PARCOURS DANS SAS INSIGHT

André Wielki

Le logiciel Sas s'utilise en mode batch et en mode interactif. Le mode interactif œuvre par l'intermédiaire de soumissions répétées d'instructions Sas saisies sur la "fenêtre programme".

Sas Institute propose pour ce faire

- des procédures générales (PLOT,PRINT,TABULATE etc) ;
- des procédures statistiques bien nombreuses ;
- des procédures graphiques (GPLOT,GCHART et GMAP).

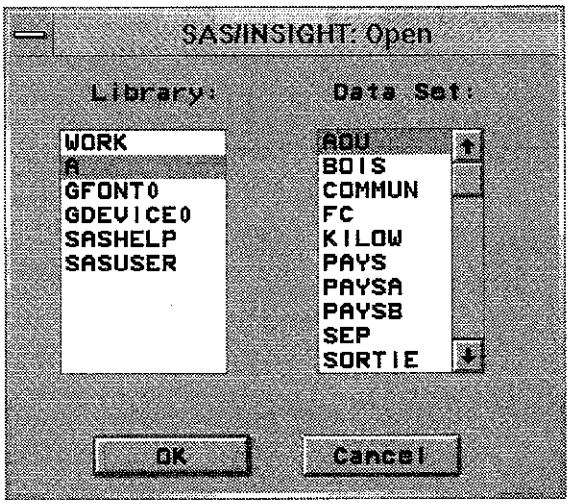
Modifier un programme nécessite de resoumettre les procédures après rectifications.

Récemment Sas Institute a proposé un nouveau produit intégré à l'INSEE dans l'offre *Sas micro sous Windows :* **SAS INSIGHT** ou le module sas d'analyse exploratoire de données.

Il est d'un niveau d'interactivité bien plus élevé que ce que nous avons connu jusqu'à présent.

Comment y accéder ?

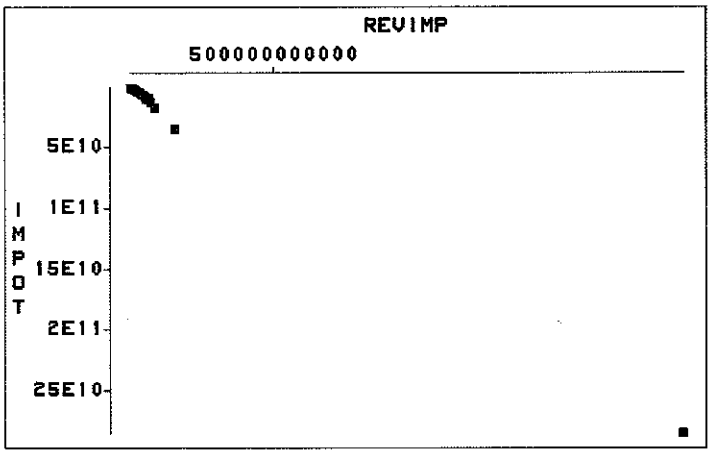
- préparer les accès à l'aide des "Libname" et "Filename" nécessaires
- ouvrir le produit en passant, par exemple, la commande Insight (il existe d'autres manières de faire)
- sélectionner ensuite la table concernée dans la fenêtre Open de Sas Insight en sélectionnant la Library et le Dataset



Exemple 1 : Des possibilités de manipulations sur les variables et sur les observations

