Etude de la segmentation d'une série de pollution en niveau d'ozone

J. Rynkiewicz, P. Letremy
SAMOS-MATISSE, Université de ParisI

3 octobre 2002

1 Introduction.

La pollution photochimique est devenue un problème crucial dans la plupart des cités modernes de grande taille. La croissance importante des principales sources d'émission (transport, centrales thermiques, chauffage urbain, etc...) alliée à des conditions météorologiques favorables conduisent à l'apparition dans ce milieu de concentrations atmosphériques de polluants supérieures aux normes de la qualité de l'air. C'est notamment le cas en France, des villes comme Paris, Lyon ou Strasbourg enregistrent pendant plusieurs jours dans l'année des concentrations dépassant les seuils d'alerte fixés par les pouvoirs publics. Il devient important de mieux connaitre les mécanismes de ces phénomènes pour, par exemple, pouvoir prévoir ou éviter les pics de pollution. Nous nous intéressons ici une série temporelle de pollution en niveau d'ozone à Paris (13ème arrondissement). On sait déjà que les modèles statistiques donnent en moyenne de bonnes prévisions de ces séries (cf Bel et al [1]),bien qu'ils échouent pour les parties les plus importantes et les plus difficiles à prévoir de la série : les pics de pollutions.

En fait, la dynamique de cette série semble non seulement fortement non-linéaire lors de l'apparition de pics, mais présente certainement un changement notable de comportement à l'approche de ces grandes valeurs de pollution en niveau d'ozone. Le comportement complexe de cette série nous a donné l'idée d'utiliser des modèles adaptés aux séries stationnaires par morceaux pour introduire des possibilités de changement de régimes : les modèles autorégressifs à changements de régime markoviens. Outre une amélioration significative du pouvoir prédictif, nous obtenons à l'aide de ces régimes une segmentation claire de notre série de pollution en niveau d'ozone.

Le but de cette étude est d'explorer la signification de ces deux régimes grâce à des outils de statistique exploratoire. Dans une première partie nous introduisons les modèles autorégressifs à changements de régime markoviens. Ensuite nous présentons la segmentation obtenue grâce à ce modèle. Nous appliquons ensuite les techniques classiques de l'analyse factorielle et discriminante pour interpréter les deux classes obtenues. Nous affinons aussi nos interprétations par une classification obtenue par l'algorithme de Kohonen.

2 Modèles autorégressifs à changements de régimes markoviens

2.1 Introduction

La modélisation des séries temporelles repose généralement sur l'hypothèse contraignante de la stationnarité du modèle. Une généralisation simple possible est de tenir compte des séries stationnaires par morceaux. Hamilton [7] par exemple a étudié de tels modèles afin de modéliser des séries temporelles sujettes à des changements discrets de régime pour analyser la série GNP (gross national product) aux Etats-Unis. On peut ainsi utiliser ces modèles par exemple pour des séries ayant un certain régime pour les périodes de croissance économique, un autre pour les périodes de récession.

Ce modèle est plus général, mais il conserve encore de nombreuses hypothèses contraignantes. D'abord le nombre de régimes possibles doit être fini. Ensuite, bien qu'il puisse y avoir des changements de régime, on suppose que ces changements interviennent de façon stationnaire, afin d'avoir une loi des grands nombres et de pouvoir ainsi faire des statistiques.

2.2 Le modèle

Les différents régimes sont modélisés par des états dans un espace de dimension fini et conditionnellement à ces états, le modèle suit une dynamique autorégressive classique.

2.2.1 Chaînes de Markov dans un espace discret

On considère $(X_t)_{t\in\mathbb{Z}}$, une chaîne de Markov homogène à valeurs dans un espace d'état fini $\mathbb{E} = \{e_1, \cdots, e_N\}$, $N \in \mathbb{N}^*$. Sans perte de généralité, on identifie l'espace d'état $\mathbb{E}$ avec le simplexe de $\mathbb{R}^N$, où e_i est un vecteur unité de $\mathbb{R}^N$ avec 1 sur la i-ème composante et 0 partout ailleurs. La chaîne X_t est caractérisée par sa matrice de transition $A = (a_{ij})_{1\leq i,j\leq N}$ qui est telle que :

$$P\left(X_{t+1} = e_i \,|X_t = e_j\right) = a_{ij}$$

Si, de plus, on définit : $V_{t+1} := X_{t+1} - AX_t$, on obtient l'écriture suivante :

$$X_{t+1} = AX_t + V_{t+1}.$$

2.2.2 Equations du modèle

On suppose que la série temporelle observée (Y_t) vérifie les équations suivantes :

$$\left\{\begin{array}{c} X_{t+1} = AX_t + V_{t+1} \\ Y_{t+1} = F_{X_{t+1}}(Y_t, Z_t) + \varepsilon_{t+1}\left(X_{t+1}\right) \end{array}\right.$$

où $(Z_t)_{t\in\mathbb{Z}}$, $Z_t \in \mathbb{R}^d$ est une suite de variables aléatoires exogènes, $\{F_{e_1}, \cdots, F_{e_N}\}$ sont des fonctions de $\mathbb{R}^{d+1} \longrightarrow \mathbb{R}$ qui seront représentées dans notre cas par des perceptrons multicouches (MLP) ou des modèles de régression linéaire qui sont des cas particuliers de MLP. De plus, pour tout $e_i \in \mathbb{E}$, $(\varepsilon_t\left(e_i\right))$ est une suite de variables aléatoires indépendantes et identiquement distribuées.

Ce modèle permet d'utiliser plusieurs MLP (on parle de mélange d'experts) et d'utiliser la chaîne de Markov (X_t) pour spécifier à un temps "t" quel est le MLP qui fait la prévision la plus pertinente. On remarquera que l'on observe seulement la série (Y_t), il faudra donc trouver un moyen de retrouver la suite des états de la chaîne (X_t), grâce au comportement de (Y_t).

Pour ajuster un tel modèle aux observations, on estime les paramètres (les poids des MLP F_{e_i} , la variance des bruits $\varepsilon\,(e_i)$, et la matrice de transition A) grâce à la méthode du maximum de vraisemblance. On pourra trouver l'étude des propriétés théoriques de cet estimateur dans Ryden et Krishnamurthy [9] et Douc, Moulines, Ryden [5].

3 Etude des pics de pollution en ozone

Les données de base de l'étude sont constituées, pour la chimie, des mesures enregistrées par le réseau de surveillance AIRPARIF et pour les données météorologiques observées, de celles des stations parisiennes de Météo-France. Il est important de posséder une base de données suffisamment grande qui contienne le plus de situations possible. La périodicité des variations des concentrations d'ozone, ainsi que la faible fréquence d'occurrence des pics de pollution sur l'ensemble de l'année, nous impose de travailler sur une base pluriannuelle la plus large possible. Les données retenues seront ici les maxima des moyennes horaires d'une journée pour les concentrations des polluants enregistrées de 1994 à 1997. Afin de prédire le maximum journalier du taux de pollution en niveau d'ozone (OZ) durant la période d'avril à septembre inclus, nous utilisons comme régresseurs le maximum du taux de pollution en niveau d'ozone de la veille (OZ24) et les observations météorologiques suivantes :

- La rayonnement globale (RAY)
- La vitesse moyenne du vent du jour (VENT)
- La température maximale de la journée (TEMP)
- Le gradient de température sur un jour (GRAD)

La modélisation statistique de l'ozone et plus particulièrement par des régressions a été beaucoup étudiée. Les modèles linéaires ne semblent pas capturer toute la complexité du phénomène. C'est pourquoi il faut employer des modèles plus riches (cf Chen, Islam and Biswas [2] ou Gardner et Dorling [6]). Parmi ces modèles, les MLP semblent donner de meilleurs résultats que les modèles linéaires bien qu'ils demandent souvent beaucoup plus d'efforts de mise en oeuvre pour obtenir seulement une modeste amélioration des prédictions. (cf Comrie [3]).

