Analyse discriminante en prévision des crues

Stage réalisé à la DIREN Nord Pas-de-Calais
Master 2 Recherche de Mathématiques appliquées
Université des Sciences et Technologies de Lille

Anthony RIDARD
Septembre 2008

Table des matières

Remerciements

Je tiens avant tout à remercier Christophe Biernacki, mon responsable pédagogique à l'université, pour son enseignement de qualité, sa proposition de stage, sa disponibilité et tous ses précieux conseils.

Je souhaite aussi témoigner ma reconnaissance à Patrice Garnier, mon responsable à la DIREN, pour sa confiance et sa formidable curiosité scientifique qu'il partage avec tant de passion.

Un grand merci également à toute l'unité *Prévision des Crues, Hydrologie et Risques Naturels* pour son accueil, ainsi qu'à son directeur, Julien Henique, qui m'a permis de réaliser ce stage dans des conditions quelque peu particulières.

Je remercie enfin Oum-Salma Azahaf, ex-stagiaire du M2 Recherche en CDD à la DIREN, pour son travail sur les réseaux de neurones et sa générosité, sans oublier l'équipe de MIXMOD et en particulier Florent Langrognet, ingénieur de recherche CNRS, pour sa réactivité et son efficacité.

Introduction

Depuis maintenant deux ans, la DIrection Régionale de l'ENvironnement (DIREN) accueille des stagiaires du Master 2 Recherche de Mathématiques appliquées de Lille 1. L'objectif étant de développer des modèles de prévision de nature statistique, complétant ainsi les modèles de nature physique déjà existants.

Ainsi, des modèles autorégressifs ont d'abord été testés, sans succès contrairement aux réseaux de neurones mis en place l'année suivante et utilisés aujourd'hui. Cette année, un nouveau modèle sera étudié : l'analyse discriminante.

Après une présentation rapide de la DIREN et de son Service de Prévision des Crues (SPC), nous rentrerons dans le vif du sujet. Nous justifierons d'abord cette nouvelle approche avant d'en exposer les principales idées. Nous comparerons ensuite l'analyse discriminante aux réseaux de neurones puis, nous discuterons le choix des variables. Enfin, nous rendrons ce nouveau modèle opérationnel sous Excel avant d'en évoquer quelques perspectives.

I. La DIREN et son Service de Prévision des Crues

1. La DIREN Nord Pas-de-Calais

La DIREN est un service déconcentré du Ministère de l'écologie, du développement et de l'aménagement durables. Elle conduit des politiques en faveur de l'environnement et concourt à son intégration dans toutes les autres politiques. Elle assure ses missions au cœur du service public de l'environnement dans les départements, la région, le bassin Artois-Picardie et le district hydrographique international de l'Escaut.

Missions principales

- Déclinaison aux contextes locaux des engagements nationaux et européens.
- Développement et diffusion des connaissances et de l'évaluation environnementale.
- Protection et réhabilitation des ressources et des milieux naturels.

Gestion du risque

La région connaît des inondations à l'origine de dommages considérables pour les personnes, les biens et les activités. Aussi, la surveillance des débits, la prévision des inondations et la prévention des risques naturels sont des priorités de la DIREN.

2. Le Service de Prévision des Crues

Le SPC Artois-Picardie est basé à Lille, au sein de la DIREN Nord Pas-de-Calais. Il comprend 4 prévisionnistes à temps plein, et fait partie de l'unité *Prévision des Crues, Hydrologie et Risques Naturels*, au sein du Service de l'Eau, des Milieux Aquatiques et des Risques Naturels (SEMARN).

Missions principales

- Surveillance, prévision et transmission de l'information sur les crues des tronçons de cours d'eau surveillés par l'Etat, via la procédure de vigilance mise en place depuis le 11 juillet 2006.
- Capitalisation d'informations et expertise dans le domaine des inondations.
- Appui aux collectivités souhaitant mettre en place, pour leurs besoins propres et sous leur responsabilité, une surveillance des crues sur des cours d'eau non surveillés par l'Etat.

Procédure de vigilance

Le SPC Artois Picardie a en charge 7 tronçons de cours d'eau (l'Aa, la Liane, la Sambre, l'Helpe mineure, l'Helpe majeure, la Solre, la Somme) pour l'application de la procédure de vigilance[1].

Sur ces tronçons, un bulletin d'information est produit, attribuant à chaque tronçon une couleur de vigilance (verte, jaune, orange ou rouge) et indiquant le niveau de crue auquel on peut s'attendre dans les prochaines 24 h. Le vert correspond à une situation normale, le jaune à un risque de crue modeste, l'orange à un risque de crue importante et le rouge à un risque de crue exceptionnelle. Le bulletin comporte également des commentaires sur la situation en cours et son évolution prévue, les conséquences possibles ainsi que des conseils de comportement.

La prévision est envoyée a minima deux fois par jour au SCHAPI (Service Central d'Hydro-météorologie et d'Appui à la Prévisions des Inondations), week-end et jours fériés compris, par un bulletin en début de matinée et en début d'après-midi, dans le but d'alimenter le dispositif national. Le SCHAPI collecte l'ensemble des bulletins des différents SPC de France afin de publier la carte nationale de vigilance[2] à 10h et à 16h sur le site internet http ://www.vigicrues.ecologie.gouv.fr/.

[1] cf Fig. 1 en Annexes
[2] cf Fig. 2 en Annexes

3. Comment le SPC prévoit-il aujourd'hui ?

Les prévisions de précipitations produites par Météo-France sont reçues et analysées par le SPC le matin et en début d'après-midi. Elles sont une base fondamentale de la prévision, car bien sûr, les précipitations sont l'élément déclencheur essentiel des crues. Le SPC dispose également des images radar et des images de satellite dans le but de mieux comprendre et d'appréhender le déplacement des masses pluvieuses.

Le SPC dispose de différents modèles numériques adaptés à chaque bassin, permettant d'obtenir des simulations de débits ou niveaux futurs. Ces modèles, calés à partir de crues anciennes, donnent une idée du comportement des cours d'eau et de leurs réactions habituellement observées aux précipitations. Ils utilisent des données de pluies, d'évapotranspiration, de débits à une station amont, des relations mathématiques plus ou moins simples.

