Analyse discriminante en prévision des crues

Stage réalisé à la DIREN
Nord Pas-de-Calais

La DIREN et son service de prévision des crues

La DIREN

- 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

Le Service de Prévision des Crues

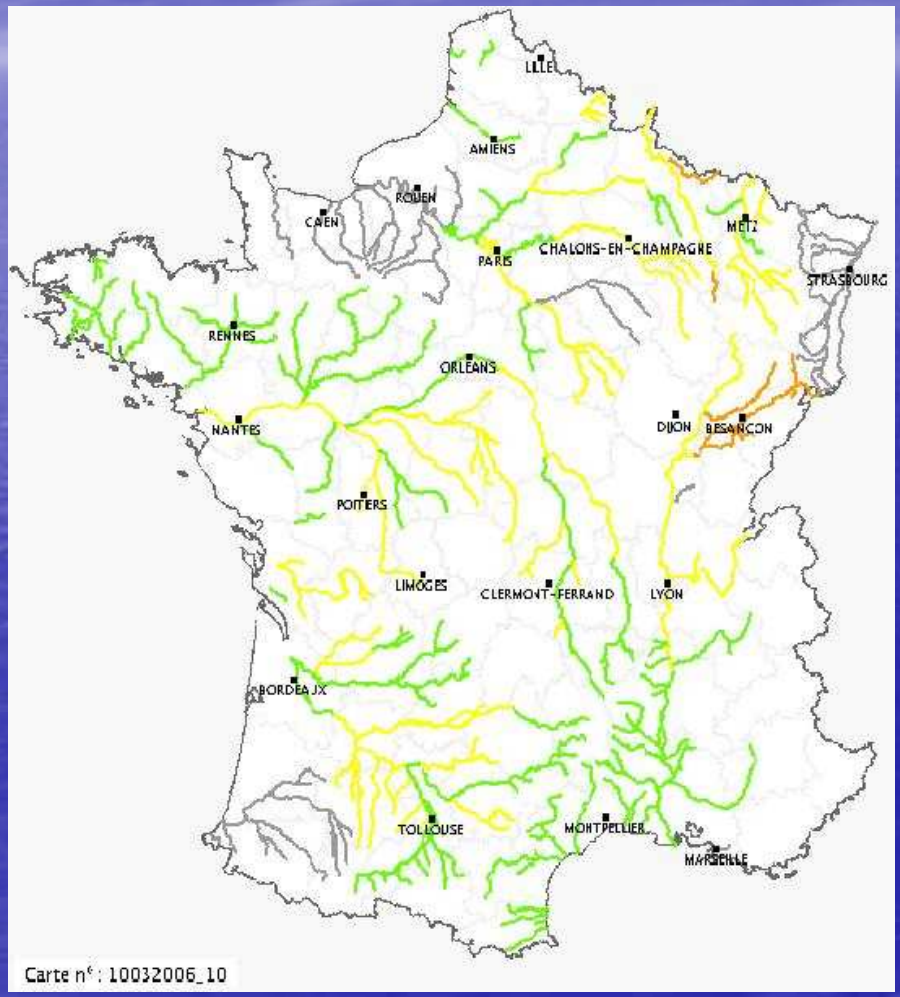
- Missions principales
  - Surveillance, prévision et transmission de l'information sur les crues via la procédure de vigilance
  - Capitalisation d'informations et expertise dans le domaine des inondations
  - Appui aux collectivités souhaitant mettre en place, pour leurs besoins propres et sous leur 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 pour l'application de la procédure de vigilance
  - 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
  - La prévision est envoyée a minima deux fois par jour au SCHAPI qui collecte l'ensemble des bulletins des différents SPC de France afin de publier la carte nationale de vigilance à 10h et à 16h sur le site internet http://www.vigicrues.ecologie.gouv.fr/

Couleurs de vigilance

Couleur	Définition	Qualification de la situation
Vert	Pas de vigilance particulière requise	Situation normale
Jaune	Risque de crue ou de montée rapide des eaux n'entraînant pas de dommages significatifs, mais nécessitant une vigilance particulière dans le cas d'activités saisonnières et/ou exposées.	Débordements localisés, coupures ponctuelles de routes, maisons isolées touchées, perturbation des activités liées au cours d'eau.
Orange	Risque de crue génératrice de débordements importants susceptibles d'avoir un impact significatif sur la vie collective et la sécurité des biens et des personnes.	Débordements généralisés, circulation fortement perturbée, évacuations
Rouge	Risque de crue majeure. Menace directe et généralisée de la sécurité des personnes et des biens.	Crue rare et catastrophique

Carte nationale de vigilance crue



Comment le SPC prévoit-il aujourd'hui ?

