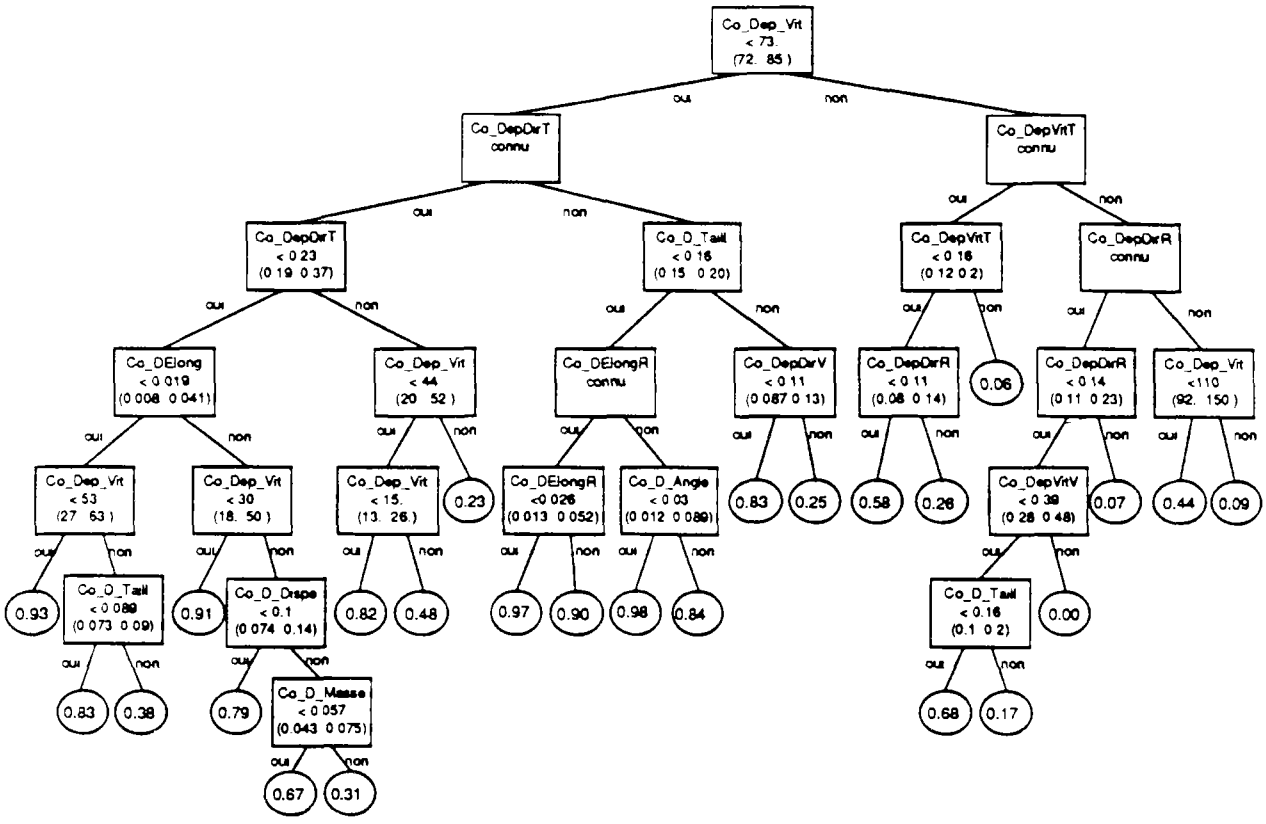


Figure IV.10: Le meilleur arbre de décision *ADD.S* généré par IAD.S

IV.3 Application de l'arbre de décision sélectionné à l'appariement des échos

IV.3.1 Taux d'erreur de l'appariement

L'arbre de décision sélectionné a été appliquée par l'algorithme III.7 à l'échantillon des 20 pluies de la base de données. Les 10 pluies de l'ensemble P_2 sont considérées comme ensemble de test de la règle, simulant son application opérationnelle, car elles n'étaient pas connues à l'algorithme de l'apprentissage. Pour deux images successives I_1 et I_2, l'évaluation de l'algorithme est effectué comme suit:

(1) Les appariements sont réalisés par l'algorithme.

(2) Pour chaque écho e_1^i de l'image I_1, l'appariement par l'algorithme est comparé à l'appariement manuel.

Nous notons un appariement effectué par la règle ADD_{APP} avec $(e_1^i,e_2^j)^A$, et un appariement manuel $(e_1^i,e_2^j)^M$. Un écho e_1^k, qui n'est pas apparié à un écho de I_2, est noté $(e_1^k,\text{-})$. Pour un écho e_1^i, trois types de cas sont distingués:

(1) cas de ***bon appariement***:
- l'écho associé à e_1^i par la règle est identique à l'écho choisi manuellement $((e_1^i,e_2^k)^A \wedge (e_1^i,e_2^k)^M)$,
- e_1^i n'est pas apparié ni manuellement ni par l'algorithme $((e_1^i,\text{-})^A \wedge (e_1^i,\text{-})^M)$;

(2) cas de ***mauvais appariement***:
- l'écho associé à e_1^i par la règle est différent de l'écho choisi manuellement $((e_1^i,e_2^k)^A \wedge (e_1^i,e_2^{l \neq k})^M)$,
- e_1^i n'est pas apparié manuellement, mais par l'algorithme $((e_1^i,e_2^k)^A \wedge (e_1^i,\text{-})^M)$;

(3) cas d'***appariement non reconnu***:
- e_1^i est apparié manuellement, mais non par l'algorithme $((e_1^i,\text{-})^A \wedge (e_1^i,e_2^l)^M)$.

Appariement automatique	Appariement manuel		
	(e_1^i, e_2^l)	$(e_1^i,\text{-})$	(e_1^i, e_2^2)
(e_1^i, e_2^l)	bon appariement	mauvais appariement	mauvais appariement
$(e_1^i,\text{-})$	non reconnu	bon appariement	(non reconnu)

Tableau IV.5: Tableau d'évaluation des appariements effectués par l'algorithme III.7

Ce mode d'évaluation est résumé dans le tableau IV.5. Remarquons que pour toute méthode de prévision automatique de l'advection des cellules, un mauvais appariement pose plus de problèmes qu'un appariement non reconnu, car dans le premier cas les vecteurs sont faussés sans que cette erreur puisse être détectée, tandis que dans le deuxième cas le dysfonctionnement peut être constaté automatiquement.

Les échos imaginaires posent un problème pour l'évaluation des appariements automatiques. Le fait, que 37% des échos imaginaires utiles ne sont pas retrouvés par l'algorithme III.1, tandis qu'environ 4 fois plus d'échos inutiles que échos utiles sont définis, peut biaisé l'évaluation de l'algorithme de l'appariement pour trois raisons:

- Si des échos imaginaires sont appariés manuellement, et non par l'algorithme, ceci peut être dû à une insuffisance de la définition automatique des échos imaginaires utiles.
- Si des échos imaginaires sont appariés par l'algorithme, et non manuellement, ceci peut être dû à la définition des échos imaginaires inutiles. Mais leur appariement n'est pas nécessairement mauvais, car un appariement biunivoque des parties des échos imaginaires peut éventuellement être correct.
- Si l'écho simple e_1^1 est apparié manuellement à un écho e_2^1, mais l'algorithme apparie l'écho e_1^1 à un écho imaginaire, qui comprend les deux échos simples e_2^1 et e_2^2, dont e_2^1 présente une cellule beaucoup plus importante que e_2^2, ceci est une erreur qui reste généralement sans influence à la prévision.

En conséquence, l'évaluation de la performance de l'algorithme III.7 et de l'arbre ADD_{APP} a été effectuée sous l'hypothèse, que tous échos imaginaires utiles soient connus à l'algorithme. L'étape (1) de l'algorithme III.7 a été modifié pour reprendre uniquement les échos imaginaires définis manuellement. L'évaluation de la performance de l'algorithme avec la définition automatique des échos imaginaires sera entreprise dans le chapitre prochain, où le taux d'erreur de la prévision des lames d'eau avec les appariements par l'algorithme sera comparé au taux d'erreur avec les appariements corrects.

Les résultats de l'évaluation sont montrés dans le tableau IV.6. La performance de la règle sélectionnée ADD_{APP} est comparée à celle de la règle ADD_{INI}. Ces résultats montrent, qu'avec une règle trop générale, comme ADD_{INI}, le taux de mauvais appariements est très élevé. Environ un quart des échos sont incorrectement appariés; par conséquent une observation des cellules est impossible.

La règle ADD_{APP}, générée par l'apprentissage automatique, permet un appariement correct de plus de 93% des échos. Le taux moyen des mauvais appariements est de 1.6%; il reste pour toutes les pluies inférieur à 4%. La différence entre les résultats pour les pluies utilisées pour l'apprentissage de la règle, et les pluies de test est négligeable. La qualité prédictive de la règle est ainsi démontrée.

Pour cinq pluies, le taux de bons appariements est inférieur à 90%: dans trois cas à cause d'un taux d'échos non reconnus relativement élevé (7.3.1989, 27.6.1989, et 21.9.1990), et dans deux cas aussi à cause d'un taux relativement élevé d'échos mal appariés (19.9.1989, et 30.9.1990). Examinons les raisons de ces "contre-performances".

*** La pluie du 7.3.1989**

Cette pluie est caractérisée par une zone large de pluie stratiforme (cf. annexe 3). L'intensité de la pluie est très faible; pour 80% de la surface des échos simples elle est proche du seuil de définition d'échos. Par conséquent, de petites variations de l'intensité de la pluie d'une image à l'autre provoquent des changements importants des caractéristiques des échos. De ce fait, les échos sont en partie mal reconnus sur les premières cinq images. La définition des échos avec un seuil d'intensité plus élevé pourra améliorer l'identification des cellules (cf. § V.3.4).

	ADD_{INI}			ADD_{APP}		
	bon appar.	non reconnu	mauvais appar.	bon appar.	non reconnu	mauvais appar.
Pluies de l'apprentissage (ensemble P_1)						
4.4.89	76.9	5.1	17.9	94.7	5.3	0.0
24.4.89	63.1	11.9	25.0	95.2	3.6	1.2
27.4.89	67.3	9.6	23.1	94.3	5.7	0.0
10.5.89	66.9	7.3	25.8	95.9	1.6	2.4
2.6.89	80.8	3.8	15.4	94.1	5.9	0.0
27.6.89	65.2	9.8	25.0	87.2	10.5	2.3
10.7.89	70.9	7.3	21.8	100.0	0.0	0.0
7.8.89	65.9	9.9	24.2	93.3	5.6	1.1
12.9.89	63.2	8.6	28.2	96.9	1.8	1.2
19.9.89	58.1	13.3	28.6	87.6	8.6	3.8
moyenne	67.8	8.7	23.5	93.9	4.9	1.2
Pluies de test (ensemble P_2)						
7.3.89	67.0	9.8	23.2	83.8	14.4	1.8
3.6.89	70.7	7.3	22.0	96.3	2.4	1.2
6.6.89	61.1	11.1	27.8	96.1	2.6	1.3
23.4.90	64.5	10.2	25.3	96.3	2.9	0.8
14.5.90	64.4	9.6	26.0	93.2	4.1	2.7
9.6.90	65.0	9.5	25.5	97.5	2.0	0.5
26.6.90	62.1	10.7	27.2	92.2	4.9	2.9
21.9.90	64.0	11.8	24.2	87.6	10.6	1.9
24.9.90	64.4	8.9	26.7	92.6	5.1	2.2
30.9.90	60.8	10.6	28.6	88.0	8.0	4.0
moyenne	64.4	10.0	25.7	92.4	5.7	1.9
moyenne totale	66.1	9.3	24.6	93.1	5.3	1.6

Tableau IV.6: Performance de l'algorithme III.7 avec la règle initiale de l'apprentissage ADD_{INI} et avec la règle générée par l'apprentissage ADD_{APP}

* **La pluie du 27.6.1989**

La pluie du 27 juin 1989 est caractérisée par une bande pluvieuse étroite, dans laquelle des cellules intenses sont imbriquées, et dont l'axe principale est sud/nord (cf. annexe 3). La bande se déplace ouest en est, tandis que les cellules intenses sont advectées sud/nord à l'intérieur de la bande. Les échos simples sont généralement des tronçons de la bande et non des cellules intenses.

Le taux relativement élevé d'échos non reconnus est dû à un déplacement du centre de gravité à l'intérieur des échos, lui-même provoqué par le déplacement et la croissance/décroissance des cellules intenses. Il en résulte une différence entre l'advection des cellules et le déplacement des centres de gravité des échos, utilisé par l'algorithme pour estimer l'advection, qui influence cinq des huit attributs utilisés dans ADD_{APP}. Une résolution de ce problème pourra être la définition des échos avec un seuil d'intensité plus élevé, afin d'identifier les cellules intenses et leur advection (cf. § V.3.4).

* **La pluie du 21.9.1990**

L'advection de la pluie du 21 septembre 1990 est supérieure à celle de toutes les autres pluies. Elle est de 80 km/h en moyenne, mais pour certaines cellules elle dépasse parfois les 100 km/h. Aucun des événements utilisés pour l'apprentissage de la règle ne possède ces caractéristiques. La plus grande partie des échos non reconnus ont une advection très forte. La règle est alors trop spécifique, ce qui indique une insuffisance de l'ensemble des exemples d'apprentissage.

* **La pluie du 19.9.1989**

Le taux des scissions et fusions des cellules est plus important pour la pluie du 19 septembre 1989, que pour les autres pluies. Dans ces cas il existent deux sources d'erreurs:

- L'écho imaginaire défini manuellement, afin de tenir compte d'une fusion, est non reconnu à cause de différences principales entre les caractéristiques des échos imaginaires et des échos simples. Notamment les paramètres de forme (angle de l'axe principale, dispersion) n'ont parfois que peu de signification pour les échos imaginaires.
- Un écho simple représentant une grande partie de la masse d'un écho imaginaire est préféré à l'écho imaginaire pour l'appariement avec l'écho créé par la fusion, souvent à cause de la proximité des centres de gravité.

Ce dysfonctionnement peut être dû à un nombre insuffisant d'appariements avec des échos imaginaires dans l'ensemble des exemples d'apprentissage. Toutefois, la résolution de ce dysfonctionnement ne sera pas tentée, car elle est à la fois non triviale (comment caractériser la forme des échos imaginaires ?) et de peu d'importance sur l'erreur de prévision (cf. § V.3.1).

* **La pluie du 30.9.1990**

La pluie du 30 septembre 1990 est un événement convectif, qui est caractérisé par un développement rapide de cellules d'une très forte intensité. Il en résulte la fusion de petits échos, ce qui contribue à rendre difficile la reconnaissance automatique des échos imaginaires reconnus par l'homme.

IV.3.2 Réflexion sur l'importance des résultats obtenus pour les objectifs de cette étude

IV.3.2.1 La prévision des lames d'eau

Les mauvais appariements ont deux effets négatifs sur la prévision des lames d'eau:

- l'altération des vecteurs individuels de déplacement,
- un mauvais calcul des vecteurs moyens.

Le vecteur moyen sera employé pour la prévision du déplacement des cellules, pour lesquelles un vecteur individuel n'a pas pu être établi. L'influence des mauvais appariements s'étend alors à toutes les cellules non reconnues.

En revanche, la non-reconnaissance des échos influe uniquement sur la prévision des cellules non reconnues, ce qui a pour conséquence de diminuer la qualité de la prévision exclusivement lorsque l'advection des cellules non reconnues diffère sensiblement de l'advection moyenne des cellules bien appariées.

Nous pensons à priori, que le taux moyen des mauvais appariements de 1.6% obtenu avec la règle ADD_{APP}, est assez bas pour ne pas influer sur la qualité de la prévision des lames d'eau. Néanmoins, l'influence de la définition automatique incomplète de l'ensemble des échos imaginaires utiles sera à vérifier dans le chapitre suivant.

IV.3.2.2 L'observation du développement des cellules

La prise en compte du cycle de vie des cellules de pluie pour la prévision nécessite une observation correcte des cellules. Cette observation est interrompue et en cas de mauvais appariement, et en cas de non-reconnaissance d'échos. Néanmoins, le premier type d'erreur est plus important que le deuxième, car l'observation est fausse dans le cas d'un mauvais appariement, tandis que dans l'autre cas elle n'existe pas et l'erreur peut être identifiée.

L'observation correcte du cycle de vie est surtout importante pour les pluies présentant un fort développement des cellules. Ces pluies sont en grande majorité les pluies orageuses. Parmi les pluies traitées, 12 ont un caractère convectif. Pour ces pluies, le taux de bons appariements est de 95.0% en moyenne, avec un écart-type de 3.1. Autrement dit, la probabilité d'observer une cellule de pluie convective d'une durée de vie assez longue correctement pendant un intervalle de 30 minutes est d'environ 70%, elle décroît à 50% pour l'intervalle de 60 minutes. Cette probabilité représente cependant une estimation pessimiste: en fait, les deux événements

- "la cellule C a été appariée correctement de l'image I_1 à l'image I_2"
- "la cellule C a été appariée correctement de l'image I_2 à l'image I_3"

ne sont pas indépendants, car la probabilité d'un appariement correct est plus grande, si les valeurs des attributs historiques sont correctes. La probabilité d'un mauvais appariement après une observation correcte d'une certaine durée est alors plus petite qu'après un appariement incorrect.

IV.3.3 Comparaison des résultats obtenus avec ceux d'autres systèmes de prévision

La comparaison des résultats obtenus par l'application de la méthode proposée dans cette étude avec ceux obtenus par les autres méthodes employées pour l'appariement des cellules sur l'image radar se heurte aux problèmes des différentes définitions des échos et de la vérification des appariements effectués. Il semble, que la vérification manuelle appliquée dans cette étude soit le seul moyen d'évaluation des techniques.

Récemment, la performance du système SCOUT (Einfalt et *al.* 1990) dans son application opérationnelle a été évaluée par Jacquet et Neumann (1991). Ce système emploie une méthode structurée de prévision proche de celle appliquée dans cette étude. L'appariement des échos dans SCOUT est basée sur des heuristiques concernant la variabilité des caractéristiques de cellules de pluie. Jacquet et *al.* ont analysé les résultats de la prévision par SCOUT pour 9 des pluies utilisées dans cette étude.

Un des critères de cette évaluation était le taux de reconnaissance d'échos, sans prise en compte de l'exactitude des appariements effectués. Ils montrent, que 64 à 78% de la totalité des échos définis par SCOUT sont reconnus, avec une valeur moyenne de 73%. Après visualisation des appariements à l'aide de l'outil développé dans cette étude, ils estiment que le taux de mauvais appariements est d'environ 10%. Pour une estimation plus exacte du taux d'erreur, une vérification manuelle systématique des appariements, comme elle a été effectuée dans cette étude, aurait été nécessaire, qui présente cependant un travail dépassant le cadre de cette étude.

Toutefois, à cause des différences dans la technique de l'identification d'échos, la comparaison des résultats semble difficile; par exemple le nombre d'échos simples définis par SCOUT est souvent plus élevé que le nombre d'échos définis par la méthode appliquée dans cette étude à cause des seuils variables utilisés par SCOUT, tandis que le nombre d'échos imaginaires définis par SCOUT est limité à 10.

IV.4 Conclusion

La méthodologie de l'apprentissage d'un arbre de décision dans le contexte d'appariement des échos, qui a été développée au chapitre précédent, a été appliquée aux données radar de Trappes. Un échantillon de 20 pluies a été sélectionné, qui représente environ 77 heures de mesure. Pour ces images, les ensembles d'échos simples ont été définis. Leur appariement a été effectué manuellement à l'aide d'un logiciel de visualisation des images, qui a été développé dans le cadre de cette étude. L'ensemble des échos imaginaires utiles, qui sont nécessaires pour le suivi correct des cellules, a été défini également par cet outil.

Une partie des pluies nous ont servi pour la génération d'un ensemble d'exemples de l'appariement. Les exemples positifs étant donnés par les appariements manuels, un ensemble d'exemples négatifs a été généré automatiquement par l'algorithme proposé au chapitre précédent (algorithme III.2). La visualisation des exemples choisis nous a permis de constater le bon fonctionnement de l'algorithme dans cette application: la plus grande partie des exemples négatifs générés sont des presque-instances, qui sont proche de bons appariements. L'algorithme III.2 peut être utile dans tous contextes de classification de deux classes, où le nombre d'exemples potentiels d'une classe est beaucoup plus important que le nombre d'exemples potentiels de l'autre classe.

L'ensemble d'exemples a été divisé aléatoirement en un ensemble d'apprentissage et un ensemble de test. A partir de l'ensemble d'apprentissage, 50 arbres de décision ont été générés par chacune des deux algorithmes proposés (IAD.O et IAD.S). Leur performance a été testée par classification de l'ensemble de test. Cette classification a montré une plus grande variabilité des taux d'erreur pour les arbres générés par IAD.O que pour ceux générés par IAD.S. La performance des meilleurs arbres de chaque technique est cependant pratiquement équivalente, malgré le fait que l'algorithme IAD.S permet une meilleure utilisation des attributs incomplètement valués. Nous avons expliqué ce résultat inattendu avec la surévaluation de l'importance de ces attributs pour la classification correcte. Il serait toutefois intéressant de comparer les deux algorithmes dans d'autres contextes d'apprentissage.

Car le meilleur arbre généré par IAD.O est plus efficace à cause de sa taille inférieure, il a été sélectionné comme arbre ADD_{APP}, qui sera utilisé ultérieurement dans cette étude. Le taux d'erreur de l'appariement d'échos par cet arbre a été évalué par comparaison des appariements effectués automatiquement à ceux définis manuellement. Le taux d'erreur est en moyenne inférieur à 2%, et pour les pluies utilisées pour l'apprentissage de l'arbre, et pour les autres pluies; ce qui démontre la haute performance prédictive de l'arbre. Les mauvaises appariements effectués, et les cas d'échos non reconnus, sont dus à une description insuffisante de l'advection des cellules, et au problème de la description des échos imaginaires. Une trop grande spécificité de l'arbre est révélée par une pluie d'une advection très forte: dans une partie de cas, où un appariement n'est pas reconnu pour cette pluie, ce dysfonctionnement peut être dû au fait, que l'ensemble d'exemples de l'apprentissage n'inclue pas de cellules d'une advection très forte. L'importance de ce résultat est à évaluer par son influence sur le taux d'erreur de la prévision (cf. § V.3.1).

L'évaluation de l'algorithme de l'appariement a été effectuée sous l'hypothèse, que tous échos imaginaires utiles soient connus à l'algorithme. En application opérationnelle, ceci n'est pas le fait. L'évaluation du fonctionnement de l'algorithme avec définition automatique des échos imaginaires est possible uniquement par le critère de l'augmentation du taux d'erreur de la prévision, comparé à la prévision avec les appariements manuels. Cette évaluation sera entreprise au chapitre prochain.

V

LA PRÉVISION DE PLUIE BASÉE SUR LA SEULE ADVECTION DES CELLULES DE PLUIE

Ce chapitre a pour sujet l'élaboration du système de prévision dans sa forme complète, et l'évaluation de sa performance. Le système, que nous baptisons ***PROPHETIA*** (**P**révision de la pluie par **R**adar p**O**ur les **P**roblèmes de l'**H**ydrologie urbaine **E**mployant des **T**echniques de l'**I**ntelligence **A**rtificielle), est intégré dans un structure informatique, qui comprend les outils de définition des exemples de l'apprentissage, le système de l'apprentissage automatique, ainsi que les différentes bases de données et un module d'évaluation des prévisions. La figure V.1 montre schématiquement le flux des données entre les différentes parties de cette structure.

PROPHETIA est un système de prévision automatisé et structuré. Le principe de son fonctionnement est alors décrit par l'algorithme I.2. Cet algorithme comprend quatre étapes, qui se présentent avec les notions introduites comme suit:

- définition des échos,
- appariement des échos,
- caractérisation des séquences strictes d'échos,
- prévision des lames d'eau.

Les réalisations des étapes de la définition et de l'appariement des échos ont été examinées aux chapitres précédents. Dans ce chapitre nous présenterons dans un premier temps la réalisation des étapes de la caractérisation des cellules et de la prévision des lames d'eau, qui est basée sur la seule advection des cellules. Dans un deuxième temps nous développerons un critère hydrologique pour l'évaluation des prévisions de lames d'eau. Dans la troisième partie les prévisions effectuées par PROPHETIA pour les pluies de la base de données seront analysées. Les résultats seront ensuite comparés aux résultats obtenus pour les mêmes données avec d'autres méthodes de prévision.

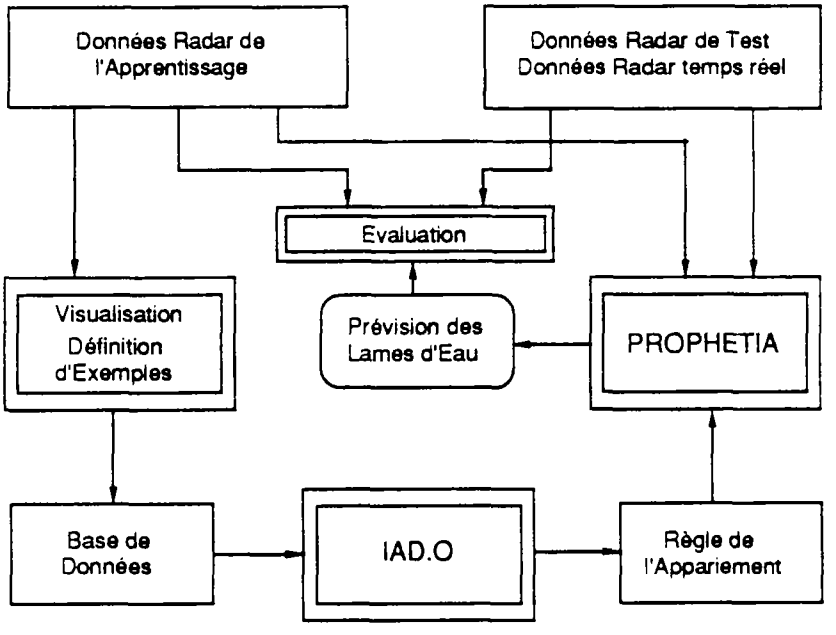


Figure V.1: Présentation schématique du flux des données entre les différents systèmes développés

V.1 Le système de prévision PROPHETIA

V.1.1 La caractérisation des séquences strictes d'échos

Pendant l'étape de la caractérisation, les caractéristiques des cellules de pluie sont identifiées à l'aide des séquences strictes d'échos définis par les appariements. Cette caractérisation permet ensuite d'estimer l'état des cellules à échéance de la prévision. Dans l'étude présentée dans ce chapitre, la prévision reposera sur la seule caractéristique de l'advection des cellules. L'advection observée sera extrapolée dans l'avenir proche, afin de prévoir les régions touchées par la pluie, et les lames d'eau attendues.

V.1.1.1 La définition de l'advection pour un seul pas de temps

Nous considérons deux images I_1 et I_2, mesurées aux instants t_1 et t_2, sur lesquelles une cellule de pluie C est représentée par les échos $e_1 \in E(I_1)$, $e_2 \in E(I_2)$. L'advection de la cellule dans l'intervalle (t_1,t_2) peut être caractérisée de deux manières différentes.

La première méthode envisageable est la définition de l'advection basée sur une comparaison des formes et/ou des distributions des intensités de pluie des deux échos. Il s'agit de chercher le vecteur $\vec{v}$, qui donne la meilleure concordance entre l'écho e_1, déplacé par $\vec{v}$, et l'écho e_2. Une technique pour définir cette concordance est la méthode de la corrélation croisée, appliquée à deux sous-images comprenant uniquement les deux échos. L'avantage de cette technique est, qu'elle permet de tenir compte de la structure intérieure des cellules. Sa complexité de calcul est cependant importante. Un autre inconvénient est l'influence d'éventuelles variations à l'intérieur de la cellule dans l'intervalle (t_1,t_2) sur le résultat obtenu.

La méthode appliquée dans PROPHETIA est le calcul de l'advection comme déplacement du centre de gravité des échos. L'advection dans l'intervalle (t_1,t_2) est alors défini comme

$$\vec{v}_{(t_1,t_2)}(C) = cg(e_2) - cg(e_1)$$

où $cg(e)$ signifie le centre de gravité de l'écho e. L'avantage de cette méthode est sa simplicité de calcul. Comme pour la première méthode, le vecteur $\vec{v}$ de l'advection peut être influencé par des changements de la pluie entre t_1 et t_2. Cette influence est d'autant plus grande, que les changements sont hétérogènes dans la cellule.

V.1.1.2 L'extrapolation de l'advection des cellules de pluie

Nous cherchons à estimer l'advection des cellules de pluie dans l'intervalle de prévision $(t_0, t_k = t_0+\Delta_P t)$ par l'extrapolation de l'advection observée dans l'intervalle $(t_{-n} = t_0-\Delta_O t\ ,\ t_0)$. Pour une cellule, qui est représentée par la séquence stricte d'échos $(e_{-n} \in E(I_{-n}),..,e_0 \in E(I_0))$, cette advection observée est déterminée par le déplacement des centres de gravité $cg(e_{-n}),..,cg(e_0)$.

Comme mentionné ci-dessus, ce déplacement est influencé par l'advection de la cellule d'une part, et par le développement de la pluie à l'intérieur de la cellule d'autre part. Afin de limiter l'influence du développement, l'extrapolation doit reposer sur une interpolation des advections sur plusieurs images. Si la durée de l'interpolation est trop courte, les variations aléatoires à l'intérieur de la cellule auront une grande influence, tandis que si la durée est trop longue il en résultera une négligence d'éventuelles changements de l'advection.

Afin de déterminer le meilleur intervalle de l'interpolation, nous avons effectué une analyse des séquences strictes d'échos de la base de données qui sont définies par les appariement manuels.

Uniquement les séquences ($e_{-n} \in E(I_{-n}),..,e_0 \in E(I_0),..,e_m \in E(I_m)$), qui remplissent les conditions suivantes, ont été utilisés:

- le centre de gravité de l'écho e_0 se trouve dans un carré de 100 km de coté, centré sur le radar,
- la durée de la séquence est d'au moins de 30 minutes, avec $t_0 - t_{-n} > 15$ min et $t_m - t_0 > 15$ min.

669 séquences ont ainsi été retenues. Pour les échos e_0 de ces séquences, nous pouvons déterminer les vecteurs individuels antérieurs, qui sont disponibles pour la prévision à l'instant t_0, ainsi que les vecteurs individuels postérieurs, qu'on cherche à estimer pour la prévision. Les vecteurs antérieurs ont été définis comme suit:

$$adv_k(e_{-n},..,e_0) = \frac{1}{\max(n,|k|)} \sum_{i=\max(-n,k)+1}^{0} (cg(e_i) - cg(e_{i-1}))$$

Nous avons pris en considération les valeurs de k égales à -1, -3, -6, -9, et -12, correspondant à des moyennes des vecteurs sur 5, 15, 30, 45, et 60 minutes. Outre les vecteurs antérieurs individuels, nous avons aussi considéré le vecteur antérieur moyen de l'image I_0 qui est défini comme

$$adv_{moy}(I_0) = \frac{1}{\sum_{e \in E(I_0)} \text{masse}(e)} \sum_{e \in E(I_0)} adv_{-12}(e) \cdot \text{masse}(e)$$

et un vecteur moyenné, défini comme

$$adv_x(e_{-n},..,e_0) = (adv_{-9}(e_{-n},..,e_0) + 2adv_{-6}(e_{-n},..,e_0) + 3adv_{-3}(e_{-n},..,e_0))/6$$

Ces vecteurs antérieurs ont été comparés au vecteur individuel postérieur:

$$adv_k(e_0,..,e_m) = \frac{1}{\max(m,k)} \sum_{i=1}^{\max(m,k)} (cg(e_i) - cg(e_{i-1}))$$

pour la valeur de k égale 12, correspondant à des moyennes des vecteurs sur 60 minutes. Les figures V.2.a et V.2.b montrent les différences en vitesse (figure V.2.a) et en direction (figure V.2.b) des différents vecteurs antérieurs et du vecteur postérieur interpolé pour 60 minutes; présentés sont les valeurs moyennes et l'écart-type des différences absolues pour les 669 cas.

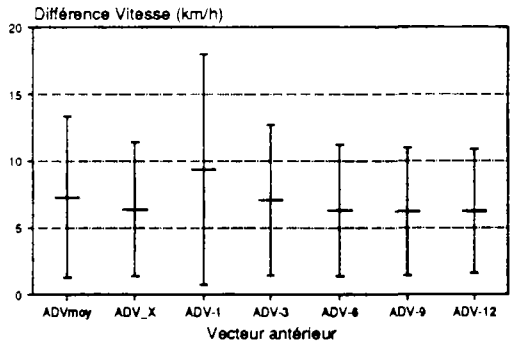


Figure V.2.a: Différence de la vitesse entre différentes vecteurs antérieurs et le vecteur postérieur de 60 minutes $adv_{12}(.)$

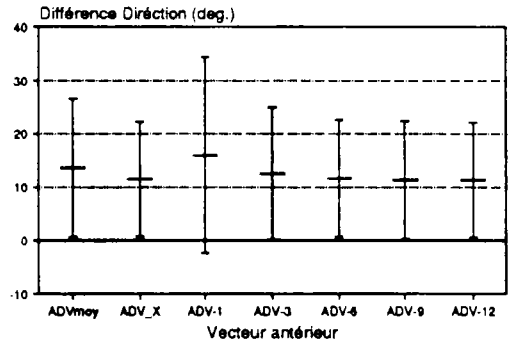


Figure V.2.b: Différence de la direction entre différentes vecteurs antérieurs et le vecteur postérieur de 60 minutes $adv_{12}(.)$

Il résulte de cette analyse, que la différence entre les vecteurs individuels antérieurs et les vecteurs individuels postérieurs est, pour des durées d'interpolation plus longues que 15 minutes, inférieure à celle entre le vecteur moyen antérieur et les vecteurs individuels postérieurs. Le maximum de corrélation des vecteurs individuels est atteint pour des intervalles d'interpolation de 45 et 60 minutes. Ces derniers vecteurs correspondent aussi mieux au vecteurs postérieurs que le vecteur moyenné $adv_x(.)$.

Ce résultat montre l'intérêt d'observer l'advection individuelle des cellules pour l'estimation de leur déplacement dans le futur proche. La corrélation relativement faible, qui a été découverte entre les vecteurs individuels postérieurs et les vecteurs individuels antérieurs d'une durées d'interpolation inférieure à 15 minutes, provient de l'influence du développement non-systématique de la pluie à l'intérieur des cellules sur le déplacement des centres de gravité. Cette influence est réduite par le calcul des moyennes des vecteurs sur plusieurs images. Par conséquent, les vecteurs utilisés par PROPHETIA comme prévision de l'advection des cellules seront définis de la façon suivante:

- pour un écho $e \in E(I_0)$, qui fait partie d'une séquence dont la durée d'observation est plus longue que 15 minutes, le vecteur de déplacement prévu est égal au vecteur de l'advection observée, moyenné sur au plus 60 minutes ($adv_{prévue}(e)=adv_{.12}(e)$),
- pour un écho $e \in E(I_0)$, qui fait partie d'une séquence dont la durée d'observation est plus courte que 15 minutes, le vecteur de déplacement prévu est égal au vecteur moyen de l'advection de l'image I_0 ($adv_{prévue}(e)=adv_{moy}(I_0)$).

V.1.2 La prévision des lames d'eau

L'algorithme V.1 décrit la méthode de PROPHETIA pour prévoir les lames d'eau de l'intervalle $(t_0,t_0+\Delta_P t)$ pour un bassin B, qui est identifié par la liste de pixels $(p_1,...,p_n)$ couvrant la surface du bassin. Dans cet algorithme, un intervalle de calcul d'une minute a été adapté, afin de réduire les erreurs dues à une advection rapide des cellules de pluie.

La longueur $\Delta_P t$ de l'intervalle d'échéance de la prévision est limitée par la taille de l'image radar et par l'emplacement du bassin sur l'image par rapport à la vitesse et la direction de l'advection de la pluie. Ainsi, une prévision pour un bassin en région parisienne basée sur l'image radar de Trappes est limitée à un intervalle d'échéance $\Delta_P t$ de deux heures environ pour une pluie se déplaçant avec 60 km/h en direction de l'est, et à un intervalle d'une heure environ pour une pluie de la même vitesse se déplaçant vers l'ouest. Toutefois, la plupart des pluies de cette région se déplace en direction de l'est; pour toutes les pluies étudiées, un intervalle de prévision d'une heure est ainsi possible.

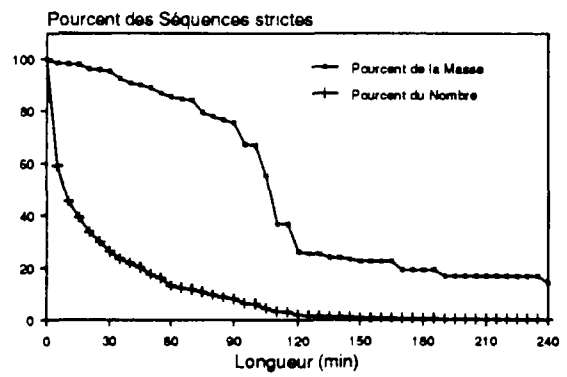


Figure V.3: Durée de vie des séquences strictes de la base des données

Une autre limite est posée à l'intervalle d'échéance par le cycle de vie des cellules. La prévision par extrapolation des observations est naturellement contrainte par la durée de vie des cellules observées. La figure V.3 présente une analyse de la longueur des séquences strictes de la base de données. Sont uniquement considérées les séquences, dont au moins un écho est situé à une distance inférieure de 50 km du radar, afin de ne tenir compte que des cellules, qui ont une

importance pour la prévision des lames d'eau en région parisienne. Sont présentées le pourcentage du nombre et de la masse pluvieuse totale des séquences dépassant une certaine durée de vie.

De toutes les séquences, seulement 8% atteignent une durée de vie supérieure à 90 minutes. Ces séquences représentent 75% de la masse pluvieuse. En principe, un intervalle d'échéance d'une heure serait alors justifié pour la région étudiée. Il ne peut cependant pas être exclu, que pour certaines pluies l'intervalle valide soit plus petit, ce qui s'exprimera par un taux d'erreur de la prévision particulièrement élevé.

Donné : Une image radar I_0 et la liste $(p_1,\ldots,p_n)$ des pixels couvrant le bassin B.

Cherché : La prévision $L_B(t_0,t_0+\Delta_P t)$ des lames d'eau attendu pour le bassin B dans l'intervalle $(t_0,t_0+\Delta_P t)$.

Algorithme:

(0) (a) Déterminer le vecteur d'extrapolation $adv(e_i)$ pour chaque $e_i \in E(I_0)$:

- $adv(e)=adv_{12}(e)$ si e a été observé pendant plus de 15 minutes,
- $adv(e)=adv_{moy}(I_0)$ sinon.

(b) $L_B := 0,\ k := 0$.

(1) Prévoir l'image I_0+k minutes:

(a) Déplacer les échos de I_0 par k minutes selon leurs vecteurs $adv(e)$.

(b) Déplacer les pixels de I_0 ne faisant pas partie d'un écho par le vecteur moyen $adv_{moy}(I_0)$.

(2) $L_B := L_B + \sum_{i=1}^{n} (\text{intensité}((I_0+k)(p_i))/60)$

(3) $k := k+1$

(4) Si $k < \Delta_P t$ continuer avec (1)

Algorithme V.1: Algorithme de prévision des lames d'eau de PROPHETIA

V.2 L'évaluation des prévisions

V.2.1 Critères de l'évaluation de la prévision par radar

L'évaluation d'une prévision, qui est fournie à l'instant t_0 pour l'intervalle d'échéance $(t_0,t_0+\Delta_p t)$, consiste en une comparaison des grandeurs prévues avec des observations faites entre t_0 et $t_0+\Delta_p t$. Les moyens, qui peuvent être employés pour l'observation, sont multiples; par exemple le radar, le pluviomètre, les mesures des débits et de crues, et l'observation par l'homme.

Les critères utilisés pour l'évaluation varient selon le contexte de l'application de la prévision. De la même façon, le taux de tolérance des erreurs dépend de la manière dont la prévision est utilisée. Par exemple, une prévision de la pluie, qui est très utile comme visualisation de la situation météorologique sur une grande zone, peut être inutile comme prévision des débits d'un bassin versant précis, et vice versa.

Denoeux (1989) a fait un résumé des méthodes d'évaluation de la prévision de pluie par radar. Il distingue entre trois types de critères:

- des critères basés sur une comparaison des vecteurs de l'advection prévus et mesurés,
- des critères basés sur une comparaison de la carte de pluie prévue et mesurée,
- des critères basés sur une comparaison des hyétogrammes prévus et mesurés.

Tandis que les deux premiers types de critères présentent surtout un intérêt dans des études atmosphériques, le troisième type est particulièrement intéressant pour les applications en hydrologie. Denoeux constate cependant qu'aucun critère critiquant la qualité de la prévision en fonction de son influence sur le coût de la gestion optimisée des réseaux d'assainissement n'a été développé.

Dans la même étude on montre que l'évaluation d'une prévision avec des critères différents donne souvent des résultats contradictoires, même au sein des critères d'une seule catégorie. Le choix d'un critère adapté précisément à l'utilisation de la prévision est alors indispensable afin de pouvoir évaluer la qualité relative des prévisions effectuées.

Denoeux propose un critère baptisé NMP (Nombre des Mauvaises Prévisions), qui est adapté à l'utilisation des prévisions en hydrologie urbaine. Il est basé sur une recherche de l'influence des erreurs de la prévision sur la qualité de la gestion d'un réseau d'assainissement exemplaire. Pour une prévision à l'instant t_0 des lames d'eau pour l'intervalle $(t_0,t_0+\Delta_p t)$, avec $\Delta_p t$=30 minutes ou $\Delta_p t$=60 minutes, le critère NMP est défini comme suit:

(1) Dans un cercle d'un diamètre de la moitié de la taille de l'image radar, qui est centré sur le radar, on choisit au hasard n pixels $p_1,..,p_n$ dont la lame mesurée par radar dans l'intervalle $(t_0,t_0+\Delta_p t)$ est supérieure à un seuil s.

(2) Le NMP de la prévision est le nombre de pixels, dont les lames sont surestimées de plus de 150% ou sous-estimées de plus de 50%:

$$\mathrm{NMP} := \left| \left\{ p \in \{p_1,..,p_n\} \text{ avec } \frac{HR(p)-HP(p)}{HR(p)} > 0.5 \text{ ou } \frac{HR(p)-HP(p)}{HR(p)} < -1.5 \right\} \right|$$

où $HR(.)$ et $HP(.)$ sont les lames d'eau mesurées et prévues pour la durée $(t_0,t_0+\Delta_p t)$.

Denoeux propose un nombre de pixels n=10 et un seuil s=1mm. Il détermine les valeurs de référence $HR(.)$ par les mesures par radar. Les caractéristiques du critère NMP sont les suivants:

- Il est basé sur le calcul des erreurs sur les lames d'eau, pour les points où la lame d'eau mesurée est importante hydrologiquement.

- Il applique des seuils de tolérance de sous-estimation et de surestimation, qui ont été déterminés par l'analyse de l'influence des erreurs de la prévision sur la gestion optimisée.
- Il tient compte de l'asymétrie entre l'influence des sous-estimations et des surestimations, qui a été démontrée par l'analyse de Denoeux.

Denoeux considère une prévision comme étant bonne si la valeur de NMP est égale à 0, et de qualité moyenne si NMP est égal à 1 ou 2. Des prévisions d'une valeur supérieure à 2 sont considérées comme étant mauvaises.

V.2.2 Proposition du nouveau critère d'évaluation TMP

Bien que le critère NMP soit développé en vue des applications en hydrologie urbaine, son emploi dans cette étude présente plusieurs inconvénients:

- Les seuils de sous-estimation et de surestimation ont été déterminés par l'étude d'un seul réseau exemplaire. Il n'est pas exclu, que ces seuils soient différents pour d'autres réseaux ou d'autres types de gestion.
- Le choix au hasard de 10 pixels d'évaluation permet une évaluation statistique de la qualité des prévisions. Une comparaison de deux prévisions précises, qui ont été effectuées à partir des mêmes données par des techniques différentes, est cependant difficile, car le nombre des points est petit et les points d'évaluation sont différents pour chaque évaluation.
- Les données utilisées par Denoeux ont une résolution spatiale de 1.6 km, un point couvrant ainsi 2.56 km^2. Les données utilisées dans cette étude ont une résolution de 0.8 km avec la surface d'un point égale à 0.64 km^2. De ce fait, l'évaluation par NMP pénaliserait les petits erreurs spatiales de façon plus importante.
- Le choix des points d'évaluation parmi les points d'une lame d'eau mesurée, qui est supérieur à 1 mm, ne permet pas d'évaluer correctement toutes les surestimations. Dans le cas extrême, une prévision d'une pluie forte est toujours jugée comme étant bonne, si la pluie mesurée ne dépasse pas le seuil donné.
- Le choix des points d'évaluation dans la partie centrale de l'image radar est basé sur l'hypothèse, que la qualité de la prévision ne dépend pas des caractéristiques topographiques du sol. Une telle influence ne peut cependant pas être exclue. Elle pourrait égarer l'évaluation, si la région d'intérêt représente des conditions météorologiques, qui sont différentes de son environ. Ceci est généralement le cas pour les zones urbanisées.