Le SPC est par ailleurs doté d'un outil d'aide à la décision basé sur des abaques calculés à partir de crues passées, sur les cumuls de pluies observés et prévus, sur l'état hydrique des sols, et d'autres paramètres, qui permet d'estimer un pic de crue pour les stations de prévision des différents tronçons dont il a la charge.

Enfin, pour faire la part des choses et choisir le modèle le plus efficace pour la situation donnée, l'expérience des prévisionnistes chevronnés est un atout majeur.

Le SPC est aussi doté d'un superviseur d'alerte appelé SCAPIN (Système Centralisé d'Alertes et de Prévention des Inondations), principal outil de surveillance de l'ensemble des bassins. Il capitalise l'information de toutes les collectes des stations surveillées, en terme de cote ou de débit. Il prévient le prévisionniste par des alertes téléphoniques en cas de dépassement de seuils indicatifs particuliers sur certaines stations, ce qui permet d'être averti, lors d'évènements non prévus, ou en cas d'aggravation de la situation. De plus, il alerte également le prévisionniste en cas de défaillance au niveau des stations de mesure, car sans mesure de la quantité de pluie tombée, du niveau ou du débit, il ne peut pas apprécier et suivre la situation.

Il est très difficile de prévoir exactement la cote d'un cours d'eau plusieurs heures à l'avance, et ce même lorsque la prévision de Météo France est très précise. De nombreux paramètres interviennent, beaucoup d'incertitudes s'ajoutent. En effet, la réaction du cours d'eau à une pluie sur son bassin versant va dépendre de plusieurs paramètres :

- Le niveau initial d'humidité du sol (qui dépend de la pluie tombée des derniers jours). Notamment, la réaction d'un cours d'eau à une même pluie l'hiver ou l'été pourra être très différente.
- Le niveau des nappes souterraines qui parfois alimente en débit le cours d'eau.
- L'intensité des pluies : une même quantité de pluie qui s'abat sur 2 heures ou sur 24 heures provoquera une réaction du cours d'eau très différente.
- L'hétérogénéité spatiale de la pluie tombée sur le bassin versant : lorsque la pluie qui tombe n'est pas uniforme sur le bassin versant, les affluents ne réagissent pas tous de la même manière et la réaction globale du cours d'eau pourra être différente d'un évènement à l'autre.

De plus il arrive parfois que les prévisions de Météo France ne se vérifient pas, la prévision des pluies étant un exercice difficile, or ce sont les pluies qui font les crues. Le travail du SPC est donc tributaire de celui de Météo France.

Pour améliorer ses prévisions le SPC essaie d'acquérir des outils plus complexes, intégrant de nouvelles variables comme l'évapotranspiration potentielle ou la saturation des sols, et modélisant les bassins de façon différente (modèles à réservoirs), en faisant appel à des partenaires extérieurs tels que des bureaux d'étude, le BRGM, le CEMAGREF, des universités[3]...

[3]Les réseaux de neurones, utilisés aujourd'hui, ont par exemple été mis en place par Oum-Salma AZAHAF sous la direction de Patrice Garnier à la DIREN et Christophe Biernacki à l'université de Lille 1.

II. Une nouvelle approche par l'analyse discriminante

1. Un nouveau modèle, pourquoi ?

Plus qu'un nouveau modèle, il s'agit là d'une nouvelle approche. Rappelons que l'objectif premier de la procédure de vigilance est la production d'un bulletin d'information attribuant, à chaque tronçon, une couleur de vigilance (Vert, Jaune, Orange ou Rouge). Actuellement, les modèles utilisés ont pour but d'estimer la hauteur future d'un cours d'eau de manière à en déduire la couleur de vigilance. Le nouveau modèle, quant à lui, propose une estimation *directe* de cette couleur. Fondamentalement, il s'agit d'estimer une variable catégorielle (la couleur de vigilance) au lieu d'une grandeur continue (la hauteur du cours d'eau). D'un point de vue statistique, le problème est alors ramené à un contexte d'apprentissage supervisé, appelée aussi analyse discriminante. Une approche probabiliste, s'appuyant sur des modélisations gaussiennes multivariées de classes, permettra alors d'estimer directement non seulement la zone de vigilance mais aussi la probabilité de se trouver dans chacune des zones.

2. L'analyse discriminante en quelques mots

L'idée est de voir les données comme des réalisations indépendantes d'une loi mélange dont on estimera les paramètres pour définir une règle de classement permettant d'associer un nouvel individu à sa classe.

Considérons alors un échantillon $x = \{x_1, \ldots, x_n\}$ d'une loi mélange de densité

$$f(x_i|\theta) = \sum_{k=1}^{K} p_k h(x_i|\lambda_k)$$

où p_k désigne la proportion de la k^{e} composante du mélange, $h(.|\lambda_k)$ sa distribution paramétrée par λ_k et $\theta = (p_1, \ldots, p_K, \lambda_1, \ldots, \lambda_K)$ le paramètre mélange.

L'apprentissage étant supervisé, nous supposerons connue la classe de chaque individu x_i c'est à dire la composante du mélange dont il est issu. Soit donc $z = \{z_1, \ldots, z_n\}$ la partition correspondante où $z_i = (z_{i1}, \ldots, z_{iK})$ avec $z_{ik} = 1$ ou 0 selon que x_i provienne de la k^{e} composante ou non.

Dans ces conditions, les $(x_1, z_1), \ldots, (x_n, z_n)$ sont des réalisations indépendantes de la loi de densité

$$f(x_i, z_i|\theta) = \prod_{k=1}^{K} p_k^{z_{ik}} [h(x_i|\lambda_k)]^{z_{ik}}.$$

La méthode du maximum de vraisemblance permet alors d'obtenir une estimation du paramètre mélange

$$\hat{\theta} = \arg\max_{\theta} L(\theta|x, z)$$

où la log-vraisemblance $L(\theta|x, z)$ est définie par

$$L(\theta|x, z) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln(p_k h(x_i|\lambda_k)).$$

Nous pouvons enfin, grâce à cet estimateur, classer un nouvel individu x_{n+1} moyennant le maximum a posteriori (MAP). Il s'agit de calculer, pour chaque k, la probabilité pour que x_{n+1} provienne de la classe k c'est à dire la probabilité conditionnelle

$$t_k(x_{n+1}|\hat{\theta}) = \frac{\hat{p}_k h(x_{n+1}|\hat{\lambda}_k)}{\sum_{l=1}^{K} \hat{p}_l h(x_{n+1}|\hat{\lambda}_l)}$$

puis d'associer à x_{n+1} la classe maximisant cette probabilité autrement dit

$$\hat{z}_{n+1\,k} = \begin{cases} 1 \text{ si } k = \arg\max_l t_l(x_{n+1}|\hat{\theta}) \\ 0 \text{ sinon} \end{cases}.$$

3. Mise en place à l'aide de MIXMOD

MIXMOD est un logiciel permettant de modéliser un mélange (MIXture MODelling) de composantes gaussiennes multivariées (resp. multinomiales) à partir de données quantitatives (resp. qualitatives). Il peut donc se révéler utile non seulement en analyse discriminante (classement) mais aussi en classification ou estimation de densité.

