MASTER : Traitement de l'Information et Exploitation des Données
Fouad Badran, Cécile Mallet, Carlos, Mejia,
Charles Sorror, Sylvie Thiria

TPA05 : ACP appliquée à des températures locales et effet de serre

I - Objectifs

Ce TP est dédié à la mise en œuvre de l'Analyse en Composantes Principales. Il s'agit d'appliquer cette méthode pour mener à bien l'interprétation des données.

Nous travaillerons ici avec des données spatio-temporelles de température pour lesquelles nous avons prévu, deux parties:

- Une étude où les lieux géographiques seront utilisés comme variables et où les pas d'évolution temporelle formeront les individus. Vu sous forme de tableau, on aura en colonne les variables-villes et en ligne les individu-mois. Cela constituera une 1ère partie de TP.
- Dans la seconde partie, les rôles des dimensions devront être inversés. Autrement dit, les pas d'évolution temporelle seront les variables, et les lieux géographiques les individus. On aura donc en colonne des « variables-mois » et en ligne des « individus-villes »

==========

Le rapport de TP devra être synthétique. Il doit montrer la démarche suivie, et ne faire apparaître que les résultats nécessaires. Il s'agit de quantifier les résultats tout en rédigeant un rapport qui les analyse et les commente. Les paramètres utilisés devront être indiqués. Les graphiques des expériences doivent être insérés dans le rapport. Pour toutes les figures que vous présenterez, essayez de les compléter avec des éléments nécessaires à leur compréhension (titre, légende, colorbar, label des axes, etc...).

II - Les Données

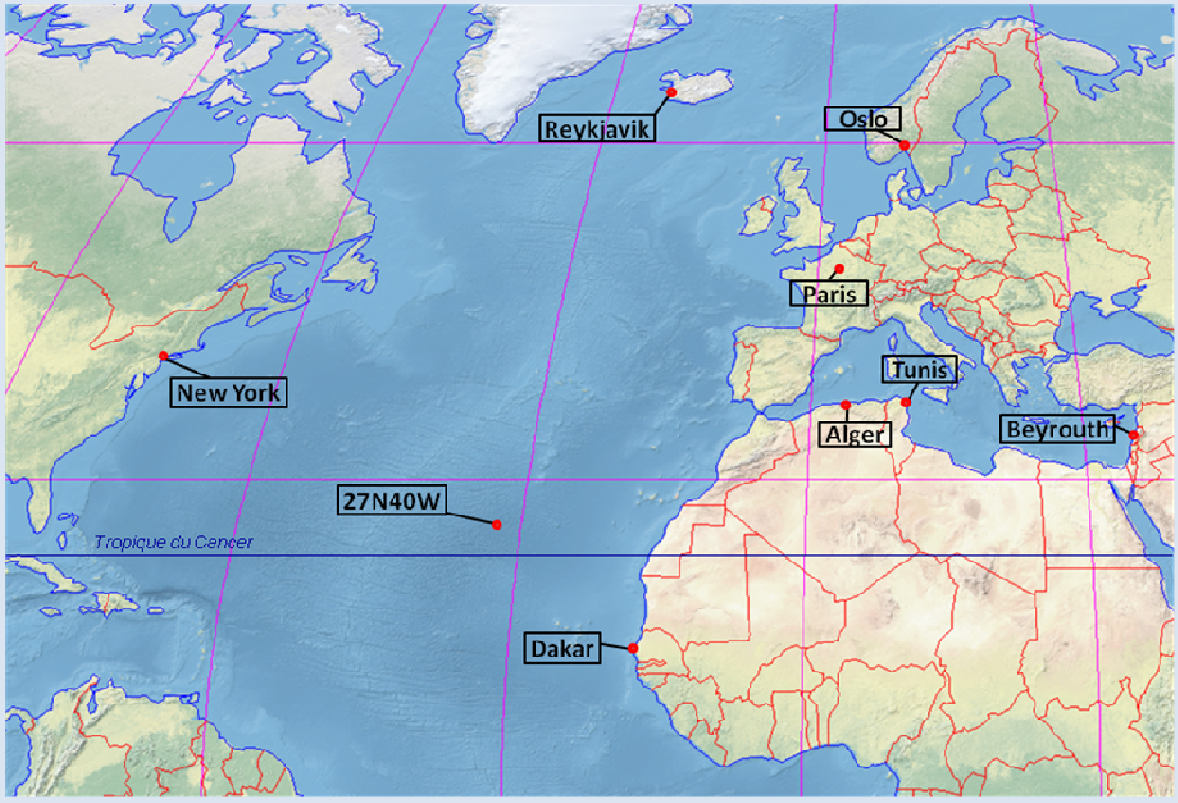
Pour la mise en œuvre de ce TP, nous utiliserons deux variables :