- Les prévisions de précipitations produites par Météo-France sont une base fondamentale de la prévision, car bien sûr, les précipitations sont l'élément déclencheur essentiel des crues

- Le SPC dispose de différents modèles numériques, calés à partir de crues anciennes, donnant 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

Une nouvelle approche par l'analyse discriminante

Un nouveau modèle, pourquoi ?

- Plus qu'un nouveau modèle, il s'agit là d'une nouvelle approche
- 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 au lieu d'une grandeur continue. 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

L'analyse discriminante en quelques mots

- L'idée est de voir les données comme des réalisations indépendantes d'une loi mélange

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

- dont on estimera les paramètres

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

- pour définir une règle de classement permettant d'associer un nouvel individu à sa classe

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

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

Mise en place à l'aide de MIXMOD

- MIXMOD est un logiciel permettant de modéliser un mélange de composantes gaussiennes multivariées à partir de données quantitatives
- Pour mener à bien notre analyse discriminante sous MIXMOD, nous disposons d'un échantillon d'apprentissage et d'un échantillon test
- MIXMOD procède en deux étapes :
  - Définition de la règle de classement à partir de l'échantillon d'apprentissage
  - Classement des nouveaux individus de l'échantillon test
- En réalité, l'échantillon test correspond à des données complètes autrement dit le classement de chaque individu est bien connu. 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

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

Analyse discriminante
VS
réseaux de neurones

Conditions expérimentales

- 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
- Les variables utilisées sont celles choisies pour les réseaux de neurones
- Trois modèles gaussiens ont été testés :
  - Modèle général
  - Modèle diagonal
  - Modèle sphérique
- Le modèle retenu pour la comparaison est le modèle général

Comparaison des modèles

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		

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 mais aussi sur le classement lui-même

$$R(r) = \sum_{k=1}^{3} p_k \sum_{l \neq k} C(l,k)\mathbf{P}(l|k)$$

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

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

La notion de coûts (2)

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		

Retour sur le choix de variables

Les variables

- Variable du présent
  - H : Hauteur du cours d'eau (à l'instant t)
- 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
  - W10995 : Humidité

