

AFD /ACP AFDM sur données environnementales Travail en binôme TP

Dans cette séance il s'agit de d'acquérir un savoir-faire de l'ACP, l'AFDM et l'AFD adaptées aux données et leur interprétation. Dans ce TP, les différentes étapes sont de :

- Faire une étude statistique préparatoire des données (moyenne, écart type, médiane, corrélation entre variables) : **attention aux données manquantes ou égale à 0**
 - Mettre en œuvre une ACP avec analyse et interprétation des résultats (avec représentativité de l'ACP)
 - Mettre en œuvre une AFD avec analyse et interprétation des résultats (avec représentativité de l'AFD)
 - Mettre en œuvre une AFDM avec analyse et interprétation des résultats (avec représentativité de l'AFDM)
 - Apporter des réponses aux questions posées sur ce cas de décision.
-
- ✓ Données fournies des 6 campagnes: BF2, BF3, CA1, CA2, CA3, CA4. BF pour bruit de fond et CA pour campagne avant installation du site. Un fichier TP4_covC1234_DS19_20.xls sous campus.
 - ✓ Vous disposez de vos programmes (*my_PCA_standardized.r* et *my_AFD.r*) réalisés dans les TD/TP précédents.
 - ✓ Mais surtout vous utiliserez l'outil de R : FactoMiner pour AFDM
-

Le travail à rendre se fait sous la forme d'un rapport de type *note de synthèse* contenant les choix, les traitements et analyses faites sur ce type de données avec les codes sources développés + résultats dans un même fichier.zip (sous format de : *noms_binome.zip*) sur campus pour le **29 octobre** .

Objet d'étude :

Dans le cadre de projet de recherche industrielle, on s'intéresse à la contribution d'un site industriel de traitement de déchets verts par compostage lors de la mise en exploitation, localisé dans la Loire. En effet, un tel processus dans certaines conditions de fonctionnement (entrant important à différentes périodes de l'année, conditions de fermentation anaérobie au lieu de dégradation aérobie, mauvaise gestion du site) peut entraîner l'émission de composés chimiques avec des risques sanitaires potentiels au niveau des populations avoisinantes.

Afin de discriminer la contribution du site par rapport à la présence éventuelle de ces composés avant installation (que l'on appelle bruit de fond) des campagnes de mesure de ces composés ont été effectuées avant (dans *le labels* les 2 lettres BF) et après la mise en activités du site (lettre CA dans le *labels*) à différentes périodes de l'année (H pour hiver et E été).

On cherche donc à répondre à certains questionnements comme :

- la localisation des m points de mesure autour du site, montre-elle des regroupements de comportement (composés chimiques atmosphériques d'origine industrielle, automobile, milieu urbain, milieu rural...)?
- existe-il une différence entre les campagnes hiver/ été ?
- existe- il une signature entre les individus avant et après la mise en activité du site ?

Description des données fournies

On dispose de 6 campagnes de mesure effectuées sur m points de mesure, pour un certain nombre de COV (composés organiques volatiles) : Liste des p variables ($p=14$) : p composés (ou familles de composés) : familles de composés COV

B	T	E	X	9_ane	10_ane	13_ane	14_ane	1_M_2_PA	BTM	FormicAcid	aceticacid
				NonaDecanoicAc	Tot_OcNoDecana						

9_ane	Nonane 9		B T E X	Benzene B
10_ane	Decane 10			Toluène T
13_ane	Tridecane			Ethylbenzene E
14_ane	Tetradecane			Xylene X
1_M_2_PA	1 Methoxy-2-propyl acetate			
BTM	Benzene x,y,z triméthyl			
FormicAcid	Acetic acid, butyl ester			
aceticacid	acide acétique			
NonaDecanoicAc	acide nonanoïque	acide décanoïque		
	octanal	nonanal	decanal	

Effectués sur m points (localisation) donnés ci-dessous (plusieurs mesures sur certains mêmes points) :

