

# MASTER : Traitement de l'Information et Exploitation des Données

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

## TPA03 : Données environnementales et effet de serre (Indicateurs bidimensionnels)

### I - Objectifs

L'objectif est d'acquérir la maîtrise de méthodes statistiques nécessaires à une première analyse de données. Il s'agit ici de données environnementales qui entrent en jeu dans l'étude de l'effet de serre.

Deux parties sont proposées :

- La 1ère s'intéresse plus spécifiquement aux données de température. On aura tout d'abord une approche statique dans le cadre de laquelle on sera amené à calculer des corrélations et on représentera les données par des nuages de points (diagramme de dispersion). Par la suite, on étudiera l'évolution des températures à l'aide de régressions linéaires à valider par des intervalles de confiance.
- Dans la seconde partie, nous verrons comment étudier les relations entre les variables de température et de CO<sub>2</sub>.

Nous vous donnons à titre d'exemple le document « FOCUS-PSDR3-Clim Chgt (4) -Vfinale.pdf ». Il s'agit d'une étude climatologique sur l'évolution du climat au Sud de la France sur la période 1950-2009. Ce document a été rédigé par le PSDR3 Midi-Pyrénées.

=====

*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. Ces figures doivent intégrer des éléments de compréhension nécessaires (légende, titre, bar de couleur, ...). Les résultats présentés devront être analysés et commentés.*

### 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 donnée est issue de la base ERA-Interim du centre européen ECMWF. Il s'agit d'une donnée modèle au sortir de l'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"E
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"E
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"E
Dakar .....	14°39'46.09"N	17°26'13.65"E

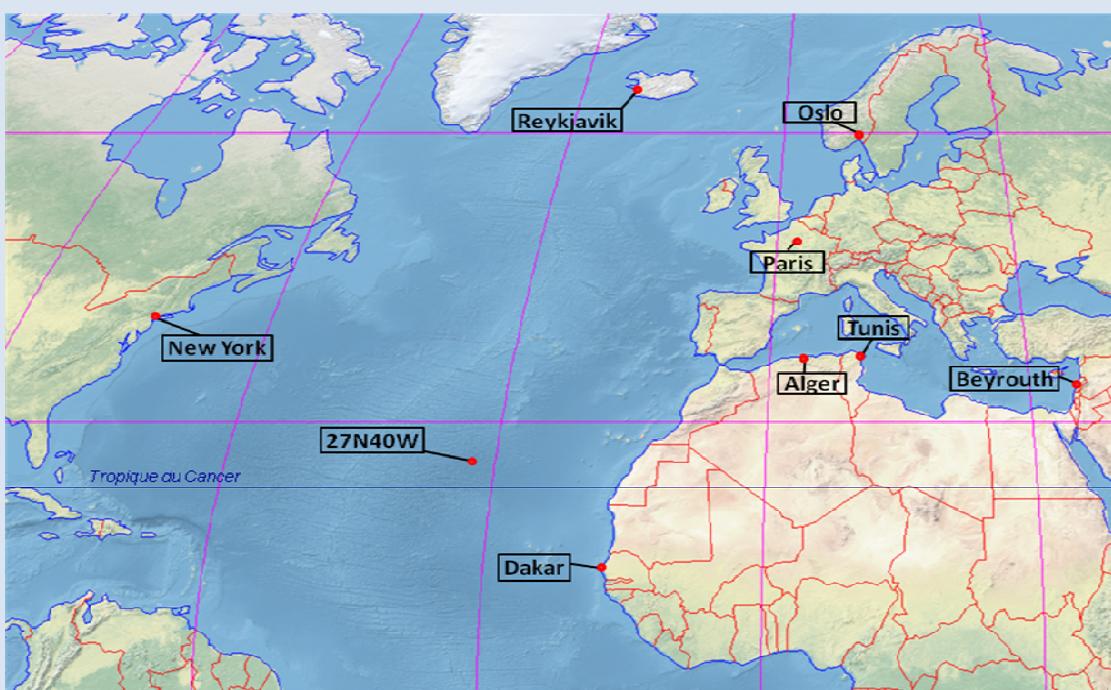
- Le  $\text{CO}_2$  en molfrac ppm (parties par million) dont les mesure de concentration ont été réalisées sur le mont Mauna Loa à Hawaii. Ces données qui proviennent de la NOAA.

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  $\text{CO}_2$ .

La 1<sup>ère</sup> colonne de ces fichiers correspond à l'année, la seconde au mois. Les deux fichiers sont 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 3<sup>ème</sup> colonne du fichier **clim\_co2\_J1982D2010.mat** contient la valeur de concentration du  $\text{CO}_2$ .