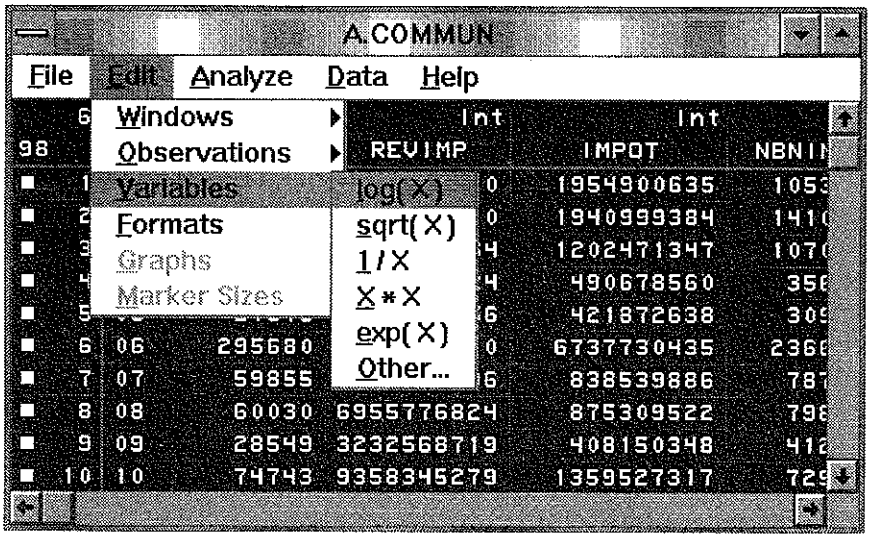
À la réception d'une table sur l'état de la récolte des impôts sur le revenu, j'ai voulu positionner les départements ;

- la rubrique *scatter plot (YX)* du menu Analyze permet de représenter les variables REVIMP et IMPOT

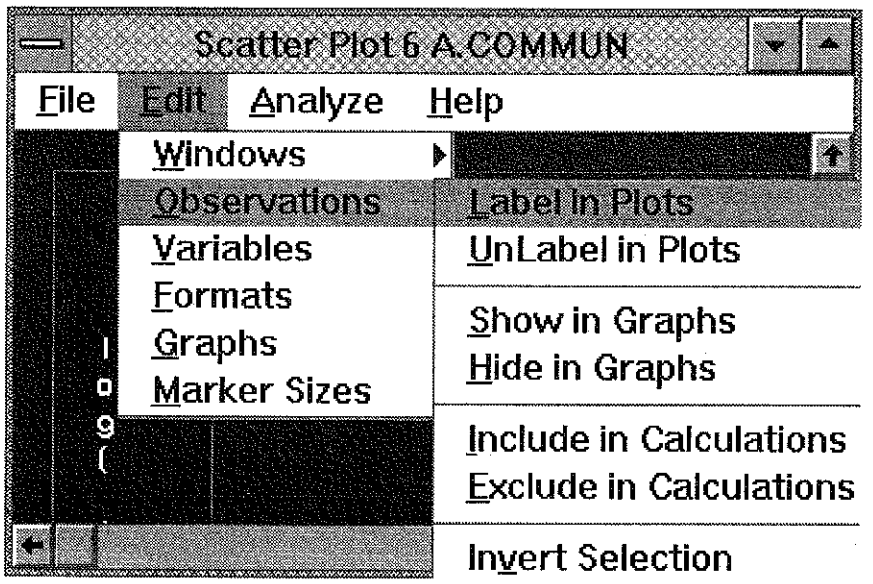


Le résultat est peu satisfaisant.

