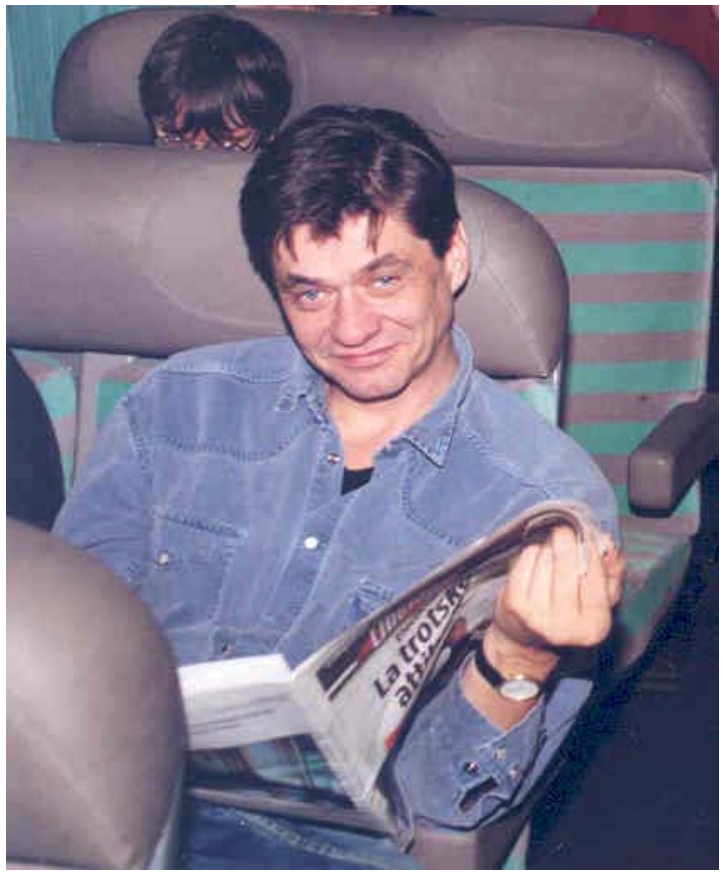
INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Classification de séries temporelles
Applications à la prévision et la désaisonnalisation

JMS – 23 mars 2009

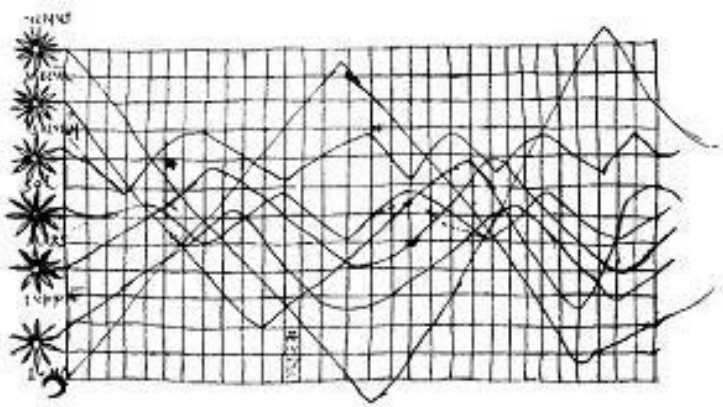
Dominique Ladiray
DSCT

- En hommage à Jean-Michel. Sans lui, ces JMS ont un petit goût amer ….



Classification et séries temporelles

Quelque part entre 768 et 814



Macrobius (Saturnales, 395)

- Deux disciplines très anciennes ….. dont les chemins se sont rarement croisés …. jusqu'à récemment.

Introduction

- La classification
  - Regrouper les individus en groupes homogènes
  - 2 individus du même groupe se ressemblent ; 2 individus de 2 groupes distincts sont différents.
  - Une méthode « naturelle » et exploratoire ancienne pour les données d'enquête
  - Principes de la CAH : Andanson, Histoire naturelle du Sénégal, 1757. Mais, bien avant, Charlemagne déjà ….
- Comment transposer ces méthodes aux cas des séries temporelles et pourquoi faire ? (prévision, désaisonnalisation)
- Exposé volontairement peu technique

Plan

- Comment classer des individus ?
  - Distances et stratégies
- Le cas spécifique des séries temporelles
  - Les problèmes
  - Quelques idées et méthodes pour y remédier
- Application à la désaisonnalisation
  - Les différentes facettes de la saisonnalité
- Application à la prévision
  - A la recherche d'un modèle

Distances et stratégies

- Regrouper les individus en classes homogènes
  - Les individus d'une même classe sont « proches »
  - Notion de distance-similarité
  - ∃ centaines de distances/similarités (dont euclidienne)
- Rendre les classes les plus différentes possibles
  - Stratégie d'agrégation de classes
  - ∃ dizaines de stratégies
  - Dont Ward: maximiser la variance inter-classes
- Le choix de la distance et de la stratégie dépend du problème à traiter …. Un petit exemple

Distance-similarité …. Un concept très relatif

Restons politiquement correct …..

Quelques méthodes

- Méthodes de partitionnement
  - On définit a priori un nombre de classes et on cherche une partition de la population
  - Centres mobiles, nuées dynamiques
- Méthodes hiérarchiques
  - Ascendantes, descendantes
  - Arbre de classification (dendogramme)

- Peut-on adapter ces méthodes au cas des séries temporelles ?