Pour l'application dans cette étude, nous modifions le critère NMP dans plusieurs aspects, afin de tenir compte des inconvénients mentionnés. Le nouveau critère sera baptisé ***TMP*** (Taux des **M**auvaises **P**révisions). Nous le définissons comme suit:

(1) On choisit un rectangle couvrant la zone urbanisée, pour laquelle les prévisions sont fournies. On divise le rectangle régulièrement en n bassins $BV = \{bv_i, i=1,n\}$ de forme carrée d'une surface de t km^2. Une fois choisis, les n bassins restent inchangés pour toute l'étude.

(2) Pour l'évaluation d'une prévision effectuée à t_0 pour l'intervalle $(t_0, t_0+\Delta_p t)$ avec $\Delta_p t \leq 120$ minutes, on considère le sous-ensemble BV' de BV, qui est formé par les bassins, dont la lame d'eau prévue ou la lame d'eau mesurée entre t_0 et $t_0+\Delta_p t$ dépasse un seuil d'intensité s:

$$BV' = \{ bv \in BV \text{ avec } HP(bv) > \frac{s}{\Delta_p t} \text{ ou } HR(bv) > \frac{s}{\Delta_p t} \}$$

où $HR(.)$ et $HP(.)$ sont, comme avant, les lames d'eau mesurées et prévues pour la durée $(t_0, t_0+\Delta_p t)$. La valeur de TMP est déterminée comme suit:

$$\text{TMP} := \frac{1}{|BV'|} \cdot \left| \left\{ bv \in BV' \quad \text{avec} \quad \frac{HR(bv)-HP(bv)}{HR(bv)} > 0.5 \quad \text{ou} \quad \frac{HR(bv)-HP(bv)}{HR(bv)} < -1.5 \right\} \right|$$

La plus grande partie des inconvénients du critère NMP a été ainsi écartée. Faute d'études plus générales, une modification des seuils de tolérance de la sous-/surestimation n'a pas été envisagée. Un carrée de 48 km de coté (60 pixels), couvrant la région parisienne, a été choisi comme zone d'intérêt, dans laquelle les bassins versants idéalisés sont situés.

L'influence de la taille t des bassins et du seuil s de l'intensité sur l'évaluation a été analysée de la façon suivante: Pour les 20 pluies de la base des données, des prévisions pour un temps d'échéance de 60 minutes ont été effectuées par PROPHETIA basées sur les appariements manuels des échos. La figure V.4.a montre le TMP moyen en fonction de la taille t des bassins de l'évaluation, avec un seuil s de l'intensité égal à 1.0 mm/h. Comme attendu, le taux d'erreur est décroissant, si la taille des bassins augmente. Ceci est dû au plus grand lissage des lames d'eau dans les grands bassins. Le seuil de l'évaluation n'a cependant pas une grande influence sur l'évaluation des prévision, comme le montre la figure V.4.b. Dans cette figure, le TMP moyen pour une taille des bassins égale à 2.56 km^2 est indiqué en fonction du seuil s de l'intensité. Bien que le nombre des bassins touchés par une pluie ou une prévision supérieure au seuil s est fortement décroissant si le seuil devient plus grand, le taux de prévisions incorrectes reste presque constant.

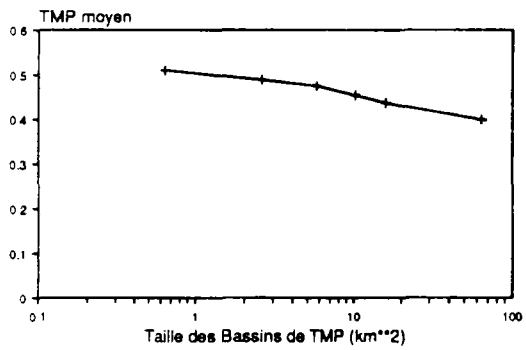


Figure V.4.a: Influence de la taille des bassins de l'évaluation sur le taux d'erreur (TMP moyen de 20 pluies)

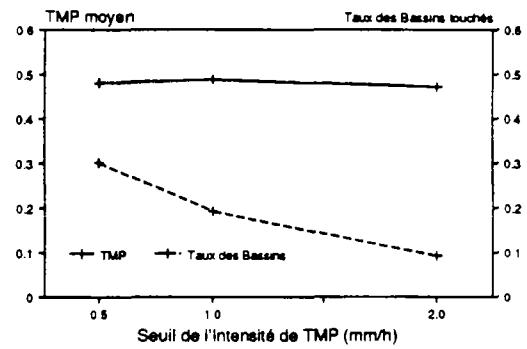


Figure V.4.b: Influence du seuil de l'intensité de l'évaluation sur le taux d'erreur (TMP moyen de 20 pluies)

Pour cette étude, nous avons choisi les valeurs suivantes:

- la taille t des bassins a été fixée à 2.56 km^2 (2 sur 2 pixels), afin d'évaluer l'exactitude spatiale de la prévision à une échelle adaptée aux besoins en hydrologie urbaine,
- le seuil s a été choisi égal à 1 mm/h, afin de ne tenir compte que des pluies provocant (ou des prévisions indiquant) un ruissellement assez important pour la gestion des réseaux d'assainissement.

Le critère TMP sera appliqué pour toutes les évaluations dans cette étude. La lame d'eau de référence $HR(.)$ sera calculée à partir des images radar dans l'intervalle d'échéance de la prévision. Seulement les prévisions, dont les données de cet intervalle sont disponibles, seront effectués. Pour les 20 pluies de la base de données, le nombre de telles prévisions est de 831 pour un temps d'échéance $\Delta_p t = 30$ min et de 711 pour $\Delta_p t = 60$ min.

Pour exprimer l'erreur moyenne des prévisions pour un événement de pluie ev, pour lequel un nombre n de prévisions p_i est effectuée, le taux suivant est calculé:

$$\mathrm{TMP}(ev) = \frac{\sum_{i=1}^{n} N^{-}(p_i)}{\sum_{i=1}^{n} N^{-}(p_i) + \sum_{i=1}^{n} N^{+}(p_i)}$$

où

- $N^{+}(p_i)$ est le nombre des bassins, dont la pluie mesurée ou la pluie prévue entre $t_0(p_i)$ et $t_0(p_i)+\Delta_p t$ dépasse le seuil de l'intensité de TMP, et dont la différence entre la lame prévue et la lame mesurée est comprise dans les seuils de tolérance de TMP;
- $N^{-}(p_i)$ est le nombre des bassins, dont la pluie mesurée ou la pluie prévue entre $t_0(p_i)$ et $t_0(p_i)+\Delta_p t$ dépasse le seuil de l'intensité de TMP, et dont la différence entre la lame prévue et la lame mesurée dépasse les seuils de tolérance de TMP.

Afin de pouvoir comparer la performance globale de différentes méthodes, nous présenterons des "courbes d'efficacité". Un point (x,y) d'une telle courbe indique le nombre x des prévisions effectuées, dont le TMP était inférieur à la valeur ordonnée y. Ce type de présentation ne permet alors pas de comparer les résultats des différentes méthodes pour une même prévision, mais il présente mieux les différences pour l'ensemble des prévisions.

V.3 La prévision de pluie par le système PROPHETIA

V.3.1 Résultats des prévisions avec différentes règles de l'appariement

Le système de prévision PROPHETIA a été appliqué aux pluies de la base de données. Trois méthodes de l'appariement ont été comparés, afin d'étudier l'influence du taux d'erreur de l'appariement sur le taux d'erreur de la prévision:

- l'appariement manuel,
- l'appariement par l'algorithme III.7 avec l'arbre de décision très simple ADD_{INI},
- l'appariement par l'algorithme III.7 avec l'arbre de décision généré par l'apprentissage ADD_{APP}.

La figure V.5 présente les courbes d'efficacité des trois méthodes pour les 20 pluies étudiées et la prévision de 60 minutes. Dans la figure V.6 est comparé le TMP moyen par événement de la prévision de 60 minutes, comme il résulte des appariements manuels, aux TMP, comme il résulte des appariements par les deux arbres de décision.

Les courbes d'efficacité montrent très peu de différences entre la qualité de la prévision avec les appariements manuels et la qualité de la prévision avec les appariements par l'arbre de décision ADD_{APP}. Comme le montre la figure V.6, seul pour la pluie du 7.3.1989 une différence significative peut être observée. Cette différence est provoquée par les mauvais appariements au début de l'événement, dont les raisons ont déjà été évoquées au chapitre précédent.

Si la règle triviale ADD_{INI} est appliquée pour les appariements, le taux d'erreur est entre 10% et 40% plus élevé qu'avec les appariements par ADD_{APP}. La qualité de la prévision est alors sensiblement diminuée, si le taux d'appariements incorrects est trop élevé.

La méthode de l'apprentissage automatique, qui a été appliquée, a mené à la génération d'une règle de l'appariement, qui permet une prévision de pratiquement la même qualité que l'appariement subjectivement correct. Dans le contexte de l'application dans PROPHETIA, la règle ADD_{APP} peut alors être considérée comme étant optimale.

Ce résultat montre aussi, que la définition automatique des échos imaginaires est suffisant pour déterminer l'advection des cellules.

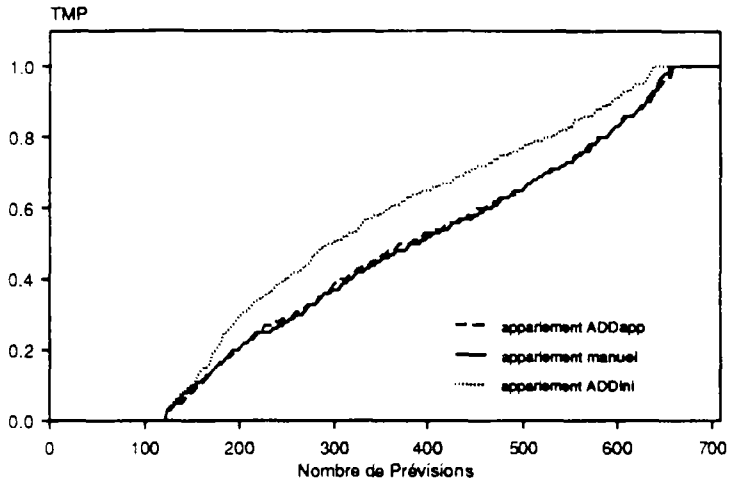


Figure V.5: Courbes d'efficacité de la prévision avec différentes méthodes d'appariement (prévisions de 60 minutes)

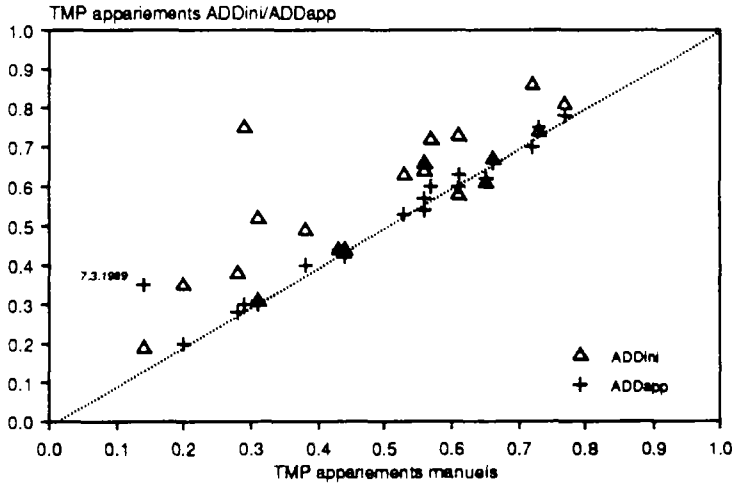


Figure V.6: Taux d'erreur de la prévision avec les appariements manuels comparé aux taux de la prévision avec les deux arbres de décision (prévisions de 60 min)

V.3.2 Analyse des sources d'erreurs de la prévision

Le taux d'erreur très élevé des prévisions pour les intervalles d'échéance supérieurs à une heure est partiellement expliqué par le fait, que la taille limitée de l'image radar et la durée de vie des cellules ne permettent parfois pas une durée d'échéance supérieure à une heure. Afin d'écarter cette source d'erreur, nous travaillerons par la suite de cette étude uniquement avec des prévisions d'un temps d'échéance de 30 minutes et de 60 minutes. Pour ces durées, nous analyserons par la suite les sources d'erreurs de la prévision automatique par PROPHETIA.

Le critère d'évaluation TMP a été introduit dans le paragraphe précédent. Les figures sur les prochaines pages montrent le taux d'erreur de chaque prévision effectuée. Pour une prévision p_i, qui est fournie à l'instant t_0 pour la durée $(t_0(p_i),t_0(p_i)+\Delta_P t)$, les figures sont à interpréter comme suit:

Si BV est l'ensemble des bassins idéalisés, qui ont été choisis pour l'évaluation, et $BV' \subset BV$ le sous-ensemble des bassins, dont la pluie mesurée ou la pluie prévue entre $t_0(p_i)$ et $t_0(p_i)+\Delta_P t$ dépasse le seuil de l'intensité de TMP, les figures présentent:

- le nombre $N^+(p_i)$ des bassins $bv \in BV'$, dont la différence entre la lame prévue et la lame mesurée entre $t_0(p_i)$ et $t_0(p_i)+\Delta_P t$ est comprise dans les seuils de tolérance de TMP (en blanc);
- le nombre $N^-(p_i)$ de bassins $bv \in BV'$, dont la différence entre la lame prévue et la lame mesurée entre $t_0(p_i)$ et $t_0(p_i)+\Delta_P t$ n'est pas comprise dans les seuils de tolérance de TMP (en noir).

$$\text{d'où} \quad \mathrm{TMP}(p_i) = \frac{N^-(p_i)}{N^-(p_i) + N^+(p_i)}.$$

Les périodes au début et à la fin des événements, pendant lesquelles aucun bassin n'était concerné ($BV'=\emptyset$), ne sont pas représentées pour toutes les pluies. Rappelons que le nombre des prévisions effectuées est moins élevé pour le temps d'échéance de 60 minutes, afin de rendre possible l'évaluation de toutes les prévisions. Remarquons aussi que les données n'étaient pas toujours disponibles pour toute la durée des périodes pluvieuses; ceci est notamment le cas pour la pluie du 2.6.1989.

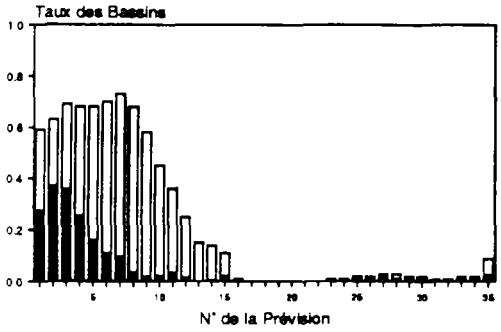


Fig. V.7.1(a): Pluie du 7.3.89 (prév. de 30 min)

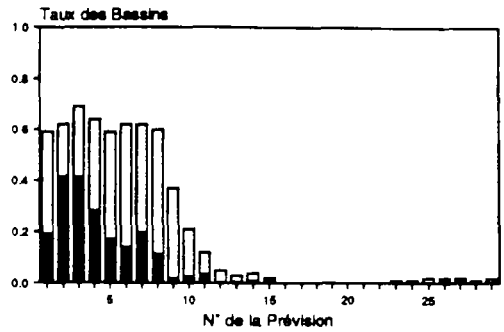


Fig. V.7.1(b): Pluie du 7.3.89 (prév. de 60 min)

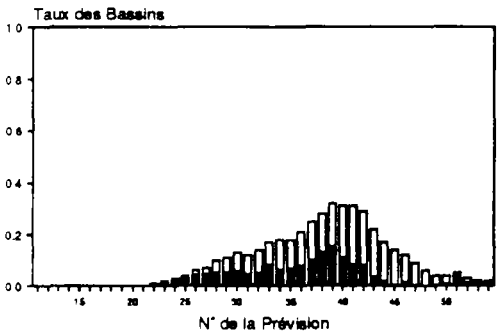


Fig. V.7.2(a): Pluie du 4.4.89 (prév. de 30 min)

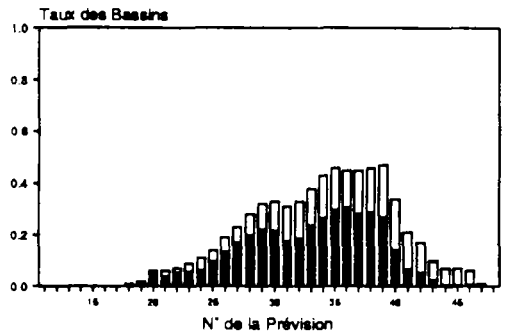


Fig. V.7.2(b): Pluie du 4.4.89 (prév. de 60 min)

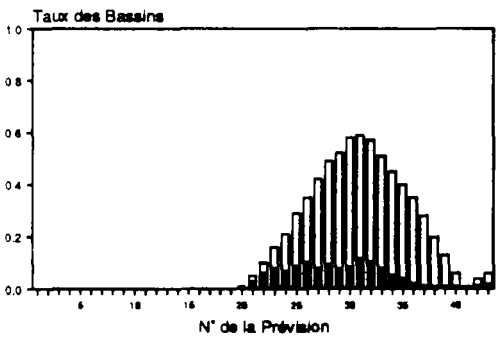


Fig. V.7.3(a): Pluie du 24.4.89 (prév. de 30 min)

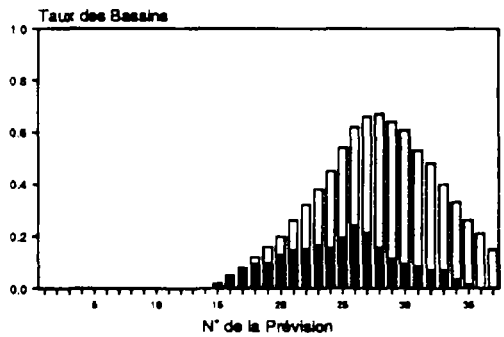


Fig. V.7.3(b): Pluie du 24.4.89 (prév. de 60 min)

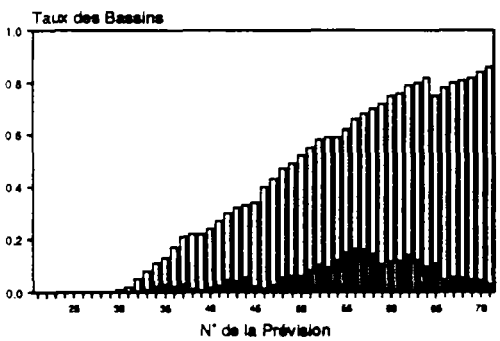


Fig. V.7.4(a): Pluie du 27.4.89 (prév. de 30 min)

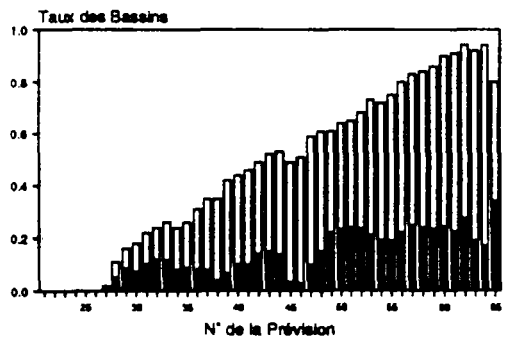


Fig. V.7.4(b): Pluie du 27.4.89 (prév. de 60 min)

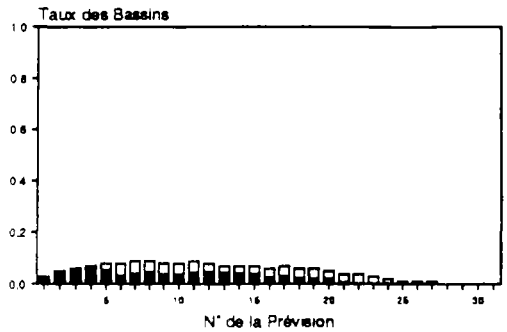


Fig. V.7.5(a): Pluie du 10.5.89 (prév. de 30 min)

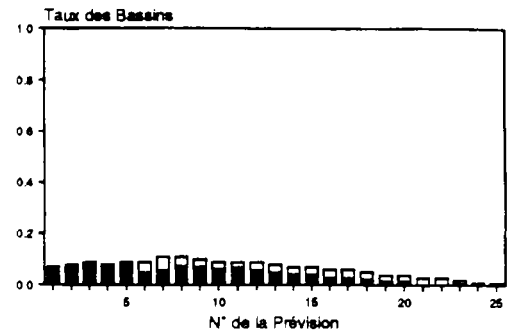


Fig. V.7.5(b): Pluie du 10.5.89 (prév. de 60 min)

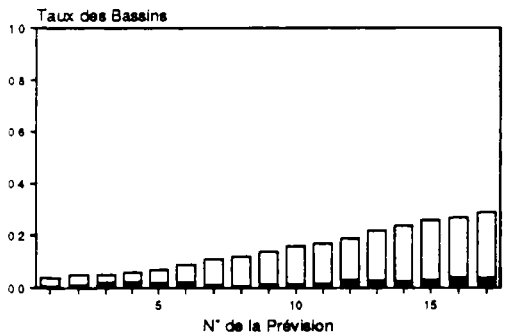


Fig. V.7.6(a): Pluie du 2.6.89 (prév. de 30 min)

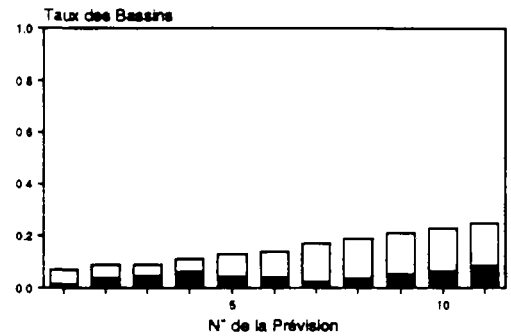


Fig. V.7.6(b): Pluie du 2.6.89 (prév. de 60 min)

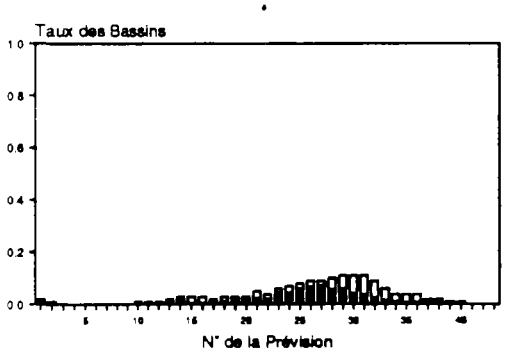


Fig. V.7.7(a): Pluie du 3.6.89 (prév. de 30 min)

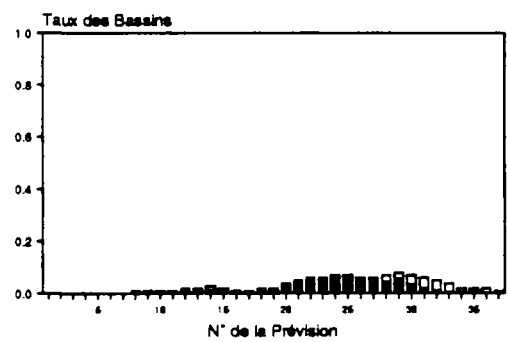


Fig. V.7.7(b): Pluie du 3.6.89 (prév. de 60 min)

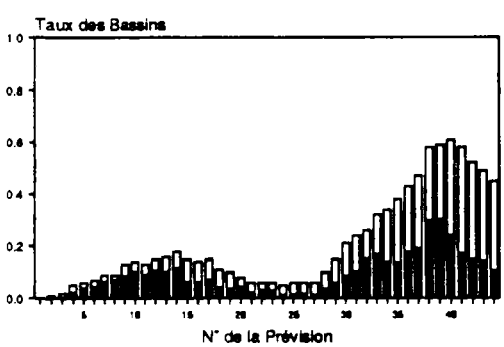


Fig. V.7.8(a): Pluie du 6.6.89 (prév. de 30 min)

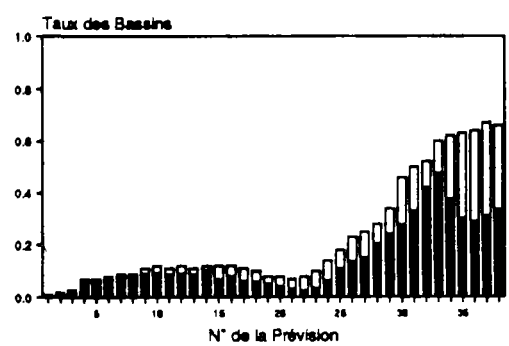


Fig. V.7.8(b): Pluie du 6.6.89 (prév. de 60 min)

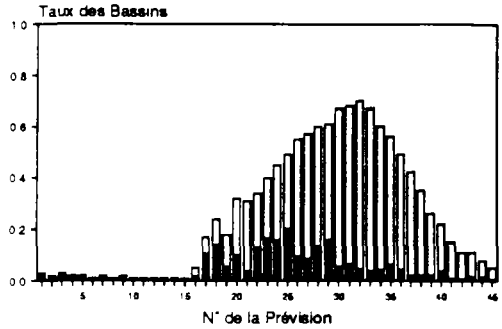


Fig. V.7.9(a): Pluie du 27.6.89 (prév. de 30 min)

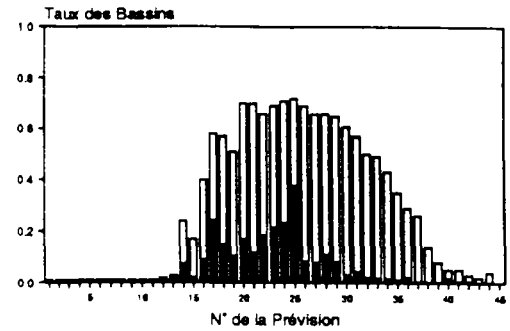


Fig. V.7.9(b): Pluie du 27.6.89 (prév. de 60 min)

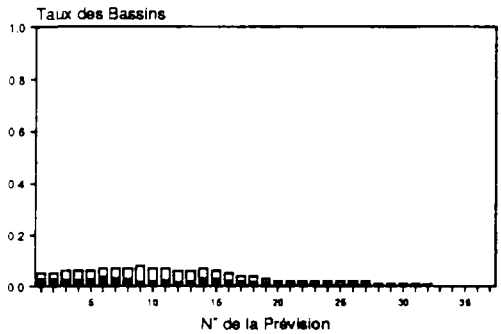


Fig. V.7.10(a): Pluie du 10.7.89 (prév. de 30 min)

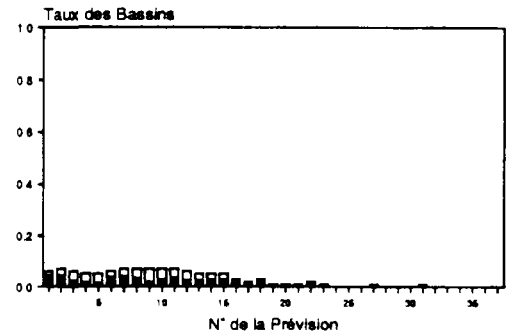


Fig. V.7.10(b): Pluie du 10.7.89 (prév. de 60 min)

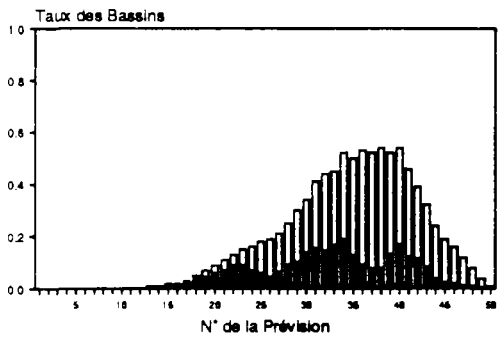


Fig. V.7.11(a): Pluie du 7.8.89 (prév. de 30 min)

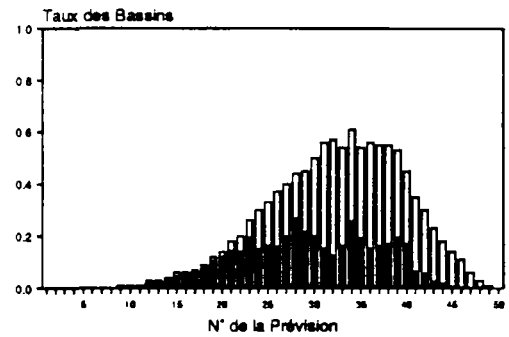


Fig. V.7.11(b): Pluie du 7.8.89 (prév. de 60 min)

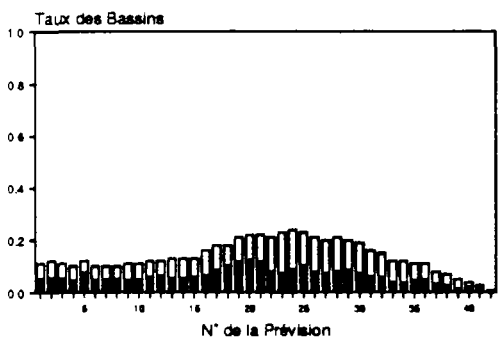


Fig. V.7.12(a): Pluie du 12.9.89 (prév. de 30 min)

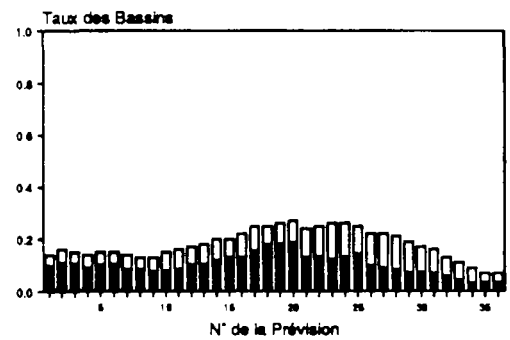


Fig. V.7.12(b): Pluie du 12.9.89 (prév. de 60 min)

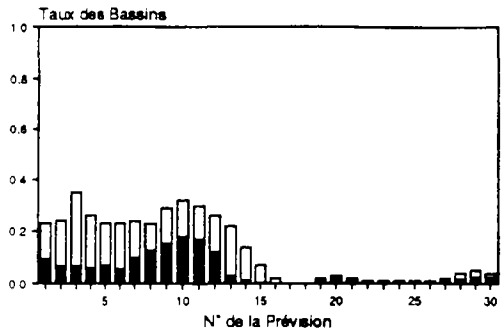


Fig. V.7.13(a): Pluie du 19.9.89 (prév. de 30 min)

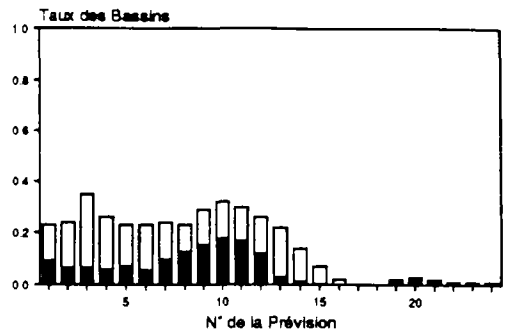


Fig. V.7.13(b): Pluie du 19.9.89 (prév. de 60 min)

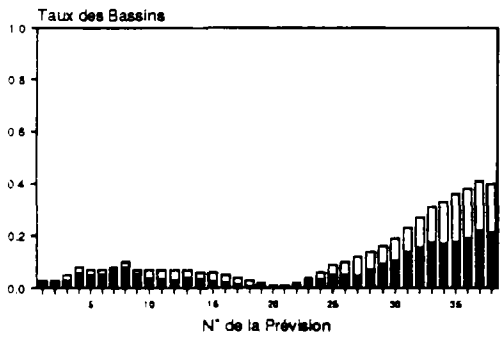


Fig. V.7.14(a): Pluie du 23.4.90 (prév. de 30 min)

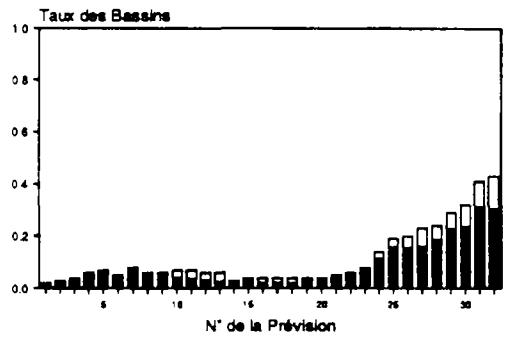


Fig. V.7.14(b): Pluie du 23.4.90 (prév. de 60 min)

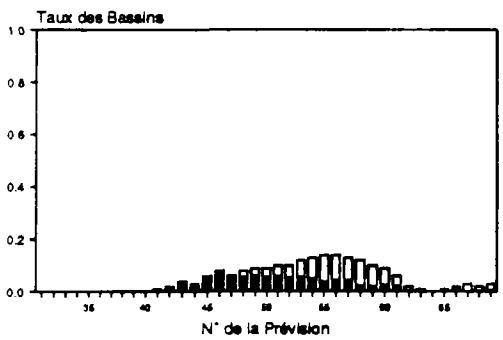


Fig. V.7.15(a) : Pluie du 14.5.90 (30 min)

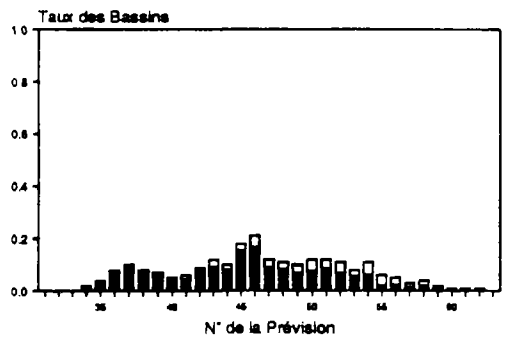


Fig. V.7.15(b) : Pluie du 14.5.90 (60 min)

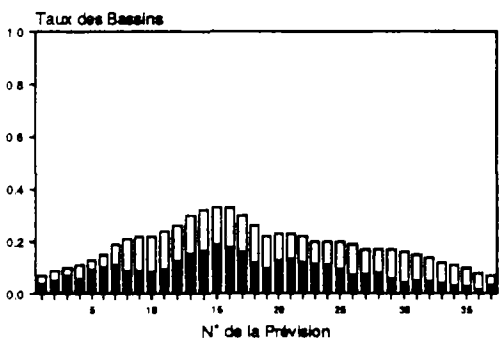


Fig. V.7.16(a): Pluie du 9.6.90 (prév. de 30 min)

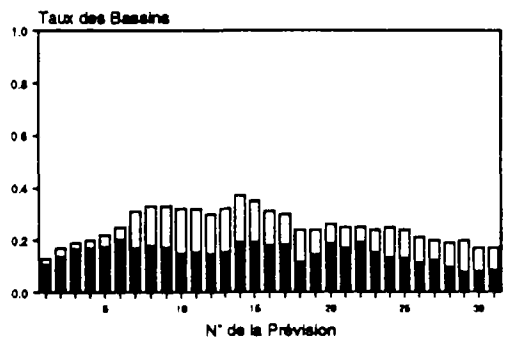


Fig. V.7.16(b): Pluie du 9.6.90 (prév. de 60 min)

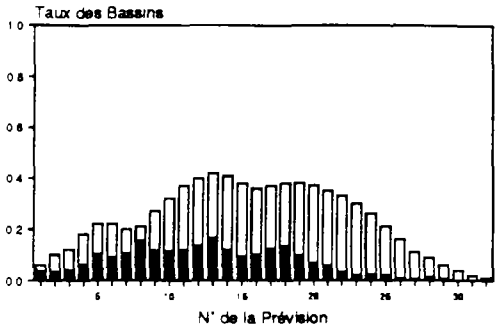


Fig. V.7.17(a): Pluie du 26.6.90 (prév. de 30 min)

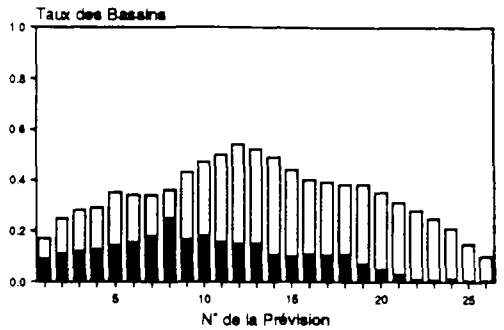


Fig. V.7.17(b): Pluie du 26.6.90 (prév. de 60 min)

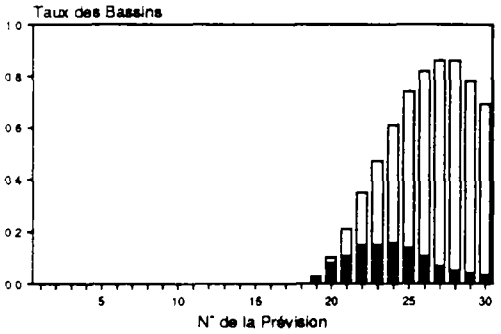


Fig. V.7.18(a): Pluie du 21.9.90 (prév. de 30 min)

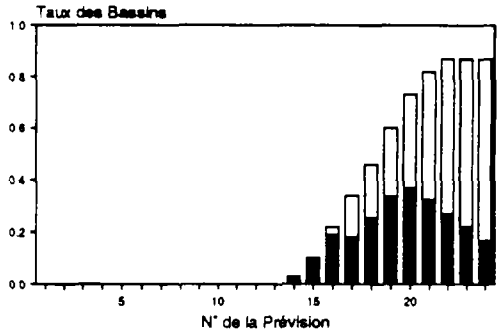


Fig. V.7.18(b): Pluie du 21.9.90 (prév. de 60 min)

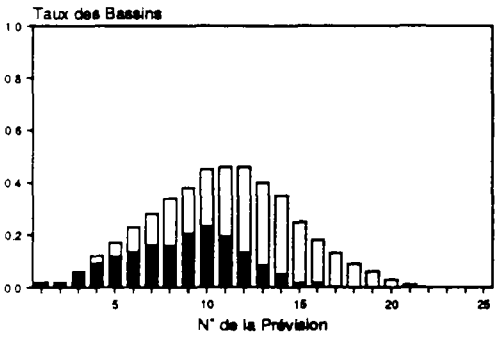


Fig. V.7.19(a): Pluie du 24.9.90 (prév. de 30 min)

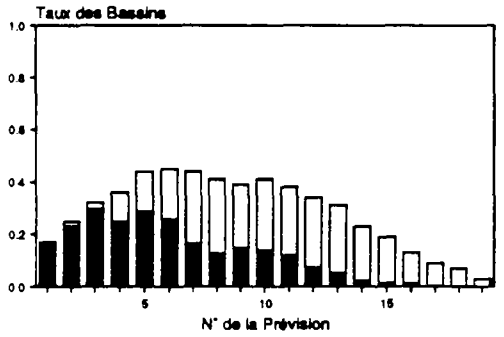


Fig. V.7.19(b): Pluie du 24.9.90 (prév. de 60 min)

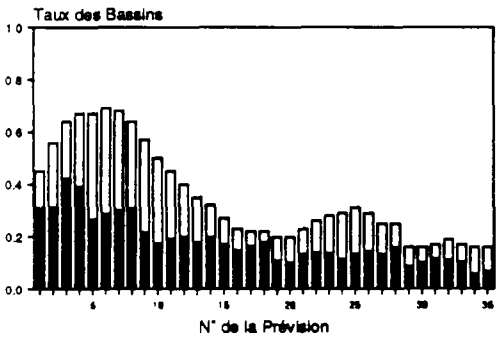


Fig. V.7.20(a): Pluie du 30.9.90 (prév. de 30 min)

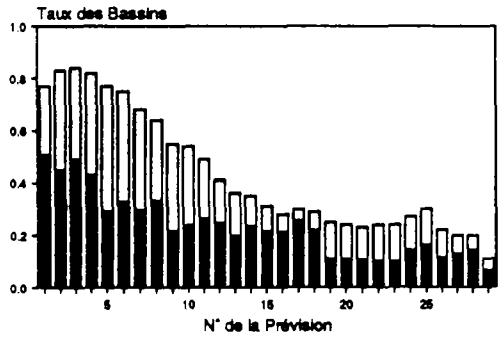


Fig. V.7.20(b): Pluie du 30.9.90 (prév. de 60 min)

Les erreurs dues à la mesure de la pluie étant écartées par le choix des données et par la méthode de l'évaluation, les erreurs observées pour les pluies étudiées sont de deux types. Premièrement, il s'agit de différences entre les lames d'eau prévues et mesurées, qui sont dues à une prévision incorrecte de l'advection des cellules; deuxièmement, ces différences sont dues à une croissance ou une décroissance de la pluie pendant l'intervalle de la prévision.

Concernant le premier type d'erreur, il est provoqué ou par une mauvaise définition des séquences d'échos, ou par une description insuffisante de l'advection observée. Des erreurs, qui sont de façon prépondérante dues à un changement de l'advection des cellules dans l'intervalle de la prévision, n'ont pas été observées. Il semble que de tels changements sont des processus suffisamment lents pour ne pas affecter les prévisions d'une échéance courte, comme celles considérées dans cette étude.

Date de la pluie	Erreurs provoquées par la prévision incorrecte de l'advection de la pluie		Erreurs provoquées par le développement de la pluie	
	faux appariements	advection mal décrit	croiss./décroiss. locale	croiss./décroiss. globale
7.3.89	(1-5) –			(6-fin) +
4.4.89			+	+
24.4.89			–	
27.4.89		(25-50) –		+/–
10.5.89			+/–	
2.6.89		–		
3.6.89			+/–	
6.6.89			–	
27.6.89		–	–	
10.7.89			+/–	
7.8.89			+/–	
12.9.89			+/–	
19.9.89				–
23.4.90			–	
14.5.90				+
9.6.90			+/–	
26.6.90			+/–	
21.9.90		(15-17) –		(18-fin) –
24.9.90			–	
30.9.90			+/–	

Tableau V.1: Analyse des sources principales des erreurs de prévision, provoquant des surestimations (+) et des sous-estimations (–) des lames d'eau

Le développement de la pluie peut être classé en développements à petite échelle, qui sont influencés par la convection locale, et développements à grande échelle, qui sont provoqués par les flux des masses d'air autour du système frontal. Les variations non systématiques de l'intensité de la pluie à l'intérieur des cellules constituent une troisième classe, qui affecte toutes les prévisions. Car l'influence de ces variations aléatoires sur le taux d'erreur est cependant difficile à déterminer, les erreurs dues à cet effet ne seront pas considérées dans l'analyse suivante.

Pour les pluies étudiées, le tableau V.1 montre une estimation des sources principales des erreurs de prévision. Par la suite, nous analyserons les différents types d'erreurs.

V.3.2.1 Erreurs dues à un appariement incorrect des échos

Les seules prévisions, qui sont affectées par des appariements incorrects, sont les prévisions n° 1-5 de la pluie du 7.3.1989. Les raisons qui, pour cette pluie, mènent à un dysfonctionnement de l'algorithme de l'appariement, ont déjà été examinées (cf. § IV.3.1). Car les images concernées sont les premières de l'événement, le vecteur moyen de la prévision est erroné à cause des mauvais appariements, qui ont une influence d'autant plus importante, que le taux d'échos non reconnus est élevé pour cette période. Par cet effet, une zone pluvieuse située au sud de la région parisienne, qui se déplace en direction nord-est, est prévue pour se déplacer vers l'est (zone marquée "*" sur la figure A.3.1). Il en résulte une sous-estimation de la pluie pour la zone urbanisée.

V.3.2.2 Erreurs dues à une mauvaise description de l'advection des cellules

Dans certains cas, l'advection des cellules est mal décrite par le déplacement des centres de gravité des échos. La source principale de ce problème est le seuil fixe d'intensité appliqué pour la définition des échos simples.

Pour les pluies frontales, le seuil fixe ne permet pas de détecter les déplacements individuels de cellules intenses, qui sont imbriquées dans des champs larges de pluie faible. Ceci est le cas pour les pluies du 27.4.1989 (prévisions n° 25-50) et du 27.6.1989 (prévisions n° 17-25) (zones marquées "*" sur les figures A.3.4 et A.3.9). Pour ces deux pluies, il en résulte une sous-estimation des lames d'eau, car la traversée des cellules intenses sur la région urbanisée n'est pas prévue correctement. La définition des échos simples à un seuil plus élevé devrait améliorer ces prévisions (cf. § V.3.4).