Pour mener à bien notre analyse discriminante sous MIXMOD, nous disposons d'un échantillon d'apprentissage $(x_1, z_1), \ldots, (x_n, z_n)$ où x_i est un vecteur de $\mathbf{R}^d$ décrivant, selon d variables choisies, le i^{e} individu tandis que z_i en fournit la classe.

Evoquons déjà une première difficulté liée au caractère exceptionnel de la couleur rouge. Notre échantillon est en effet dépourvu du moindre représentant de cette classe, empêchant ainsi MIXMOD de considérer le rouge, et donc de le prévoir. Nous reviendrons plus tard sur ce formidable défi, prévoir ce qui ne s'est jamais produit, mais en attendant nous nous limiterons à 3 classes : V, J et O respectivement numérotées 1, 2 et 3.

Supposons alors que la densité de la loi mélange soit de la forme

$$f(x_i|\theta) = \sum_{k=1}^{3} p_k h(x_i|\lambda_k)$$

où $h(.|\lambda_k)$ désigne la densité gaussienne multivariée de la k^{e} composante paramétrée par $\lambda_k = (\mu_k, \Sigma_k)$ avec μ_k sa moyenne et Σ_k sa matrice de variance.

Notons ici que le logiciel permet d'imposer certaines contraintes sur le mélange, pouvant traduire des informations que nous aurions sur les données, et ainsi en limiter les paramètres. Nous pouvons, par exemple, imposer à Σ_k d'être diagonale (resp. scalaire) et ainsi choisir le modèle gaussien diagonal (resp. sphérique).

Précisons enfin les deux étapes de l'analyse discriminante sous MIXMOD :

1. Définition de la règle de classement à partir de l'échantillon d'apprentissage.
2. Classement de nouveaux individus décrits selon les d variables choisies.

En réalité, l'échantillon test composé des nouveaux individus à classer correspond à des données complètes autrement dit le classement de chaque individu est bien connu. Aussi, pouvons-nous évaluer notre règle en calculant non seulement les probabilités conditionnelles $\mathbf{P}(V|V)$, $\mathbf{P}(J|V)$, $\mathbf{P}(O|V)$, $\mathbf{P}(V|J)$, ... mais aussi l'erreur de classement définie par

$$e(r) = \sum_{k=1}^{3} p_k \sum_{l \neq k} \mathbf{P}(l|k).$$

Chaque règle de classement sera alors accompagnée de son évaluation :

```
P(V|V)   P(J|V)   P(O|V)
P(V|J)   P(J|J)   P(O|J)
P(V|O)   P(J|O)   P(O|O)

Erreur de classement
```

III. Analyse discriminante VS réseaux de neurones

1. Conditions expérimentales

Echantillons

L'échantillon d'apprentissage (10573 lignes) et l'échantillon test (31720 lignes) sont ceux utilisés pour les réseaux de neurones. Notons ici les proportions des trois classes dans l'échantillon test :
- $p1 = 0.9765448$
- $p2 = 0.0205864$
- $p3 = 0.0028689$

Choix de variables

Les variables[4] utilisées sont celles choisies pour les réseaux de neurones :
- Prévision à 3h : Pp6, Pp3, Pf3, Hm3, W10995, Hp6, H
- Prévision à 6h : Pp3, Pf3, Pf6, Hm3, W10995, Hp3, H
- Prévision à 24h : Pp12, Pp2, Pf15, Pf24, W10995, Hm2, H

Choix de modèle

Trois modèles gaussiens ont été testés[5] :
- $p_k L_k C_k$
- $p_k L_k B_k$
- $p_k L_k I$

Modèle général $p_k L_k C_k$:

```
94.131027    5.820726    0.048248
19.656236   76.068753    4.275011
 0.000000   30.875576   69.124424

7.503153
```

Modèle diagonal $p_k L_k B_k$:

```
93.676121    6.092980    0.230899
18.378140   75.055090    6.566770
 0.000000   43.317972   56.682028

8.162043
```

Modèle sphérique $p_k L_k I$:

```
90.209188    9.642623    0.148189
30.409872   63.596298    5.993830
 5.069124   41.244240   53.686636

12.194199
```

Le modèle retenu pour la comparaison est le modèle général $p_k L_k C_k$.

[4] Nous y reviendrons plus en détail par la suite.
[5] Les variables utilisées sont celles de la prévision à 24h.

2. Comparaison des modèles

Prévision à 3h

Analyse discriminante :

```
95.541298    4.455446    0.003257
 0.227273   98.068182    1.704545
 0.000000   16.911765   83.088235

4.441992
```

Réseaux de neurones :

```
99.830641    0.169359    0.000000
11.818182   88.181818    0.000000
 0.000000   16.176471   83.823529

0.561160
```

Prévision à 6h

Analyse discriminante :

```
91.197762    8.374465    0.427772
 3.896104   83.463203   12.640693
 0.000000   22.857143   77.142857

9.161412
```

Réseaux de neurones :

```
99.516288    0.483712    0.000000
21.298701   77.835498    0.865801
 0.000000   30.285714   69.714286

1.437579
```

Prévision à 24h

Analyse discriminante :

```
94.131027    5.820726    0.048248
19.656236   76.068753    4.275011
 0.000000   30.875576   69.124424

7.503153
```

Réseaux de neurones :

```
98.569804    1.416411    0.013785
30.101366   69.105333    0.793301
 0.000000   53.225806   46.774194

4.246532
```

Commentaires

Au regard de l'erreur, l'AD est moins performante que les réseaux. Pourtant, le nouveau modèle offre une bien meilleure gestion des J et des O. En fait, l'AD prévoit du J au lieu du V plus souvent que les réseaux, ce qui est très pénalisant au niveau de l'erreur compte tenu de la proportion de V. Notons maintenant que parmi les mauvaises prévisions, celle qui fait défaut à l'AD est en réalité la moins dommageable. Il serait peut-être alors intéressant d'introduire des coûts de manière à pénaliser au niveau de l'erreur ce qui est le plus impactant dans la "vraie vie".