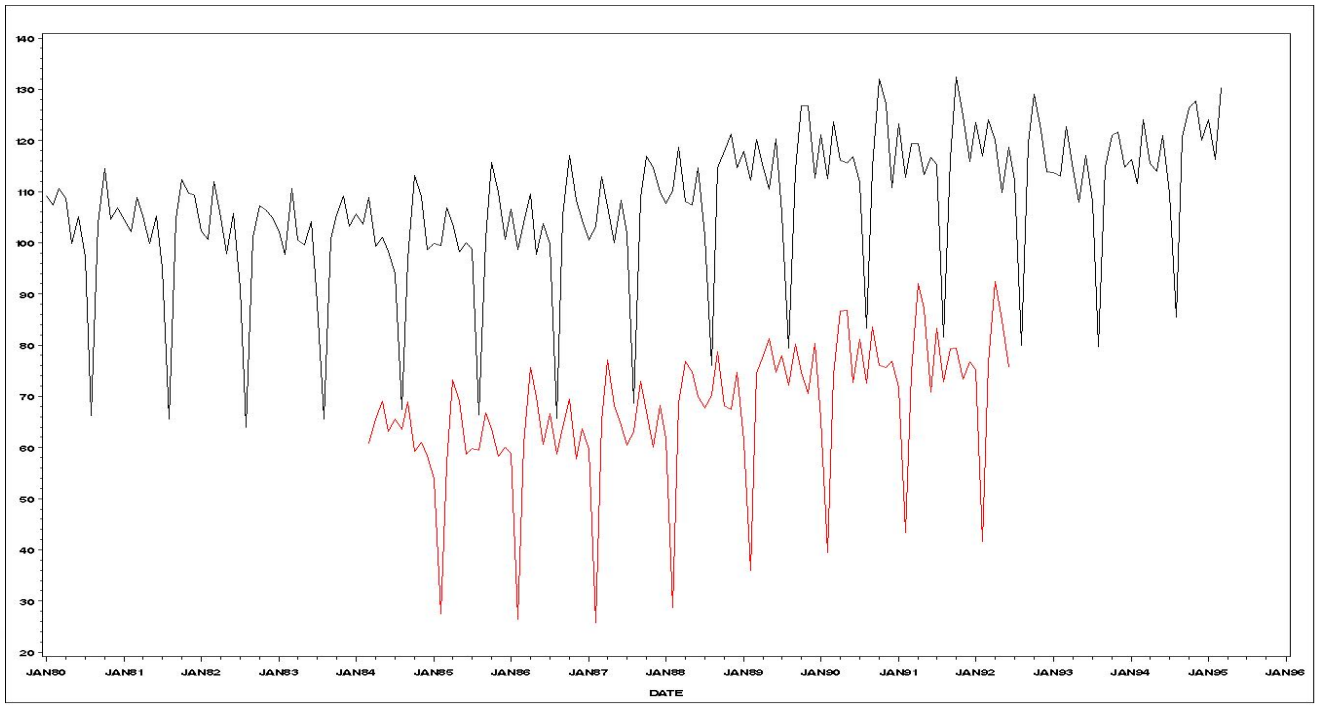
Une explosion d'intérêt ….

- Depuis 20 ans arrivée d'immenses bases de séries temporelles
  - Génome humain, consommation d'énergie, météorologie, reconnaissance de la parole, de l'écriture etc.
- Besoin de méthodes pour explorer ces bases
- Depuis 20 ans, des centaines de papier pour indexer, classer, discriminer les séries temporelles

- Passage à l'économie plus récent (séries plus courtes et moins nombreuses)

Le problème fondamental

- Similarité ?