- La température (**t2**) à 2 mètres du sol (en degré celsius) pour 9 lieux géographiques. Cette données est issue et de la base ERA-Interim du centre européen ECMWF. Il s'agit d'une donnée calculée par un modèle, après assimilation des données en se positionnant à midi. Les lieux pour lesquels nous avons extrait les valeurs sont dans l'ordre du nord au sud :

Reykjavik	64°08'07.14"N	21°53'42.63"O
Oslo	59°54'49.85"N	10°45'08.18"E
Paris	48°51'12.03"N	2°20'55.59"E
New York	40°42'51.67"N	74°00'21.50"O
Tunis	36°49'07.72"N	10°09'57.46"E
Alger	36°45'10.39"N	3°02'31.37"E
Beyrouth	33°53'19.06"N	35°29'43.72"E
Atlan27N40W	27°00'00.00"N	40°00'00.00"O
Dakar	14°39'46.09"N	17°26'13.65"O

- Le CO_2 en molfrac ppm (parties par million) dont les mesures de concentration on été réalisées sur le mont Mauna Loa à Hawaii. Ces données qui proviennent de la NOAA.

Pour ces deux variables nous avons réalisé une moyenne mensuelle de la période allant de janvier 1982 à décembre 2010, soit 29 années complètes. Les fichiers qui contiennent ces données sont **clim_t2C_J1982D2010.mat** pour la température et **clim_co2_J1982D2010.mat** pour le CO_2.
La 1ère colonne de ces fichiers correspond à l'année, la seconde au mois. Les deux fichiers sont en en correspondance sur ces deux premières colonnes, elles contiennent donc le même nombre de lignes (**N=348**).
- Pour le fichier **clim_t2C_J1982D2010.mat**, les colonnes 3 à 11 contiennent les valeurs des températures pour les 9 lieux dans l'ordre où on les a énumérés
- La 3ème colonne du fichier **clim_co2_J1982D2010.mat** contient la valeur de concentration du CO_2.



Rappels partiels et notations pour l'Analyse en Composantes Principales (ACP)

L'ACP est une méthode statistique qui consiste à effectuer un changement de base (projection dans un nouveau repère) pour réduire le nombre d'axes nécessaires à la compréhension des données tout en maximisant la variance projetée. Elle consiste à déterminer $C = XU$ où X sont les données centrées de n individus (en ligne) et p variables (en colonne), U est la matrice de passage. Elle est composée des vecteurs propres qui définissent les axes principaux qu'il convient de trouver. La matrice C résultantes est constituée des nouvelles variables (dans la nouvelle base) appelées composantes principales (CP). On établit que :

$X^tX\boldsymbol{u}_k = \lambda_k\boldsymbol{u}_k$. avec $\boldsymbol{u}_k$ le *kième* vecteur colonne de U.

Les inconnues à déterminer sont les vecteurs propres de X^tX.

Les nouvelles variables (CP) étant des combinaisons linéaires des variables initiales, l'interprétation d'une ACP peut être délicate. Pour nous y aider, on est amené à s'intéresser aux éléments suivants :

- Le rapport d'une valeur propre λ_k à la somme des autres ($\lambda_k / \Sigma_i\lambda_i$) est la part de l'inertie (ou variance expliquée) par l'axe k. L'étude se réduit alors aux plans formés à l'aide des k premiers axes qui cumulent suffisamment d'inertie ou qui offrent un intérêt particulier.

- On détermine les corrélations entre les nouvelles et les anciennes variables ($r(C_k,X_h)$). En prenant les composantes 2 à 2, on peut reporter ces corrélations sur un cercle (appelé cercle des corrélations). Cette représentation aide à l'interprétation des données. Lorsque les données initiales sont centrées et réduites on a :

$$r(C_k,X_h) = \boldsymbol{u}_{h,k}\ \sqrt{\lambda_k}\ .$$

- Le nuage des individus : il s'agit de représenter graphiquement les coordonnées des individus sur les nouveaux axes pris 2 à 2.

- La qualité de représentation d'un individu ($\boldsymbol{o}_i$), de norme $\|o_i\|$, par un axe k est donnée par :

$qlt_k(\boldsymbol{o}_i)= c_{ik}\ / \|o_i\|^2$ avec c_{ik} la coordonnée de l'individu i sur l'axe k.

Un individu mal représenté sur un axe ne devrait pas trop intervenir dans l'interprétation de cet axe.