3. La notion de coûts

L'idée est d'associer, à chaque erreur, un coût et ainsi, prendre en compte les conséquences, plus ou moins graves, d'une mauvaise prévision. Notons que ce raffinement a une influence non seulement sur le calcul de l'erreur de classement, appelée ici risque[6], mais aussi sur le classement lui même[7].

Dans notre cas, la matrice "coût" utilisée est la suivante :

$$C = \begin{pmatrix} 0 & 3 & 50 \\ 1 & 0 & 10 \\ 50 & 5 & 0 \end{pmatrix}$$

où $C(i,j)$ représente le coût de prévoir i au lieu de j.

Précisons que les coûts sont à interpréter de manière relative et non absolue. Par exemple, il est trois fois plus dommageable de prévoir du V au lieu du J que d'annoncer du J au lieu du V. En interprétant ainsi la matrice C, nous pouvons ranger les erreurs par ordre décroissant de gravité :
- Se tromper de deux classes (50)
- Râter un O en mettant du J (10)
- Râter un J en mettant du O (5)
- Râter un J en mettant du V (3)
- Râter un V en mettant du J (1)

Reprenons maintenant, dans ce cadre, la comparaison précédente.

Prévision à 3h

Analyse discriminante :

```
95.052762     4.943981     0.003257
 0.113636    96.704545     3.181818
 0.000000    13.970588    86.029412

155.069645
```

Réseaux de neurones :

```
99.830641     0.169359     0.000000
11.818182    88.181818     0.000000
 0.000000    16.176471    83.823529

195.547501
```

[6]Le risque est défini par $R(r) = \sum_{k=1}^{3} p_k \sum_{l \neq k} C(l,k)\mathbf{P}(l|k)$ où $C(l,k)$ représente le coût de prévoir l au lieu de k

[7]Le classement n'est plus réalisé moyennant le MAP mais bien de manière à minimiser le risque

Prévision à 6h

Analyse discriminante :

```
89.746627    9.700559    0.552813
 3.116883   81.212121   15.670996
 0.000000   17.714286   82.285714

262.243309
```

Réseaux de neurones :

```
99.516288    0.483712    0.000000
21.298701   77.835498    0.865801
 0.000000   30.285714   69.714286

366.947426
```

Prévision à 24h

Analyse discriminante :

```
93.042010    6.882173    0.075818
13.750551   80.343764    5.905685
 0.000000   22.350230   77.649770

287.070822
```

Réseaux de neurones :

```
98.569804    1.416411    0.013785
30.101366   69.105333    0.793301
 0.000000   53.225806   46.774194

612.682250
```

Commentaires

Nous pouvons déjà noter que cette notion de coûts améliore les prévisions de l'AD, au moins en ce qui concerne le O (et le J à 24h). Quant au risque, il apparaît nettement moins important en analyse discriminante qu'avec les réseaux. Aussi, les performances du nouveau modèle sont-elles bien mises en valeur qu'il s'agisse d'une prévision à 3h, 6h ou 24h.

Nous aimerions par ailleurs attirer l'attention du lecteur sur la subjectivité engagée dans le choix de la matrice "coût". C'est donc dans un souci d'objectivité que nous conserverons, tout au long du rapport, les deux points de vue, avec et sans coût. Cette précaution devrait en prime nous fournir des indications quant au gain (ou au danger) de cette notion de coûts. Nous présenterons alors les résultats sous la forme suivante :

```
P(V|V)   P(J|V)   P(O|V)
P(V|J)   P(J|J)   P(O|J)
P(V|O)   P(J|O)   P(O|O)

Risque
```

```
P(V|V)   P(J|V)   P(O|V)
P(V|J)   P(J|J)   P(O|J)
P(V|O)   P(J|O)   P(O|O)

Erreur de classement
```

IV. Retour sur le choix de variables

1. Les variables

Revenons ici sur les variables existantes et remettons en question le choix des réseaux.

Les données disponibles concernant la Liane (cours d'eau choisi pour la mise en place du modèle) couvrent la période du 01/01/1998 à 00 :30 au 30/03/2007 à 23 :30.

Variable du présent

– H : Hauteur du cours d'eau (à l'instant t)[8]

Variables du passé

– Hpi : Hauteur du cours d'eau il y'a i heures (à l'instant $t - i$)
– Ppi : Pluie passée des i dernières heures
– Hmi : Hauteur moyenne du cours d'eau des i dernières heures

Variables du futur

– Hfi : Hauteur du cours d'eau dans i heures (à l'instant $t + i$)
– Pfi : Pluie future des i prochaines heures

Variables saisonnières

– ETP : Evapotranspiration Potentielle[9]
– $W10995$: Humidité

2. Un choix minimaliste

Regardons notre modèle ($p_k L_k C_k$) en dimension 3 avec une variable présente, une passée et une future.

Prévision à 3h

Les variables H, $Pp6$ et $Pf1$ fournissent :

96.707269	3.253648	0.039083
1.704545	95.795455	2.500000
0.000000	8.088235	91.911765
97.827361		

97.202319	2.761855	0.035826
2.840909	96.136364	1.022727
0.000000	10.294118	89.705882
2.859395		

Prévision à 6h

Les variables H, $Pp6$ et $Pf3$ fournissent :

95.860480	4.060546	0.078973
6.580087	90.129870	3.290043
0.000000	14.285714	85.714286
177.277514		

96.640342	3.287266	0.072392
8.571429	89.350649	2.077922
0.000000	22.285714	77.714286
3.729508		

[8] Il s'agit en fait d'une variable du passé "récent" étant donné le temps de réaction du bassin (environ 7 heures).
[9] Somme des volumes d'eau utilisés par les plantes (eau de constitution, eau de végétation) et évaporés à la surface du sol

Prévision à 24h

Les variables H, $Pp6$ et $Pf21$ fournissent :

94.089672	5.703553	0.206775
15.689731	77.434993	6.875275
0.000000	24.884793	75.115207
322.429764		