P19	P21	P18	P17	P20	P10	P02	P01	P05	P06	P07	P03	P08	P13
P14	P15	P09	P16	P11	P19	P21	P18	P20	P02	P05	P06	P07	P03
P08	P13	P14	P15	P04	P09	P16	P11	P10	P12	P19	P21	P18	P17
P20	P05	P06	P07	P03	P08	P13	P14	P15	P04	P09	P16	P10	P02
P01	P12	P22	P19	P21	P18	P17	P20	P05	P06	P07	P03	P08	P13
P14	P15	P04	P09	P16	P32	P33	P10	P29	P11	P26	P30	P27	P28
P25	P02	P01	P12	P19	P21	P17	P20	P05	P06	P07	P03	P08	P13
P14	P15	P09	P16	P32	P29	P11	P26	P27	P28	P25	P02	P01	P12
P19	P21	P18	P17	P20	P05	P06	P07	P03	P08	P13	P14	P15	P04
P09	P16	P32	P33	P29	P11	P26	P30	P27	P28	P25	P02	P01	P12

Données météorologiques durant les campagnes de mesures

Campagne	Date de début	Date de fin	nombre de jours	Période	T° moyenne (°C)	Cumul de pluie (mm)	% de temps sans vent	Principales directions des vents (%)			% de vents supérieur à 2 m.s-1
BF1	09/05/2007	16/10/2007	160	Estivale	16,6	386	30,6	N : 23,7%	NNW : 6,4%	SE : 4,4%	17,57
BF2	16/10/2007	05/11/2007	20	Hivernale	7,6	2	31,1	N : 41%	NNW : 16,3%	NW : 4,2%	23,2
BF3	25/01/2008	08/02/2008	14	Hivernale	2,9	11,9	52,5	N : 12,5%	SSE : 6%	S : 5,4%	17,8
CA1	15/02/2009	05/03/2009	18	Hivernale	4,6	6	49,5	N : 19,5%	SE : 11,7%	NNW : 11,7%	8,0
CA2	09/09/2009	23/09/2009	14	Estivale	15,6	11	47,4	NNW : 26,3%	N : 17,8%	E : 2,35%	9,0
CA3	02/03/2010	16/03/2010	14	Hivernale	1,2	2,5	36,5	N : 30,4%	NNW : 24,9%	NW : 4%	25,6
CA4	05/07/2010	16/07/2010	11	Estivale	23,2	28,7	54,9	N : 19,3%	SE : 8,3%	NNW : 3,3%	7,3

Vous devez choisir une **stratégie de traitement de données multivariées**:

- ✓ En effet vous disposez de données qui sont dans des ordres de grandeurs différentes (des concentrations (ng/m³)): pour l'ACP vous devez au moins centrer vos données ; si vous voulez vous affranchir du problème des échelles vous devez réduire vos données (pour visualiser sur le cercle de corrélation mais pas seulement... il est recommandé de centrer et de réduire les données initiales).
- ✓ Les données disponibles présentent des données manquantes : si vous voulez comparer plusieurs périodes il faut choisir des données soit : en remplaçant quelques valeurs même si vous biaiserez votre approche, mais vous conservez les variables les plus échantillonnées que possibles ; soit, vous pouvez aussi prendre une approche complémentaire en faisant une ACP sur des données avec moins de variables mais complètes et comparer (idem en AFD).
- ✓ Vous disposez de données pour chaque campagne de mesure d'un échantillon de mesure effectuées en ces $n =$ nombre des points localisés autour du site et p variables = nombre de types de composés mesurés :

- un individu X_i est un vecteur ligne, X_i^j une mesure des $j=1$ à p composés, en un point donné pour une période de temps donnée. chaque nouvelle campagne de mesure effectuée au niveau de la même station de mesure est un re échantillonnage en ce point : il constitue un nouvel individu si l'on décide que cela ne constitue pas une redondance d'échantillonnage : $n = \text{nombre de campagne} \times \text{nombre de points de mesures}$ individus au total

Vous disposez aussi de 4 variables qualitative que sont : TYPE : environnement du site soit urbain, industriel, rural, le site de compostage ; la SAISON : (hiver - été) ; Campagne : (4 campagnes en été et 2 campagne hivernale) ; Localisation : un label de point.

On vous demande successivement de réaliser les étapes suivantes :

Etape :1^{ère}

- 1) Le traitement statistique des données permettra d'évaluer la variabilité :
 - a. sur l'ensemble de l'échantillon - c. des 6 différentes campagnes.
 - b. pour les campagnes de mesure en hiver et les campagnes en été d. sur les deux campagnes avant ouverture du site (BF) et après ouverture du site (CA)
- 2) L'affichage des corrélations possibles entre les différentes variables : pour chacune des 4 périodes (hiver, été, avant activité, après activité)
- 3) Des traitements statistiques que pouvez-vous en déduire sur les différents périodes hiver/été et avant et après ouvertures du site ?

