DIVERSES MACROS SAS : Analyse Exploratoire des Données, Analyse des Séries Temporelles.

Dominique LADIRAY *

Septembre 2010

*Institut National de la Statistique et des Études Économiques, Département des Statistiques de court-terme, 18 boulevard Adolphe Pinard, 75014 Paris, FRANCE, email : dladiray@free.fr.

Table des matières

Introduction

Au fil des ans, j'ai écrit un assez grand nombre de macros SAS, que ce soit dans le cadre de cours donnés à l'ENSAE et à l'ENSAI, ou plus simplement pour accomplir mon travail de statisticien-économiste. Certains de ces programmes pourront peut être vous aider et je les mets bien volontiers à votre disposition.

Les macros SAS présentées dans les pages qui suivent sont regroupées en deux grandes catégories : celles liées à l'Analyse Exploratoire des Données (une branche de la Statistique malheureusement bien mal connue) et celles liées à l'Analyse des Séries Temporelles (mon travail de tous les jours).

Toutes ces macros ont été testées sous les versions 8 et 9 de SAS et elles ont l'air de bien fonctionner.

Chaque macro fait l'objet d'une présentation assez standard :

- Quelques mots sur le contexte et l'objet de la macro et une bibliographie sommaire au cas où vous souhaiteriez en savoir plus sur le sujet ;
- Une présentation détaillée des paramètres et la liste des modules SAS utilisés dans le programme ;
- Quelques exemples mettant en évidence les possibilités du programme.

Les macros sont livrées sous fome compilée dans un catalogue SAS. Pour les utiliser, vous devez donc allouer ce catalogue en utilisant par exemple des instructions SAS du type :

```
LIBNAME store  "D:\Dominique\SASstore";
OPTIONS MSTORED SASMSTORE=store;
%boxplot(DATA=monde,VAR=esper_h esper_h,TYPE=4,GROUP=conti);
```

Bien entendu, il doit rester des erreurs dans ces programmes. Si vous en rencontrez, merci de me le signaler. De même, si vous avez des idées de nouveaux paramètres, de développements possibles, n'hésitez pas à me contacter par courrier électronique (dladiray@yahoo.fr).

Cette documentation est régulièrement mise à jour sur le site de l'association MIRAGE (Mouvement International pour le développement de la Recherche en Analyse Graphique et Exploratoire) :

http ://www.unige.ch/ses/sococ/mirage/.

Enfin, si vous utilisez ces programmes, merci de citer vos sources, et en particulier le lien internet ci-dessus !

1 Analyse Exploratoire des Données

Les méthodes d'Analyse Exploratoire des Données (EDA : Exploratory Data Analysis) sont assez présentes dans le logiciel SAS, soit évidemment dans le module SAS/INSIGHT, soit dans quelques procédures. Vous trouverez une présentation générale en français de l'EDA dans Destandeau, Ladiray, Le Guen ([7]) [1] et une présentation complète de SAS/INSIGHT dans Destandeau, Le Guen ([8]).

1.1 %BOXPLOT : Boîtes à Pattes (Box and Whiskers Plot)

1.1.1 Brève définition

La Boîte à Pattes (BàP, Box Plot) est la figure emblématique de l'Analyse Exploratoire des Données. C'est en fait une simple représentation graphique des résumés de base de la variable étudiée : médiane, quart haut, quart bas, minimum et maximum. La portée inter-quarts (proche de l'intervalle inter-quartiles classique) sert aussi de mesure de base pour détecter les points adjacents, proches et lointains qui sont alors représentés par des symboles particuliers.

SAS propose des BàP dans les modules SAS/INSIGHT et SAS/IML. La version 8 contient une PROC BOXPLOT qui permet de réaliser de beaux graphiques. Malheureusement, elle est assez limitée dans la mesure où elle ne permet que de faire des BàP par groupes (une sorte d'analyse de la variance graphique). La macro %BOXPLOT utilise cette procédure et permet de faire d'autres types d'analyses : BàP simple, parallèles (plusieurs variables), par groupes et parallèles par groupe. En outre, vous pouvez aussi visualiser l'échelle simple de Tukey qui permet de trouver la transformation qui symétrise votre variable.

Les boîtes à pattes et l'échelle simple de Tukey sont décrites dans tout livre d'analyse exploratoire des données, comme Hoaglin et Velleman ([19]), Hoaglin, Mosteller et Tukey ([17]) ou bien même Tukey ([29]). La documentation du module SAS/STAT présente aussi de façon détaillée la boîte à pattes dans le chapitre consacré à la PROC BOXPLOT ([27]).

1.1.2 Paramètres et mise en oeuvre

La macro %BOXPLOT utilise plusieurs modules : SAS/BASE, SAS/GRAPH, SAS/STAT et SAS/IML. Elle s'appelle par l'instruction générale suivante :

```
%boxplot(DATA=,TYPE=,VAR=,GROUP=,ID=,FORMATID=,BOXSTY=,
         NOTCHES=,LABEL=,TITLE=);
```

Elle a donc 10 paramètres.

DATA : Nom de la table SAS où figurent les variables à analyser. Par défaut la dernière table créée (_LAST_).

1. Ce numéro consacré à l'EDA contient aussi une bibliographie complète.

TYPE : Type de BàP. Cinq possibilités :

0 : BàP simple, pour une variable ;

1 : BàP d'une variable par groupe (paramètre GROUP) ;

2 : BàP parallèles (plusieurs variables) ;

3 : BàP parallèles avec standardisation, les variables sont représentées avec la même étendue ce qui permet de mieux comparer les formes des BàP ;

4 : BàP parallèles et par groupe (paramètre GROUP) ;

5 : BàP pour réexpression (échelle simple de Tukey).

VAR : Variables analysées. Elles doivent être numériques. Par défaut, on prend toutes les variables numériques de la table SAS en entrée. Dans ce cas le paramètre GROUP est mis à blanc. On peut utiliser les conventions de liste SAS usuelles.

GROUP : Variable de groupe, obligatoire pour TYPE=1 ou 4. Attention de bien choisir une variable à peu de modalités !

ID : Variable pour identifier les points extrêmes. Non valable pour TYPE=5.

FORMATID : Format utilisé pour l'impression des identifiants (paramètre ID).

BOXSTY : Style de la BàP (voir la documentation de la PROC BOXPLOT) :

1 : schematic (le défaut si ID est à blanc),

2 : schematicid (le défaut si ID est renseigné),

3 : schematicidfar.

LABEL : Si renseigné par une valeur quelconque, on met le label des variables dans les graphiques, sur l'axe des abscisses. Valide seulement pour TYPE=2. Par défaut rien.

TITLE : Si renseigné à YES, le titre courant est introduit dans le graphique. Par défaut TITLE=no.

1.1.3 Quelques exemples

Boîte à Pattes simple avec identification des extrèmes

L'instruction suivante demande la BàP de la variable espérance de vie des hommes (variable ESPER_H) avec identification des points proches et lointains :

```
%boxplot(DATA=monde,VAR=esper_h,TYPE=0,ID=pays);
```

La BàP est représenté dans la figure 1. La Sierra Leone (SIER) apparaît comme ayant une espérance de vie des hommes très basse.

Boîte à Pattes multiple par groupe

L'instruction suivante demande les BàP des variables espérance de vie des hommes (ESPER_H) et espérance de vie des femmes (ESPER_F) par continent :

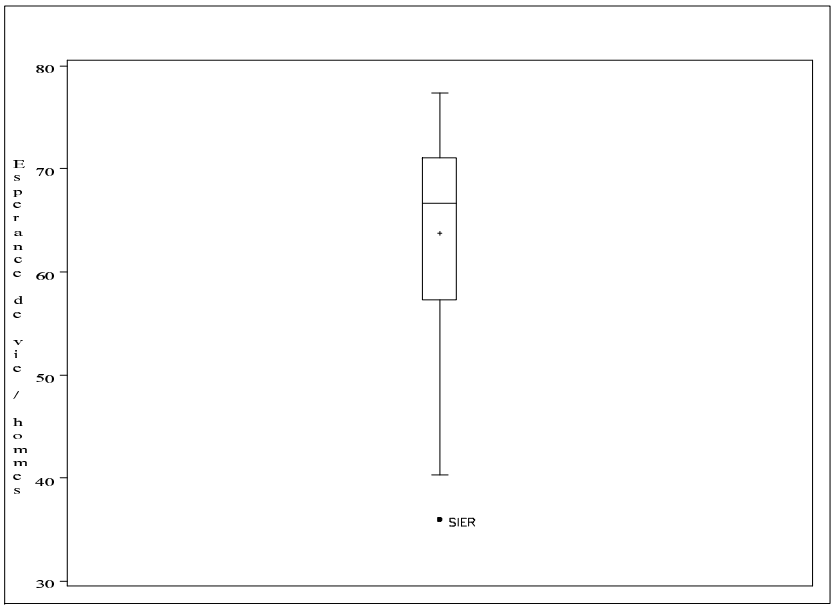


FIGURE 1 – BàP de la variable espérance de vie des hommes (ESPER_H) : les points proches et lointains sont repérés par la variable PAYS.

```
%boxplot(DATA=monde,VAR=esper_h esper_h,TYPE=4,GROUP=conti);
```

Les BàP sont représentés dans la figure 2. L'espérance de vie des femmes apparaît plus élevée que celle des hommes dans tous les continents.

échelle simple de Tukey

L'instruction suivante demande la reexpression de la variable PNB par habitant selon les puissances de l'échelle simple de Tukey :

```
%boxplot(DATA=monde,VAR=pnbhab,TYPE=4);
```

La figure 3 montre les BàP des différentes transformations. Le logarithme du PNB par habitant (puissance 0 dans l'échelle de Tukey) semble avoir une distribution symétrique.

1.2 %CLUSTER : Classification ascendante hiérarchique

1.2.1 Brève définition

Il est souvent très utile de regrouper les individus présentant des caractéristiques communes afin par exemple de comprendre comment se structure un paquet

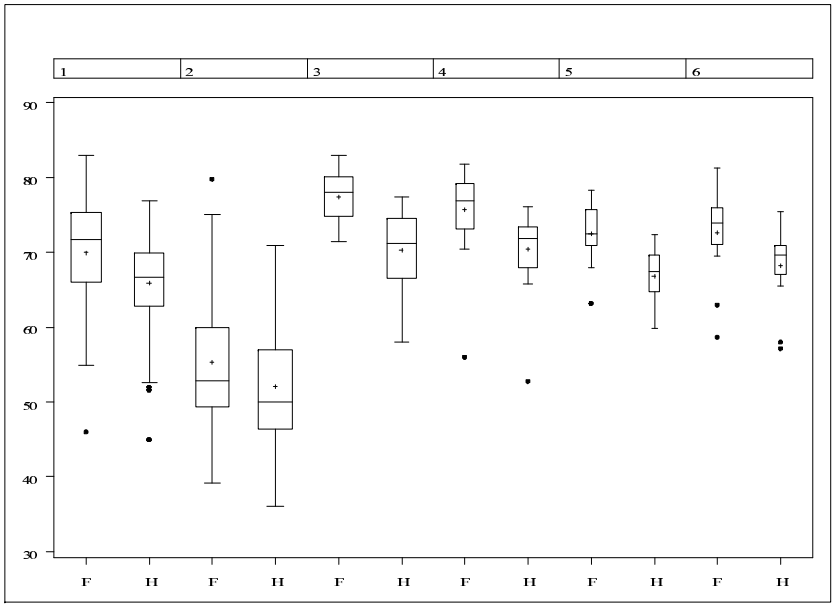


FIGURE 2 – BàP des variables espérance de vie des hommes (ESPER_H) et espérance de vie des femmes (ESPER_F) par continent.

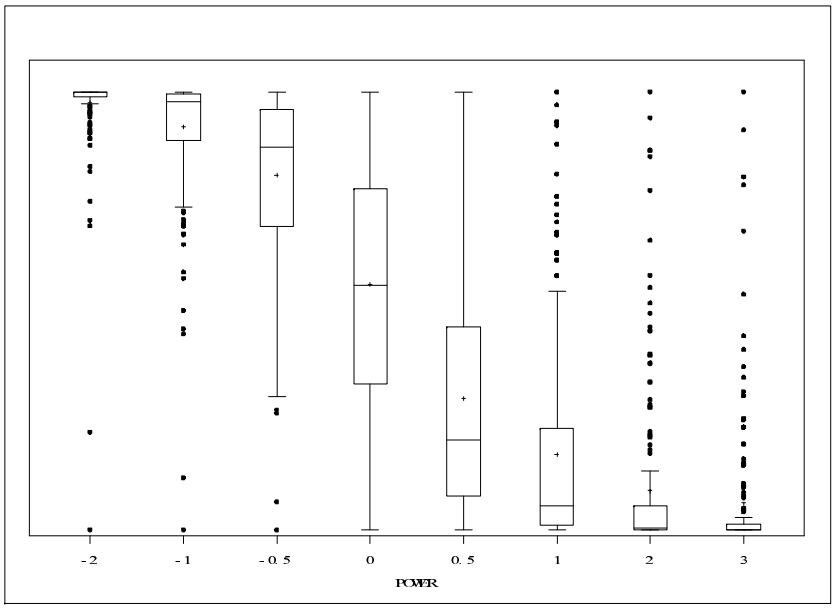


FIGURE 3 – échelle de Tukey du PNB par habitant (PNBHAB).

de données. C'est l'objectif de la macro %Cluster qui réalise une classification ascendante hiérarchique sur un ensemble d'individus. En partant d'une population de n individus, la détermination des classes se fait itérativement. Les deux individus les "'plus proches"' sont regroupés dans une classe représentée par son centre de gravité ; puis les deux individus (ou classes) les plus proches sont agrégés et ainsi de suite jusqu'à reconstituer la population totale. Cette méthode permet donc de déterminer un arbre d'agrégation ou chaque classe est emboîtée dans une autre. La macro permet ensuite de calculer les caractéristiques des classes en fonction des variables numériques ayant servi à déterminer les classes mais aussi en fonstion de toute autre variable, qualitative ou non.

1.2.2 Paramètres et mise en oeuvre

La macro %CLUSTER utilise les modules SAS/BASE, SAS/STAT, SAS/GRAPH et SAS/IML et se met en oeuvre par l'instruction générale suivante :

```
%Cluster(DATA=,VAR=,VAREXP=,IDOBS=,FREQ=,STANDARD=,DCHI2=,
         OUTNEW=,OUTCAH=,CLUSTERS=,MAXOBS=,NFAST=,NBCLUS=,
         CLUSNAME=,MEDOIDS=,GMAX=,PRINT=,FORMAT=);
```

Elle a 18 paramètres.

DATA : Nom de la table SAS où figurent les variables à analyser. Par défaut la dernière table créée (_LAST_).

VAR : Noms des variables à traiter. Ces variables doivent être numériques. Si ce paramètre est à blanc, toutes les variables numériques de la table seront utilisées.

VAREXP : Noms des variables explicatives qui seront utilisées pour interpréter les classes. Elles peuvent être numériques ou caractères. Si ce paramètre est à blanc, toutes les variables actives (paramètre VAR) seront utilisées.

IDOBS : Nom de la variable identifiant les observations. Si ce paramètre est à blanc, on prend le numéro de l'observation (OBxx).

STANDARD : Si "'yes"', alors les variables actives seront standardisées avant de faire l'analyse. Ce paramètre est utile si les variables ont des moyennes et variances très différentes. Par défaut STANDARD=no.

DCHI2 : Si "'yes"', alors on utilise la distance du Chi2. Par défaut DCHI2=no.

OUTNEW : Nom de la table SAS en sortie contenant toutes les variables de la table en entrée (paramètre DATA) et la variable de classe. Par défaut OUTNEW=_outnew_.

OUTCAH : Nom de la table SAS en sortie contenant tous les résultats de la classification (agrégations successives). Par défaut OUTCAH=_outcah_.

CLUSTERS : Nom de la table SAS en sortie avec le contenu de chacune des classes. Par défaut CLUSTERS=_clusters_.

MAXOBS : Si vous avez plus de MAXOBS observations, la macro va d'abord faire une classification par centres mobiles (FASTCLUS) en NFAST classes puis une CAH sur les centres de ces NFAST groupes. Par défaut MAXOBS = 2000.

NFAST : Nombre de classes demandées pour la classification rapide. Par défaut, 500. Plus ce nombre est élevé, plus les classes seront a priori homogènes.

NBCLUS : Nombre de classes demandées pour la CAH finale. Par défaut, 5.

CLUSNAME : Nom de la variable de classe. Par défaut, cluster.

MEDOIDS : Nombre de medoids (individus les plus proches du centre de gravité de la classe) souhaités en sortie. Si le paramètre est à blanc, tous les individus de la classe seront listés.

GMAX : Nombre maximum de classes représentées dans l'arbre de classification. Par défaut, 50.

PRINT : Si "'yes"', alors le contenu des classes est imprimé. Par défaut PRINT=no.

1.2.3 Un exemple

L'instruction suivante demande une classification des pays de la table SAS Monde2008 en fonction d'un certain nombre de variables démographiques (PopGrowth, Fertility, BirthRate, SexRatio, HIVrate, LifeExp, MedianAge et MortInf). 5 classes sont demandées ainsi que le calcul des moyennes par classe des variables actives et d'autres variables explicatives (Cont, GDPgrowth, GDPind, GDPagri, GNIpcPPP, LFagri, LFind et Literacy).

```
%Cluster(DATA=eda.monde2008,
         VAR=PopGrowth Fertility BirthRate SexRatio
             HIVrate LifeExp MedianAge MortInf,
         VAREXP=cont GDPgrowth GDPind GDPagri GNIpcPPP
                LFagri LFind Literacy,IDOBS=pays,
         STANDARD=yes,OUTCAH=,CLUSTERS=,MAXOBS=1000,
         NFAST=500,NBCLUS=5,MEDOIDS=5,GMAX=,PRINT=yes);
```

L'arbre de classification est représenté à la figure 4 et les principales statistiques par classe à la figure 5.

1.3 %LETTER : Boîtes à Lettres (Letter Values Display)

1.3.1 Brève définition

Une Boîte à Lettres (BàL, Letter Value Display) est une représentation semi-graphique des résumés numériques d'une variable. Ces résumés sont basés sur les fractiles de la variable, représentés par des lettres, et sont donc fonction des statistiques d'ordre des observations.

SAS permet d'obtenir assez facilement des résumés numériques d'une variable, par exemple en utilisant la PROC UNIVARIATE ou le module SAS/INSIGHT, ces

	Total	CL01	CL02	CL03	CL04	CL05	st1	st2	st3	st4	st5
Frequency	111	60	22	9	18	2					
PopGrowth	1,16	1,67	0,59	0,40	0,59	0,77	++	---	--- -	---	--
Fertility	2,48	3,15	1,82	1,44	1,61	2,00	+	-	--	--	-
BirthRate	19,82	25,31	15,66	12,61	11,16	11,45	+		-	--	-
SexRatio	1,00	1,00	0,96	0,98	1,06	0,98		-		++	
HIVrate	1,51	2,39	0,77	0,25	0,24	0,15	+	-	--	--	--
LifeExp	70,16	64,97	72,98	77,37	79,37	79,50			+	+	+
MedianAge	29,23	23,78	32,19	36,54	39,07	39,00			+	+	+
MortInf	28,07	43,19	16,69	5,80	5,41	4,12	+	-	-	--	--
GDPagri	12,66	19,84	6,42	3,54	2,16	1,70	+	-	--	--	--
GDPgrowth	6,04	6,86	6,60	4,60	3,56	4,10					
GDPind	31,02	28,97	36,33	33,09	30,68	27,45		+			
GNIpcPPP	13540,90	4046,67	13274,55	23413,33	35587,22	58455,00					++
LFagri	29,33	45,81	16,39	7,29	4,19	2,50	+	-	-	--	--
LFind	19,82	14,80	25,17	25,27	27,52	17,50					
Literacy	85,52	77,35	93,25	95,32	96,82	100,00	--	++	++	+++	++++

FIGURE 4 – Classification ascendante hiérarchique : arbre de classification.