La pluie du 2.6.1989 est caractérisée par une grande zone de pluie faible, qui tourne autour de son centre de gravité. L'advection détectée par PROPHETIA est alors très faible, tandis qu'elle est en réalité d'une vitesse d'environ 20 km/h dans les parties extérieures de la zone. La méconnaissance de cette advection provoque une sous-estimation de la pluie en région parisienne. Contrairement aux pluies du 27.4.1989 et du 27.6.1989, la montée du seuil de définition des échos simples ne devrait pas permettre une amélioration des prévisions de cette pluie, car les zones provoquant l'erreur sont des zones de pluie faible, qui ne seraient pas repérées comme échos à un seuil supérieur (cf. § V.3.4).

La détermination de l'advection des cellules à l'aide du déplacement des centres de gravité des échos pose aussi un problème lorsque la cellule traverse le bord de l'image radar. Pendant ce passage, le déplacement détecté ne correspond pas au déplacement réel, ce dernier étant généralement d'une vitesse plus grande. Aussi la direction de l'advection, qui est déterminée par PROPHETIA, peut être erronée dans de telles situations. Pour la pluie du 27.6.1989, toutes les prévisions sont affectées par cet effet.

La pluie du 21.9.1990 finalement est marquée par une cellule d'une surface d'environ 2000 km^2 (marquée "*" sur la figure A.3.18), qui s'approche de la région parisienne en direction sud-est avec une vitesse d'environ 80 km/h. L'advection déterminée par PROPHETIA pour cette cellule est légèrement erronée par l'effet du bord de l'image, car, pendant l'apparition de la cellule sur l'image, son centre de gravité se déplace dans une direction différente de l'advection réelle. A cause de la forte vitesse, cet effet provoque déjà pendant ce passage une sous-estimation des lames d'eau en région parisienne. Néanmoins, pour cette pluie l'influence du développement de la pluie est beaucoup plus forte que cette erreur.

V.3.2.3 Erreurs dues à un développement local de la pluie

La plus grande partie des erreurs est due au développement de la pluie dans l'intervalle de la prévision. Contrairement aux développements globaux, les développements locaux sont des processus qui sont limités à des régions d'un ordre de grandeur de quelques dizaines de km^2. Pour toutes les pluies convectives, les prévisions sont affectées par une erreur de l'estimation des lames d'eau provoquée par la croissance et la décroissance des cellules. A un moindre degré, les pluies frontales peuvent être concernées, si des cellules de pluie convective y sont imbriquées. Ceci est le cas pour les pluies du 4.4.1989 et du 27.6.1989 (zones marquées "*" sur les figure A.3.2 et A.3.9).

Pour la pluie du 4.4.1989, l'advection d'une grande cellule détectée par PROPHETIA diffère de l'advection réelle à cause du développement local à l'intérieur de cette cellule. Une partie de l'erreur des prévisions est aussi due à ce phénomène (prévisions n° 20-30).

Compte rendu de l'importance de cette source d'erreurs, le chapitre prochain sera consacré à l'analyse des développements convectifs et à l'étude des améliorations possibles de la prévision dans ces cas.

V.3.2.4 Erreurs dues à un développement global de la pluie

Nous désignons un développement comme global, si toute la zone pluvieuse est affectée par la croissance ou la décroissance de l'intensité de pluie. Ce type d'erreur est observé surtout pour les pluies frontales; il est provoquée par une altération des caractéristiques des flux de masses d'air autour de la perturbation. Néanmoins, cette altération est généralement lente; les erreurs dues au développement global sont moins élevées que ceux dues au développement local. Parmi les prévisions étudiées, celles du 4.4.1989 et du 14.5.1990 sont affectées de façon importante par un développement global, qui provoque dans ces cas une forte sous-estimation des lames d'eau due à une décroissance rapide de l'intensité de la pluie. Pour les pluies du 19.9.1989 et du 21.9.1990 les lames d'eau sont sous-estimées à cause d'une croissance globale de la pluie.

Dans le cas d'un développement global de la pluie, une amélioration de la qualité de la prévision nécessite la connaissance des caractéristiques de flux d'air, qui se manifestent à une échelle beaucoup plus grande que celle fournie par un seul radar. Dans le cadre de cette étude, une réduction de l'importance de ce type d'erreur semble alors être exclue.

V.3.3 Comparaison de la performance de PROPHETIA avec celle d'autres méthodes de prévision

Dans la suite nous comparerons les résultats de la prévision par PROPHETIA, qui ont été analysés au paragraphe précédent, avec les prévisions par d'autres méthodes de prévision par radar. Trois méthodes seront examinées:

- une technique très simple, qui ne tient pas compte de l'advection de la pluie,
- une méthode structurée,
- une méthode non structurée.

La technique de prévision la plus simple est de considérer l'image radar mesurée à l'instant t_0 comme étant représentative pour l'intervalle de prévision $(t_0, t_0+\Delta_p t)$. La méthode, que nous baptisons ***PERSIST***, consiste simplement en un calcul des lames d'eau par multiplication des intensités, mesurées instantanément à t_0, par la durée $\Delta_p t$. Bien que la faible qualité de telles prévisions soit évidente, il nous semble intéressant de comparer le taux d'erreur par cette méthode aux erreurs de PROPHETIA, car un intérêt possible de l'application de PERSIST consiste en son extrême simplicité, qui permet la fourniture de la prévision pratiquement sans délai temporel.

La deuxième technique examinée est le système ***SCOUT II.0***, qui a été proposé par Einfalt et *al.* (1990). SCOUT est une méthode structurée de prévision, qui utilise un approche heuristique pour l'appariement des échos. Les différences entre SCOUT et PROPHETIA se trouvent notamment dans la méthode de la définition des échos, dans la méthode de l'appariement des échos et dans la définition des vecteurs de l'advection. Un ensemble d'échos de base est défini par SCOUT en appliquant un seuil d'intensité, qui est déterminé en fonction de la structure de la pluie. De ce fait, la valeur du seuil peut changer pendant un événement de pluie. Généralement, les échos définis par SCOUT sont plus petits et d'une intensité plus grande que ceux définis par PROPHETIA.

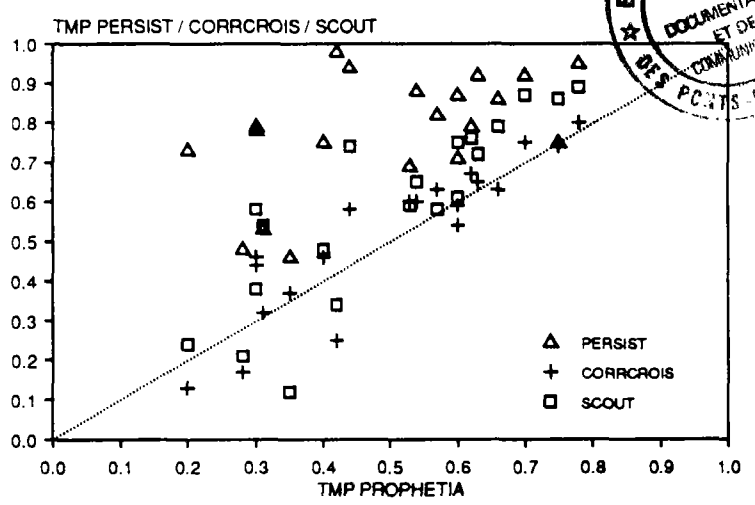


Figure V.8: Taux d'erreur de PROPHETIA comparé au taux d'erreur des autres techniques, et ligne de TMP égale ($\Delta_p t$ = 60 min)

Pour deux images successives I_1 et I_2, mesurées aux instants t_1 et t_2, l'appariement par SCOUT repose en grande partie sur l'advection moyenne observée avant t_1, afin de garantir une certaine stabilité des vecteurs. Un vecteur initial doit être défini manuellement pour un appariement correct des échos sur les premières images. Pour cette étude, le vecteur initial a été défini égal à l'advection moyenne des cellules, qui résulte des appariements manuels effectués.

SCOUT a été développé initialement pour des données radar d'une résolution spatiale de 1.6 km, et d'une résolution temporelle de 15 minutes. Récemment le système a été adaptée pour l'application opérationnelle, qui travaille avec le même type de données que cette étude (Jacquet et Neumann 1991). Cette version modifiée a été utilisée pour la comparaison dans cette étude.

Date	TMP de la Prévision de 30 min				TMP de la Prévision de 60 min			
	PER SIST	CORR CROIS	SCOUT	PRO PHETIA	PER SIST	CORR CROIS	SCOUT	PRO PHETIA
7.3.89	0.28	0.23	0.10	0.25	0.46	0.37	0.12	0.35
4.4.89	0.59	0.34	0.42	0.35	0.71	0.54	0.61	0.60
24.4.89	0.68	0.25	0.32	0.19	0.79	0.44	0.38	0.30
27.4.89	0.29	0.11	0.10	0.13	0.48	0.17	0.21	0.28
10.5.89	0.79	0.48	0.68	0.54	0.86	0.63	0.79	0.66
2.6.89	0.26	0.10	0.29	0.15	0.53	0.32	0.54	0.31
3.6.89	0.80	0.56	0.71	0.52	0.92	0.75	0.87	0.70
6.6.89	0.84	0.43	0.54	0.44	0.92	0.65	0.72	0.63
27.6.89	0.53	0.14	0.25	0.20	0.73	0.13	0.24	0.20
10.7.89	0.49	0.42	0.44	0.43	0.69	0.60	0.59	0.53
7.8.89	0.61	0.34	0.42	0.31	0.75	0.46	0.48	0.40
12.9.89	0.65	0.48	0.49	0.45	0.82	0.63	0.58	0.57
19.9.89	0.68	0.38	0.54	0.38	0.79	0.67	0.76	0.62
23.4.90	0.89	0.59	0.70	0.56	0.95	0.80	0.89	0.78
14.5.90	0.58	0.48	0.58	0.47	0.75	0.74	0.86	0.75
9.6.90	0.82	0.44	0.64	0.48	0.87	0.59	0.75	0.60
26.6.90	0.63	0.37	0.47	0.29	0.78	0.46	0.58	0.30
21.9.90	0.78	0.28	0.17	0.17	0.98	0.25	0.34	0.42
24.9.90	0.88	0.44	0.61	0.38	0.94	0.58	0.74	0.44
30.9.90	0.82	0.53	0.60	0.52	0.88	0.60	0.65	0.54

Tableau V.2: Comparaison des taux d'erreur de quatre méthodes de prévision

La troisième technique est une technique non structurée, qui est basée sur la recherche du vecteur de déplacement pour l'image entière à l'aide de la corrélation croisée entre deux images successives. A l'instant t_k, le vecteur d'extrapolation $\vec{v}_k$ est déterminé comme suit:

$$\vec{v}_k = (\vec{v}_{k-1} + \vec{v}(I_{k-1}, I_k)) / 2$$

où $\vec{v}_{k-1}$ est le vecteur d'extrapolation de la prévision précédente, et $\vec{v}(I_{k-1}, I_k)$ est le vecteur déterminé par la méthode de la corrélation croisée décrite au premier chapitre, appliquée aux images I_{k-1} et I_k. Cette technique, qui est similaire à celle appliquée par le système SHARP (Bellon et Austin 1978), sera nommée ***CORRCROIS***.

Le TMP des prévisions des quatre techniques est, pour des intervalles d'échéance de 30 et 60 minutes, représenté dans le tableau V.2. Sont indiqués les TMP moyens des événements. Le coefficient ρ de la corrélation croisée était en moyenne de 0.68, avec un écart-type de 0.11.

Les différences entre les taux d'erreur de PROPHETIA et ceux des autres techniques sont comparées graphiquement dans la figure V.8. A l'exception de quelques cas, la performance de PROPHETIA est meilleure que celle des autres méthodes. Néanmoins, pour certaines pluies, le taux d'erreur par SCOUT et/ou CORRCROIS est inférieur à celui de PROPHETIA. Pour toutes les pluies, la performance de PERSIST est largement inférieure à celle de PROPHETIA.

Les figures V.9.a et V.9.b montrent les courbes d'efficacité des quatre techniques selon le type de la pluie. La performance de PROPHETIA est la meilleure pour les pluies convectives, tandis qu'elle est inférieure à celle de CORRCROIS pour les pluies frontales. Cette observation est confirmée par la figure V.10, qui montre la différence entre les TMP des deux techniques en fonction de la taille moyenne des échos. Toutes les pluies, pour lesquelles la méthode de la corrélation croisée donne des résultats significativement meilleurs, sont caractérisées par une taille moyenne des échos supérieure à 200 km^2.

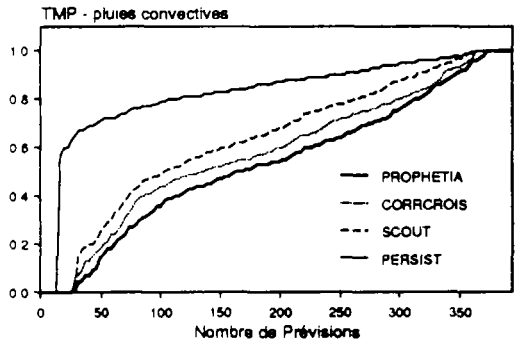


Figure V.9.a: Courbes d'efficacité pour les pluies convectives ($\Delta_p t$ = 60 min)

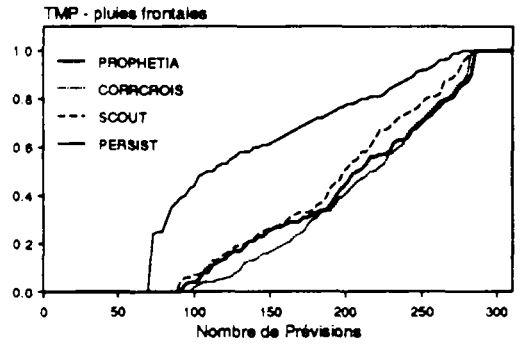


Figure V.9.b: Courbes d'efficacité pour les pluies frontales ($\Delta_p t$ = 60 min)

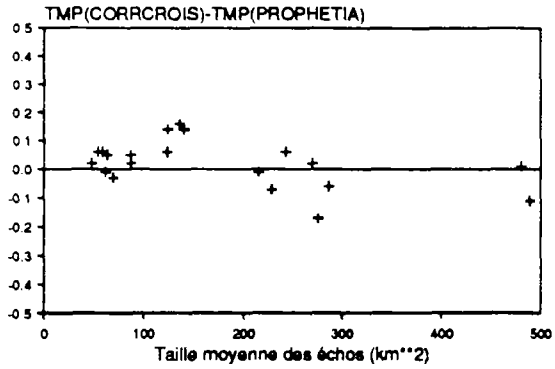


Figure V.10: Différence entre TMP de CORRCROIS et TMP de PROPHETIA en fonction de la taille moyenne des échos ($\Delta_p t$ = 60 min)

Plusieurs sources d'erreurs de PROPHETIA dans des situations frontales ont déjà été mentionnées. Les techniques appliquées par CORRCROIS et par SCOUT ne réagissent pas de la même façon à ces phénomènes:

- L'influence du bord de l'image augmente avec la taille des échos, car le temps de passage du bord est plus longue pour les grands échos. Pendant le temps de passage, l'advection de la cellule ne correspond pas au déplacement du centre de gravité de l'écho. Dans ce cas, la méthode de la corrélation croisée est mieux adaptée à la détermination de l'advection correcte. Ceci est notamment le cas pour la pluie 27.6.1989.
- Dans des situations frontales, la définition d'échos par SCOUT est généralement effectuée à partir d'un seuil plus haut. Les échos étant plus petits, l'effet du bord de l'image est moins important et l'advection individuelle des cellules intérieurs est mieux décrite. Néanmoins, les cellules ainsi repérées par SCOUT font l'objet de variations rapides, ce qui rend leur appariement difficile. En effet, pour les pluies frontales, la prévision par SCOUT repose beaucoup plus sur l'advection moyenne que sur les advections individuelles des cellules. La seule pluie frontale, pour laquelle SCOUT montre une performance supérieure et à celle de CORRCROIS, et à celle de PROPHETIA, est la pluie du 7.3.1989.
- Lorsque l'advection des cellules est mal décrite par le déplacement des centres de gravité des échos définis par PROPHETIA, l'advection moyenne de l'image détectée par CORRCROIS correspond parfois mieux à l'advection réelle. D'où la différence des taux d'erreur pour la pluie du 27.4.1989.
- Pour la pluie du 4.4.1989, le changement de l'advection de la cellule principale, provoquée par le développement local de la pluie à son intérieur, influe plus sur la prévision par PROPHETIA que sur celle par CORRCROIS, qui est en grande partie basée sur l'advection détectée sur deux images seulement. CORRCROIS réagit alors plus vite à ce changement, d'autant plus que la cellule en question est la seule cellule importante sur toute l'image, et l'advection détectée par CORRCROIS est alors en effet l'advection individuelle de cette cellule.

La pluie du 21.9.1990 présente un cas exceptionnel, le taux d'erreur de la prévision de 60 minutes de CORRCROIS étant d'environ 40% inférieur à celui de PROPHETIA. L'infériorité de PROPHETIA est provoquée par plusieurs sources. L'effet du bord de l'image au début de cette pluie a déjà été mentionné. Néanmoins, les prévisions du n° 18 jusqu'à la fin de cet événement ne souffrent pas de cet effet. La prévision est cependant influencée par un déplacement du centre de l'intensité à l'intérieur d'une cellule de taille croissante (marquée "*" sur la figure A.3.18). L'advection de la cellule, qui est déterminée par PROPHETIA, est fortement influencée par ce déplacement.

La figure V.11 montre la cellule en question à deux instants t_0 et t_0+62min, en même temps que les advections prévues pour la cellule à l'instant t_0 (la situation montrée correspond à la prévision n° 18 de cette pluie). Subjectivement, l'advection prévue par PROPHETIA est plus correcte que celle déterminée par CORRCROIS, car elle représente bien le déplacement des fortes intensités. Néanmoins, cette dernière méthode prévoit un déplacement de la cellule vers la région parisienne, qui est en fait touchée par la pluie de la zone de croissance à l'est de la cellule. La supériorité importante de CORRCROIS dans ce cas est alors partiellement due à l'effet d'un hasard.

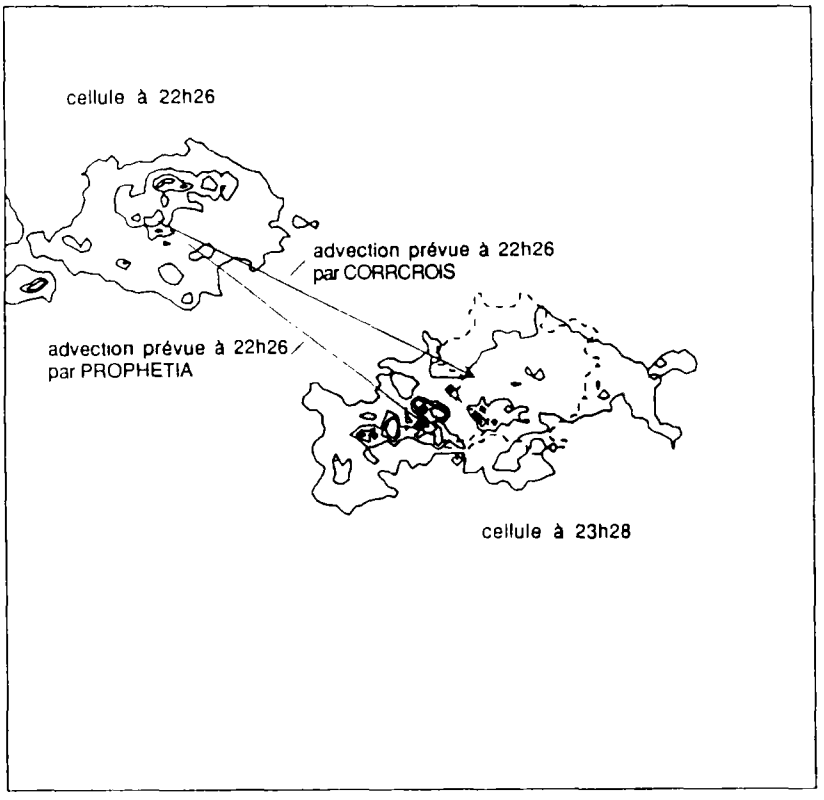


Figure V.11: Explication de l'erreur de la prévision pour la pluie du 21.9.1990. La région parisienne est indiquée par la ligne en tirets. Présentés sont les isohyètes de 2,10,15 et 20 mm/h.

V.3.4 Examen des améliorations possibles par la modification du mode de définition des échos simples

Pour certaines pluies frontales, l'analyse des sources d'erreur de la prévision de PROPHETIA a démontré une insuffisance du mode d'identification des échos simples à partir d'un seuil fixe d'intensité. Notamment pour les pluies du 27.4.89 et du 27.6.89 cette méthode ne permet pas d'identifier correctement l'advection des cellules intenses, tandis que, pour la pluie du 2.6.1989, c'est le déplacement des zones de faible pluie, qui est mal décrit. De ce fait, le taux d'erreur des prévisions par CORRCROIS est inférieur à celui de PROPHETIA pour les premiers deux de ces événements. Les meilleurs résultats de SCOUT pour les pluies du 7.3.89 et du 21.9.90 indiquent, que pour ces événements aussi une amélioration par la modification du mode de définition des échos pourrait être possible.

Trois modifications de la méthode de définition des échos simples seraient envisageables afin d'améliorer la performance du système pour les pluies frontales:

- la définition des échos à partir de maxima locaux d'intensité,
- l'adaptation du seuil de définition au type de la pluie,
- la définition des échos à plusieurs niveaux à la fois.

La définition des échos à partir des maxima locaux est proposée par Crane (1979) et appliquée par Rosenfeld (1987) et Brémaud (1991) pour l'identification des cellules par leur centre de convection. Il nous semble cependant, que la discrétisation des mesures utilisées dans cette étude en 16 niveaux n'offre peut être pas une échelle assez fine pour l'application de cette technique.

L'adaptation du seuil de définition des échos à l'intensité de la pluie est appliquée dans SCOUT, afin d'identifier les cellules les plus menaçantes pour l'hydrologie urbaine. Cette méthode implique un changement du seuil pendant l'événement de pluie selon les caractéristiques visibles sur l'image radar. L'examen des sources d'erreurs de SCOUT par Jacquet et Neumann (1990) a mis en évidence, que ce changement du seuil pendant la pluie pose des problèmes de reconnaissance des cellules et de l'observation de la pluie, dû au développement plus rapide et à la durée de vie plus courte des cellules définies à un niveau d'intensité plus haut. Vue la performance relativement faible de SCOUT pour les pluies convectives, nous n'avons pas envisagé d'introduire la méthode de définition d'échos de SCOUT dans le système PROPHETIA.

Une troisième possibilité est l'application systématique de plusieurs seuils à la fois. Par rapport à l'adaptation d'un seul seuil, cette méthode présente plusieurs avantages:

- le changement du seuil pendant l'événement de pluie est évité,
- les cellules intenses imbriquées dans des champs plus larges sont identifiables,
- la prévision ne tient compte des cellules intenses qu'au cas où les cellules intenses sont repérées <u>et</u> appariées.

L'application immédiate de cette méthode n'est cependant pas possible, car les caractéristiques différentes des cellules définies à un niveau d'intensité plus haut imposent la nécessité de suivre la même démarche que celle suivie dans cette étude pour les cellules définies au niveau de 25 dBZ:

- la sélection d'une masse minimale pour la définition des échos simples,
- la vérification du bon fonctionnement de l'algorithme de définition de échos imaginaires,
- l'apprentissage automatique d'une règle d'appariement.

Cette procédure nécessite l'appariement manuel d'un grand nombre d'échos pour chaque seuil, comme il a été effectué pour les échos du niveau 25 dBZ. Ce travail dépasse le cadre de cette étude.

Afin d'estimer les améliorations, qui seraient possible par la méthode proposée, nous avons entrepris quelques modifications heuristiques, guidé par le souci de ne pas augmenter le taux d'erreur par rapport à la prévision par PROPHETIA:

- Les échos simples sont définis aux niveaux de discrétisation 2, 3, 4, 5, et 6 (valeurs de r_s de 25, 30, 34, 38, et 41 dBZ) avec la taille minimale de 3.2 km^2, et la masse minimale $i(r_s) \cdot 2.5 \cdot 10^4$ m^3/h, où $i(r_s)$ est l'intensité moyenne du niveau de seuil r_s.
- Le facteur c de la distance maximale δ_{max} utilisé par l'algorithme de définition des échos imaginaires a été fixé à $\frac{2 \cdot i(r_s)}{i(r_s)^2}$ / min.
- Comme règle de l'appariement l'arbre de décision ADD_{APP} a été utilisé.
- Le vecteur moyen de l'image reste calculé comme la moyenne des vecteurs de déplacement des seuls échos du niveau 2. Il n'est donc pas changé par cette méthode.
- La caractérisation de l'advection pour les cellules définies à un niveau plus haut que 2 est effectuée de la même façon que pour les échos du niveau 2; avec une différence: si un vecteur individuel ne peut pas être établi pour une cellule C_i définie par un seuil r_s>25 dBZ, l'advection de C_i est estimée récursivement par le vecteur de la cellule d'un niveau plus bas, dans laquelle C_i est imbriquée.

Ces modifications n'affectent pas le fonctionnement de l'algorithme pour les échos du niveau 2. Pour les échos d'un niveau plus haut, la définition de la masse minimale en fonction du valeur de seuil réduit le nombre de cellules repérées. Le facteur c de la distance maximale prend les valeurs 0.53, 0.29, 0.18, et 0.13 pour les niveaux 3, 4, 5, et 6 respectivement, et réduit ainsi le nombre d'échos imaginaires définis. Ces deux modifications semblent nécessaires pour tenir compte de l'intensité moyenne plus élevée des cellules intenses.

Faute d'exemples d'apprentissage, la génération d'arbres de décision pour l'appariement des échos d'un seuil haut nous est impossible. Nous appliquons alors la règle ADD_{APP}, générée à partir des exemples d'appariements au niveau 2. La visualisation des appariements effectués par l'algorithme a permis de constater que les appariements sont en grande partie corrects. Le taux d'échos non reconnus semble cependant être plus important pour les échos d'un niveau supérieur à 2, ce qui peut être dû à la plus grande variabilité de ces cellules.

Avec la méthode de la caractérisation de l'advection appliquée, l'advection prévue est différente de celle obtenu par la seule définition des échos au niveau 2 uniquement pour les cellules intenses, qui sont appariées pour des intervalles supérieurs à 15 minutes.

La figure V.12 montre l'amélioration obtenue par cette méthode de définition de plusieurs ensembles d'échos simples pour les cinq pluies en question: elle permet de montrer le taux de réduction des TMP des prévisions de 60 minutes avec les seuils (2), (2,3), (2,3,4), (2,3,4,5), et (2,3,4,5,6).

Pour la pluie du 7.3.1989, l'amélioration obtenue par la définition des échos au niveau 3 s'explique par le fait, que ces échos sont, au début de l'événement, mieux appariés que ceux au niveau 2. La sous-estimation des lames d'eau est alors réduite, mais reste supérieure à celle par SCOUT à cause des mauvais appariements des cellules du niveau 2.

Pour la pluie du 27.4.1989, une réduction du taux d'erreur de plus de 30% peut être observée, si les échos sont définis aux niveaux 2, 3 et 4. L'advection d'une cellule intense, qui est imbriquée dans un champs large de pluie faible, est bien identifiée par les échos définies aux niveaux 3 et 4.

La réduction du taux d'erreur pour les autres trois pluies reste faible:

Pour la pluie du 2.6.1989, la qualité de la prévision reste pratiquement stable. Ce résultat était attendu, car l'advection individuelle des cellules est, pour cette pluie, déjà bien identifiée par les cellules du niveau 2 (cf. § V.3.2.2). C'est surtout la caractérisation incorrecte des zones de faible pluie, qui provoque la plus grande partie de l'erreur de prévision.

Pour la pluie du 27.6.1989, les cellules intenses sont bien reconnues. Toutefois l'erreur (exprimée par le TMP moyen de l'événement) n'est pas réduite, car elle est provoquée par l'évaluation de l'advection des cellules de niveau 2 qui est perturbée par l'effet du bord de l'image. L'utilisation pour le déplacement moyen prévu du vecteur moyen observé sur les échos du niveau 3 (ou supérieur) améliore l'erreur: le TMP moyen passe de 0.18 à 0.12. Cette solution n'est toutefois pas à retenir, car elle augmente l'erreur de la prévision sur d'autres pluies, comme celle du 2.6.1989, où le déplacement moyen est mal décrit par l'advection des échos du niveau 3.

Pour la pluie du 21.9.1990 finalement, la qualité de la prévision n'est pas améliorée remarquablement. Comme dans le cas du 2.6.1989, l'advection individuelle des cellules est assez bien identifiée par les cellules du niveau 2. La supériorité de SCOUT pour cette pluie semble être due au même hasard que celui qui favorise aussi la méthode CORRCROIS (cf. figure V.11 et § V.3.3), car la principale cellule n'est pas reconnue et son déplacement est estimé par le déplacement moyen.

Cet examen d'un échantillon de pluies frontales démontre, qu'une amélioration de la qualité de prévision par la définition multiple d'échos simples est possible. La vérification de la méthodologie appliquée reste à entreprendre.

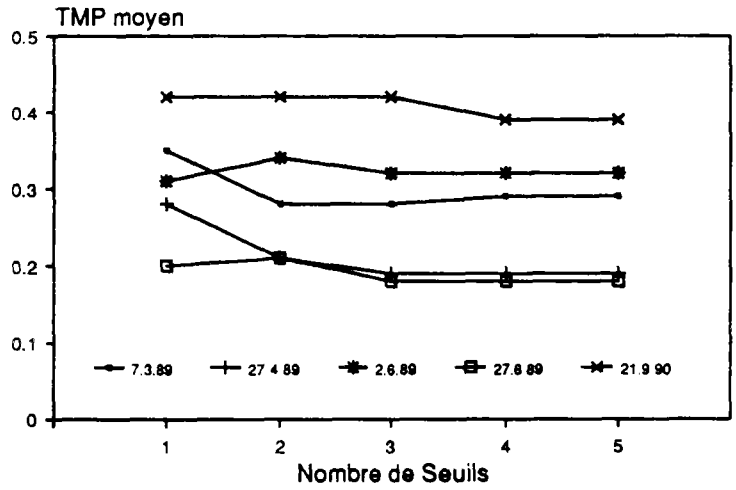


Figure V.12: Amélioration du taux d'erreur par définition multiple des échos pour cinq pluies frontales

V.3.5 Analyse de la complexité du système PROPHETIA

La complexité d'un système informatique s'exprime en termes de besoin de mémoire vive d'une part, et de temps de calcul de l'autre part. Le système PROPHETIA est installé sur un micro-ordinateur du type PC 386/20 sous le système d'exploitation DOS. La version utilisée dans cette étude n'est pas optimisée dans tous ses modules, afin de permettre une flexibilité concernant des modifications des paramètres de caractérisation éventuellement nécessaires. Toutefois, la complexité d'une version opérationnelle peut être estimée à partir de cette version de recherche.

Vu la grande quantité de données, que représente une image radar, le besoin de mémoire vive était un obstacle important pour les premiers systèmes de prévision. Le progrès obtenu dans la construction des micro-processeurs permet cependant aujourd'hui facilement des applications qui utilisent plusieurs centaines de kO. PROPHETIA nécessite environ 400 kO de mémoire vive, ce qui ne devrait pas poser de problèmes pour une application opérationnelle.

Un critère d'évaluation important est le temps de calcul nécessaire pour une prévision. L'utilité d'un système de prévision à courts termes décroît rapidement, si le temps de calcul augmente. Afin d'être utile, une prévision d'un temps d'échéance d'une heure doit être accessible en quelques minutes seulement. On s'intéresse surtout au délai maximal, à cause des contraintes posées par l'intégration dans un système de gestion en temps réel.

Le tableau V.3 présente les durées de l'exécution des quatre étapes de l'algorithme de PROPHETIA. Sont montrées les valeurs moyennes et maximales des 711 prévisions d'une échéance d'une heure pour la région parisienne. Toutefois, la somme des durées maximales des quatre étapes de l'algorithme ne correspond pas au temps de calcul maximal de l'exécution du système, car la complexité de la définition des échos, de leur appariement, et de la caractérisation des cellules dépendent du nombre d'échos sur l'image, tandis que le temps de calcul de la prévision des lames d'eau est surtout fonction de la taille des champs de pluie, qui est généralement plus grande pour les pluies stratiformes d'un petit nombre d'échos. Le temps de calcul maximal de la prévision par PROPHETIA est de 31 secondes. La définition multiple des échos augmente ce temps d'environ 100% pour quatre niveaux supplémentaires de définition d'échos. Si ce temps de calcul est acceptable ou non pour l'application opérationnelle doit être jugé en fonction des contraintes posées par l'intégration de PROPHETIA dans un système de gestion.

Etape de l'algorithme de PROPHETIA	Temps de calcul moyen (sec CPU)	Temps de calcul maximal (sec CPU)
Définition des échos simples	0.82	2.04
Définition des échos imaginaires	0.34	3.62
Appariement des échos	1.94	12.91
Caractérisation des cellules	0.36	1.82
Prévision des lames d'eau	5.07	14.33
Ensemble du système	8.53	26.37

Tableau V.3: Valeurs moyennes et maximales du temps de calcul du système PROPHETIA pour des prévisions de 60 minutes

V.4 Conclusion

Nous avons analysé la performance du système de prévision PROPHETIA pour 20 événements pluvieux, qui ont touché la région parisienne. Les pluies considérées représentent un échantillon, qui a été choisi uniquement selon des critères de diversité des types de pluie, de qualité des données, et d'importance de la pluie pour l'hydrologie urbaine dans la région étudiée.

Les prévisions automatiques par PROPHETIA ont été comparées aux prévisions semi-automatiques, qui sont basés sur un appariement manuel des échos de pluie. Cette comparaison a démontré, que la même qualité de prévision peut être atteinte avec la méthode automatique qu'avec l'intervention de l'homme dans le processus de l'appariement. La faible qualité de la prévision basée sur une règle triviale de l'appariement a mis en évidence la qualité de la règle de l'appariement, qui a été générée par l'apprentissage à partir d'exemples. Cette application montre, que la méthode de l'apprentissage automatique est un moyen performant de découverte de règles dans des domaines d'une haute complexité, dont les données sont perturbées par des incertitudes et du bruit.

La confrontation avec deux techniques automatiques de prévision, qui sont appliquées d'une manière opérationnelle, a démontrée une supériorité de PROPHETIA, qui n'est cependant pas homogène pour toutes les pluies étudiées. En fait, la méthode basée sur la corrélation croisée montre une meilleure performance pour une partie des pluies frontales, tandis que les prévisions de PROPHETIA sont meilleures surtout dans des conditions convectives.

L'analyse des sources d'erreur des prévisions par PROPHETIA a révélé deux problèmes principaux:

- l'identification des cellules intenses dans les pluies frontales,
- le développement des cellules pour les pluies convectives.

Une méthode d'identification des cellules à plusieurs niveaux d'intensité a été proposée afin de mieux identifier les cellules intenses des pluies frontales. A cause du cadre limité de cette étude, il ne nous était cependant pas possible d'optimiser le fonctionnement de cette méthode. Les résultats d'un test d'échantillon ont démontré, que pour une partie des pluies frontales une amélioration considérable est possible.

Le développement de la pluie est l'obstacle le plus important pour la prévision à court terme. La grande majorité des erreurs est due à un changement de la pluie dans l'intervalle de l'échéance. Ce problème sera examiné dans le chapitre prochain.

VI

LA PRISE EN COMPTE DU DÉVELOPPEMENT DES CELLULES DE PLUIE CONVECTIVE POUR LA PRÉVISION DE PLUIE

Dans ce dernier chapitre, nous examinerons la réalisation du deuxième objectif de cette étude: la prise en compte du développement de la pluie pour la prévision. Nous avons mis en évidence que la plus grande partie des erreurs de la prévision basée sur la seule advection des cellules est due à la croissance ou décroissance de la pluie pendant l'intervalle d'échéance. Ces erreurs sont d'autant plus graves, qu'elles biaisent surtout les prévisions des pluies orageuses: à cause de leur hétérogénéité spatiale, ces pluies offrent le plus grand potentiel pour la gestion des réseaux d'assainissement. La prévision des lames d'eau avec un faible taux d'erreur peut alors être particulièrement utile dans ces cas. Plusieurs auteurs ont adressé ce problème (par exemple Browning 1978, Bellon et Austin 1978); néanmoins, les solutions proposées à ce jour n'ont que faiblement réduit les erreurs et n'ont jamais été appliquées de manière opérationnelle.

Dans la première partie nous concrétiserons les deux aspects de la prévision du développement des cellules de pluie convective: la prévision du taux de croissance/décroissance des cellules d'une part, et la prévision de la répartition spatiale de la croissance/décroissance d'autre part.

Les mécanismes du développement de cellules de pluie convective seront examinés en deuxième partie. Nous analyserons les facteurs météorologiques, qui influent sur la convection et sur l'intensité de la pluie. Un modèle très simplifié du développement des cellules de pluie convective sera proposé, qui permettra d'identifier les paramètres qui sont à déterminer pour atteindre une amélioration de la prévision.

Dans un troisième temps, la répartition spatiale de la croissance/décroissance des cellules sera déterminé à l'aide d'une analyse statistique des cellules observables sur les images radar de la base de données. Les résultats de cette analyse permettront la définition d'une méthode de prise en compte du taux de développement des cellules pour la prévision.

Le gain qui peut apporter cette méthode sera ensuite examiné sous l'hypothèse, que le taux de croissance/décroissance des cellules soit connu. Ce taux sera déterminé à l'aide des séquences d'échos sur les images radar dans l'intervalle d'échéance de la prévision.

La dernière partie sera consacrée à la recherche d'une méthode de prévision du taux de développement des cellules. Plusieurs approches seront examinées. A cause d'insuffisances des données et des incertitudes importantes concernant les hypothèses faites pour la définition du modèle des cellules, la proposition d'une solution définitive à ce problème n'a pas été possible dans le cadre de cette étude.

VI.1 Formulation du problème

La méthode de prévision par PROPHETIA, qui a été examinée au chapitre précédent, est basée sur la caractérisation des cellules et de leur advection. La prise en compte du développement des cellules de pluie convective nécessite en plus la prévision de deux aspects du développement:

- le taux de croissance/décroissance des cellules de pluie convective,
- la répartition spatiale de la croissance/décroissance des cellules de pluie convective.

Soulignons que notre objectif n'est pas la prévision du déclenchement de la convection, ni de l'initiation des précipitations. L'information fournie par l'image radar ne serait pas suffisante pour une telle tentative. Nous cherchons exclusivement à prévoir le développement des cellules existantes, dont la précipitation est déjà assez importante pour provoquer un écho sur l'image radar.

Nous exprimons le développement d'une cellule dans un intervalle de temps (t_i,t_j) comme la variation relative de sa masse par rapport à l'instant t_i:

Définition VI.1:

Soit C une cellule, qui est représentée sur les images radar $(I_1,..,I_n)$, mesurées aux instants $t_1,..,t_n$, par la séquence stricte d'échos $se=(e_1 \in E(I_1),..,e_n \in E(I_n))$. Nous définissons le ***taux de croissance/ décroissance*** $tcd_{(t_i,t_j)}(C)$ de la cellule dans l'intervalle (t_i,t_j) $(j>i)$ comme suit:

$$tcd_{(t_i,t_j)}(C) = tcd(e_i,e_j) = \frac{\text{masse}(e_j) - \text{masse}(e_i)}{\text{masse}(e_i)}$$

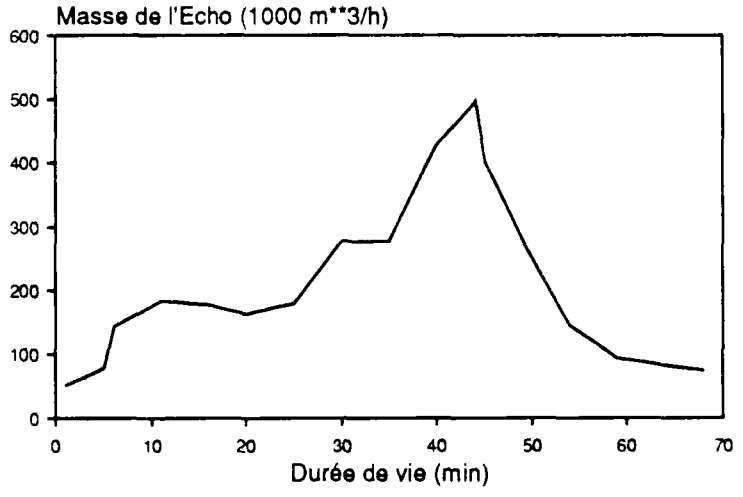


Figure VI.1: Développement de la masse d'une cellule convective (cellule de la pluie du 6.6.1989)

La figure VI.1 présente graphiquement le développement de la masse d'une cellule convective pendant son cycle de vie d'une longueur de 68 minutes. Cet exemple met en évidence que le développement des cellules n'est pas toujours uniforme, mais parfois d'une forte variabilité. Afin de tenir compte de cette variabilité, le *tcd* doit être prévu à une échelle temporelle assez fine. Car l'échelle temporelle des données utilisées dans cette étude est de 5 minutes, nous exprimons le développement des cellules dans ces intervalles. Le premier sous-problème de la prévision du développement des cellules ce présente alors comme suit:

(1) **Pour une prévision effectuée à l'instant t_0 pour la durée $(t_0,t_0+\Delta_p t)$, prévoir pour chaque cellule C_i représentée par l'écho $e_i \in E(I_0)$ la suite des taux de croissance/décroissance sur 5 minutes dans l'intervalle d'échéance de la prévision:** $(\, tcd_{(t_0,t_0+5)}(C_i), tcd_{(t_0+5,t_0+10)}(C_i),\ldots, tcd_{(t_0+\Delta_p t-5,t_0+\Delta_p t)}(C_i)\,)$

Nous cherchons à diminuer le taux d'erreur de la prévision par rapport à celui de la prévision sans prise en compte du développement, évaluée au chapitre précédent. Pour la prévision sans prise en compte du développement, la suite des *tcd* était en fait "prévue" égale à (0,...,0) pour toutes les cellules.

Le deuxième problème à résoudre pour la prise en compte du développement des cellules est la prévision de la localisation de cette croissance/décroissance:

(2) **Pour une prévision effectuée à l'instant t_0 pour la durée $(t_0,t_0+\Delta_p t)$, prévoir pour chaque cellule C_i représentée par l'écho $e_i \in E(I_0)$, pour laquelle une prévision de la suite des taux de croissance/décroissance à été possible, la localisation de ce développement pour les intervalles $(t_0,t_0+5),\ldots,(t_0+\Delta_p t-5,t_0+\Delta_p t)$.**

Afin de pouvoir proposer des solutions à ces deux problèmes, nous examinerons par la suite les facteurs qui influent sur le cycle de vie des cellules de pluie convective, et nous proposerons un modèle simplifié du développement des cellules.

VI.2 Examen des mécanismes du développement des cellules de pluie convective

VI.2.1 Facteurs influant sur le développement des cellules de pluie convective

VI.2.1.1 La convection comme source des précipitations

La pluie est désignée comme convective, si la convection est le mécanisme principal de son fonctionnement. A l'origine de la convection il y a l'instabilité verticale de l'atmosphère, qui est définie comme suit (cf. Triplet et Roche 1977):

Définition VI.2:
Soit A une particule d'air à un niveau initial de pression P_0. Lorsqu'elle est déplacée verticalement jusqu'au niveau de pression P_1, son comportement indique

- ***stabilité verticale*** de l'atmosphère entre P_0 et P_1, si A tend spontanément à revenir à P_0,
- ***instabilité verticale*** de l'atmosphère entre P_0 et P_1, si A tend spontanément s'éloigner davantage de P_0.

Dans le cas de l'instabilité, il y a une amplification progressive des déplacements verticaux. Si cette amplification reste limitée, et A trouve son équilibre à un niveau P_2 proche de P_1, on parle d'un ***instabilité sélective***: il y a stabilité pour les petites perturbations verticales, mais il peut y avoir instabilité pour certaines perturbations suffisamment importantes.

Si T_0 est la température de la particule A au niveau P_0, et T_A est la température prise par A, si elle est soulevée adiabatiquement (sans échange thermique avec son environnement) au niveau P_1 avec la température atmosphérique T_1, on montre (Triplet et Roche 1977):

- Si $T_A>T_1$, la particule poursuit spontanément son ascension: il y a instabilité verticale.
- Si $T_A<T_1$, la particule est sollicitée vers le bas et retourne au niveau P_0: il y a stabilité verticale.