Etape :2^{ème}

- 4) On cherche à savoir si l'on peut identifier une réduction du nombre de variables par ACP: recherche de composantes principales et si les individus se regroupent ou pas selon ce nouvel espace R ($q < p$ avec $p=14$). **Vous devez mettre en œuvre l'ACP en justifiant les résultats (Inertie expliquée/inertie totale, qualité de la réduction de dimension et de la qualité des projections de individus sur les 14 variables quantitatives. Quelles variables sont les mieux expliquées ?**
- 5) On recherche une signature des composants pour chaque période (été, hiver, 1 avant_activité, 1 apres_activité) ; une réduction du nombre de variables par ACP serait-elle une méthode adaptée pour tenter de répondre et comment la mettre en œuvre ?

Compte tenu de vos résultats de cette étape : quels sont vos principaux constats, quelles propositions de traitements faites-vous pour chaque période, quelles sont les variables marqueurs ?

Etape : 3^{ème}

Vous disposez maintenant d'information sur 4 variables de type catégorie : AFD s'intéresse à une (des) variable(s) de type qualitative (période (avant ou après installation du site), saison (été/hiver), localisation, ou le type d'environnement de proximité).

- 6) Avant de faire une AFD , les statistiques par variable sur les données selon les deux types de modalités hiver/été de la variable saison : y-a-t-il des différences entre les groupes (H/E) ?
- 7) Afin de discriminer au mieux les groupes a) *saison (hiver/été)* on vous propose de mettre en œuvre une AFD sur la variable qualitative concernée: qu'observez-vous ? que vaut le critère donné η (variance interclasse/ variance totale) pour les axes discriminants Y de valeurs propres λ principaux : que reprenez-vous pour l'interprétation ?
- 8) Les statistiques sur les données selon les deux types de modalités (avant installation (BF) ou après installation du site de compostage (CA)) montraient-elles des différences entre les groupes en termes de variance totale et variances inter et intraclasse?

- 9) Afin d'évaluer la contribution d'un site au niveau de la qualité de l'air ambiant, on vous propose de mettre en œuvre une AFD sur la variable qualitative *période* (CA et BF) et donner les résultats de l'AFD (qualité de la réduction, fonction linéaire discriminante, critère de et votre interprétation en terme de variables contribuant le plus à la discrimination).
- 10) Vous avez alors deux résultats ACP et AFD sur les données : les deux réductions ne sont pas faites selon le même critère mais pouvez-vous conclure sur l'effet sur de l'activité du site sur l'environnement ou non ?

Etape : 4^{ème}

Il existe une généralisation de l'AFDM qui intègre ACP et AFC pour plusieurs var. quantitatives et qualitatives, appelée Analyse factorielle des données mixtes.

- 11) Préparer les données au format attendus par le package FactoMineR
- 12) Mettre en œuvre cette méthode à partir des packages disponibles pour avoir une réduction de dimension sur l'espace de 14 var. quantitatives et des 4 variable qualitatives
- 13) Interprétation des résultats obtenus intégrant ces 18 var. au total
- 14) Cela vous apporte il des éléments complémentaires à la première ACP ?
- 15) Que pourriez-vous suggérer pour établir les éléments de comparaison entre groupes campagne hiver/été, groupe avant et après installation industrielle, groupe en fonction de la localisation du point de mesure (urbain, rural, site industriel, sur site de compostage).
- 16) Vous avez alors deux résultats ACP et AFDM sur les données : les deux réductions ne sont pas faites selon le même critère mais pouvez-vous conclure sur l'effet sur de l'activité du site sur l'environnement ou non ?

17) Questions complémentaires (compter en plus si réaliser)

On s'intéresse maintenant à la signature de profils *i* de concentration de chaque individu (*i* point échantillonné, parmi les *n*) : une première estimation faite par les chimistes est d'attribuer un type (*rural, urbain, compostage, site industriel*) à chaque individu : pouvez-vous à partir d'une statistique de type AFD sur cette variable qualitative 'type' proposer une réduction et une analyse de la discrimination des groupes : pensez-vous que ce regroupement empirique initiale est cohérente avec la localisation effective du point de mesure dans son environnement immédiat ?

- 18) Enfin un individu n'est pas typé par son environnement (?) pouvez l'extraire et refaire l'AFD et faire la prévision d'appartenance à sa classe en utilisant AFD en mode prédictif ?

Les points A et B sont renommés par P (ex : A01 et B01 sont le même point P01 dans le fichier)