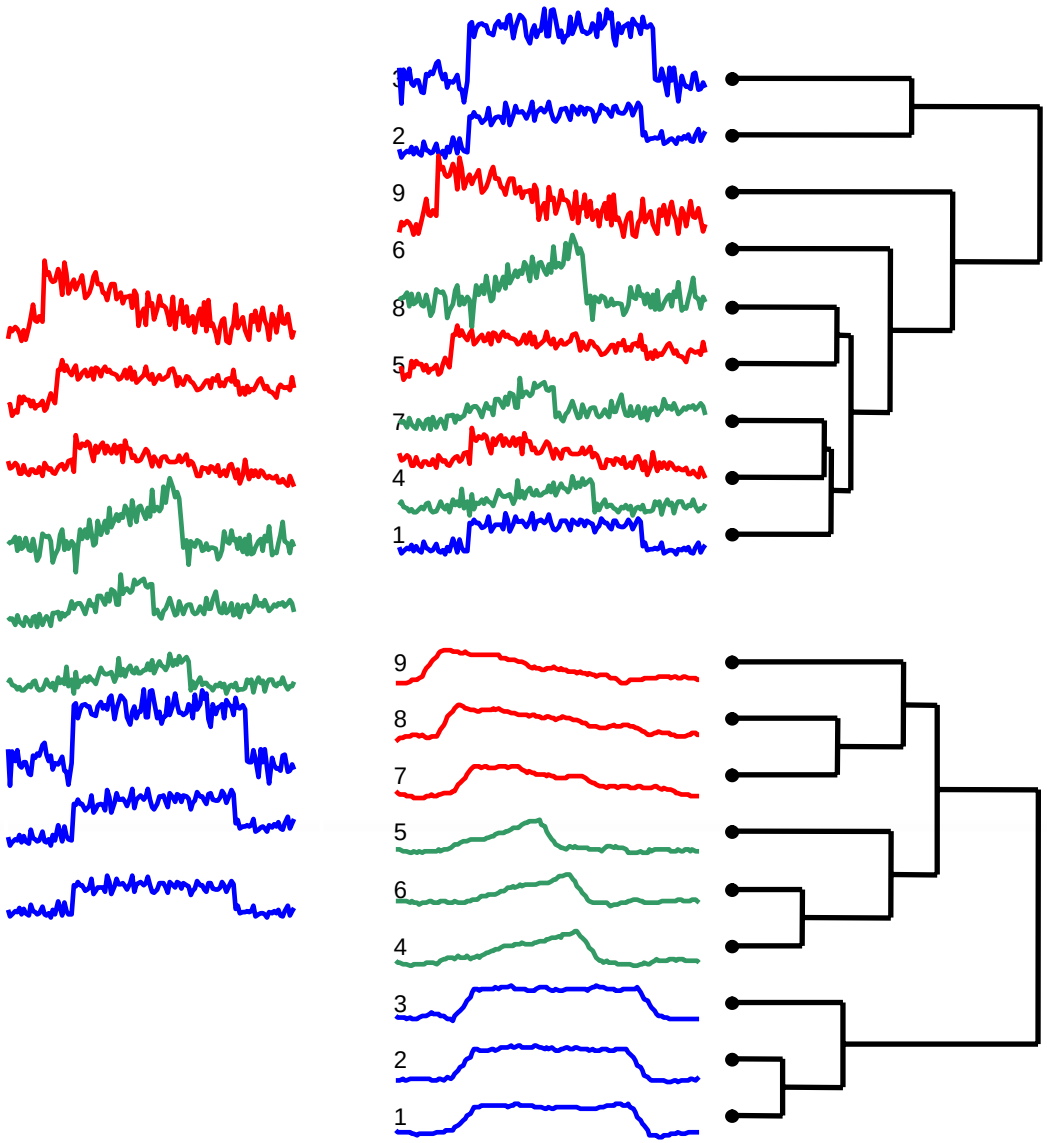


- Distance euclidienne pas toujours adaptée.
- Que faire ?

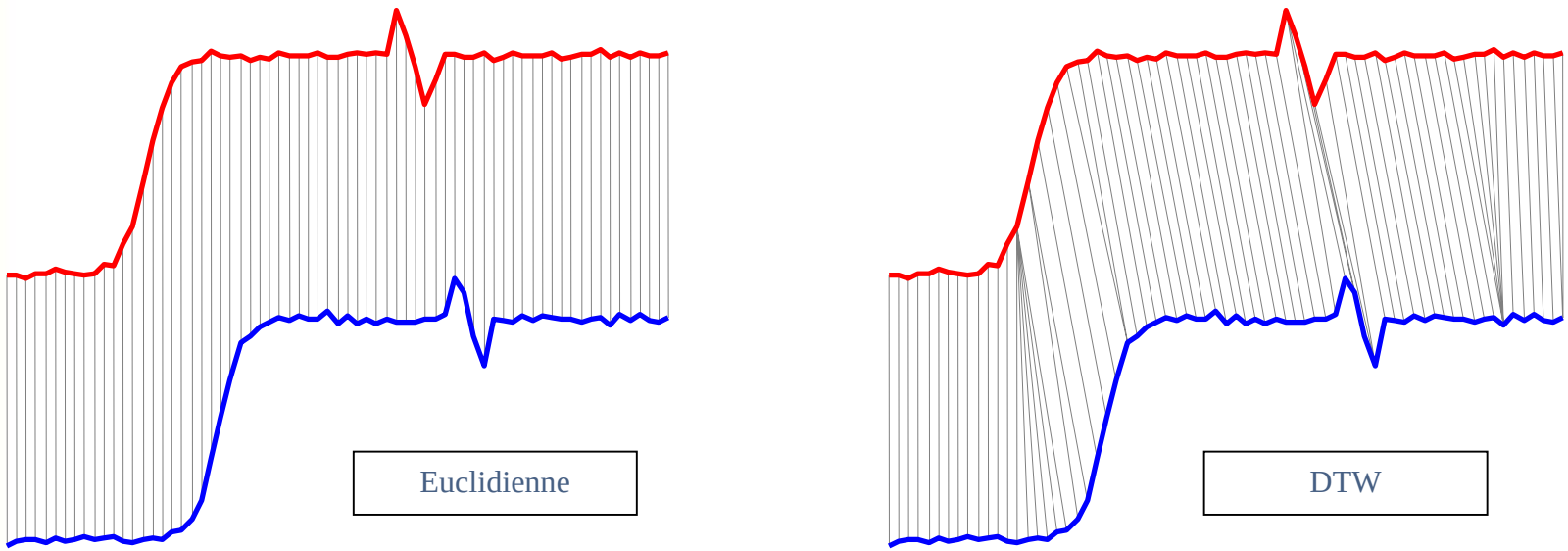
Quelques idées

- Préparer les données à la classification
  - Lissage, changement d'échelle, enlever la tendance etc.
- Inventer de nouvelles distances
  - Exemple : Dynamic Time Warping
    (To warp : voiler, gondoler, courber)
- Décrire les données de façon plus « économe »
  - Fonctions d'autocorrélation,Spectre, ondelettes etc.
- Modélisation stochastique les données
  - Modèles AR ou ARIMA, markovien caché (HMM) etc.
  - Distances spéciales (Corduas et Piccolo, 2008).

Préparer les données



Dynamic Time Warping

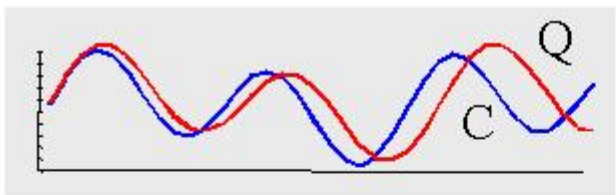


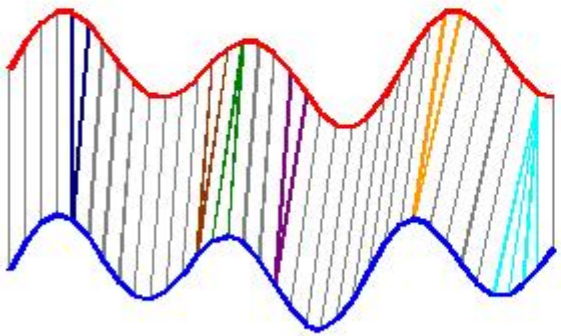
- Enquêtes de conjoncture
- Mais coûteux en temps de calcul