95.430265	4.428439	0.141296
25.473777	69.501983	5.024240
0.230415	30.645161	69.124424
6.784363		

Commentaires

Dans le cas d'une prévision à 24h, remarquons que la perte d'informations (trois variables au lieu de sept) dégrade la prévision du J et du O, augmentant du même coup le risque associé. Profitons d'ailleurs de cette occasion pour justifier, une fois de plus, la pertinence de la notion de coûts. Elle fournit en effet, via le risque, un critère de choix en parfait accord avec les priorités fixées : râter le moins possible de O et de J. Pour s'en convaincre, il suffit de regarder l'erreur qui se trouve améliorée par une légère progression de $\mathbf{P}(V|V)$, compte tenu de la proportion de V, alors que le risque, lui, ne se laisse pas abuser.
Dans le cas d'une prévision à 3h ou 6h, le constat est beaucoup plus surprenant. La réduction des variables semble en effet profiter à la gestion du O (et du J à 6h) améliorant ainsi les performances du modèle.

3. Apport de l'humidité

Voyons maintenant ce qu'apporte la variable saisonnière $W10995$.

Prévision à 3h

96.769151	3.185253	0.045597
1.818182	96.363636	1.818182
0.000000	12.500000	87.500000
138.781032		

97.309797	2.644607	0.045597
2.727273	96.022727	1.250000
0.000000	18.382353	81.617647
2.793190		

Prévision à 6h

95.949325	4.030931	0.019743
6.320346	89.870130	3.809524
0.000000	17.142857	82.857143
207.317875		

96.630471	3.366239	0.003291
8.398268	89.523810	2.077922
0.000000	21.714286	78.285714
3.729508		

Prévision à 24h

94.386050	5.507117	0.106834
15.028647	78.845306	6.126047
0.000000	26.036866	73.963134
328.030751		

95.795568	4.114829	0.089603
24.107536	71.220802	4.671662
0.000000	32.949309	67.050691
6.355612		

Commentaires

Au regard du risque, l'humidité semble dégrader les trois prévisions.
Dans le cas d'une prévision à 3h ou 6h, il s'agit peut-être d'un excès d'informations[10].
Quant à la prévision à 24h, plus gourmande en variables, cela nous interroge sur l'humidité même.

[10] Suivant l'idée que "trop d'infos tue l'info"

4. Notre choix de variables

En imaginant que les variables à choisir dépendent moins du modèle utilisé que du phénomène à prévoir, revenons un instant sur les variables des réseaux en remplaçant néanmoins l'humidité, remise en question précédemment, par l'autre variable saisonnière soit l'ETP.

Prévision à 3h

95.013679	4.986321	0.000000
0.113636	98.068182	1.818182
0.000000	13.970588	86.029412
148.436758		

95.528270	4.471730	0.000000
0.227273	98.522727	1.250000
0.000000	17.647059	82.352941
4.445145		

Prévision à 6h

94.402764	5.524844	0.072392
3.549784	91.601732	4.848485
0.000000	7.428571	92.571429
107.880583		

94.840408	5.090490	0.069102
4.155844	92.467532	3.376623
0.000000	10.285714	89.714286
5.274275		

Prévision à 24h

92.880036	6.899404	0.220560
13.089467	80.343764	6.566770
0.000000	18.894009	81.105991
254.827265		

94.113795	5.696661	0.189544
18.422212	76.597620	4.980167
0.000000	31.105991	68.894009
7.484237		

Commentaires

Dans tous les cas, toujours au regard du risque, l'ETP apparaît plus pertinente que l'humidité. Notons cependant que la prévision à 3h n'aura jamais été aussi bonne qu'en dimension trois[11].

Conclusion

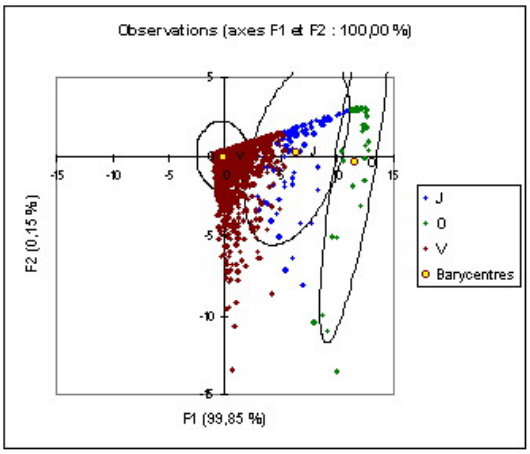
Finalement, les variables choisies sont :
- Prévision à 3h : H, Pp6, Pf1
- Prévision à 6h : Pp3, Pf3, Pf6, Hm3, ETP, Hp3, H
- Prévision à 24h : Pp12, Pp2, Pf15, Pf24, ETP, Hm2, H

5. Analyse factorielle discriminante avec XLSTAT

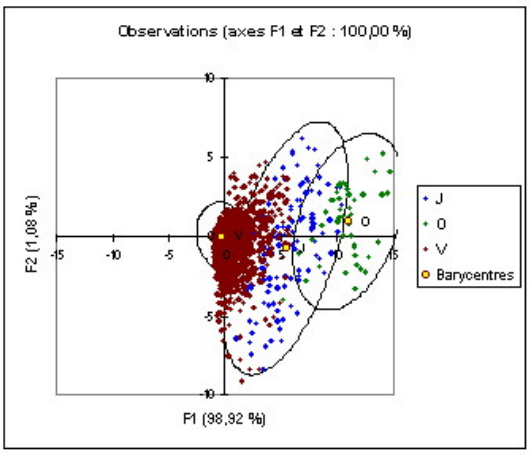
Terminons par une illustration, réalisée avec XLSTAT, représentant les différentes classes suivant les deux facteurs les plus discriminants. Un échantillon de 6000 données décrites par les variables choisies précedemment fournit :

[11] Les modèles physiques utilisés pour la prévision à 3h comptent, eux aussi, trois variables