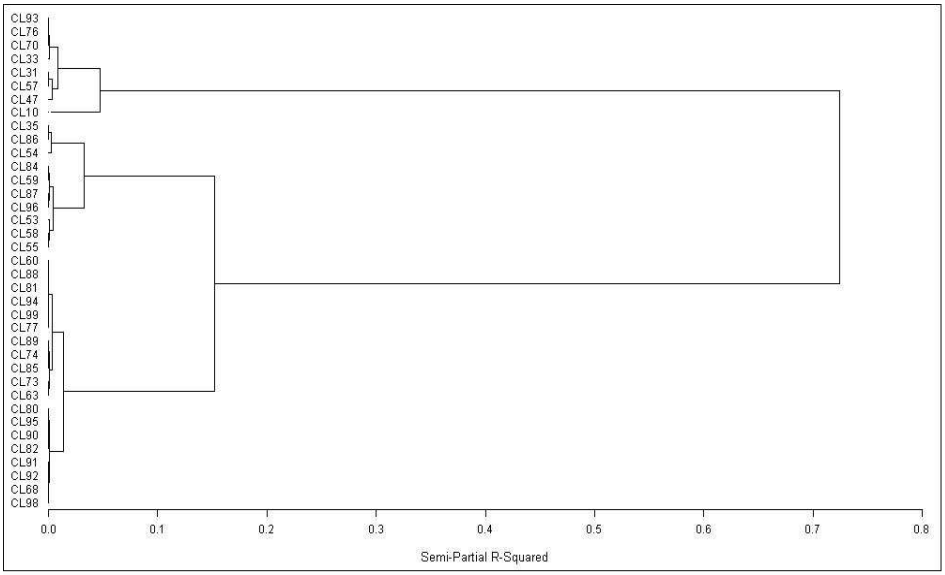


FIGURE 5 – Classification ascendante hiérarchique : statistiques par classe.

deux approches fournissant même les quantiles de la variable (notion assez proche des fractiles). On peut même, dans la version 8, obtenir des estimateurs robustes de tendance centrale et de dispersion avec la PROC STDIZE.

La macro %LETTER permet en outre d'obtenir les portées (spreads) et les centrages (midspreads) de la variable ainsi que les profondeurs associées aux différents fractiles. Toutes ces notions sont définies, par exemple, dans Hoaglin, Mosteller et Tukey ([17]).

1.3.2 Paramètres et mise en oeuvre

La macro %LETTER utilise les modules SAS/BASE et SAS/IML et se met en oeuvre par l'instruction générale suivante :

```
%letter(DATA=,VAR=,ID=,DEC=,SPREAD=,DETAIL=,MEAN=,TRONQ=);
```

Elle a 8 paramètres.

DATA : Nom de la table SAS où figurent les variables à analyser. Par défaut la dernière table créée (_LAST_).

VAR : Noms des variables à traiter. Ces variables doivent être numériques. Si ce paramètre est à blanc, toutes les variables numériques de la table seront analysées.

ID : Nom de la variable identifiant les observations. Si ce paramètre est à blanc, on prend le numéro de l'observation.

DEC : Nombre de décimales pour l'édition des résultats. Par défaut 2.

SPREAD : Paramètre demandant l'édition des portées (spreads) et des centrages (midspreads). Par défaut SPREAD = oui.

DETAIL : Nombre de résumés à éditer :

0 : tous les résumés possibles sont édités,

3 : Médiane et Extrêmes,

5 : Médiane, Extrêmes et Quarts (le défaut),

7 : Médiane, Extrêmes, Quart et Huitièmes.

MEAN : Calcul de la moyenne et variance. Par défaut MEAN = non.

TRONQ : Calcul des moyennes tronquées et winsorisées. Par défaut TRONQ = non.

1.3.3 Un exemple

L'instruction suivante demande la boîte à lettres, la plus complète possible, de la variable espérance de vie des hommes (variable ESPER_H) :

```
%letter(DATA=monde,VAR=esper_h,ID=pays,DEC=2,SPREAD=oui,
        DETAIL=,MEAN=oui,TRONQ=oui);
```

Résumés Numériques de la variable ESPER_H

198 observations
188 observations non manquantes

	Ident				Spread	Midspread	TRIMEAN
M(94)	SYRI	POLO		66.70			65.43
F(47)	PAPO	BERM	57.20	71.10	13.90	64.15	
E(24)	TOGO	IRLA	48.80	74.00	25.20	61.40	
D(12)	MOZA	MACA	45.50	75.10	29.60	60.30	
C(6)	GUIB	ISRA	42.40	75.70	33.30	59.05	
B(3)	OUGA	SUED	40.40	76.20	35.80	58.30	
O(1)	SIER	ISLA	36.00	77.40	41.40	56.70	

MEAN	VARIANCE
63.76	98.05

	TRONQ								
	5	10	15	20	25	30	35	40	45
Tronquées	64.26	64.70	65.19	65.64	66.09	66.47	66.59	66.67	66.78
winsorisées	63.89	64.00	64.18	64.52	65.14	65.92	66.49	66.50	66.66

FIGURE 6 – Boîte à Lettres de la variable espérance de vie des hommes (ESPER_H).

La boîte à lettres est représentée à la figure 6. Les moyennes tronquées et winsorisées sont éditées en fonction du pourcentage de troncature (TRONQ = oui). De même, on a les résumés classiques moyenne et variance (MEAN = oui). Les observations sont repérées par le nom court du pays. Il y a 188 valeurs renseignées et la médiane correspond donc à la profondeur 94, soit aux observations 94 (Syrie, SYRI) et 95 (Pologne POLO). De même, les quarts correspondent à la profondeur 47 (94/2) et donc à la Papouasie (PAPO, quart bas) et aux Bermudes (BERM, quart haut). La distribution est plutôt dissymétrique à droite comme le montre la décroissance des centrages.

1.4 %MPOLISH : Polissage d'un tableau croisé par médianes (Median Polish)

1.4.1 Brève définition

La macro %MPOLISH ajuste un modèle additif à un tableau croisé qui donne les valeurs d'une variable de réponse Y en fonction d'un facteur ligne A et d'un facteur colonne B.

Le modèle s'écrit de façon générale sous la forme : $Y_{ij} = m + A_i + B_j + \epsilon_{ij}$.

La macro %MPOLISH estime les différents paramètres et fournit certains diagnostics [2]. Ceux ci permettent non seulement de juger de l'adéquation du modèle

2. La macro %SQCOMB, présentée au paragraphe 1.8, ajuste selon une méthode différente le même type de modèle.

mais aussi de tester la pertinence d'un modèle à effets croisés. Pour en savoir plus, on peut consulter l'ouvrage collectif de Hoaglin, Mosteller et Tukey ([17]).

1.4.2 Paramètres et mise en oeuvre

La macro %MPOLISH utilise les modules SAS/BASE et SAS/IML. Le module SAS/GRAPH est en outre nécessaire aux graphiques de diagnostic. Elle fait aussi appel aux macros %STEM (voir paragraphe 1.9) et %RESLINE (voir paragraphe 1.7). Elle s'appelle par l'instruction générale suivante :

```
%mpolish(DATA=,COLS=,LIGNES=,OUT=,DIAGN=,INT=,ITER=,
         PRINT=,FORMAT=,POWER=,CRES=,CREG=);
```

Elle a donc 10 paramètres.

DATA : Nom de la table SAS où se trouve le tableau à analyser. Les colonnes du tableau sont des variables numériques de la table SAS.

COLS : Nom des colonnes du tableau. Ce sont nécessairement des variables numériques de la table précisée ci-dessus. Par défaut, on prend toutes les variables numériques.

LIGNES : Variable identifiant des lignes du tableau. Par défaut on prendra LIG1, LIG2, ..., LIGn.

OUT : nom de la table SAS en sortie avec les différents effets. Par défaut : _mpolish_.

DIAGN : Paramètre gérant l'édition de diagnostics. Dans le cas où ce paramètre est renseigné, un branchage des résidus et un "diagnostic plot" sont édités. Ces diagnostics permettent de statuer sur le caractère additif du modèle (transformation suggérée) et la présence d'un effet croisé. Par défaut, pas de diagnostic.

PRINT : Permet d'obtenir l'impression des résultats. Coder OUI, NON ou ALL. Si ALL, tous les résultats intermédiaires sont aussi imprimés. Par défaut PRINT=oui.

ITER : Nombre maximal d'itérations. Par défaut 5.

FORMAT : Format SAS numérique valide utilisé pour impression des résultats. Par défaut FORMAT=5.1

POWER : Puissance de la transformation sur la variable de réponse. Par défaut, pas de transformation : POWER=1.

CRES : Couleur du nuage des résidus. Par défaut RED.

CREG : Couleur de la ligne résistante. Par défaut BLACK.

Pour ces 2 derniers paramètres, choisissez des valeurs acceptables par SAS.

1.4.3 Un exemple

Soit la table SAS créée par le programme suivant :

```
DATA a;
 INPUT b1-b4;
 CARDS;
11.7  8.7 15.4  8.4
18.1 11.7 24.3 13.6
26.9 20.3 37.0 19.3
41.0 30.9 54.6 35.1
66.0 54.3 71.1 50.0
;
RUN;
```

Les relations entre lignes et colonnes du tableau sont analysées grâce à l'instruction :

```
%mpolish(DATA=a,DIAGN=oui);
```

La plupart des options par défaut sont utilisées. Le paramètre DIAGN indique qu'on souhaite obtenir les diagnostics. La figure 7 présente le résultat du lissage, les estimations des différents effets, et le branchage des résidus. La figure 8 présente le graphique de diagnostic. La structure linéaire du nuage de point montre un effet croisé.

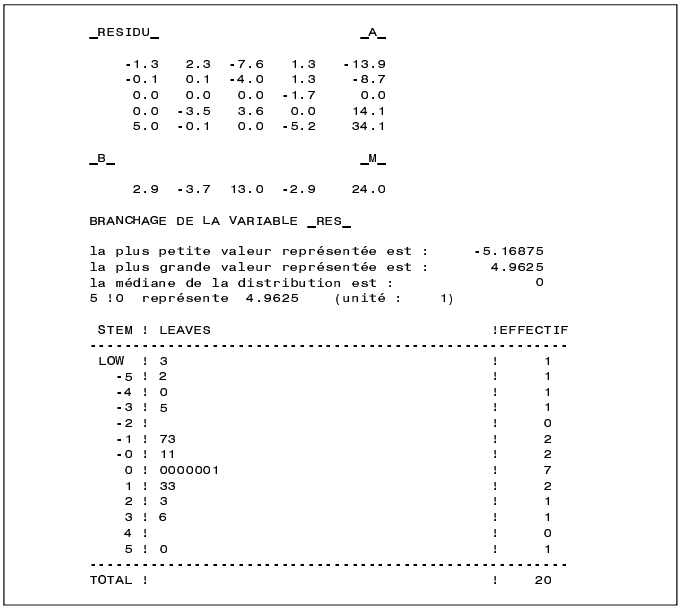
1.5 %QQPLOT : Adéquation graphique à une loi théorique (Q-Q Plots)

1.5.1 Brève définition

La macro %QQPLOT permet de tester l'adéquation de la loi d'une variable à une loi théorique à l'aide d'un graphique comparant les quantiles observés aux quantiles théoriques. Ce type de graphique est disponible dans la PROC UNIVARIATE (instruction QQPLOT) et dans le module SAS/INSIGHT (menu DISTRIBUTION). Dans ces deux cas, on peut obtenir de jolis graphiques et tester des lois non usuelles (Weibull notamment). L'intérêt de la macro %QQPLOT réside dans le fait qu'elle vous donne accès à des diagnostics (droite et limites de confiance). Pour en savoir plus, on peut consulter l'ouvrage collectif de Chambers, Cleveland, Keiner et Tukey ([6]) ou le livre édité par Fox et Long ([14]).

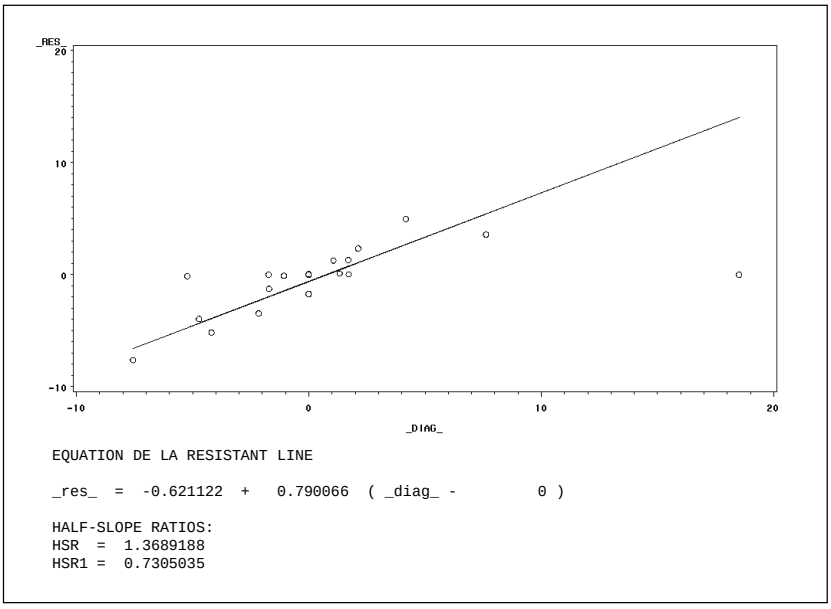
1.5.2 Paramètres et mise en oeuvre

La macro %QQPLOT utilise les modules SAS/BASE et SAS/GRAPH. Elle s'appelle par l'instruction générale suivante :



```
_RESIDU_                          _A_

    -1.3   2.3  -7.6   1.3   -13.9
    -0.1   0.1  -4.0   1.3    -8.7
     0.0   0.0   0.0  -1.7     0.0
     0.0  -3.5   3.6   0.0    14.1
     5.0  -0.1   0.0  -5.2    34.1

_B_                               _M_

     2.9  -3.7  13.0  -2.9    24.0

BRANCHAGE DE LA VARIABLE _RES_

la plus petite valeur représentée est :      -5.16875
la plus grande valeur représentée est :       4.9625
la médiane de la distribution est :                0
5 !0  représente  4.9625    (unité :    1)

 STEM ! LEAVES                                          !EFFECTIF
-------------------------------------------------------------------
 LOW  ! 3                                               !      1
   -5 ! 2                                               !      1
   -4 ! 0                                               !      1
   -3 ! 5                                               !      1
   -2 !                                                 !      0
   -1 ! 73                                              !      2
   -0 ! 11                                              !      2
    0 ! 0000001                                         !      7
    1 ! 33                                              !      2
    2 ! 3                                               !      1
    3 ! 6                                               !      1
    4 !                                                 !      0
    5 ! 0                                               !      1
-------------------------------------------------------------------
TOTAL !                                                 !     20
```

FIGURE 7 – Polissage d'un tableau par médiane : estimation des effets et branchage des résidus




```
EQUATION DE LA RESISTANT LINE

_res_  =  -0.621122  +   0.790066  ( _diag_ -        0 )

HALF-SLOPE RATIOS:
HSR  =  1.3689188
HSR1 =  0.7305035
```

FIGURE 8 – Graphique de diagnostic du polissage d'un tableau par médiane et ligne résistante associée.

```
%qqplot(DATA=,VAR=,LOI=,METH=,POWER=,PARA=,REG=,IC=,
        CQQ=,CREG=,CLIM=);
```

Elle a donc 11 paramètres.

DATA : Nom de la table SAS où figure la variable à analyser. Par défaut la dernière table créée (_LAST_).

VAR : Variable à traiter. Elle doit être numérique. Ce paramètre est obligatoire.

LOI : Désigne la famille de lois à laquelle on veut comparer la distribution empirique. Les valeurs possibles sont :

1 : loi normale (le défaut),

2 : loi uniforme,

3 : loi gamma (de paramètre PARA),

4 : loi du chi2 (PARA degrés de liberté),

5 : loi exponentielle à 1 paramètre,

6 : loi exponentielle à 2 paramètres,

7 : loi de gumbel.

METH : Il y a plusieurs méthodes possibles pour estimer la fonction de répartition empirique. Sont disponibles ici les méthodes de :

1 : TUKEY (le défaut),

2 : HAZEN,

3 : CHEGODAIEV.

POWER : Nombre réel quelconque qui permet de transformer la variable de départ par une fonction puissance. 0 pour le logarithme népérien (cas d'une loi lognormale).

PARA : Pour les lois Gamma et Chi2, il est nécessaire de donner un paramètre supplémentaire : paramètre de forme pour la loi Gamma, nombre de degrés de liberté pour la loi du Chi2.

REG : Afin de juger du caractère "droit" de la courbe dessinée, on peut demander une droite de régression. La droite calculée est celle une droite voisine de la ligne résistante. Ce paramètre est par défaut pris égal à NON.

IC : Enfin, il est possible de demander le tracé de deux "lignes de confiance" qui permettent de tenir compte du caractère aléatoire des estimateurs des quantiles empiriques. Il est alors souhaitable que la courbe soit comprise entre ces deux lignes pour être considérée comme "droite". Ce paramètre est par défaut égal à NON.

CQQ : Couleur du QQplot dans le graphique (défaut BLACK).

CREG : Couleur de la ligne de régression (défaut RED).

CLIM : Couleur des limites de confiance (défaut BLUE).

Pour ces 3 derniers paramètres, choisissez des valeurs acceptables par SAS.

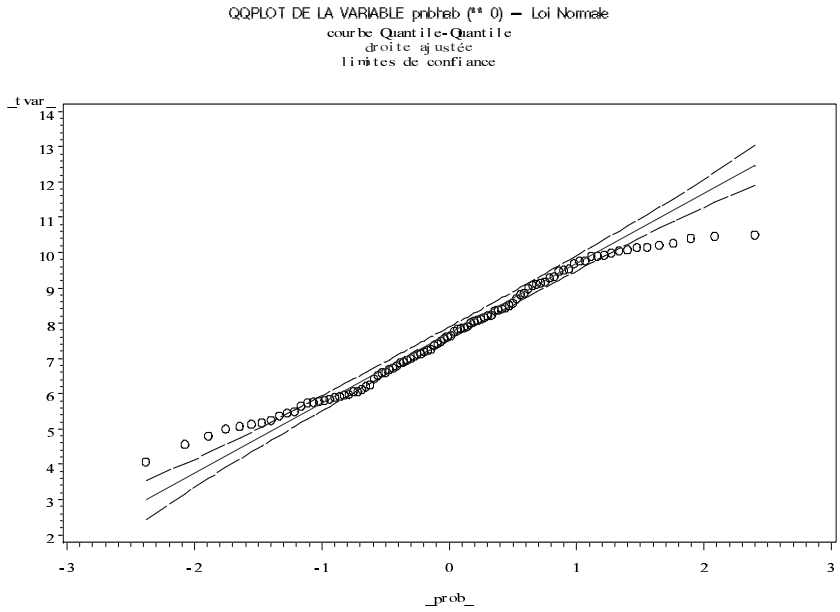


FIGURE 9 – Un exemple de QQ Plot : la variable PNBHAB suit-elle une loi lognormale ?