### III - Eléments pour la réalisation du TP

La plupart des fonctions nécessaires à la réalisation de ce TP sont des fonctions Matlab. Souvent nous vous donnerons des indications sur les fonctions à utiliser sans pour autant entrer dans le détail de la programmation qu'il vous incombera de prendre à votre charge. Nous avons par ailleurs jugé utile de fournir 3 fonctions ad hoc supplémentaires pour ne pas compliquer d'avantage le TP. Il s'agit des fonctions :

- **regrtrace** : Régression linéaire : Coefficients, et optionnellement traçage, de la droite de régression linéaire :  $y = b_0 + b_1 \cdot x$
- **regtest** : Intervalle de confiance (Test de Student) pour la pente d'une régression linéaire et la prévision, selon cette régression en un point  $x_0$  fixé.
- **dblyplot** : Plot avec 2 ordonnées ayant des unités différentes.

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

## Régression (rappels partiels)

On est en présence d'un échantillon de  $n$  points décrit par 2 variables  $x$  et  $y$ . On cherche les coefficients  $b_1$  et  $b_0$  d'une droite ( $y_c = b_1x + b_0$ ) qui décrit au mieux la tendance de l'échantillon. La régression linéaire, dite droite des moindres carrés, consiste à minimiser la somme des écarts verticaux ( $e_i$ ) entre les ordonnées des points de l'échantillon et leurs correspondants sur la droite recherchée. On trouve les coefficients  $b_1$  et  $b_0$  en annulant leurs dérivées partielles dans l'expression de cette somme.

Les formules qui permettent d'obtenir les valeurs sont :

$$b_1 = \frac{\sum_{i=1}^n ((x_i - \bar{x})(y_i - \bar{y}))}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

où  $x_i$  et  $y_i$  représentent les points qui constituent l'échantillon, et  $\bar{x}$  et  $\bar{y}$  les moyennes de ces points

Remarque :  $b_1 = \text{cov}(X, Y) / \text{var}(X)$

### Intervalle de confiance du coefficient directeur (pente) $b_1$

Sous l'hypothèse que les erreurs  $e_i$  suivent une loi normale de moyenne nulle, et en notant  $\alpha$  le seuil de risque, il est possible d'obtenir un intervalle de confiance  $(1-\alpha)$  du coefficient de la pente  $b_1$  en déterminant  $t_{\alpha, n-2}$  avec une table de Student à  $(n-2)$  degrés de liberté. L'intervalle s'établit alors ainsi :

$$b_1 \pm t_{\alpha, n-2} e.t(b_1)$$

$$\text{avec } e.t(b_1) = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad \text{ou } s = \sqrt{\frac{\sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2}{n-2}} \quad (\text{une estimation de l'écart type } \sigma)$$

### Intervalle de prévisibilité

Avec les mêmes hypothèses, on peut également calculer pour chaque valeur  $x^0$  un intervalle de confiance au niveau  $\alpha$  sur la prévision de  $y^0$  tel que :

$$y_c^0 - t_{\alpha, n-2} \sqrt{\text{VAR}(y_c^0 - y^0)} < y^0 < y_c^0 + t_{\alpha, n-2} \sqrt{\text{VAR}(y_c^0 - y^0)}$$

$$\text{avec } \text{VAR}(y_c^0 - y^0) = s^2 \left[ 1 + \frac{1}{n} + \frac{(x^0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

## IV - 1<sup>ère</sup> partie : Etude statique des températures

### 1°) Etude statique des températures

Après avoir chargé le fichier des températures (**clim\_t2C\_J1982D2010.mat**), les données (y compris les années et les mois) récupérées sont identifiées par la variable **clim\_t2**. Nous vous demandons alors :

1.1°) De présenter, sur une même figure, les courbes des températures par ville (soit 9 courbes), avec une couleur différente. Une légende devra permettre d'associer la ville et la couleur. L'abscisse devra indiquer les années.

-> 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 1<sup>er</sup> mois de l'année (janvier) de la façon suivante (par exemple ici avec **clim\_t2**) :

**llab = find(clim\_t2(:,2)==1);** % Indices des mois de janvier

**Xlab = clim\_t2(llab,1);** % en label, l'année de ces indices

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

**set(gca,'XTick',llab, '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°) De calculer l'ensemble des coefficients de corrélation linéaires<sup>1</sup> des températures entre les villes prises deux à deux. Faire ensuite une représentation graphique de la matrice de ces coefficients en faisant apparaître l'échelle de couleur. Fonction à utiliser : **corrcoef**, **imagesc**, **colorbar**.

