MASTER : Traitement de l'Information et Exploitation des Données

Fouad Badran, Cécile Mallet, Carlos, Mejia,
Charles Sorror, Sylvie Thiria

TPA04 : ACP appliquée à des variables climatiques liées à lœffet de serre

I - Objectifs

Ce TP porte sur des données géophysiques environnementales qui sont habituellement prises en compte dans l'étude de l'effet de serre. Nous nous intéresserons plus spécifiquement à l'étude de ces données à l'aide de l'ACP (Analyse en Composantes Principales). Le but du TP sera d'apprendre à mettre en œuvre une ACP et à se familiariser à l'interprétation des résultats qu'elle produit. Les données dont nous disposons correspondent à différents villes, comme cela est décrit ci-après. Des ACP différentes peuvent être menées pour chacune des villes prise individuellement. Nous nous limiterons cependant à trois d'entre elles. Nous avons découpé le TP en 2 parties. Dans la 1ère, une climatologie mensuelle devra être présentée suivie de 2 ACP pour la ville de Reykjavik, l'une de type saisonnière, l'autre de type interannuelle. Dans la 2ème partie ces mêmes types d'ACP seront à réaliser sur les villes d'Alger et de Dakar.

==========

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

Les données que nous utiliserons sont issues d'une sélection de la base de données ERA-Interim du centre européen ECMWF. Il s'agit de données modèles pour 5 variables sur 9 lieux géographiques. Nous disposons également de mesures de CO_2 réalisées sur le mont Mauna Loa à Hawaii, données qui proviennent de la NOAA. Pour chacune de ces variables nous avons calculé une **moyenne mensuelle** de la période allant de janvier 1982 à décembre 2010, soit 29 années complètes. Pour les 9 lieux, il y a deux sortes de données :

- les données analysées proviennent des modèles au sortir de l'assimilation des données en se positionnant à midi.
- les données de prévision (dites « Forecast ») sont obtenues en faisant fonctionner le modèle 24 heures après une assimilation, elles sont donc aussi positionnées à midi. Elles présentent plus d'incertitude que les premières.

Liste des variables:	noms des fichiers fournis
1) Données Analysées à midi :	
t2 : Temperature at 2 meters (degC) (Température à 2 mètres)	**clim_t2C_J1982D2010.mat**
tcc : Total cloud cover (0-1) (Couverture nuageuse total)	**clim_tcc_J1982D2010.mat**

2) Données de Prévisions à midi (assimilation 24h avant)

lsp : Large scale precipitation (m) (Précipitation à large échelle) **clim_lsp_J1982D2010.mat**

cp : Convective precipitation (m) (Précipitation convective) **clim_cp_J1982D2010.mat**

ssr : Surface solar radiation ((W/m^2)s) (Radiation solaire de surface) **clim_ssr_J1982D2010.mat**

3) **co_2** : molfrac ppm (parties par million) **clim_co2_J1982D2010.mat**

Excepté le CO_2, les lieux pour lesquels nous avons extrait les valeurs des variables sont dans l'ordre du nord au sud :