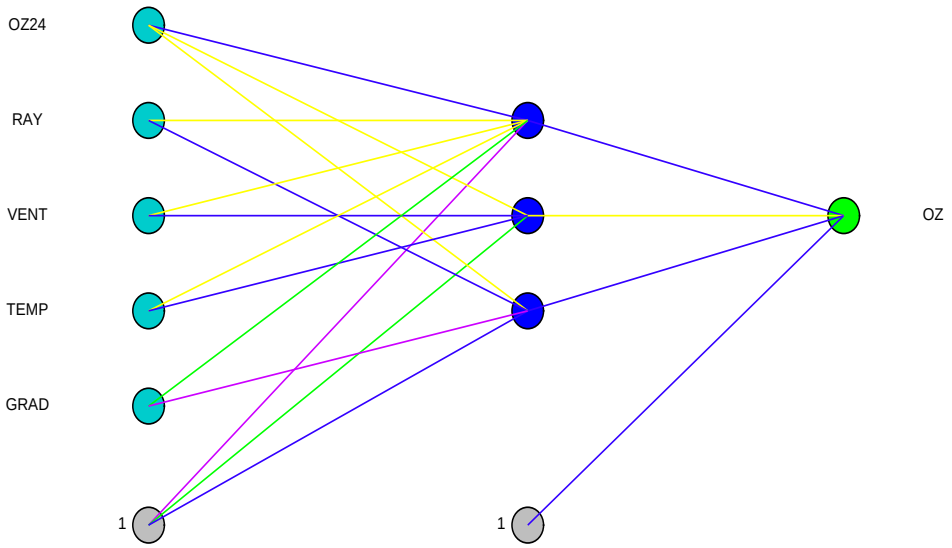
Pour cette étude, nous disposons des observations météorologiques et du taux de pollution d'ozone pour les années de 1994 jusqu'à 1997 inclus. Nous utilisons les données de 1994 jusqu'à 1996 pour estimer nos modèles (données "in sample"), et nous testons ce modèle sur les données 1997 (données "out of sample), pour contrôler les phénomènes de sur-apprentissage.

3.1 Performance du modèle hybride sur la série d'ozone

Nous avons choisi d'utiliser un modèle avec deux états possibles. En effet, puisque les modèles linéaires sont capables de modéliser correctement les faibles valeurs de pollutions en niveau d'ozone, nous choisissons d'utiliser pour notre modèle hybride un modèle de régression linéaire et un modèle MLP ayant une architecture déterminée par une précédente étude sur ces mêmes données. Nous espérons ainsi que le modèle linéaire se spécialise dans la prédiction des faibles

valeurs en niveau d'ozone, alors que le MLP se spécialisera dans la prédiction des fortes valeurs. Ce choix est guidé par le souci de garder un nombre de paramètres le plus raisonnable possible. Son architecture est représentée par la figure 1.

FIG. 1 – Architecture du MLP



Après estimation des paramètres, nous obtenons la matrice de transition suivante pour la chaîne de Markov cachée :

$$\hat{A} = \begin{pmatrix} 0.97 & 0.02 \\ 0.03 & 0.98 \end{pmatrix}$$

On remarque que les termes diagonaux, qui représente la probabilité de rester dans le même état, sont très proches de la probabilité maximale 1. Cela signifie que le modèle reste pendant de longues plages dans le même état, c'est le signe qu'il a bien identifié deux régimes distincts. Les déviations standards pour l'expert linéaire σ_1 et le MLP σ_2 sont les suivantes :

$$\begin{cases} \sigma_1 = 0.11 \\ \sigma_2 = 0.20 \end{cases}$$

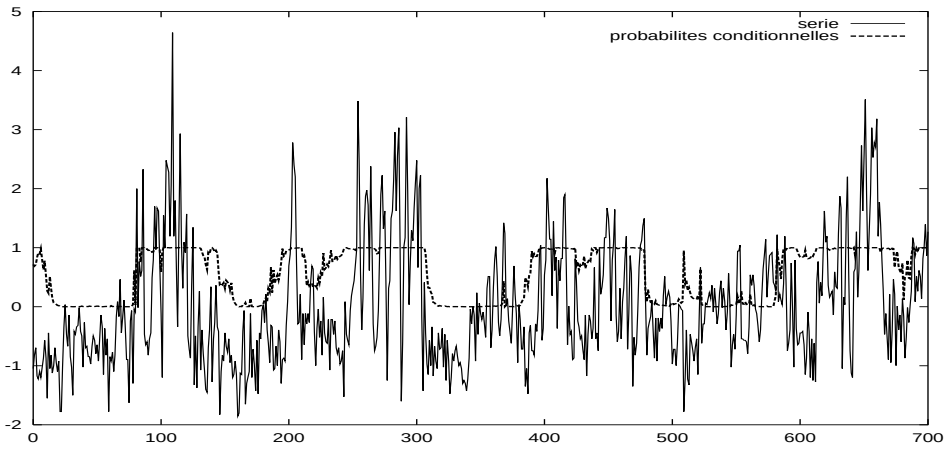
Ce résultat est cohérent avec l'intuition, puisque nous verrons que le modèle linéaire s'est spécialisé dans la partie facile de la série, les valeurs moyennes ou basses, ce qui lui permet de faire de bonnes prédictions, alors que le MLP est spécialisé dans la partie difficile, les fortes valeurs.

Finalement, les résultats en terme d'erreurs de prédiction sont les suivants :

Années	1994-1996	1997
RMSE	$16.51 \mu g/m^3$	$16.75\ \mu g/m^3$

Ces prévisions sont relativement bonnes puisque la racine carrée de l'erreur quadratique de prévision est significativement en dessous des 20 microgrammes d'ozone par mètre cube. De plus le modèle hybride donne des informations plus riches que le modèle de régression simple. On obtient en effet une segmentation de la série suivant la probabilité conditionnelle des deux régimes (cf figure 2). On remarque que les fortes probabilités du régime associé au MLP sont plutôt associées aux fortes valeurs, en outre il n'y a jamais de pic de pollution lorsque que cette probabilité est faible.

FIG. 2 – Série centrée normée et probabilité de l'état associé au MLP



Les figures 3 et 4 montre le 'scatterplot' des prédictions des deux experts par rapport aux vraies valeurs sur l'ensemble des données (in et out of sample)

On remarque sur ces figures que l'expert linéaire est meilleur que le MLP sur les valeurs basses, cependant, pour les fortes valeurs, ses prédictions sont nettement plus petites que les vraies valeurs. Le comportement de l'expert MLP est l'opposé, il surestime les faibles valeurs mais il estime beaucoup mieux les fortes. Si on utilise uniquement ce MLP pour faire les prédictions, l'erreur quadratique moyenne sera moins bonne celle du modèle autorégressif simple, mais elle sera meilleure sur la partie intéressante de la série : les pics de pollution en niveau d'ozone.