La formation de nuages convectifs est liée à la présence de l'instabilité verticale, qui est alors déterminée par le profil vertical de la température et de l'humidité relative de l'atmosphère. La détection de l'instabilité nécessite la connaissance complet de ce profil dans son extension tridimensionnelle, tandis que les données disponibles sont généralement contraintes à des mesures ponctuelles par radiosondages et par des stations au sol. Plusieurs ***indices d'instabilité*** ont été proposés, qui permettent d'estimer la vraisemblance de l'existence d'une instabilité verticale à partir de ces mesures, dont celui de Showalter, et celui de Galway (Triplet et Roche 1977).

L'***indice de Showalter*** est défini comme

$$I_S = T_{500} - T_{A850}$$

où T_{500} est la température au niveau de pression 500 mb, et T_{A850} est la température prise par une particule à 850 mb, qui est amenée adiabatiquement au niveau de pression 500 mb.

L'***indice de Galway*** est défini de façon similaire comme

$$I_G = T_{500} - T_{Asol}$$

où T_{Asol} est la température prise par une particule situé près du sol, qui est amenée adiabatiquement au niveau de pression 500 mb. Galway propose de définir A comme une particule ayant la température maximale attendu ou mesurée pour une région.

Pour l'interprétation des deux indices, les seuils suivants ont été proposés:

$+3 < I$	pas de convection
$+1 < I < +3$	risques d'orages
$-3 < I < +1$	orages probables
$-6 < I < -3$	orages forts probables
$I < -6$	tornades

L'importance de la pluie provoquée par l'instabilité verticale dépend de plusieurs facteurs, dont Chalon (1987) souligne les suivants:

- la vitesse de l'ascendance,
- la hauteur de la couche atmosphérique instable,
- la chaleur et l'humidité des masses d'air soulevées,
- le cisaillement vertical des vents horizontaux.

Nous nous intéressons au facteurs, qui influent sur le développement des cellules de pluie convective après le déclenchement des précipitations. La figure VI.2 montre schématiquement les flux des masses d'air autour d'une telle cellule. Une descente de masses d'air froid est associée aux précipitations, qui rencontre les masses d'air chaud au sol et crée ainsi une zone frontale (pseudo-front). Les masses d'air soulevé par la convection, qui alimentent la cellule, parviennent des basses couches de l'atmosphère en avant de cette zone frontale. Chalon (1978) constate que le pseudo-front, et avec lui la zone d'alimentation, peut être repoussé plusieurs kilomètres à l'avant de la précipitation, si la descente des masses d'air froid possède une composante horizontale importante.

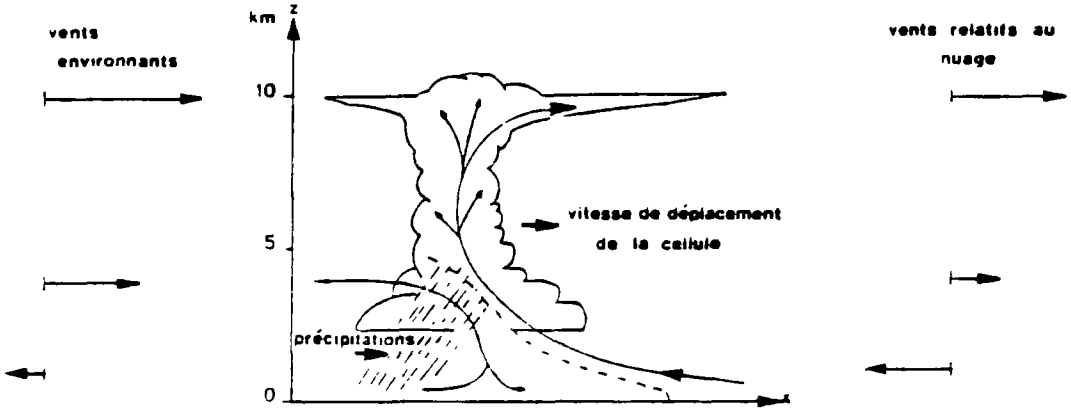


Figure VI.2: Modèle schématique d'une cellule convective, présentant aussi les vents autour du système. La ligne en pointillé schématise le pseudo-front (d'après Chalon 1978)

Nous appelons l'origine des masses d'air soulevés la ***source chaude***, et la couche atmosphérique, où la précipitation est générée, la ***source froide*** du nuage convectif. Dans la plus grande partie des nuages convectifs produisant de la pluie, la source chaude est située près de la surface du sol, où le contact avec le sol, chauffé par le rayonnement solaire, augmente la température des masses d'air, et l'évapotranspiration augmente son humidité relative.

Pendant le déplacement de la cellule, c'est surtout le changement des caractéristiques de la source chaude qui détermine son développement, car les caractéristiques de la source froide ne font pas l'objet de changements brusques. Par exemple Zawadzki et *al.* (1981) démontrent une forte corrélation statistique entre les conditions de surface, et l'intensité maximale de la pluie convective. Ils constatent que les cellules ont la tendance de s'intensifier lorsqu'elles traversent des zones d'une forte instabilité verticale de l'atmosphère, mais ils remarquent, qu'à cause des interactions entre les cellules cette correspondance spatiale n'est pas toujours facile à établir pour chaque cellule individuelle.

Nous appelons la zone au sol, d'où proviennent les masses d'air alimentant une cellule convective à un instant donné, la ***zone d'alimentation*** de la cellule. La prévision du développement de la cellule nécessite la connaissance de l'emplacement et des caractéristiques de cette zone à échéance de la prévision.

VI.2.1.2 L'influence des contrastes locaux sur la convection

Car la zone d'alimentation est d'une extension spatiale limitée, les contrastes locaux jouent un rôle important dans le cycle de vie de la cellule. A part le relief, c'est le type de couverture de la surface qui provoque ces contrastes, notamment:

- des sources d'humidité (rivières, lacs, forêts,...)
- des sources de chaleur (zones urbanisées, larges champs de céréales après rayonnement solaire intense,...)

Nous nous intéressons en premier lieu à l'effet des zones urbanisées, car c'est le développement de la pluie pendant son passage sur ces zones, qui provoque en grande partie les erreurs des prévisions constatées dans cette étude. La différence principale entre le climat urbain et le climat rural est ce qu'on appelle l'***îlot de chaleur*** des villes. Par exemple l'excès moyen annuel de la température entre la station météorologique de la Tour Saint-Jacques au centre de Paris et la station de Saint-Maur, à la limite de la région urbanisée, est de +1.06°C (Pedelaborde 1957). Des différences du même ordre de grandeur ont été observées pour d'autres grandes zones urbanisées. Arya (1988) donne les raisons suivantes pour cet excès:

- la radiation solaire est plus absorbée dans les villes à cause de la pollution de l'atmosphère urbain et à cause de la grande surface, crée par les immeubles, qui est exposé à la radiation;
- la chaleur est conservée à cause des propriétés thermiques des matériaux urbains;
- les activités humaines produisent de la chaleur (chauffage, climatisation, transports, industrie,...);
- le refroidissement par l'évapotranspiration est réduite à cause de l'imperméabilisation des surfaces et à cause du manque de végétation.

Deuxième différence entre le climat urbain et le climat rural, l'humidité relative est moins élevée dans les villes, à cause de la température plus élevée et de la réduction de l'évapotranspiration. Ainsi, Pedelaborde (1957) mentionne une étude selon laquelle l'humidité relative à la Tour Saint-Jacques serait en moyenne de 6 à 7% inférieure à celle à Saint-Maur. L'extension verticale de cette anomalie serait d'au moins 300 m (sommet de la Tour Eiffel).

L'importance de ces différences entre le climat urbain et le climat rural a été démontrée par Changnon et *al.* (1976), qui ont étudié la pluviométrie de la ville de St.-Louis (États-Unis) pendant une période de trois années. Leurs résultats indiquent, que l'atmosphère au-dessus de la zone urbanisée et sous le vent de celle-ci est favorable et à l'initiation de la pluie, et à son intensification. Une autre étude par Changnon (1976) démontre, que les cellules de pluie convective, qui traversent la ville, atteignent une plus grande altitude, une durée de vie plus longue, et une intensité plus forte que les cellules des mêmes événements, qui restent sur des zones rurales.

VI.2.2 Proposition d'un modèle du développement des cellules de pluie convective

Sous l'angle des réflexions menées ci-dessus, nous pouvons préciser les facteurs, qui influent sur le développement des cellules de pluie convective:

- Le développement des cellules à un instant donné est déterminé par les caractéristiques de leur zone d'alimentation, dont proviennent les masses d'air soulevé par la convection.
- Les caractéristiques le plus déterminantes de cette zone sont l'humidité relative et la température des masses d'air près du sol.
- Outre les conditions météorologiques générales, les contrastes locaux influent de façon importante sur les caractéristiques des masses d'air près du sol. Ces contrastes sont provoqués par les différences à petite échelle de la surface, notamment l'altitude, la végétation et le degré de l'urbanisation.

Afin de pouvoir simuler le développement des cellules de pluie convective dans une intervalle de temps (t_0, t_1), nous proposons un modèle basé sur les hypothèses suivantes:

- l'emplacement et la forme de la cellule à l'instant t_0 sont connus,
- l'advection de la cellule dans l'intervalle (t_0, t_1) est connue,
- la zone d'alimentation de la cellule ne s'étend pas au-delà de 2 km des limites de son écho radar,
- les masses d'air situées sur des régions touchées par une pluie supérieure à 0.5 mm dans l'intervalle (t_0-15 min, t_0) ne peuvent pas alimenter la cellule.
- le développement de la cellule dans l'intervalle (t_0,t_1) peut être expliqué par le changement des caractéristiques de la zone d'alimentation.

Ces hypothèses mènent à une modélisation très simplifiée, qui néglige plusieurs aspects du développement des cellules, notamment:

- L'extension spatiale de la zone d'alimentation est très variable; elle dépend entre autres de la topographie locale, des vents autour de la cellule, et de l'intensité des mouvements verticaux des masses d'air.
- La capacité des masses d'air près du sol d'alimenter la cellule après avoir été touchées par une chute de pluie dépend entre autres de la température au sol et en altitude, et du rayonnement solaire après la chute. Le seuil choisi de 0.5 mm pourrait être trop bas dans certaines situations météorologiques.
- Le développement de la cellule est déterminé et par les caractéristiques de la zone d'alimentation près du sol, et par les caractéristiques de l'atmosphère en altitude, notamment par le cisaillement vertical des vents horizontaux (Chalon 1978). Bien qu'une interaction existe entre ces influences, le caractère tourbillonnaire des cellules de pluie convective introduit une forte composante aléatoire.

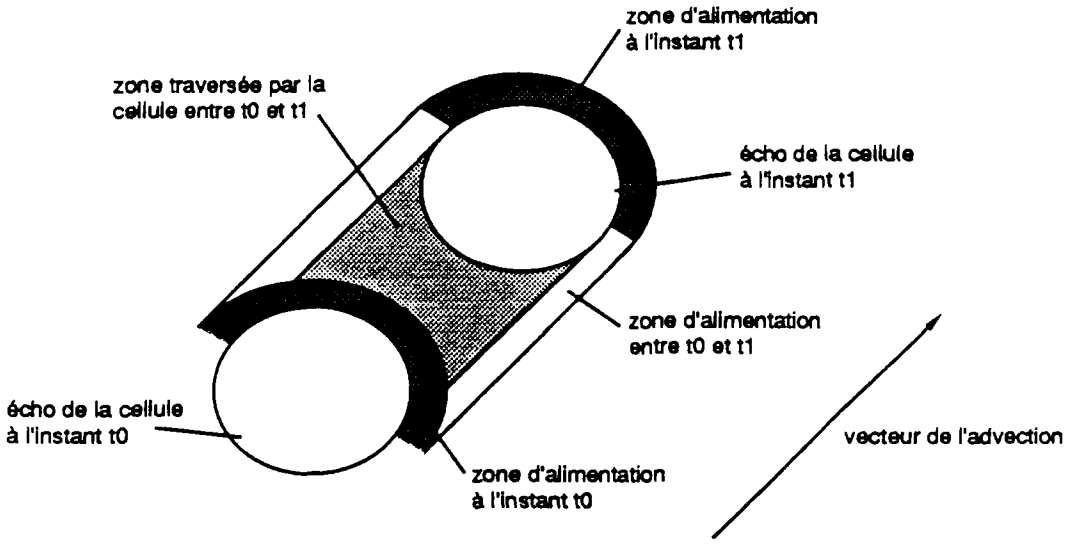


Figure VI.3: Présentation schématique du modèle de développement des cellules de pluie convective

Néanmoins, les données accessibles pour cette étude ne permettent pas l'utilisation d'un modèle plus exact. La figure VI.3 présente schématiquement le modèle des zones, dont les caractéristiques déterminent le développement d'une cellule pendant l'intervalle (t_0, t_1):

- la zone couverte par l'écho de la cellule à l'instant t_0,
- la zone couverte par l'écho de la cellule à l'instant t_1,
- la zone traversée par la cellule pendant l'intervalle (t_0, t_1),
- la zone d'alimentation de la cellule à l'instant t_0,
- la zone d'alimentation de la cellule à l'instant t_1,
- la zone d'alimentation de la cellule pendant l'intervalle (t_0, t_1).

Ce modèle nous permettra d'étudier les résolutions possibles aux deux problèmes de la prévision du taux et de la localisation de la croissance/décroissance des cellules.

VI.3 Étude de la répartition spatiale de la croissance/décroissance dans les cellules de pluie convective

VI.3.1 La croissance des cellules de pluie convective

Pour une cellule croissante C, qui est représentée par les deux échos e_0 et e_1 sur deux images I_0 et I_1 mesurées aux instants t_0 et $t_1 = t_0$+5 min, nous cherchons à déterminer la localisation de la masse supplémentaire de l'écho e_1:

$$\Delta masse(e_0,e_1) = \text{masse}(e_1) - \text{masse}(e_0) > 0$$

On applicant le modèle proposé du développement des cellules, nous définissons quatre zones autour de l'écho e_1, qui sont susceptibles d'avoir des caractéristiques d'alimentation différentes (cf. figure VI.4):

- la zone définie par les limites de l'écho e_0, advecté pour l'intervalle (t_0,t_1) (zone 1),
- la zone traversée par la cellule dans l'intervalle (t_0,t_1) (zone 2),
- la zone 1 km autour de la zone 1, hors zone 2 (zone 3),
- la zone 2 km autour de la zone 1, hors zones 2 et 3 (zone 4).

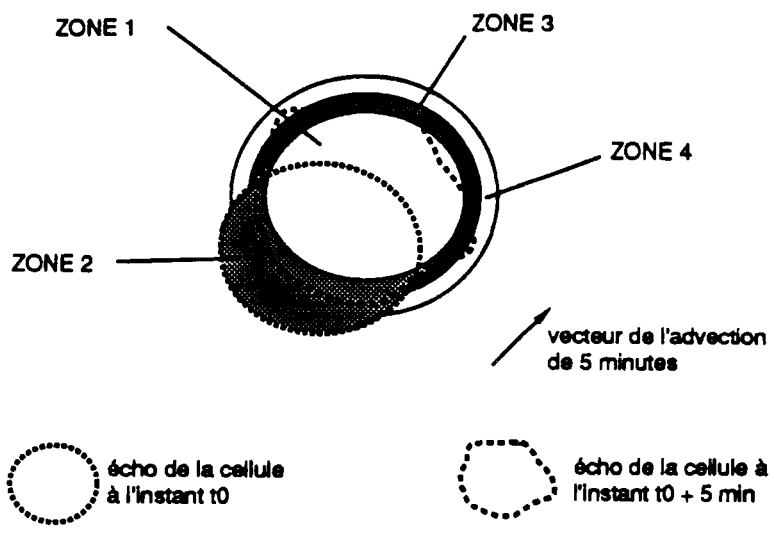


Figure VI.4: Les différentes zones autour d'une cellule en croissance

La base de données radar, avec les appariements manuels d'échos, a été utilisée afin d'établir statistiquement une rélation entre la croissance de la masse et les quatre zones ainsi définies. 532 couples d'échos (e_0,e_1) ont été sélectionnés selon les critères suivantes:

- les échos font partie d'une pluie convective d'une advection moyenne supérieure à 20 km/h (10 pluies),
- le centre de gravité de l'écho e_0 est compris dans un carré de 100 km de coté, centré sur le radar,
- les échos e_0 et e_1 sont des échos simples, qui ont été appariés manuellement,
- le taux de croissance est non négligeable: $tcd(e_0,e_1) > 0.02$/min.

Ainsi, uniquement les cellules, dont le déplacement est assez important pour la détermination des quatre zones, et qui ne sont pas affectées par l'effet du bord de l'image, ont été considérées. Afin d'écarter le biais, qui peut être introduit par l'influence du développement de la cellule sur son advection individuelle, le vecteur moyen de l'image I_0 a été utilisé comme vecteur de l'advection. Pour chaque couple (e_0,e_1), la localisation de la masse supplémentaire a été déterminée comme suit:

- l'écho e_0, advecté par le vecteur moyen de l'image I_0, a été superposé sur l'image I_1,
- les quatre zones ont été définies sur l'image I_1,
- le taux de croissance a été déterminé pour chaque zone:

$$tcd^{\,z}(e_0,e_1) = \frac{\text{masse}\,(e_1 \cap \text{zone}\,z) - \text{masse}\,(e_0 + adv_{moy}(I_0) \cap \text{zone}\,z)}{\text{masse}(e_0)}$$

(on a $e_0 + adv_{moy}(I_0) \cap \text{zone}\,z = \varnothing$ pour $z \neq 1$)

En moyenne seule 13% de la masse supplémentaire est située hors les quatre zones; cette masse a été négligée dans cette analyse.

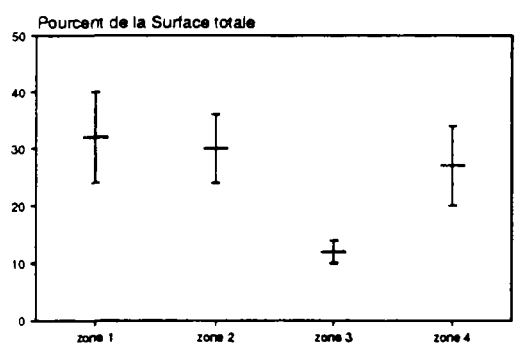


Figure VI.5.a: Surface des quatre zones comme pourcentage de la surface totale (moyenne +/- écart-type)

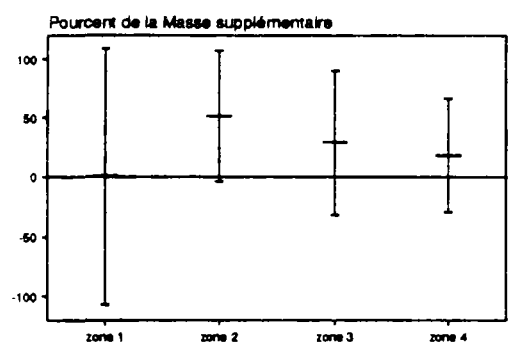


Figure VI.5.b: Répartition de la masse supplémentaire sur les quatre zones (moyenne +/- écart-type)

La figure VI.5.a montre la taille des surfaces des quatre zones comme pourcentage de la somme des quatre surfaces. Sont présentées les valeurs moyennes des 532 cas et leur écart-type. Les zones 1, 2, et 4 ont en moyenne une surface pratiquement équivalente, tandis que la zone 3 ne représente en moyenne que 12% de la surface totale. La localisation moyenne de la masse supplémentaire est montrée par la figure VI.5.b, qui présente la valeur moyenne des $tcd^z(e_0,e_1)$ comme pourcentage du tcd total moyen, ainsi que l'écart-type de cette valeur.

En moyenne, environ 50% de la masse supplémentaire sont situés dans la zone 2, qui ne représente que 30% de la surface considérée, et 30% de la masse supplémentaire se trouvent dans la zone 3, qui représente environ 10% de la surface. La croissance dans la zone 1 est en moyenne négligeable, en dépit du fait que la zone représente 30% de la surface. L'écart-type de ces valeurs est cependant extrêmement fort, surtout en ce qui concerne la croissance dans la zone 1, ce qui indique une forte variabilité de la répartition spatiale de la croissance. L'écart entre l'advection moyenne appliquée, et l'advection réelle des cellules peut expliquer une partie de cette variabilité. Nous attribuons une autre partie au caractère tourbillonnaire, et donc aléatoire, de la convection.

Le fait, qu'en moyenne la moitié de la masse supplémentaire se trouve dans la zone 2, peut être expliqué comme suit: si la partie de la zone d'alimentation de la cellule pendant l'intervalle (t_0,t_1), qui n'est pas touchée par la pluie (la partie qui se trouve sur les cotés de la cellule, cf. figure VI.3), continue à alimenter la cellule après le passage de son "noyau", alors la cellule s'étend à l'arrière par rapport à son déplacement et dévient plus allongée. Cette allongement des cellules peut souvent être observée sur les images radar des pluies convectives.

Une corrélation entre la localisation de la masse supplémentaire et le taux total de croissance $tcd(e_0,e_1)$ a été observée seulement pour la zone 1. Pour cette zone, le *tcd* de la zone est généralement positif, si le *tcd* total est très grand, tandis qu'il est négligeable ou légèrement négatif pour les cas d'un *tcd* total moins important. Afin d'alléger la méthode, nous négligeons cette corrélation.

Nous avons alors pu démontrer, que la croissance des cellules de pluie convective dans un intervalle de temps (t_0, $t_1=t_0+5$ min) est un processus non-uniforme, qui suit certaines régularités:

- Dans les limites de l'écho mesuré à l'instant t_0, la masse reste en moyenne stable.
- Environ 50% de la croissance s'expriment par un allongement de la cellule à l'arrière par rapport à son déplacement.
- Environ 30% de la croissance consistent en une extension spatiale de la cellule de moins d'un kilomètre des limites de l'écho à l'instant t_0.

Compte rendu de la grande variabilité des résultats obtenus, ils sont à interpréter comme une tendance observée sur un grand échantillon de cas.

VI.3.2 La décroissance des cellules de pluie convective

Afin d'étudier la localisation de la décroissance des cellules, nous avons suivi une démarche similaire à celle suivi concernant la croissance. Nous considérons une cellule décroissante C, qui est représentée par les deux échos e_0 et e_1 sur deux images I_0 et I_1 mesurées aux instants t_0 et $t_1=t_0+5$ minutes, et nous cherchons à déterminer la localisation dans l'écho e_1 de la perte de masse:

$$\Delta masse(e_0,e_1) = \text{masse}(e_1) - \text{masse}(e_0) < 0$$

Deux zones dont l'alimentation par les basses couches peut être différente ont été définies pour une cellule décroissante (figure VI.6):

- la zone définie par l'intersection des limites des échos e_0 et e_1, qui peut être alimentée par la même zone d'alimentation que l'écho e_0 (zone 1a),
- la zone définie par les limites de l'écho e_1, hors zone 1a, qui n'est pas alimentée par la même zone d'alimentation que l'écho e_0 (zone 1b).

533 couples d'échos d'un taux de décroissance non négligeable ($tcd(e_0,e_1) < -0.02$/min) ont été choisis selon les mêmes critères appliqués pour l'étude de la croissance.

Les figures VI.7.a et VI.7.b montrent les résultats de l'analyse de la décroissance de la même façon que les figures VI.5.a et VI.5.b pour la croissance. En moyenne, la zone 1b représente environ deux tiers de la surface, tandis que la zone 1a représente l'autre tiers. La répartition spatiale de la

décroissance correspond exactement à cette répartition des surfaces. Nous devons alors conclure, que la décroissance des cellules de pluie convective est un processus uniforme, qui s'étend à toute la surface de la cellule.

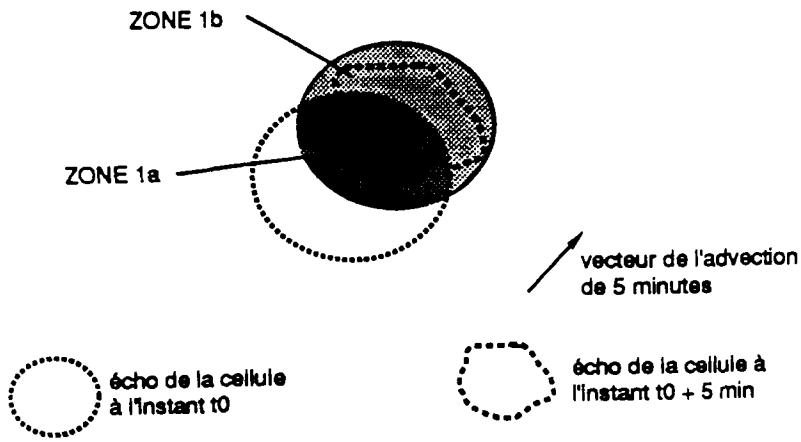


Figure VI.6: Les différentes zones autour d'une cellule décroissante

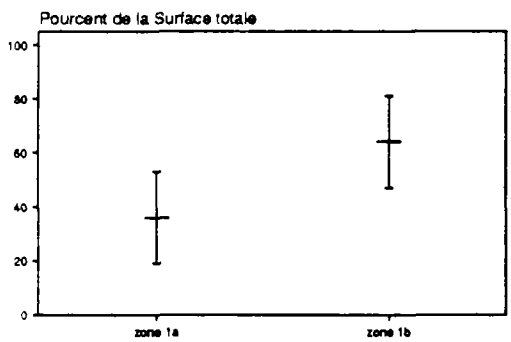


Figure VI.7.a: Surface des deux zones comme pourcentage de la surface totale (moyenne +/- écart-type)

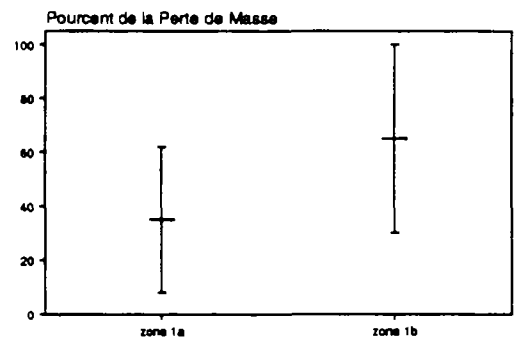


Figure VI.7.b: Répartition de la masse diminuée sur les deux zones (moyenne +/- écart-type)

VI.4 Examen de l'amélioration possible de la prévision par prise en compte du développement des cellules de pluie convective

Les résultats obtenus par l'étude du processus du développement des cellules de pluie convective nous permettent la mise en place d'une méthode d'application du taux de croissance/décroissance pour la prévision des lames d'eau. Nous examinerons le gain, que peut apporter la prise en compte du développement des cellules pour la prévision, par l'application des *tcd* exactes, qui seront déterminés à l'aide des séquences d'échos sur les images dans l'intervalle d'échéance de la prévision.

L'algorithme de PROPHETIA (algorithme V.1) a été modifié afin d'adapter les échos de l'image I_0 selon la suite des *tcd* déterminée dans l'intervalle $(t_0, t_0+\Delta_p t)$. Rappelons que, pour un écho $e_i \in E(I_0)$, qui représente une cellule C_i sur l'image I_0 cette suite est la suite des *tcd* sur 5 min

$$(tcd_{(t_0,t_0+5)}(C_i), tcd_{(t_0+5,t_0+10)}(C_i),..., tcd_{(t_0+\Delta_p t-5,t_0+\Delta_p t)}(C_i))$$

qui, pour cet examen, peut être déterminé par les échos de la séquence dont e_i fait partie.

L'adaptation de la masse des échos est effectuée dans l'étape (1) de l'algorithme, qui prévoit l'image pour l'instant $t=t_0+k$ minutes ($k=1,..,\Delta_p t$). Dans la nouvelle version de l'algorithme, qui est nommé ***PROPHETIA.II***, la masse des échos est modifiée dans les cycles de $k=2,7,12,...,\Delta_p t-3$. Les échos sont modifiés selon les valeurs des $tcd_{(..)}(C_i)$ avec la répartition spatiale découverte par l'analyse de la croissance/décroissance des cellules:

(1) Si $tcd_k(C_i) := tcd_{(t_0+k-2,t_0+k+3)}(C_i) < 0$, alors la masse de l'écho e_i est diminuée de la différence

$$\Delta\text{masse}_k(e_i) = \text{masse}(e_i) \cdot tcd_k(C_i)$$

par choix aléatoire de pixels de l'écho et réduction de leur intensité par un niveau de réflectivité. Si le niveau de réflectivité d'un pixel sélectionné correspond au niveau de seuil de définition des échos, alors le pixel est enlevé de l'écho.

(2) Si $tcd_k(C_i) = tcd_{(t_0+k-2,t_0+k+3)}(C_i) > 0$, alors les quatre zones de croissance sont déterminées pour l'écho e_i, et le gain de masse

$$\Delta\text{masse}_k(e_i) = \text{masse}(e_i) \cdot tcd_k(C_i)$$

est réparti entre les zones 2, 3, et 4, avec un coefficient de 0.5, 0.3, et 0.2 respectivement, par choix aléatoire des pixels et augmentation de leur intensité au niveau correspondant à l'intensité moyenne de l'écho e_i.

La décroissance de la masse est alors reproduite par diminution des intensités uniformément sur toute la surface de l'écho, tandis que la croissance est modélisée non-uniformément dans les quatre zones autour de l'écho.

Les figures VI.8.a-VI.8.l présentent, pour les 12 pluies convectives de la base de données, le changement du taux d'erreur de la prévision d'un intervalle d'échéance de 60 minutes obtenu avec PROPHETIA.II par cette prise en compte du développement, comparé aux taux d'erreur de la prévision par PROPHETIA sans prise en compte du développement. Afin de mettre en évidence l'importance des améliorations obtenues, le taux de bassins touchés par la pluie ou la prévision par PROPHETIA est rappelé sur ces mêmes figures.

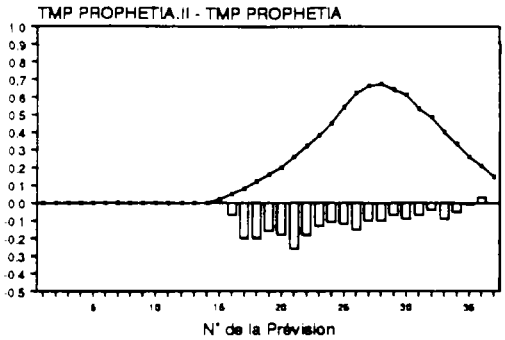


Fig. VI.8.a: Taux de bassins touchés et réduction du taux d'erreur pour la pluie du 24.4.89

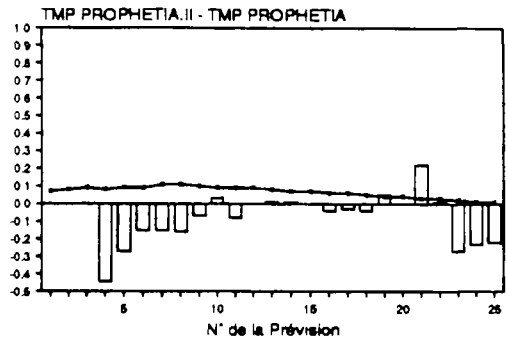


Fig. VI.8.b: Taux de bassins touchés et réduction du taux d'erreur pour la pluie du 10.5.89

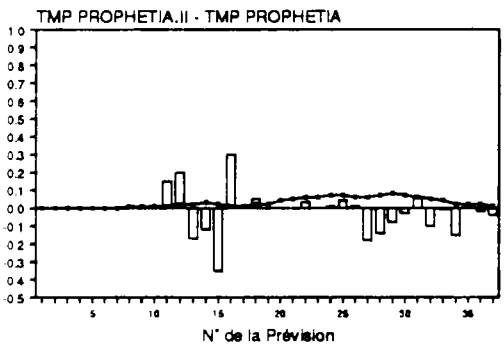


Fig. VI.8.c: Taux de bassins touchés et réduction du taux d'erreur pour la pluie du 3.6.89

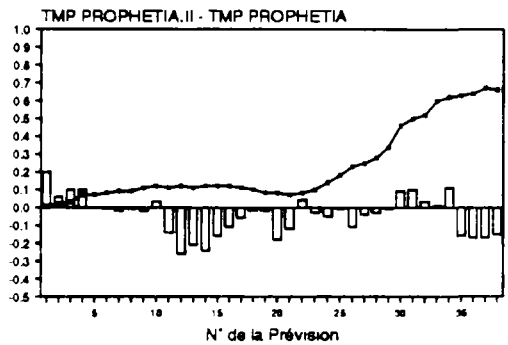


Fig. VI.8.d: Taux de bassins touchés et réduction du taux d'erreur pour la pluie du 6.6.89

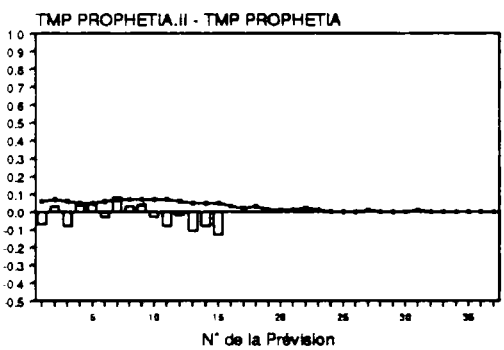


Fig. VI.8.e: Taux de bassins touchés et réduction du taux d'erreur pour la pluie du 10.7.89

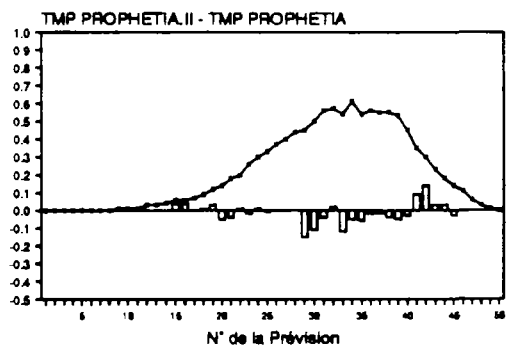


Fig. VI.8.f: Taux de bassins touchés et réduction du taux d'erreur pour la pluie du 7.8.89

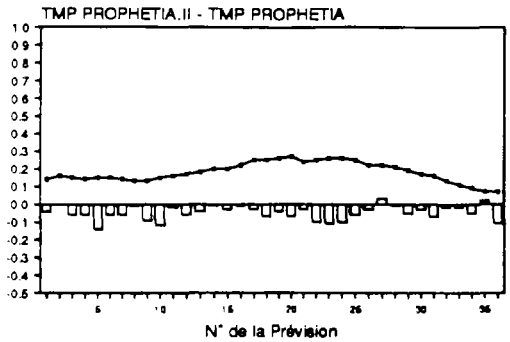


Fig. VI.8.g: Taux de bassins touchés et réduction du taux d'erreur pour la pluie du 12.9.89

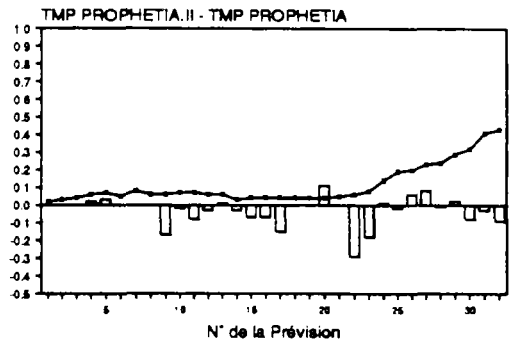


Fig. VI.8.h: Taux de bassins touchés et réduction du taux d'erreur pour la pluie du 23.4.90

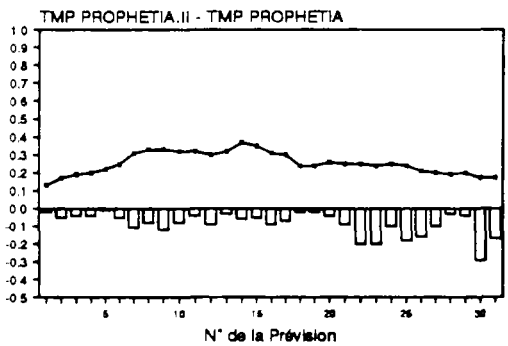


Fig. VI.8.i: Taux de bassins touchés et réduction du taux d'erreur pour la pluie du 9.6.90

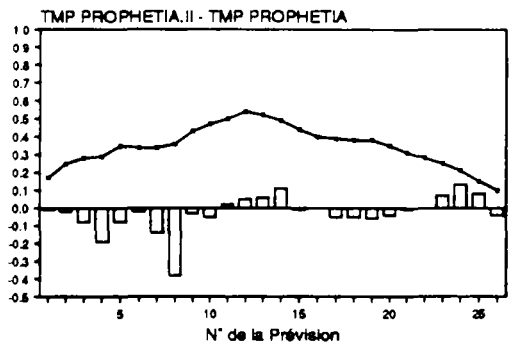


Fig. VI.8.j: Taux de bassins touchés et réduction du taux d'erreur pour la pluie du 26.6.90

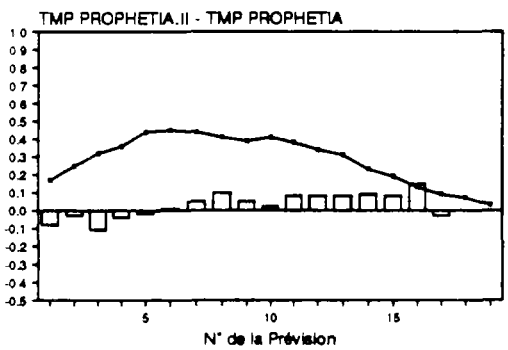


Fig. VI.8.k: Taux de bassins touchés et réduction du taux d'erreur pour la pluie du 24.9.90

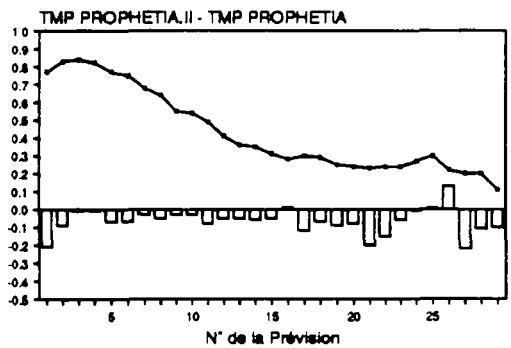


Fig. VI.8.l: Taux de bassins touchés et réduction du taux d'erreur pour la pluie du 30.9.90

Pour la plus grande partie des prévisions, qui touchent la région parisienne, la prise en compte du développement des cellules permet la réduction du taux d'erreur. Le taux d'erreur n'est évidemment pas changé pour les prévisions, où l'erreur est provoquée par des cellules croissantes qui n'existent pas encore sur l'image I_0 (par exemple les prévisions n° 1-4 de la pluie du 10.5.89, et n° 1-8 de la pluie du 23.4.90).

Pour certaines périodes, le taux d'erreur est augmenté par l'application du *tcd*. Dans certains cas, cette augmentation est due à un nombre très petit de bassins touchés par la pluie dans l'intervalle d'échéance. Dans ces cas, la composante aléatoire du développement est parfois trop importante et ne permet pas une amélioration de la prévision (notamment les prévisions n° 19 et 21 de la pluie du 10.5.89, n° 11, 12, et 16 de la pluie du 3.6.89, n° 1 et 2 de la pluie du 6.6.89, et n° 20 de la pluie du 23.4.90).

Toutefois il existent d'autres cas d'augmentation du taux d'erreur, pour lesquels un nombre considérable de bassins est touché par la pluie dans l'intervalle d'échéance. La visualisation du fonctionnement de PROPHETIA.II a révélé deux origines majeures à ces cas:

- la décroissance d'une cellule repérée comme écho sur l'image I_0, associée à l'initiation et à la croissance d'une cellule près de cette première, qui n'existe pas encore sur l'image I_0;
- la croissance de cellules sous une forme, qui ne correspond pas au modèle formulé.

La première source d'erreur a été observée pour 15 prévisions (les prévisions n° 3 et 4 de la pluie du 6.6.89, n° 2, 4, 5, et 7-9 de la pluie du 10.7.89, n° 15-19 de la pluie du 7.8.89, et n° 26, 27, et 29 de la pluie du 23.4.90). Elle conduit à une sous-estimation des lames d'eau.

La deuxième source d'erreur, conduisant à une surestimation des lames d'eau, concerne les cellules, pour lesquelles le mécanisme du développement ne correspond pas au modèle développé:

- Pour certaines cellules, la croissance ne s'exprime pas par un allongement dans le sens du déplacement, mais par la génération d'une front perpendiculaire au déplacement formée par l'apparition de plusieurs cellules de pluie convective (prévisions n° 30-34 de la pluie du 6.6.89, et n° 41-44 de la pluie du 7.8.89).
- Pour d'autres cellules, la croissance n'affecte que des petites cellules convectives d'une très forte intensité à l'intérieur de cellules plus grandes (> 200 km^2). Le mode de définition par seuillage des échos est la source principale de ce problème: il ne permet pas de déterminer correctement le taux de croissance des cellules (prévisions n° 11-14 et n° 23-25 de la pluie du 26.6.90, n° 7-16 de la pluie du 24.9.90, et n° 26 de la pluie du 30.9.90).

La figure VI.9 présente les courbes d'efficacité de la prévision avec et sans prise en compte du développement; la réduction des TMP moyens des événements est présentée dans le tableau VI.1. La réduction de l'erreur est de 9% en moyenne; pour cinq pluies elle dépasse 10% (plus de 35% pour la pluie du 24.4.89).

Néanmoins, la prévision des *tcd* exacts, comme ils étaient appliqués ci-dessus, est certainement impossible. Afin d'estimer la sensibilité des résultats obtenus à une différence entre le *tcd* prévu et le *tcd* réel, des erreurs de deux types ont été introduites:

(1) Une segmentation du *tcd* exact en cinq classes, puis l'application d'une valeur unique pour chaque classe:
   - "fortement décroissant" ($tcd < -0.3$) $\rightarrow$ $tcd = -0.4$
   - "décroissant" ($-0.3 \leq tcd < -0.1$) $\rightarrow$ $tcd = -0.2$
   - "stable" ($-0.1 \leq tcd < 0.1$) $\rightarrow$ $tcd = 0$.
   - "croissant" ($0.1 \leq tcd < 0.3$) $\rightarrow$ $tcd = 0.2$
   - "fortement croissant" ($0.3 \leq tcd$) $\rightarrow$ $tcd = 0.4$

(2) Un bruit aléatoire entre -50% et +50% de la valeur exacte du *tcd*.

Les résultats des prévisions avec ces valeurs incorrectes sont présentés dans les colonnes 3 et 4 du tableau VI.1. Par rapport à la prévision avec le *tcd* correct, la réduction du taux d'erreur est moins importante, mais reste dans le même ordre de grandeur.

Date de la pluie	PROPHETIA sans prise en compte du *tcd*	PROPHETIA.II avec prise en compte du *tcd* exact	PROPHETIA.II avec prise en compte du *tcd* segmenté en 5 classes	PROPHETIA.II avec prise en compte du *tcd* bruité
24.4.89	0.30	0.19	0.21	0.19
10.5.89	0.66	0.59	0.61	0.60
3.6.89	0.70	0.66	0.67	0.68
6.6.89	0.63	0.57	0.59	0.57
10.7.89	0.54	0.52	0.52	0.53
7.8.89	0.40	0.36	0.37	0.37
12.9.89	0.57	0.52	0.53	0.52
23.4.90	0.78	0.74	0.75	0.74
9.6.90	0.60	0.52	0.53	0.53
26.6.90	0.30	0.28	0.29	0.27
24.9.90	0.44	0.45	0.44	0.47
30.9.90	0.54	0.47	0.50	0.48

Tableau VI.1: Comparaison des taux d'erreur de la prévision sans et avec prise en compte du développement avec la connaissance exacte et erronée du *TCD*

L'importance de l'amélioration obtenue par la prise en compte du développement des cellules doit être jugée dans le contexte précis de l'application des prévisions (cf. Denoeux 1989). Sans connaissance de ce contexte, l'appréciation objective des résultats est impossible. Il nous semble cependant, que la réduction possible du taux d'erreur d'environ 10% justifie la poursuite de la recherche d'une méthode de prévision du développement des cellules; d'autant que la prise en compte du développement est le seul moyen de dépasser les limites actuelles de la qualité des prévisions.