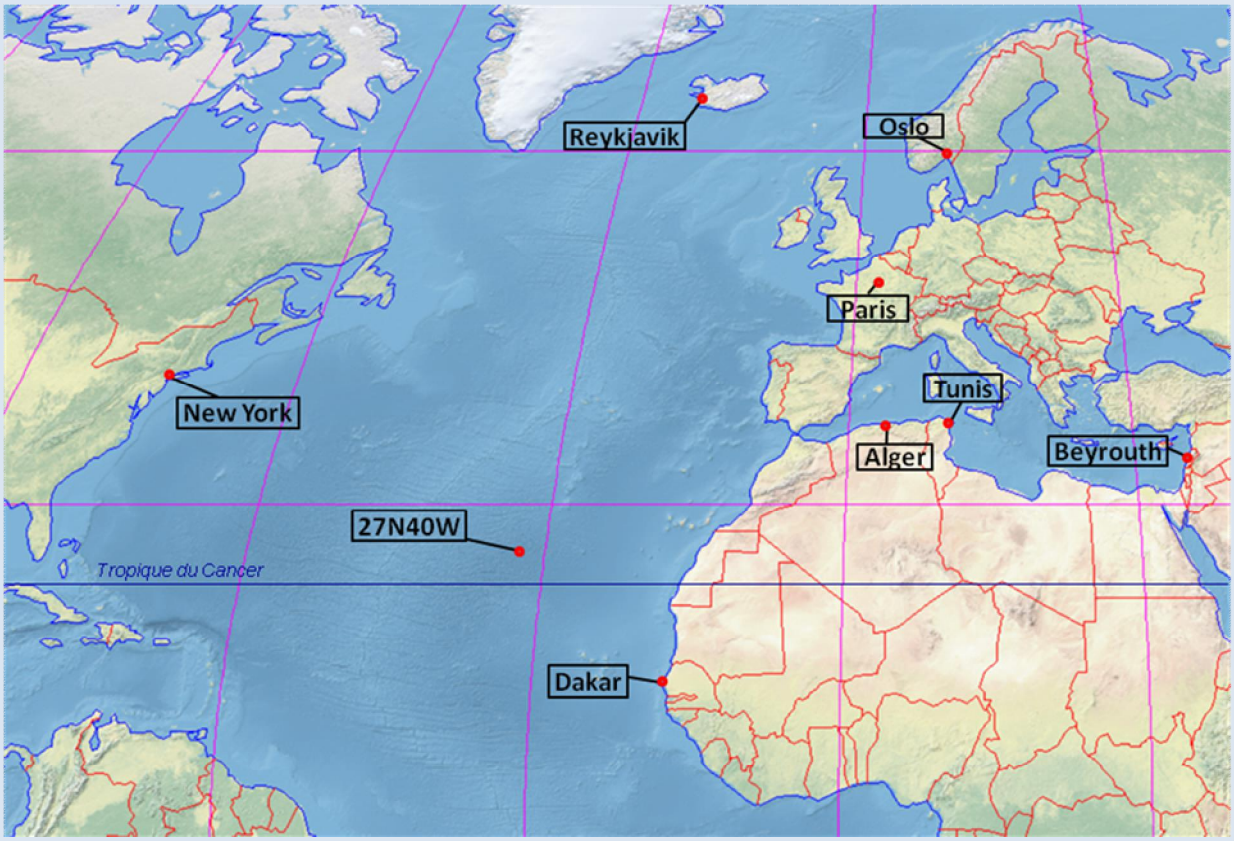
Reykjavik	64°08'07.14"N	21°53'42.63"O
Oslo	59°54'49.85"N	10°45'08.18"E
Paris	48°51'12.03"N	2°20'55.59"E
New York	40°42'51.67"N	74°00'21.50"O
Tunis	36°49'07.72"N	10°09'57.46"E
Alger	36°45'10.39"N	3°02'31.37"E
Beyrouth	33°53'19.06"N	35°29'43.72"E
Atlan27N40W	27°00'00.00"N	40°00'00.00"O
Dakar	14°39'46.09"N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

Contenu des fichiers :

- 1ère colonne : l'année
- 2ème colonne : le mois
- Colonnes 3 à 11 : valeur de la variable pour les 9 lieux dans l'ordre où on les a énumérés.

Pour le fichier de CO_2 on retrouve les mêmes deux 1ère colonnes et une 3ème colonne pour la valeur de concentration du CO_2.

A noter que tous les fichiers fournis, pour ce TP, sont en correspondance sur les deux premières colonnes, elles contiennent donc le même nombre de lignes (**N=348**).



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 :

$$X^tX\boldsymbol{u}_k = \lambda_k\boldsymbol{u}_k. \qquad \text{avec } \boldsymbol{u}_k \text{ le } k\text{ième vecteur colonne de } U.$$

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 :

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

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 :

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

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 « *ACPrappels* »

III - Eléments pour la réalisation du TP

- **xclimmens** : Charge les données puis produit des tracés par ville des climatologies mensuelles pour les variables t2, tcc, lsp, cp, ssr et co_2.

- **centred** : Normalisation par centrage réduction.

- **phinertie** : Calculs et cumuls des pourcentages d'inertie des 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 c'est le cas pour ce TP, il suffira de ne passer que les 5 premiers paramètres, et d'omettre les 2 derniers.

- **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).

- **moyan** : Calcul de moyennes annuelles.

- **acp** : ACP

Ces fonctions seront, selon les cas, fournies ou pas.
Toute fonction dont le nom n'est pas indiqué ci-dessus est, a priori, fournie (module python ou fichier triedacp.py).

IV – 1ère partie : Climatologie mensuelle et ACP de Reykjavik

1) Présentation des données : climatologies mensuelles par ville des variables t2, tcc, lsp, cp, ssr et CO_2.

Une visualisation des données de moyenne mensuelle, telles qu'elles sont enregistrées dans les fichiers ne seraient pas d'une grande aide pour leur compréhension. A la place, vous présenterez pour chaque ville, une climatologie mensuelle des variables (c'est-à-dire la moyenne de chaque mois) en valeur centrée et réduite.
Pour cela, après avoir chargé les données, vous devrez pour chaque ville :

- Normaliser les variables par centrage et réduction (fonction **centred**).
- Pour les variables normalisées, calculer les climatologies mensuelles. Cela consiste à faire la moyenne sur les 29 années mois par mois. Présenter les courbes de cette climatologie mensuelle sur un même repère. Chaque variable devra être repérable par une couleur ou un marqueur différent.

Au final, vous devriez obtenir 9 repères, chacune comportant 6 courbes constituées de 12 points (i.e. 1 point par mois). Vous pourrez soit faire 1 figure par ville (9 figures), soit une seule figure avec 9 subplot (fonction **xclimens**).