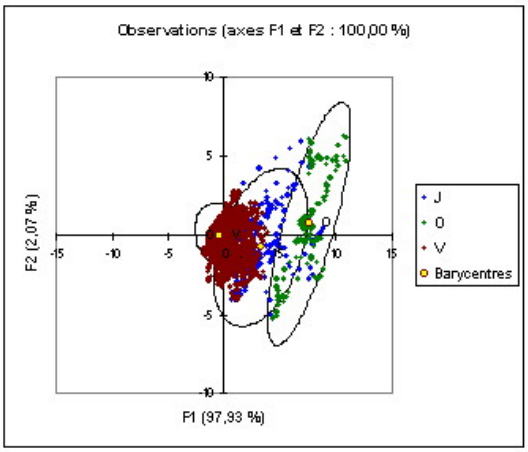
Prévision à 3h



Prévision à 6h



Prévision à 24h



V. Mise en place du modèle sous Excel

1. Les besoins opérationnels

Afin de rédiger son bulletin d'information, le prévisionniste s'appuie sur une plate-forme Excel, appelée "Cassandre", regroupant des abaques et les réseaux de neurones. Plus précisément, à chaque modèle correspond une feuille qui affiche la couleur de vigilance, déterminée à partir des prévisions de précipitations produites par Météo-France. Aussi, l'objectif est-il de créer une feuille "Analyse discriminante" fournissant non seulement la couleur de vigilance mais aussi la probabilité de chaque couleur. Notons que ces probabilités participent à la mise en valeur du nouveau modèle en exprimant la confiance que l'on peut avoir dans notre prévision[12].

2. Les calculs et le résultat

Avec les notations du II, il s'agit d'abord de calculer, pour chaque k, la probabilité pour que x_{n+1} provienne de la classe k c'est à dire la probabilité conditionnelle

$$t_k(x_{n+1}|\hat{\theta}) = \frac{\hat{p}_k h(x_{n+1}|\hat{\lambda}_k)}{\sum_{l=1}^{K} \hat{p}_l h(x_{n+1}|\hat{\lambda}_l)}$$

où $h(x_{n+1}|\hat{\lambda}_k)$ désigne la densité gaussienne multivariée paramétrée par $\hat{\lambda}_k = (\hat{\mu}_k, \hat{\Sigma}_k)$ c'est à dire

$$h(x_{n+1}|\hat{\lambda}_k) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det\hat{\Sigma}_k}} \exp\Big[-\frac{1}{2}(x-\hat{\mu}_k)'\hat{\Sigma}_k^{-1}(x-\hat{\mu}_k)\Big]$$

et où les estimateurs $\hat{p}_k$, $\hat{\mu}_k$ et $\hat{\Sigma}_k$ sont fournis par MIXMOD.

Attention, il convient de rappeler ici que les coûts ne sont pas égaux et donc que le classement ne s'effectue pas suivant la probabilité maximale mais selon le risque minimal. Nous devons alors déterminer, pour chaque k, le risque d'associer x_{n+1} à la classe k c'est à dire

$$R_k(x_{n+1}|\hat{\theta}) = \sum_{l\neq k} C(k,l) t_l(x_{n+1}|\hat{\theta})$$

Etant donné un individu x à classer, voici alors ce qu'affiche la feuille Excel :

	A	B	C	D	E
1		x		Risques	Probas
2	Pp12	5,066000		20,7258753	0,000000
3	Pp2	5,033000		3,77146285	0,622854
4	Pf15	18,635000		3,11426896	0,377146
5	Pf24	40,400002			
6	ETP	0,000000			
7	Hm2	0,864333			
8	H	0,652000			
9					

Ici[13], le modèle indique au prévisionniste que le risque minimal est associé à la couleur orange malgré une probabilité d'appartenance non maximale $(37,7\%)$. Notons que dans la réalité, cet individu correspond bien à une alerte orange.

[12] Contrairement aux abaques, il est possible de fournir des intervalles de confiance pour les réseaux de neurones. D'ailleurs, cette étude sera peut-être l'objet d'un prochain travail de recherche entre l'université de Lille 1 et la DIREN

[13] Il s'agit d'une prévision à 24h sur la Liane

VI. Extension à deux autres cours d'eau

Le nouveau modèle ayant fait ses preuves sur la Liane, la DIREN a exprimé le désir de l'étendre à d'autres cours d'eau. L'idée est donc de mettre en place une procédure simple pemettant de réaliser cette extension. Nous allons pour cela nous appuyer sur le travail réalisé pour les réseaux de neurones. En effet, les données ont déjà été corrigées et les variables déjà choisies. Nous ne remettrons donc pas en cause, dans un premier temps en tout cas, le choix de variables proposé pour les réseaux[14].

Notons qu'un Mémento (cf Annexes) a été rédigé de manière à rendre la procédure accessible à tous.

1. La Hem

Cette extension, réalisée par mes soins, a été l'occasion de tester la procédure.

2. La Solre

C'est Mathieu Floquet, stagiaire en Master 1 Pro à l'université de Lille 1, qui s'en est chargé.

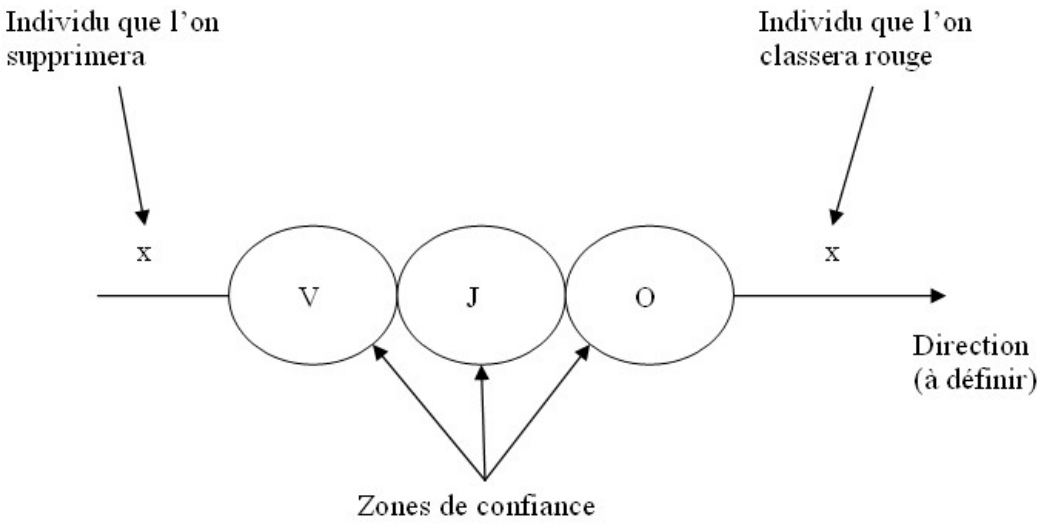
VII. Perspectives

Pour terminer, revenons un instant sur la classe rouge dont la prévision représente un formidable défi.