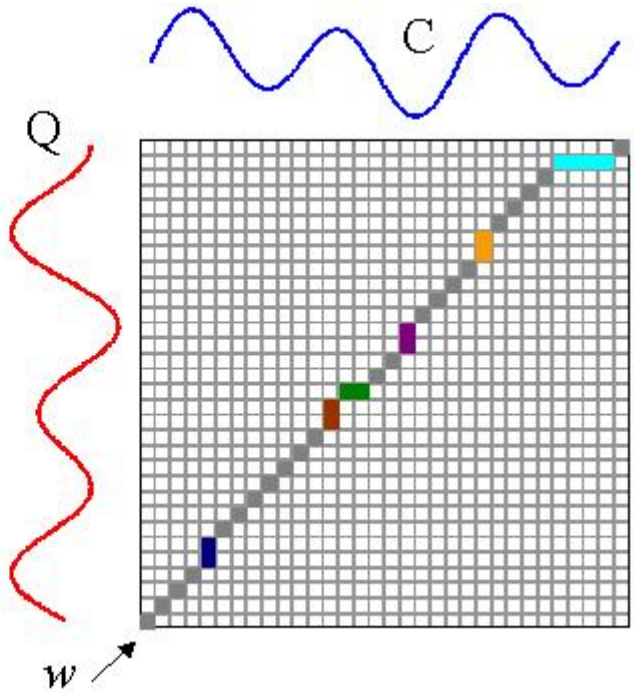
Calcul de la distance DTW

- Tout chemin dans la matrice des distances est une « courbure du temps ». On choisit la « meilleure », la « plus courte »

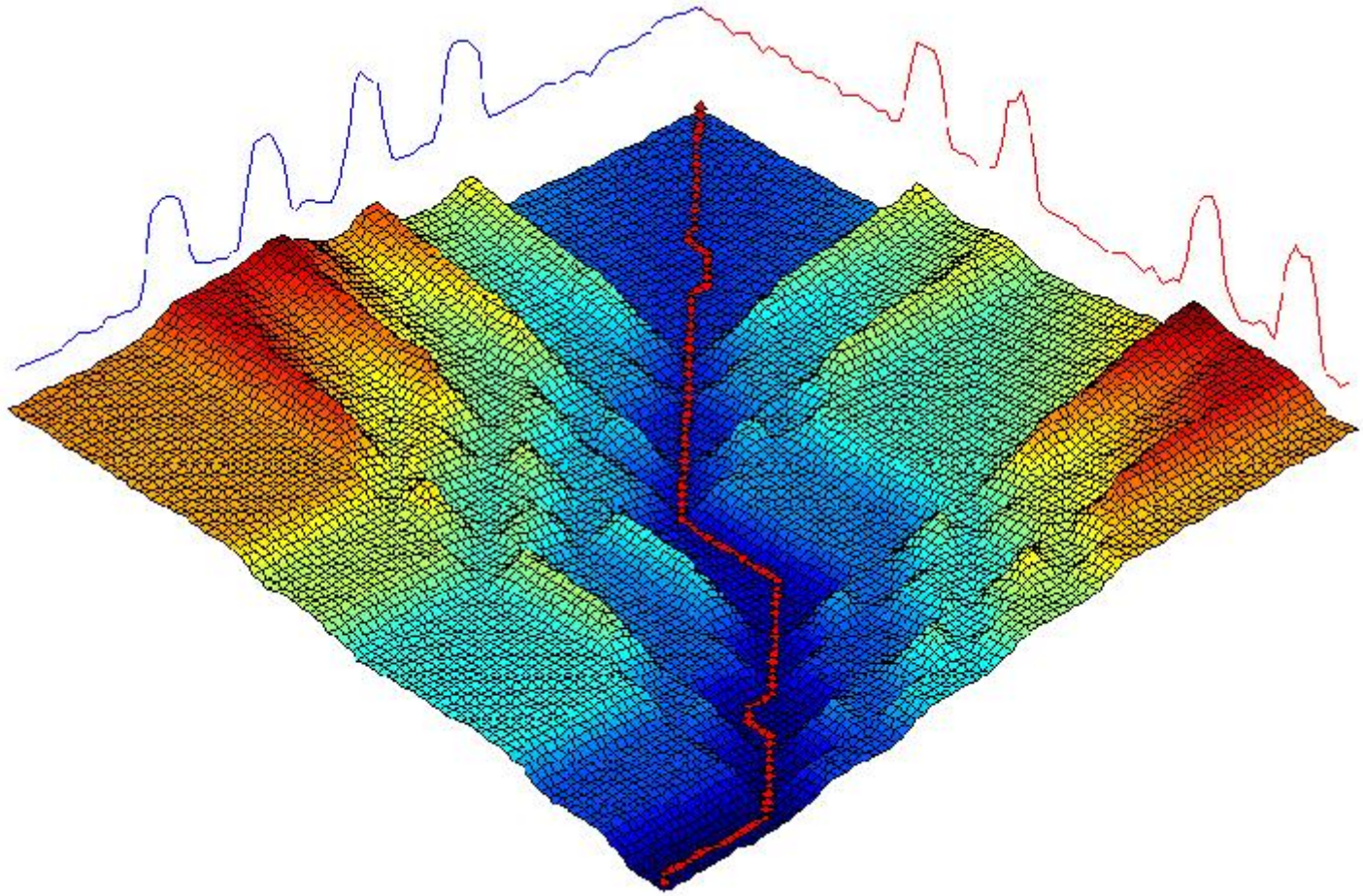
$$DTW(Q,C) = \min_{w} \frac{1}{N} \sum_{k=1}^{k=N} w_k$$





