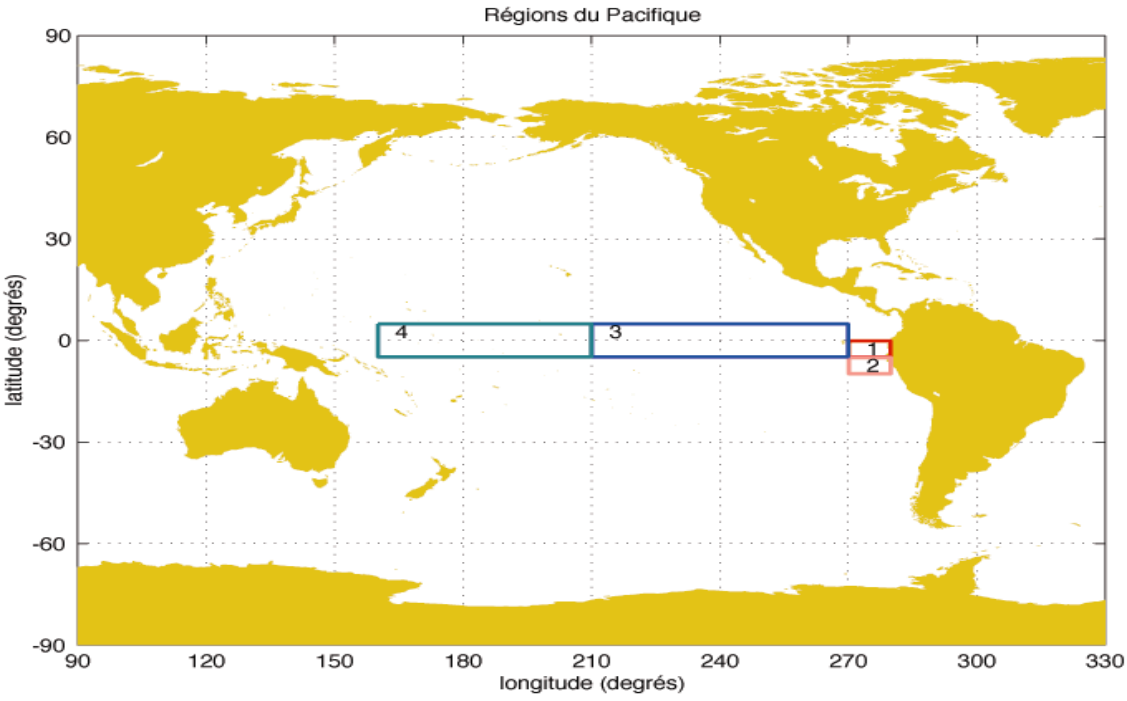
TPC04 : Classification : Application des cartes topologiques à l'étude des évènements El Niño

I – Description du phénomène El Niño

El Niño est un phénomène météorologique qui se présente périodiquement dans la planète dans l'océan Pacifique équatorial et qui affecte le climat global. Différents signaux révèlent la présence de ce phénomène tel qu'une forte variation de la température à la surface de la mer. Le vent est un des facteurs physiques qui peuvent causer ces variations de température. La force du vent crée une tension en surface qui se traduit en un changement de température.



REGION	LONGITUDE	LATITUDE
NINO1	90°W - 80°O	0° - 5°S
NINO2	90°W - 80°O	5°S - 10°S
NINO3	150°W - 90°O	5°N - 5°S
NINO4	160°W - 150°O	5°N - 5°S

Fig. 1 : Régions d'étude dans le Pacifique. Les coordonnées sont données en degrés et en Longitude Est ou Ouest et Latitude Nord ou Sud.

Pour l'étude de ce phénomène, quatre zones géographiques ont été définies par les physiciens dans la région équatorial du pacifique telles que représentées par 4 boites sur la figure, respectivement appelées NINO1, NINO2, NINO3 et NINO4.

...

Une des signatures des évènements El Niño est les anomalies de température de surface de la mer (écart à la moyenne) (ou SST en anglais pour Sea Surface Temperature) près de la côte d'Amérique du Sud (NINO1, NINO2). Il peut exister d'un évènement à l'autre des déphasages différents entre NINO1+2 et NINO3.

Avant de commencer ce TP, nous vous encourageons à consulter un document de cours qui donne une description plus complète du phénomène physique de l'El Niño et qui est accessible par le lien suivant : http://www.ifremer.fr/lpo/cours/elnino/index.html.

II - Les Objectifs

1ère Partie : Les 4 SST permettent de caractériser les évènements « El Niño ». Après une étude descriptive des données on devra le vérifier à l'aide d'une carte topologique (CT) qui prendra en entrée ces 4 indices de température. On se servira de certaines périodes extrêmes connues pour être représentatives de l'événement « El Niño » afin d'appréhender l'organisation de la carte.
Ce travail devra être complété par une classification ascendante hiérarchique (CAH) qui permettra de confronter le critère de la Norm(SST1)>1 qui se propose d'identifier les évènements El Niño.