1.5.3 Un exemple

On s'intéresse par exemple à la variable PNB par habitant (PNBHAB) et on cherche si elle suit une loi lognormale. L'instruction suivante construit le graphique Quantile-Quantile approprié, avec les diagnostics associés.

```
%qqplot(DATA=monde,VAR=pnbhab,LOI=1,METH=,POWER=0,PARA=,
        REG=oui,IC=oui,CQQ=,CREG=,CLIM=);
```

Le graphique Quantile-Quantile est présenté à la figure 9. On y remarque immédiatement le comportement "atypique" des extrémités.

1.6 %QQPLOTMAT : Matrice de QQ-Plots

1.6.1 Brève définition

La macro %QQPLOTMAT compare les distributions de plusieurs variables en traçant les graphiques quantile-quantile (QQ Plots) empiriques : on représente en abscisse les quantiles de l'une des variables et en ordonnée les quantiles de l'autre variable. L'ensemble des QQPlots est alors regroupé dans une matrice.

1.6.2 Paramètres et mise en oeuvre

La macro %QQPLOTMAT utilise les modules SAS/BASE et SAS/INSIGHT. Elle s'appelle par l'instruction générale suivante :

```
%qqplotmat(DATA=,VARX=,VARY=,PAS=,POWER=);
```

Elle a donc 5 paramètres.

DATA : Nom de la table SAS où figure les variables à comparer. Par défaut la dernière table créée (_LAST_).

VARX : Premier ensemble de variables. Elles doivent être numériques et appartenir à la table DATA. Par défaut, toutes les variables numériques sont utilisées.

VARY : Second ensemble de variables. Elles doivent être numériques et appartenir à la table DATA. Par défaut, VARY=&varx.

PAS : Pour tracer les graphiques, on doit calculer les quantiles. Ce paramètre précise ceux que l'on calcule. Par défaut PAS=5 et, dans ce cas, on calcule les quantiles d'ordre 0 5 10 15 20 ..... 90 95 et 100.

POWER : Vous pouvez ici préciser une transformation puissance que vous voulez appliquer à toutes les variables avant de faire le graphique. Si POWER=0, on prend le logarithme des variables.

1.6.3 Un exemple

On veut par exemple comparer les distributions des variables ESPER96 (Espérance de vie) ESPER_F (Espérance de vie des femmes) ESPER_H (Espérance de vie des hommes) MORTINF (taux de mortaité infantile) NAT (taux de natalité) PNBHAB (PNB par habitant). L'instruction suivante construit la matrice des QQ plots :

```
%qqplotmat(DATA=monde,
           VARX=ESPER96 ESPER_F ESPER_H MORTINF NAT PNBHAB,
           VARY=ESPER96 ESPER_F ESPER_H MORTINF NAT PNBHAB);
```

Cette matrice est présentée à la figure 10. On y remarque immédiatement la similitude des lois des variables liées à l'espérance de vie. Comme la matrice est éditée par SAS/INSIGHT, on peut alors utiliser les transformations interactives des variables pour affiner le diagnostic.

1.7 %RESLINE : Ligne résistante (Resistant Line)

1.7.1 Brève définition

La macro %RESLINE permet de faire une régression linéaire simple robuste, de la forme $Y = a(X - m) + b$ où m est un paramètre de centralité de la variable

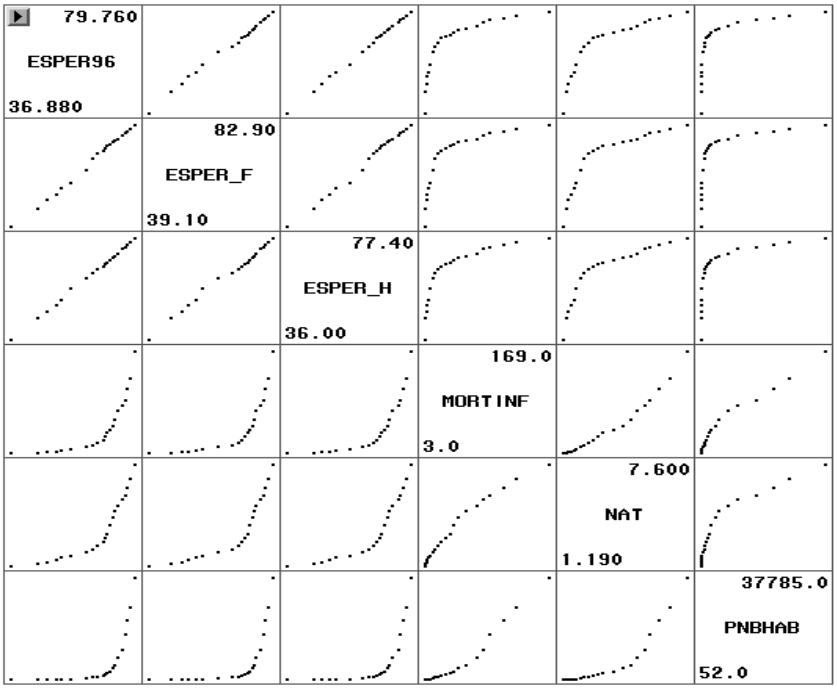


FIGURE 10 – Un exemple de matrice de QQ Plots.

X (médiane). Elle calcule les paramètres de la droite résistante, ajuste la droite au nuage de points et produit plusieurs diagnostics permettant de juger de la qualité de l'ajustement (les demi-pentes) et de savoir quelle transformation de la variable à expliquer pourrait, le cas échéant, conduire à la linéarité.

Pour une définition précise de la droite et des diagnostics, on pourra consulter Hoaglin et Velleman ([19]) ou bien Hoaglin, Mosteller et Tukey ([17]). Dans ces livres figure en outre l'algorithme de calcul.

1.7.2 Paramètres et mise en oeuvre

```
%resline(DATA=,XX=,YY=,DIAGN=,OUT=,OUTDIAG=,AJUST=,
         RESID=,NITER=,GRAPH=);
```

La macro %RESLINE utilise les modules SAS/BASE et SAS/IML ; elle fait appel à SAS/GRAPH pour les graphiques de diagnostics et utilise aussi la macro %STEM (voir paragraphes 1.9). Elle a 10 paramètres.

DATA : Nom de la table SAS où figurent les variables à traiter Par défaut, la dernière table SAS est utilisée (_LAST_).

XX : Nom de la variable explicative. Cette variable doit être numérique. Ce paramètre est obligatoire.

YY : Nom de la variable à expliquer. Cette variable doit être numérique. Ce paramètre est obligatoire.

DIAGN : Permet de demander des diagnostics. Dans ce cas (DIAG=oui) un branchage des résidus est édité avec une indication sur une transformation puissance possible de la variable à expliquer pour améliorer la relation. Par défaut DIAG=non.

OUT : Nom de la table SAS dans laquelle vous souhaitez récupérer les résultats. Cette table contient les variables explicative et à expliquer, la variable ajustée et le résidu. Par défaut, le nom de la table est _result_.

OUTDIAG : Nom de la table SAS dans laquelle vous souhaitez récupérer les résultats pour faire le graphique de diagnostics. Par défaut, le nom de cette table est _diag_.

AJUST : Nom de la variable SAS de la table en sortie contenant la variable ajustée. Par défaut _fit_.

RESID : Nom de la variable SAS de la table en sortie contenant la variable résidu. Par défaut _resid_

NITER : Nombre maximum d'itérations pour calculer les paramètres de la droite. Par défaut 5.

GRAPH : Paramètre permettant d'éditer les graphiques. Ces graphiques utilisent SAS/GRAPH. Par défaut GRAPH=oui (autre valeur possible : non).

1.7.3 Exemple

On étudie la relation qui pourrait exister entre le taux de croissace du PNB (variable CPNB96) et taux de croissance de la population urbaine (variable TXURB). L'instruction suivante demande d'estimer une ligne résistante et de produire les différents diagnostics et graphiques :

```
%resline(DATA=monde,XX=txurb,YY=cpnb96,DIAGN=oui,GRAPH=oui);
```

La figure 11 montre le nuage de points, l'équation de la ligne résistante et les valeurs des demi-pentes (Half Slope Ratios).

La figure 12 présente un branchage des résidus et le nuage de points de ces mêmes résidus en fonction de la variable explicative (TXURB).

Enfin, la figure 13 présente le graphique de diagnostic : l'équation de la ligne résistante donne une indication sur la transformation puissance de la variable à expliquer (CPNB96) qui pourrait éventuellement amener à la linéarité.

1.8 %SQCOMB : Polissage d'un tableau par combinaisons croisées (Square Combining Table)

1.8.1 Brève définition

La macro %SQCOMB, comme la macro %MPOLISH (voir paragraphe 1.4), ajuste un modèle additif à un tableau croisé qui donne les valeurs d'une variable de

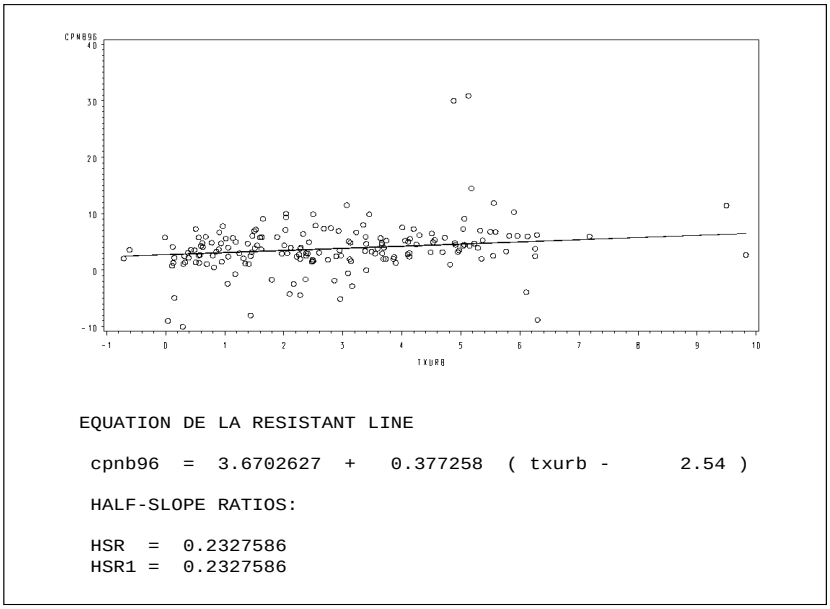


FIGURE 11 – Nuage de points des variables TXURB et CPNB96 et équation de la ligne résistante.

réponse Y en fonction d'un facteur ligne A et d'un facteur colonne B.

Le modèle s'écrit de façon générale sous la forme : $Y_{ij} = m + A_i + B_j + \epsilon_{ij}$.

La macro %SQCOMB estime les différents paramètres et fournit certains diagnostics. Ceux ci permettent non seulement de juger de l'adéquation du modèle mais aussi de tester la pertinence d'un modèle à effets croisés. Pour en savoir plus, on peut consulter l'ouvrage collectif de Hoaglin, Mosteller et Tukey ([18], chapitre 2 de Katherine Godfrey).

1.8.2 Paramètres et mise en oeuvre

La macro %SQCOMB utilise les modules SAS/BASE et SAS/IML. Le module SAS/GRAPH est en outre nécessaire aux graphiques de diagnostic. Elle fait aussi appel aux macros %STEM (voir paragraphe 1.9) et %RESLINE (voir paragraphe 1.7). Elle s'appelle par l'instruction générale suivante :

```
%sqcomb(DATA=,COLS=,LIGNES=,DIAGN=, FORMAT=,
        POWER=,CRES=,CREG=);
```

Elle a donc 9 paramètres.

DATA : Nom de la table SAS où se trouve le tableau à analyser. Les colonnes du tableau sont des variables numériques de la table SAS.

BRANCHAGE DE LA VARIABLE _RESID_

la plus petite valeur représentée est : -7.972175597
la plus grande valeur représentée est : 7.6397905679
la médiane de la distribution est : 0.1313688247
7 !6 représente 7.6397905679 (unité : 1)

STEM	LEAVES	EFFECTIF
LOW	30 170 28 107 48 138	6
-8	0	1
-7	77	2
-6	70	2
-5	6521	4
-4	40	2
-3	975	3
-2	976533211111000	15
-1	999998766665554433332211111	28
-0	887777766665544433333221111	27
0	022223444556666777778999	23
1	000111122223334456789999	25
2	01345556788899	14
3	0114677779	10
4	02457	5
5	13899	5
6	25	2
7	16	2
HIGH	102 24 62	3
TOTAL		179

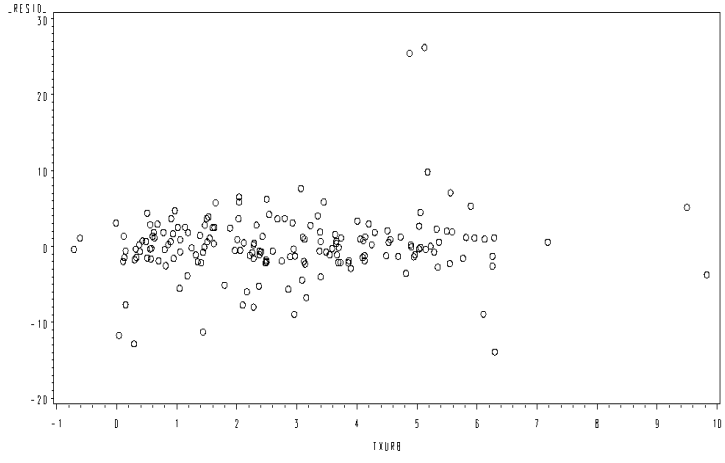


FIGURE 12 – Analyse des résidus : branchage et représentation des résidus en fonction de la variable explicative TXRUR.

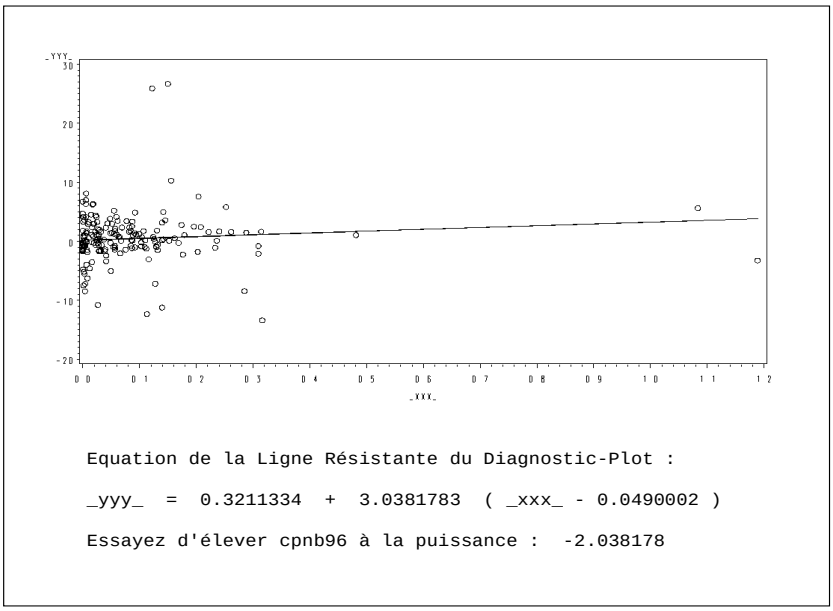


FIGURE 13 – Graphique de diagnostic et suggestion de transformation.

COLS : Nom des colonnes du tableau. Ce sont nécessairement des variables numériques de la table précisée ci-dessus. Par défaut, on prend toutes les variables numériques.

LIGNES : Variable identifiant des lignes du tableau. Par défaut, on prendra LIG1, LIG2, ..., LIGn.

DIAGN : Permet de demander des diagnostics. Dans le cas où ce paramètre est renseigné, un branchage des résidus et un "diagnostic plot" sont édités. Ces diagnostics aident à statuer sur le caractère additif du modèle (transformation suggérée) et la présence d'un effet croisé. Indiquer oui ou non. Par défaut pas de diagnostic DIAGN=non.

FORMAT : Format SAS numérique valide utilisé pour l'impression des résultats. Par défaut FORMAT=5.1

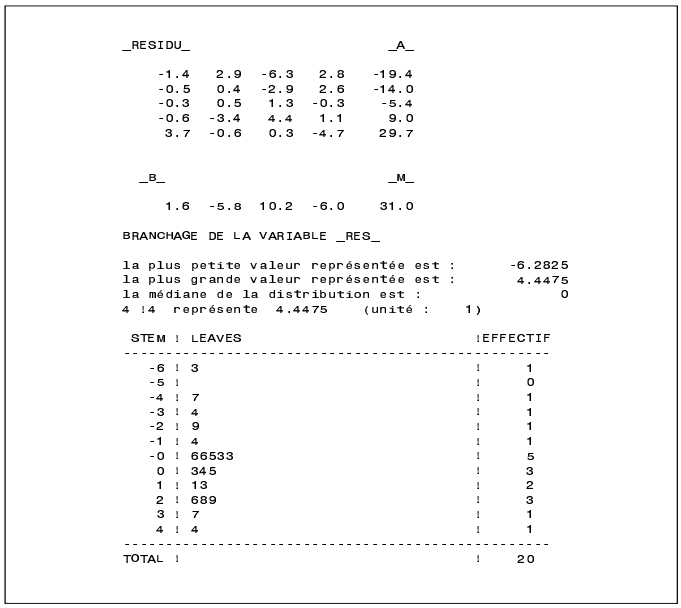
POWER : Puissance de la transformation sur la variable de réponse. Par défaut, pas de transformation : POWER=1.

CRES : Couleur du nuage des résidus. Par défaut RED.

CREG : Couleur de la ligne résistante. Par défaut BLACK.
Pour ces 2 derniers paramètres, choisissez des valeurs acceptables par SAS.