FIG. 3 – Prédictions de l'expert linéaire, en fonction des vraies valeurs

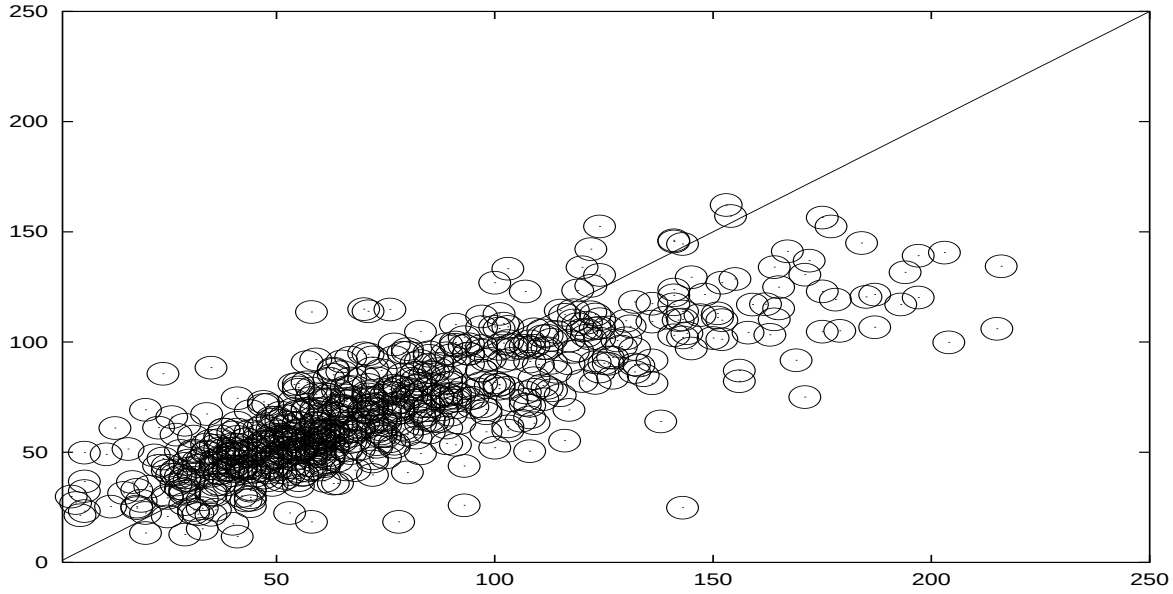
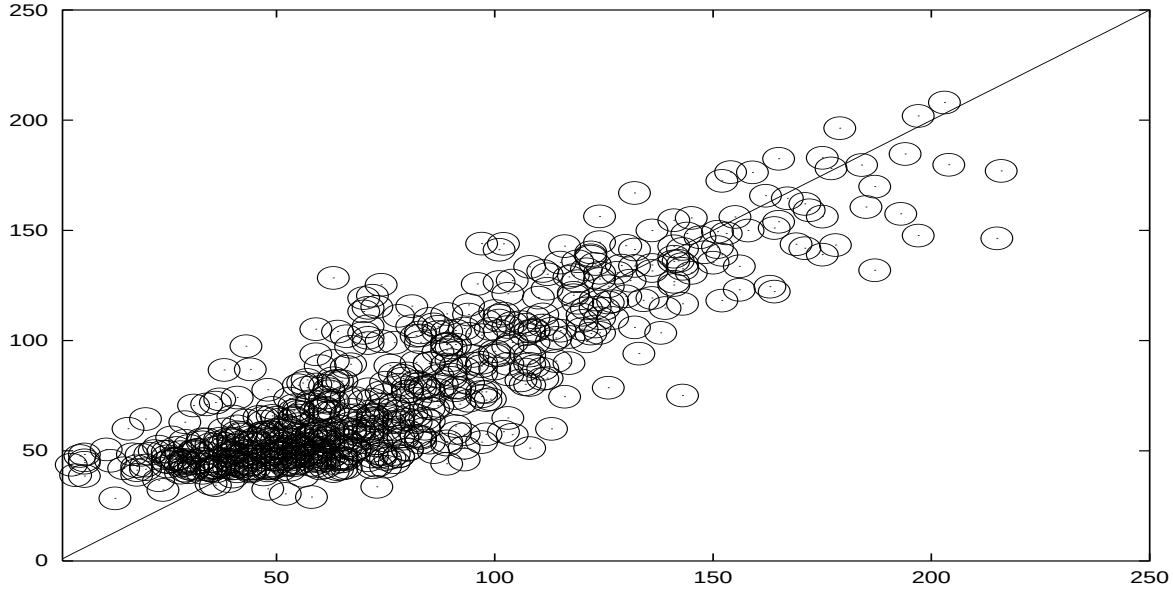


FIG. 4 – Prédictions de l'expert MLP, en fonction des vraies valeurs



4 Analyse de la segmentation

Nous essayons ici d'affiner les premières constatations de la section précédente. Il ne faut pas perdre de vue que la segmentation obtenue ne différencie pas seulement les hauts niveaux des bas niveaux de pollution, mais caractérise plutôt un comportement linéaire vis-à-vis d'un comportement non-linéaire de notre modèle. Il s'agit ici d'obtenir une description de ces deux régimes grâce aux différentes variables explicatives et du taux d'ozone. Dans toute la suite les séries considérées seront toujours centrées et normées comme celles utilisées pour estimer les paramètres de notre modèle.

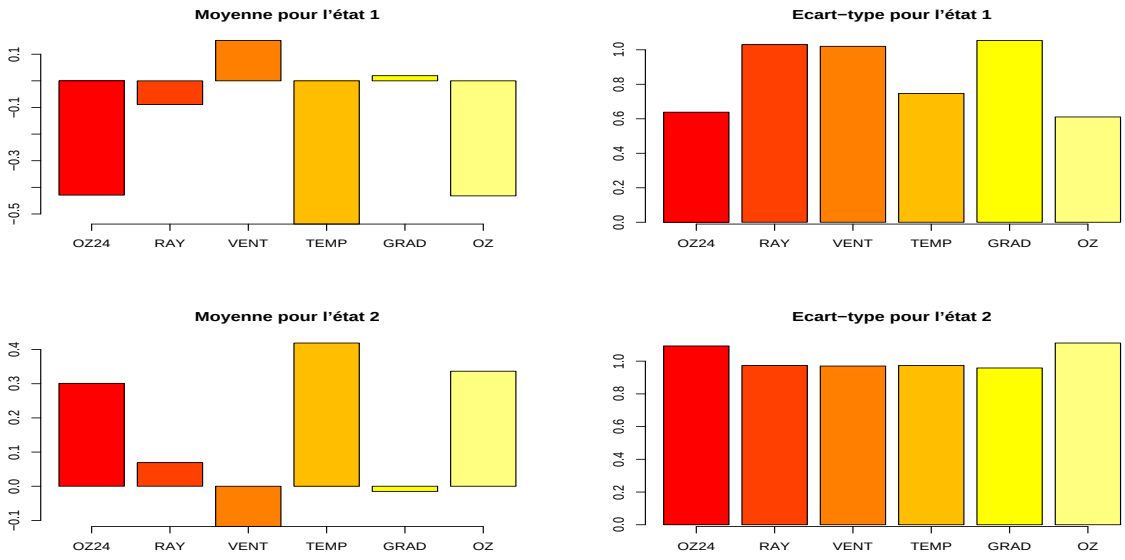
4.1 Analyse exploratoire

Les deux nuages de points associés aux deux états sont de dimension 6 (5 Variables explicatives et 1 variable expliquée). Nous commençons par décrire leur profil moyen puis par réduire la dimension pour les visualiser.

4.1.1 Statistiques descriptives

Nous donnons tout d'abord (cf figure 5) la moyenne est la variance de l'ozone et de ses prédicteurs en fonctions de l'état le plus probable de la chaîne de Markov cachée. La symétrie

FIG. 5 – Moyenne et variance des variables suivant les états



par rapport à l'axe des abscisses des deux profils est normale puisque toutes les séries sont centrées. L'état "1" (associé au modèle linéaire) correspond en moyenne à un niveau faible en ozone et en température, mais à un plus fort niveau de vent. Pour l'état "2" (associé au MLP) c'est l'inverse. Par contre le rayonnement et le gradient de temperature sont proche de zéro ce qui indique qu'ils sont quasi équi-répartis entre les deux états. On remarque aussi que l'état du

modèle linéaire est associé à des variances plus faibles pour l'ozone et la température. Le modèle linéaire correspond donc à une partie moins chaotique de la série.

4.1.2 Analyse en composantes principales (ACP)

Nous allons maintenant visualiser le nuage suivant différents axes factoriels de l'analyse en composantes principales. Rappelons que nous travaillons avec des données centrées et normées.

Les axes factoriels Les vecteurs propres ont pour cordonnées :

	axe1	axe2	axe3	axe4	axe5	axe6
OZ24	0.426	0.542	-0.077	-0.027	-0.312	0.647
RAY	0.375	-0.299	-0.432	0.756	0.095	0.030
VENT	-0.310	0.147	-0.890	-0.295	0.005	-0.035
TEMP	0.503	0.0662	-0.044	-0.347	0.786	-0.0265
GRAD	0.208	-0.761	-0.072	-0.429	-0.251	0.352
OZ	0.532	0.099	-0.082	-0.187	-0.459	-0.673

Les pourcentages cumulés de variance expliquée sont les suivants :

	axe1	axe2	axe3	axe4	axe5	axe6
% cumulés	0.485	0.693	0.828	0.922	0.973	1.000

On remarque que les trois premiers facteurs expliquent déjà près de 83% de la variance. Nous allons donc projeter le nuage sur les axes 1, 2 et 3.

Les projections La figure 6 montre la projection des deux nuages de points associés aux états les plus probables sur les axes 1 et 2, ainsi que la projection des différentes variables sur le plan engendré par ces axes. On remarque que les fortes valeurs d'ozone et de température sont associées à l'état "2". Le vent apparaît opposé au fort niveau d'ozone et de température. On peut faire pratiquement les mêmes remarques pour la projection sur l'axe 1 et 3 (figure 7).

Par contre, la projection sur les axes 2 et 3 (figure 8) ne permet pas de distinguer les deux états. On observe que les variables ozone et température ne sont pas prises en compte par ces deux axes, elles déterminent donc principalement les deux régimes. On a vu que pour les fortes températures et les forts niveaux d'ozone le phénomène n'est plus linéaire et le MLP devient alors le modèle privilégié comme modéle de régression. L'ACP confirme donc les statistiques descriptives. L'ACP nous a montré aussi qu'il est extrêmement difficile de distinguer les deux régimes sans au moins une de ces deux variables. Nous allons maintenant caractériser plus précisément le régime non-linéaire grâce à une analyse par carte de Kohonen.