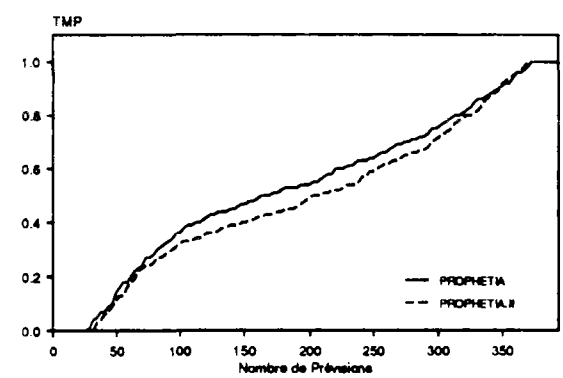


Figure VI.9: Courbes d'efficacité de la prévision avec et sans prise en compte du développement (pluies convectives, $\Delta_p t$ = 60 min)

VI.5 La prévision du taux de croissance/ décroissance pour les cellules de pluie convective

L'étude menée a démontré l'intérêt d'une prise en compte du développement des cellules de pluie convective pour la prévision des lames d'eau. Après avoir développé une technique d'application des taux de croissance/décroissance des cellules, nous examinerons par la suite différentes méthodes de prévision de ce taux.

VI.5.1 L'identification de régions favorables à l'intensification et à l'affaiblissement de la pluie

L'influence de certaines caractéristiques topographiques sur la pluie est bien connue: par exemple des chaînes de montagne provoquent une intensification de la pluie sur le versant au vent, et une affaiblissement sur le versant sous le vent. En présence de tels effet orographiques, une amélioration des prévisions par leur prise en compte semble évident.

Bellon et Austin (1978) ont tenté d'identifier de telles régions aux environs de Montréal (Canada). Dans leur étude, la méthode suivante a été appliquée:

- prévision des images radar (méthode de la corrélation croisée),
- comparaison des images prévues aux images mesurées (images CAPPI à 2000 m d'altitude),
- identification des pixels d'une forte différence des intensités entre les deux images.

Ils ont repéré plusieurs régions, qui étaient statistiquement plus favorables soit à une intensification, soit à un affaiblissement de la pluie. Pour une partie de ces régions, des effets topographiques ont pu être identifiés (montagnes, lacs). Les auteurs soulignent cependant, que leurs résultats ne sont pas assez significatifs pour permettre une amélioration des prévisions dans la région étudiée.

Pour la prévision du développement des cellules de pluie convective en région parisienne, cet approche ne promet pas une amélioration, car les différences orographiques dans cette région sont faibles et des grandes surfaces d'eau n'existent pas. Quant à l'effet des zones urbanisées, leur influence sur la convection est trop diversifiée pour permettre d'établir une tendance générale.

VI.5.2 L'extrapolation du développement observé des cellules

Plusieurs auteurs ont tenté d'améliorer la qualité de la prévision par l'extrapolation du développement observé des cellules:

Huff et *al.* (1980) ont appliqué une technique d'extrapolation des développements observés pendant l'intervalle (t_0-10 min, t_0) pour l'intervalle de prévision (t_0, $t_0+\Delta_P t$). Ils ne donnent pas les résultats de cette prise en compte du développement a amélioré la prévision.

Tsonis et Austin (1981) ont mené une étude sur l'extrapolation du développement de cellules tropicales. Ils ont examiné quatre techniques d'extrapolation; l'extrapolation linéaire a montré la meilleure performance. Les taux de croissance/décroissance ainsi établis ont été appliqués par l'ajout ou la suppression aléatoire de pixels au limites des échos. Mais ils n'ont pas

trouvé une amélioration de la qualité de prévision de cette technique par rapport à celle de prévision sans prise en compte du développement.

Afin d'examiner le gain qui peut apporter une technique d'extrapolation des *tcd* observés pour la prévision par PROPHETIA, nous avons analysé la variabilité du *tcd* des cellules de pluie convective. Des séquences ($e_{-n} \in E(I_{-n}),..,e_0 \in E(I_0),..,e_m \in E(I_m)$) de cellules de pluie convective ont été sélectionnés selon les critères suivantes:

- le centre de gravité de l'écho e_0 se trouve dans un carré de 100 km de coté, centré sur le radar,
- la durée de la séquence est d'au moins de 30 minutes avec $t_0 - t_{-n} > 15$ min et $t_m - t_0 > 15$ min.

492 séquences ont été retenues, dont les *tcd* sur 5, 10 et 15 minutes avant et après t_0 ont été déterminés. La comparaison de ces valeurs est présentée graphiquement dans les figures VI.10.a-VI.10.c. Pour aucun des intervalles il n'y a de corrélation entre les *tcd* observés avant t_0 et les valeurs d'après t_0. La prévision des *tcd* par extrapolation des observations semble alors impossible pour les pluies étudiées dans cette étude.

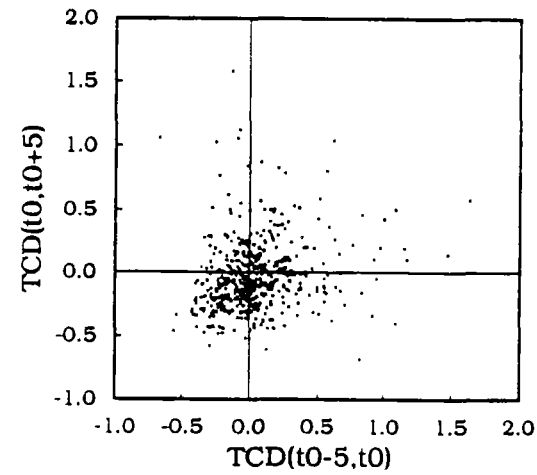


Figure VI.10.a: *tcd* sur 5 minutes avant et après t_0 pour 492 cellules de pluie convective

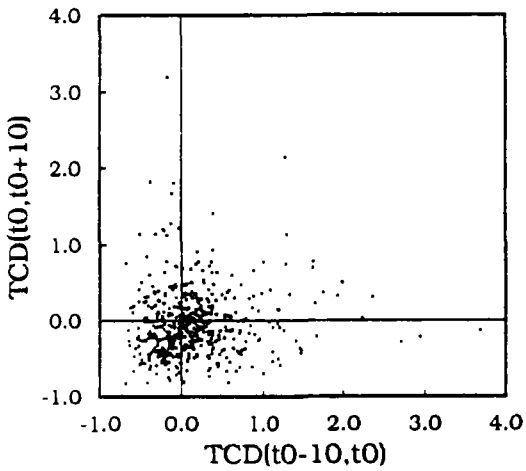


Figure VI.10.b: *tcd* sur 10 minutes avant et après t_0 pour 492 cellules de pluie convective

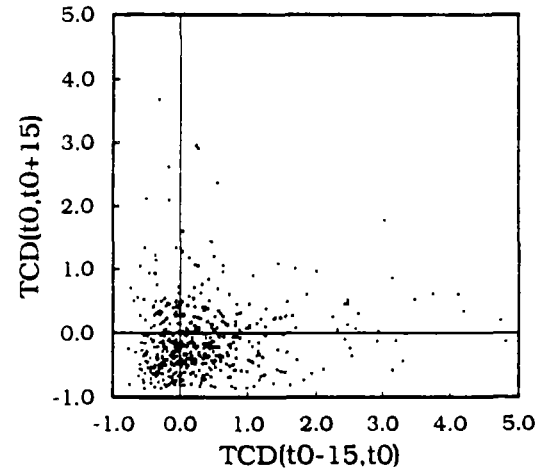


Figure VI.10.c: *tcd* sur 15 minutes avant et après t_0 pour 492 cellules de pluie convective

VI.5.3 Application de l'apprentissage automatique à la prévision du taux de croissance/décroissance des cellules

Le problème de la prévision du *tcd* peut être formalisé comme contexte de classification de plusieurs classes. L'ensemble des objets de ce contexte est formé par les séquences d'échos $s=(e_{-k}\in E(I_{-k}),..,e_0\in E(I_0))$, qui peuvent être décrites par des attributs caractérisant le développement de la cellule, son endroit sur l'image radar, *etc.* L'ensemble des problèmes du contexte est l'estimation du taux de croissance/décroissance $tcd_{(t_0,t_0+\delta)}(C)$ de la cellule C représentée par s dans les 5 minutes suivant l'instant t_0. Une partition des valeurs de *tcd* en n classes $(c_1,..,c_n)$ est nécessaire, chaque classe correspondant à un intervalle (r_i^1,r_i^2) de croissance/décroissance.

La prévision de la suite des *tcd* sur 5 minutes pour tout l'intervalle d'échéance de la prévision peut être effectuée par classification récursive des séquences

$$(e_{-k},..,e_0)\,,\,(e_{-k},..,e_0,e_1)\,,\ldots,\,(e_{-k},..,e_0,e_1,..,e_p)$$

où $e_1,..,e_p$ sont les échos obtenus par l'application des *tcd* prévus.

Nous avons tenté de générer un arbre de décision comme base de connaissances de ce contexte par application de l'algorithme IAD.O. L'ensemble d'exemples de l'apprentissage a été généré à partir de la base de données avec les appariements manuels des échos. Plusieurs partitions des valeurs de *tcd* en classes, ainsi qu'une multitude d'attributs caractérisant les séquences ont été développés. La génération d'une règle fiable n'était cependant pas possible, car tous les noeuds de l'arbre générés par l'algorithme ont été refusés par le test de χ^2, appliqué dans IAD.O pour la vérification de la signification des attributs de test sélectionnés. Par suppression du test de χ^2 la génération d'un arbre peut être forcée, mais les résultats ainsi obtenus semblent aléatoires. Cette observation a été vérifiée par l'introduction d'attributs dont la valeur est aléatoire: ces attributs ont été utilisés pour la classification dans l'arbre, ce qui démontre la faible signification des autres attributs pour la classification des objets. De même une approche d'apprentissage avec des réseaux de neurones (Laroussinie 1990) n'a pas permis la génération d'une base de connaissance classifiant plus de 50% des cas correctement.

L'échec des méthodes de l'apprentissage automatique doit être expliqué par l'insuffisance des attributs définis pour la description des objets. En fait, il semble que les seuls paramètres observables sur l'image radar, sans prise en compte de la topographie, ne permettent pas la détermination en avance du cycle de vie des cellules.

VI.5.4 La prise en compte des contrastes locaux pour la prévision du développement des cellules de pluie convective

L'analyse des influences sur la convection, qui a été menée au début de ce chapitre, a révélée la forme complexe du processus de l'alimentation des cellules. Les résultats obtenus par les approches de l'apprentissage automatique à partir de descriptions des cellules, qui sont basées sur les paramètres observées sur l'image radar seule, mettent en évidence la nécessité de tenir compte de ce processus. Ceci nécessite la connaissance de:

- la zone d'alimentation des cellules,
- les caractéristiques (température et humidité relative) de la zone d'alimentation,
- les caractéristiques atmosphériques (température et humidité relative) de la source froide de la cellule.

Le modèle du développement des cellules de pluie convective, proposé en deuxième partie de ce chapitre, permet d'identifier la zone d'alimentation des cellules. Des données météorologiques supplémentaires aux mesures par radar sont nécessaires afin de déterminer les caractéristiques

de ces zones à la surface et de l'atmosphère en altitude. Les données suivantes ont été disponibles pour cette étude:

- le profil vertical de la température et de l'humidité relative au site du radar à une échelle mi-journalière (radiosondages de la Météorologie Nationale),
- la température et l'humidité au sol à six stations en région parisienne et à quatre stations aux environs de Paris à une échelle trihoraire (réseaux de la Météorologie Nationale).

L'extension spatiale et temporelle de ces mesures ponctuelles est nécessaire pour l'application du modèle. L'extension spatiale à été faite sous les hypothèses suivantes:

- la mesure par radiosondage en altitude (niveau de pression de 500 mb) à Trappes est représentative de la source froide de toute la région couverte par l'image radar,
- les mesures au sol sont représentatives de des régions de la même altitude et des mêmes caractéristiques de la couverture de surface autour des stations météorologiques, jusqu'au prochaines barrières climatologiques (rivières), et à la distance maximale de 25 km.

Les surfaces ainsi définies pour chaque station météorologique sont présentées dans la figure VI.11. Pour chaque région, la mesure de la station correspondante est considérée représentative.

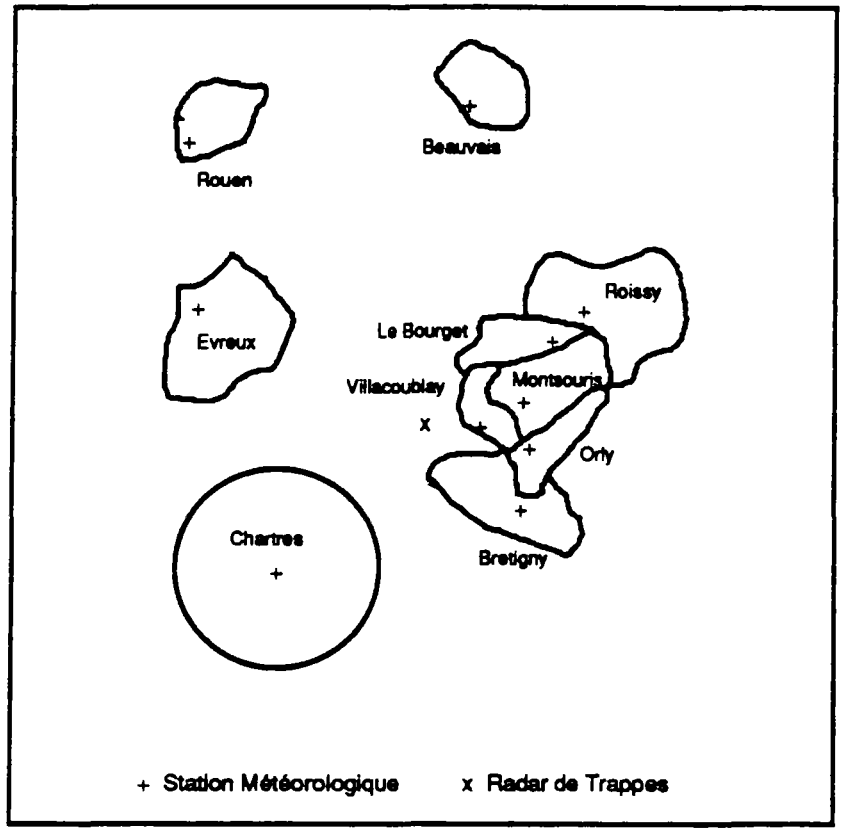


Figure VI.11: Les régions de l'extension spatiale des mesures météorologiques au sol

Pour l'extension temporelle à un instant t_0 de la mesure, les hypothèses suivantes ont été appliquées:

- l'état de l'atmosphère en altitude est déterminé par le dernier radiosondage avant t_0,
- l'état des masses d'air au sol est donné par la dernière mesure avant t_0, s'il n'y avait pas précipitation dans la région entre l'instant de la mesure et t_0, et par la prochaine mesure après t_0 s'il y avait précipitation.

Les figures VI.12.a-VI.12.h présentent le développement de quelques cellules de pluie convective pendant leur passage en région parisienne, ainsi que la répartition spatiale de leur zone d'alimentation sur les six régions définies à la surface. Les cellules choisies sont celles, dont le développement provoque une grande partie de l'erreur de prévision pour certaines pluies (cellules marquées *A-H*, figures A.3.1-A.3.20). Pour les cellules *D*, *E*, *F*, et *H*, l'observation jusqu'à la fin de leur cycle de vie n'est pas possible à cause de fusions avec d'autres cellules.

Pour toutes les cellules présentées, on peut observer une simultanéité des transitions de croissance (arrêt de croissance - inflexion de croissance) avec les transitions de l'alimentation:

Ainsi l'arrêt de croissance des cellules *A*, *B*, et *C* se produit lorsqu'elles sortent des zones urbanisées, représentées par les stations de Montsouris, Orly, Le Bourget, et lorsqu'elles pénètrent dans la plaine de France représentée par la station Roissy. Leur décroissance se poursuit pendant que s'accroît le part d'alimentation de la plaine de France (Roissy).

Ainsi le départ de la croissance des cellules *D*, *F*, *G*, et *H* est concomitant avec le passage dans la zone urbaine du Bourget.

Ainsi le départ de croissance de la cellule *E* a pour origine le passage dans la zone d'alimentation urbaine d'Orly et son maximum (arrêt de croissance) coïncide avec le passage dans la zone plus verte et périurbaine de Bretigny.

La prévision du développement des cellules pendant leur passage d'une certaine région nécessite la connaissance de la capacité de la région d'alimenter la cellule. Une possibilité d'estimer cette capacité est la détermination de l'instabilité verticale de l'atmosphère dans la région, qui est exprimé par les indices d'instabilité. L'indice de Galway est plus utile dans cette application que celui de Showalter, car il exprime l'instabilité entre la couche atmosphérique près du sol et la couche au niveau de 500 mb, et permet ainsi de tenir compte des contrastes locaux entre les masses d'air des basses couches.

Dans les figures VI.12.a-VI.12.h, l'indice de Galway est indiqué pour les régions importantes. Rappelons que l'instabilité, et ainsi la capacité d'alimentation, d'une région est plus grande pour des indices plus petits. On s'attend alors à ce que l'augmentation de l'influence sur une cellule d'une région d'indice relativement petit se traduise par une croissance de la cellule. L'étude des exemples présentés montre, que cette rélation n'est pas toujours valable: pour les cellules *B*, *C*, *G*, et *H*, l'indice de la région d'influence croissante est pratiquement égal à celui de la région de plus grande influence avant la phase de croissance, et pour la cellule *E* il est même plus grand. Pour les cellules *D* et *F* l'indice de la région déterminant le développement de la cellule avant leur phase de croissance est inconnu (car situé hors des régions, pour lesquelles les données sont disponibles), mais peut être supposé d'être plus grand que les indices en région parisienne, comme il est indiqué par les stations aux environs de Paris (Beauvais, Rouen, Chartres et Evreux). L'indice de Galway n'est donc pas représentatif de l'influence sur la croissance des cellules d'une zone d'alimentation.

Nous expliquons le comportement inattendu de la plus grande partie des cellules étudiées par les incertitudes, qui sont comprises dans les hypothèses faites, notamment:

- La validité de l'extension spatiale des mesures est limitée par les contrastes locaux à une échelle très petite. Cinq des six stations en région parisienne sont situés sur des

terrains d'aéroports, qui sont caractérisées par des surfaces libres et plates, et donc différentes des régions urbanisées à proximité. La sixième station se trouve dans le parc de Montsouris, qui présente, malgré sa petite surface, sans doute un microclimat différent des surfaces urbanisées de Paris. En absence de vents près du sol, il ne peut être exclu que ces différences à petite échelle introduisent un biais dans les résultats.

- La validité de l'extension temporelle des mesures dépend de facteurs, qui n'ont pas été pris en compte, notamment du rayonnement solaire entre la mesure et t_0, et des vents à la surface.

Une autre incertitude est introduite par le modèle des zones d'alimentation, car l'hypothèse, que cette zone ne s'étend pas au-delà de 2 km des limites de l'écho de la cellule, peut être fausse en présence d'une convection très forte. Le développement de la cellule *D* par exemple indique une influence de la région Le Bourget sur la cellule déjà 10 minutes avant l'entrée de sa zone d'alimentation dans cette région. Compte rendu de la vitesse de l'advection, qui est d'environ 35 km/h pour cette cellule, cet écart correspond à une distance d'environ 5 km. Si l'on suppose, que l'ascendance verticale des masses d'air soulevé par la convection est d'environ 10 m/s, le début de la croissance environ 5 minutes après le début de l'influence de la région serait attendu.

Rappelons enfin que notre modèle d'alimentation des cellules est très simplificatrice; il ne tient pas compte d'influences comme le cisaillement du vent, les interactions entre les cellules, et d'autres.

A cause de ces incertitudes, l'étude d'un grand nombre de cellules serait indispensable pour la vérification des hypothèses faites. Un nombre suffisant de cellules n'est cependant pas disponible dans cette étude, à cause des contraintes suivantes:

- Le nombre de cellules traversant la région parisienne pendant les pluies convectives étudiées est limité.
- Pour une partie des pluies, les cellules individuelles forment des orages comprenant plusieurs cellules, dont les interactions ne permettent pas l'identification des zones d'alimentation des cellules individuelles (notamment pour les pluies du 24.4.1989, du 26.6.1990 et du 24.9.1990).
- L'observation d'une partie de ces cellules pendant une période assez longue est rendu difficile par des fusions et scissions des échos simples (notamment pour les pluies du 12.9.1990 et du 30.9.1990).
- Pour la pluie du 10.7.1989, les données météorologiques ne sont pas accessibles.

Légende des figures suivantes: rayures des six régions en région parisienne

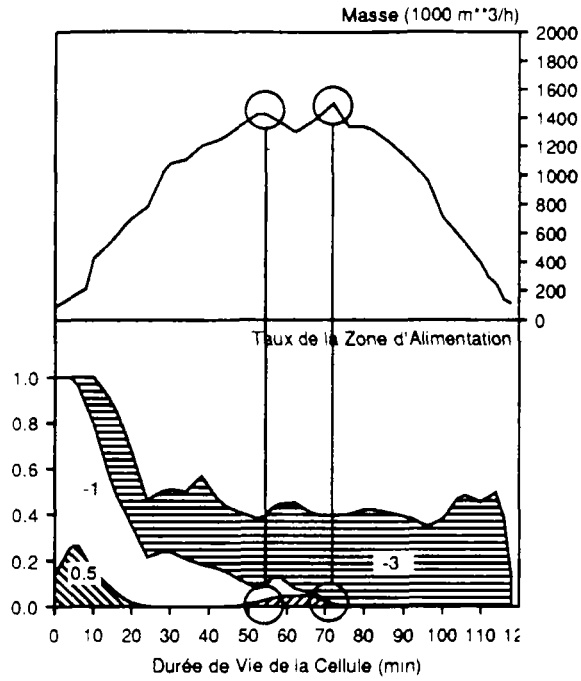


Figure VI.12.a: Cellule *A* (10.5.1989)

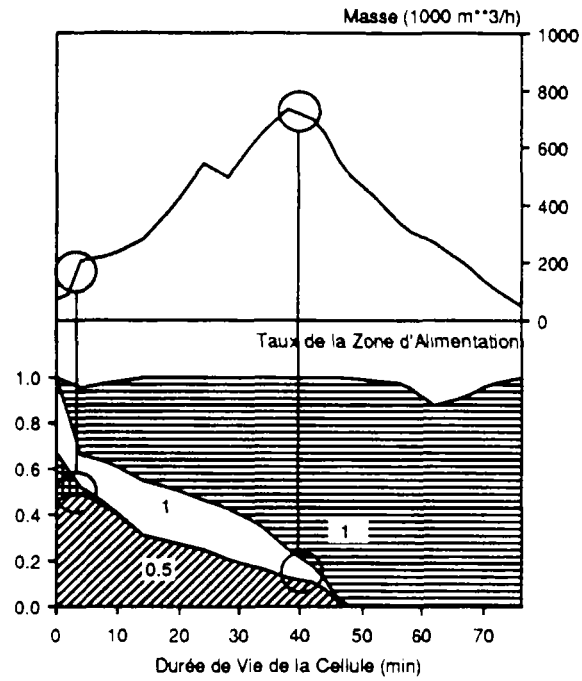


Figure VI.12.b: Cellule *B* (6.6.1989)

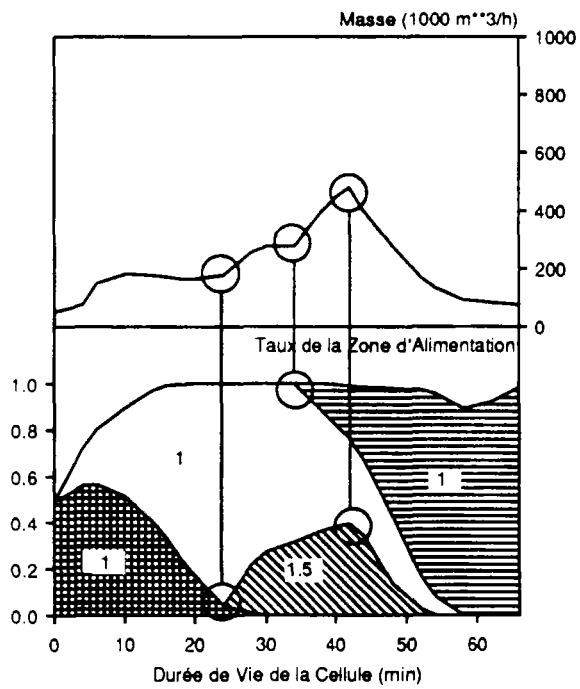


Figure VI.12.c: Cellule *C* (6.6.1989)

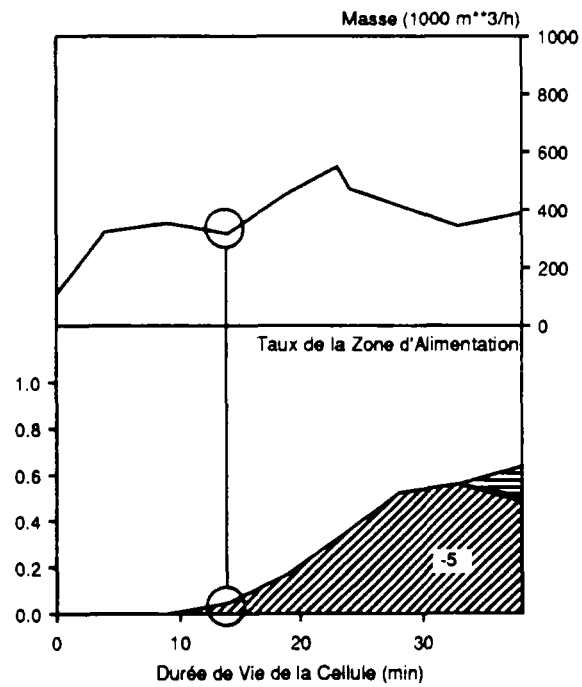


Figure VI.12.d: Cellule *D* (7.8.1989)

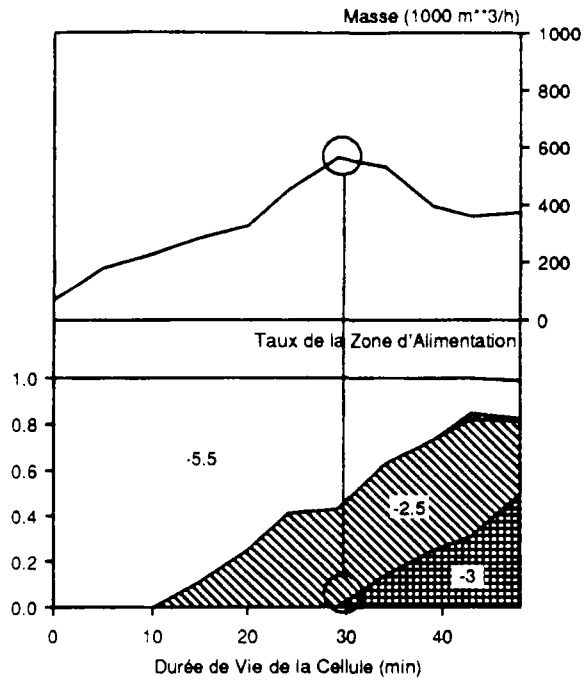


Figure VI.12.e: Cellule *E* (23.4.1990)

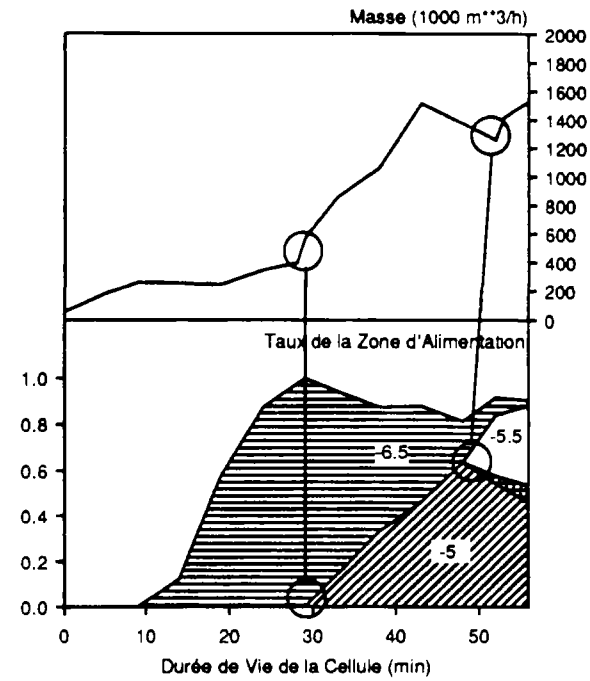


Figure VI.12.f: Cellule *F* (23.4.1990)

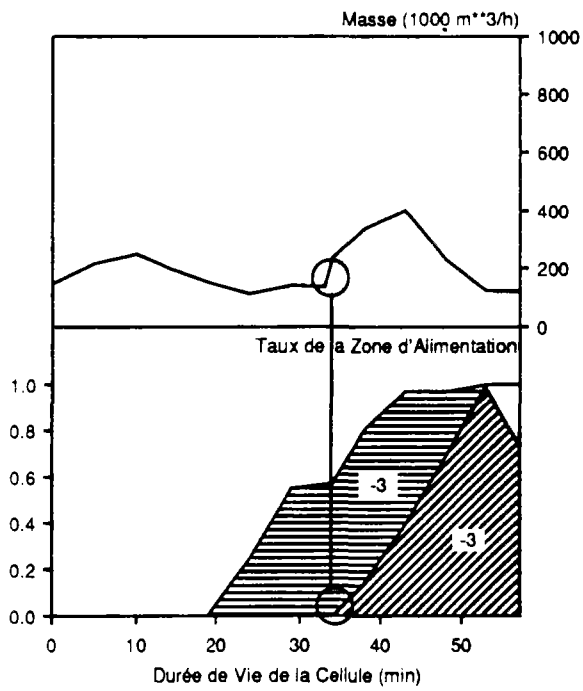


Figure VI.12.g: Cellule *G* (9.6.1990)

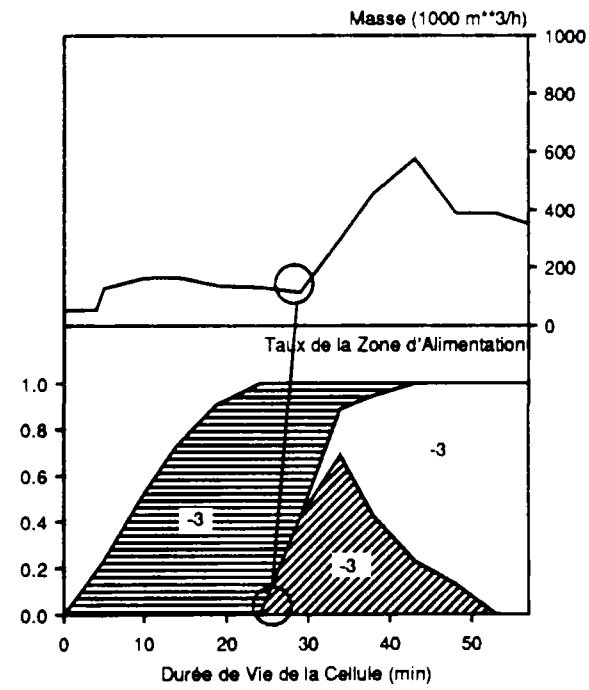


Figure VI.12.h: Cellule *H* (9.6.1990)

VI.6 Conclusion

Nous avons tenté de développer une méthodologie, qui permet la prévision avec prise en compte du développement des cellules pour les pluies convectives. Ce problème a été divisé en deux sous-problèmes:

- la prévision de la suite des taux de croissance/décroissance (*tcd*) pour chaque cellule, avec une résolution de 5 minutes à échéance de la prévision,
- la prévision de la répartition spatiale du développement des cellules.

Un modèle des masses d'air influant sur le développement a été proposé, qui nous a permis d'identifier les zones, qui déterminent le développement. A l'aide de ce modèle, la localisation des pertes ou des gains de masse a pu être précisée statistiquement par l'analyse des échos de la base de données. Il a été démontré, que la croissance des cellules est un processus non-uniforme, qui se concentre sur quelques zones précises autour de la cellule, tandis que la décroissance est un processus uniforme, qui s'étend à toute la cellule.

Basé sur ces résultats, une méthode d'application du *tcd* à la prévision a été développée. Il a été démontré, qu'une amélioration considérable de la prévision est possible, si les *tcd* sont parfaitement connus. Pratiquement les mêmes résultats sont obtenus, si l'on introduit un bruit artificiel de 25%.

Plusieurs techniques de prévision du *tcd* de cellules ont été examinées. Nous avons montré, que les techniques basées sur un simple extrapolation ne peuvent pas apporter une amélioration. L'analyse de quelques cellules exemplaires a mis en évidence, que la prise en compte des contrastes locaux est indispensable pour la prévision du développement.

Dans le cadre de cette étude, ils ne nous était pas possible de proposer une méthode définitive pour la prévision du *tcd* des cellules. Pour une poursuite de cette recherche, il serait nécessaire:

- de disposer d'un réseau météorologique au sol, qui soit assez dense pour pouvoir tenir compte des contrastes locaux dans une région d'une surface assez importante. Le manque de mesures aux environs des zones urbanisées par exemple ne nous a pas permis de déterminer le changement des caractéristiques d'alimentation des cellules s'approchant des zones urbanisées.
- d'avoir accès à une grande quantité de mesures radar de pluies convectives. La détermination de l'influence des caractéristiques au sol au développement des cellules nécessite une population assez importante de cellules, qui ne font pas l'objet de scission ou fusions. Les 12 pluies convectives étudiées dans cette étude n'ont pas permis d'établir des statistiques significatives.
- d'utiliser des modèles fiables des interactions entre la convection et les masses d'air au sol et en altitude. Les indices utilisées dans cette étude introduisent des incertitudes trop importantes. On pourrait supposer que l'emploi de modèles numériques à mesoéchelle, avec une résolution spatiale de 1 à 2 km, promet des résultats plus fiables. L'initiation du modèle serait possible en présence d'un réseaux dense de mesures au sol, ensemble avec l'information fournie par radar. Un tel approche permettrait aussi la vérification du modèle de l'alimentation des cellules.

CONCLUSION GÉNÉRALE

L'étude présentée dans ce mémoire était motivée par le besoin en hydrologie urbaine, de disposer des prévisions des lames d'eau à des échéances courtes, avec une haute résolution spatiale et une haute fiabilité. L'hétérogénéité de la plus grande partie des pluies importantes dans ce domaine nous a amené au développement d'un système structuré de prévision, baptisé PROPHETIA. Le système est basé sur l'observation des structures météorologiques, qui possèdent des caractéristiques uniformes et une persistance dans le temps permettant leur extrapolation à échéance de la prévision. Notre objectif était l'étude des gains accessibles par une telle approche plutôt que le développement d'un système définitif et opérationnel. En travaillant en temps différé, nous avons donc écarté tous problèmes liés aux erreurs de mesure, tel que le calage de la mesure, le traitement des échos fixes, *etc.* Le développement d'une version opérationnelle de PROPHETIA, et son intégration dans un système de gestion des réseaux urbains, nécessiterait le traitement préalable de ces erreurs.

Notre approche est fondée sur une base de données, qui a été soigneusement constituée à l'aide d'un logiciel multi-fonctionnel de traitement des images radar. Outre les images vérifiées et libérées des erreurs visibles, cette base comprend aussi la description des échos des cellules, les définitions manuelles des échos imaginaires, et les appariements manuels des échos. La base de données nous a permis la vérification et l'optimisation de chaque étape de l'algorithme de PROPHETIA. L'analyse détaillée des sources d'erreurs de la prévision nous était possible par la comparaison des résultats de l'algorithme avec les résultats optimaux des définitions manuelles.

Cette analyse a démontré le bien-fondé de l'application de l'apprentissage automatique au problème de l'appariement des échos: le suivi automatique des cellules de pluie est obtenu avec une haute fiabilité. La confrontation avec les appariements optimaux, tels qu'ils ont été définis manuellement, a prouvé que la prévision par PROPHETIA n'est pratiquement pas affectée par des erreurs de l'observation. Grâce au suivi correct des cellules, la performance de PROPHETIA est plus élevée que celle d'autres systèmes de prévision pour toute les pluies caractérisées par une structure hétérogène. L'objectif principal de cette étude a été ainsi atteint.

L'analyse des erreurs de la prévision par PROPHETIA à révélé deux sources principales d'écart entre lames d'eau prévues et lames d'eau réelles:

- pour une partie des pluies frontales, l'identification des cellules par les échos définis avec un seuil fixe et unique est insuffisante,
- pour la plus grande partie des pluies convectives, la prise en compte du développement de la pluie est nécessaire.

Concernant la première source d'erreur, nous avons proposé une modification du système de prévision, qui est basée sur l'identification hiérarchique des cellules à plusieurs niveaux d'intensité. Cet approche combine l'avantage de l'observation à un seuil bas, qui consiste en la plus grande stabilité des cellules observées, avec la prise en compte des caractéristiques individuelles des cellules intenses, où celles-ci ont une persistance dans le temps. Un test a montré que la technique proposée permet une amélioration considérable de la qualité de prévision pour certaines situations. Le cadre limité de cette étude ne nous a cependant pas permis d'optimiser la technique proposée. Ainsi plusieurs questions persistent: la sélection des seuils nécessaires, le mode optimal de la définition des échos imaginaires pour les seuils hauts, et la validité au niveau haut de l'appariement par l'arbre de décision généré uniquement à partir des exemples d'appariements au niveau bas.

Bien que la prise en compte des caractéristiques individuelles des cellules intenses ait une importance pour certaines pluies, la source principale d'erreur de la prévision par extrapolation est l'hypothèse d'absence du développement local de la pluie convective à l'horizon de cette prévision. Le problème de la prévision de ce développement a été divisé en deux sous-problèmes: la prévision du taux de développement, et la prévision de la localisation de ce développement par rapport à la cellule. A l'aide de la base de données, qui nous fournit la description exacte de bon nombre de cellules de pluie convective pendant leur cycle de vie, nous avons pu étudier de façon très détaillée les facteurs influant sur leur développement.

Nous avons proposé un modèle des cellules, qui permet l'identification de la zone d'alimentation déterminant leur développement. A l'aide de ce modèle, la répartition de l'intensification et de l'affaiblissement de la masse pluvieuse des cellules lors de leur croissance et décroissance a été déterminée statistiquement. Les résultats de cette analyse ont amené à une méthode de prévision, qui tient compte du développement, s'il est connu d'avance. Nous avons démontré, que la connaissance de l'ordre de grandeur du taux de croissance/décroissance des cellules peut apporter un gain considérable par rapport à la prévision basée sur la seule advection.

La recherche de méthodes de prévision du taux de croissance/décroissance des cellules était le but de la dernière partie de cette étude. Nous avons mis en évidence, que cette prévision est impossible sans prise en compte des contrastes locaux de la basse atmosphère. L'analyse de quelques cas exemplaires a démontré, que la méthodologie développée, qui est basée sur la recherche des caractéristiques de l'alimentation des cellules, est prometteuse. La détermination des caractéristiques des masses d'air près du sol nécessite cependant des informations complémentaires, qui étaient insuffisamment disponibles pour cette étude. Un réseau plus dense de mesures au sol serait souhaitable, ainsi qu'un modèle plus développé, afin d'avoir une meilleure représentativité spatiale des conditions thermodynamiques au sol.

La poursuite de cette recherche nécessite l'approfondissement de l'étude du mécanisme d'alimentation des cellules. Là, encore, se pose la question de l'échelle de l'observation: l'attention pourrait être portée aux noyaux intenses (c'est à dire les cellules convectives au sens strict donné par les physiciens de l'atmosphère) plutôt qu'au cellules définies par un niveau d'intensité de pluie relativement bas. Mais la réduction de la stabilité et aussi de la qualité de caractérisation des cellules définies aux niveaux plus hauts impose des limites à ce choix, sans pourtant résoudre le problème pour tous les cas: le phénomène de deux cellules convectives au sein d'une même cellule par exemple peut être observé à tous les niveaux. Une nouvelle réflexion sur l'utilité de la technique du seuillage pour l'étude du développement est alors indispensable.

La limite principale des systèmes de prévision par extrapolation est imposée par la durée de vie des structures observées. Toute prévision d'une échéance dépassant cette durée de vie est en effet aléatoire. La prise en compte du développement des cellules ne permet pas de franchir cette limite, qui peut restreindre la validité de la prévision à une échéance bien inférieure d'une heure dans certaines situations convectives. On observe cependant souvent une organisation des cellules dans des structures plus larges, dont l'identification et la description permettraient une extension de l'horizon de la prévision au-delà de la durée de vie des cellules individuelles. L'exactitude spatiale d'une telle prévision est certes limitée, mais elle peut être utile pour une estimation préalable des lames d'eau attendues, qui est après concrétisée par la prévision détaillée. Il faut cependant vérifier, si la "fenêtre", qui représente une image radar, est assez large pour l'identification de telles supra-structures. L'utilisation des images radar composites, telles qu'elles sont produites par la Météorologie Nationale à partir des mesures du réseau ARAMIS, peut s'avérer plus utile dans ce contexte.

L'application d'une telle technique nécessite la résolution de plusieurs problèmes, notamment de l'identification météorologique des supra-structures par leur description et par leur appariement, et de la transition entre la prévision indicative et la prévision détaillée. Dans le cadre de cette étude, l'approfondissement de cette problématique n'était pas possible. Nous espérons, que la continuation de notre travail aboutira à des solutions de ces problèmes dans un avenir proche.

RÉFÉRENCES BIBLIOGRAPHIQUES

Andrieu, H., 1986. Interprétation de mesures du radar Rodin de Trappes pour la connaissance en temps réel des précipitations en Seine-St.Denis et Val-de-Marne. Thèse de Docteur-Ingénieur, École Nationale des Ponts et Chaussées, Paris.

Agostini-Blanchet, B., 1988. Incertitudes liées à la mesure de la pluie. Rapport de DEA, École Nationale des Ponts et Chaussées, Paris.

Arya, S.P., 1988. Introduction to micrometeorology. Academic Press, New York.

Austin, G.L., et Bellon, A., 1974. The use of digital weather radar records for short-term precipitation forecasting. Quart. J. R. Met. Soc., 100, pp. 658-664.

Austin, G.L., et Bellon, A., 1982. Very-short-range forecasting of precipitation by the objective extrapolation of radar and satellite data. *In*: Browning (éd.) Nowcasting. Academic Press, London.

Austin, G.L., Kilambi, A., et Bellon, A., 1990. On the relative merit of forecasting precipitation by image extrapolation and mesoscale numerical models. Conf. on operational Precip. Estimation and Prediction, Anaheim, California.

Austin, P.M., et Houze, R.A., 1972. Analysis of the structure of precipitation patterns in New England. J. Appl. Met., Vol. 11, pp. 926-935.

Avissar, R., 1990. The impact of soil moisture and vegetation on evapotranspiration and regional atmospheric processes. Prepr. 8th Conf. on Hydrometeorology, Kananaskis Park.

Barklay, P.E., et Wilk, K.E., 1970. Severe thunderstorm radar echo motion and related weather events hazardous to aviation operations. ESSA Techn. Memo No. 46, NSSL.

Barr, A., et Feigenbaum, E.A., 1981. The Handbook of Artificial Intelligence. W. Kaufmann.

Battan, L.J., 1973. Radar observation of the atmosphere. The University of Chicago Press, Chicago.

Bellon, A., et Austin, G.L., 1978. The evaluation of two years of real-time operation of a short-term precipitation forecasting procedure (SHARP). J. of Appl. Met., Vol. 17, pp. 1778-1787.

Bellon, A., et Austin, G.L., 1984. The accuracy of short-term radar rainfall forecasts. J. of Hydrol., Vol. 70, pp. 35-49.

Bjerkaas, C.J., et Forsyth, D.E., 1980. An automated real-time storm analysis and storm tracking program (WEATRK). NSSL Technical Report No. AFGL-TR-80-0316, Norman, Oklahoma.

Blackmer, R.H., Duda, R.O., et Reboh, R., 1973. Application of pattern recognition techniques to digitized weather radar data. Report N. 36072, Stanford Research Institute, Menlo Park, California.

Blanchet, B., Neumann, A., Jacquet, G., et Andrieu, H., 1989. Improvement on rainfall measurements due to accurate synchronisation of raingauges and due to advection use in calibration. Int. Symp. on hydrol. Appl. of Weather Radar, Salford.

Bougeault, P., Ducrocq, V, Imbard, M., et Tardieu, J., 1989. Prévoir les très violents orages. La Recherche, No. 216, pp. 1526-1528.

Breiman, L., Friedman, J.H., Olshen, R.A., et Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Pacific grove.

Brémaud, P., 1991. Suivi et prévision automatique du déplacement des nuages précipitants par radar météorologique. Thèse de Docteur en Physique de l'Atmosphère, Université Blaise Pascal, Clermont-Ferrand.

Breuer, L.J., 1976. Anmerkungen zur Erforschung von Niederschlagsstrukturen mit elektronischen Pulsmeßgeräten. Kleinheubacher Berichte, Nr.19, FTZ Darmstadt.

Browning, K.A., 1978. Meteorological applications of radar. Rep. Prog. Phys., Vol. 41, pp. 779-806.