- La contribution d'un individu ($\boldsymbol{o}_i$) à la fabrication d'un axe k est donnée par :

$ctr_k(\boldsymbol{o}_i) = q_ic_{ik}^2/\lambda_k$ où q_i est le poids de l'individu i.

C'est la part de la variance de l'axe k qui est due à l'individu i, q_i représentant le poids de cet individu dans l'analyse. La contribution permet de s'assurer qu'un individu n'est pas prépondérant dans la définition d'un axe. Elle permet de repérer des valeurs extrêmes si trop peu de données ont des contributions significatives. Par la suite, on omettra q_i en considérant qu'il s'agit d'un poids uniforme et égal à 1. Il est possible d'introduire ces poids si, on a des connaissances sur la significativité des individus.

Pour un résumé partiel un peu plus détaillé sur l'ACP, vous pouvez vous reporter au document « *RappelACP* »

III - Eléments pour la réalisation du TP

- **centred.m** : Normalisation par centrage réduction.
- **phinertie** : Calculs et cumuls des pourcentages d'inertie de valeurs propres positives et représentation graphique par histogramme.
- **corcer** : Cercle des corrélations : représentation sur un cercle des coefficients de corrélation entre deux nouvelles variables et l'ensemble des variables initiales. Cette fonction permet d'afficher dans une même couleur les vecteurs qui ont un même label de variable, ce qui la rend un peu compliquée à utiliser. Dans le cas le plus simple, comme pour la 1ère partie du TP, il suffit de ne passer que les 5 premiers paramètres, et d'omettre les 2 derniers. Pour la seconde partie, les 2 derniers paramètres devront être utilisés en fonction du résultat attendu. Il conviendra donc de bien lire la documentation de cette fonction.
- **qltctr** : Qualité de représentation et contribution.
- **acpnuage** : Nuage des individus d'une ACP sur le plan de 2 composantes, dont les points peuvent être associés à une couleur selon un vecteur de valeurs à construire. Chacun des points doit ainsi avoir son niveau de couleur associé dans ce vecteur. Chaque point peut être représenté par 2 triangles orientés selon l'abscisse et l'ordonnée et dont les tailles peuvent être proportionnées (par les composantes d'une matrice à 2 colonnes). Les triangles sont orientés de la façon suivante :
- avec Matlab : vers la droite pour l'abscisse et vers le haut pour l'ordonnée,
- avec Octave3.2.4 : vers le haut pour l'abscisse et vers le bas pour l'ordonnée.
- **regtrace** : Régression linéaire : Coefficients, et optionnellement traçage, de la droite de régression linéaire : y = b0 + b1*x
- **colmois** : Une proposition de map de couleur pour les mois.

- La fonction Matlab qui fait l'ACP s'appelle **princomp**. Selon les versions de Matlab, les noms des variables de sortie (en particulier, la 1ère) de cette fonction peuvent être différents :

[PC, SCORE, LATENT, TSQUARE] = princomp(X)

ou [COEFF, SCORE, LATENT, TSQUARE] = princomp(X) ...

- Attention, la 1ère sortie (PC ou COEFF) correspond à la matrice des vecteurs propres U (il ne s'agit donc pas des composantes principales, mais des coefficients qui permettent de les obtenir)
- SCORE sont les composantes principales C (CP).
- LATENT sont les valeurs propres λ.
- TSQUARE renvoie au test statistique d'Hotteling pour chaque point (nous n'en aurons pas l'usage ici)

Les différentes versions de Matlab pouvant avoir des noms de variables de sortie différents, il conviendra de s'assurer à quoi ils correspondent.

- **ocp** : ACP utilisable pour Octave-3.2.4 (qui ne dispose pas de **princomp**).

Vous pouvez vous référer à la fonction d'aide (**help**) de ces fonctions pour avoir des précisions sur leur utilisation.

IV – 1ère partie : ACP des températures
(« individus-mois » « variables-villes »)

Dans cette 1ère partie, vous devrez utiliser les données de température (fichier **clim_t2C_J1982D2010.mat**) en prenant les villes en variables et les années-mois en individus. Il s'agira donc d'étudier l'évolution des températures entre ces villes.