FIG. 6 – Projection des points et des variables sur les axes 1 et 2

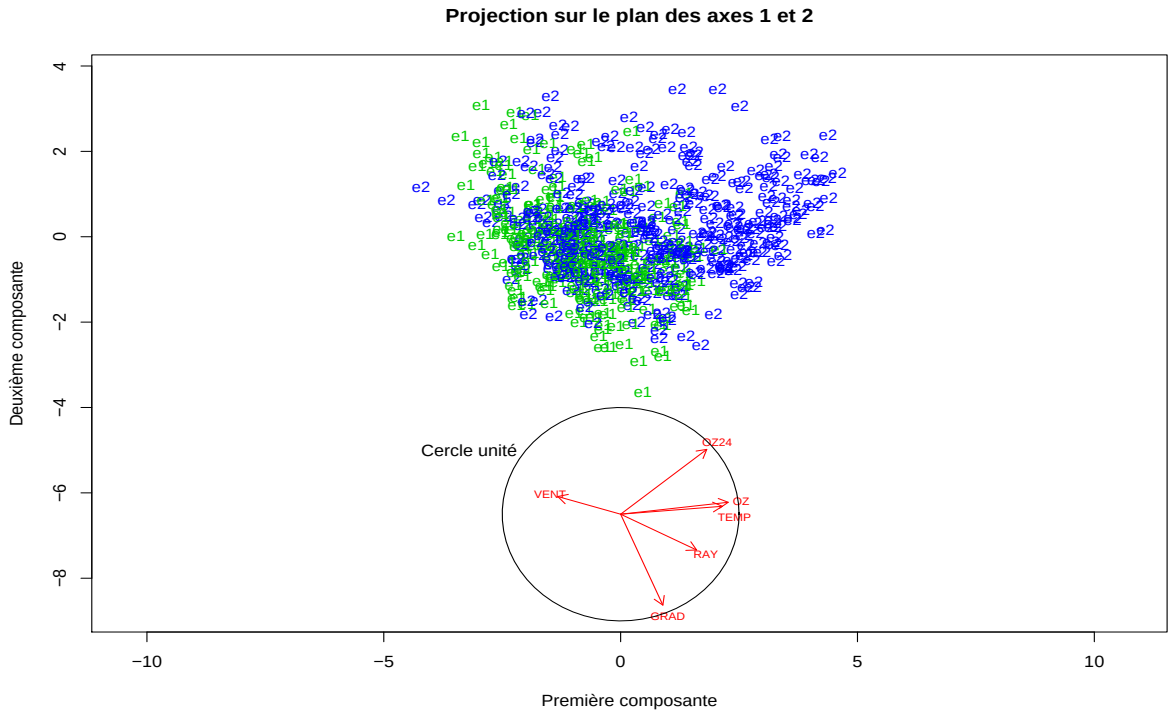


FIG. 7 – Projection des points et des variables sur les axes 1 et 3

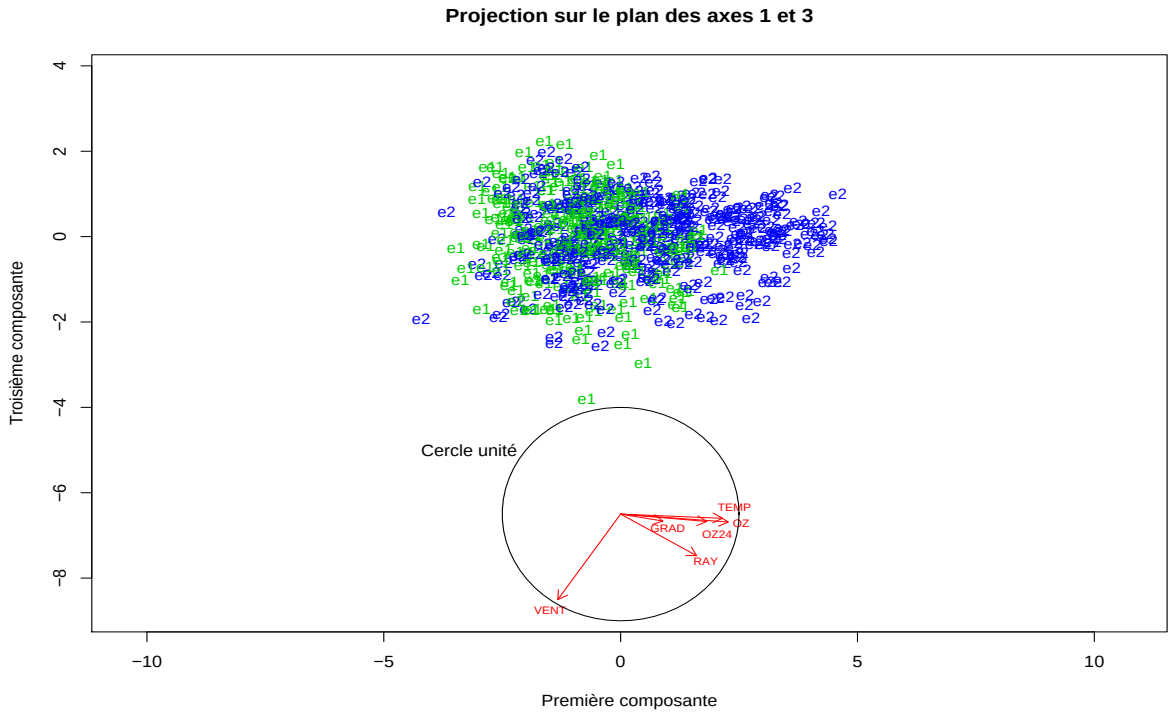
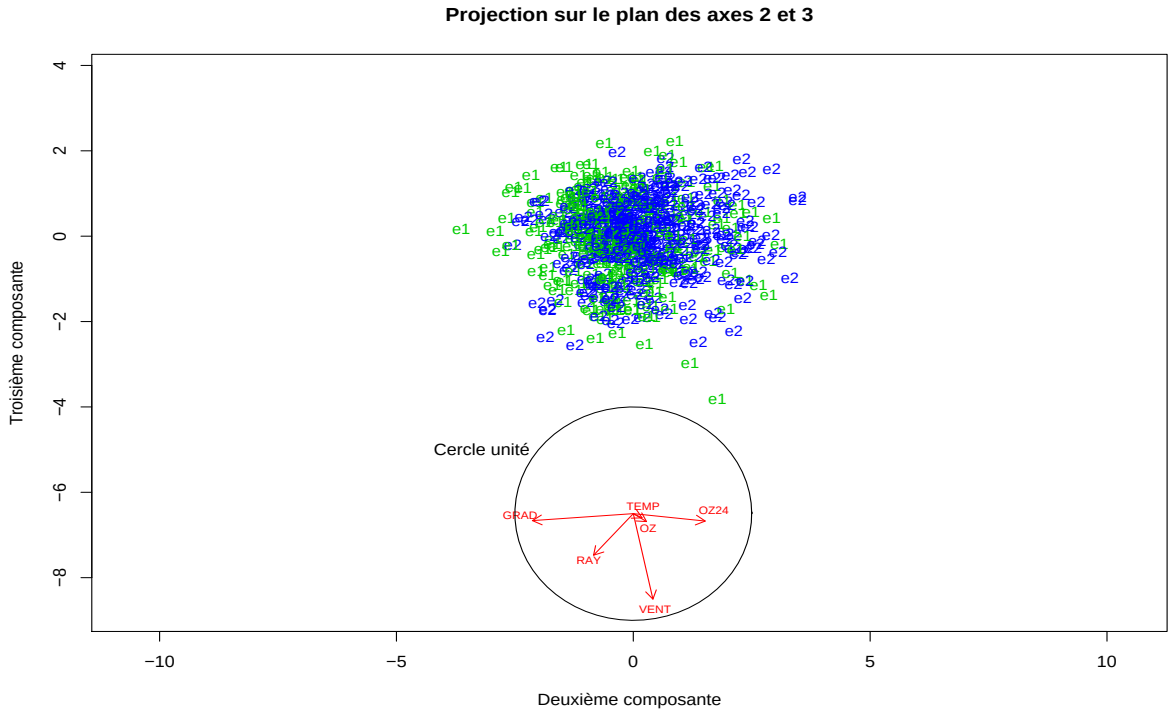


FIG. 8 – Projection des points et des variables sur les axes 2 et 3



4.1.3 analyse par une carte de Kohonen

Principes L'algorithme de Kohonen (Kohonen [8]) est un algorithme d'apprentissage non-supervisé qui est apparu dans le domaine des réseaux de neurones (cf Cottrel et Fort [4]). Il s'agit d'une extension des algorithmes de classification non supervisée, du type des centres mobiles. Le réseau de neurones est formé d'un ensemble d'unités. Chaque unité est caractérisée par un vecteur-code (ou centroïde), appartenant au même espace que les données. Chaque unité possède donc une localisation dans cet espace. Chaque unité est également caractérisée par ses voisins dans le réseau de Kohonen, le nombre de voisins décroît au cours de l'apprentissage pour terminer à 0. Chaque unité représente en fait un groupe d'individus (les individus pour lesquels, dans l'espace d'entrée, le vecteur-code de l'unité en question est le plus proche, au sens de la distance euclidienne).