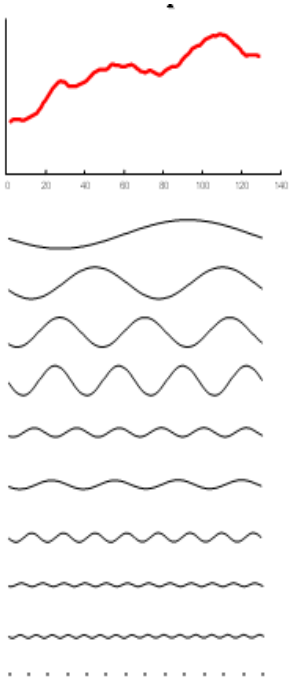


Un exemple réel

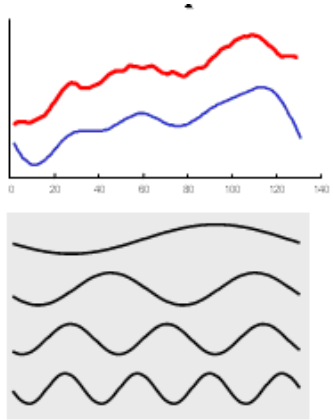


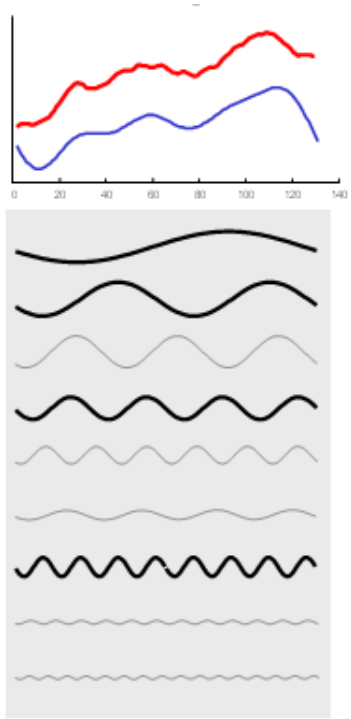
« Réduire » les données : un exemple

Série brute, coefficients de Fourier et fonctions associées	Décomposition sur les 4 premières fonctions	Décomposition sur les 4 plus importantes fonctions



0.4995	1.5698
0.5264	1.0485
0.5523	0.7160
0.5761	0.8406
0.5973	0.3709
0.6153	0.4670
0.6301	0.2667
0.6420	0.1928
0.6515	0.1635
0.6596	0.1602
0.6672	0.0992
0.6751	0.1282
0.6843	0.1438
0.6954	0.1416
0.7086	0.1400
0.7240	0.1412
0.7412	0.1530
0.7595	0.0795
0.7780	0.1013
0.7956	0.1150
0.8115	0.1801
0.8247	0.1082
0.8345	0.0812
0.8407	0.0347
0.8431	0.0052
0.8423	0.0017
0.8387	0.0002
...	...