1.3°) De faire une figures des diagrammes de dispersion des températures par ville deux à deux. Chaque diagramme devra faire l'objet d'un **subplot** (vous devriez en obtenir 36 puisqu'il y a 9 lieux différents). Les couleurs des points des nuages devront être associés aux mois. Nous vous suggérons de créer une table de couleurs à partir de la map « **hot** ». Pour chaque diagramme, la mention des villes devra apparaître ainsi que le coefficient de corrélation (déjà calculé lors de la question précédente).

-> Indications (très partielles) pour la réalisation : vous devriez avoir 3 boucles imbriquées à programmer : pour i = 1 à nombre\_de\_villes  
 pour j = i+1 à nombre\_de\_villes  
 pour mois = 1 à 12  
 plotter les points de ce mois dans la couleur du mois  
 finpour, finpour, finpour

Vous aurez aussi besoin de gérer un indice de **subplot**. Pour une appréciation équilibrée des diagrammes nous vous suggérons d'utiliser **axis equal**.

---

<sup>1</sup> Pour information, le coefficient de corrélation linéaire  $r_{xy}$  entre 2 variables  $x$  et  $y$  est donné par :  $r_{xy} = S_{xy}/(S_x S_y)$  où  $S_{xy}$  est la covariance entre  $x$  et  $y$  :  $S_{xy} = 1/n \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$  et  $S_x$  et  $S_y$  sont les écarts types de  $x$  et  $y$ .  $r_{xy}$  indique le degré de dépendance **linéaire** entre les 2 variables. Si les 2 variables sont centrées réduites, il s'interprète géométriquement comme le cosinus de l'angle  $\theta_{xy}$  entre les 2 variables :  $\text{Cos}(\theta_{xy}) = \frac{\langle x; y \rangle}{\|x\| \|y\|} = \frac{S_{xy}}{S_x S_y}$ . On a l'égalité suivante  $b_1 = r_x * S_y / S_x$ .

## 2) Etude de l'évolution temporelle des températures

2.1) Nous nous intéressons maintenant à l'évolution de la température sur la période des données disponibles. Pour simplifier, plutôt que de travailler avec chaque ville individuellement, nous vous demandons d'établir la variable des températures moyennes des villes pour chaque mois de la période. Nous appellerons **t2moy** cette variable (qui doit donc être de dimension 348x1).

Par la suite, vous devrez :

- Présenter une figure avec :

- Le tracé de la courbe de **t2moy** en fonction du temps (i.e.: `[1:length(t2moy)]'`) (fonction **plot**)
- Le tracé de la droite de régression de **t2moy** en fonction du temps. Vous pouvez utiliser la fonction **regtrace** en lui passant :
  - le vecteur temps qui correspond aux points d'abscisses
  - **t2moy**
  - une chaîne de caractère qui indique le marqueur et la couleur à utiliser pour tracer droite (comme avec la fonction **plot**).

Les coefficients de la droite (de régression), **b0** (l'ordonnée à l'origine) et **b1** (la pente), rendus par **regtrace**, devront être indiqués sur la figure (ou dans le rapport).

- Déterminer l'augmentation de la température sur la période selon la régression obtenue (Pente \* nombre de pas, le nombre de pas étant égal au nombre de points - 1). Faites ensuite une prévision de température à 100 ans (pour l'année 2110) dans l'hypothèse où le changement climatique poursuivrait la même évolution (selon la régression obtenue). Indiquez à la fois l'augmentation et la valeur.

- Donner l'intervalle de confiance sur la pente.

2.2) On souhaite synthétiser davantage l'information, on va supprimer les effets de la variation saisonnière. Pour cela, vous devez créer, à partir de **t2moy**, un nouveau vecteur : **t2moyan** qui contient les moyennes annuelles de **t2moy**, soit 29 valeurs.

2.2.a) Comme précédemment, nous vous demandons de présenter la courbe de **t2moyan** ainsi que la droite de régression obtenue en indiquant les valeurs des coefficients de la droite de régression (**b0** et **b1**).

Si le changement climatique poursuit la même évolution (selon la régression obtenue) quelle pourrait être l'augmentation et la valeur des températures en 2110 (soit 100 ans après).

2.2.b) On veut maintenant restreindre notre étude aux 10 dernières années. On vous demande donc de refaire le même travail en utilisant les moyennes annuelles de **t2moyan** pour les années 2001 à 2010.

Commenter l'ensemble des résultats.

2.3) Intervalle de confiance (Test de Student) sur les coefficients directeurs (pentes) et la prévision d'une régression. Quelle confiance peut-on accorder aux régressions obtenues ? Pour le savoir, nous vous demandons de déterminer, avec un risque d'erreur inférieur à 5% et à l'aide de fonction **regtest** :

2.3.a) Pour les cas 2.2.a et 2.2.b : un intervalle de confiance sur les pentes.

2.3.b) Pour le cas 2.2.a uniquement vous encadrerez l'accroissement pour la période des 100 années considérées en vous servant des bornes de l'intervalle de confiance sur la pente.