1.8.3 Un exemple

L'exemple est le même que celui présenté au paragraphe consacré au polissage par médiane (voir paragraphe 1.4). Soit la table SAS créée par le programme



```
_RESIDU_                        _A_

    -1.4   2.9  -6.3   2.8   -19.4
    -0.5   0.4  -2.9   2.6   -14.0
    -0.3   0.5   1.3  -0.3    -5.4
    -0.6  -3.4   4.4   1.1     9.0
     3.7  -0.6   0.3  -4.7    29.7


  _B_                           _M_

     1.6  -5.8  10.2  -6.0    31.0

BRANCHAGE DE LA VARIABLE _RES_

la plus petite valeur représentée est :       -6.2825
la plus grande valeur représentée est :        4.4475
la médiane de la distribution est :                 0
4 !4  représente  4.4475    (unité :    1)

 STEM ! LEAVES                          !EFFECTIF
-------------------------------------------------
   -6 ! 3                               !     1
   -5 !                                 !     0
   -4 ! 7                               !     1
   -3 ! 4                               !     1
   -2 ! 9                               !     1
   -1 ! 4                               !     1
   -0 ! 66533                           !     5
    0 ! 345                             !     3
    1 ! 13                              !     2
    2 ! 689                             !     3
    3 ! 7                               !     1
    4 ! 4                               !     1
-------------------------------------------------
TOTAL !                                 !    20
```

FIGURE 14 – Polissage de tableau par combinaisons croisées : estimation des effets et branchage des résidus.

suivant :

```
DATA a;
 INPUT b1-b4;
 CARDS;
11.7  8.7 15.4  8.4
18.1 11.7 24.3 13.6
26.9 20.3 37.0 19.3
41.0 30.9 54.6 35.1
66.0 54.3 71.1 50.0
;
RUN;
```

Les relations entre lignes et colonnes du tableau sont analysées, avec les options par défaut, par l'instruction suivante :

```
%sqcomb(DATA=a,DIAGN=oui);
```

La figure 14 présente le résultat du lissage, les estimations des différents effets, et le branchage des résidus. La figure 15 montre le graphique de diagnostic. La structure linéaire du nuage de point suggère un effet croisé.

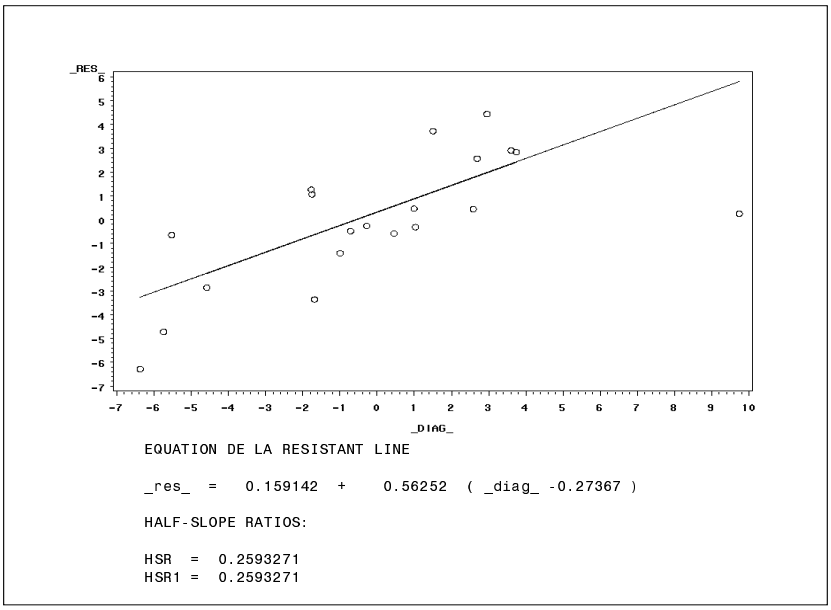


FIGURE 15 – Graphique de diagnostic du polissage de tableau par combinaisons croisées et ligne résistante associée.

1.9 %STEM : Branchage (Stem & Leaf Display)

1.9.1 Brève définition

Le branchage (BàF, Branche à Feuilles), cousin de l'histogramme, est une représentation graphique d'une variable numérique construite à partir des données ordonnées et permettant de nombreuses analyses. Ce graphique est disponible dans la PROC UNIVARIATE, mais sous une forme très rudimentaire. Le branchage, un graphique de base de l'EDA, est défini en général dans tout livre traitant de l'Analyse Exploratoire. La plupart sont malheureusement en anglais et on pourra par exemple consulter Hoaglin et Velleman ([19]), Hoaglin, Mosteller et Tukey ([17]) ou bien même Tukey ([29]).

La macro %STEM permet de construire plusieurs variantes de branchages.

1.9.2 Paramètres et mise en oeuvre

```
%stem(DATA=,VAR=,ID=,IDEXTR=,FORM=,BORNE=,COLS=,
      MAXLIG=,MINLIG=,NGROUP=);
```

La macro %STEM utilise les modules SAS/BASE et SAS/IML. Elle a 10 paramètres.

DATA : Nom de la table SAS où figurent les variables à représenter.

VAR : Noms des variables à traiter. Ces variables doivent être numériques. Si ce paramètre est à blanc, toutes les variables numériques de la table seront analysées.

ID : Variable identifiant. Si ce paramètre est renseigné le graphique représentera ces identifiants et non les valeurs de la variable. Le nombre de positions utilisées dépend de l'option FORMAT. Si ID est numérique, il peut y avoir "reformattage".

IDEXTR : Variable identifiant qui servira à représenter les points "lointains" uniquement, les autres étant représentés par la valeur de la variable (ou ID). Le nombre de positions utilisées dépend de l'option FORMAT. Si IDEXTR est numérique, il peut y avoir "reformattage".

FORM : Nombre de caractères utilisés pour éditer les valeurs de ID ou IDEXTR.

BORNE : Nombre réel positif qui permet de juger du caractère "lointain" d'une valeur. Est lointain un individu dont la valeur de la variable est à plus de b intervalles Inter-Quartiles du premier ou du dernier quartile. Par défaut 2.

COLS : Nombre indiquant la largeur (nombre de colonnes) du graphique. Par défaut 90.

MAXLIG : Nombre maximum de branches du graphique. Si on en a plus, on prend une unité plus grande.

MINLIG : Nombre minimum de branches du graphique. Si on en a moins, les lignes sont dédoublées.

NGROUP : Nombre de sous-divisions souhaitées de chaque branche. Les valeurs permises sont "blanc", 1, 2 ou 5. Si blanc, alors on gère par défaut : le nombre de sous-divisions dépendra du nombre de lignes du graphique (MAXLIG et MINLIG).

1.9.3 Quelques exemples

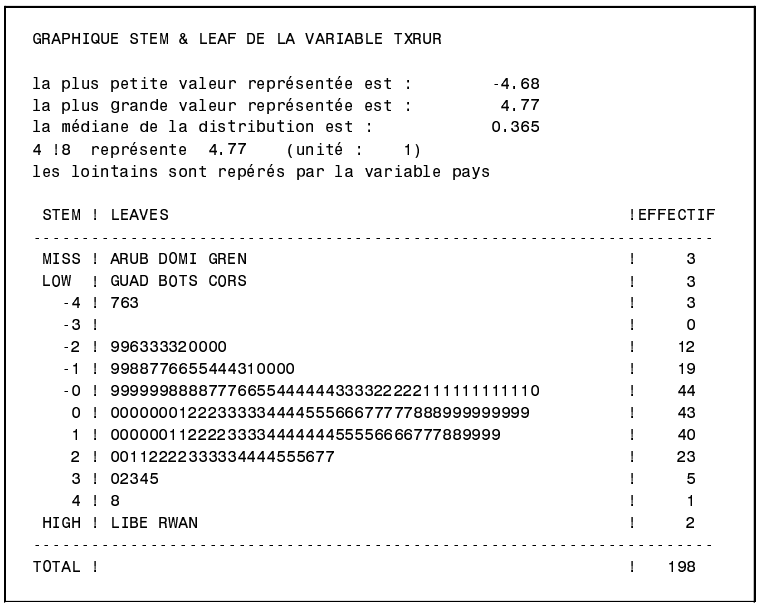
Pour ces exemples, nous utiliserons quelques variables socio-démographiques de 198 pays regroupées dans la table SAS "monde".

Un exemple de branchage simple

L'instruction suivante demande le branchage de la variable taux de croissance de la population rurale (variable TXRUR) :

```
%stem(DATA=monde,VAR=txrur,IDEXTR=pays,COLS=70);
```

Les valeurs "lointaines" sont repérées par le nom du pays (variable PAYS) ; le branchage est représenté dans la figure 16. Comme on peut le noter, outre quelques statistiques simples (minimum, maximum et médiane de la variable), la macro met en évidence les valeurs manquantes (pays ARUB, DOMI, GREN) ainsi que les points lointains bas (pays GUAD, BOTS, CORS) et hauts (pays LIBE, RWAN).



```
GRAPHIQUE STEM & LEAF DE LA VARIABLE TXRUR

la plus petite valeur représentée est :        -4.68
la plus grande valeur représentée est :         4.77
la médiane de la distribution est :            0.365
4 !8  représente  4.77    (unité :    1)
les lointains sont repérés par la variable pays

 STEM ! LEAVES                                                    !EFFECTIF
-------------------------------------------------------------------------
 MISS ! ARUB DOMI GREN                                            !     3
 LOW  ! GUAD BOTS CORS                                            !     3
   -4 ! 763                                                       !     3
   -3 !                                                           !     0
   -2 ! 996333320000                                              !    12
   -1 ! 9988776655444310000                                       !    19
   -0 ! 99999988887776655444444333322222111111111110             !    44
    0 ! 000000012223333344445556667777788899999999               !    43
    1 ! 00000011222233334444445555666677788999                   !    40
    2 ! 00112222333334444555677                                  !    23
    3 ! 02345                                                    !     5
    4 ! 8                                                        !     1
 HIGH ! LIBE RWAN                                                !     2
-------------------------------------------------------------------------
TOTAL !                                                          !   198
```

FIGURE 16 – Exemple simple de branchage par la macro %STEM : variable taux de croissance de la population rurale (TXRUR).

Visualisation d'une autre variable

L'instruction suivante demande le branchage de la variable espérance de vie des hommes (variable ESPER_H) ; les pays sont repérés par le code du continent :

```
%stem(DATA=monde,VAR=esper_h,IDEXTR=pays,ID=conti,COLS=70);
```

Le branchage est représenté dans la figure 17. Les feuilles sont les codes des continents et il apparaît nettement que le continent 2 (l'Afrique) est associé aux espérances de vie les plus faibles.

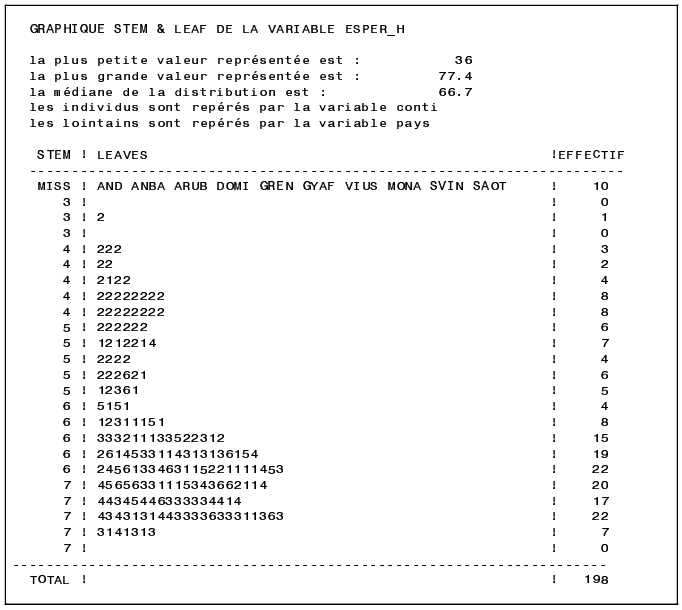
1.10 %SUPERSM : Lissage non paramétrique (Super Smoother)

1.10.1 Brève définition

Le "Super Smoother" est un lisseur non paramétrique, mis au point par Friedman, qui permet d'estimer numériquement des modèles de la forme :

$Y = f(X_1, X_2, \ldots, X_n) + \epsilon$.

SAS propose une méthode semblable dans sa PROC LOWESS ([28]). La macro %SUPERSM permet d'appliquer cette méthode pour étudier la liaison entre deux variables. Pour en savoir plus sur ce lisseur, on pourra par exemple consulter le livre de Härdle ([16]).



```
GRAPHIQUE STEM & LEAF DE LA VARIABLE ESPER_H

la plus petite valeur représentée est :          36
la plus grande valeur représentée est :        77.4
la médiane de la distribution est :            66.7
les individus sont repérés par la variable conti
les lointains sont repérés par la variable pays

 STEM ! LEAVES                                                    !EFFECTIF
-------------------------------------------------------------------------
 MISS ! AND ANBA ARUB DOMI GREN GYAF VIUS MONA SVIN SAOT          !    10
    3 !                                                           !     0
    3 ! 2                                                         !     1
    3 !                                                           !     0
    4 ! 222                                                       !     3
    4 ! 22                                                        !     2
    4 ! 2122                                                      !     4
    4 ! 22222222                                                  !     8
    4 ! 22222222                                                  !     8
    5 ! 222222                                                    !     6
    5 ! 1212214                                                   !     7
    5 ! 2222                                                      !     4
    5 ! 222621                                                    !     6
    5 ! 12361                                                     !     5
    6 ! 5151                                                      !     4
    6 ! 12311151                                                  !     8
    6 ! 333211133522312                                           !    15
    6 ! 2614533114313136154                                       !    19
    6 ! 245613346311522111145 3                                   !    22
    7 ! 4565633111534366 2114                                     !    20
    7 ! 44345446333334414                                         !    17
    7 ! 43431314433336333113 63                                   !    22
    7 ! 3141313                                                   !     7
    7 !                                                           !     0
-------------------------------------------------------------------------
 TOTAL !                                                          !   198
```

FIGURE 17 – Branchage de la variable espérance de vie des hommes (ESPER_H) : les feuilles sont les codes des continents.

1.10.2 Paramètres et mise en oeuvre

La macro %SUPERSM utilise les modules SAS/BASE, SAS/ETS et, pour l'édition des graphiques, SAS/GRAPH. Elle s'appelle par l'instruction générale suivante :

```
%supersm(DATA=,OUTSM=,VARY=,VARX=,WEIGHT=,EPS=,
         SPAN=,TMW=,ALPHA=,GRAPH=);
```

Elle a donc 10 paramètres.

DATA : Table SAS contenant les variables à analyser.

OUTSM : Table SAS en sortie contenant les variables de départ et le résultat du lissage (SMOOTH).

VARY : Variable numérique à expliquer.

VARX : Variable numérique explicative.

WEIGHT : Variable numérique de poids des observations. Si elle n'existe pas, les observations sont équipondérées.

EPS : Critère numérique d'arrêt de l'algorithme. Si l'accroissement du coefficient linéaire est inférieur à inférieur à EPS, on arrête. Par défaut 0.001.

SPAN : Part (en fenêtre du lisseur. Entre 0 et 1. Si SPAN est égal à 0, le défaut, le programme optimise automatiquement le choix.

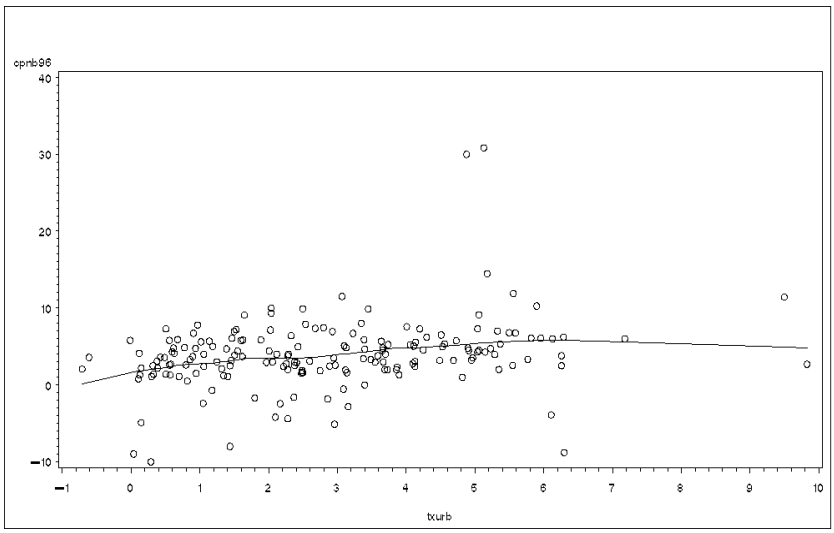


FIGURE 18 – Nuage de points des variables TXURB et CPNB96 et lissage par Super Smoother automatique.

TMW : Ordre des 3 lisseurs que le programme utilise lorsque SPAN=0. Par défaut 0.05 0.2 0.5 :

0.05 correspond au lisseur Tweeter,

0.2 correspond au lisseur Midrange,

0.5 correspond au lisseur Woofer.

ALPHA : Paramètre de pondération utilisé dans la recherche du lissage optimal (SPAN=0) Doit être compris entre 0 et 10.

GRAPH : Option pour l'impression du graphique. Coder oui ou non. Par défaut oui.

1.10.3 Un exemple

La commande suivante demande une analyse automatique de la relation entre le taux de croissace du PNB (variable CPNB96) et taux de croissance de la population urbaine (variable TXURB) :

```
%supersm(DATA=monde,VARX=txurb,VARY=cpnb96);
```

La macro édite le graphique représenté à la figure 18. Cette analyse confirme les résultats obtenus par la macro %RESLINE (voir figure 11).

1.11 %TABCOD : Codage d'un tableau (Coded Display)

1.11.1 Brève définition

Un tableau de chiffres, même de dimensions réduites, est toujours difficile à lire. Les listes codées sont des représentations semi-graphiques d'un tableau de variables. Le principe est très simple. Pour chaque variable X, on situe chaque observation par rapport à la médiane de X, en lui attribuant un code composé de '+' ou de '-' selon qu'elle est au dessus ou en dessous de la médiane. Le nombre de '+' ou de '-' varie en fonction de l'éloignement.

1.11.2 Paramètres et mise en oeuvre

La macro %TABCOD utilise les modules SAS/BASE et SAS/IML. Elle s'appelle par l'instruction générale suivante :

```
%tabcod(DATA=,OUT=,VAR=,IDOBS=,TYPE=,PAS=,NCAT=,
        FUZZ=,PRINT=)
```

Elle a donc 9 paramètres.

DATA : Nom de la table SAS où figurent les variables à analyser. Par défaut la dernière table créée (_LAST_).

OUT : Table SAS en sortie contenant le tableau codé.

VAR : Liste des variables numériques à analyser. Par défaut, toutes les variables numériques sont analysées.

IDOBS : Variable identifiant les observations (lignes). Par défaut, on prend le numéro de l'observation.

TYPE : option permettant de savoir si la distribution de référence est la colonne (type=1), la ligne (type=2) ou la table entière (type=3).

PAS : unité, en nombre d'intervalles inter-quartile, pour la discrétisation des variables. Par défaut un demi IQR.

NCAT : nombre de catégories souhaitées. Choisir un nombre impair. Par défaut 7 : '—' '–' '-' ' ' '+' '++' '+++'.

FUZZ : intervalle dans lequel les observations sont considérées comme égales à la médiane. Par défaut, plus ou moins un demi IQR.

PRINT : Impression de la table des résultats. Si renseigné (valeur quelconque), on imprime.