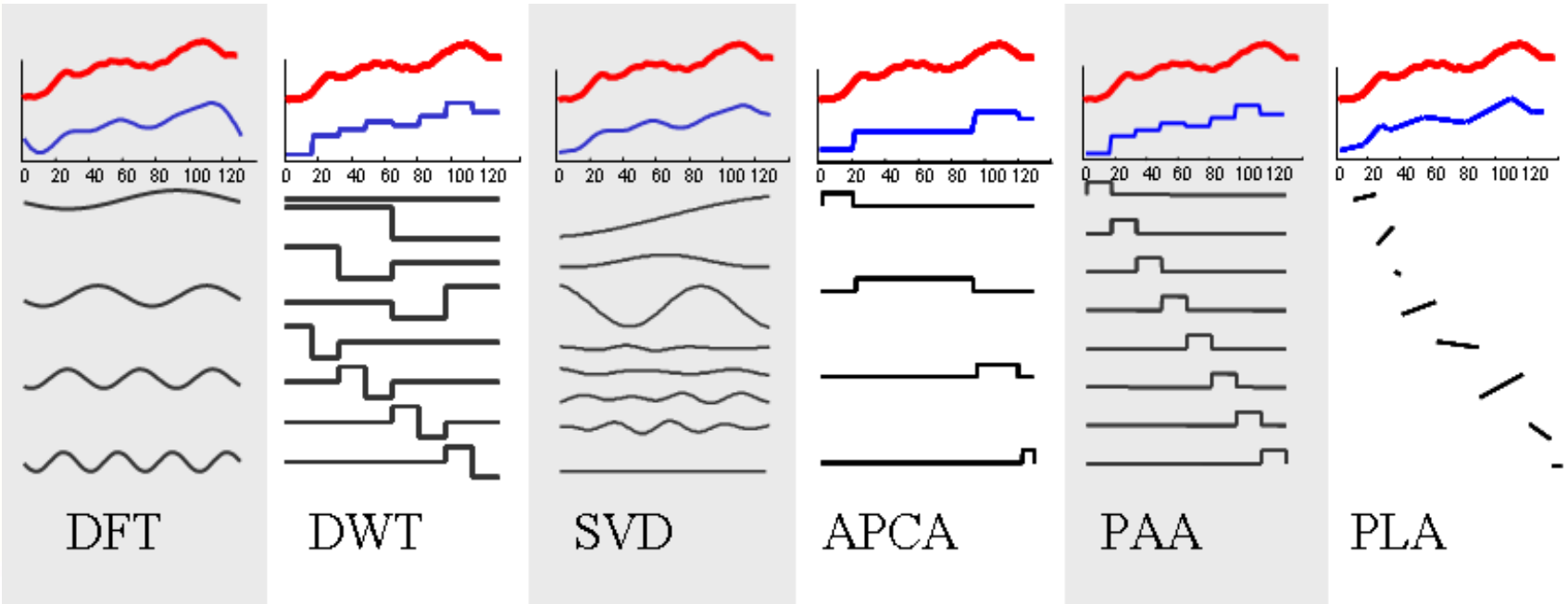




De nombreuses autres idées

- Fonctions d'autocorrélation directe, inverse, partielle (ACF, PACF, IACF) (Maballée and Maballée, 1911; Wang and Wang., 2000);
- Transformée de Fourier discrète (DFT) (Agrawal et al., 1993);
- Transformées par ondelettes, utilisant les bases de Daubechies, Haar (DWT) ou autres (Huntala et al., 1997);
- Polynômes de Chebyshev (Ng and Cai, 2004)
- Codage du Cepstrum (LPC), (Kalpakis et al., 2001);
- Décomposition en valeurs singulières via une ACP par exemple (Korn et al., 1997; Cleveland, 2004);
- Approximations linéaires par morceaux (Morikane et al., 2001);

Un petit dessin pour mieux comprendre



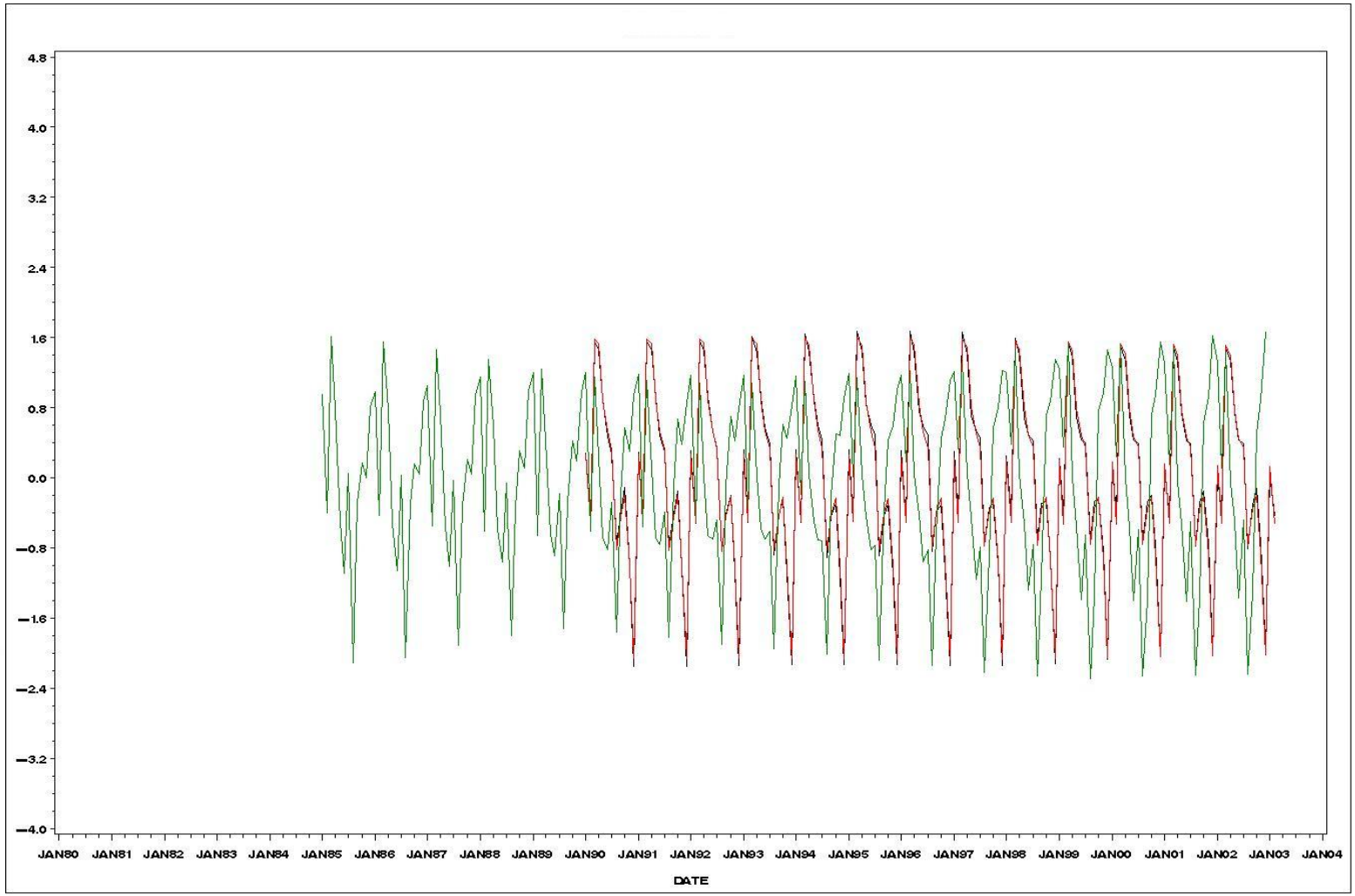
La désaisonnalisation

- Un modèle complexe ….