2ème Partie : On exploite les résultats de la 1ère partie pour explorer, dans cette deuxième partie, la possibilité de prévoir un El Niño à partir d'une série chronologique de SST1 à construire.

==========

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. Les résultats présentés devront être analysés et commentés.

III - Les Données

Le fichier ascii de données **'el_nino.mat'** contient 13 colonnes :

date	SST1	SST2	SST3	SST4	Tx1	Tx2	Tx3	Tx4	Ty1	Ty2	Ty3	Ty4
6101	-9999.00	-9999.00	-9999.00	-9999.00	-5.37	-1.09	2.64	9.62	-2.46	12.44	0.98	63.69
6102	-9999.00	-9999.00	-9999.00	-9999.00	-6.58	-1.63	3.84	-16.44	-8.30	-6.53	4.12	1.21
6103	-9999.00	-9999.00	-9999.00	-9999.00	16.25	14.76	16.26	21.68	-13.31	-11.81	-9.29	-5.09
6104	-9999.00	-9999.00	-9999.00	-9999.00	-17.27	-3.42	-5.58	2.53	7.49	-0.06	3.41	-7.80

Extrait des données

La première colonne est la date de la mesure donnée sur 4 chiffres, au format AAMM (année et mois).

Les 4 colonnes SSTx suivantes correspondent à des indices de température moyenne de surface de la mer dans les 4 régions du Pacifique (Il s'agit en réalité des anomalies inter-annuelles de la température. C'est la série chronologique des températures moyennes auxquelles on a retiré la composante saisonnière.)

Les 8 colonnes suivantes sont les moyennes des anomalies inter-annuelles de tension du vent dans les mêmes 4 régions. Les 4 premières la tension zonale (dans le sens des Longitudes) et les 4 autres la tension méridienne (dans le sens des Latitudes). Tx et Ty sont donc des pseudo-stress, un paramètre proche de la vitesse du vent, exprimées en $(m/s)^2$.

Les données vont du mois de janvier de l'année 1961 à avril de 1997. Avant l'année 1970 nous n'avons pas de valeur de température moyenne, la valeur est donc -9999.00.

IV - Eléments pour le déroulement du TP

Pour réaliser ce TP, nous utiliserons la librairie Matlab « SOM_Toolbox version 2.0beta », qui implémente le logiciel d'apprentissage par cartes topologiques, développé par l'équipe de Kohonen. C'est un produit du domaine public que vous pouvez par ailleurs trouver sur le site Web de l'Université de Helsinky à l'adresse http://www.cis.hut.fi/research/software.

Voici les principales fonctions de la SOM_Toolbox dont vous aurez ou pourriez avoir besoin :

- **som_data_struct** : Création d'une structure de donnée qui, en plus des données elles-mêmes, comporte, entre autre, un champ de labellisation. Passer par ce type de structure, ce qui n'est pas toujours nécessaire, permet l'utilisation de certaines fonctions comme, par exemple, celles qui se rapportent à la labellisation.
- **som_map_struct** : Crée une structure de carte topologique. On y trouve entre autre sa taille, les vecteurs référents, sa topologie (taille, connexion, forme), des informations sur les paramètres d'apprentissage par étape (algorithme, température initiale et finale, nombre d'itération, ...). On distingue éventuellement une étape d'initialisation, et 2 étapes d'apprentissage la seconde étant un affinement de la 1ère utilise normalement une température plus faible.
- **som_lininit** : Crée (ou modifie) une CT avec une initialisation linéaire des vecteurs référents (il y a aussi **som_randinit** qui le fait de façon aléatoire)
- **som_batchtrain** : Algorithme batch d'apprentissage de la CT, et **som_seqtrain** pour la version séquentielle.
- **som_quality** : Calcule et renvoie l'erreur de quantification (distance moyenne des données à leurs référents) et l'erreur topographique (proportion des données dont les 2 référents les plus proches ne sont pas adjacent sur la carte).
- **som_grid** : Visualisation de la grille de la carte topologique qui peut aussi être projeté dans l'espace de coordonnées des référents.
- **som_show** : Affichage de la matrice U et de la CT variable par variable. (py34: ctk.showmap)
- **som_show_add** : Permet l'ajout d'information additionnelle à associer aux neurones de la CT comme des labels par exemple.
- **linkage** (ou **som_linkage**, **som_cllinkage**) : Création de regroupements (clusters) hiérarchique de données. (py34 : from scipy.cluster.hierarchy import dendrogram, linkage, fcluster)
- **dendrogram** (ou **som_dendrogram**) : Affichage de l'arbre hiérarchique des clusters de données (cette visualisation est appelée un dendrogramme).
- **cluster** : Affectation d'un indice de classe (dont le nombre est passé en paramètre) aux données de sortie de linkage. (py34 : fcluster)
- **som_label** : Permet de mettre à jour les labels des données d'une structure de données ou des référents d'une structure de CT.
- **som_autolabel** : Mise à jour automatique des labels selon un mode spécifié (fréquence ou vote majoritaire par exemple)

...