1.11.3 Un exemple

On considère l'ensemble des pays d'amérique du nord et centrale pour lesquels on analyse les 6 variables suivantes : taux de natalité, taux de chômage, taux de croissance de la population urbaine, densité de population et espérance de vie à la naissance. Le codage de la table s'obtient avec la commande suivante :

NOM	NAT	CHOMAGE	TXURB	PNBHAB	DENS96	ESPER96
Antigua-et-Barbuda	.	.				
Antilles Néerlandaises				+		
Aruba	.	.	.	.		.
Bahamas				+		
Barbade					+	
Belize	+					
Bermudes		.		+++	++++	
Canada				+	-	
Costa Rica			+			
Cuba		.				
Dominique	.	.	.			
El Salvador						
Etats-Unis				++		
Grenade	.	.	.			.
Guadeloupe						
Haïti	++	.	+			---
Honduras	+	-	+			-
Iles Vierges US				.		
Jamaïque						
Martinique		.				
Mexique		-				
Nicaragua	+		+			-
Panama						
Porto Rico				+		
République Dominicaine						
Saint Vincent		.	+			
Saint-Kitts-et-Nevis		.	-			
Sainte Lucie	+	.				
Trinidad et Tobago						

FIGURE 19 – Un exemple de table codée : quelques variables observées sur les pays d'amérique du nord et centrale.

```
%tabcod(DATA=monde,
        VAR=nat chomage txurb pnbhab dens96 esper96,
        IDOBS=nom,OUT=_code_,NCAT=11,
        TYPE=1,FUZZ=0.5,PAS=1,PRINT=oui);
```

La table codée résultant de ce programme est présentée dans la figure 19. On y remarque immédiatement la position atypique des Bermudes (PNB par habitant et densité élevés) et de Haiti (espérance de vie très faible).

1.12 %TWOWAYS : Construction et analyse d'un tableau croisé

1.12.1 Brève définition

La macro TwoWays permet de créér un tableau à deux dimensions en précisant simplement les variables colonne et ligne et les formats, ou regroupements, de ces variables et analyse le tableau par polissage (voir macro %Mpolish).

1.12.2 Paramètres et mise en oeuvre

La macro %TWOWAYS utilise les modules SAS/BASE et SAS/IML et se met en oeuvre par l'instruction générale suivante :

```
%twoways(DATA=,OUT=,VARS=,FORMATS=,GROUPS=,RESPVAR=,STAT=);
```

Elle a 7 paramètres.

DATA : Nom de la table SAS où figurent les variables à analyser. Par défaut la dernière table créée (_LAST_).

VAR : Noms des 2 variables à traiter. Ces variables peuvent être numériques ou caractères.

OUT : Nom de la table SAS qui contiendra le tableau créé et le résultat du polissage. Par défaut OUT=_twoways_.

FORMATS : Formats associés avec les différentes variables. Le premier format est associé à la première variable etc. Si vous ne voulez pas associer de format à une variable, précisez 0. Attention, il est difficile de vérifier des formats. Soyez donc attentif à la syntaxe.

GROUPS : Nombre de groupes requis pour chaque variable. Si 0, il n'y a pas de regroupement (cas d'une variable caractère). La première valeur est associée à la première variable etc. Si un format est précisé pour une variable, il l'emporte sur la valeur de l'option GROUPS mise alors à 0 pour cette variable. Pour une variable caractète, GROUPS=0.

RESPVAR : Nom de la variable de réponse qui est utilisée pour calculer la valeur de chaque case du tableau.

STAT : Nom de la statistique calculée dans chaque case pour la variable de réponse. Le paramètre doit être égal à N, SUM, MEAN ou MEDIAN.

1.12.3 Un exemple

L'instruction suivante demande de créér un tableau croisant le continent (Conti) et la variable taux de natalité (Fertility), cette dernière étant divisée en 5 groupes de taille égale. Dans chaque case, on calcule alors le PNB par habitant (GDPpc) moyen.

```
%TwoWays(DATA=eda.Monde2003,OUT=,VARS=Conti Fertility,
         FORMATS=,GROUPS= 0 5,RESPVAR=GDPpc,STAT=mean);
```

Le tableau initial est représenté à la figure 20. Le résultat du polissage par médianes à la figure 21

1.13 %UPLOTS : Quelques graphiques univariés (Dot Plots)

1.13.1 Brève définition

Cette macro trace, pour une variable numérique, des graphiques exploratoires simples : dot plots, stacked plots, jittered plots. Ces différents graphiques sont définis par exemple dans Wilkinson ([30]).

Fertility	Afrique	Amérique N & C	Amérique du Sud	Asie	Europe	Océanie
.					10220,00	
0		35310,00		37980,00	23011,33	
1	11323,33	18081,67	11815,00	13706,00	27278,18	29840,00
2	6647,50	11135,71	8215,00	12363,57		3530,00
3	4755,00	4864,00	4346,67	4880,00		2818,00
4	2174,12	1150,00		7960,00		

FIGURE 20 – Tableau croisé : PNB par habitant moyen selon le continent et le taux de natalité.

Fertility	Afrique	Amérique N & C	Amérique du Sud	Asie	Europe	Océanie	RowEffect
.					0		798,33
0		0		0	-12189,67		25779,33
1	-2336,75	2411,26	-3338,08	-4634,41	11716,77	16215,59	6139,74
2	-74,51	2403,38	0	961,24		-3156,33	-798,33
3	1901,33	0	0	-2654,00		0	-4666,67
4	74,51	-2959,94		1180,06			-5420,73
ColEffect	-1697,16	313,17	-204,17	2983,17	204,17	-1732,83	9217,50

FIGURE 21 – Polissage par médiane du tableau croisé : PNB par habitant moyen selon le continent et le taux de natalité.

1.13.2 Paramètres et mise en oeuvre

La macro %UPLOTS utilise les modules SAS/BASE et SAS/GRAPH. On la met en oeuvre par l'instruction générale suivante :

```
%uplots(DATA=,VAR=,WORD=);
```

Elle a donc 3 paramètres.

DATA : Nom de la table SAS où figurent les variables à représenter. Par défaut la dernière table créée (_LAST_).

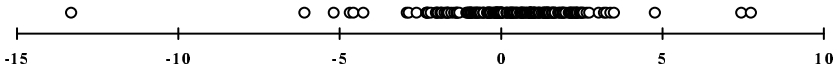
VAR : Noms des variables à représenter. Ces variables doivent être numériques. Si ce paramètre est à blanc, toutes les variables numériques de la table seront représentées.

WORD : Si ce paramètre est renseigné (valeur quelconque), on récupère dans le fichier OUTPUT de SAS les différentes valeurs utiles au tracé des graphiques, séparées par des points virgules et donc facilement utilisables dans WORD ou EXCEL pour de plus jolis graphiques.

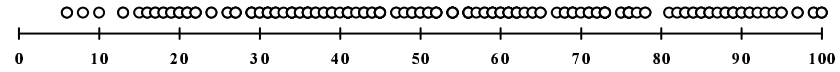
1.13.3 Un exemple global

Cette macro ne pose pas de difficultés de mise en oeuvre. La figure 22 montre quelques exemples de graphiques.

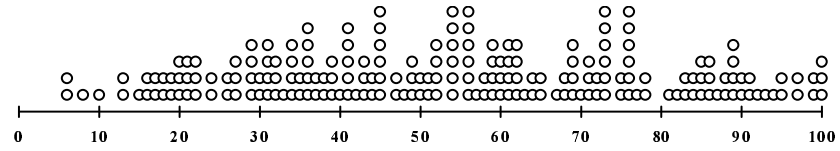
- Dot Plot de la variable TXRUR



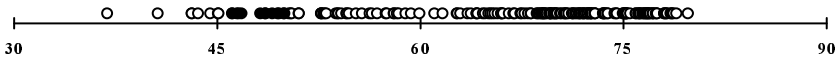
- Dot Plot de la variable POPURB



- Stacked Plot de la variable POPURB



- Dot Plot de la variable ESPER96



- Jittered Plot de la variable ESPER96

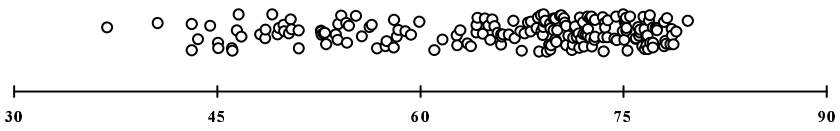


FIGURE 22 – Exemples de graphiques univariés exploratoires.

2 Analyse de Séries Temporelles

Les méthodes d'analyse des séries temporelles sont très présentes dans le logiciel SAS, dans les modules SAS/ETS, SAS/IML et SAS/INSIGHT mais aussi dans les applications "Time Series Forecasting System" et "Time Series Viewer". Les macros présentées ici permettent, soit de mettre en oeuvre des méthodes pour l'instant absentes de SAS, soit d'utiliser plus simplement les outils de SAS.

2.1 %ARIMA : Ajustement automatique de modèles ARIMA non saisonniers

2.1.1 Brève définition

La macro %ARIMA cherche des modèles ARIMA non saisonniers (p, d, q) succeptibles d'ajuster au mieux une série temporelle. Ces modèles sont obtenus par des procédures automatiques de détermination des ordres p et q : les méthodes ESACF et SCAN (disponibles dans la PROC ARIMA, instruction IDENTIFY, voir [25]) mais aussi la méthode ODQ (Hu-Ming, Ping [20]) et la méthode du coin (Beguin, Gouriéroux, Monfort [2]). Pour chaque modèle possible, des statistiques sont éditées qui permettent de poursuivre la modélisation : test de Ljung-Box sur les résidus, coefficients AR et MA etc.

2.1.2 Paramètres et mise en oeuvre

La macro %ARIMA utilise les modules SAS/BASE, SAS/ETS et SAS/IML. Elle s'appelle par l'instruction générale suivante :

```
%arima(DATA=,OUT=,VAR=,DATE=,MAXPQ=,DIFF=,ALPHA=,
       METHOD=,MAXITER=,PRINT=);
```

Elle a donc 10 paramètres.

DATA : Nom de la table SAS où figure la série à analyser. Par défaut la dernière table créée (_LAST_).

OUT : Nom de la table SAS en sortie avec les modèles proposés et leurs caractéristiques. Par défaut OUT=_result_.

VAR : Nom de la variable à modéliser. Elle doit être numérique.

DATE : Nom de la variable date. Ce doit être une date SAS. Attention, ce point est non contrôlable. Cette variable est utilisée pour les tests de LjungBox sur les résidus et pour les prévisions.

MAXPQ : Ce paramètre définit les valeurs possibles maximales des ordres p et q. Par défaut MAXPQ=8.

DIFF : Vous pouvez différencier la série avant de chercher le modèle. DIFF indique l'ordre de différenciation. Les valeurs possibles sont 0 (le défaut), 1, 2 ou 3.

Description	Model1	Model2	Model3	Model4	Model5	Model6	Model7
Arima	(0,1,2)	(0,1,8)	(1,1,2)	(2,1,0)	(4,1,4)	(8,1,6)	(8,1,8)
Selmod	*	*	*	*	*	*	
LjungBox	1	1	1	1	1	1	0
PLjung	0.2380	0.1310	0.3560	0.2850	0.0900	0.2580	0.0170
ToLags	12	12	12	12	12	18	18
SumAR	.	.	-0.342	-0.583	0.233	-2.982	0.674
SumMA	0.476	0.346	0.274	.	0.516	-1.38	0.793
MU	17.153	17.346	17.181	17.257	17.313	17.268	17.364
AR1	.	.	-0.342	-0.307	-0.431	-0.359	1.2
AR2	.	.	.	-0.276	0.605	-0.636	0.331
AR3	.	.	.	.	0.061	-0.236	-0.618
AR4	.	.	.	.	-0.001	-0.41	-0.915
AR5	.	.	.	.	.	-0.024	0.994
AR6	.	.	.	.	.	-0.861	-0.197
AR7	.	.	.	.	.	-0.241	0.128
AR8	.	.	.	.	.	-0.215	-0.249
MA1	0.308	0.331	-0.022	.	-0.106	-0.036	1.523
MA2	0.168	0.201	0.296	.	0.954	-0.311	0.155
MA3	.	-0.089	.	.	-0.13	-0.029	-1.066
MA4	.	-0.022	.	.	-0.201	-0.26	-0.697
MA5	.	0.009	.	.	.	0.128	1.511
MA6	.	-0.05	.	.	.	-0.872	-0.524
MA7	.	-0.007	.	.	.	.	-0.037
MA8	.	-0.026	.	.	.	.	-0.074
Methods	CORNER ESACF SCAN	CORNER	ESACF	CORNER SCAN	ESACF	ESACF	ODQ

FIGURE 23 – Modélisation ARIMA automatique de l'Indice de la Production Industrielle française.

ALPHA : Niveau de significativité des tests utilisés dans les différentes méthodes. Par défaut ALPHA=0.05. Ce paramètre doit être compris strictement entre 0 et 1.

METHOD : Méthode d'estimation utilisée. Les valeurs possibles sont CLS (le défaut), ULS et ML.

MAXITER : Nombre maximal d'itérations pour l'estimation.

Par défaut MAXITER=100.

PRINT : Demande l'impression de la table en sortie avec les modèles et leurs différentes caractéristiques. Coder YES ou NO. Par défaut PRINT=yes.

2.1.3 Un exemple

Cette macro ne pose pas de problème particulier d'utilisation. Les paramètres sont assez simples et la commande suivante analyse l'indice mensuel de la production industrielle français (données désaisonnalisées de 1956 à 1993) :

```
%arima(DATA=ocdem,VAR=fraiip,DATE=date,MAXPQ=8,
       DIFF=1,ALPHA=0.05,OUT=Models);
```

La série est différenciée avant la recherche de modèles. Les modèles proposés par les différentes méthodes sont listés, avec leurs caractéristiques, à la figure 23. D'après ces résultats, les modèles 1, 3 et 4 méritent d'être examinés plus en détail, les autres pouvant être écartés.

2.2 %BKING : Lissage avec le filtre de Baxter-King

2.2.1 Brève définition

La macro %BKING permet de lisser une série temporelle à l'aide d'un filtre "passe-bas" ou d'extraire le cycle de la série à l'aide d'un filtre "passe bande". Les différents filtres sont calculés selon la méthodologie mise au point par Baxter et King (voir [1] ou [13]). Cette macro vous permet de récupérer les valeurs de la série lissée mais aussi les coefficients des filtres utilisés.

2.2.2 Paramètres et mise en oeuvre

La macro %BKING utilise les modules SAS/BASE, SAS/IML et SAS/GRAPH pour les graphiques. Elle s'appelle par l'instruction générale suivante :

```
%bking(DATA=,OUT=,BKOUT=,VAR=,DATE=,LOW=,HIGH=,
       ORDER=,ENDRULE=,PRINT=,WORD=,FORMAT=,
       GRAPH=);
```

Elle a donc 13 paramètres.

DATA : Nom de la table SAS où figurent les séries à analyser. Par défaut la dernière table créée (_LAST_).

VAR : Nom de la variable à lisser. Elle doit être numérique.

DATE : Nom de la variable date. Ce doit être une date SAS. Attention, ce point est non contrôlable. Cette variable est utilisée pour les graphiques et par défaut, le nom est DATE.

LOW, HIGH : Les paramètres LOW et HIGH précisent la bande de fréquences utilisée pour définir le cycle de la série. Elles doivent être exprimées en années. Par exemple LOW=1.5, HIGH=6 : si la série est trimestrielle, cela correspond à une longueur de cycle entre 6 et 24 trimestres. Si LOW et HIGH sont tous les deux renseignés, LOW doit être plus petit que HIGH.

Si un seul des deux paramètres (LOW ou HIGH) est renseigné, le filtre passe-bas associé est calculé, et la série lissée avec ce filtre.

ENDRULE : Ce paramètre précise la stratégie à utiliser pour estimer les valeurs associées aux premiers et derniers points de la série.

0 : on ne fait rien (les valeurs pour les dates les plus récentes ne seront pas estimées),

1 : on utilise des filtres de longueur décroissante. C'est la valeur par défaut.

ORDER : Précise l'ordre des filtres (Baxter-King, Low et High). Ce doit être un entier impair. Si ORDER=$2p+1$, les filtres utiliseront p points dans le passé, le point courant, et p points dans le futur.

OUT : Nom de la table SAS qui, en sortie, contiendra la variable de départ et la (ou les) série(s) lissée(s) demandée(s). Ces nouvelles séries prendront le nom du filtre utilisé : LOW, HIGH et BKING. Par défaut OUT=_result_.

PRINT : Demande l'impression des différentes tables en sortie. Coder YES ou NO. Par défaut PRINT=yes.

WORD : PRINT=YES et WORD=YES, permet l'impression des tables avec les valeurs séparées par des point-virgules, ce qui les rend facilement utilisables dans WORD. Coder YES ou NO. Par défaut WORD=no.

FORMAT : Format numérique pour les impressions. Vous pouvez préciser 2 formats. Le premier sera utilisé pour les coefficients des filtres (défaut : 7.4), le second pour imprimer la série et son cycle (défaut : 10.2). Par exemple : FORMAT= 6.3 8.5.

GRAPH : Permet d'obtenir un graphique de la série brute et de son cycle (ou de la série lissée). Coder YES ou NO. Par défaut GRAPH=yes.

2.2.3 Un exemple

Cette macro ne pose pas de problème particulier d'utilisation. Les paramètres sont assez simples et la commande suivante extrait le cycle de l'indice mensuel de la production industrielle français (données désaisonnalisées de 1956 à 1993) :

```
%bking(DATA=ocdem,OUT=sortie,BKOUT=filters,
       VAR=fraiip,DATE=date,LOW=3,HIGH=6,
       ORDER=71,PRINT=no,GRAPH=yes);
```

Le cycle est défini par les fréquences associées aux périodes de 3 à 6 ans. Le filtre porte sur 71 termes, c'est à dire environ 3 années d'observations de part et d'autre. Les tables SORTIE (série brute et cycle) et FILTERS (coefficients des filtres) sont créées et un graphique édité. La série brute et le cycle associé sont représentés à la figure 24.

2.3 %BNELSON : Décomposition d'une série en tendance et cycle par la méthode de Beveridge-Nelson

2.3.1 Brève définition

La macro %BNELSON permet de décomposer une série temporelle en tendance et cycle. L'idée de base est d'exprimer la série différenciée sous forme d'un processus moyenne mobile infini. Ce processus est alors décomposé en une partie permanente, la tendance, supposée être une marche aléatoire, et une composante cyclique stationnaire. La méthode utilisée est détaillée dans un article de Beveridge et Nelson ([3]). On peut aussi consulter, pour une référence en français, à Doz, Rabault, Sobczak ([11]).