D'un point de vue pratique, il semble délicat de déclencher une alerte rouge mais détenir des informations allant dans ce sens peut néanmoins s'avérer déterminant en cas de crise. D'un point de vue plus théorique, prévoir ce qui ne s'est jamais produit relève moins du bon sens que de la recherche qui trouvera alors matière dans ce problème.

Voici par exemple une première réflexion. Notre modèle actuel nous permet de définir pour le vert, le jaune et le orange une zone de confiance, par exemple à 95%. Un individu, décrit par un certain nombre de variables, peut n'appartenir à aucune des zones ce qui lui laisse alors deux alternatives :
– soit il s'agit d'une erreur
– soit il s'agit de la classe rouge

On peut alors imaginer une direction (à définir) nous permettant de choisir entre ces deux possibilités :



[14] Même s'il peut être amélioré (cf étude précédente), ce choix est déjà satisfaisant

Conclusion

En résumé, disons que l'analyse discriminante a tenu toutes ses promesses.

Que ce soit à 3h, 6h ou 24h, les prévisions fournies par ce nouveau modèle sont tout à fait satisfaisantes. En comparaison avec les réseaux de neurones, nous retiendrons par exemple que l'analyse discriminante assure une bien meilleure gestion des jaunes et des oranges. Cette différence est d'ailleurs accentuée grâce à la notion de coûts, introduite pour mieux prendre en compte les conséquences d'une mauvaise prévision. Notons à ce sujet que le risque associé, critère de choix réaliste, exprime parfaitement la pertinence du nouveau modèle.

Toujours en faveur de l'analyse discriminante, notons que sa mise en place est aisée, grâce à MIXMOD, et que son utilisation est confortable pour le prévisionniste. Il dispose en effet, sous EXCEL, de la confiance qu'il peut accorder à sa prévision ou encore du risque qu'il prend.

Espérons enfin que ce nouveau modèle apporte entière satisfaction aux prévisionnistes du Nord Pas-de-Calais et d'ailleurs...

Annexes

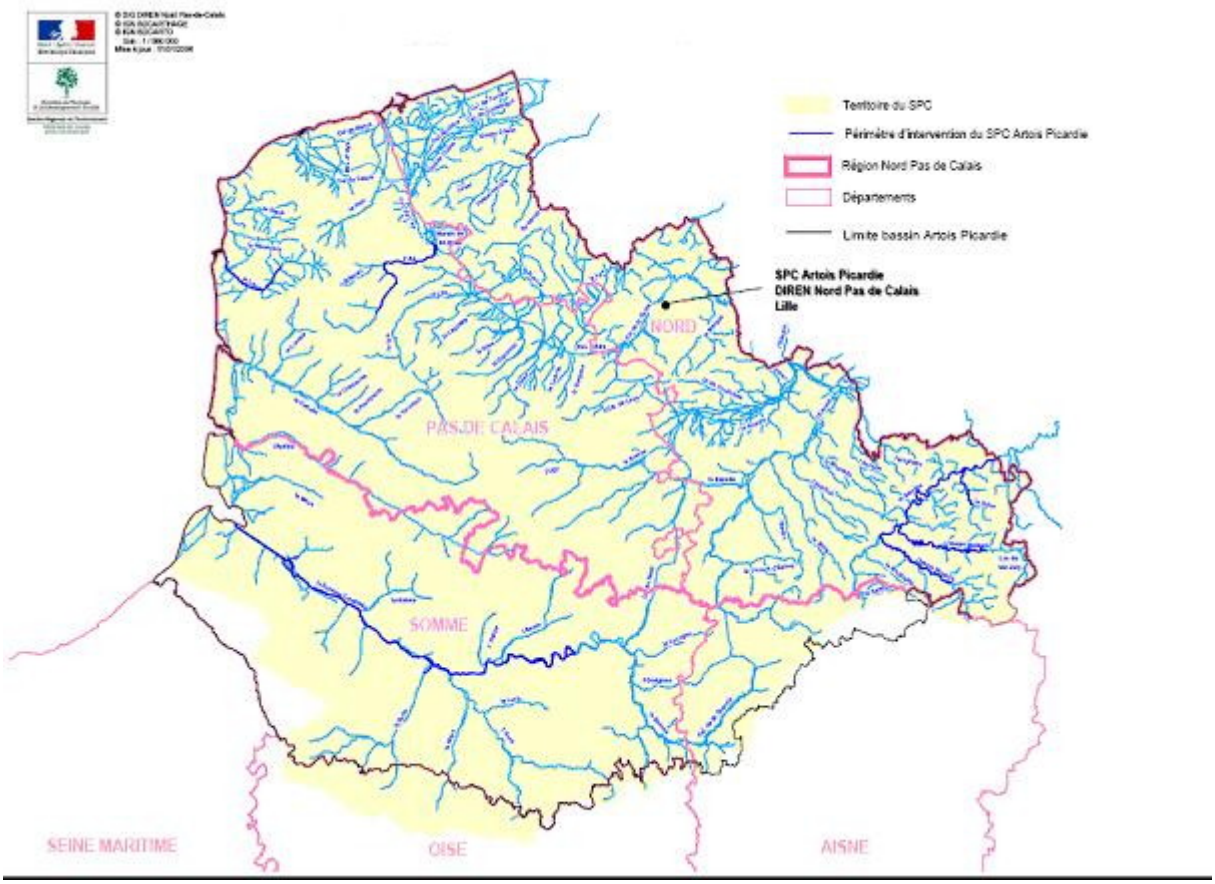


Fig. 1 – Cartographie du territoire et des périmètres d'intervention du SPC Artois-Picardie