Browning, K.A., 1979. The FRONTIERS plan: a strategy for using radar and satellite imagery for very-short-range precipitation forcasting. Met. Off. Radar Res. Lab. Res. Report, No. 11.

Browning, K.A., 1980. Local weather forecasting. Proc. R. Soc. London, A371, pp. 179-211.

Browning, K.A., 1985. Conceptual models of precipitation systems. The Meteorological Magazine, Vol. 114, No. 1359.

Browning, K.A., et Collier, C.G., 1982. An integrated radar-satellite nowcasting system in the UK. Browning (éd.) : Nowcasting. Academic Press, London.

Bundy, A., Silver, B., et Plummer, D., 1985. An analytical comparison of some rule-learning programs. Artificial Intelligence, Vol. 27, pp. 137-181.

Carbonell, J.G., Michalski, R.S., et Mitchell, T.M., 1983. An overview of machine learning. *In*: Michalski et *al.* (éds.), Machine Learning - an artificial intelligence approach. Springer, New York.

Cestnik, B., Konenko, I., et Bratko, I., 1987. ASSISTANT 86: A knowledge-elication tool for sophisticated users. *In*: Bratko et Lavrac (éds.), Progress in Machine Learning. Sigma Press, Wilmslow, Yug..

Changnon, S.A., 1976. Effects of urban areas and echo merging on radar echo behavior. J. Appl. Met., Vol. 15, pp. 561-570.

Changnon Jr., S.A., Semonin, R.G., et Huff, F.A., 1976. A hypothesis for urban rainfall anomalies. J. Appl. Met., Vol 15, pp. 544-560.

Chalon, J.P., 1978. Dynamique des nuages convectifs. Influence du cisaillement de vent sur l'évolution des cumulonimbus. La Météorologie, VI° série, n° 13.

Cheng, J., Fayyad, U.M., Irani, K.B., et Qian, Z., 1988. Improved decision trees: a generalized version of ID3. 5th Int. Conf. on Machine Learning, Ann Arbor.

Cheze, J.-L., Tardieu, J., et Gilet, M., 1991. The French weather radar network. 25th Conf. Radar Met., Paris.

Cluckie, I., Tilford, K., et Shepherd, G., 1989. Radar signal quantisation and its impact on rainfall-runoff models. Int. Symp. on hydrol. Appl. of Weather Radar, Salford.

Collier, C.G., 1981. Objective rainfall forecasting using data from the United Kingdom weather radar network. IAMAP Symp., Hamburg.

Collier, C.G., 1989. Applications of weather radar systems. Ellis Horwood, Chichester.

Crane, R.K, 1979. Automatic cell detection and tracking. IEEE Trans. Geosc. Elec., Vol. GE-17, No. 4, pp. 250-262.

Dalezios, N.R., 1988. Objective rainfall evaluation in radar hydrology. ASCE J. of Water Planning and Management, Vol. 114, No. 5.

Damant, C., Austin, G.L., Bellon, A., Osseyrane, M., et Nguyen, N., 1983. Radar rain forecasting for wastewater control. J. of hydrol. Eng., Vol. 109, No. 2, pp. 293-297.

Davis, R., et King, J.J., 1977. An overview over production systems. *In*: Elcock et Michie (éds.), Machine Intelligence 8, Ellis Horwood, Chichester.

Delattre, J.M., Bachoc, A., et Jacquet, G., 1986. Performance of hardware components for real time management of sewer systems. *In*: Torno, H.C., Marsalek, J., et Desbordes, M. (éds.), Urban Runoff Pollution, Springer.

Denoeux, T., 1989. Fiabilité de la prévision de pluie par radar en hydrologie urbaine. Thèse de Docteur en Sciences et Techniques de l'Environnement, École Nationale des Ponts et Chaussées, Paris.

Denoeux, T., Einfalt, T., et Jacquet, G., 1990. Determination in real time of the reliability of radar rainfall forecasts. J. of Hydrol., Vol. 122, pp. 353-371.

Dietterich, T.G., 1989. Limitations on inductive learning. 6th Int. Workshop on Machine Learning, Ithaka, N.Y..

Dong-Jun, S., Krajewski, W.F., et Bowles, D.S., 1990. Stochastic interpolation of rainfall data from rain gages and radar using cokriging - 1. Design of experiments. Water Resources Research, Vol. 26, No. 3, pp. 469-477.

Doswell, C.A., 1986. Short-range forecasting. *In*: Ray, P. (éd.), Mesoscale Meteorology and Forecasting, AMS, Boston.

Einfalt, T., 1988. Recherche d'une méthode optimale de prévision de pluie par radar en hydrologie urbaine. Thèse de Docteur en Sciences et Techniques de l'Environnement, École Nationale des Ponts et Chaussées, Paris.

Einfalt, T., et Denoeux, T., 1987. Radar rainfall forecasting for real-time control of a sewer system. 4th Int. Conf. on Urban Storm Drainage, Lausanne.

Einfalt, T., Denoeux, T., et Jacquet, G., 1990. A radar rainfall forecasting method designed for hydrological purposes. J. of Hydrol., Vol. 114, pp. 229-244.

Elvander, R.C., 1976. An evaluation of the relative performance of three weather radar echo forecasting techniques. 17th Conf. Radar Met., Seattle.

Fisher, D.H., 1987. Knowledge acquisition via incremental conceptual clustering. Machine Learning, Vol. 2, pp. 139-172.

Fohmann, L., 1985. Wissenserwerb und maschinelles Lernen. *In*: Savory, S. E. (éd.), Künstliche Intelligenz und Expertensysteme, Oldenbourg, München, Wien.

Foufoula-Georgiou, E., et Kumar, P., 1990. Monitoring and short-term forecasting of precipitation fields using fourier domain shape analysis methods. Conf. on operational Precip. Estimation and Prediction, Anaheim, California.

Frérot, A., 1987. Procédures d'optimisation des consignes de gestion d'un réseau d'assainissement automatisée. Thèse de Docteur-Ingénieur, École Nationale des Ponts et Chaussées, Paris.

Ganascia, J.-G., 1987. AGAPE et CHARADE: deux mécanismes d'apprentissage symbolique appliqués à la construction de bases de connaissance. Thèse d'État, Université Paris-Sud.

Golding, B.W., 1987. The UK meteorological office mesoscale model. Boundary-layer Met., Vol. 41, pp. 97-107.

Hall, C.D., 1986. Verification of precipitation forecasts from the UK operational limited-area model. *In*: Matsuno, T. (éd.), Short- and Medium-Range Numerical Weather Prediction, Tokyo.

Harrold, T.W., 1973. Mechanisms influencing the distribution of precipitation within baroclinic disturbances. Quart. J. R. Met. Soc., Vol. 99, pp. 232-251.

Harrold, T.W., et Austin, P.M., 1974. The structure of precipitation systems - a review. J. de Recherches Atmosphériques, Vol. 8, pp. 41-57.

Harrold, T.W., Nicholass, C.A., et Collier, C.G., 1975. The measurement of heavy rainfall over small catchments using radar. Bull. des Sc. Hydrol., XX, 1, pp. 69-76.

Hayes-Roth, F., Waterman, D.A., et Lenat, D.B. (éds.), 1983. Building expert systems. Addison-Wesley.

Herremans, L., 1990. L'impact des problèmes pluviaux sur l'alimentation en eau potable. La Gestion de l'Eau, Colloque Européen, Paris.

Hitschfeld, W.F., 1986. The invention of radar meteorology. Bull. Am. Met. Soc., Vol. 67, No. 1.

Houze, R.A., et Austin, P.M., 1972. Analysis of precipitation patterns in New England. J. Appl. Met., Vol. 11, pp. 926-935.

Hudlow, M., Smith, J., Walton, M., et Shedd, R., 1989. NEXRAD - New era in hydrometeorology in the United States. Int. Symp. on hydrol. Appl. of Weather Radar, Salford.

Huff, F.A., 1986. Urban hydrometeorology review. Bull. Am. Met. Soc., Vol. 67, No. 6, pp. 703-711.

Huff, F.A., Changnon, S.A., et Vogel, J.L., 1980. Convective rain monitoring and forecasting system for an urban area. 19th Conf. Radar Met., Miami.

Hunt, E.B., Marin, J., et Stone, P.J, 1966. Experiments in induction. Academic Press, New York.

Imbard, M., Juvanon du Vachat, R., Joly, A., Durand, Y., Craplet, A., Geleyn, J.F., Audoin, J.M., Marie, N., et Pairin, J.M., 1986. The PERIDOT fine-mesh numerical weather prediction system - description, evaluation and experiments. *In*: Matsuno, T. (éd.), Short- and Medium-Range Numerical Weather Prediction, Tokyo.

Jacquet, G., et Neumann, A., 1991. Analyse à postériori des résultats de la prévision automatique. Rapport CERGRENE, ENPC.

K.N.M.I. - Gemeentewerken Rotterdam, 1989. Weerkundige Begleiding van de Rioolwaterbeheersing in de Gemeente Rotterdam. Rapport de projet.

Khelil, A., 1990. Adaptation of an expert system for the real-time control of a sewerage network: case of Bremen left side of the Weser. 5th Int. Conf. on Urban Storm Drainage, Osaka.

Kodratoff, Y., 1986. Leçons d'Apprentissage Symbolique Automatique. Cepadues-Editions, Toulouse.

Kohonen, T., 1988. An introduction to neural computing. Neural Networks, Vol. 1, pp. 3-16.

Laroussinie, F., 1990. Application des k-plus proches voisins (knn) et des réseaux de neurones à la prédiction de l'évolution d'échos radar de nuages de pluie. Rapport de stage, École Nationale Supérieure d'Informatique et de Mathématiques Appliquées de Grenoble.

Lebowitz, M., 1987. Experiments with incremental concept formation: UNIMEM. Machine Learning, Vol. 2, pp. 103-138.

Lippmann, R.P., 1987. An introduction to computing with neural nets. IEEE ASSP Magazine, April 1987.

Lipschutz, R.C., Pratte, J.F., et Smart, J.R., 1986. An operational Z_{DR}-based precipitation type/intensity product. 23rd Conf. on Radar Met. and Cloud Physics, Colorado.

Lopez, R.E., Blanchard, D.O., Rosenfeld, D., Hiscox, W.L., et Casey, M.J., 1984. Population characteristics, development processes and structure of radar echos in south Florida. Monthly Weather Review, Vol. 112, pp. 56-75.

Marshall, J.S., et Palmer, W.McK., 1948. The distribution of raindrops with size. J. of Met., Vol. 5, No. 4, pp. 165-166.

Matheus, C.J., et Hohensee, W.E., 1987. Learning in artificial neural systems. Computer Intelligence, Vol. 3, pp. 283-294.

McDermott, D., 1987. A critique of pure reason. Computer Intelligence, Vol. 3, pp. 151-160.

Michalski, R.S., 1983. A theory and methodology of inductive learning. Artificial Intelligence, Vol. 20, pp. 111-161.

Michalski, R.S., et Stepp, R., 1982. Revealing conceptual structure in data by inductive inference. *In*: Hayes et Michie (éds.), Machine Intelligence 10. Ellis Horwood, Chichester.

Michalski, R.S., et Stepp, R., 1983. Learning from observation: conceptual clustering. *In*: Michalski et *al.* (éds.), Machine Learning - an artificial intelligence approach. Springer, New York.

Michie, D., 1986. Machine intelligence: the first 2400 years. *In*: Michie (éd.), On Machine Intelligence, 2e éd.. Ellis Horwood.

Minsky, M., 1975. A framework for representing knowledge. *In*: Winston (éd.), The Psychology of Computer Vision. McGraw-Hill, New York.

Mitchell, T., 1982. Generalization as search. Artificial Intelligence, Vol. 18, pp. 203-226.

Newell, A., et Simon, H.A., 1963. GPS, a program that simulates human thought. *In*: Feigenbaum et Feldman (éds.), 1981, Computers and Thought, McGraw-Hill, New York.

Newsome, D. et Collier, C., 1989. COST-73: Weather radar in western Europe - possible hydrological applications. Int. Symp. on hydrol. Appl. of Weather Radar, Salford.

Nilsson, N.J., 1980. Principles of Artificial Intelligence. Morgan Kaufmann, Los Altos.

Ostlund, S.S., 1974. Computer software for rainfall analyses and echo tracking of digitized radar data. Tech. Memo NOAA-74052009, Boulder, Colorado.

Pedelaborde, P., 1957. Le climat du bassin Parisien. Génin, Paris.

Pedelaborde, P., 1982. Introduction à l'étude scientifique du climat. SEDES, Paris.

Quinlan, J.R., 1984. Learning efficient classification procedures and their application to chess end games. *In*: Michalski et *al.* (éds.), Machine Learning - an artificial intelligence approach. Springer, New York.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learning, Vol. 1, pp. 81-106.

Quinlan, J.R., 1987a. Decision trees as probabalistic classifiers. 4th Int. Workshop on Machine Learning, Irvine, Cal..

Quinlan, J.R., 1987b. Simplifying decision trees. Int. J. Man-Machine Studies, Vol. 27, pp. 221-234.

Quinlan, J.R., 1988. Generating production rules from decision trees. 10th Int. Joint Conf. on Artificial Intelligence, Milan.

Quinlan, J.R., 1989. Unknown attribute values in induction. 6th Int. Workshop on Machine Learning, Ithaka, N.Y..

Rogers, R.R., 1979. A short course in cloud physics. Pergamon Press, Oxford.

Rogers, R.R., 1984. A review of multiparameter radar observations of precipitation. Radio Science, Vol. 19, No. 1, pp. 23-36.

Rosenfeld, D., 1987. Objective method for analysis and tracking of convective cells as seen by radar. J. Atmosph. and Oceanic Tech., Vol. 4, pp. 422-434.

Russchenberg, H.W.J., et Baptista, J.P.V., 1990. Doppler-polarimetric research of precipitation with the Delft atmospheric research radar. Rapport ESA.

Samuel, A.L., 1963. Some studies in machine learning using the game of checkers. *In*: Feigenbaum et Feldman (éds.), Computers and Thought, McGraw-Hill, New York.

Sauvageot, H., 1982. Radarmétéorologie. Eyrolles, Paris.

Schilling, W., et Petersen, S.O., 1987. Real time operation of urban drainage systems - validity and sensitivity of optimization techniques. *In*: Beck, M.B. (éd.), Systems Analysis in Water Quality Management, Pergamon Press, Oxford.

Schlimmer, J.C., et Fisher, D., 1986. A case study of incremental concept induction. 5th Nat. Conf. on Artificial Intelligence, Philadelphia, PA..

Schlimmer, J.C., et Granger, R.H., 1986. Incremental learning from noisy data. Machine Learning, Vol. 1, pp. 317-354.

Spangler, S., Fayyad, U.M., et Uthurusamy, R., 1989. Induction of decision trees from inconclusive data. 6th Int. Workshop on Machine Learning, Ithaka, N.Y..

Sutton, G., et Conway, B.J., 1989. Automatic precipitation nowcasts based on satellite and radar imagery with numerical model products. COST-73 Symp..

Tatehira, R., Makino, Y., et Hitsuma, M., 1981. Combined use of radar with mesoscale surface network for very-short-range prediction of precipitation. IAMAP Symp., Hamburg.

Tilford, K.A., et Cluckie, I.D., 1990. Hydrological utilisation of weather radar data in the United Kingdom. Int. Symp. on Remote Sensing of Precip. and its Appl. to Hydrol., Sao Paulo.

Torterotot, F., 1988. Les limites de la prévision météorologique. Bull. de la Soc. Fr. de Phys., No. 68, pp. 3-7.

Triplet, J.P., et Roche, G., 1977. Météorologie générale. École Nationale de la Météorologie, 2e éd., Paris.

Tsonis, A.A., et Austin, G.L., 1981. An evaluation of extrapolation techniques for the short-term prediction of rain amounts. Atmosphere-Ocean 19, No. 1, pp. 54-65.

Utgoff, P.E., 1988. ID5: An incremental ID3. 5th Int. Conf. on Machine Learning, Ann Arbor.

Utgoff, P.E., 1989. Incremental induction of decision trees. Machine Learning, Vol. 4, 161-186.

Van de Velde, W., 1989. IDL, or taming the multiplexer. 4th Europ. Working Session on Learning, Montpellier.

Volle, M., 1985. Analyse des données. 3e éd, Economica, Paris.

Viers, G., 1968. Éléments de climatologie. Nathan, Paris.

Vogel, J.L., 1980. Real time measurement of convective precipitation over an urban area. Symp. hydrol. Forecasting, Oxford.

Waterman, D.A., et Hayes-Roth, F., 1978. Pattern-Directed Inference Systems. Academic Press, New York.

Weizenbaum, J., 1976. Computer Power and Human Reason. Freeman, San Francisco, London.

Winston, P.H., 1975. Learning structural descriptions from examples. *In*: Winston (éd.), The Psychology of Computer Vision. McGraw-Hill, New York.

Winston, P.H., 1977. Artificial Intelligence. Addison-Wesley, Reading. 2e éd. 1981, traduit en français: Intelligence artificielle, InterEditions, Paris, 1988.

Wirth, J., et Catlett, J., 1988. Experiments on the costs and benefits of windowing in ID3. 5th Int. Conf. on Machine Learning, Ann Arbor.

Wilson, J.W., et Brandes, E.A., 1979. Radar measurement of rainfall - a summary. Bull. Amer. Met. Soc., Vol. 60, No. 9, pp. 1048-1058.

Zawadzki, I., 1984. Factors affecting the precision of radar measurements of rainfall. 22nd Conf. on Radar Met., Zurich.

Zawadzki, I., Torlaschi, E., et Sauvageau, R., 1981. The relationship between mesoscale thermodynamic variables and convective precipitation. J. Atmospheric Sci., Vol. 38, No. 8, pp. 1535-1540.

Zawadzki, I., et Calheiros, R.V., 1987. Reflectivity-rain rate relationships for radar hydrology in Brazil. J. of Climate and Appl. Met., Vol. 26, No. 1, pp. 118-132.

Zipser, E.J., 1990. Rainfall predictability: when will extrapolation-based algorithms fail? 8th Conf. on Hydrometeorology, Kananaskis Park.

Index terminologique

ADDAPP 90
ADDINI 86
Appariement non reconnu 93
Apprentissage automatique de connaissances . 35
Arbre 44
Arbre de décision 44
Attribut
  linéaire 33
  nominal 33
  structuré 33
Attribut de test 44
Attribut incomplètement valué 54
Attributs historiques 64
Bande brillante 17
Bandes pluvieuses étroites 8
Bandes pluvieuses larges 8
Base de connaissances 34
Bon appariement 93
CAPPI 14
Cellule 58
Classe 44
Concept 35
Connaissances procédurales 32
Connaissances spécifiques 32
Contexte de classification 44
Contexte de connaissances 33
CORRCROIS 123
Corrélation croisée 26
Échelle synoptique 6
Écho 59
Échos de pluie 59
Échos de sol 15, 59
Échos fixes 17
Échos imaginaires utiles 61
Ensemble d'échos 61, 83
Ensemble d'échos imaginaires 60, 85
Ensemble d'échos simples 60
Ensemble d'exemples d'apprentissage 45
Entropie 46
Facteur de réflectivité 13
Frames 36
Front
  anabatique 8
  chaud 7
  froid 7
  katabatique 8
  occlus 7
Fusions 60
Groupement conceptuel 43
IAD.L 74
IAD.O 68
IAD.S 72
ID3 . . . . . 44, 45
Image radar 14
Indices d'instabilité 135
Instances 42
Langage de description 34
Mauvais appariement 93
Mesoéchelle 6
Mesure d'impureté 46
Microéchelle 6
Moteur d'inférences 37
Neurones 39
Noeud terminal 44
Non-instances 42
Objets 33
Ordre de complexité 75
PERSIST 121
Pixel 14
PPI 14
Presque-instance 65, 86
Problème admissible 34
Productions 37
Propagation anormale 15
PROPHETIA 101
PROPHETIA.II 144
Racine 44
Scissions 60
SCOUT II.0 121
Séquence d'échos 59
Séquence stricte d'échos 62
Seuils souples 53
Source chaude 137
Source froide 137
Système à base de connaissances 34
  système complet 34
  système consistant 34
Taux de croissance/ décroissance 133
Techniques de prévision
  automatisées non structurées 25
  automatisées structurées 26
TMP 107
VIL 14
Zone d'alimentation 137

Glossaire

$|X|$ ..... cardinal d'un ensemble X

$X \setminus X'$ ..... complément de X' dans X

$\boldsymbol{A}, \boldsymbol{b}$ ..... paramètres de la relation Z-R

$\boldsymbol{A}$ ..... ensemble d'attributs

$\vec{a}(o)$ ..... liste des valeurs de l'objet o

$\boldsymbol{AC}$ ..... apprentissage automatique de connaissances

$\boldsymbol{ADD}$ ..... arbre de decision

$\boldsymbol{adv}_{moy}(I)$ .. advection moyenne des cellules sur l'image I

$\boldsymbol{adv}_k(e_1,..,e_n)$ advection moyenne d'une cellule observée sur k pas de temps

$\boldsymbol{a}^n$ ..... attribut de test du noeud n

$\boldsymbol{BC}$ ..... base de connaissances

$\boldsymbol{c}$ ..... classe d'un contexte de classification

$\boldsymbol{C}$ ..... cellule de pluie

$\boldsymbol{cg}(e)$ ..... centre de gravité de l'écho e

$\boldsymbol{c}^n$ ..... classe associée à un noeud terminal n

$\boldsymbol{CT}$ ..... contexte de connaissances

$\boldsymbol{CT}_{EX}$ ..... contexte exemplaire (classification d'animaux)

$\boldsymbol{CT}_{AP}$ ..... contexte de classification de l'appariement d'échos

$\boldsymbol{EX}$ ..... ensemble d'exemples

$\boldsymbol{d}_{max}$ ..... distance maximale entre deux pixels connectés

$\delta(e_1,e_2)$ .... distance entre les échos e_1 et e_2

$\delta_{max}(I_1,I_2)$. distance maximale de la définition des échos imaginaires

$\Delta\Phi(n,a_i)$ .. mesure de la réduction de l'impureté par l'attribut a_i

$\Delta\Phi'(n,a_i)$ .. mesure modifiée de la réduction de l'impureté par l'attribut a_i

$\Delta_P t$ ..... intervalle d'échéance de la prévision

$\Delta_O t$ ..... intervalle de l'observation

$\boldsymbol{e}$ ..... écho

$\boldsymbol{E}(I)$ ..... ensemble des échos de l'image I

$\boldsymbol{es}$ ..... séquence des échos

$\boldsymbol{E}_s(I)$ ..... ensemble des échos simples de l'image I

$\boldsymbol{E}_i(I)$ ..... ensemble d'échos imaginaires de l'image I

$\boldsymbol{F}_{ADD}$ ..... fonction de classification définie par l'arbre ADD

$\boldsymbol{f}^n$ ..... fonction de test associée au noeud n

$\boldsymbol{I}$ ..... image radar

$\boldsymbol{IC}$ ..... système d'interprétation de connaissances

$\boldsymbol{I}_G$ ..... indice d'instabilité de Galway

$\boldsymbol{K}$ ..... système à base de connaissances

$\boldsymbol{n}$ ..... noeud d'un arbre de decision

$\boldsymbol{o}$ ..... objet d'un contexte de connaissances

$\boldsymbol{O}$ ..... ensemble d'objets

$O(f)$ ..... ordre de complexité de la fonction f

$\boldsymbol{p}^n$ ..... facteur de probabilité associé à un noeud terminal n

Φ ..... mesure d'impureté

$\boldsymbol{R}$ ..... intensité de la précipitation

ρ ..... coéfficient de la correlation croisée

$\boldsymbol{r}_s$ ..... seuil de réflectivité de la définition des échos

$\boldsymbol{s}^+, \boldsymbol{s}^-$ ..... seuils souples

$\boldsymbol{SE}(I_1,..,I_k)$. ensemble des séquences strictes d'échos des images $I_1,..,I_k$

$\boldsymbol{tcd}_{(t_i,t_j)}(C)$ taux de croissance/ décroissance de la cellule C dans l'intervalle (t_i,t_j)

$\mathbf{TMP}(p)$ .. taux d'erreur de la prévision p

$\mathbf{TMP}(ev)$ .. taux moyen d'erreur des prévisions pour un événement ev

$\boldsymbol{V}$ ..... volume d'air

$\boldsymbol{X}_A$ ..... ensemble d'exemples de l'apprentissage

$\boldsymbol{X}^n$ ..... ensemble d'exemples associé au noeud n

$\boldsymbol{X}_T$ ..... ensemble d'exemples du test interne de signification des attributs

$\boldsymbol{X}^+, \boldsymbol{X}^-$ .... ensemble d'exemples positifs/négatifs

$\boldsymbol{Y}$ ..... ensemble d'exemples de test

$\boldsymbol{Z}$ ..... facteur de réflectivité radar

ANNEXES

A.1 Définitions des attributs du contexte CT_{AP}

Par la suite nous précisons les paramètres utilisés dans cette étude pour la description des échos et des séquences, ainsi que les attributs du contexte de l'appariement, qui ont été utilisés pour la description des objets $(es,e) \in O_{AP}$. Soit $(I_1,..,I_n)$ une suite d'images radar, mesurées aux instants $t_1,..,t_n$; soit $es = (e_1 \in E(I_1),..,e_{n-1} \in E(I_{n-1}))$ une séquence stricte d'échos, et $e_n \in I_n$ un écho de l'image I_n.

Pour un <u>écho</u> e nous définissons:

- $\boldsymbol{A(e)}$ la taille (surface) de l'écho (km^2)
- $\boldsymbol{M(e)}$ la masse (flux) de l'écho (10^3 m^3/h)
- $\boldsymbol{R_{moy}(e)}$ l'intensité moyenne de l'écho (mm/h)
- $\boldsymbol{R_{max}(e)}$ l'intensité maximale de l'écho (mm/h)
- $\boldsymbol{R_{var}(e)}$ la variance de l'intensité de l'écho
- $\boldsymbol{C_x(e)}$, $\boldsymbol{C_y(e)}$ les coordonnés du barycentre de l'écho
- $\boldsymbol{I_{max}(e)}$, $\boldsymbol{I_{min}(e)}$ l'inertie maximale et minimale de l'écho
- $\tau(e)$ l'angle de l'axe d'inertie maximale de l'écho ($\in [0,\pi)$)
- $\boldsymbol{D(e)}$ la dispersion de l'écho: $D(e) = \dfrac{I_{max}(e) - I_{min}(e)}{A(e)^2}$
- $\boldsymbol{E(e)}$ l'élongation de l'écho: $E(e) = \dfrac{I_{max}(e) - I_{min}(e)}{I_{max}(e) + I_{min}(e)}$ ($\in [0,1]$)

Pour un <u>couple</u> d'échos (e_1,e_2) nous définissons:

- $\boldsymbol{V(e_1,e_2)}$ la vitesse de déplacement correspondant à la distance des deux barycentres (km/h)
- $\boldsymbol{D(e_1,e_2)}$ la direction de déplacement correspondant aux deux barycentres (rad)

Pour une <u>séquence</u> d'échos es nous définissons:

- $\boldsymbol{V_{moy}(es)}$ la vitesse moyenne de déplacement dans l'intervalle (t_1,t_{n-1}) (km/h)
- $\boldsymbol{D_{moy}(es)}$ la direction moyenne de déplacement dans l'intervalle (t_1,t_{n-1}) (rad)
- $\boldsymbol{DA_{moy}(es)}$ la moyenne de la différence relative de la taille dans l'intervalle $(t_{max(1,n-4)},t_{n-1})$
- $\boldsymbol{DM_{moy}(es)}$ la moyenne de la différence relative de la masse dans l'intervalle $(t_{max(1,n-4)},t_{n-1})$
- $\boldsymbol{DE_{moy}(es)}$ le changement moyen de l'élongation dans l'intervalle $(t_{max(1,n-4)},t_{n-1})$

Pour une <u>image</u> I nous définissons:

- $\boldsymbol{ImV_{moy}(I)}$ la moyenne des vitesses $V_{moy}(es)$ des séquences $es = (e_1^{es},..,e_k^{es})$ avec $e_k^{es} \in E(I)$
- $\boldsymbol{ImD_{moy}(I)}$ la moyenne des directions $D_{moy}(es)$ des séquences $es = (e_1^{es},..,e_k^{es})$ avec $e_k^{es} \in E(I)$

Pour la description des couples (es,e) (objets du contexte CT_{AP}), les attributs suivants ont été définis:

(1) la vitesse de déplacement correspondant à la distance des barycentres des échos e_{n-1} et e_n

$$\boldsymbol{Co_Dep_Vit(es,e)} = \frac{60}{t_n - t_{n-1}} \, ||(C_x(e_{n-1}),C_y(e_{n-1})) - (C_x(e_n),C_y(e_n))|| \quad [km/h]$$

(2) la différence relative de la taille des échos e_{n-1} et e_n

$$\boldsymbol{Co_D_Taill(es,e)} = \frac{1}{t_n - t_{n-1}} \frac{A(e_n) - A(e_{n-1})}{A(e_{n-1})}$$

(3) la différence relative de la masse des échos e_{n-1} et e_n

$$\boldsymbol{Co_D_Masse(es,e)} = \frac{1}{t_n - t_{n-1}} \frac{M(e_n) - M(e_{n-1})}{M(e_{n-1})}$$

(4) la différence relative de l'intensité moyenne des échos e_{n-1} et e_n

$$\boldsymbol{Co_D_IntMo(es,e)} = \frac{1}{t_n - t_{n-1}} \frac{R_{moy}(e_n) - R_{moy}(e_{n-1})}{R_{moy}(e_{n-1})}$$

(5) la différence relative de l'intensité maximale des échos e_{n-1} et e_n

$$\boldsymbol{Co_D_IntMa(es,e)} = \frac{1}{t_n - t_{n-1}} \frac{R_{max}(e_n) - R_{max}(e_{n-1})}{R_{max}(e_{n-1})}$$

(6) la différence relative de la variance de l'intensité des échos e_{n-1} et e_n

$$\boldsymbol{Co_D_IntVa(es,e)} = \frac{1}{t_n - t_{n-1}} \frac{R_{var}(e_n) - R_{var}(e_{n-1})}{R_{var}(e_{n-1})}$$

(7) la différence des angles de l'axe principale des échos e_{n-1} et e_n

$$\boldsymbol{Co_D_Angle(es,e)} = \frac{1}{t_n - t_{n-1}} \, | \tau(e_n) - \tau(e_{n-1}) |$$

(8) la différence de la dispersion des échos e_{n-1} et e_n

$$\boldsymbol{Co_D_Dispe(es,e)} = \frac{1}{t_n - t_{n-1}} \, | D(e_n) - D(e_{n-1}) |$$

(9) la différence de l'élongation des échos e_{n-1} et e_n

$$\boldsymbol{Co_D_Elong(es,e)} = \frac{1}{t_n - t_{n-1}} \, | E(e_n) - E(e_{n-1}) |$$

(10) la distance des échos e_{n-1} et e_n relatif à la vitesse moyenne de déplacement de l'image I_{n-1}

$$\boldsymbol{Co_DepVitV(es,e)} = \begin{cases} \dfrac{1}{t_n - t_{n-1}} \dfrac{Co_Dep_Vit(es,e) - ImV_{moy}(I_{n-1})}{ImV_{moy}(I_{n-1})} & si \; ImV_{moy}(I_{n-1}) \; connu \\ inconnu & sinon \end{cases}$$

(11) la direction de déplacement correspondant au barycentres des échos e_{n-1} et e_n relatif à la direction moyenne de déplacement de l'image I_{n-1}

$$\boldsymbol{Co_DepDirV(es,e)} = \begin{cases} \dfrac{1}{t_n - t_{n-1}} \, |D(e_{n-1},e_n) - ImD_{moy}(I_{n-1})| & si\ ImD_{moy}(I_{n-1})\ connu \\ inconnu & sinon \end{cases}$$

(12) la distance des échos e_{n-1} et e_n relatif à la vitesse moyenne de déplacement de la séquence *es*

$$\boldsymbol{Co_DepVitR(es,e)} = \begin{cases} \dfrac{1}{t_n - t_{n-1}} \, \dfrac{Co_Dep_Vit(es,e) - V_{moy}(es)}{V_{moy}(es)} & si\ n > 2 \\ inconnu & sinon \end{cases}$$

(13) la direction de déplacement correspondant au barycentres des échos e_{n-1} et e_n relatif à la direction moyenne de déplacement de la séquence *es*

$$\boldsymbol{Co_DepDirR(es,e)} = \begin{cases} \dfrac{1}{t_n - t_{n-1}} \, |D(e_{n-1},e_n) - D_{moy}(es)| & si\ n > 2 \quad (\in [0,\pi)) \\ inconnu & sinon \end{cases}$$

(14) la différence relative de la taille des échos e_{n-1} et e_n relatif au taux moyen de développement de la taille de la séquence *es*

$$\boldsymbol{Co_DTaillR(es,e)} = \begin{cases} |Co_D_Taill(es,e) - DA_{moy}(es)| & si\ n > 2 \\ inconnu & sinon \end{cases}$$

(15) la différence relative de la masse des échos e_{n-1} et e_n relatif au taux moyen de développement de la masse de la séquence *es*

$$\boldsymbol{Co_DMasseR(es,e)} = \begin{cases} |Co_D_Masse(es,e) - DM_{moy}(es)| & si\ n > 2 \\ inconnu & sinon \end{cases}$$

(16) la différence relative de l'élongation des échos e_{n-1} et e_n relatif au changement moyen de l'élongation de la séquence *es*

$$\boldsymbol{Co_DElongR(es,e)} = \begin{cases} |Co_D_Elong(es,e) - DE_{moy}(es)| & si\ n > 2 \\ inconnu & sinon \end{cases}$$

(17) la distance des échos e_{n-1} et e_n relatif à la vitesse moyenne de déplacement de la séquence *es*, si cette dernière est assez fiable

$$\boldsymbol{Co_DepVitT(es,e)} = \begin{cases} Co_DepVitR & si\ t_{n-1} - t_1 \geq 15\ \text{min} \\ inconnu & sinon \end{cases}$$

(18) la direction de déplacement correspondant au barycentres des échos e_{n-1} et e_n relatif à la direction moyenne de déplacement de la séquence, si cette dernière est assez fiable

$$\boldsymbol{Co_DepDirT(es,e)} = \begin{cases} Co_DepDirR & si\ t_{n-1}-t_1 \geq 15\ \text{min} \\ inconnu & sinon \end{cases}$$

(19) la différence relative de la taille des échos e_{n-1} et e_n relatif au taux moyen de développement de la taille de la séquence *es*, si cette dernière est assez fiable

$$\boldsymbol{Co_DTaillT(es,e)} = \begin{cases} Co_DTaillR & si\ t_{n-1}-t_1 \geq 15\ \text{min} \\ inconnu & sinon \end{cases}$$

(20) la différence relative de la masse des échos e_{n-1} et e_n relatif au taux moyen de développement de la masse de la séquence *es*, si cette dernière est assez fiable

$$\boldsymbol{Co_DMasseT(es,e)} = \begin{cases} Co_DMasseR & si\ t_{n-1}-t_1 \geq 15\ \text{min} \\ inconnu & sinon \end{cases}$$

(21) la différence relative de l'élongation des échos e_{n-1} et e_n relatif au changement moyen de l'élongation de la séquence *es*, si cette dernière est assez fiable

$$\boldsymbol{Co_DElongT(es,e)} = \begin{cases} Co_DElongR & si\ t_{n-1}-t_1 \geq 15\ \text{min} \\ inconnu & sinon \end{cases}$$

Les valeurs des attributs n° 1-9 sont connues pour tous les couples (*es*,*e*), tandis que les valeurs des attributs n° 10 et 11 sont connues uniquement, s'il existent des échos sur l'image I_{n-1}, qui font partie de séquences observées avant t_{n-1}, et les valeurs des attributs n° 12-16 ne sont connues qu'au cas où la séquences *es* est non-trivial. Afin de tenir compte de l'instabilité des valeurs des attributs n° 12-16, nous avons introduit les attributs n° 17-21, qui sont définis uniquement lorsque la séquence *es* est d'une longueur supérieure à 15 minutes.

Dans l'arbre de décision ADD_{APP}, les attributs n° 1, 2, 3, 6, 7, 8, 17, et 18 sont utilisés (cf. la figure IV.9).

A.2 Présentation du système de visualisation des images radar, d'appariement manuel et d'analyse des résultats de la prévision

L'étude présentée dans ce mémoire est en grande partie basée sur une analyse détaillée des images radar, du développement de la pluie, et des sources d'erreurs de la prévision. Cette analyse a été rendu possible par le développement d'un système, qui possède les fonctions suivantes:

(1) La visualisation
- des images radar (image par image ou animation rapide avec possibilité d'un zoom sur une zone),
- des échos simples et des échos imaginaires,
- des séquences d'échos définies par les appariements,
- des caractéristiques des échos et des séquences,
- des caractéristiques topographiques de la région couverte par les mesures radar,
- des lames d'eau provoquées par une cellule choisie ou par une zone pluvieuse,
- des zones de croissance/décroissance de la pluie
- de l'arbre de décision et du fonctionnement de l'algorithme de l'appariement automatique,
- des vecteurs de déplacement des cellules de pluie,
- des vecteurs de déplacement prévues par PROPHETIA,
- des lames d'eau prévues et des écarts entre lames prévues et lames mesurées.

(2) La définition manuelle
- des échos de base et des échos imaginaires,
- des appariements d'échos.

(3) La suppression manuelle des échos de sol.

(4) Le calcul et la sortie sur fichier des caractéristiques des échos et des séquences d'échos.

(5) La visualisation des échos et des appariements définis par SCOUT et la vérification de chaque étape des systèmes de prévision PROPHETIA, CORRCROIS, et PERSIST.

Le logiciel est géré par des menus. La sélection des fonctions, des échos, des zones de zoom et de zones de prévision est effectuée par une souris.

Ce système a été d'une extrême utilité pour cette étude. La compréhension des sources d'erreurs de la prévision et des interactions entre le développement de la pluie et les caractéristiques de la surface du sol n'aurait pas été possible sans cet outil. Aussi la définition des exemples de l'apprentissage a pu être effectuée seulement grâce à la visualisation des échos. Sur les pages suivantes nous présentons graphiquement quelques fonctions de cet outil.