Au terme de l'apprentissage, deux caractéristiques sont observées. Les centroïdes viennent se positionner au centre de gravité (ou barycentre) des classes d'individus qu'ils définissent. Il s'agit d'un résultat classique des algorithmes de type competitive learning ou k-means. En outre, et c'est là que se situe l'une des spécificités principales des algorithmes d'auto-organisation, la correspondance entre les caractéristiques des individus et les unités respecte (plus ou moins bien) la structure de l'espace d'entrée : des individus qui présentent des caractéristiques communes correspondent à des unités identiques ou voisines sur la carte de Kohonen. L'état final de la carte préserve en ce sens la topologie de l'espace d'entrée.

Résultats Nous avons utilisé une grille de Kohonen avec 49 centroïdes. Une fois la classification obtenue, nous avons regroupé par classification hiérarchique les centroïdes en 5 classes. Les profils pour ces 5 classes (figure 9) représentent, dans l'ordre, les variables : OZ24, RAY, VENT, TEMP, GRAD, OZ.

La répartition des deux états suivant ces classes est représentée par les diagrammes de la figure 10. La classe violette comporte très majoritairement des profils correspondant à l'état 2 (modèle non linéaire). Le profil moyen de cette classe se caractérise par un niveau du taux d'ozone de la veille fort, un rayonnement moyen, un vent faible, une température élevée, un gradient de température moyen et un taux d'ozone actuel fort. On peut finalement remarquer qu'il reste toujours un mélange des deux états pour les autres profils. La classe qui représente le mieux l'état linéaire est la classe verte juste sous la classe violette. La prédominance de l'état "1" pour cette classe est réelle même si elle n'est pas très marquée. Le profil moyen de cette classe correspond à des valeurs quasiment moyennes pour toutes les variables. Cela explique notamment pourquoi le modèle linéaire est un bon prédicteur pour les comportements sans fortes valeurs. Finalement on remarquera que les faibles niveaux d'ozone et de température ne sont pas aussi déterminants pour le régime linéaire que peuvent l'être les fort niveaux pour le régime non-linéaire.

Les figures 11, 12, 13, 14, 15 et 16 montrent la répartition des variables suivant les différentes classes.