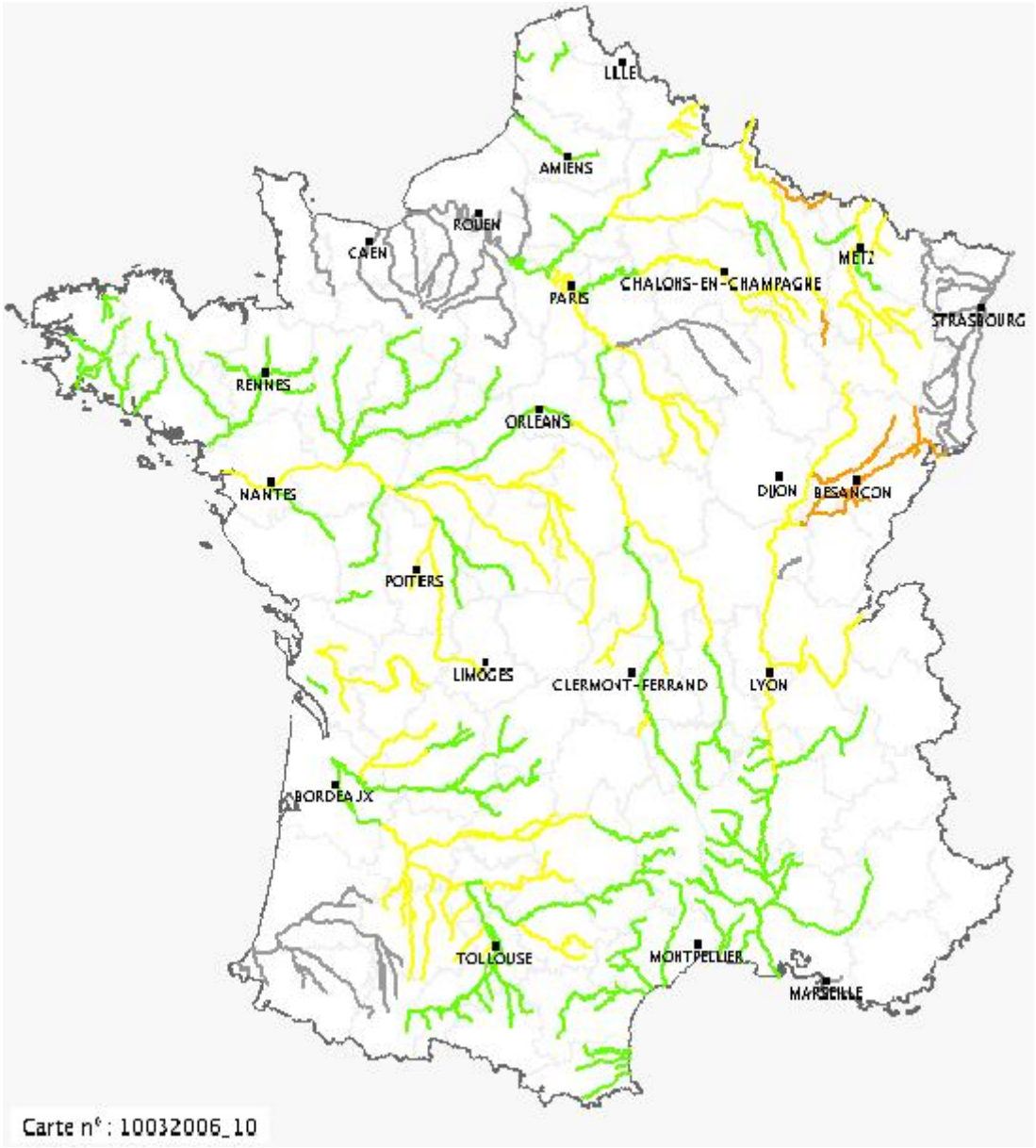


FIG. 2 – Carte nationale de vigilance crue

Mémento

Analyse Discriminante avec MIXMOD

Notons déjà que chaque cours d'eau possède ses propres scripts[15] qu'il est facile d'adapter suivant les consignes laissées dans les encadrés.

Chargement des données

– Exécuter sous Scilab[16], dans l'ordre, les scripts *Extension mémoire.sci* et *Chargement données.sce*[17]

Génération des fichiers nécessaires à MIXMOD

– Exécuter le script *AvantMixmod.sci*[18]

Utilisation de MIXMOD

– Changer de répertoire sous Scilab : `"C:\...\Cours d'eau\Mixmod\GUI"`[19]
– Exécuter le script *initMixmod.sci* situé dans `"C:\Program Files\Mixmod-2.1"`[20]
– Répondre *No* pour le changement de répertoire
– Taper sous Scilab la commande *mixmodGraph()* pour lancer l'assistant de MIXMOD
– Choisir *Discriminant Analysis*
– Entrer le nombre de variables (dimension du problème) puis le nombre de classes
– Charger le fichier *Training.dat* puis le fichier *Training.part* pour l'apprentissage
– Sélectionner *Start Discriminant Analysis* puis patienter ...
– Sélectionner *CONTINUE DISCRIMINANT ANALYSIS*
– Charger le fichier *Remaining.dat* pour le test puis patienter ...
– Sélectionner *Save output variable* pour sauvegarder les résultats
– Enregistrer le fichier dans `"C:\...\Cours d'eau\Mixmod"` avec l'extension *.sav*

Génération des fichiers d'évaluation et des données pour Excel

– Charger sous scilab le fichier *AD.sav*
– Exécuter le script *AprèsMixmod.sci*[21]

Mise en place sous Excel

– Dans la feuille *CalculsAD* du fichier *AD.xls*, remplacer les parties grisées[22]
– Sélectionner éventuellement des exemples[23]

[15] Rangés dans le dossier *Scripts* du cours d'eau en question
[16] Grâce au menu déroulant *Fichier*
[17] A récupérer dans le répertoire *AZAHAF* suivant le cours d'eau considéré
[18] Les fichiers *Training.dat*, *Training.part* et *Remaining.dat* sont ainsi générés
[19] C'est là que seront stockés tous les fichiers générés par MIXMOD lors de son utilisation
[20] Vous aurez ainsi accès à la fonction *mixmodGraph()*
[21] Le modèle peut alors être évalué grâce aux fichiers *ErreurAD.dat* et *ErreurCoûtAD.dat*
[22] Cf par exemple les commentaires du fichier *AD24h.xls* de la Hem
[23] On pourra pour cela utiliser les fichiers *PartitionT.dat* et *Remaining.dat*

Références

Mc Lachlan (1992). Discriminant Analysis and Statistical Pattern Recognition. Wiley

MIXMOD User's Guide (2007)
MIXMOD Statistical Documentation (2007)

http ://www-math.univ-fcomte.fr/mixmod/index.php
http ://www.nord-pas-de-calais.ecologie.gouv.fr