Un choix minimaliste

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		

Apport de l'humidité

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		

Notre choix de variables

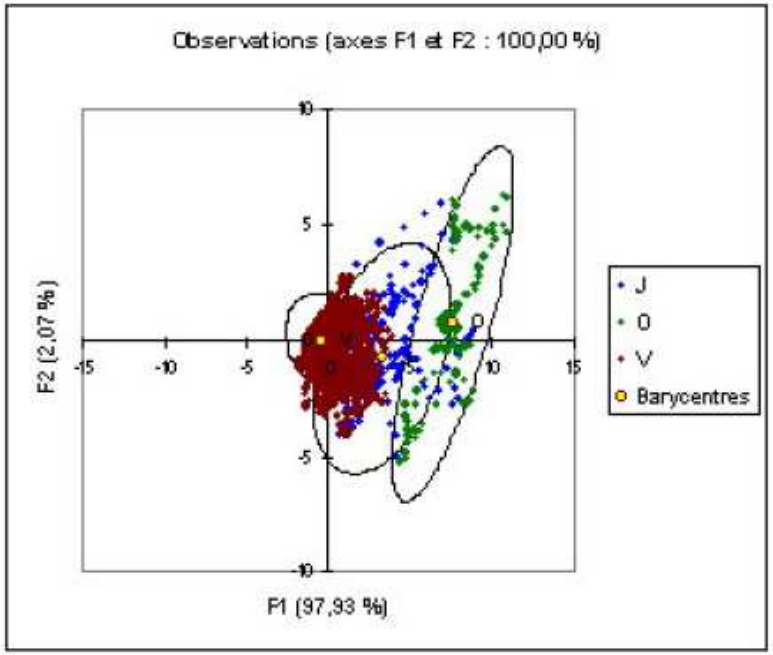
Prévision à 24h

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		

Analyse factorielle discriminante avec XLSTAT

Prévision à 24h



Mise en place du modèle sous Excel

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

Les calculs et le résultat

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

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

	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					

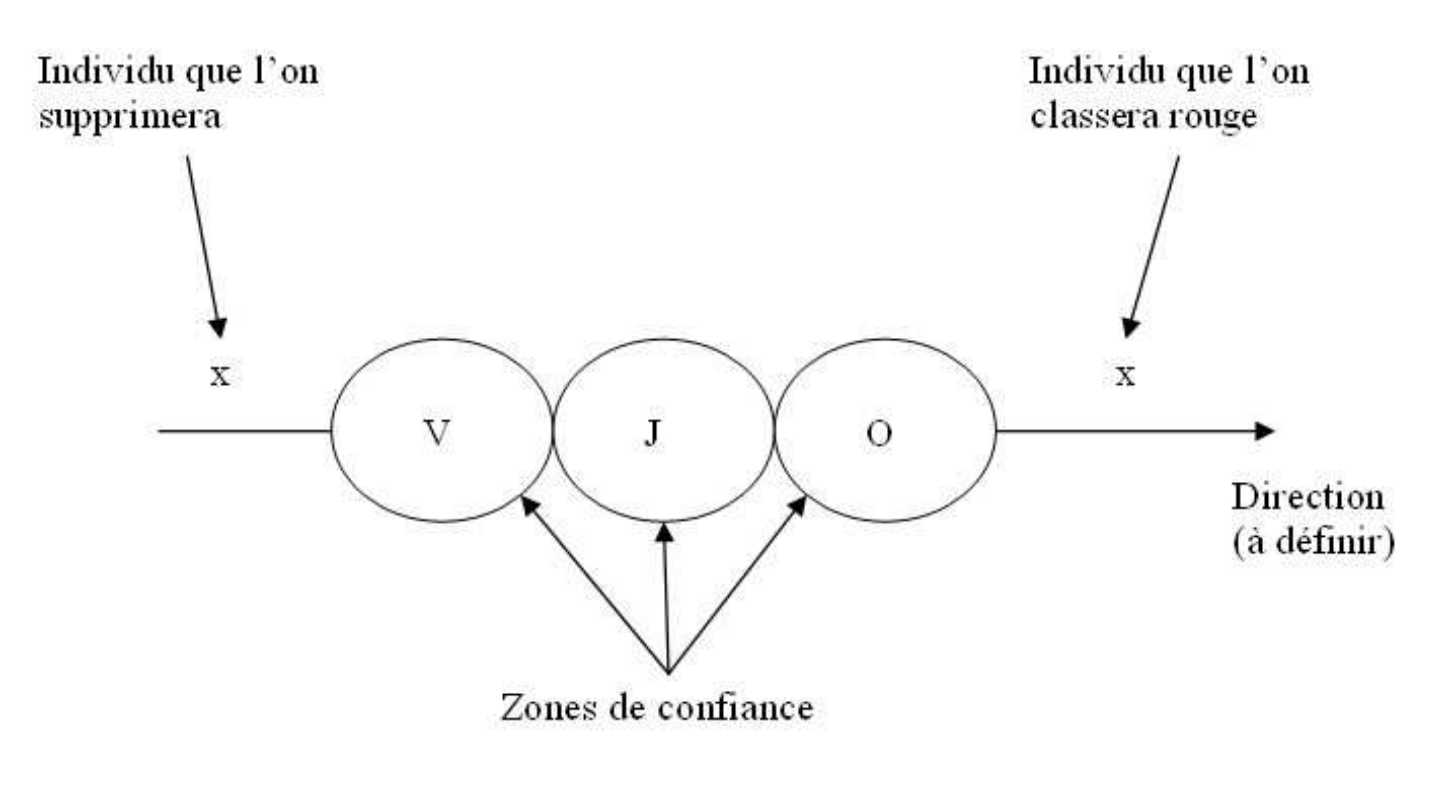
Extension à deux autres cours d'eau

La Hem

La Solre

Perspectives

Gestion de la classe rouge



CONCLUSION

- 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...