1°) Représentation des données de température : Nous vous demandons en premier lieu de représenter graphiquement les données brutes contenues dans le fichier **clim_t2C_J1982D2010.mat**. Pour cela vous devez :

1.1°) Faire une figure avec une courbe des températures, de couleur différente, pour chaque ville (soit 9 courbes). L'abscisse devra mentionner les années, et l'ordonnée les températures.
-> Indications pour les couleurs des courbes : Vous pouvez créer votre propre table de couleurs extraite à intervalle régulier d'une « map » de couleur donnée de la façon suivante :

```
Col = jet(nombre_de_courbes);          % votre map personnalisée, ici, avec la map « jet » par exemple
```

Par la suite, pour tracer une courbe i avec l'instruction **plot**, vous pouvez utiliser comme paramètre : 'color', Col(i, :)

-> Indications pour l'abscisse avec les années : Vos labels pourront être créés en repérant l'indice du 1er mois de l'année (janvier) de la façon suivante (par exemple ici avec clim_t2) :

```
Ilab  = find(clim_t2(:,2)==1 );        % Indices des mois de janvier
Xlab = clim_t2(Ilab,1);                % en label, l'année de ces indices
```

Pour labéliser un axe courant, vous pourrez alors saisir l'instruction :

```
set(gca,'XTick',Ilab, 'XTickLabel',Xlab) ;
```

Si vous butez sur cette représentation graphique et pour vous éviter de perdre trop de temps, vous pourrez utiliser la fonction **plotclimt2**.

1.2°) Sur 2 autres figures (ou une seule avec 2 subplot) tracer la courbe des valeurs moyennes des villes puis celle des écarts types.

2°) Analyse en Composante Principale (ACP).
Pour réaliser l'étude des données par une ACP, vous devrez neutraliser les niveaux différents de température des villes pour ne pas en favoriser certaines au détriment d'autres. Pour cela l'ACP devra donc être réalisée sur les valeurs normalisées par centrage et réduction (fonction à utiliser : **centred**).
L'ACP elle-même pourra être calculée avec la fonction matlab **princomp** ou **ocp**. Pour rendre compte des résultats, nous vous demandons de réaliser les tâches suivantes :

2.1°) Présenter une figure qui montre les pourcentages d'inertie de chaque axe. Fonction à utiliser : **phinertie**.
Ne retenir, pour la suite de l'étude, que les 2 axes principaux qui ensemble recueillent au moins 94% de l'inertie.

2.2°) Calculer les contributions des individus pour les 2 premiers axes et les représenter graphiquement. S'assurer qu'il n'y a pas d'individus qui contribueraient trop fortement à la formation des axes par rapport aux autres. La formule de calcul à utiliser est : $ctr_k(\boldsymbol{o}_i) = q_i c_{ik}^2/\lambda_k$ ce qui peut se traduire en Matlab :

```
CTR = (1/(size(CP,1)-1))*(CP.^2) ./ (ones(size(CP,1),1)*λ');
```

2.3°) Présenter et commenter le cercle des corrélations entres les variables et les 2 premières composantes principales (fonction à utiliser : **corcer**).

2.4°) Représentation du nuage des individus
En préparation de la visualisation vous calculerez :
- La température moyenne (**t2moy**) calculée sur l'ensemble des villes, c'est un vecteur de la même dimension que celui des individus (348).

- La qualité de représentation :
$qlt_k(\boldsymbol{o}_i) = c_{ik}^2 \;/\; \|o_i\|^2$, que l'on peut programmer en Matal comme suit :

```
dist = (sum((CP').^2))';                % où CP sont les composantes principales
QLT = (CP.^2) ./ (dist*ones(1,size(CP,2)));
```

- Présenter et commenter le nuage des « individus-mois » sur les plans principaux 1 et 2 à l'aide de la fonction **acpnuage** avec les paramètres suivants (outre la matrice des composantes principales):
  - Une échelle de couleur associée à la température moyenne **t2moy**.
  - Une labellisation des individus par leur mois
  - La qualité de représentation. Choisir empiriquement un facteur de « zoom » convenable qui rende la figure exploitable.

2.5°) Pour vous aider à l'interprétation du 2ème axe en particulier, vous devrez réaliser une figure des courbes de climatologie mensuelle pour chaque ville avec les valeurs normalisées. C'est-à-dire que pour chaque ville, vous devrez calculer la moyenne de chaque mois. Puisqu'il y a 9 villes et 12 mois, vous devrez donc obtenir 9 courbes de 12 points que vous représenterez sur la même figure mais de façon à, pouvoir distinguer chacune des villes, par des couleurs ou marqueurs différents par exemple. Complétez alors vos commentaires.