- **som_label2num :** Donne un codage numérique d'un ensemble de labels (par classe de label dans l'ordre ou ils se présentent)
- **som_bmus** : Permet d'associer les données aux référents.
- **som_cplane :** Visualisation d'une carte topologique (2D) à laquelle on peut associer différentes couleurs aux référents. (copié de TPC02).

Cette liste n'est pas nécessairement exhaustive, et vous pourriez être amené à repérer d'autres fonctions par vous-même.

Autres fonctions dédiées pour ce TP

ctk_bar (en alternative à **som_barplane**) pour la visualisation des référents de la carte sous forme de barre.
serielabset : Mise en forme des séries et labels associés dans des structures des données pour l'apprentissage et le test
drasticperf2 : Calcul de performance en classification pour une CT et une classe c. Elle est calculée pour l'ensemble d'apprentissage avec lequel on effectue la labellisation (vote majoritaire) et pour l'ensemble de test. La formule de calcul utilisée est (exprimée pour la classe « niño ») :

$$Performance = 1 - \frac{nombre\ de\ cas\ prévus\ niño\ à\ tort + nombre\ de\ cas\ niño\ non\ prévus}{nombre\ total\ de\ cas\ niño\ réels}$$

ctk_confus : Etablit la matrice de confusion pour une carte topologique et des données, les deux devant être déjà labellisées.
ctk_profils (en alternative à **som_plotplane**) : Trace sur la CT les courbes des référents et/ou les données qui seront associées à chaque référent.

Nous rappelons que vous pouvez vous aider des commandes **help** ou **type** pour obtenir des informations sur les fonctions ainsi que du document Contents de la SOM Toolbox pour celles qui en relèvent.

V - 1ère Partie : Etude des données du phénomène El Niño

Nous proposons, dans cette 1ère partie, une étude du phénomène El Niño uniquement à partir des moyennes d'anomalies mensuelles des températures des 4 zones géographiques préalablement identifiées SST1, SST2, SST3, SST4 (les données de tension du vent ne seront pas utilisées). On abordera cette thématique d'abord par une exploration statistique avant de la poursuivre avec l'outil de CT.

Dans le passé, plusieurs périodes ont connus des évènements extrêmes correspondant au phénomène El Niño. Nous utiliserons plus particulièrement les données des années 1972 et 1983 comme éléments de référence de ce phénomène.

On rappelle que les figures et plus généralement les résultats présentés doivent être commentés.

Travail à faire

1) Statistique descriptives :
- Présenter tout d'abord un graphe des données et donner leurs moyennes et écarts types ou tout autre indicateur que vous jugerez intéressant.
- Faire une figure d'histogramme pour chacune des 4 variables.
- Etablir également les diagrammes de dispersion et donner les coefficients de corrélation linéaire des variables prisent 2 à 2. On fera ressortir, par des couleurs différentes, les points de données des années 72 et 83.

2) Etude par carte topologique :
2.1) Quantification vectorielle
Réaliser une quantification vectorielle par CT (toujours avec les 4 SST). Une dimension de carte 7x7 peut convenir ; étant donné le peu de données disponibles, il ne devrait pas être nécessaire d'aller au-delà de 10x10.
Faites état des résultats la quantification obtenue (déploiement de la carte dans l'espace des données, erreur de quantification et topographique). Indiquez dans votre rapport, les différents paramètres de la configuration d'apprentissage que vous aurez utilisés (taille de la carte, température initiale et finale, nombre d'itérations).
Présenter chacune des variables SSTx sur la carte avec, dans la mesure du possible, une échelle de couleur commune.
Faire apparaitre sur la carte, les données des années 72 et 83 en les labellisant sous la forme « aamm » (où aa représente l'année et mm le mois). Que constatez-vous ?

...

2.2) Classification ascendante hiérarchique (CAH) des neurones de la carte

Compléter l'étude en réalisant une CAH des référents en 3 classes. Indiquer la mesure de distance et le critère d'agrégation que vous aurez utilisés.

On labellisera les données selon 2 classes (à priori) :

- la classe « El Niño » : ce sont les évènements des années 72 et 83.
- la classe « NON El Niño » : ce sont tous les autres

Reporter sur la CT, le résultat de la classification par vote majoritaire selon cette labellisation. On pourra indiquer ~~la fréquence~~ le nombre des occurrences de chaque classe pour chaque neurone Que peut-on dire des évènements des années 72 et 83 par rapport à cette classification.