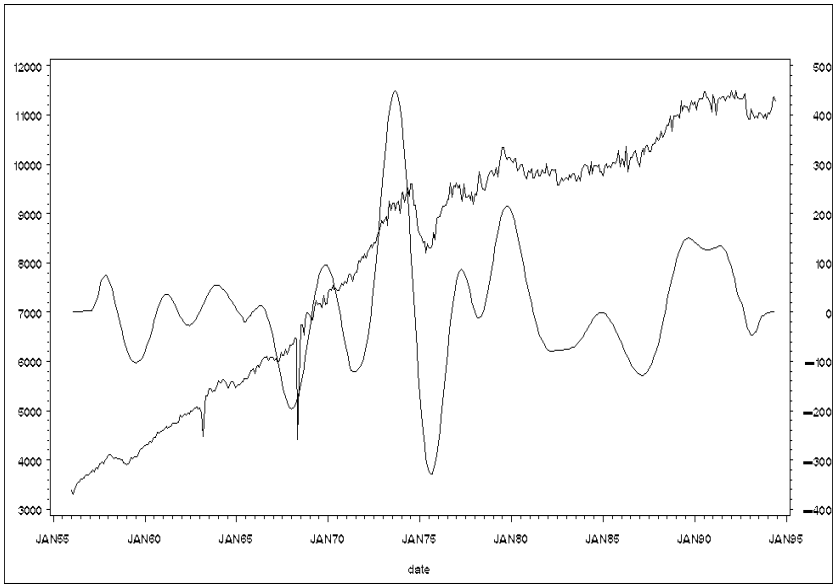


FIGURE 24 – Extraction du cycle de l'Indice de la Production Industrielle française avec le filtre de Baxter-King.

Cette macro vous permet de récupérer les composantes de la série et quelques informations sur la modélisation ARIMA effectuée. Vous pouvez préciser vous même les ordres p et q du modèle ARIMA $(p, 1, q)$ ajusté, ou bien laisser la macro chercher automatiquement un modèle. Dans ce cas, elle utilise les options ESACF et SCAN de la PROC ARIMA (instruction ESTIMATE, voir [25]). Des statistiques simples sont alors calculées : test de Ljung-Box sur les résidus, coefficients des polynômes AR et MA, etc. La décomposition est faite pour tous les modèles semblant valides ; il vous reste à choisir celle qui vous plait !

2.3.2 Paramètres et mise en oeuvre

La macro %BNELSON utilise les modules SAS/BASE, SAS/ETS, SAS/IML et SAS/GRAPH pour les graphiques. Elle s'appelle par l'instruction générale suivante :

```
%BNelson(DATA=,OUT=,VAR=,DATE=,MAXPQ=,PAR=,QMA=,
         METHOD=,MAXITER=,MCOEF=,PRINT=,GRAPH=,
         CRAW=,CTREND=,CCYCLE=);
```

Elle a donc 15 paramètres.

DATA : Nom de la table SAS où figure la série à analyser. Par défaut la dernière table créée (_LAST_).

OUT : Nom de la table SAS qui, en sortie, contiendra la variable de départ et ses composantes. La tendance prendra le nom BN_Trend$n$ et le cycle le nom

BN_cycle$n$, le n faisant référence au modèle ARIMA $(p, 1, q)$ ajusté. Par défaut OUT=_result_.

VAR : Nom de la variable à décomposer. Elle doit être numérique.

DATE : Cette variable est utilisée pour les graphiques et les estimations. Elle doit correspondre à une variable date SAS. Attention, ce point est non contrôlable.

MAXPQ : Si vous demandez une recherche automatique des ordres p et q, cette recherche se fera dans la limite des valeurs entières de 0 à MAXPQ. Si MAXPQ et PAR ou QMA (voir ci après) ne sont pas précisés, alors MAXPQ est pris égal à 10 et on fait une recherche automatique.

PAR : Dans le cas où vous souhaitez vous même préciser le modèle, ce paramètre précise l'ordre du processus autoregressif. PAR doit être positif.

QMA : Dans le cas où vous souhaitez vous même préciser le modèle, ce paramètre précise l'ordre du processus moyenne mobile.

Attention : vous devez préciser à la fois PAR et QMA. Dans ce cas, le modèle ARIMA $(p, 1, q)$ sera ajusté. Dans le cas contraire, un ajustement automatique sera effectué.

METHOD : Méthode d'estimation utilisée. Les valeurs possibles sont CLS (le défaut), ULS et ML.

MAXITER : Nombre maximal d'itérations pour l'estimation.

Par défaut MAXITER=100.

MCOEF : La décomposition de la série différenciée se fait à partir d'un processus moyenne mobile infini. MCOEF précise le nombre "maximal" de coefficients du processus qui l'approxime. Par défaut MCOEF=100.

Attention : la décomposition ne sera faite que pour les modèles satisfaisant le test de Ljung-Box sur les résidus.

PRINT : Demande l'impression de quelques résultats sur la modélisation automatique de la série. Coder YES ou NO. Par défaut PRINT=yes.

GRAPH : Permet d'obtenir un graphique de la série brute, de sa tendance et de son cycle. Coder YES ou NO. Par défaut GRAPH=yes.

CRAW : Couleur de la série brute. Par défaut BLACK.

CTREND : Couleur de la tendance. Par défaut RED.

CCYCLE : Couleur du cycle. Par défaut BLUE.

Pour ces 3 derniers paramètres, choisissez des valeurs acceptables par SAS.

2.3.3 Un exemple

Cette macro ne pose pas de problème particulier d'utilisation. Les paramètres sont assez simples, pour qui connaît assez bien la modélisation ARIMA ! La commande suivante décompose l'indice mensuel de la production industrielle français

Description	Model1	Model2	Model3	Model4	Model5	Model6	Model7	Model8	Model9	Model10	Model11
Arima	(4,1,2)	(6,1,1)	(3,1,4)	(1,1,6)	(2,1,6)	(9,1,0)	(2,1,7)	(1,1,9)	(7,1,6)	(6,1,9)	(10,1,8)
LjungBox	0	0	0	1	0	0	1	1	1	1	0
PLjung	<.0001	0,0280	0,0030	0,0540	0,0290	0,0220	0,1030	0,1950	0,0880	0,0760	0,0120
ToLags	12	12	12	12	12	12	12	12	18	18	24
MU	17,259	17,499	17,357	17,545	17,472	17,583	17,506	17,619	17,475	17,534	17,339
AR1	1,495	-0,083	-0,37	-0,296	-0,421	-0,355	-1,382	-0,884	0,146	0,183	0,158
AR2	-0,322	0,015	-0,091	.	-0,235	-0,083	-0,706	.	-0,835	-1,043	0,666
AR3	-0,063	0,095	0,785	.	.	0,056	.	.	0,814	0,731	-0,358
AR4	-0,174	-0,071	.	.	.	-0,056	.	.	-0,507	-0,759	0,681
AR5	.	0,065	.	.	.	0,057	.	.	0,468	0,542	-0,194
AR6	.	0,164	.	.	.	0,168	.	.	-0,553	-0,593	-0,438
AR7	.	.	.	.	.	0,042	.	.	-0,135	.	0,224
AR8	.	.	.	.	.	0,028	.	.	.	.	0,095
AR9	.	.	.	.	.	0,095	.	.	.	.	0,06
AR10	.	.	.	.	.	.	.	.	.	.	-0,21
MA1	1,879	0,275	-0,024	0,069	-0,055	.	-1,026	-0,534	0,491	0,534	0,49
MA2	-0,923	.	-0,026	0,069	-0,119	.	-0,239	0,282	-0,972	-1,184	0,609
MA3	.	.	0,76	-0,064	0,015	.	0,15	-0,12	1,148	1,085	-0,716
MA4	.	.	-0,297	0,083	0,069	.	-0,006	0,032	-0,84	-1,016	1,025
MA5	.	.	.	-0,118	-0,101	.	-0,034	-0,046	0,615	0,692	-0,542
MA6	.	.	.	-0,188	-0,188	.	-0,278	-0,268	-0,86	-0,825	-0,641
MA7	.	.	.	.	.	.	-0,241	-0,089	.	0,052	0,688
MA8	.	.	.	.	.	.	.	0,02	.	-0,012	-0,125
MA9	.	.	.	.	.	.	.	-0,118	.	-0,117	

FIGURE 25 – Modélisation ARIMA automatique de l'Indice de la Production Industrielle française.

(données désaisonnalisées de 1956 à 1993) avec les options par défaut et, en particulier, une modélisation ARIMA automatique :

```
%bnelson(DATA=ocdem,OUT=sortie,VAR=fraiip);
```

La méthode étant très sensible à la présence de points aberrants, on a corrigé à priori les valeurs atypiques de 1968 (voir figure 24).

La figure 25 montre les caractéristiques des modèles sélectionnés par la procédure automatique. 11 modèles ont été présélectionnés par les options ESACF et SCAN de la procédure ARIMA de SAS. Seuls les modèles 4, 7, 8, 9 et 10 satisfont le test de Ljung-Box de non corrélation des résidus. La figure 26 présente la série brute et les tendance et cycle associés au modèle 8 (1,1,9).

2.4 %GRAPHICS : Graphiques exploratoires pour série temporelle

2.4.1 Brève définition

Cette macro propose un certain nombre de graphiques de base qui permettent de se faire rapidement une idée sur les principales caractéristiques d'une série temporelle **mensuelle** ou **trimestrielle** : stationnarité (en moyenne ou en variance), présence d'une saisonnalité, points atypiques, etc.

Elle présente un certain nombre de graphiques exploratoires pour l'analyse d'une série temporelle. Sont édités les graphiques suivants [3] :

3. Attention : Certains de ces graphiques sont impossibles avant la version 8 (Loess et Boxplots). Dans ce cas, la macro s'adapte automatiquement et propose d'autres graphiques.

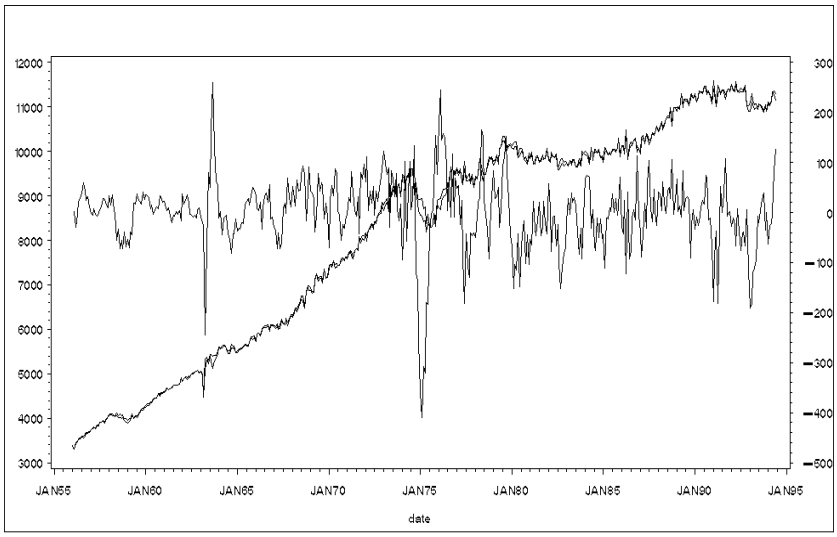


FIGURE 26 – Méthode de Beveridge-Nelson : Tendance et cycle de l'Indice de la Production Industrielle française.

- Graphique de la série brute,
- évolution de chaque année par mois, graphes superposés,
- évolution de chaque année par mois, graphes accolés,
- Box Plots par année (SASV8),
- Série lissée par Loess (SASV8),
- Spectre de la série,
- Autocorrélogramme de la série,
- Autocorrélogramme de la série différenciée (1,0),
- Autocorrélogramme de la série différenciée (0,1),
- Autocorrélogramme de la série différenciée (1,1).

Ces graphiques sont édités séparément et 8 d'entre eux sont regroupés dans deux panels de quatre graphes qui permettent une vision globale des caractéristiques de la série. Sur demande (paramètre PARTINV), les autres fonctions d'autocorrélations sont éditées :

- Autocorrélogramme partiel de la série,
- Autocorrélogramme partiel de la série différenciée (1,0),
- Autocorrélogramme partiel de la série différenciée (0,1),
- Autocorrélogramme partiel de la série différenciée (1,1),
- Autocorrélogramme inverse de la série,
- Autocorrélogramme inverse de la série différenciée (1,0),
- Autocorrélogramme inverse de la série différenciée (0,1),
- Autocorrélogramme inverse de la série différenciée (1,1).

Huit nouveaux graphiques sont édités séparément et aussi regroupés dans deux panels de quatre graphes qui permettent de compléter la vision des caractéristiques

de la série.

2.4.2 Paramètres et mise en oeuvre

La macro %GRAPHICS utilise les modules SAS/BASE, SAS/STAT, SAS/IML et bien entendu SAS/GRAPH. Elle s'appelle par l'instruction générale suivante :

```
%graphics(DATA=,DATE=,VAR=,DIF=,LOG=,PARTINV=,SMOOTH=);
```

Elle a donc 7 paramètres.

DATA : Nom de la table SAS où figurent les séries à analyser. Par défaut la dernière table créée (_LAST_).

VAR : Liste des séries à traiter. Elles doivent être numériques. Par défaut, toutes les variables numériques sont analysées.

DATE : Nom de la variable date. Ce doit être une date SAS. Attention, ce point est non contrôlable. Paramètre obligatoire. Par défaut le nom est DATE.

DIF : Lorsque ce paramètre est renseigné, par une valeur quelconque, le spectre est celui de la série différenciée (ce qui enlève la tendance).

LOG : Lorsque ce paramètre est renseigné, par une valeur quelconque, c'est le logarithme du spectre qui est représenté.

PARTINV : Lorsque ce paramètre est renseigné, par une valeur quelconque, les autocorrélations partielles et inverses sont aussi représentées.

SMOOTH : Paramètre de lissage par Loess (SAS V8). Ce paramètre est commun à toutes les séries. Vous devez spécifier un nombre compris entre 0 et 1. Par défaut SMOOTH=0.25.

2.4.3 Un exemple

Cette macro ne pose pas de problème particulier d'utilisation. Les paramètres sont assez simples et la commande suivante analyse l'indice mensuel de la production industrielle :

```
%graphics(DATA=ipimens,DATE=date,VAR=ipi);
```

Les deux panels de graphiques sont présentés aux figures 27 et 28.

2.5 %HP : Estimation de tendance avec le filtre de Hodrick-Prescott

2.5.1 Brève définition

La macro %HP extrait la tendance d'une série temporelle par la méthode proposée par Hodrick et Prescott. En sortie, vous obtenez une table SAS avec la variable de départ, l'estimation de la tendance et de la série débarrassée de cette tendance.

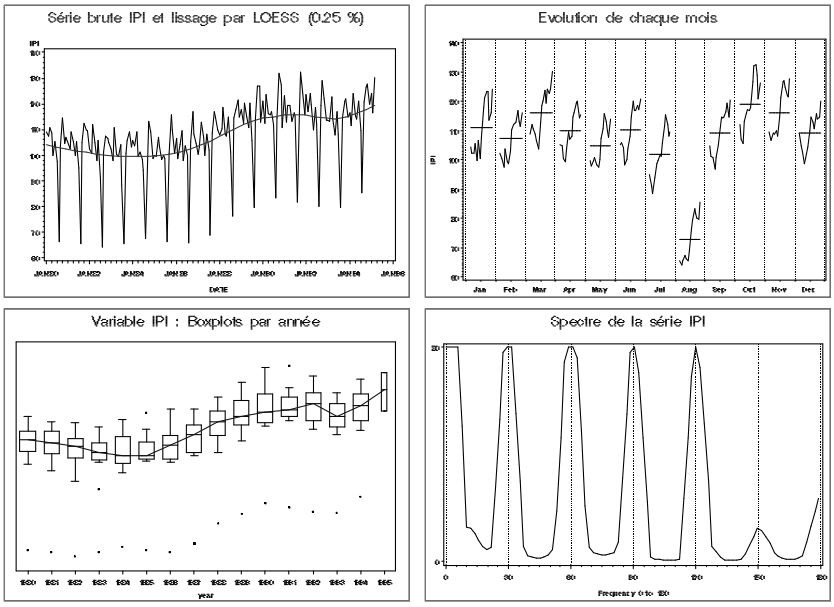


FIGURE 27 – Analyse exploratoire de la série IPI : premier panel (Loess, spectre, etc.).

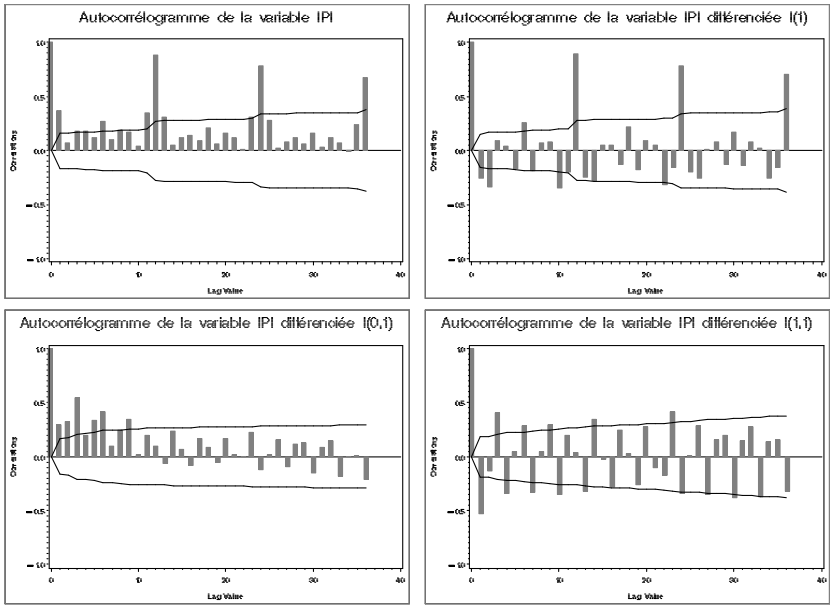


FIGURE 28 – Analyse exploratoire de la série IPI : second panel (autocorrélogrammes).

Cette macro est juste une adaptation d'un programme FORTRAN écrit par Prescott : rien n'a été changé, les instructions ont juste été converties en ordres IML.

Pour en savoir plus sur ce filtre, on peut se reporter à Hodrick, Prescott ([21]) ou, pour une référence en français, à Doz, Rabault, Sobczak ([11]).

2.5.2 Paramètres et mise en oeuvre

La macro %HP utilise les modules SAS/BASE, SAS/IML pour les calculs, et SAS/GRAPH pour le graphique. Elle s'appelle par l'instruction générale suivante :

```
%HP(DATA=,OUT=,VAR=,DATE=,LAMBDA=,PRINT=,WORD=,
    FORMAT=,GRAPH=);
```

Elle a donc 9 paramètres.

DATA : Nom de la table SAS où figure la série à décomposer. Par défaut la dernière table créée (_LAST_).

OUT : Table SAS contenant les résultats du traitement : série brute, tendance (nom de la variable de départ préfixé par HP_) et série débarrassée de sa tendance (nom de la variable préfixé par D_). Par défaut OUT=_result_.

VAR : Série à lisser. Ce doit être une variable numérique.

DATE : Variable Date servant à ordonner les observations. Elle est utilisée pour les graphiques. Attention : ce doit être une variable date SAS et ce point n'est pas contrôlé.

LAMBDA : Paramètre de lissage. Par défaut, il est pris égal à $100p^2$ où p est la périodicité de la série ($p = 4$ ou $p = 12$).

PRINT : Si PRINT est égal à YES (autre valeur NO), la table en sortie est imprimée. Par défaut PRINT=YES.

FORMAT : Format pour les impressions. Par défaut FORMAT=10.2 ; Attention de choisir un format SAS valide.