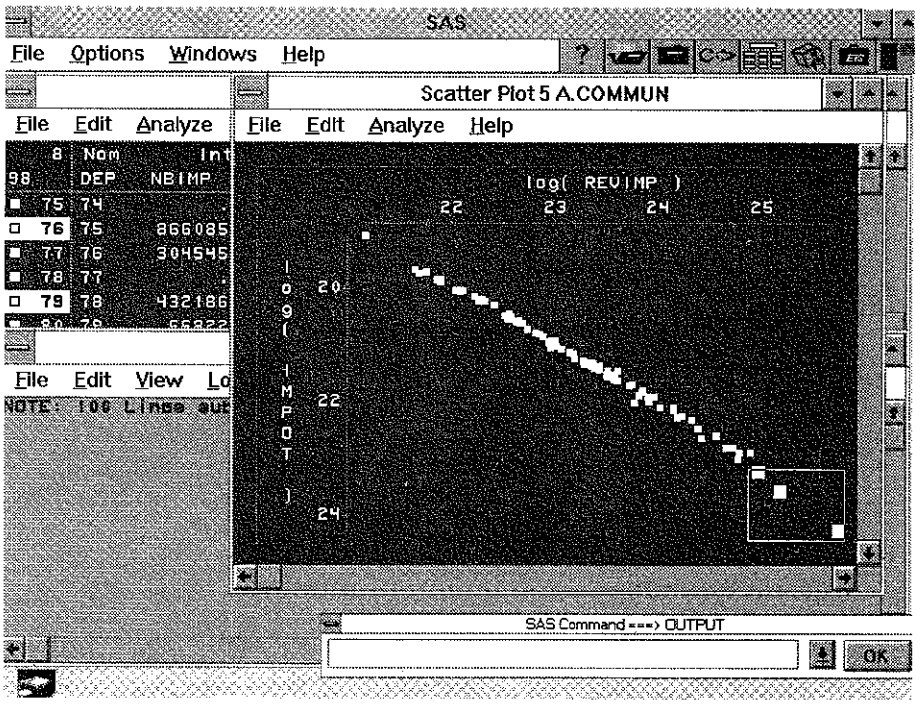
- grâce à la rubrique *Variables-Log (X)* du menu Edit, nous avons opéré la transformation des variables à leur logarithme



- grâce à la rubrique *Observations-Hide in Graphs* du menu Edit, nous avons neutralisé le point aberrant (en fait un total)



- la sélection par encadrement des 3 points les plus contributeurs révèle bien une persistence de l'Île-de-France (75, 78 et 92 surlignés dans la table)



Exemple 2 : D'un coup plusieurs procédures

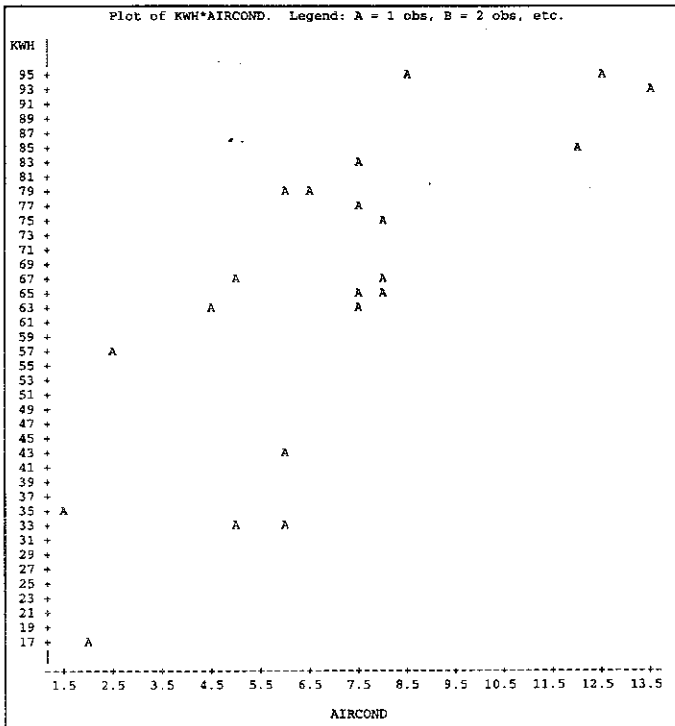
Le programme Sas suivant examine une table de consommation d'électricité en pleine canicule du mois d'août.

Avant, en Sas interactif, on soumettait le programme suivant par morceaux :

```
proc print data=a.kilow;
 var kwh airc;
run;
 proc plot data=a.kilow;plot kwh * aircond;run;
 proc reg data=a.kilow;
 model kwh=aircond;
 id aircond;
  title 'modèle KHW et Air conditionné';
 run;

  print cli;
 run;
  print clm;
 run;

  plot kwh*aircond='0' pred.*aircond='-' l95.*aircond='l'
  u95.*aircond='u'  /overlay;
 run;
 quit;
```

Avec comme résultats successifs :