Fonctions à utiliser (entre autre) : pour matlab : som_label, som_autolabel, som_cplane, som_grid,…, pour python34 : ctk.reflabfreq, ctk.cblabvmaj, ctk.cblabfreq, tls.concstrlist, ctk.showcarte.

Pour vous aider dans l'interprétation de la classification, vous pouvez présenter un diagramme de barres par référents (fonction **ctk_bar** ou **som_barplane**) (py34 : ctk.showbarcell).

Certains géophysiciens ont proposé de discriminer les évènements El Niño selon le critère :

$Norm(SST1) > 1$, selon la formule $Norm(x) = \dfrac{x - m_x}{\sigma_x}$ où x représente la variable, m_x et σ_x en sont respectivement la moyenne et l'écart type. Si pour un x donné ce critère est rempli, on considère que x traduit un El Niño,

Vous devez maintenant comparer cette hypothèse à la précédente en labellisant les données sur ce critère. On représentera cette labellisation année/mois (comme précédemment) sur la CT uniquement pour les évènements « El Niño » selon ce nouveau critère. Que peut-on en déduire ?

Fonctions à utiliser (entre autre) : pour matlab : som_show, som_label, som_show_add ; pour python34 : ctk.showmap.

VI - 2ème Partie : Utilisation de séries chronologiques pour l'anticipation du phénomène El Niño

La 1ère partie nous aura permis de mettre en évidence la proéminence de SST1 comme marqueur du phénomène El Niño. Dans cette seconde partie on va construire des séries chronologiques sur cette variable pour tenter une anticipation du phénomène de l'El Niño. On rappelle le principe de construction d'une série chronologique :

soit une variable aléatoire $X_t : x_1, x_2, x_3, x_4, x_5, x_6, \ldots, x_m$ et une taille de fenêtre n ($n>0$),

une série chronologique de taille n est le n-uplet : $(x_{t-n+1}, \ldots, x_{t-1}, x_t)$.

Avec la série X_t ci-dessus, et en prenant par exemple $n=3$, on obtient l'ensemble de séries chronologiques suivant :

$$\{ (x_1, x_2, x_3) , (x_2, x_3, x_4), (x_3, x_4, x_5), (x_4, x_5, x_6), \ldots, (x_{m-2}, x_{m-1}, x_m) \}$$

La base des séries chronologiques ainsi formée devra permettre de déterminer la classe « El niño » ou « NON El niño » de l'évènement au temps **t+1**. Une partie de la base ($\frac{3}{4}$) devra être utilisée pour l'apprentissage et le quart restant pour le test.

Travail à faire

1) Détermination d'une taille de fenêtre n ($0<n\leq10$) optimale :

- Utilisation de la fonction **serielabset.m** qui construit (pour une variable x) les ensembles de séries chronologiques ($\{ (x_{t-n+1}, \ldots, x_{t-1}, x_t) \}$) pour l'apprentissage et le test (à raison de $\frac{3}{4}$, $\frac{1}{4}$ et de façon aléatoire). Cette fonction retourne également la labellisation en 2 classes des séries selon le critère $Norm(x_{t+1})>1$ défini dans la 1ère partie.

- Apprentissage par CT des séries chronologiques qui composent l'ensemble d'apprentissage. On calculera les performances sur les ensembles d'apprentissage et de test (fonction à utiliser : **drasticperf2**).

Produire une figure qui montre ces performances selon la taille de la fenêtre.

2) Etude du cas optimal (meilleure performance sur l'ensemble de test).

- Pour le cas optimal présenter le résultat de la classification sur la CT selon le critère $Norm(SST1_{t+1})>1$ ainsi que les ~~fréquences~~ nombre de données d'apprentissage par classe.
Fonctions à utiliser (entre autre) : pour matlab : som_label, som_autolabel, som_label2num, som_cplane, som_grid ; pour python34 : ctk.reflabfreq, ctk.cblabvmaj, ctk.cblabfreq, tls.concstrlist, ctk.label2ind, ctk.showcarte.

- Donner les matrices de confusion en apprentissage et en test. On vérifiera que les deux ensembles ont une proportion à peu près équivalente de phénomènes El Niño et NON El Niño. Si ce n'était pas le cas, il faudra reprendre l'expérience jusqu'à l'obtenir.

- Compléter l'étude de la carte et de sa classification à l'aide de la fonction **ctk_profils** (et/ou som_plotplane (py34 : ctk.showprofils)). Celle-ci réalise une figure qui représente les profils des référents et éventuellement les données de l'ensemble d'apprentissage que ces derniers ont captées. Il conviendra de présenter 2 figures, l'une avec des échelles d'axes indépendantes par référent, l'autre avec des échelles d'axes communes à tous les référents.