FIG. 9 – Profils obtenus par classification hiérarchique

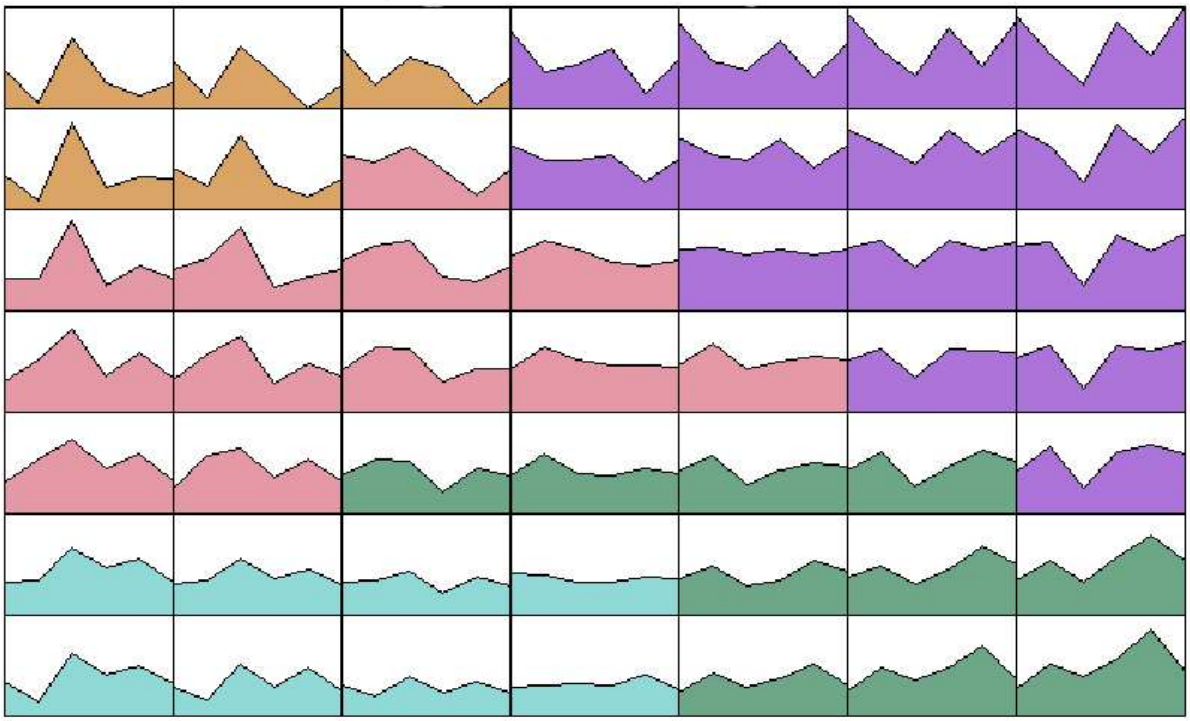


FIG. 10 – Etat “2” le plus probable : en jaune

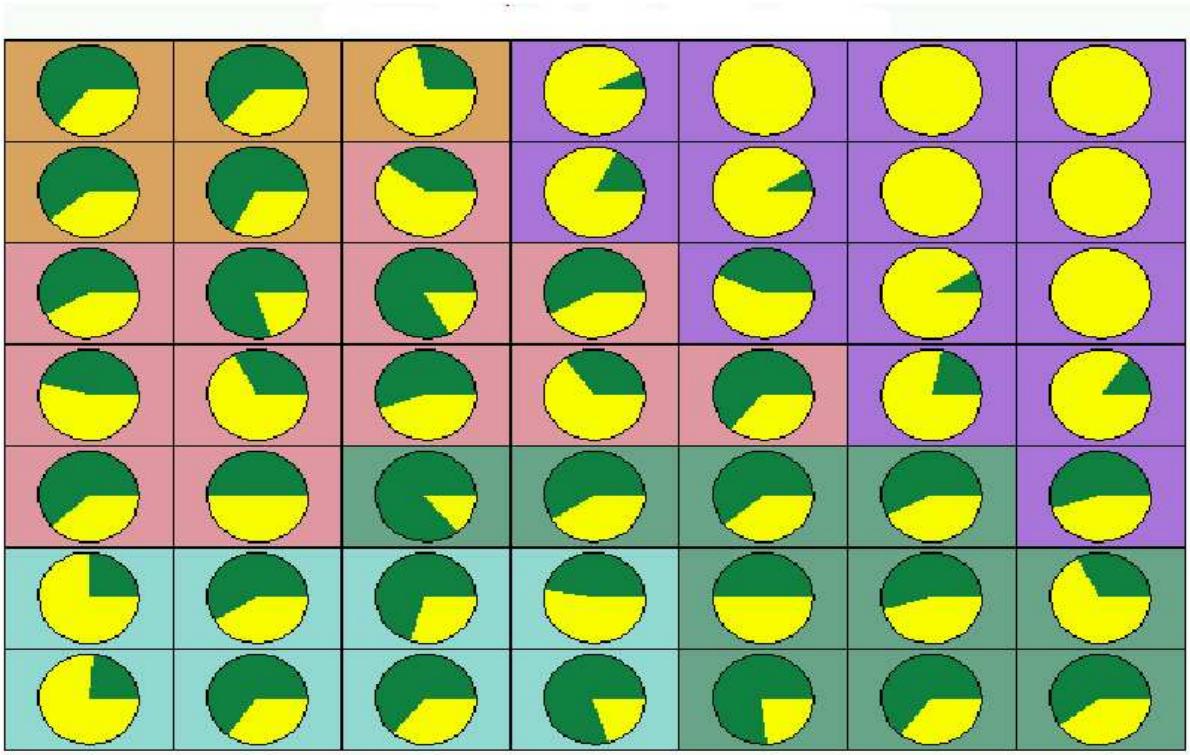


FIG. 11 – Répartition de la variable OZ24 suivant les différentes classes

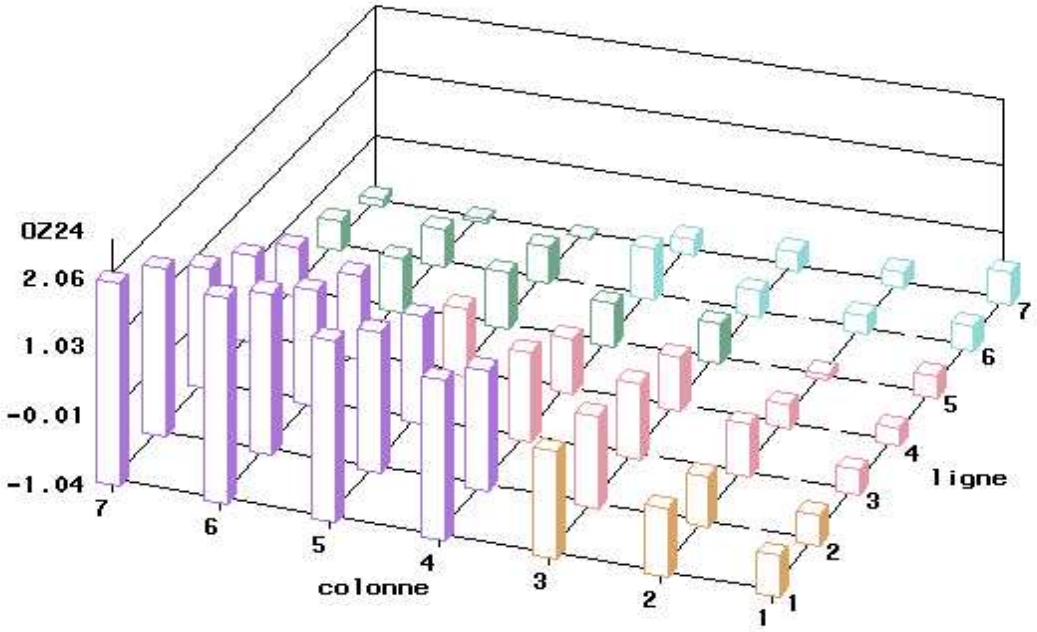


FIG. 12 – Répartition de la variable RAY suivant les différentes classes

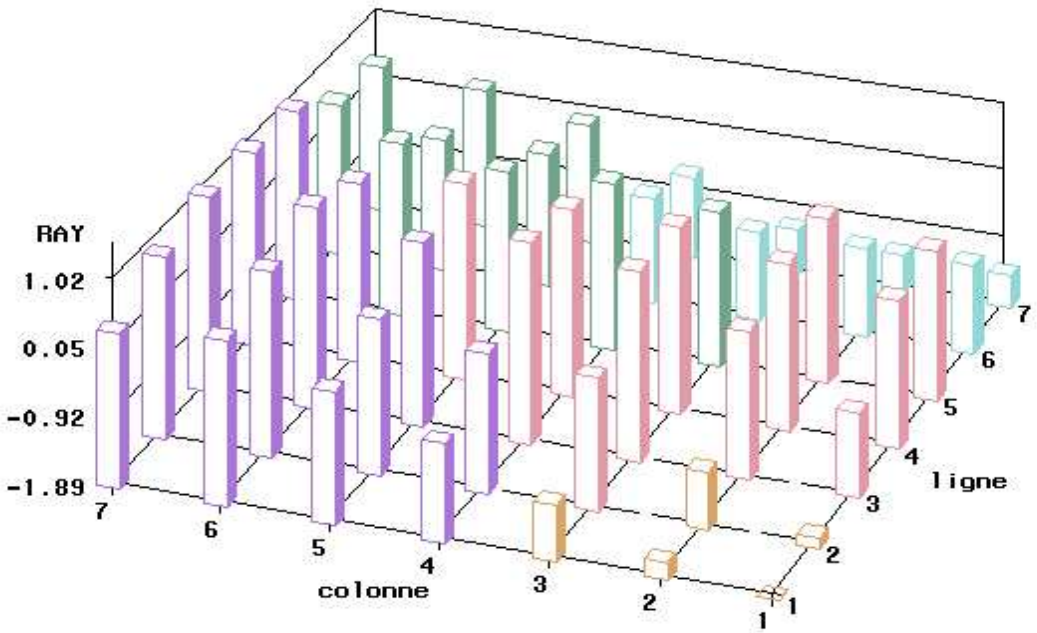


FIG. 13 – Répartition de la variable VENT suivant les différentes classes

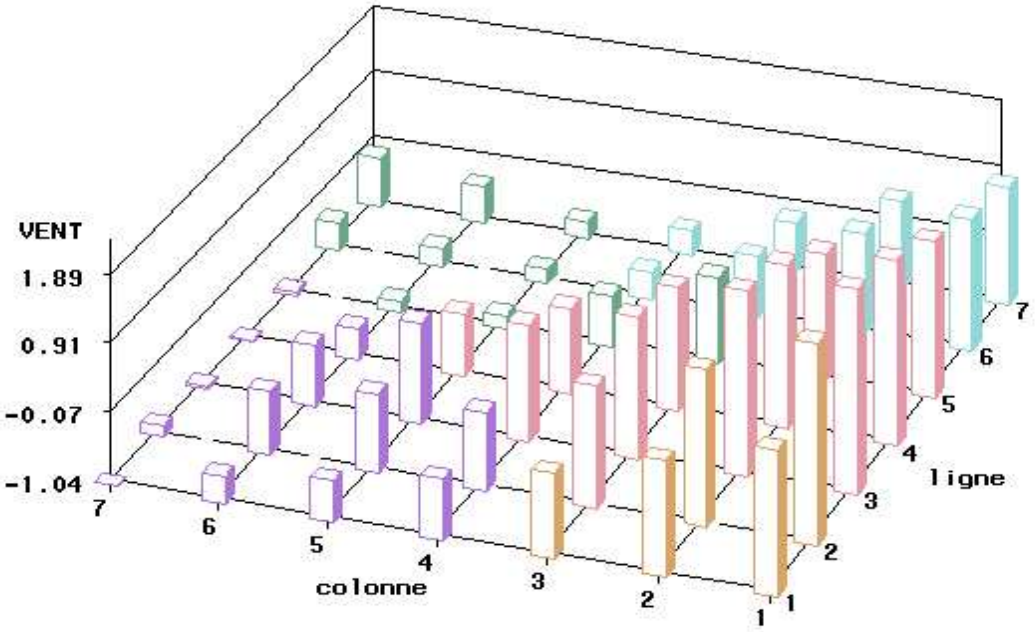


FIG. 14 – Répartition de la variable TEMP suivant les différentes classes

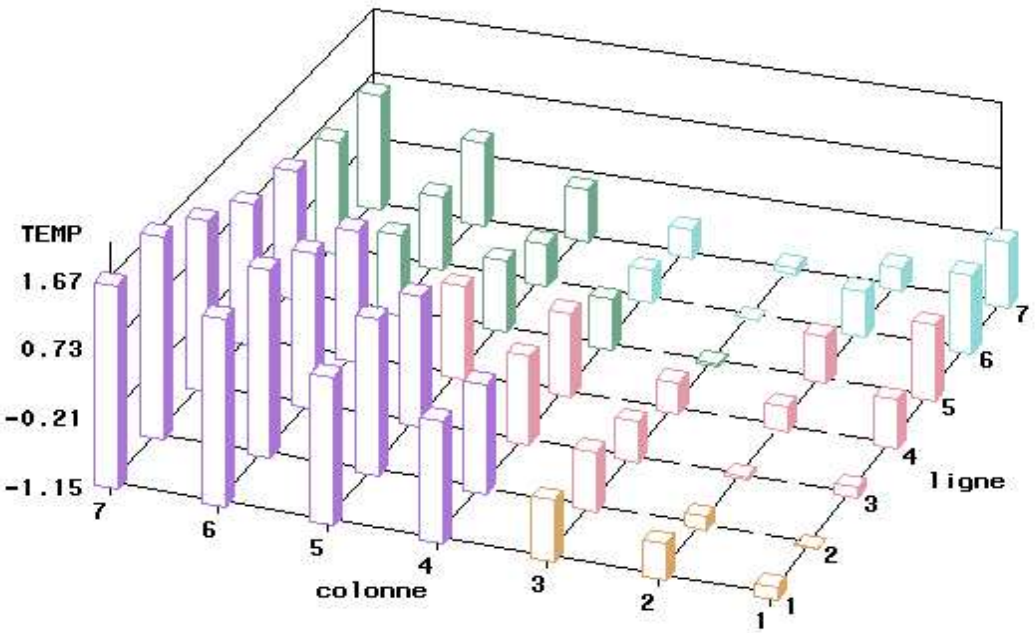


FIG. 15 – Répartition de la variable GRAD suivant les différentes classes

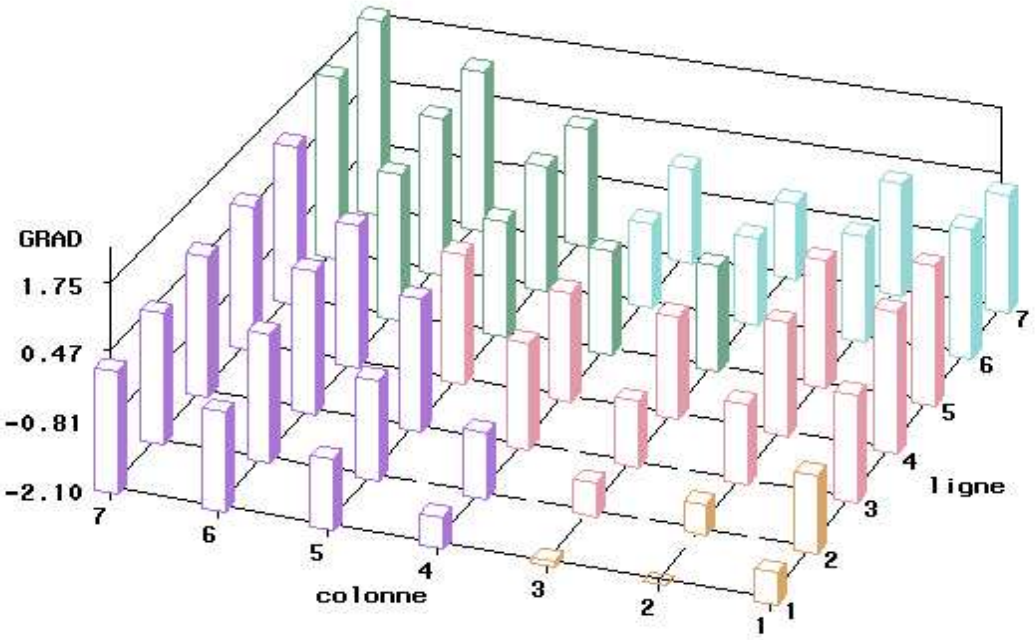
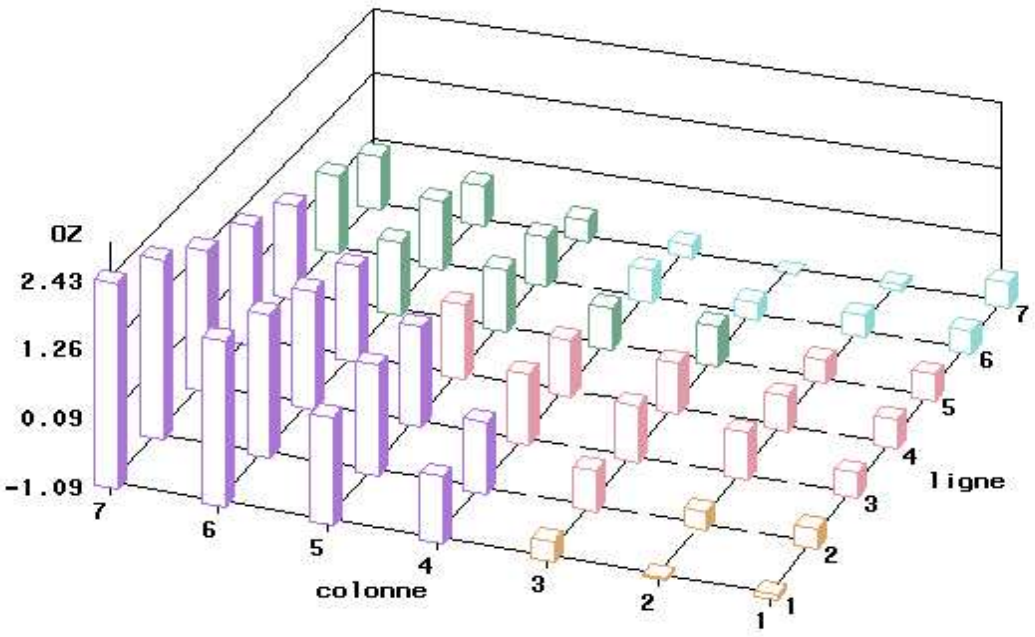


FIG. 16 – Répartition de la variable OZ suivant les différentes classes



4.2 Analyse discriminante

Nous avons une description d'une partie de notre modèle associé à l'état non-linéaire. Il est difficile cependant de dissocier clairement les deux états pour des régions du nuage où l'ozone et la température sont plus faibles. Nous allons donc employer des techniques classiques de classification pour essayer d'obtenir une séparation des deux états. Nous commençons par une analyse discriminante linéaire, puis nous continuons par une régression logistique. Cette dernière nous permettra en plus d'estimer la sensibilité de la probabilité des deux classes suivant les différentes variables.

4.2.1 Analyse discriminante linéaire

