TP3 – Corrélations et régressions linéaires

I Objectif général

L'objectif de ce TP est d'utiliser les notions de corrélation et de régression linéaire pour l'analyse géophysique d'un jeu de données d'observation.

II Théorie et formalisme mathématique

La régression permet d'ajuster une fonction à une série de données. Cette méthode sera appliquée ici au cas le plus simple où il faut trouver la droite représentative d'un phénomène linéaire à partir de données expérimentales. L'algorithme de régression linéaire est programmé dans la plupart des calculatrices modernes, il n'est donc généralement pas nécessaire de recourir aux équations explicites. Les voici néanmoins pour informations.

La fonction , ici une droite, à ajuster pour un ensemble N de points discrets (x_i; y_i) a la forme

$$y = ax+b. \qquad \textbf{Eq (1)}$$

Le problème est de trouver les valeurs des coefficients *a* et *b* de la droite (Eq 1) qui minimise la somme des carrés des écarts y*i* - (ax_i + b), où y_i sont les données mesurées et $ax_i + b$ la valeur théorique (Eq 1) de cette mesure. Concrètement, la fonction à minimiser est appelée χ^2 :

$$\chi^2(a,b) = \sum_{i=1}^{N} \left(y_i - ax_i - b\right)^2. \qquad \textbf{Eq (2)}$$

Les valeurs optimales des coefficients *a* et *b* qui minimisent la fonction χ^2 sont

$$a = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\Delta}, \qquad \textbf{Eq (3)}$$

$$b = \frac{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_i}{\Delta}, \qquad \textbf{Eq (4)}$$

avec

$$\Delta = N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 . \qquad \textbf{Eq (5)}$$

Ces équations permettent de trouver les valeurs optimales des coefficients *a* et *b*.

Le calcul du coefficient de corrélation est une méthode souvent utilisée pour évaluer si les données sont bien représentées par la droite que l'on a déterminée (Eq 1). Le coefficient de corrélation est donné par l'équation

$$r = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}}. \qquad \textbf{Eq (6)}$$

La valeur de r est comprise entre -1 et 1. Elle vaut exactement 1 si les points sont parfaitement alignés sur une droite de pente positive, ou -1 si les points sont sur une droite de pente négative. r = 0 indique que les variable x et y ne sont pas corrélées.

La technique de régression linéaire peut être utilisée pour étudier la corrélation entre deux paramètres géophysiques, mais également pour mettre en évidence une tendance à long terme d'un paramètre géophysique en fonction du temps.

Les tendances à long terme s'expriment généralement en unité de la variable par décade (10 ans), ou en % par décade. Afin d'améliorer la précision sur le calcul de tendance des variables qui présentent un cycle annuel marqué, on peut calculer la régression linéaire sur les données corrigées du cycle annuel (appelées résidus), obtenues par exemple en soustrayant les valeurs moyennes mensuelles aux données.

III Données utilisées

Dans ce TP on utilisera la matrice Station_PDD_horaire_1995_2017.mat déjà utilisée dans le premier TP. Pour rappel, les variables contenues dans le fichier sont les suivantes :

- an, mois, jour, heure : dates de la mesure
- temps_fractionne : partie entière : année et partie décimale : fraction de l'année
- O3 : ozone en ppbv (nombre de molécules d'ozone multiplié par 10^9 et divisé par le nombre de molécules d'air)

- CO : monoxyde de carbone en ppbv (nombre de molécules de monoxyde de carbone multiplié par 10^9 et divisé par le nombre de molécules d'air)
- CO2 : dioxyde de carbone en ppmv (nombre de molécules dioxyde de carbone multiplié par 10^6 et divisé par le nombre de molécules d'air)
- Temp : température de l'air en °C
- Press : pression de l'air en hPa.

IV Calculs et analyse des résultats

1) Vérifier que les séries de paramètres (O3, Temp, CO et CO2) ne comportent pas de point erroné, les supprimer s'il y en a, et sauvegarder les séries corrigées dans un nouveau fichier .mat ou .txt.
2) À partir de ce nouveau fichier, tracer les paramètres (O3, Temp, CO et CO2) les uns en fonctions des autres soit 6 figures. On appelle ces types de graphique des graphes de dispersion (scatterplot en anglais). Discuter qualitativement ces résultats.
3) Calculer les coefficients de la droite de régression et le coefficient de corrélation dans le cas du CO2 en fonction de la température, d'abord en programmant les équations (3), (4) et (6), puis en utilisant les fonctions corrcoef et polyfit. Tracer le graphe de dispersion du CO2 vs température en y superposant la droite de régression. Analyser vos résultats.
4) Tracer les paramètres en fonction du temps et calculer leur tendance à long terme. Analyser vos résultats.
5) Désaisonnaliser les données de CO2 et de température en retirant aux données de CO2 les moyennes pour chaque mois de l'année. Attention : prendre *toutes* les données d'un même mois, pour toutes les années. Calculer les coefficients de régression et de corrélation et tracer les tendances à long terme sur les résidus. Analyser vos résultats.