Vous devrez proposer un commentaire global des tracés obtenus, il vous servira par la suite à confirmer vos résultats.

2) ACPs pour la ville de Reykjavik des variables t2, tcc, lsp, cp, ssr et CO_2)

2.1) ACP « saisonnière »

Après avoir ouvert les fichiers de chacune des variables (si ce n'est déjà fait), vous devrez ranger en colonne les valeurs des variables de la ville de Reykjavik dans une matrice. Celle-ci devra donc comporter 6 colonnes correspondant au 6 variables et 348 lignes correspondant au 348 « individus-mois ». Disposant de moyennes mensuelles, cette ACP pourra être qualifiée d'analyse « ***saisonnière*** » (ce qui veut dire qu'elle sous entend nécessairement une étude du cycle saisonnier).

ŏ) Vous devez maintenant réaliser l'ACP avec la matrice des données que vous aurez préalablement centrées et réduites, à l'aide par exemple de la fonction **centred**. Le calcul de l'ACP peut dès lors être effectué avec la fonction **acp**.

ŏ) Pour faire état des résultats obtenus, vous devrez présenter :

- La figure des climatologies mensuelles (à reprendre du point 1).
- La figure des pourcentages d'inertie des valeurs propres (fonction à utiliser : **phinertie**).

Puis uniquement pour les 2 premières composantes principales :

- Le cercle des corrélations (utilisation de la fonction **corcer**).
- Le nuage des individus avec le mois en échelle de couleur et en label (fonction **acpnuage**). La taille du marqueur triangle utilisé par la fonction devra être proportionnée à la qualité de représentation (QLT). On choisira un facteur de taille adéquate qui rende lisible la figure. (La qualité pourra être déterminée à l'aide de la fonction **qltctr2** fournie.)

  Avec 348 individus, le nuage des individus risque d'être surchargé. On pourra sélectionner le nombre de points du nuage en ne retenant que les individus qui ont une qualité de représentation plus importante (supérieure à 0.5 par exemple). Cela pourra être réalisé à l'aide la méthode **where** associée aux **numpy.array**. Il conviendrait donc d'appeler **acpnuage** uniquement avec les points sélectionnés.

Pour chaque ville, l'interprétation devra être menée en considérant l'ensemble de ces figures toutes à la fois.

ŏ) Vous devrez présenter ensuite, un second nuage des individus en utilisant l'année en échelle de couleur. Ce choix d'échelle de couleur est-il pertinent ; si oui, pourquoi ?

2.2) ACP « interannuelle »

Cette ACP est dite « **interannuelle** » car elle devra être effectuée avec les données des moyennes annuelles. L'angle d'étude est cette fois celui d'une évolution globale sur la période considérée.

ŏ) Vous devrez donc commencer par le calcul des moyennes annuelles (pour chaque variable). Ce calcul devra être réalisé avant le centrage et la réduction des données. Vous devriez obtenir une matrice de données (centrées et réduites) avec 29 lignes d'individus correspondant aux années et 6 colonnes. (Pour le calcul des moyennes annuelles, vous pouvez éventuellement vous servir de la fonction **moyan**)

ŏ) A l'instar de l'ACP saisonnière, nous vous demandons (en utilisant les mêmes fonctions que pour l'étude saisonnière) :

- De faire une figure des moyennes annuelles centrées et réduites. Cette figure devra donc comporter 6 courbes de 29 points chacune.
- De faire l'ACP de ces moyennes annuelles.
- De présenter en suivant les mêmes indications que celles de l'étude saisonnière, les pourcentages d'inertie (une figure), le cercle des corrélations (une figure) et le nuage des individus (une figure). Pour ce dernier, la seule échelle de couleur à utiliser est l'année.

Commentez l'ensemble de ces résultats.

V – 2ème partie : ACP d'Alger et de Dakar

La 2ème partie de ce TP va consister à refaire, dans les mêmes conditions que pour la première partie, les ACP (saisonnière et interannuelle) faites pour la ville de Reykjavik. Cette fois-ci, les villes concernées sont d'abord, la ville d'Alger puis ensuite celle de Dakar.

Concernant la ville de Dakar, cependant, nous vous demandons de faire une étude complémentaire sur le plan factoriel des composantes principales 3 et 4, aussi bien dans le cas saisonnier qu'interannuel. Il s'agira donc, pour ce plan (3-4) et dans les deux cas, de produire (toujours à l'aide des mêmes codes) et de commenter :

- Un cercle des corrélations,
- Dans le cas de l'étude saisonnière : Présenter deux nuages des individus : l'un avec le mois en échelle de couleur, l'autre avec l'année. On pourra abaisser le seuil de sélection des individus en fonction de leurs qualités de représentation (à 0.25 par exemple), car sur ce plan, ces qualités pourraient être moins élevées.
- Dans le cas de l'étude interannuelle il n'y a qu'un nuage à présenter avec l'année en échelle de couleur :