Comme il n'y a que deux états, l'unique vecteur discriminant a pour coefficients :

$$\begin{pmatrix} OZ24 & : & -0.038 \\ RAY & : & -0.313 \\ VENT & : & 0.074 \\ TEMP & : & 1.050 \\ GRAD & : & -0.308 \\ OZ & : & 0.390 \end{pmatrix}$$

On peut voir que pour séparer au mieux linéairement les deux états, il faut tenir compte essentiellement des forts niveaux de température mais aussi des faibles niveaux de rayonnement ainsi que des faibles niveaux du gradient de température. Ce résultat peut paraitre surprenant puisque la formation d'ozone est un processus photo-chimique. Cependant cela explique la présence des pics d'ozone même si le rayonnement est faible, du moment que la température est suffisamment élevée pendant une période assez longue. Ce phénomène, qui correspond à un modèle fortement non linéaire, intervient à Paris lorsque le temps est très chaud et pourtant nuageux (avant les périodes d'orages). De plus, cela ne veut pas dire que les pics soient impossibles lorsque le rayonnement est fort, cela implique seulement que, si le rayonnement est fort, le modèle se rapproche du modèle linéaire.

Les prédictions de l'analyse discriminante linéaire sur la série sont les suivantes :

	prévisions e1	prévisions e2
vrai e1	222	85
vrai e2	92	302

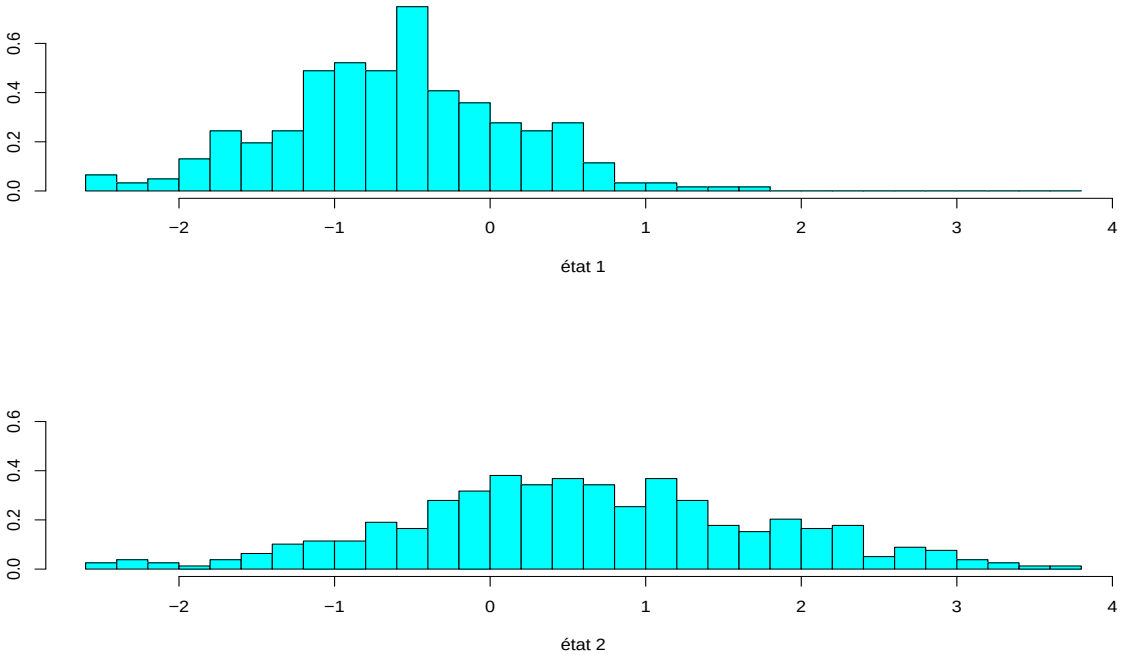
Cette méthode permet donc de bien prevoir l'état "2" dans $76,6\%$ des cas et l'état "1" dans $72,3\%$ des cas. L'histogramme suivant (figure 17) nous permet de visualiser la distribution du produit scalaire du vecteur discriminant avec les vecteurs de variables suivant l'état le plus probable de la chaîne de Markov cachée.