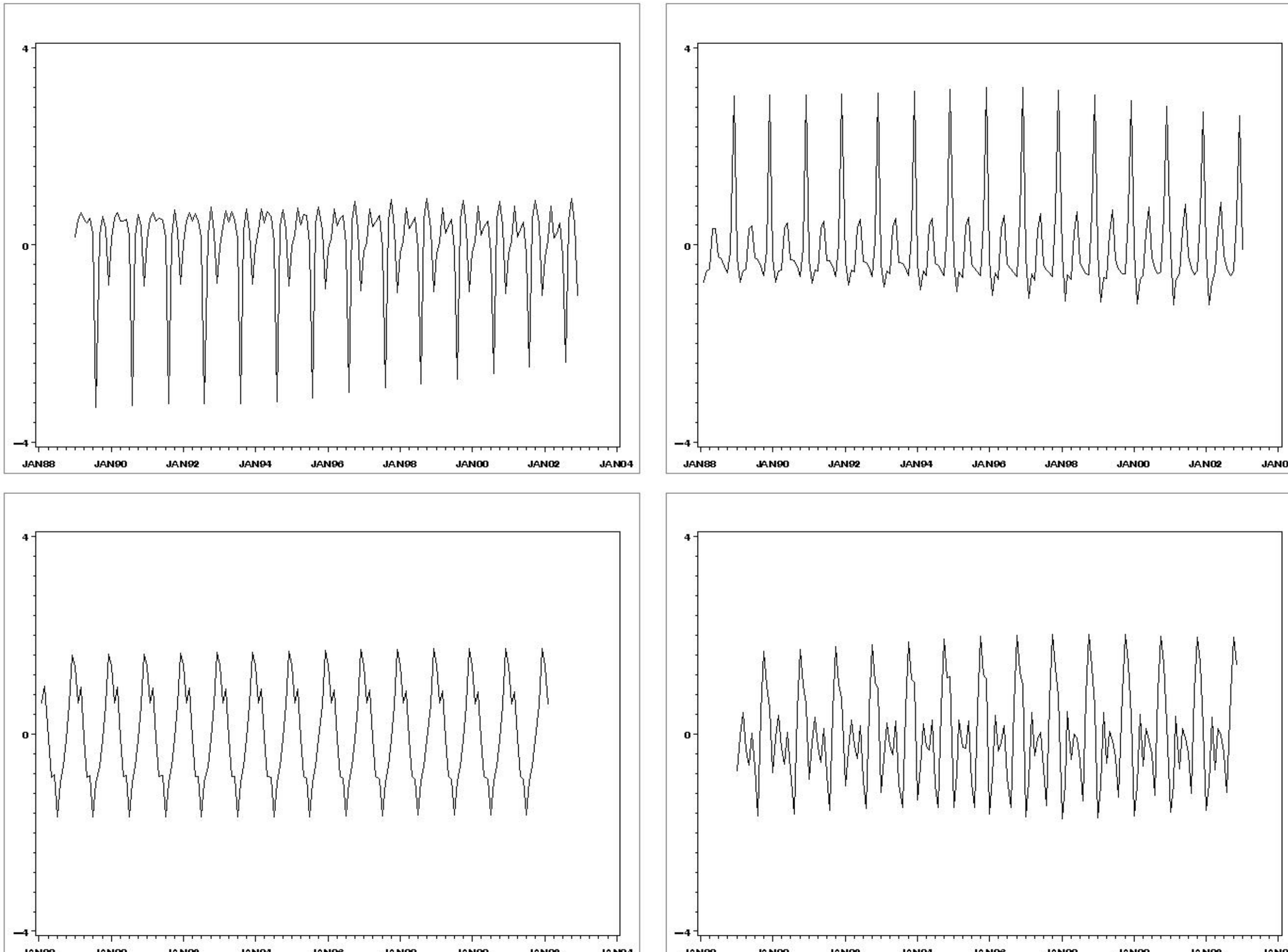
$$X_t = TC_t + S_t + TD_t + MH_t + O_t + I_t$$

- Où seule la série X_t est observable.
- A quoi ressemble la saisonnalité S_t ?
- Désaisonnalisation de 1100 séries mensuelles de la base Euro-Indicateurs avec Tramo-Seats et X12-Arima
- On classe ensuite les spectres des 2200 composantes saisonnières obtenues

Quelques parangons



Quelques saisonnalités caractéristiques



Les séries « Frankenstein »

- La classification des diverses composantes permet de définir des composantes « types » :
  - Tendance-cycles, saisonnalité, effets de jours ouvrables
- Et donc, par combinaison, de définir des séries « Frankenstein », de « fausses-vraies » séries qui permettent:
  - D'évaluer la qualité des méthodes de désaisonnalisation
  - De tester de nouvelles méthodes;
  - De proposer des valeurs par défaut adaptées pour les paramètres des logiciels.

Application à la prévision économique

- Problème : le PIB trimestriel de la zone Euro est publié à t+45, beaucoup plus tard que le PIB américain.
- Peut-on publier plus tôt ?
  - Très difficile d'accélérer la production
    - ⇒ Utiliser des modèles économétriques
- On connaît en effet beaucoup de choses sur le trimestre qui vient de s'écouler :
  - 2 mois au moins d'IPI, CA etc.
  - enquêtes de conjoncture
- Comment trouver un modèle ????

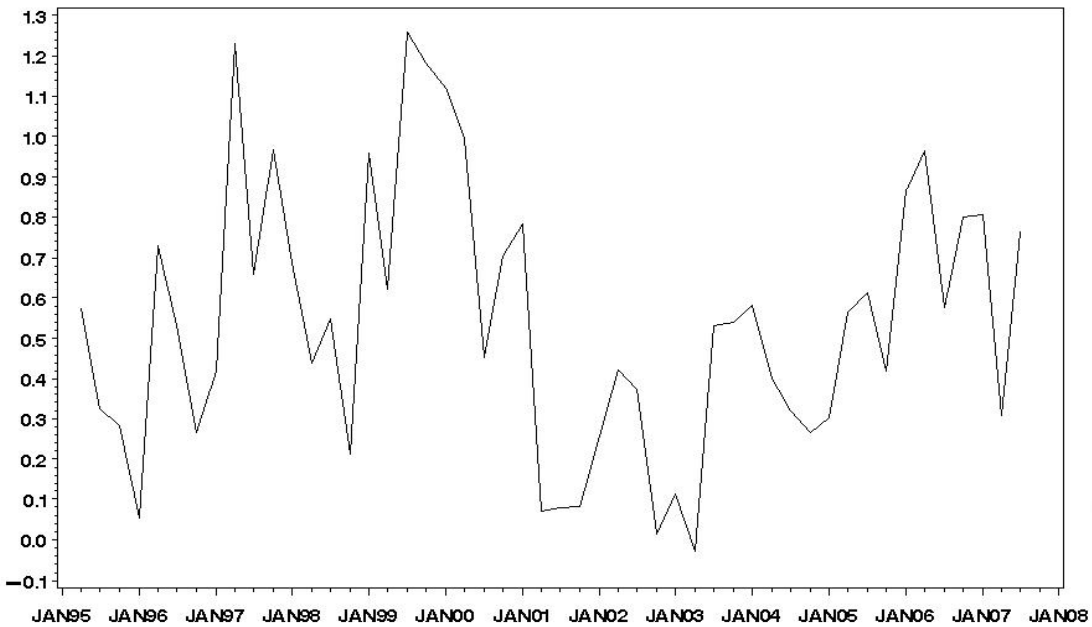
Qu’est-ce qu’un bon modèle ?