WORD : Si PRINT=YES et WORD=YES, vous obtenez une impression où les valeurs sont séparées par des points virgule et dont facilement utilisable dans WORD. Codez YES ou NO. Par défaut, WORD=NO.

GRAPH : Si GRAPH est égal à YES (autre valeur NO), on édite deux graphiques : l'un de la série brute et de la tendance estimée et l'autre de la série débarassée de sa tendance.

2.5.3 Un exemple

Cette macro est assez simple et la commande suivante décompose l'indice mensuel de la production industrielle français (données désaisonnalisées de 1956 à 1993) avec les options par défaut :

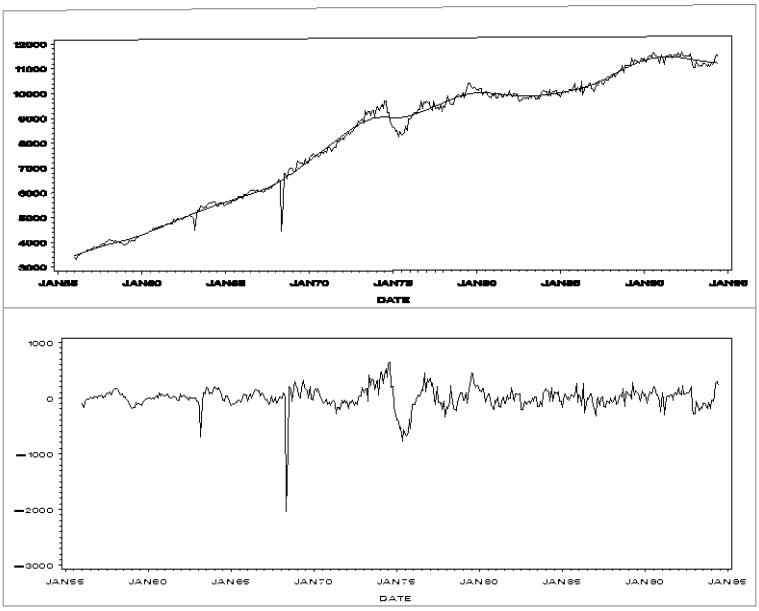


FIGURE 29 – Lissage de la série IPI par le filtre de Hodrick-Prescott : tendance et série débarrassée de sa tendance.

```
%HP(DATA=ocdem,VAR=fraiip,DATE=date,OUT=trend);
```

Les résultats du lissage sont dans une table SAS appelée "trend". La série brute, la tendance et la série débarrassée de sa tendance sont présentées à la figure 29.

2.6 %MEDMOB : Lissage par médianes mobiles (Running Medians)

2.6.1 Brève définition

Les médianes mobiles sont une alternative robuste aux moyennes mobiles. Elles sont, en particulier, peu sensibles à la présence de points atypiques. Pour une définition précise des lisseurs utilisés dans la pratique, on peut se reporter à Tukey ([29]) ou, pour une référence en français, à Ladiray, Roth ([22]).

2.6.2 Paramètres et mise en oeuvre

La macro %MEDMOB utilise les modules SAS/BASE et SAS/IML pour les calculs, et SAS/GRAPH pour le graphique. Elle s'appelle par l'instruction générale suivante :

```
%medmob(DATA=,OUT=,VAR=,DATE=,LISSEUR=,TWICE=,
        INTER=,PRINT=,GRAPH=,CBRUT=,CLIS=);
```

Elle a donc 12 paramètres.

DATA : Nom de la table SAS où figure la série à lisser. Par défaut la dernière table créée (_LAST_).

OUT : Table SAS contenant les résultats du lissage.

VAR : Série à lisser. Ce doit être une variable numérique.

DATE : Variable Date servant à ordonner les observations. Attention : ce doit être une variable date SAS et ce point n'est pas contrôlé.

LISSEUR : Lisseur robuste souhaité. Par exemple 4253h ou 43rsr2h.

TWICE : Si TWICE est égal à OUI (autre valeur NON), le lisseur est "doublé".

INTER : Si INTER est égal à OUI (autre valeur NON), tous les lissages intermédiaires figureront dans la table en sortie.

PRINT : Si PRINT est égal à OUI (autre valeur NON), la table en sortie est imprimée.

GRAPH : Si GRAPH est égal à OUI (autre valeur NON), on édite un graphique de la série brute et de la série lissée.

CBRUT : Couleur de la série brute dans le graphique. Attention de bien choisir une couleur valide.

CLIS : Couleur de la série lissée dans le graphique. Attention de bien choisir une couleur valide.

2.6.3 Un exemple

Cette macro ne pose pas de problème particulier d'utilisation, même s'il n'est pas évident a priori de savoir à quoi un lisseur comme 43RSR2H correspond ! Les paramètres sont assez simples et la commande suivante lisse l'indice mensuel de la production industrielle avec le lisseur 7RJ :

```
%MEDMOB(DATA=ipimens,VAR=ipi,DATE=date,OUT=lissage,
        LISSEUR=7Rj,TWICE=oui,INTER=oui,PRINT=oui,
        GRAPH=oui,CBRUT=black,CLIS=red);
```

Les résultats des lissages intermédiaires sont conservés dans la table en sortie appelée "lissage". Un extrait de cette table est présenté à la figure 30. La série brute et la série lissée sont présentées à la figure 31.

2.7 %MOVAV : Lissage par moyennes mobiles (Moving Averages)

2.7.1 Brève définition

Les moyennes mobiles restent de nos jours à la base de nombreuses méthodes de décomposition de séries temporelles, par exemple la célèbre méthode de désaisonnalisation X-11 (ou X-12). C'est un outil simple, souple et facile à utiliser. Par

date	ipi	L7	L7R	L7RJ	L7RJT
JAN80	109.2	109.2	109.2	109.200	109.200
FEB80	107.4	109.2	109.2	109.200	109.200
MAR80	110.6	108.7	108.7	108.700	108.700
APR80	108.7	107.4	107.4	107.000	106.647
MAY80	99.9	105.1	105.1	106.125	105.642
JUN80	105.1	103.8	104.5	105.463	104.934
JUL80	97.2	103.8	104.5	104.938	104.406
AUG80	66.2	103.8	104.5	104.575	104.086
SEP80	103.8	104.5	104.5	104.500	104.141
OCT80	114.5	104.5	104.5	104.500	104.270
NOV80	104.5	104.5	104.5	104.500	104.358
DEC80	106.8	104.5	104.5	104.500	104.448
JAN81	104.5	105.2	104.5	104.525	104.516
FEB81	102.2	104.5	104.5	104.550	104.534
MAR81	108.9	105.2	104.5	104.575	104.553
APR81	105.2	104.5	104.7	104.625	104.591
MAY81	99.9	102.2	104.7	104.650	104.600
JUN81	105.3	104.7	104.7	104.675	104.616
JUL81	95.0	104.7	104.7	104.700	104.631
AUG81	65.5	104.7	104.7	104.738	104.750
SEP81	104.7	105.3	104.7	104.775	105.064

FIGURE 30 – Lissage de la série IPI par 7RJT : contenu de la table SAS en sortie (extrait).

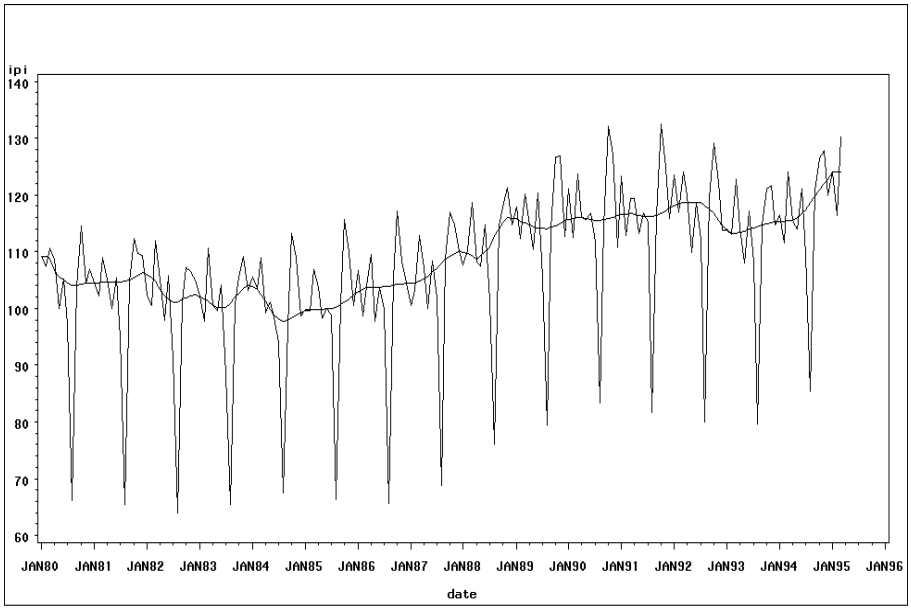


FIGURE 31 – Lissage de la série IPI par 7RJT : série brute et série lissée.

ailleurs, elles sont terriblement efficaces : il est en effet possible de construire une moyenne mobile ayant les caractéristiques souhaitées en matière d'élimination de saisonnalité, de conservation de tendance etc. La macro %MOVAV permet justement de construire toutes sortes de moyennes, avec les moyennes mobiles asymétriques associées et, accessoirement, de lisser des séries avec les moyennes ainsi définies. Ce calcul se fait à partir de la résolution d'un problème de minimisation d'une forme quadratique des coefficients recherchés sous contraintes. Pour en savoir plus sur la construction des moyennes mobiles, on peut se référer à l'article de Grun-Rehomme et Ladiray ([15]).

2.7.2 Paramètres et mise en oeuvre

La macro %MOVAV utilise les modules SAS/BASE et SAS/IML pour les calculs, et SAS/GRAPH pour les graphiques. Elle s'appelle par l'instruction générale suivante :

```
%movav(ORDER=,D=,S=,DS=,MOY=,TABMOY=,CRIT=,ALPHA=,
       ENDRULE=,PRINT=,GAIN=,DATA=,OUT=,XX=,DATE=,
       GRAPH=,CRAW=,CLIS=);
```

Elle a donc 18 paramètres.

ORDER : Ordre de la moyenne mobile. Ce doit être un nombre entier impair $2p + 1$. Dans ce cas, la valeur de la série lissée à l'instant t fait intervenir p points dans le passé, le point courant et p points dans le futur.

D : Degré du polynôme modélisant la tendance (et qui doit donc être conservé). Les valeurs possibles sont 0 (le défaut), 1, 2 ou 3.

S : Saisonnalités que doit éliminer la moyenne mobile. Vous pouvez en préciser plusieurs. Par exemple : S= 5 12. Dans ce cas, la moyenne mobile construite éliminera les saisonnalités de période 12 (mensuelles) et 5. Par défaut S=1.

DS : Les saisonnalités précisées ci dessus peuvent varier, de façon polynomiale, avec le temps. Précisez dans DS les degrés de ces polynômes. Par exemple DS= 0 1 associé à S= 5 12 indique que la saisonnalité d'ordre 5 est constante et que celle d'ordre 12 varie localement linéairement avec le temps.

MOY : Préfixe court identifiant la moyenne mobile. Par défaut MOY=M_. Ce préfixe sera utisé pour construire les noms des diverses moyennes calculées : M_p_f pour une moyenne d'ordre $p + f + 1$ avec p points dans le passé et f points dans le futur.

TABMOY : Table SAS en sortie contenant les coefficients des moyennes mobiles symétrique et asymétriques. Par défaut TABMOY=_tabmoy_.

CRIT : Critère minimisé pour calculer les coefficients de la moyenne mobile. Ce critère peut se mettre sous la forme de la somme des carrés des coefficients différenciés. CRIT est le degré de différenciation. Il peut prendre les valeurs

0 (le défaut), 1, 2 ou 3 (valeur utilisée pour les moyennes mobiles de Henderson).

ALPHA : Valeur de pondération permettant d'obtenir un critère réalisant un compromis entre la fidélité (CRIT=0) et le degré de lissage de la série lissée (CRIT différent de 0). Par exemple ALPHA=1600 et CRIT=2 correspond à une version locale du filtre de Hodrick-Prescott pour une série trimestrielle. ALPHA doit être positif et, par défaut, est égal à 0.

ENDRULE : Ce paramètre précise la stratégie utilisée pour générer les moyennes mobiles asymétriques qui viendront compléter la moyenne mobile calculée. Ces moyennes asymétriques sont utilisées pour lisser les extrémités de la série (début et fin). 4 stratégies sont possibles :

0 : on ne fait rien. Seule la moyenne mobile symétrique d'ordre $2p + 1$ est calculée. Si on lisse une série, les p premiers et derniers points ne seront pas estimés.

1 : Des moyennes mobiles asymétriques, de même ordre que la moyenne symétrique, sont calculées en résolvant le même problème de minimisation (seules les contraintes changent).

2 : Des moyennes mobiles asymétriques de Musgrave sont calculées, comme dans X-11. Ces moyennes dépendent d'une constante définie ici en fonction de la longueur du filtre. C'est la valeur par défaut du paramètre ENDRULE.

3 : Des moyennes mobiles asymétriques de Musgrave sont calculées. La constante de Musgrave est déterminée par des régressions linéaires en début et fin de série. Cette stratégie ne peut être utilisée que s'il y a une série à lisser (paramètre XX). Dans ce cas, la table TABMOY contiendra l'ensemble des moyennes calculées pour chaque série.

GAIN : Option de calcul et d'édition des fonctions de gain et de phase des filtres symétriques et asymétriques générées. Coder YES ou NO. Par défaut GAIN=yes.

PRINT : Option d'impression de la table des coefficients TABMOY. Coder YES ou NO. Par défaut PRINT=yes.

DATA : Nom de la table SAS où figurent les séries à lisser. Par défaut la dernière table créée (_LAST_).

OUT : Table SAS contenant les résultats du lissage. Une série lissée prend le nom de la série brute associée préfixé par la valeur du paramètre MOY.

XX : Séries à lisser. Ce doivent être des variables numériques.

DATE : Variable Date servant à ordonner les observations et utilisée pour les graphiques. Si elle est précisée, ce doit être une variable date SAS et ce point n'est pas contrôlé. Si elle n'est pas précisée et si GRAPH=yes, alors une variable _numero_ est automatiquement créée.

GRAPH : Si GRAPH est égal à YES (autre valeur NO), on édite un graphique, pour chaque série, de la série brute et de la série lissée.

i	M_7_0	M_7_1	M_7_2	M_7_3	M_7_4	M_7_5	M_7_6	M_7_7
-7	-0.0791	-0.0402	-0.0159	-0.0054	-0.0045	-0.0081	-0.0120	-0.0137
-6	-0.0571	-0.0388	-0.0250	-0.0184	-0.0178	-0.0203	-0.0232	-0.0245
-5	-0.0140	-0.0162	-0.0129	-0.0102	-0.0099	-0.0114	-0.0132	-0.0141
-4	0.0569	0.0341	0.0269	0.0258	0.0258	0.0254	0.0245	0.0240
-3	0.1486	0.1052	0.0875	0.0825	0.0823	0.0828	0.0830	0.0829
-2	0.2443	0.1804	0.1522	0.1433	0.1428	0.1444	0.1456	0.1459
-1	0.3249	0.2404	0.2018	0.1890	0.1881	0.1908	0.1930	0.1937
0	0.3754	0.2704	0.2213	0.2046	0.2035	0.2072	0.2103	0.2115
1	0.0000	0.2647	0.2052	0.1846	0.1832	0.1880	0.1921	0.1937
2	0.0000	0.0000	0.1590	0.1346	0.1329	0.1387	0.1439	0.1459
3	0.0000	0.0000	0.0000	0.0694	0.0674	0.0743	0.0805	0.0829
4	0.0000	0.0000	0.0000	0.0000	0.0061	0.0140	0.0211	0.0240
5	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0256	-0.0174	-0.0141
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0282	-0.0245
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0137

FIGURE 32 – Moyenne mobile de Henderson sur 15 termes et moyennes mobiles asymétriques de Musgrave associées.

CRAW : Couleur de la série brute dans le graphique. Défaut BLACK. Attention de bien choisir une couleur valide.

CLIS : Couleur de la série lissée dans le graphique. Défaut RED. Attention de bien choisir une couleur valide.

2.7.3 Un exemple

La commande suivante demande le calcul d'une moyenne mobile de Henderson (CRIT=3) sur 15 termes et des moyennes mobiles asymétriques de Musgrave associées (ENDRULE=2). Ces moyennes, dont on calcule aussi les fonctions de gain et de déphasage, sont ensuite utilisées pour lisser la série GBXFINL de la table OCDEM.

```
%MOVAV(ORDER=15,D=2,S=1,DS=,MOY=,TABMOY=,CRIT=3,
       ENDRULE=2,GAIN=yes,PRINT=yes,DATA=ocdem,
       XX=gbxfinl,DATE=date,GRAPH=yes);
```

La figure 32 montre les coefficients des différentes moyennes mobiles (contenu de la table _tabmoy_). La série brute et la série lissée sont présentées à la figure 33. Les coefficients et fonctions de gain et de déphasage de la moyenne mobile symétrique de Henderson sont présentés à la figure 34, celles de la moyenne mobile asymétrique sur 11 termes (7 dans le passé, 3 dans le futur et le point courant) sont présentés à la figure 35.

FIGURE 33 – Lissage de la série GBXFINL par une moyenne mobile de Henderson sur 15 termes.

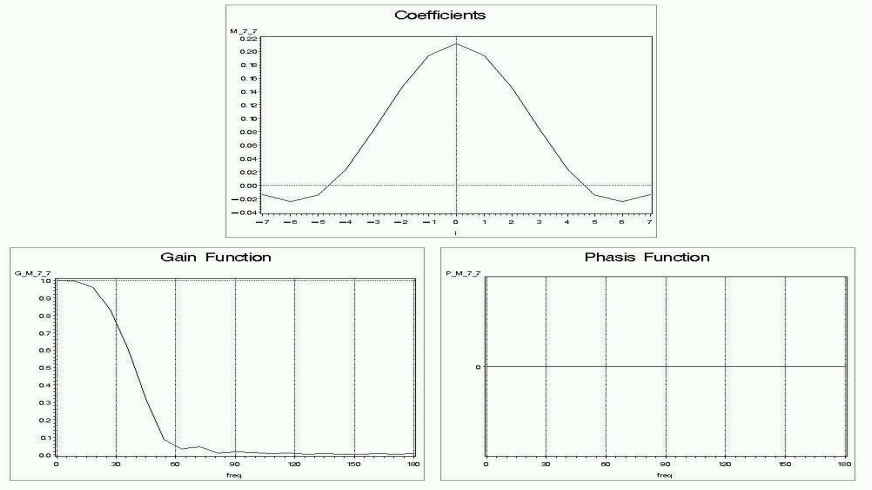


FIGURE 34 – Coefficients, fonctions de gain et de déphasage de la moyenne mobile symétrique de Henderson sur 15 termes.