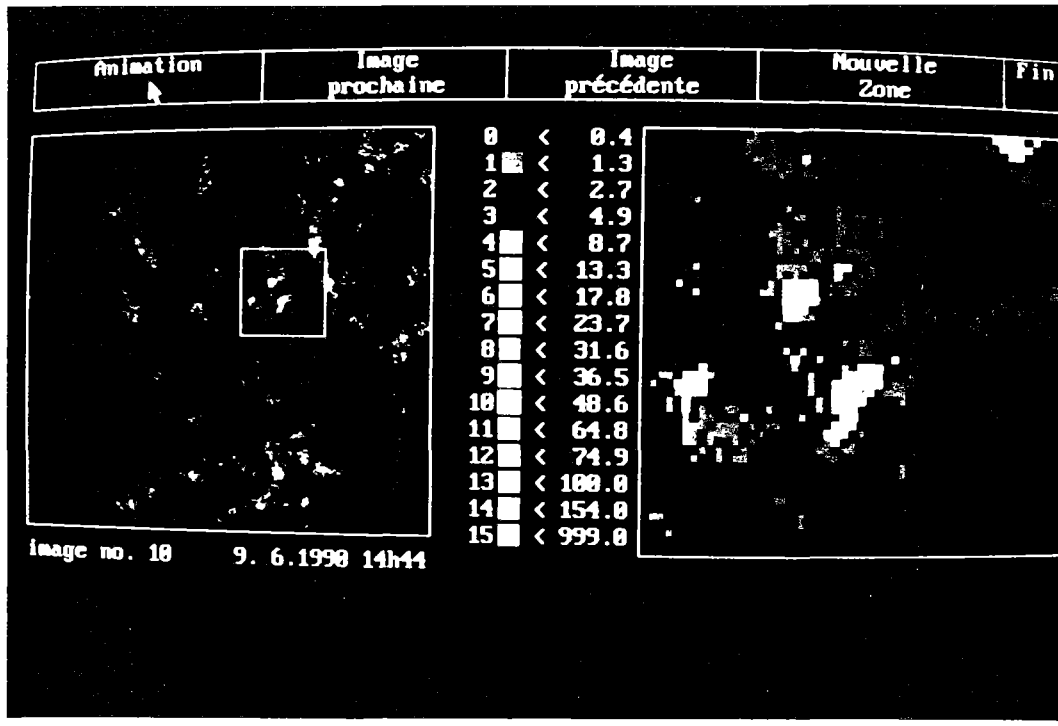


Figure A.1.1: Visualisation d'une image radar (à gauche) et zoom sur une zone (à droite) (échelle en mm/h)

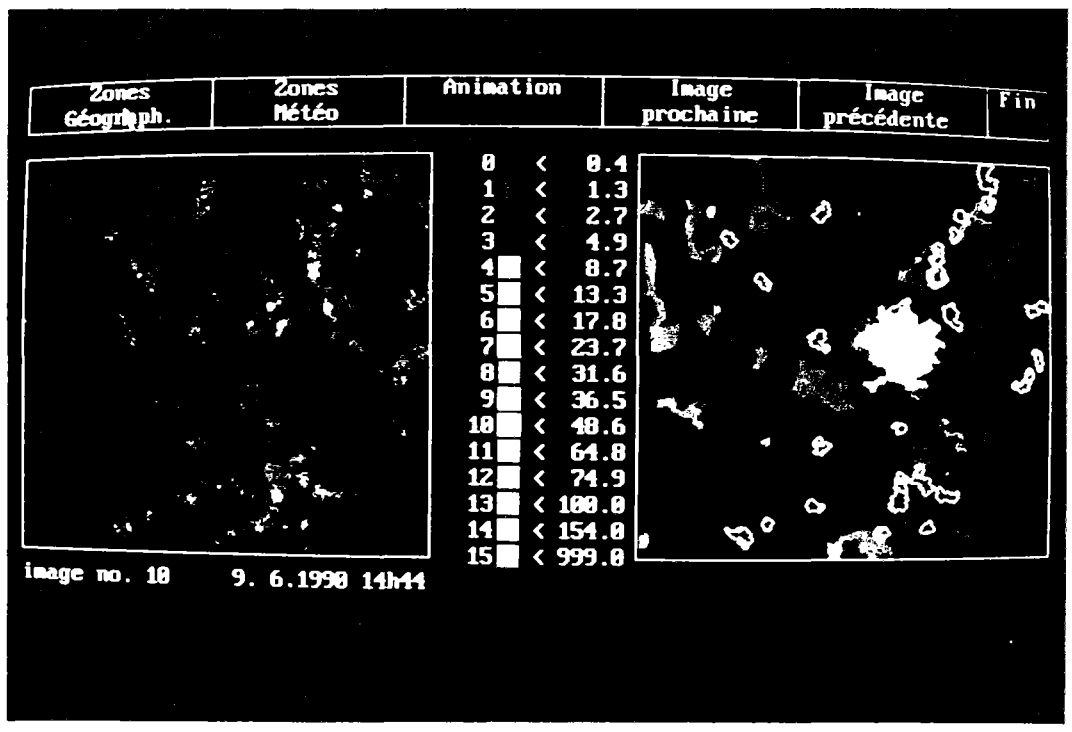


Figure A.1.2: Visualisation d'une image (à gauche) et superposition des contours des échos sur le fond de carte (à droite, zones urbanisée en orange, zones boisées en vert, fleuves en bleu)

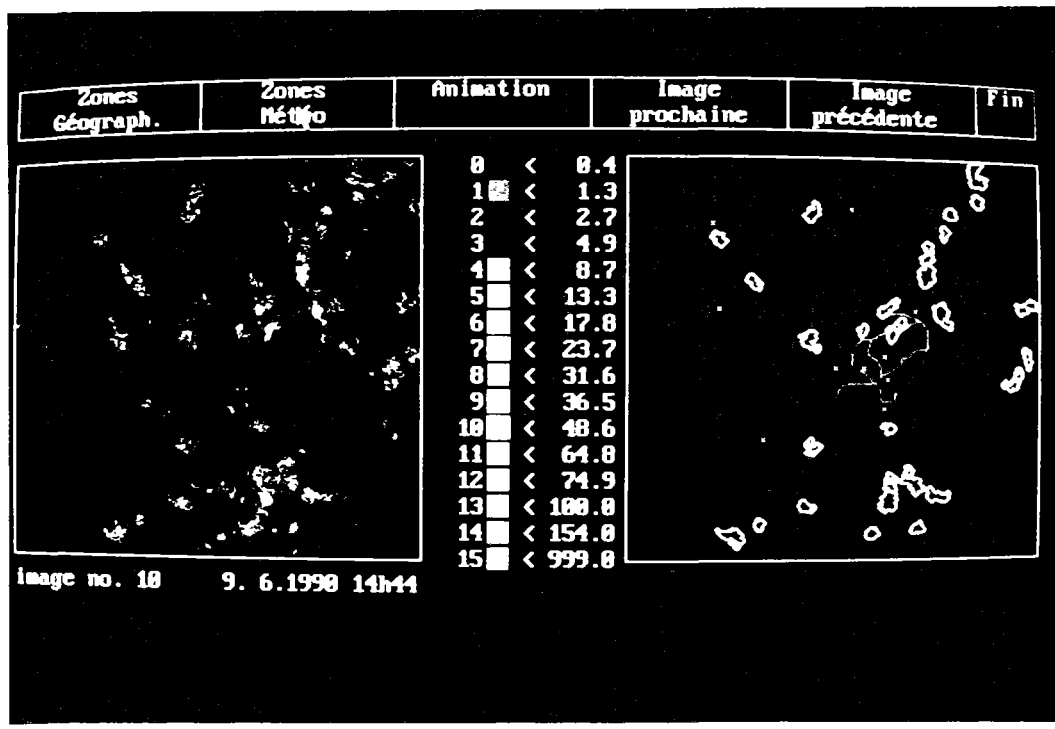


Figure A.1.3 Visualisation d'une image (à gauche) et superposition des contours des échos sur le fond de carte des régions définies pour les stations météorologiques (à droite) (cf. chapitre VI)

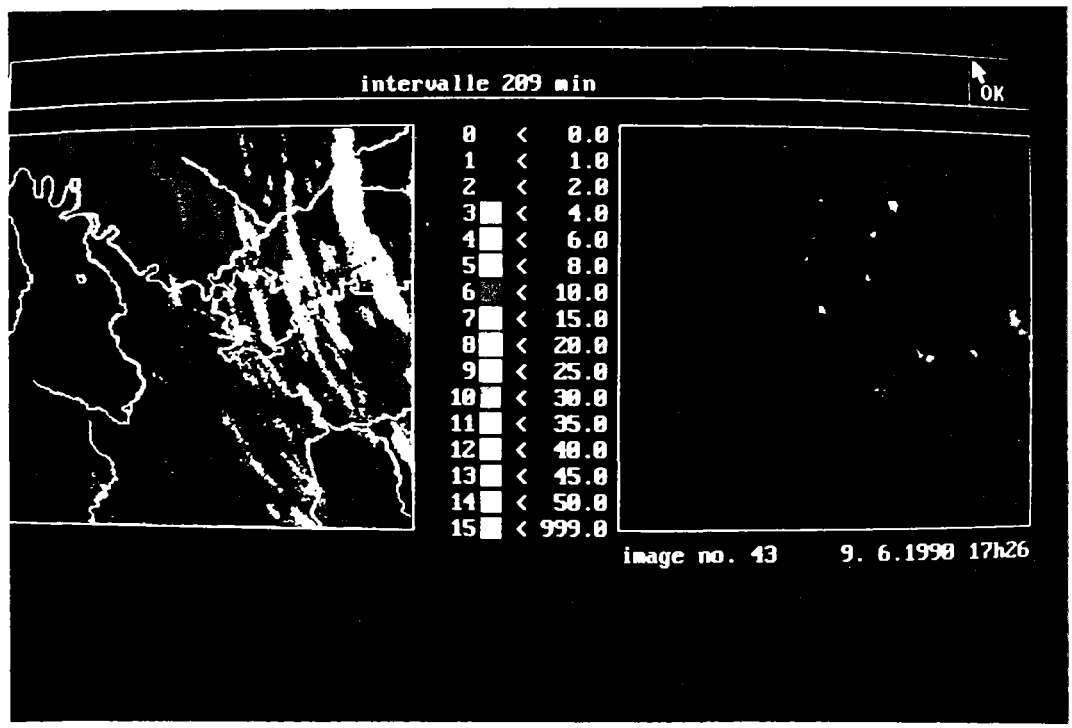


Figure A.1.4: Visualisation des lames d'eau provoquées par un événement de pluie (échelle en mm)

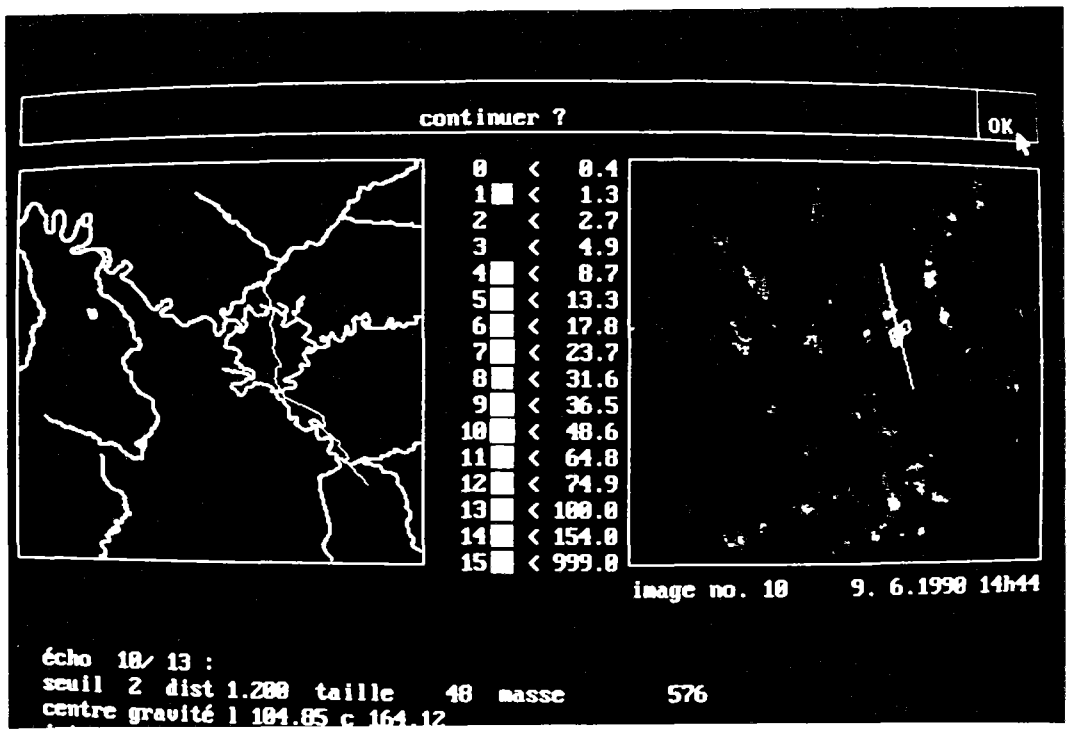


Figure A.1.5: Identification de l'écho et vecteur moyen de l'advection avant et après (à droite) et visualisation du déplacement de la cellule correspondante (à gauche)

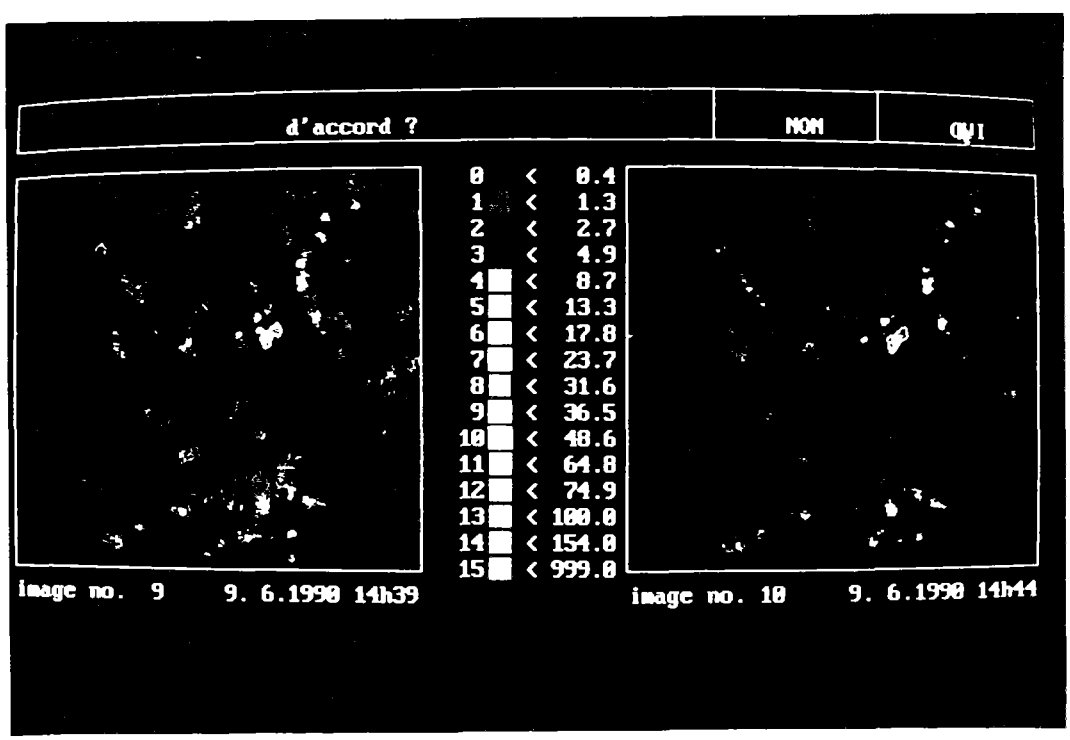


Figure A.1.6: Appariement manuel des échos sur deux images consécutives

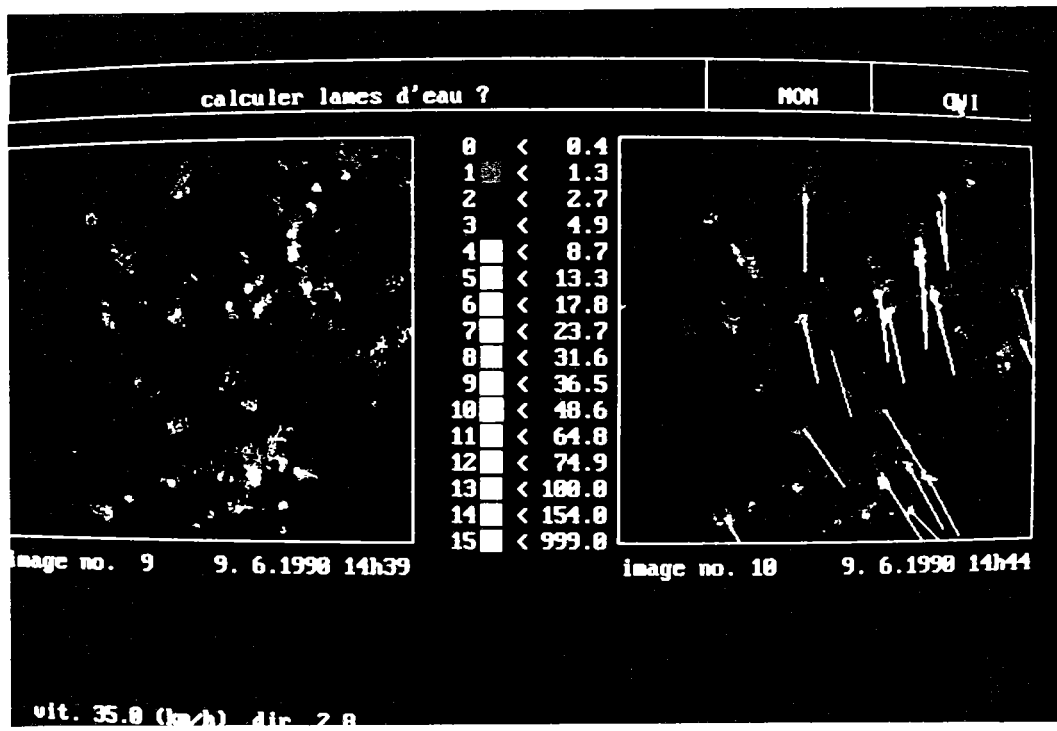


Figure A.1.7: Prévision de l'advection des cellules pour à une échéance de 60 minutes

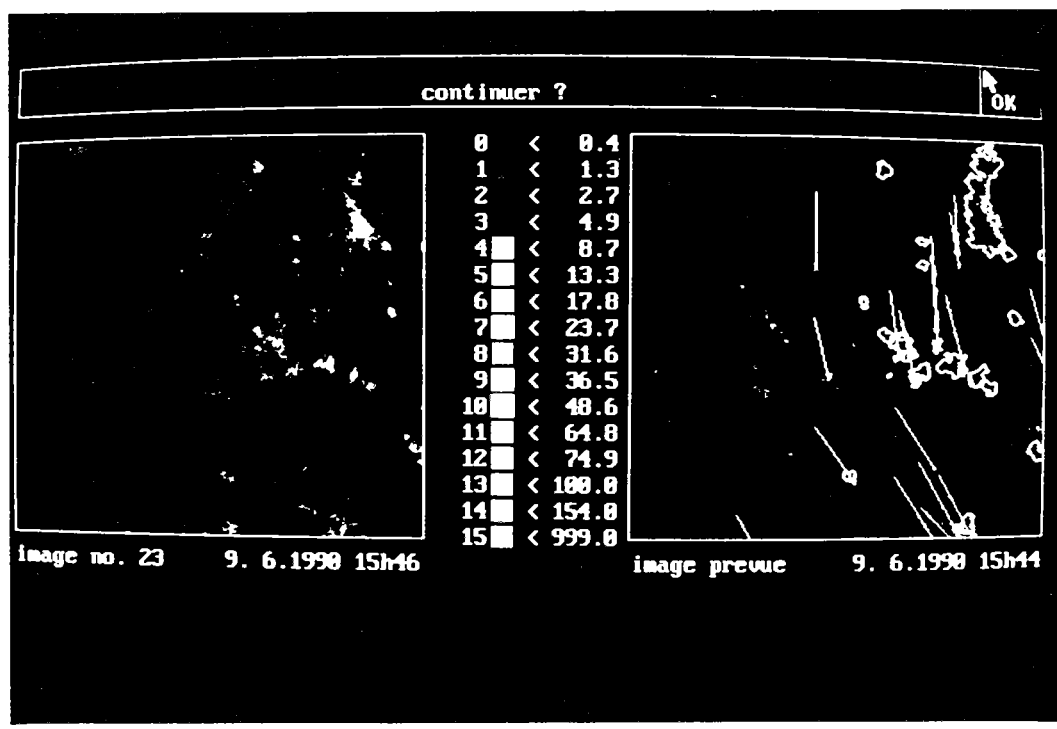


Figure A.1.8: Visualisation de l'image prévue (à droite, avec superposition des contours des échos de l'image mesurée, présentée à gauche) (prévision de 60 minutes)

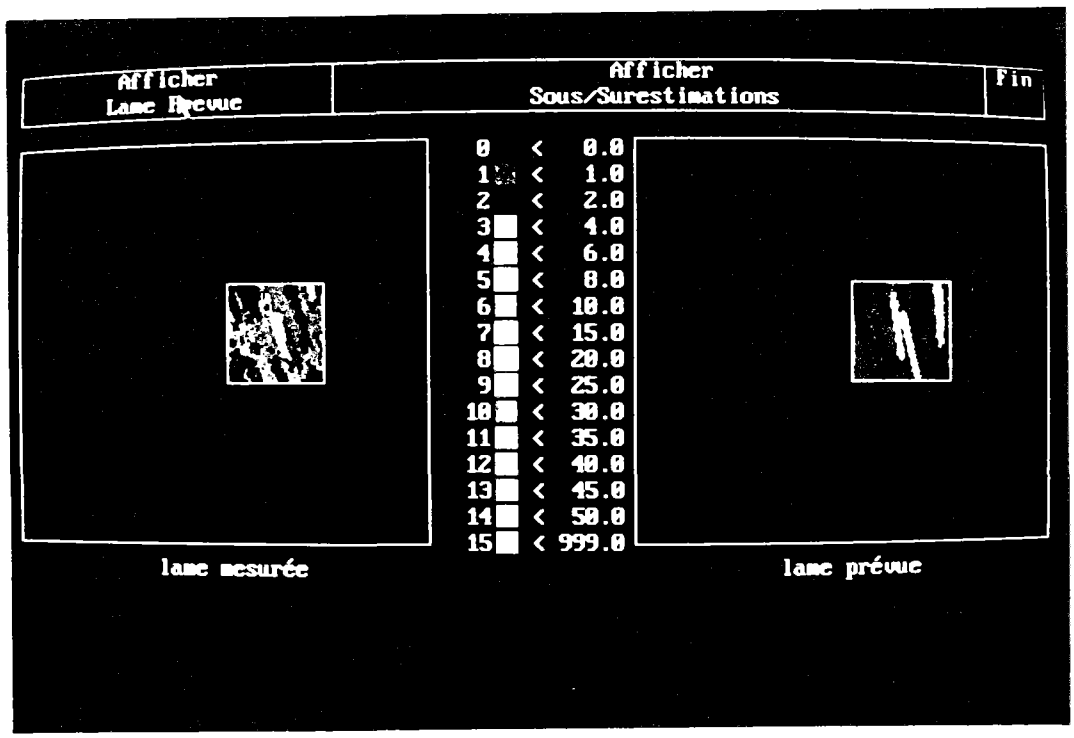


Figure A.1.9: Présentation graphique des lames d'eau prévues et des lames d'eau mesurées (échelle en mm, prévision de 60 minutes

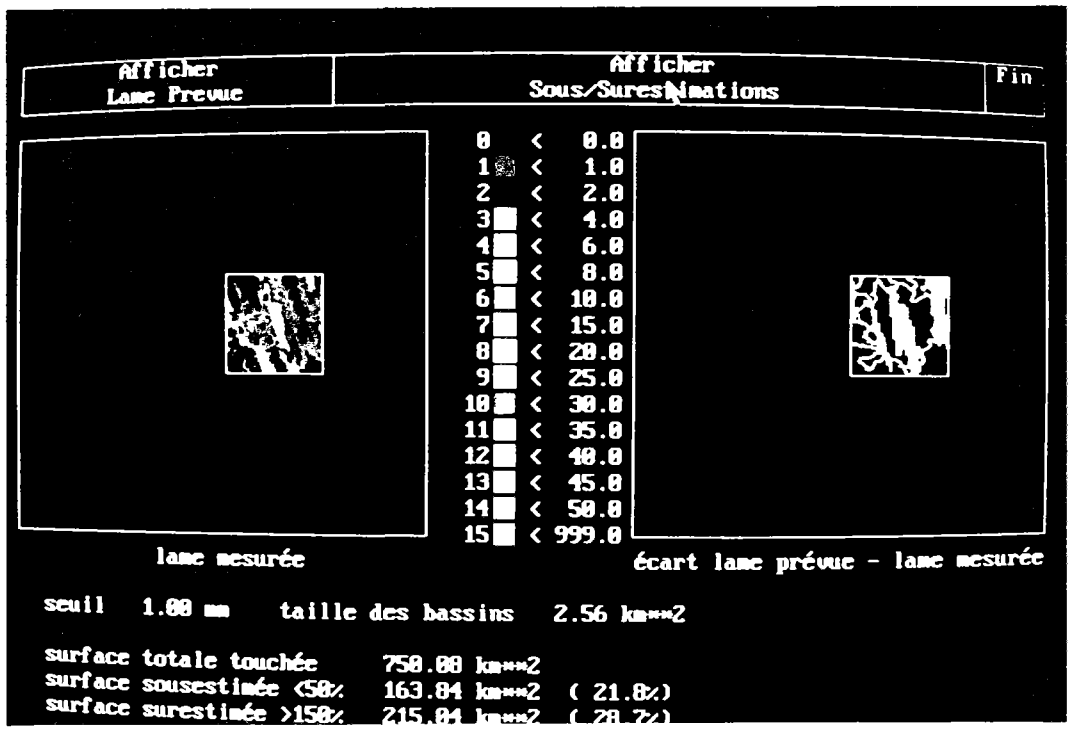


Figure A.1.10: Présentation de l'erreur de la prévision (à droite, zones sous-estimées en rouge, zones sur-estimées en rouge clair, avec en superposition le fond de carte)

A.3 Présentation graphique des événements de pluie utilisés dans cette étude

Sur les pages suivantes, nous présenterons graphiquement les 20 événements de pluie, qui ont été sélectionnés pour cette étude (cf. aussi le tableau IV.2). Pour chaque pluie, les figures présentent six images radar, à intervalles de 30 minutes environ, qui correspondent à la période la plus intéressante de l'événement. Les limites de la zone urbanisée en région parisienne sont représentées par la ligne blanche sur chaque image. La région, pour laquelle les prévisions ont été évaluées dans cette étude (critère TMP), correspond approximativement au carré circonscrivant cette zone.

L'échelle des intensités est présentée dans la figure A.2. Rappelons que les intensités sont estimées à l'aide de la relation *Z-R* de Marshall et Palmer à partir des images non calibrées discrétisées en 16 niveaux. Des comparaisons effectuées par la société RHEA avec des mesures par pluviomètres dans le département Seine-St.Denis montrent une sous-estimation des intensités réelles pour toutes les pluies de l'année 1989. La sous-estimation est d'environ 50%-60% pour la plus grande partie des pluies, et elle atteint son maximum pour la pluie du 7.8.1989 avec environ 80%. Car l'estimation correcte des intensités est d'une faible importance pour cette étude, nous n'avons pas tenté de corriger cette erreur de mesure.

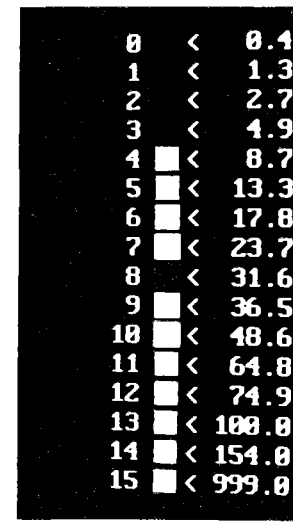


Figure A.2: Échelle des couleurs et des intensités (mm/h) des figures suivantes

Le numéro de la prévision, indiqué sur les figures V 7.1-V.7.20 et VI 8.a-VI.8.l, n'est pas égal au numéro de l'image, car pour la première image de chaque pluie les vecteurs d'advection sont inconnus, et il n'existent pas d'échos sur les premières images des pluies du 4.4.1989 et du 27.4.1989. Sur les figures suivantes, la différence entre le numéro de l'image et le numéro de la prévision est indiquée

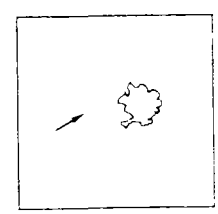
Dans l'icone figurant en bas de chaque page, les principaux déplacements sont indiquées par des flèches (la longueur des flèches correspond approximativement au déplacement par heure). Plusieurs cellules présentent des particularités, dont les effets ont été examinés au chapitres IV, V, et VI. Ces cellules sont marquées sur les graphiques, afin de permettre une meilleure compréhension des phénomènes.

image no. 3 7. 3.1989 15h 1 image no. 9 7. 3.1989 15h30

image no. 15 7. 3.1989 15h59 image no. 22 7. 3.1989 16h32

image no. 28 7. 3.1989 17h 1 image no. 34 7. 3.1989 17h30

Figure A.3.1: Événement du 7.3.1989 (n° prév. = n° image -1)
(zone marquée * : cf. § V.3.2.1)



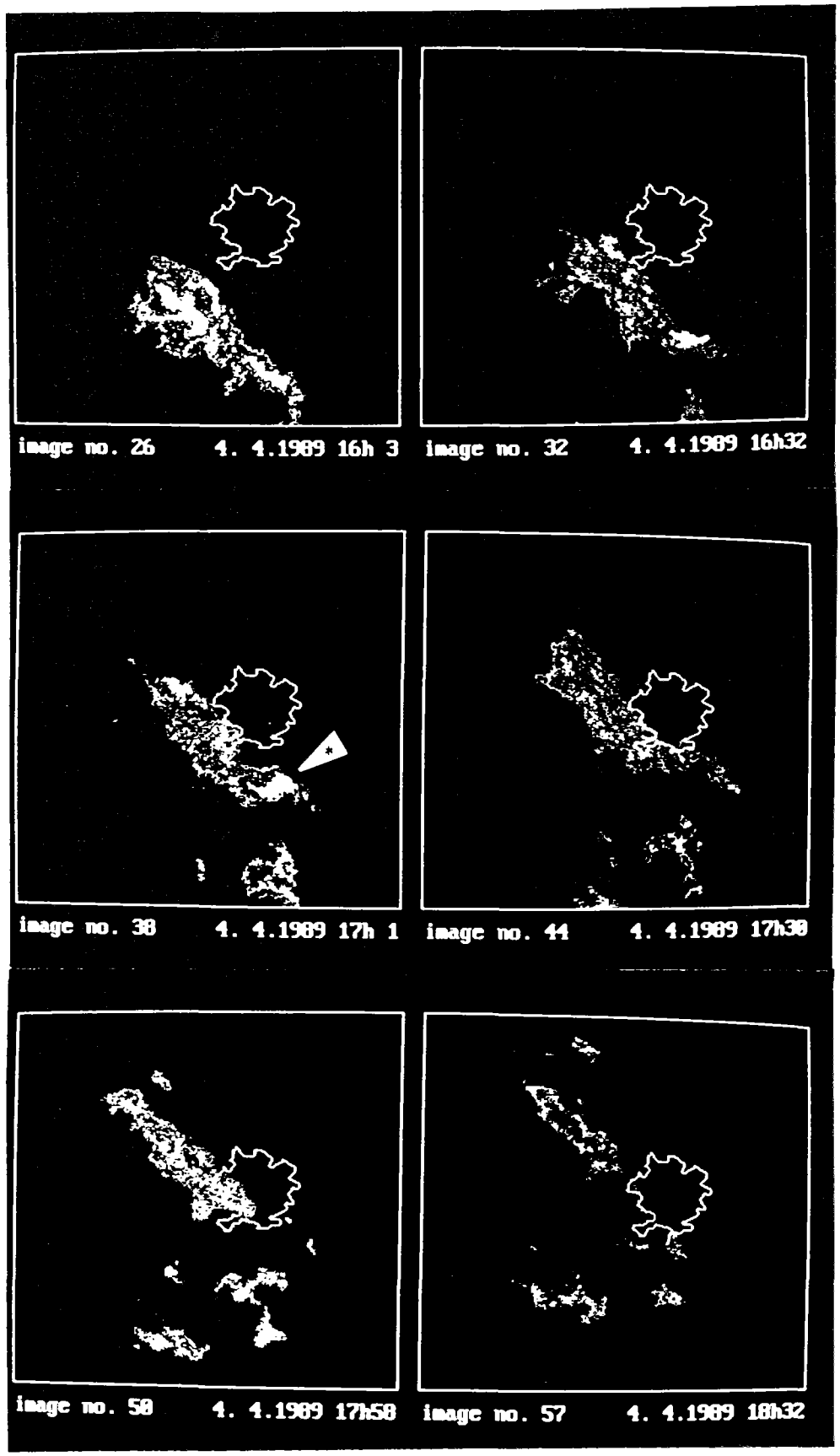
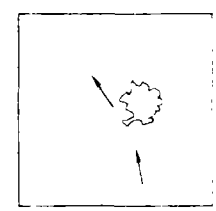


Figure A.3.2: Événement du 4.4.1989 (n° prev. = n° image -2)
zone marquée * (cf. § V 3.2.3)



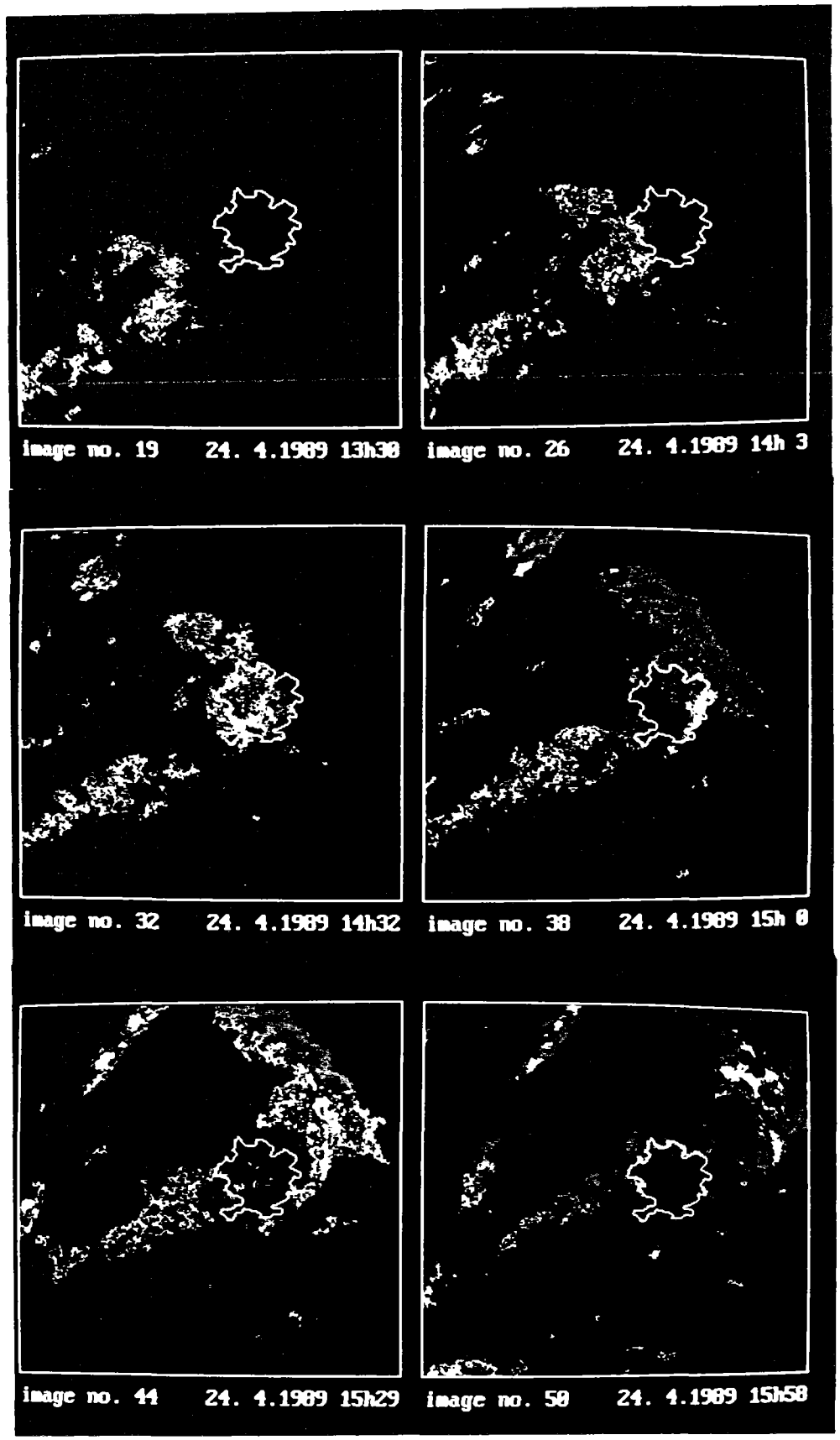
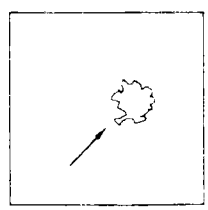


Figure A.3.3 Événement du 24.4.1989 (n° prév. = n° image -1)



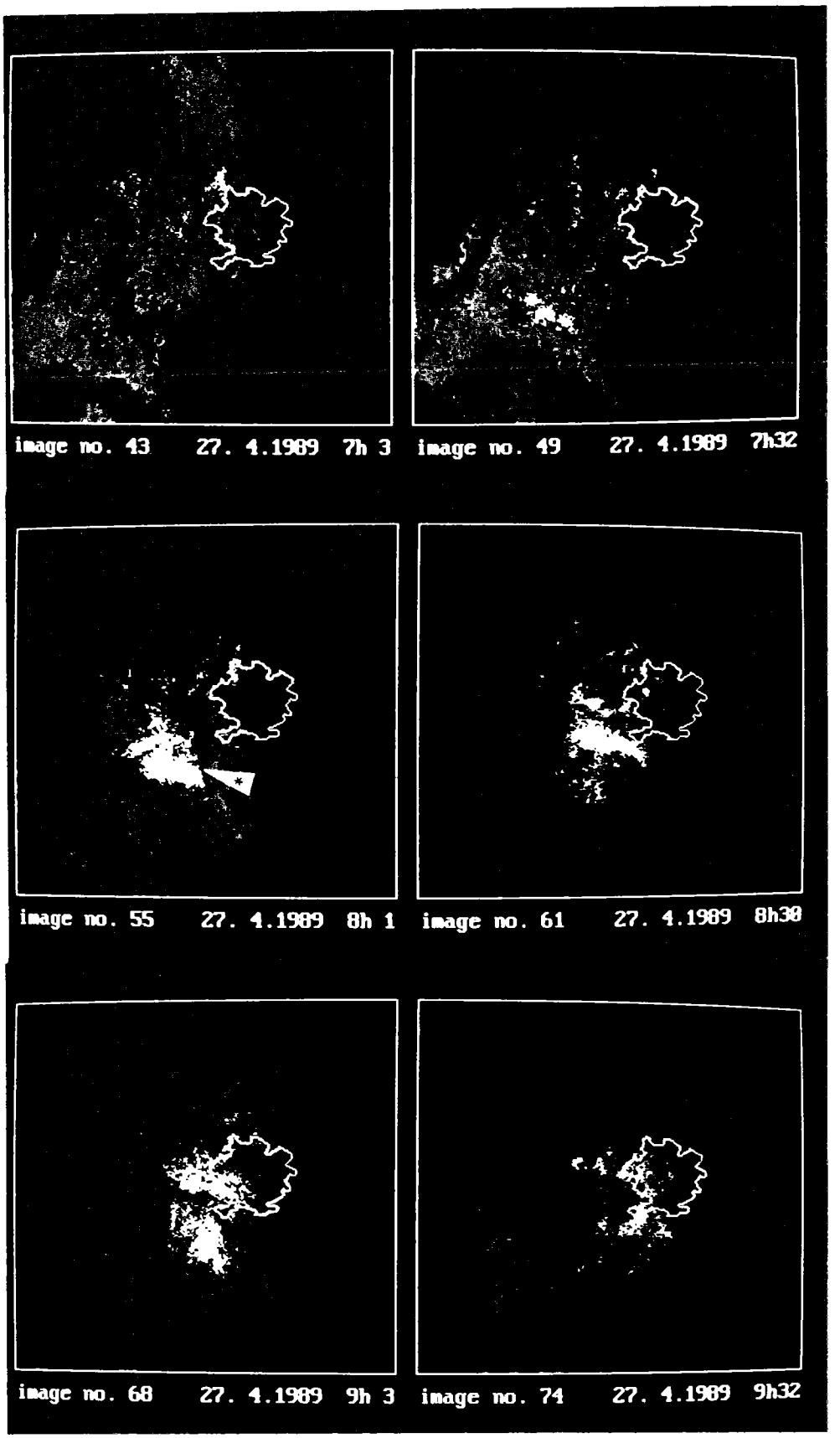
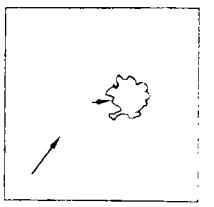


Figure A.3.4: Événement du 27.4.1989 (n° prév. = n° image -2)
(zone marquée * : cf. § V.3.2.2)



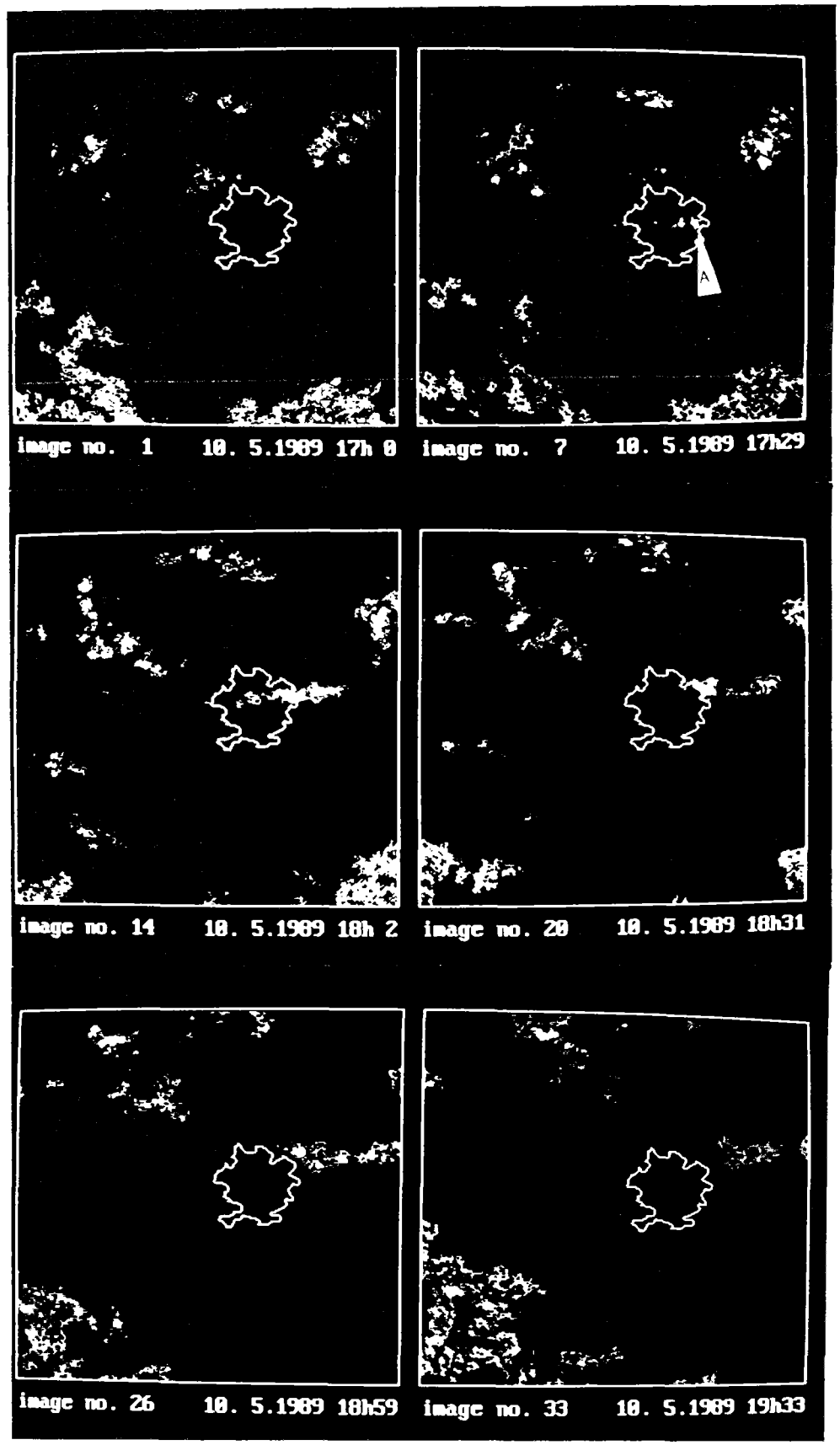
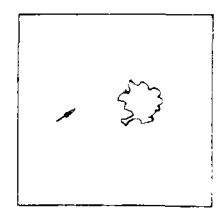


Figure A.3.5: Événement du 10.5.1989 (n° prév. = n° image -1)
(cellule A : cf § VI.5.4)



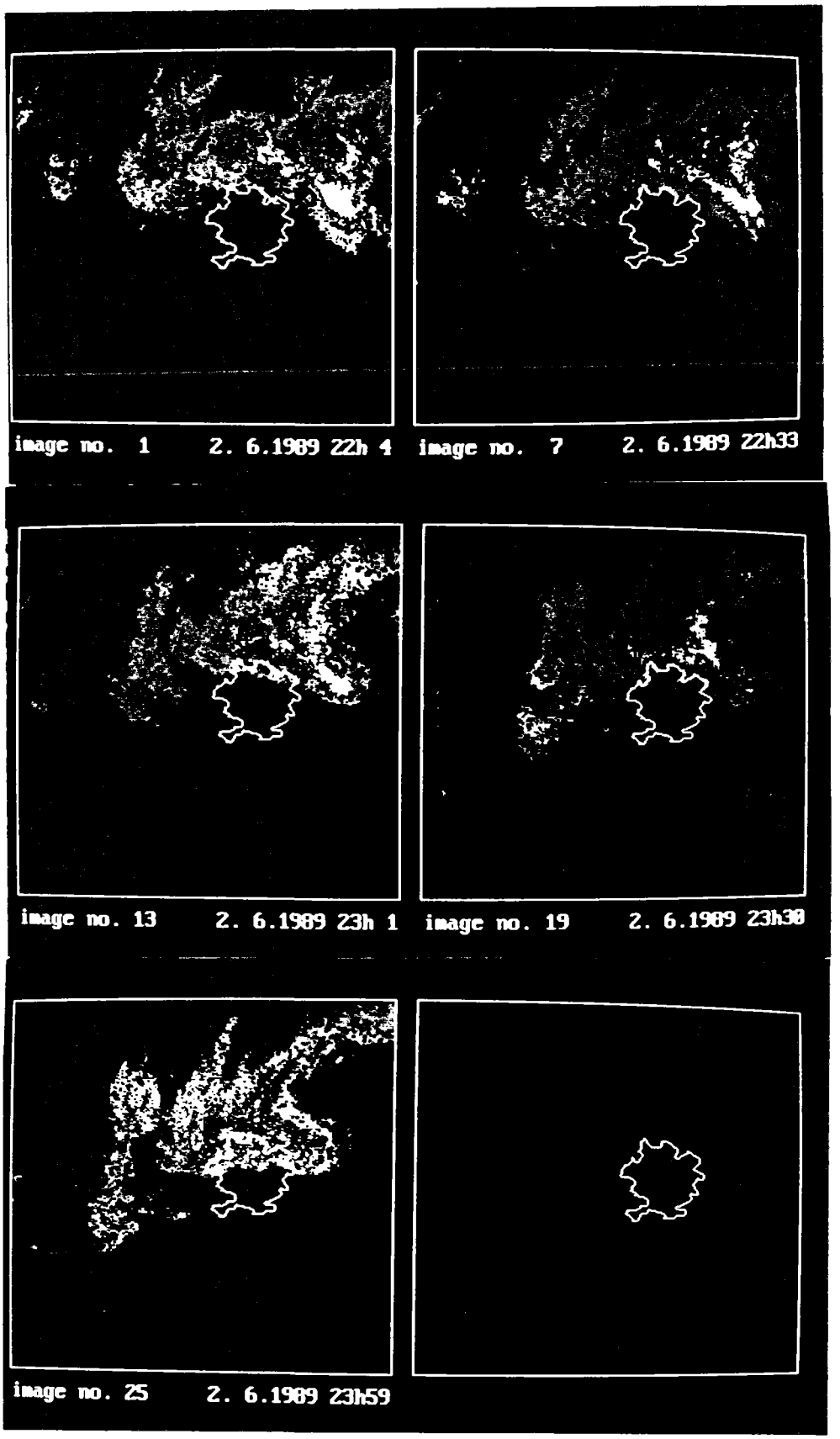
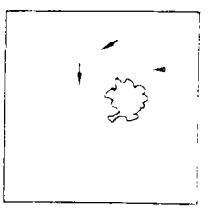


Figure A.3.6: Événement du 2.6.1989 (n° prév. = n° image -1)