- Plusieurs caractéristiques importantes :
  - il doit être simple, c'est-à-dire ne faire intervenir qu'un nombre limité de variables,
  - il doit être interprétable : les relations exprimées doivent avoir un sens économique,
  - il doit être stable dans le temps, et en particulier ne doit pas être remis en cause chaque mois,
  - et enfin, il doit avoir un bon pouvoir prédictif.

- Mélange de caractéristiques « économiques » (1,2) et « statistiques » (3,4)

La variable à expliquer

- Le glissement trimestriel du PIB de la zone Euro (EA13)



- Ajustement automatique d'un (0,1,1) par Tramo : Rmse=0.29

Les variables explicatives

- Facile de trouver une vingtaine de variables candidates :
  - IPI, CA, Commandes, Immatriculations, enquêtes de conjoncture, IPC, prix de l'énergie, construction etc.
- 20 variables, 13 pays + EA, 2 retards ?
  - 20*(13+1)*3=840 variables potentielles
- Plus de 20 milliards de modèles à 4 variables !!!!

- Comment choisir ????

La sélection de variables

- Vous devez (?) réduire le nombre d'exogènes
  - Approche usuelle par tâtonnement : ne marche pas !!!

    Principe Shadok: « *Plus ça rate, plus on a de chances que ça marche !* »
  - Approche du NIESR (J. Mitchell)

    Partir d'un ensemble très réduit de variables et évaluer tous les modèles possibles
  - Approche GETS (General to specific; Hendry)

    Partir d'un modèle sur-paramétré et utiliser des tests statistiques pour le simplifier
  - Approche par Analyse Factorielle Dynamique

    Résumer l'ensemble des variables en quelques facteurs
  - Approche par classification

Les variables explicatives

- Exemple
  - On veut “prévoir” 2007Q4 à 30 jours..
  - On regarde ce qui est disponible au 30/01/2008

- Variables de la base Euro-Indicateurs
  - 668 mensuelles; 162 trimestrielles
  - Nettoyage des variables (points atypiques)
  - Prévision des mensuelles et trimestrialisation

Classification et sélection

- Sélection des variables
  - Classification des exogènes en groupes de variables “semblables”
  - Sélection de quelques variables dans chaque classe en utilisant des tests de Causalité, des corrélations etc. Ou bien prendre les facteurs principaux de chaque classe
  - Exemple: 8 classes, 2 variables par classe, 2 retards $\rightarrow$ 48 variables potentielles $\rightarrow$ 200,000 modèles à 4 exogènes.

Evaluation des modèles

- Première sélection des modèles par régression MCO stepwise
  - R-square, Mallow's Cp, Adjusted R-square etc.
- Evaluation complète des modèles (régression avec auto-corrélation des erreurs)
  - Toute la batterie de tests traditionnels
  - Cohérence de signe avec Y, erreurs de prévisions à l'horizon 1 et 2 etc.
- Liste de *n* modèles classés par « pertinence statistique »

Choix des modèles

- Le choix final du modèle est alors basé sur des critères statistiques et sur des « critères d'expert » :
  - Le modèle a-t-il toutes les qualités statistiques requises ?
  - Est-il suffisamment simple et robuste ?
  - Est-il pertinent du point de vue économique et interprétable ? En général NON !!!
- Mais peut-être facilement amélioré en utilisant des variables de la même classe !!

Exemple

Rang	Modèle	R2	RMSE
1	ET_BAL_GR IO_PL IP_EA13 UE_NL	0,933	0,133
2	IO_PL IP_EA13 UE_NL	0,921	0,136
3	IO_PL IP_EA13 IT_RT_IE UE_NL	0,876	0,141
4	EPI_BU_PL IO_PL IP_EA13 UE_FR	0,869	0,144
5	IO_PL IP_EA13 IP_PT UE_NL	0,862	0,148
6	DIT_EA13 IO_PL IP_EA13 UE_NL	0,856	0,148
7	IO_PL IP_EA13 IT_RT_IE UE_FR	0,855	0,148
8	IO_PL IP_EA13 SV_PR_ES UE_FR	0,855	0,151
9	IOB_EA13 IO_PL IP_EA13 IT_RT_IE	0,851	0,153
100	PIB_EA13_1 PIB_EA13_2	0,257	0,296
101	PIB_EA13_1	0,226	0,296
102	PIB_EA13_2	0,040	0,298

- Modèle 2 : IO_PL, IP_EA13, UE_NL
  - IO_PL est proche de EXP_EA13
  - UE_NL est proche de UN_EA13
- Nouveau modèle : EXP_EA13, IP_EA13, UN_EA13 (Rmse=0.15)

Les 2 modèles ….