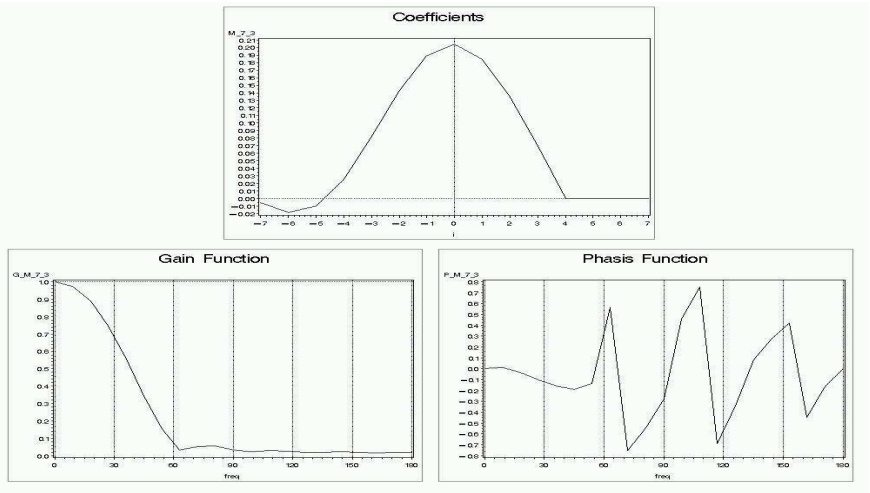


FIGURE 35 – Coefficients, fonctions de gain et de déphasage de la moyenne mobile asymétrique de Henderson sur 11 termes (7 dans le passé et 3 dans le futur).

2.8 %PAT : Estimation de tendance et chronologie des points de retournement (Phase Average Method)

2.8.1 Brève définition

Dans les années 1970, Le NBER (National Bureau for Economic Reasearch) a mis au point une méthode d'analyse du cycle économique basée sur une estimation de la tendance de la série et sur la datation des points de retournement. Cette méthode, assez complexe, est basée sur l'utilisation itérative de moyennes mobiles. Cette méthode est donc une méthode non-paramétrique de la même famille que la méthode de désaisonnalisation X-11 (et X-12).

Par rapport à la méthode originale, la macro %PAT possède beaucoup plus de paramètres : il est ainsi possible d'étudier la robustesse de la méthode en faisant varier des valeurs fixées dans le programme du NBER (ou de l'OCDE).

La littérature sur le sujet est curieusement assez rare pour une méthode encore aujourd'hui utilisée par l'OCDE. On pourra par exemple consulter les articles originaux (Boschan et Bry [4], Boschan et Ebanks [5]) ou Fayolle ([12]) et Doz et Ladiray ([10]).

2.8.2 Paramètres et mise en oeuvre

La macro %PAT utilise les modules SAS/BASE, SAS/IML et SAS/GRAPH. Elle se met en oeuvre par l'instruction générale suivante :

```
%PAT(DATA=,OUT=,XX=,DATE=,NFOR=,ITERM=,LCYCLE=,
     LPHASE=,MOYMOB=,MCD=,LOG=,NMONTH=,IRD=,
```

```
NDTR=,INV=,AMUL=,PRINT=,GRAPH=,CLIS=,
CBRUT=,CTURNP=);
```

Elle a donc 21 paramètres : ce nombre élevé reflète la complexité de l'algorithme.

DATA : Table SAS contenant les données à analyser.

OUT : Table SAS en sortie contenant tous les résultats du traitement de la série analysée. Par défaut OUT=_result_.

XX : Variable numérique à analyser.

DATE : Variable Date servant à ordonner les observations. Attention : ce doit être une variable date SAS et ce point n'est pas contrôlé. Par défaut DATE=date.

NFOR : Périodicité de la série :

1 : mensuelle,

2 : trimestrielle.

ITERM : Ordre de la moyenne mobile simple utilisée pour le tout premier lissage. Par défaut ITERM=75.

LCYCLE : Longueur minimale d'un cycle, exprimée en mois. Par défaut 15.

LPHASE : Longueur minimale d'une phase, exprimée en mois. Par défaut 5.

MOYMOB : Ordre de la moyenne mobile simple utilisée pour lisser la série corrigée des points atypiques et débarassée de sa tendance.

MCD : Ordre de la moyenne mobile utilisée pour déterminée les points de retournement. Si MCD=0, l'ordre est calculé par le programme.

LOG : Utilisation, ou non, de la transformation logarithmique dans l'estimation finale de la tendance :

0 : oui (le défaut),

1 : non.

NMONTH : Nombre de mois utilisé pour déterminer les points de retournement potentiels. Par défaut NMONTH=6.

IRD : Type de modèle de composition utilisé :

0 : multiplicatif (déviation au trend = ratio),

1 : additif (déviation au trend = différence).

Par défaut IRD=0.

NDTR : élimination ou non de la tendance :

0 : on élimine la tendance (le défaut),

1 : on travaille sur les données brutes.

INV : Inversion de la série :

0 : non (le défaut),

1 : oui.

POINTS DE RETOURNEMENT POTENTIELS SUR SERIE CORRIGEE DU TREND

Pics potentiels			Creux potentiels		
IPEAKS	DPEAKS	PEAKS	ITROUGHS	DTROUGHS	TROUGHS
25	JAN58	105.31	37	JAN59	95.37
57	SEP60	102.37	87	MAR63	86.31
97	JAN64	103.68	109	JAN65	97.88
127	JUL66	103.83	149	MAY68	68.39
155	NOV68	104.06	185	MAY71	94.47
224	AUG74	108.68	233	MAY75	93.03
249	SEP76	104.29	264	DEC77	95.89
283	JUL79	105.28	299	NOV80	98.55
312	DEC81	102.41	320	AUG82	97.79
341	MAY84	101.93	349	JAN85	98.49
359	NOV85	102.26	373	JAN87	96.47
415	JUL90	104.68	456	DEC93	93.81

FIGURE 36 – Méthode PAT : points de retournement de la série de l'indice de la production industrielle française.

AMUL : Nombre d'écart-types à partir duquel un point sera considéré comme extrême. Par défaut AMUL=3.5.

PRINT : Permet d'obtenir l'impression de tous les résultats intermédiaires. Coder OUI ou NON. Par défaut PRINT=oui.

GRAPH : Permet d'obtenir un graphique de la série brute, de la tendance et des points de retournement. Coder OUI ou NON. Par défaut GRAPH=oui.

CBRUT : Couleur de la série brute dans le graphique. Attention de bien choisir une couleur valide. Défaut BLACK.

CLIS : Couleur de la série lissée dans le graphique. Attention de bien choisir une couleur valide. Défaut BLUE.

CTURNP : Couleur des points de retournement dans le graphique. Attention de bien choisir une couleur valide. Défaut RED.

2.8.3 Un exemple

L'exemple suivant analyse la série de l'indice de la production industrielle française, sur longue période (de 1956 à 1995), en utilisant les valeurs des paramètres par défaut :

```
%PAT(DATA=base.ocdem,XX=fraiip,DATE=date,NFOR=1);
```

La figure 36 montre la chronologie proposées des points de retournement (pics et creux). La figure 37 représente la série analysée, la tendance extraite par la méthode et les points de retournement.

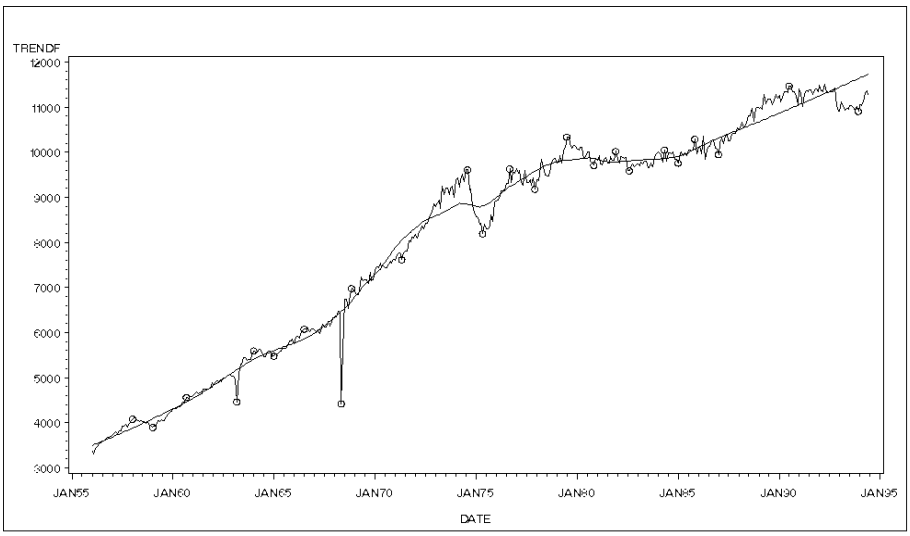


FIGURE 37 – Méthode PAT : tendance et points de retournement de la série de l'indice de la production industrielle française.

2.9 %STATIONARITY : Tests de racine unité

2.9.1 Brève définition

La stationnarité est une hypothèse implicite de nombreuses méthodes d'analyse des séries temporelles. La macro %STATIONARITY présente de façon condensée les résultats des tests de racine unité de Dickey-Fuller et de Phillips-Perron faits sur la série étudiée, le cas échéant avec 3 formes de modèles possibles : pas de terme constant (les variables sont centrées), présence d'un terme constant, présence d'un trend linéaire.

La littérature sur le sujet est abondante (voire excessive !) et on pourra par exemple consulter les différents manuels de SAS ([24], [25], [26]) ou les articles de Dickey, Fuller ([9]), et de Phillips, Perron ([23]).

2.9.2 Paramètres et mise en oeuvre

La macro %STATIONARITY utilise les modules SAS/BASE et SAS/ETS : elle repose essentiellement sur la PROC ARIMA dont elle présente différemment certains résultats. Elle s'appelle par l'instruction générale suivante :

```
%stationarity(DATA=,VAR=,MAXLAGS=,ALPHA=,
              OUTTESTS=,PRINT=,WORD=);
```

Elle a donc 7 paramètres.

DATA : Nom de la table SAS où figure(nt) la (les) série(s) à analyser. Par défaut la dernière table créée (_LAST_).

VAR : Liste des séries à traiter. Elles doivent être numériques. Par défaut, toutes les variables numériques sont analysées.

MAXLAGS : Nombre maximum de retards à prendre en compte dans les tests. Ce paramètre doit être un nombre entier strictement positif. Par défaut MAXLAGS = 8.

ALPHA : Niveau de significativité des tests, exprimé en %. Par défaut ALPHA = 5.

OUTTEST : Table SAS en sortie qui contient le résultat de tous les tests. La valeur par défaut est OUTTEST = _tests_.

PRINT : Ce paramètre vous permet d'imprimer les résultats des différentes procédures ARIMA. Codez YES ou NO. Par défaut PRINT = NO.

WORD : Ce paramètre vous permet d'obtenir un résultat facilement utilisable en WORD. Si vous codez YES, les valeurs sont séparées par un point virgule dans l'impression. Par défaut WORD = NO.

2.9.3 Un exemple

Cette macro ne pose pas de problème particulier d'utilisation. Les paramètres sont assez simples et la commande suivante teste la stationnarité des variables d'une table "sertemp" avec un maximum de 10 retards, et en présentant les résultats sous une forme facilement importable en WORD :

```
%stationarity(DATA=sertemp,VAR=,MAXLAGS=10,WORD=yes);
```

La macro sort deux tables de même format :

- La première contient les valeurs des statistiques de tests pour les différents retards (figure 38).
- La seconde, plus facile à lire, donne le résultat final de ces tests (No : la série n'est pas jugée stationnaire, Yes : la série est jugée stationnaire) sous une forme plus immédiatement interprétable (figure 39). Une variable appelée Stationary résume les résultats (égale à Yes si la variable est jugée stationnaire et à No dans le cas contraire).

On remarque que l'hypothèse de stationnarité est, dans cet exemple, rejetée par le test de Phillips-Perron pour toutes les variables alors que le test de Dickey-Fuller augmenté l'accepte pour toutes les variables ! ! ! Et oui, c'est souvent le cas ! ! !

Variable	Test	Stat	lag0	lag1	lag2	lag3	lag4	lag5	lag6
Eoutput	ADF	Tau	-2.031	-1.508	-1.704	-3.212	-3.537	-3.312	-3.882
		Pr<Tau	0.273	0.528	0.428	0.021	0.008	0.016	0.003
	PP	Tau	-2.031	-1.769	-1.813	-2.081	-2.174	-2.278	-2.421
		Pr<Tau	0.273	0.395	0.373	0.252	0.216	0.180	0.137
Eprices	ADF	Tau	-2.384	-1.628	-1.680	-3.008	-3.714	-3.431	-4.192
		Pr<Tau	0.148	0.467	0.440	0.036	0.005	0.011	0.001
	PP	Tau	-2.384	-1.996	-1.985	-2.256	-2.352	-2.442	-2.609
		Pr<Tau	0.148	0.288	0.293	0.187	0.157	0.132	0.093
Forders	ADF	Tau	-2.507	-1.990	-1.801	-2.167	-2.356	-2.477	-2.899
		Pr<Tau	0.116	0.291	0.379	0.219	0.156	0.123	0.048
	PP	Tau	-2.507	-2.170	-1.981	-2.157	-2.163	-2.173	-2.308
		Pr<Tau	0.116	0.218	0.295	0.223	0.220	0.217	0.170
Invent	ADF	Tau	-2.043	-1.904	-2.168	-3.143	-3.749	-3.735	-3.823
		Pr<Tau	0.268	0.330	0.218	0.025	0.004	0.004	0.003
	PP	Tau	-2.043	-1.962	-2.037	-2.226	-2.369	-2.472	-2.575
		Pr<Tau	0.268	0.303	0.271	0.197	0.152	0.124	0.100
Orders	ADF	Tau	-2.015	-1.968	-1.933	-2.931	-3.037	-3.178	-4.343
		Pr<Tau	0.280	0.300	0.316	0.044	0.034	0.023	0.001
	PP	Tau	-2.015	-2.007	-2.004	-2.179	-2.281	-2.358	-2.494
		Pr<Tau	0.280	0.283	0.285	0.214	0.179	0.155	0.119

FIGURE 38 – Résultats des tests de racine unité.

Variable	Test	Stationary	lag0	lag1	lag2	lag3	lag4	lag5	lag6
Eoutput	ADF	Yes	No	No	No	Yes	Yes	Yes	Yes
	PP	No	No	No	No	No	No	No	No
Eprices	ADF	Yes	No	No	No	Yes	Yes	Yes	Yes
	PP	No	No	No	No	No	No	No	No
Forders	ADF	Yes	No	No	No	No	No	No	Yes
	PP	No	No	No	No	No	No	No	No
Invent	ADF	Yes	No	No	No	Yes	Yes	Yes	Yes
	PP	No	No	No	No	No	No	No	No
Orders	ADF	Yes	No	No	No	Yes	Yes	Yes	Yes
	PP	No	No	No	No	No	No	No	No

FIGURE 39 – Résultats des tests de racine unité au seuil de 5 %.

Références

[1] Baxter, M., King, R.G. (1999), *Measuring Business Cycles : Approximate Band-Pass Filters for Economic Time Series*. Working Paper, University of Virginia.

[2] Beguin, J.M., Gouriéroux C., Monfort A. (1979), *Identification of a mixed autoregressive-moving average process : the corner method*, Document de travail, INSEE.

[3] Beveridge, S., Nelson, C.R. (1981), A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the Business Cycle, *Journal of Monetary Economics*, 7, 151-174.

[4] Boschan, C., Bry, G. (1971), Programmed selection of cyclical turning points, *Cyclical Analysis of Time Series : Selected Procedures and Computer Programmes*, NBER.

[5] Boschan, C., Ebanks, W. (1978). The Phase-Average Trend, a New way of Measuring Economic Growth, *Proceedings of the Business and Economic Statistics Section*, American Statistical Association.

[6] Chambers, J.M., Cleveland, W.S., Keiner, B., Tukey, P.A. (1983), *Graphical Methods for Data Analysis*, Wadsworth and Brooks, Cole Publishing Company.

[7] Destandeau, S., Ladiray, D., Le Guen M. (1999), Analyse exploratoire des données, *Courrier des statistiques*, 90, INSEE, Paris.

[8] Destandeau, S., Le Guen M. (1998), Analyse exploratoire des données avec SAS/INSIGHT, *INSEE guides*, 7-8, INSEE, Paris.

[9] Dickey, D. A., Fuller, W. A. (1979), Distribution of the Estimation for Autoregressive Time Series with a Unit Root, *Journal of The American Statistical Association*, 74, 427-431.

[10] Doz, C., Ladiray, D. (1996), *Décomposition tendance-cycle et datation des points de retournement : analyse de la méthode PAT employée par l'OCDE*, note commune DP (A1-96-059) INSEE (57-G120), Paris.

[11] Doz, C., Rabault, G., Sobczak, N., (1995), Décomposition tendance-cycle : estimations par des méthodes statistiques univariées, *économie et Prévision*, 120, 73-93.

[12] Fayolle, J. (1993), Décrire le cycle économique, *Revue de l'OFCE*, 45.

[13] Fournier, J.Y. (1999), *Extraction du cycle des affaires : la méthode de Baxter et King*, DESE, Document de travail G9916, INSEE, Paris.

[14] Fox J., Long J.S. (1990), *Modern Methods of Data Analysis*, Sage Publications.

[15] Grun-Rehomme, C., Ladiray, D. (1994), Moyennes mobiles centrées et non-centrées : construction et comparaison, *Revue de Statistique Appliquée*, France.

[16] Härdle, W. (1990), *Applied Non-Parametric Regression*, Cambridge University Press.

[17] Hoaglin, D. C., Mosteller, F., Tukey, J. W. (1983), *Understanding Robust and Exploratory Data Analysis*, John Wiley, New York.

[18] Hoaglin, D. C., Mosteller, F., Tukey, J. W. (1985), *Exploring Data Tables, Trends and Shapes*, John Wiley, New York.

[19] Hoaglin, D. C., Velleman, P. F. (1981), *ABC of EDA : Applications, Basics and Computing of Exploratory Data Analysis*, Duxbury Press, Boston.

[20] Hu-Ming, Z., Ping, W. (1994), A New Way to Estimate Orders in Time Series, *Journal of Time series analysis*, 15, 5.

[21] Hodrick, R., Prescott, E. (1980), *Post War US Business Cycles : An Empirical Investigation*, manuscript, Carnegie Mellon University.

[22] Ladiray, D., Roth, N.(1987), Lissage robuste de séries chronologiques : une étude expérimentale, *Annales d'économie et de statistique*, 5, 147-182.

[23] Phillips, P.C., Perron, P.(1988), Testing for a Unit Root in Time Series Regression, *Biometrika*, 75, 335-346.

[24] SAS Institute (1999), SAS Macros and functions, *SAS/ETS User's Manual*, Chapter 4.

[25] SAS Institute (1999), The ARIMA Procedure, *SAS/ETS User's Manual*, Chapter 7.

[26] SAS Institute (1999), The AUTOREG Procedure, *SAS/ETS User's Manual*, Chapter 8.

[27] SAS Institute (1999), The BOXPLOT Procedure, *SAS/STATS User's Manual*, Chapter 18.

[28] SAS Institute (1999), The LOESS Procedure, *SAS/STATS User's Manual*, Chapter 38.

[29] Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison Wesley.

[30] Wilkinson, L. (1999), Dot Plots, *The American Statistician*, 53, 3, 276-281.