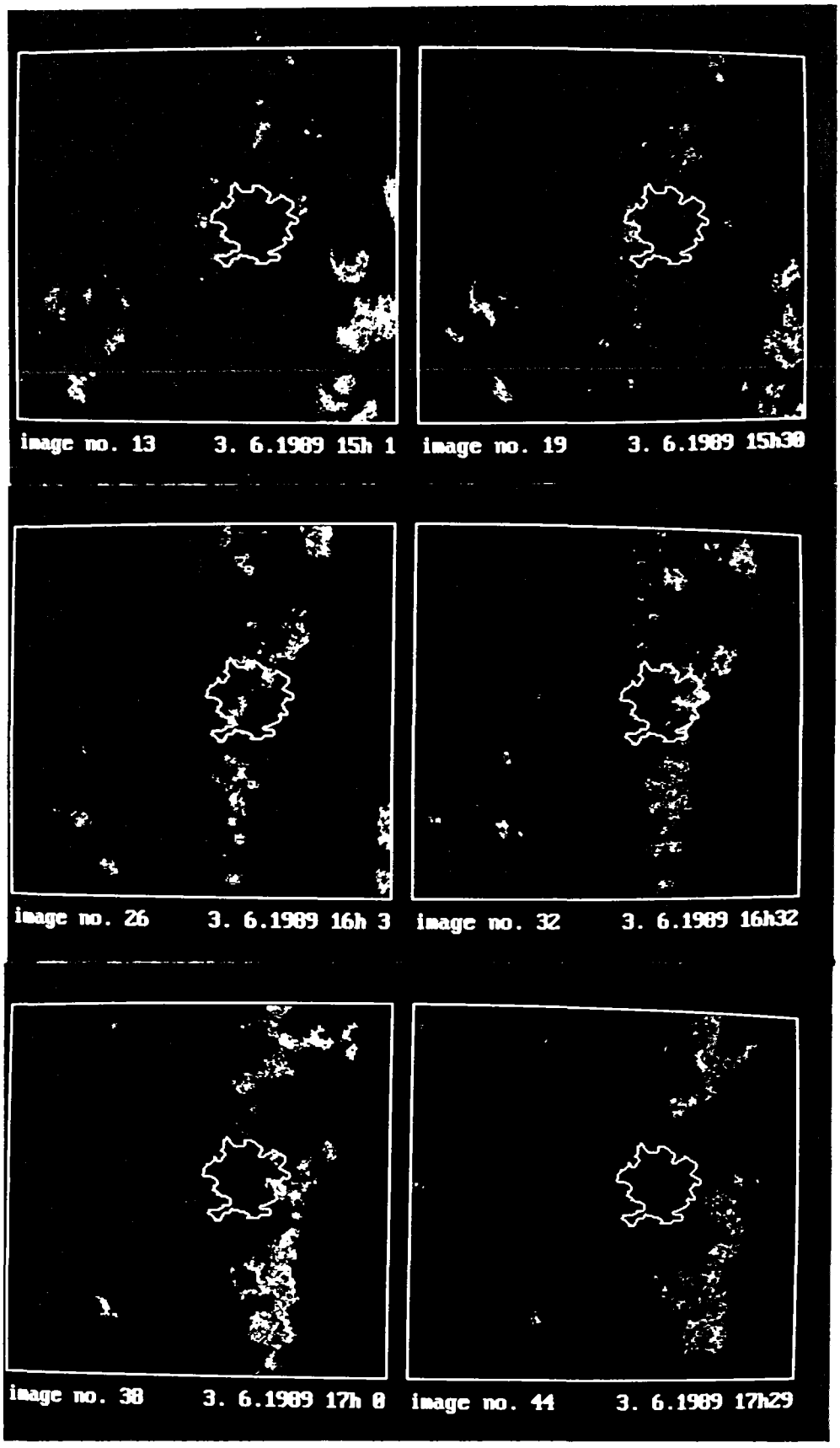
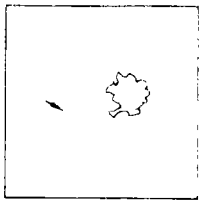


Figure A.3.7: Evenement du 3.6.1989 (n° prév. = n° image -1)



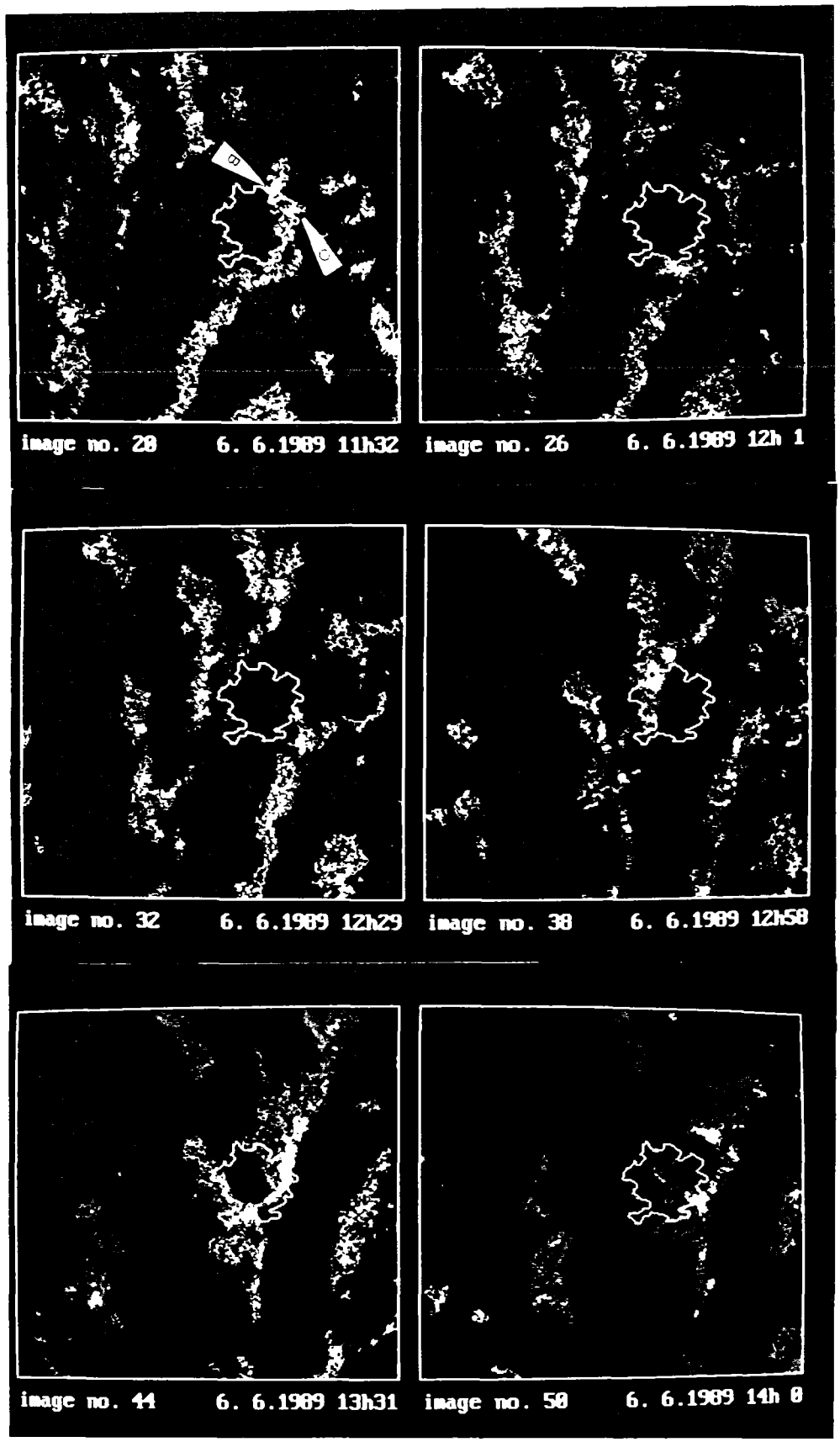


Figure A.3.8: Événement du 6.6.1989 (n° prév. = n° image -1)
(cellules B+C : cf. § VI.5.4)

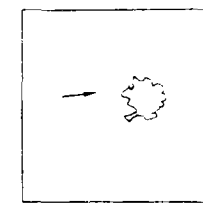


image no. 17 27. 6.1989 10h17

image no. 23 27. 6.1989 10h45

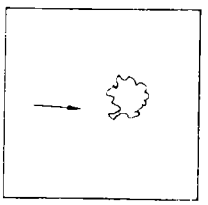
image no. 29 27. 6.1989 11h14

image no. 36 27. 6.1989 11h47

image no. 42 27. 6.1989 12h16

image no. 48 27. 6.1989 12h45

Figure A.3.9: Événement du 27.6.1989 (n° prév. = n° image -1)
(zone marquée * cf. § V.3.2.2)



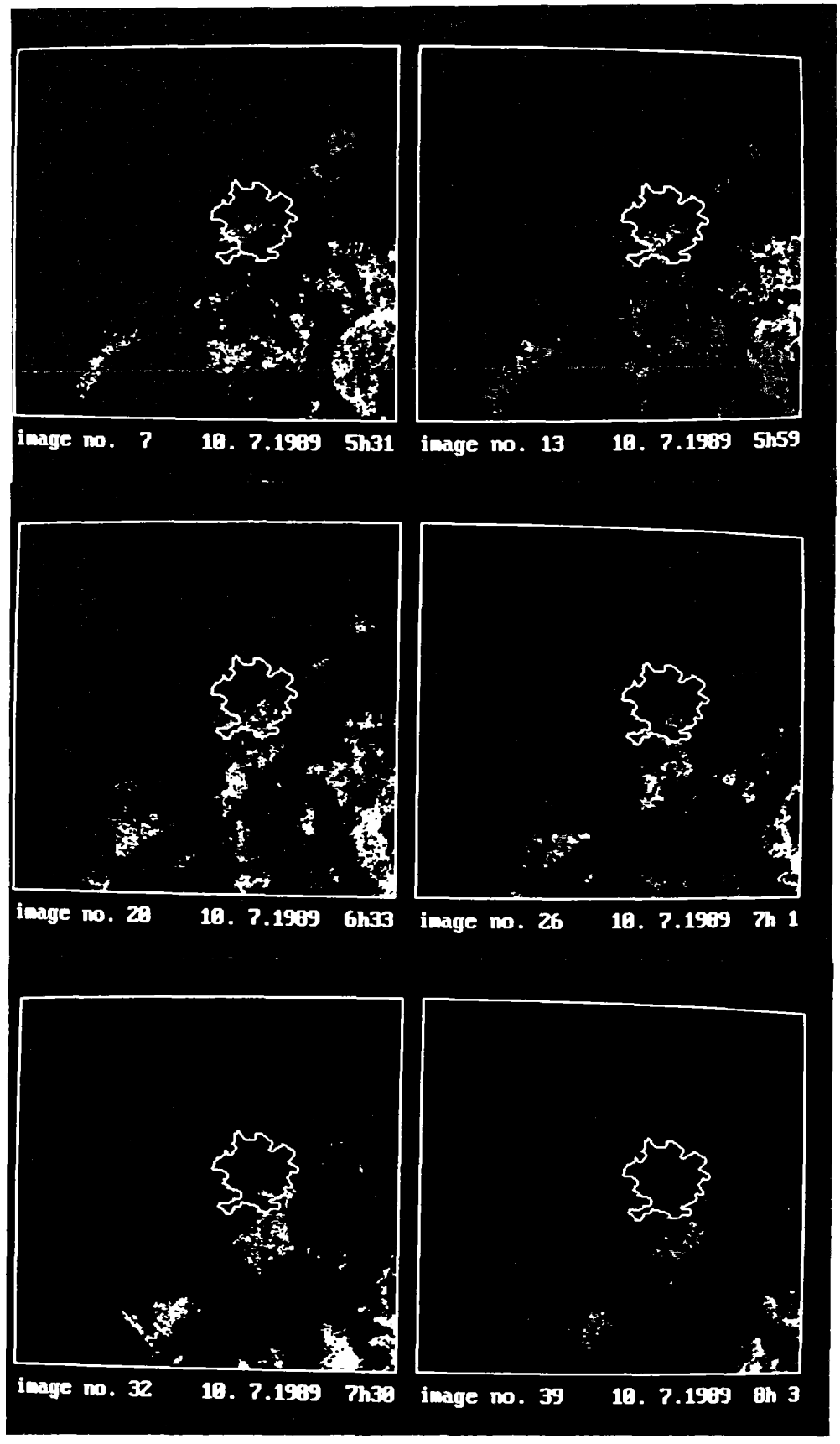


Figure A.3.10: Événement du 10.7 1989 (n° prév. = n° image -1)

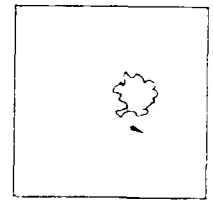
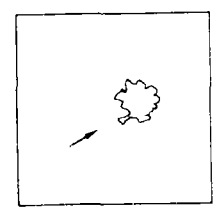


image no. 23 7. 8.1989 14h45 image no. 29 7. 8.1989 15h13

image no. 35 7. 8.1989 15h42 image no. 42 7. 8.1989 16h15

image no. 48 7. 8.1989 16h44 image no. 55 7. 8.1989 17h17

Figure A.3.11: Événement du 7.8.1989 (n° prév. = n° image -1)
(cellule D : cf. § VI.5.4)



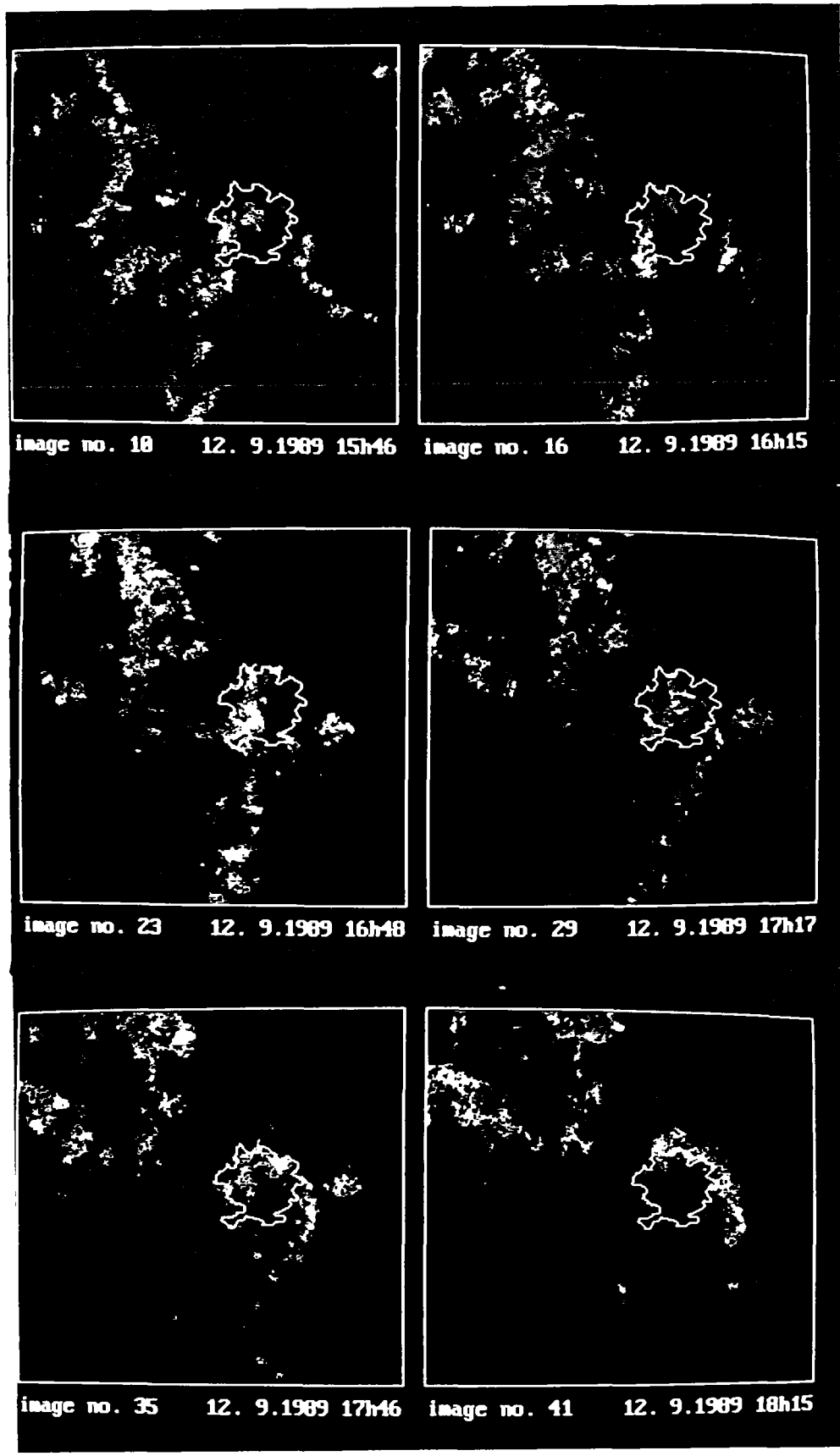
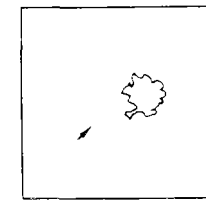


Figure A.3.12: Événement du 12.9.1989 (n° prév. = n° image -1)



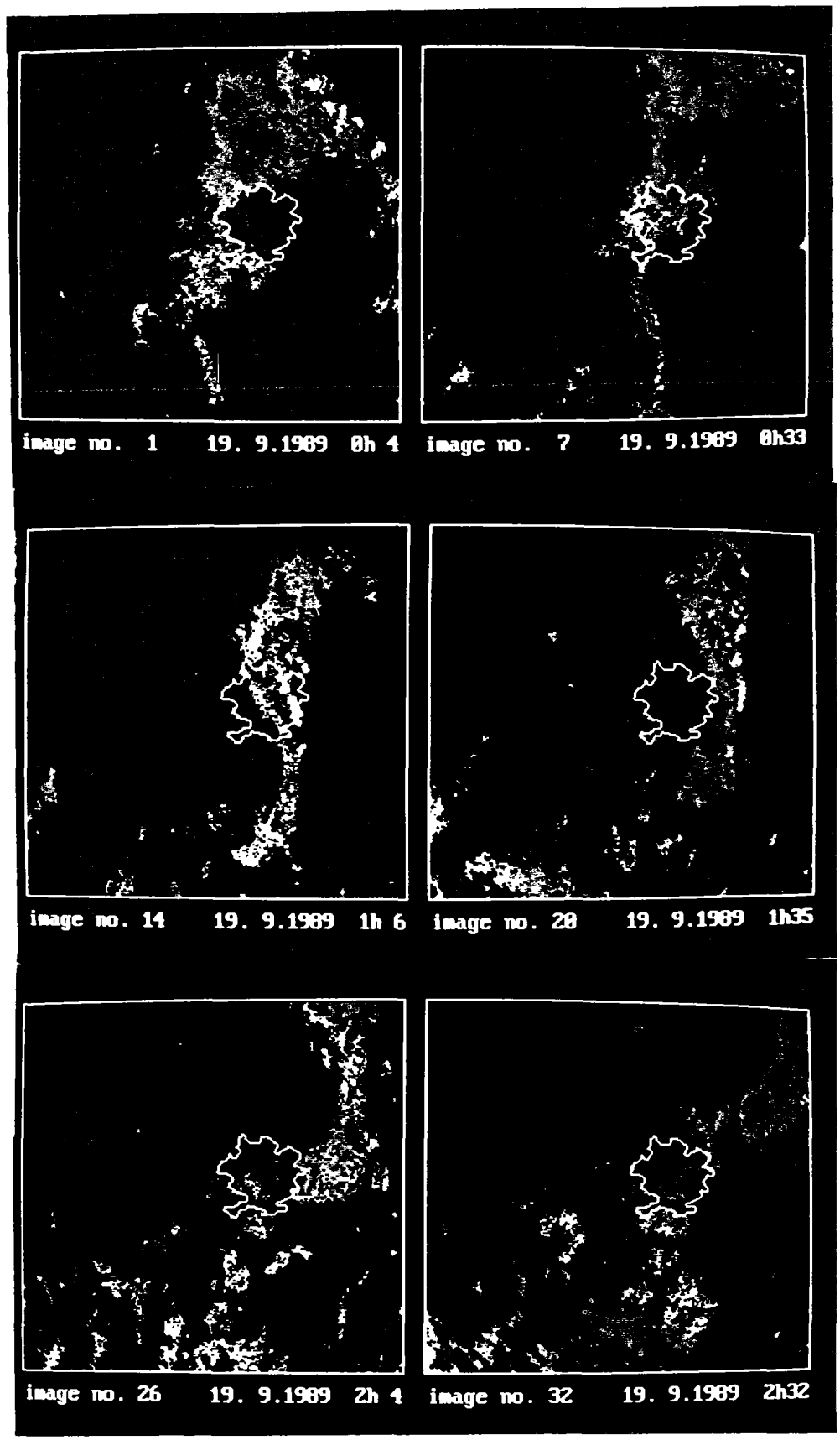
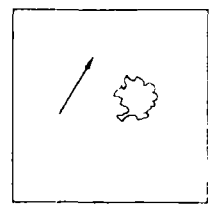


Figure A.3.13: Évenement du 19.9.1989 (n° prév = n° image -1)



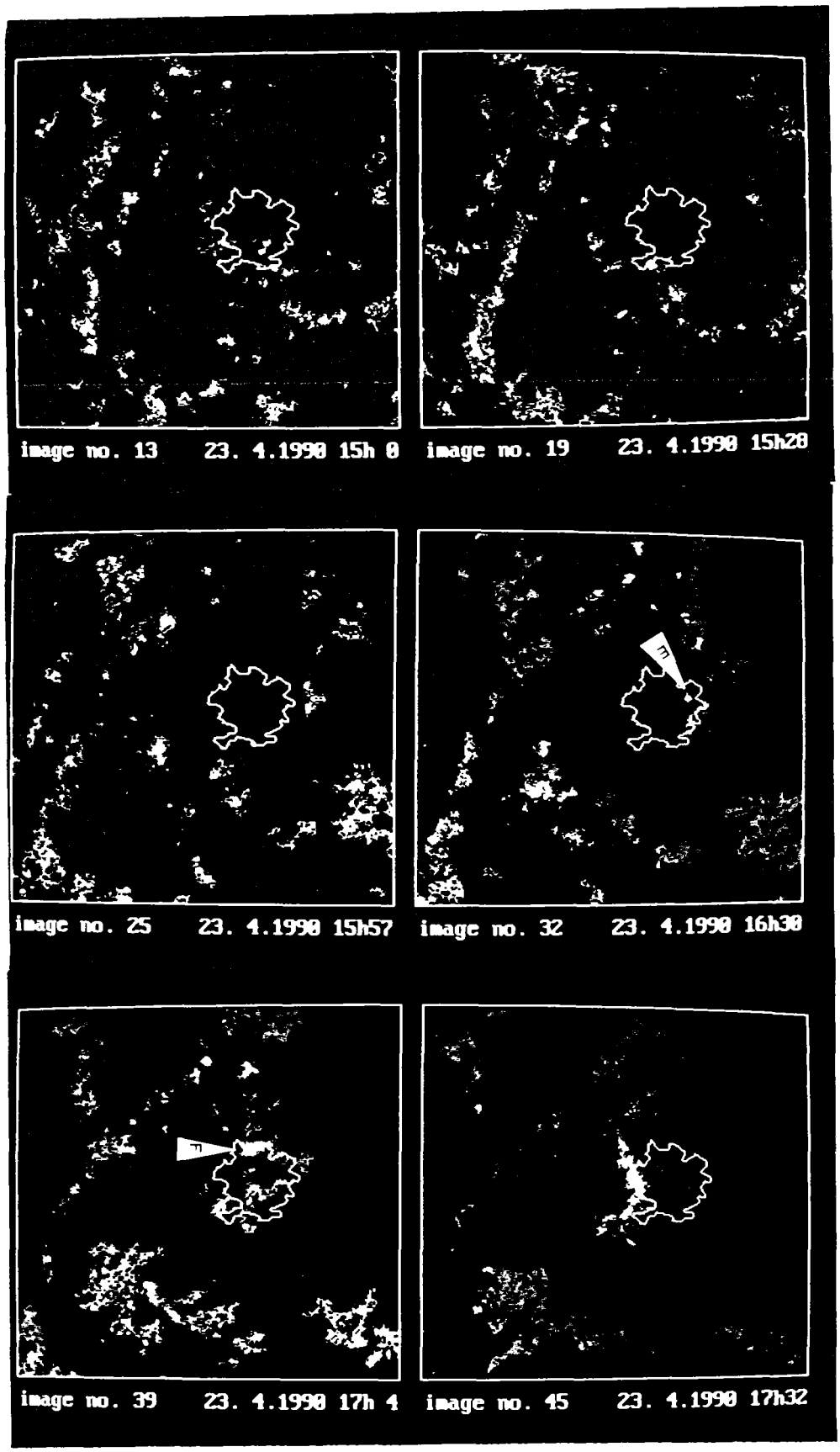
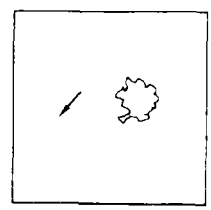


Figure A.3.14: Événement du 23.4.1990 (n° prév. = n° image -1)
(cellules E+F : cf. § VI.5.4)



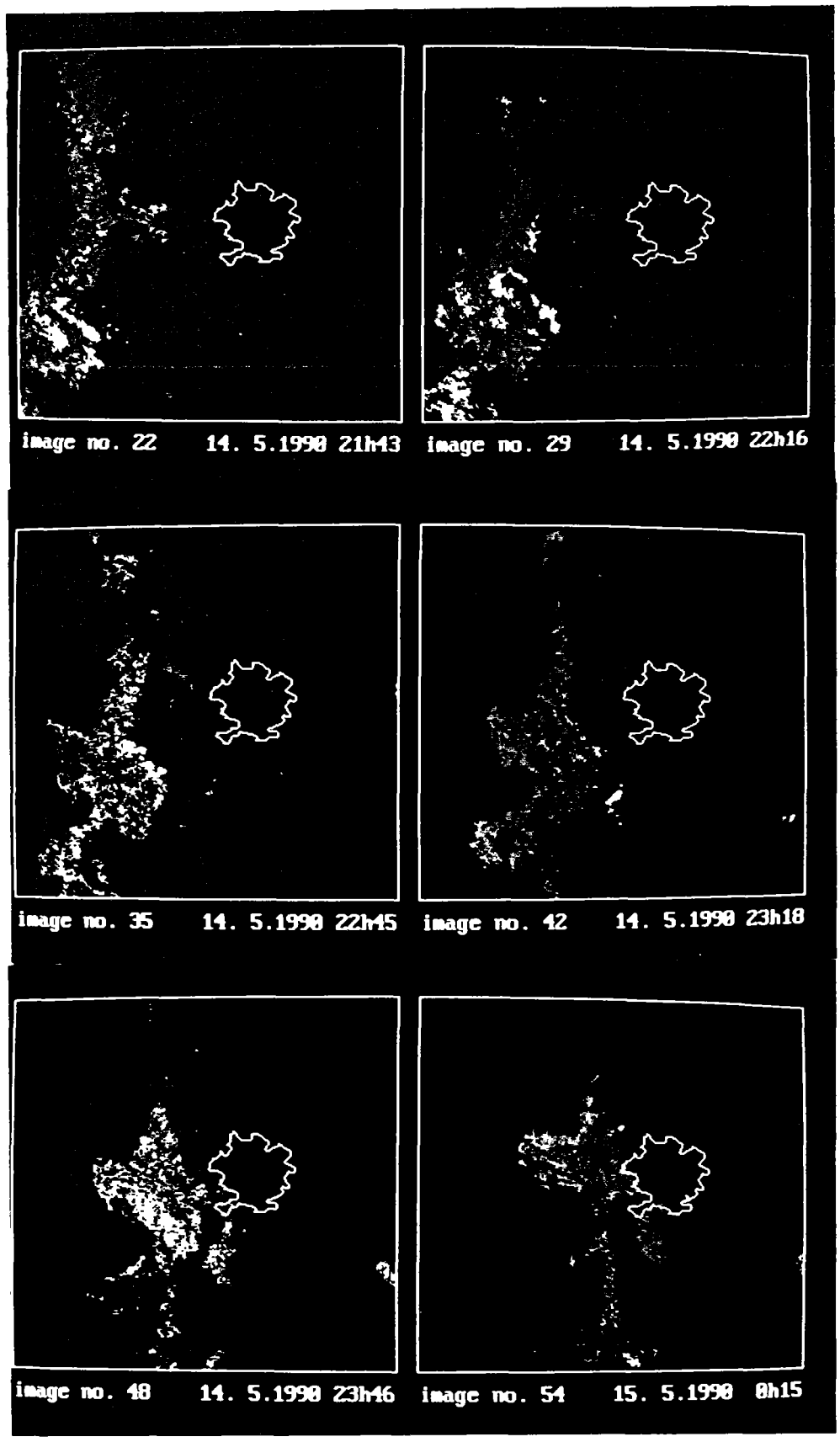
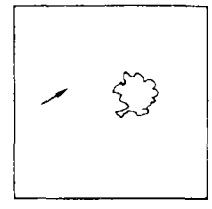


Figure A.3.15: Événement du 14.5.1990 (n° prév. = n° image -1)



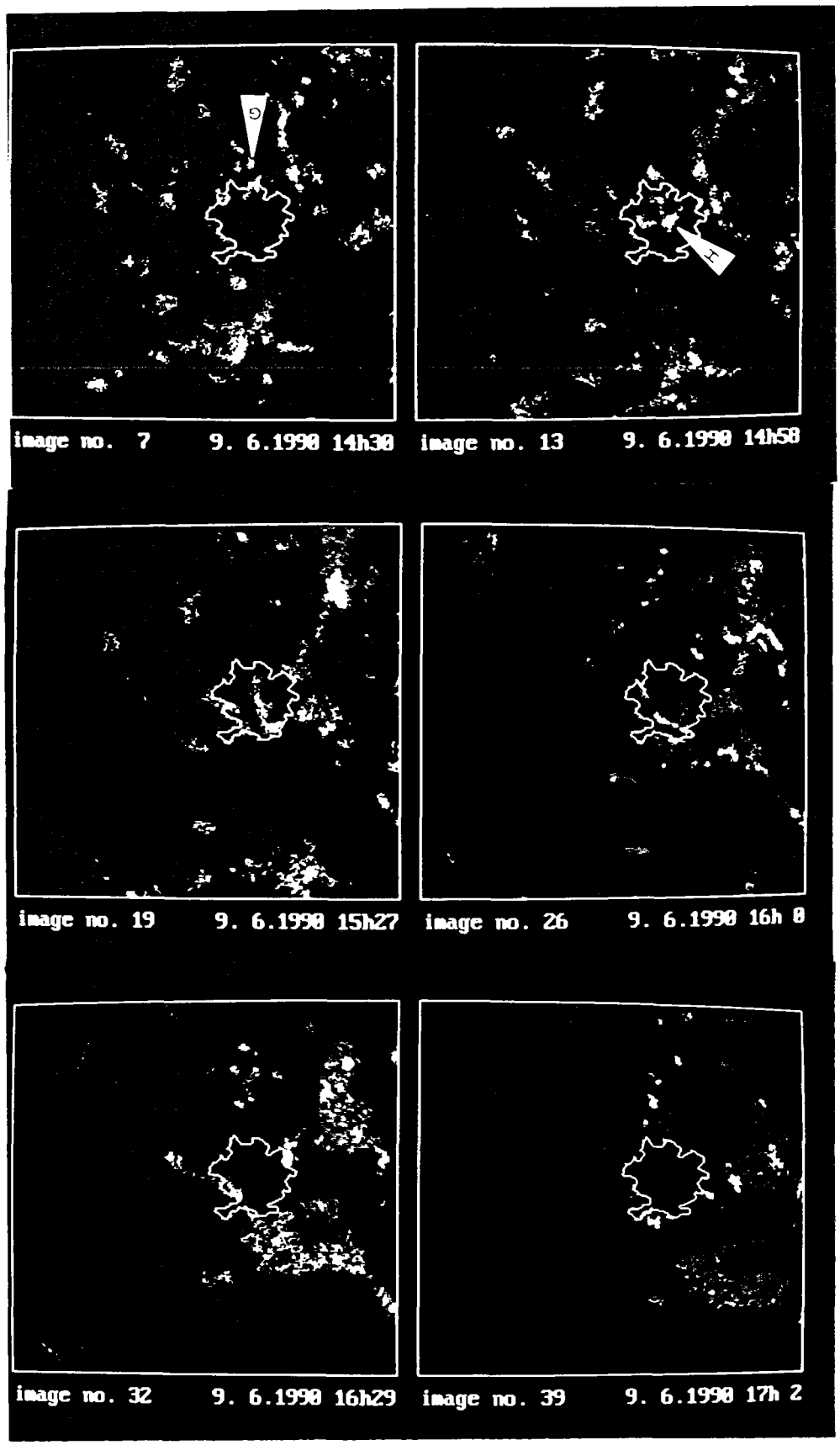
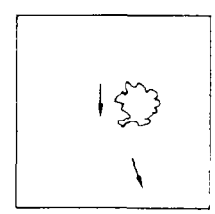


Figure A.3.16: Événement du 9.6.1990 (n° prév. = n° image -1)
(cellules G+H : cf. § VI.5.4)



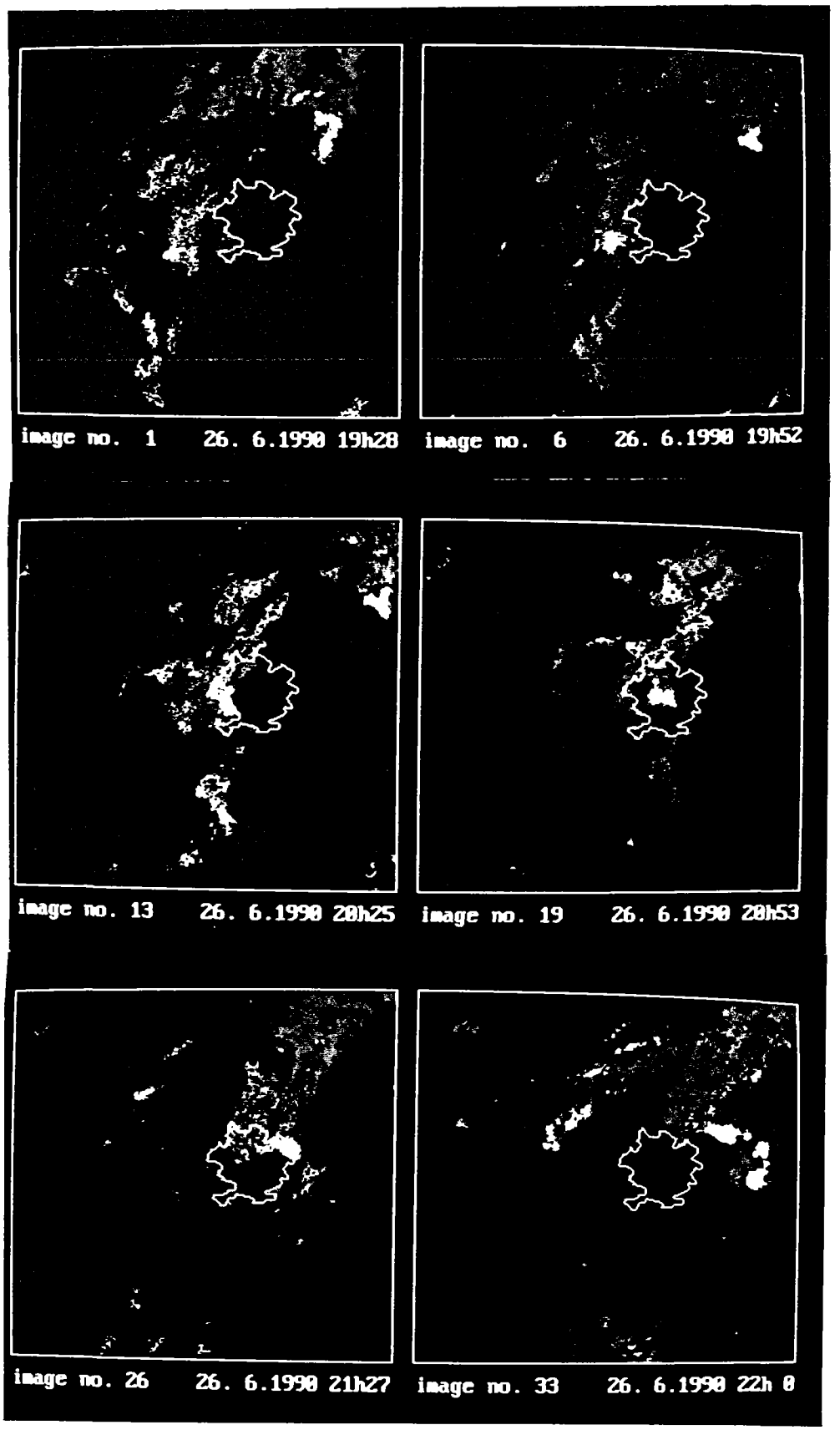


Figure A.3.17: Événement du 26.6.1990 (n° prév. = n° image -1)

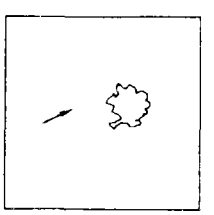
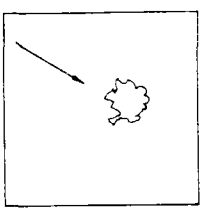


image no. 7 21. 9.1990 21h29 image no. 14 21. 9.1990 22h 2

image no. 20 21. 9.1990 22h31 image no. 26 21. 9.1990 22h59

image no. 32 21. 9.1990 23h28 image no. 38 21. 9.1990 23h56

Figure A.3.18: Événement du 21.9.1990 (n° prév. = n° image -1)
(zone marquée * : cf. § V.3.2.2 et § V.3.3)



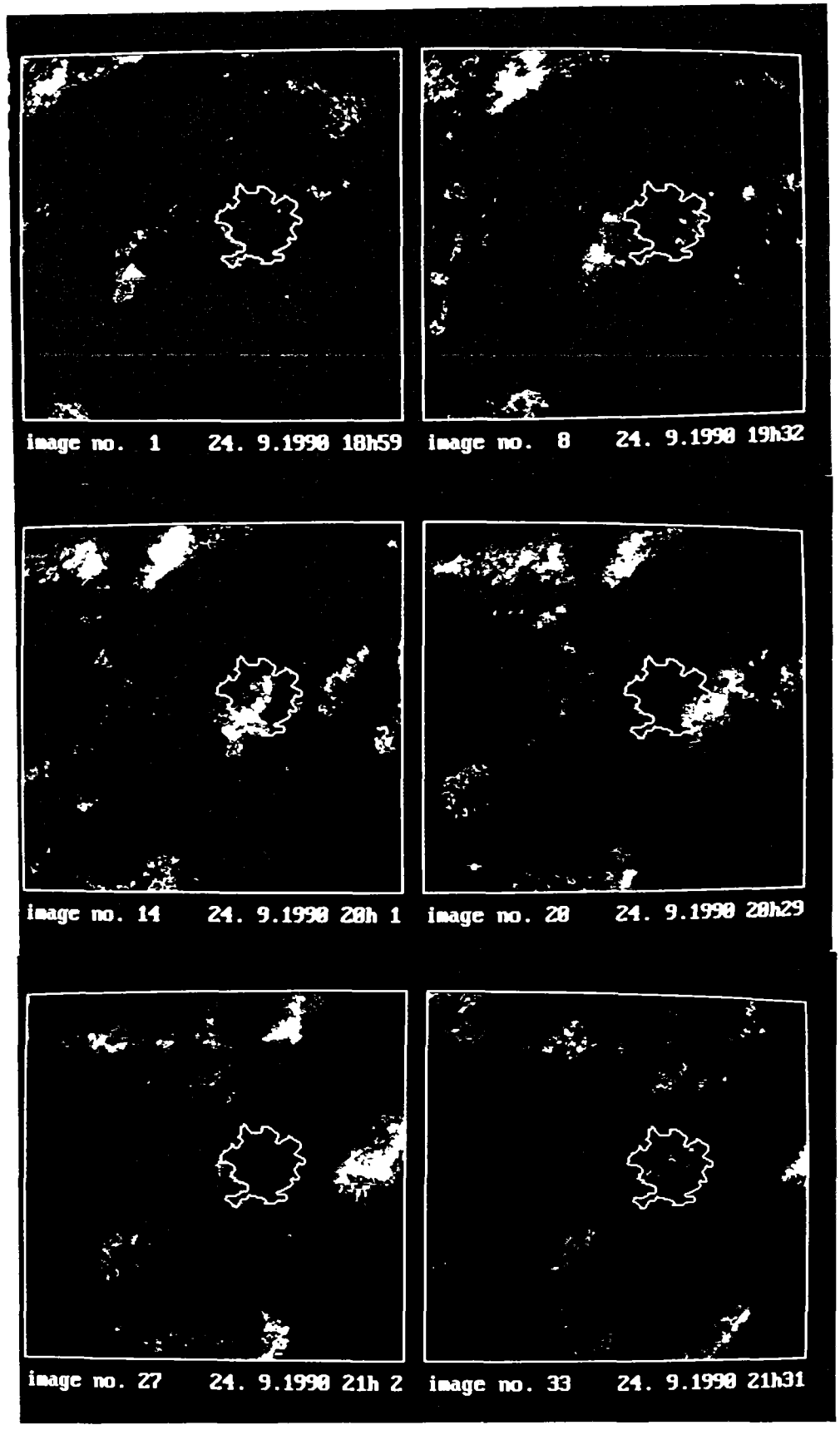
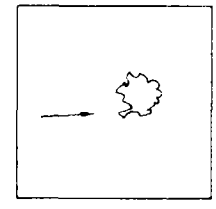


Figure A.3.19: Événement du 24.9.1990 (n° prév. = n° image -1)



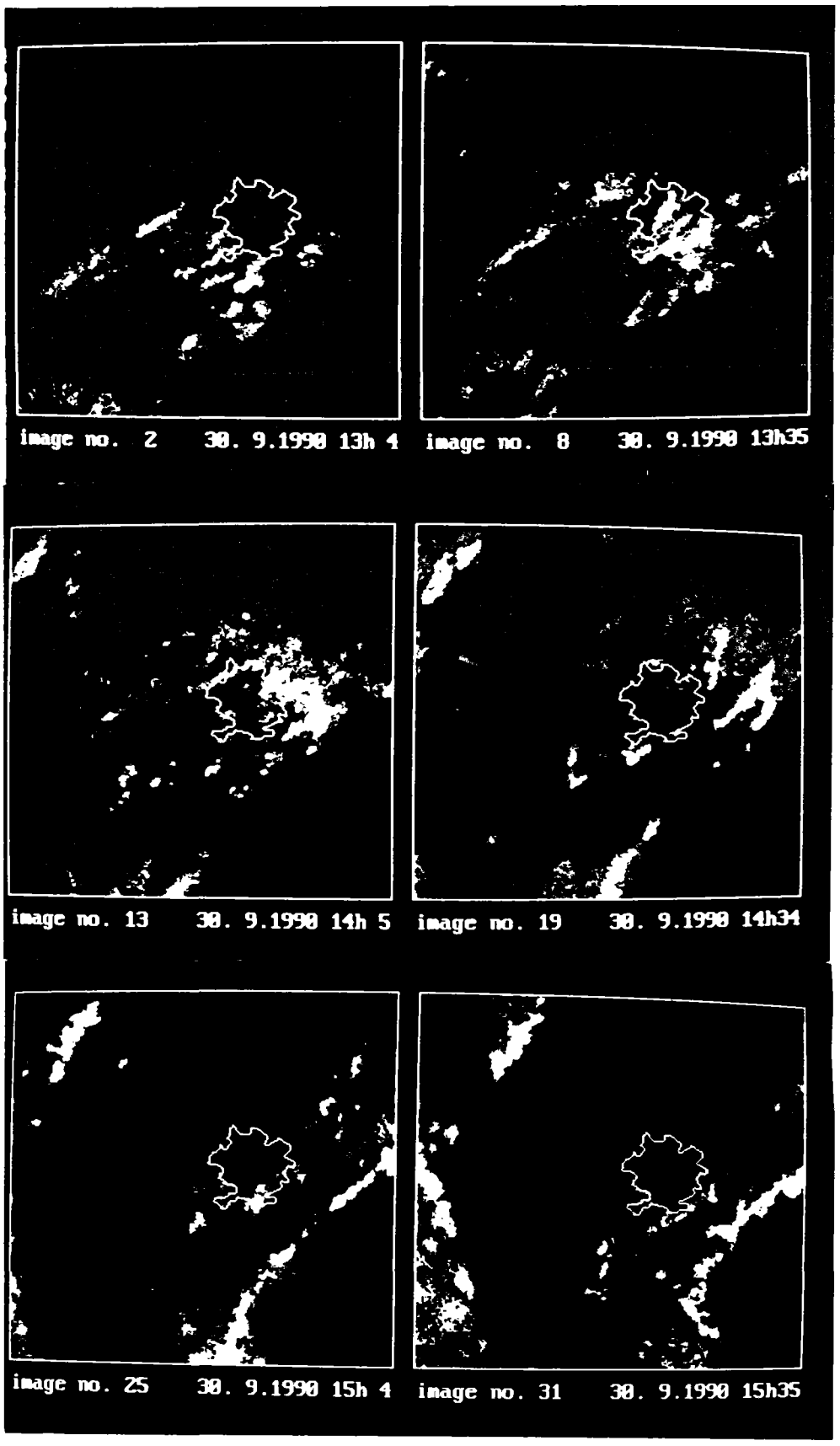
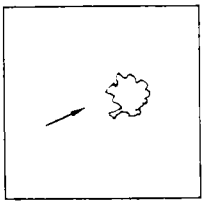



Figure A.3.20: Evénement du 30.9.1990 (n° prév. = n° image -1)