A titre de comparaison les prédictions de l'analyse discrimante quadratique sont les suivantes :

	prévision e1	prévision pe2
vrai e1	265	42
vrai e2	125	269

Elles sont meilleures pour l'état 1, car la variance associée à cet état est moindre, mais bien plus mauvaises pour l'état 2.

FIG. 17 – Historamme des deux classes de l'analyse discriminante



4.2.2 Régression logistique

Notons $\pi(x)$ la probabilité de l'état "2" pour un vecteur de variables x. Par convention $x_0 = 1$ et $x_1, \cdots, x_6$ correspondent aux valeurs de $OZ24$, RAY, $VENT$, $TEMP$, $GRAD$ et OZ. La formule du modèle logistique est alors

$$\pi(x) = \frac{\exp\left(\sum_{j=0}^{j=6} \alpha_j x_j\right)}{1 + \exp\left(\sum_{j=0}^{j=6} \alpha_j x_j\right)}$$

Les paramètres estimés par la régression logistique sont les suivants :

α_0(constante)	α_1 (OZ24)	α_2 (RAY)	α_3 (VENT)	α_4 (TEMP)	α_5 (GRAD)	α_6 (OZ)
0.450	-0.040	-0.472	0.070	1.214	-0.374	0.755

Ici encore, on remarque que les fortes probabilités de l'état associé au modèle non linéaire sont obtenues pour une forte température, un fort niveau d'ozone ainsi qu'un faible rayonnement et un faible gradient de température. Les prédictions de la régression logistique sur la série sont les suivantes :

	prévision e1	prévision e2
vrai e1	227	80
vrai e2	96	298

Les résultats sont extrêmement proches de l'analyse discriminante linéaire, puisque cette méthode permet de prévoir l'état "2" dans $75,6\%$ des cas et l'état "1" dans $76,4\%$ des cas. Finalement nous étudions la sensibilité des probabilités conditionnelles des états suivant les différentes variables prises isolément. Ici on ne fait varier qu'une seule variable, les autres sont mises à leur valeur moyenne, 0 puisqu'elles sont centrées. Les tableaux qui suivent donnent la valeur nécessaire de la variable pour obtenir les probabilités prédéterminées de l'état "2", ainsi que son écart-type.

TAB. 1 – Valeur et écart-type de OZ24 en fonction de la probabilité de l'état "2"

probabilité	valeur	écart-type
p = 0.1	66.101	297.269
p = 0.2	45.855	206.297
p = 0.3	32.399	145.835
p = 0.4	21.369	96.277
p = 0.5	11.246	50.816
p = 0.6	1.124	5.803
p = 0.7	-9.906	44.364
p = 0.8	-23.362	104.794
p = 0.9	-43.607	195.757

TAB. 2 – Valeur et écart-type de RAY en fonction de la probabilité de l'état "2"

probabilité	valeur	écart-type
p = 0.1	5.608	1.304
p = 0.2	3.890	0.908
p = 0.3	2.749	0.649
p = 0.4	1.813	0.446
p = 0.5	0.954	0.282
p = 0.6	0.095	0.204
p = 0.7	-0.840	0.305
p = 0.8	-1.982	0.535
p = 0.9	-3.700	0.921

TAB. 3 – Valeur et écart-type de VENT en fonction de la probabilité de l'état "2"

probabilité	valeur	écart-type
p = 0.1	-37.728	50.817
p = 0.2	-26.173	35.273
p = 0.3	-18.492	24.947
p = 0.4	-12.196	16.493
p = 0.5	-6.419	8.769
p = 0.6	-0.641	1.647
p = 0.7	5.654	7.717
p = 0.8	13.334	17.982
p = 0.9	24.890	33.509

TAB. 4 – Valeur et écart-type de TEMP en fonction de la probabilité de l'état "2"

probabilité	valeur	écart-type
p = 0.1	-2.179	0.278
p = 0.2	-1.511	0.200
p = 0.3	-1.068	0.151
p = 0.4	-0.704	0.115
p = 0.5	-0.370	0.090
p = 0.6	-0.037	0.080
p = 0.7	0.326	0.092
p = 0.8	0.770	0.128
p = 0.9	1.437	0.200

TAB. 5 – Valeur et écart-type de GRAD en fonction de la probabilité de l'état "2"

probabilité	valeur	écart-type
p = 0.1	7.076	2.306
p = 0.2	4.909	1.610
p = 0.3	3.468	1.152
p = 0.4	2.287	0.784
p = 0.5	1.204	0.468
p = 0.6	0.120	0.263
p = 0.7	-1.060	0.432
p = 0.8	-2.501	0.851
p = 0.9	-4.668	1.535

TAB. 6 – Valeur et écart-type de OZ en fonction de la probabilité de l'état "2"

probabilité	valeur	écart-type
p = 0.1	-3.502	0.937
p = 0.2	-2.429	0.649
p = 0.3	-1.716	0.460
p = 0.4	-1.132	0.311
p = 0.5	-0.595	0.187
p = 0.6	-0.059	0.127
p = 0.7	0.524	0.209
p = 0.8	1.237	0.382
p = 0.9	2.310	0.665

5 Conclusion

Nous avons utilisé un modèle autorégressif à changements de régime markoviens sur des données de pollution en niveau d'ozone à Paris. Ceux-ci semblent prometteurs pour prédire des phénomènes probablement associés à des changements de régime tels que les pics. Nous avons obtenu, grâce à ce modèle, une segmentation de la série que nous avons essayé d'interpréter à l'aide de techniques issues de l'analyse de données et de l'analyse discriminante. Il ressort de cette étude que ce qui distingue le mieux le régime non linéaire de la série sont les fortes températures, associées naturellement aux forts niveaux d'ozone. En outre nous avons remarqué que le modèle risque d'être d'autant plus non-linéaire que les niveaux de rayonnement et de gradient de température sont bas. Les parisiens constatent en effet chaque année que des pics interviennent lorsque le temps est "lourd", ce type de pics correspond au régime non-linéaire de notre modélisation de la série. Finalement, il faut remarquer que si, ni l'analyse exploratoire, ni l'analyse discriminante n'arrive à séparer totalement les deux états, c'est parce que la segmentation obtenue grâce aux chaînes de Markov cachées tient compte de la chronologie des évènements. Ainsi, les modèles de classification ne peuvent pas facilement supplanter cette technique puisqu'il leur manque toujours l'aspect temporel du processus.

Références

[1] L. Bel et. al. Elément de comparaison de prévisions statistiques des pics d'ozone. *Revue de Statistique appliquée*, 47 :3 :7–25, 1972.

[2] J.L. Chen, S. Islam, and P. Biswas. Nonlinear dynamics of hourly ozone concentrations : nonparametric short-term prediction. *Atmospheric Environment*, 32 :1839–1848, 1998.

[3] A.C. Comrie. Comparing neural network and regression models for ozone forecasting. *Journal of the Air and Waste Management Association*, 47 :653–663, 1997.

[4] M. Cottrell and J.C. Fort. Etude d'un algorithm d'auto-organisation. *Ann. de l'Inst. Henri Poincarré*, 23 :1–20, 1987.

[5] R. Douc, E. Moulines, and T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regimes. Technical reports 9, University of Lund, 2001.

[6] M.W. Gardner and S.R. Dorling. Statistical surface ozone models : an improved methodology to account for non-lionear behaviour. *Atmospheric environment*, 34 :21–34, 2000.

[7] J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57 :357–384, 1989.

[8] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, 1995.

[9] V. Krishnamurthy and T. Rydén. Consistent estimation of linear and non-linear autoregressive models with Markov regime. *Journal of time series analysis*, 19 :3 :291–307, 1998.