2.3.c) Pour les cas 2.1 et 2.2.a (qui correspondent à la période de 29 années), un intervalle de confiance sur la prévision en 2110.

Comparer et conclure.

## V - 2<sup>e</sup> partie : Température et CO<sub>2</sub>

Nous allons maintenant utiliser la variable des données du CO<sub>2</sub> présente dans le fichier **clim\_co2\_J1982D2010.mat** sous le nom de **clim\_co2**.

1°) CO<sub>2</sub> brut : On vous demande de présenter la courbe de cette variable ainsi que la droite de régression linéaire obtenue en prenant le temps comme variable explicative : on précisera la valeur de la pente (b1).

2°) CO<sub>2</sub> corrigé : On souhaite pouvoir comparer la température et le CO<sub>2</sub>. On constate que ces 2 variables présentent des variations saisonnières, mais avec une différence de pente importante quant à leur tendance globale.

On vous demande de créer une nouvelle variable qui corrige le CO<sub>2</sub> de sa tendance globale. On appellera « CO<sub>2</sub> corrigé » cette nouvelle variable que l'on notera **CO<sub>2cor</sub>**. Pour la déterminer, il suffit de calculer **CO<sub>2cor</sub> = CO<sub>2</sub> - b1\*tclim**, où **tclim=[1:taille des données]** correspond aux pas de temps. Vous devrez présenter la courbe du CO<sub>2</sub> corrigé.

3°) CO<sub>2</sub> corrigé décalé : On constate un décalage de phase entre les variations saisonnières de la température et celles du CO<sub>2</sub>. Pour déterminer ce décalage, on vous demande de calculer plusieurs coefficients de corrélation entre ces 2 séries chronologiques en décalant **CO<sub>2cor</sub>** de **m** pas par rapport **t2moy** (Un pas correspond ici à un mois). Vous devrez faire varier **m** de 0 à 59 (soit de 0 mois de décalage à 59 mois de décalage).

3.1°) On calculera donc les corrélations entre les deux séries temporelles **t2moy(i)** et **CO<sub>2cor</sub>(i+m)** avec **i** variant de 1 à **N-m** et ceci pour chaque valeur de **m**. On peut représenter cela schématiquement en représentant les indices de temps qui se correspondent pour chaque valeur de **m** :

$$\begin{aligned} t2moy &: \{1, 2, 3, 4, \dots, N-m\} \\ m=0: CO_2cor_0 &: \{1, 2, 3, 4, \dots, N\} \\ m=1: CO_2cor_1 &: \{2, 3, 4, 5, \dots, N\} \\ m=2: CO_2cor_2 &: \{3, 4, 5, 6, \dots, N\} \end{aligned}$$

Il y a donc 60 coefficients à calculer, avec la fonction **corrcoef**, que l'on vous demande de représenter graphiquement.

Le décalage des cycles des températures et du CO<sub>2</sub> s'explique en partie par le fait, que l'augmentation des températures, à partir du printemps, va permettre à la végétation de se développer en consommant du CO<sub>2</sub>. On constatera visuellement que le cycle des corrélations est annuel et l'on cherchera **m\*** le nombre de mois de décalage associé au coefficient qui présente la plus forte anti-corrélation pour un cycle annuel : il s'agit du mois qui correspond à la plus forte anti-corrélation obtenue.

3.2°) Pour constater visuellement ce décalage, on vous demande de faire les moyennes par mois de **t2moy** et **CO<sub>2cor</sub>** (soit 12 valeurs chacune) et de les représenter graphiquement.

4°) Afin de faire ressortir l'importance du codage des données, nous vous demandons de faire 3 diagrammes de dispersion entre :

- t2moy et le CO<sub>2</sub> brut
- t2moy et le CO<sub>2</sub> corrigé (CO<sub>2</sub>cor)
- t2moy et le CO<sub>2</sub> corrigé décalé de m\* mois (CO<sub>2</sub>cor<sub>m\*</sub>)

Pour chacun de ces diagrammes, vous devrez faire apparaître la droite de régression (fonction **regtrace**) et indiquer le coefficient de corrélation (fonction **corrcoef**). Commenter les diagrammes.

5°) Pour compléter cette étude, vous devrez comparer les évolutions de la température et du CO<sub>2</sub>. Pour cette comparaison, il faudra s'affranchir de la variabilité saisonnière. Cela nécessite donc de calculer les moyennes annuelles des températures et du CO<sub>2</sub>.

Nous vous demandons de présenter sur une même figure, à l'aide de la fonction **dblyplot**, les courbes des valeurs obtenues. Calculer le coefficient de corrélation. Qu'en pensez-vous ?