2.6°) Faire de nouveau le nuage des « individus-mois » sur les plans principaux 1 et 2 en utilisant cette fois, comme échelle de couleur, la concentration du CO_2 en valeur corrigé de sa tendance globale. Pour cela, vous devrez utiliser les données du fichier **clim_co2_J1982D2010** qu'il conviendra de corriger en retirant la tendance globale. Pour cela, on estimera cette tendance linéairement en utilisant la formule : ***CO₂cor*=*CO₂*–b1*tclim**, où **tclim**=[1:taille des données] correspond aux pas de temps et **b1** est le coefficient de la pente de régression linéaire. Vous pourrez obtenir ce coefficient en utilisant la fonction **regtrace**.
Les individus devrons toujours être labellisés par leur mois, mais il ne sera par contre pas utile cette fois de proportionner la taille du marqueur à un quelconque indicateur.

V – 2ème partie : ACP des températures
(« individus-villes » « variables- mois »)

Pour cette 2ème partie, vous devrez opérer une transposition des données de température (fichier **clim_t2_J1982D2010.mat**) de sorte que les individus soient les villes et que les variables soient les mois (en fait, il s'agit des mois pour chaque année). Chaque ville est ainsi représentée par une série chronologique. Avant d'effectuer l'ACP, ces données devront avoir été normalisées par centrage réduction (fonction **centred**). Cette normalisation à pour effet d'atténuer l'évolution interannuelle mais elle préserve une évolution saisonnière.

1°) Représentation des données
Nous vous demandons tout d'abord de représenter la série chronologique de chacune des villes. Pour apprécier visuellement l'effet de la normalisation on vous demande de reprendre la figure des données brutes (cf 1.1°), et de faire une figure du même type mais cette fois avec les données normalisées. On aura là encore une courbe par ville, soit 9 courbes qui devront avoir des couleurs différentes repérables par une légende ; les années devront être indiquées en abscisse.

Pour quantifier l'atténuation de la tendance globale induit par la normalisation, vous devrez déterminer, en utilisant la fonction **regtrace**, la valeur des pentes des courbes de chaque ville avant et après normalisation. Vous pourrez ne mentionner que les valeurs minimums et maximums obtenues (en valeur absolue).

2°) Faire le calcul de l'ACP pour les données centrées et réduites (Points « individus-villes » « variables-mois »):

2.1°) Produire une figure qui indique les pourcentages d'inertie des valeurs propres associées aux composantes. Fonction à utiliser : **phinertie**.
Ne retenir, pour la suite de l'étude, que les 2 axes principaux qui recueillent le plus d'inertie.

2.2°) Cercle des corrélations : A l'aide de la fonction **corcer** nous vous demandons de produire 3 figures du cercle des corrélations :
- La première devra représenter, sur un seul et même cercle, les vecteurs de toutes les variables.
- Avec 348 variables, la 1ère figure risque d'être (un peu) encombrée, nous vous demandons alors de produire une seconde figure contenant (en subplot) un cercle pour chaque mois (un cercle pour tous les vecteurs des mois de janvier, un pour tous les vecteurs des mois de février, etc…).
- Enfin, une 3ème figure pourra être produite avec un seul cercle des corrélations contenant les vecteurs moyens des 12 mois.

La fonction **corcer** devra donc être appelée par 3 fois ; c'est en jouant sur les paramètres passés que vous pourrez obtenir ces 3 différentes figures pour lesquelles on vous demande d'utiliser une couleur spécifique pour identifier les vecteur de chaque mois.

2.3°) Calculer les contributions des individus pour les 2 axes retenus et les représenter graphiquement. S'assurer qu'il n'y a pas d'individus qui contribueraient plus fortement à la formation des axes que les autres. Vous pouvez éventuellement utiliser la fonction **qltctr** qui calcule les contributions mais également les qualités de représentation qui vous servirons par la suite.

2.4°) Nuage des individus sur le plan principal retenu

a) Présenter le nuage des points « individus-variables », qui devront être labellisés par les noms des villes. Ils devront être représentés par des triangles proportionnés à leurs qualités de représentation. Ces dernières devront également être rapportées en clair.

b) Pour l'interprétation de l'axe 1, vous pourrez comparer la courbe des moyennes des températures brutes par ville à celle de la première composante principale :
- Représenter les éléments de cette comparaison.
- Indiquer la valeur de la corrélation linéaire entre les valeurs moyenne et la composante principale.

c) Pour vous aider dans l'interprétation de l'axe 2 qui présente certainement d'avantage de difficultés nous vous proposons de mettre en regard deux figures de climatologie mensuelle (moyenne mois par mois), l'une sur les données brutes, l'autre sur les données normalisées. Chaque figure devra comporter 9 courbes (une par ville) de 12 points (un point pour chaque mois). Il sera peut être utile aussi de se rappeler que la normalisation consiste à rapporter l'anomalie à l'écart type (l'anomalie étant la différence d'une variable à la moyenne de cette variable ($x_i - \bar{x}$)).