OBS	KWH	AIRCOND
1	35	1.5
2	63	4.5
3	66	5.0
4	17	2.0
5	94	8.5
6	79	6.0
7	93	13.5
8	66	8.0
9	94	12.5
10	82	7.5
11	78	6.5
12	65	8.0
13	77	7.5
14	75	8.0
15	62	7.5
16	85	12.0
17	43	6.0
18	57	2.5
19	33	5.0
20	65	7.5
21	33	6.0



modèle KHW et Air conditionné

Model: MODEL1
Dependent Variable: KWH

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	5609.66260	5609.66260	26.855	0.0001
Error	19	3968.90883	208.88994		
C Total	20	9578.57143			

Root MSE	14.45303	R-square	0.5856
Dep Mean	64.85714	Adj R-sq	0.5638
C.V.	22.28440		

Parameter Estimates

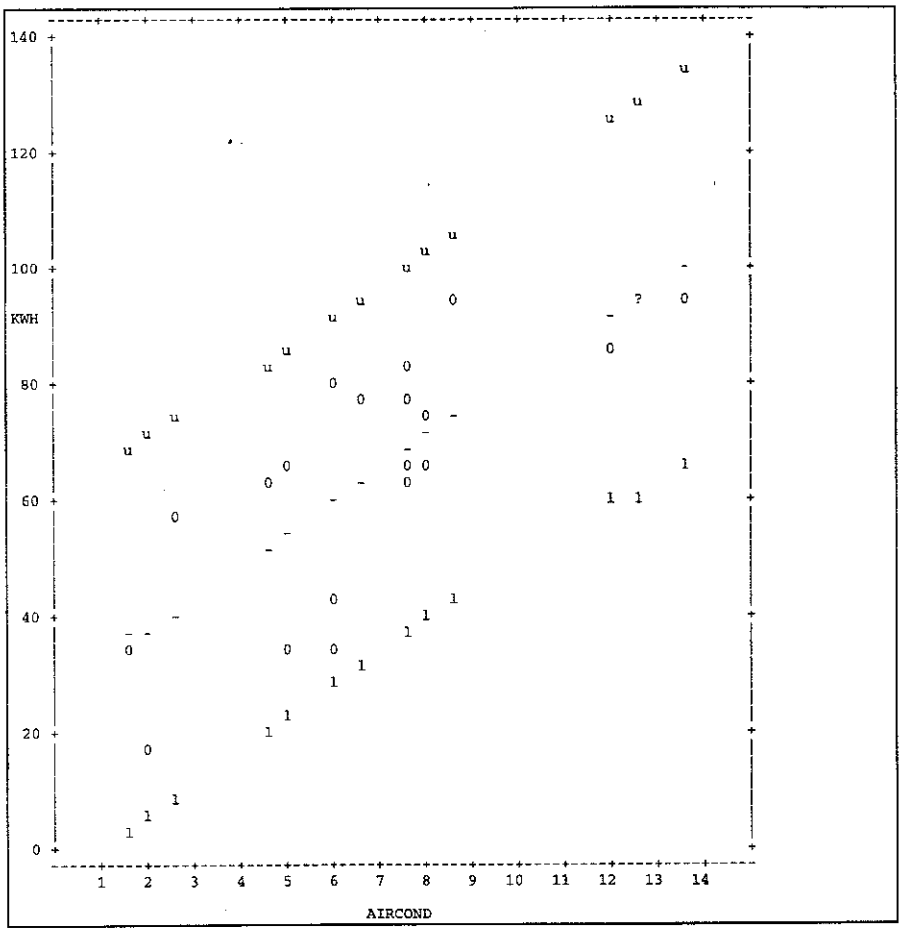
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	27.851072	7.80653827	3.568	0.0021
AIRCOND	1	5.341082	1.03067009	5.182	0.0001

Obs	AIRCOND	Dep Var KWH	Predict Value	Std Err Predict	Lower95% Predict	Upper95% Predict	Residual
1	1.5	35.0000	35.8627	6.423	2.7597	68.9657	-0.8627
2	4.5	63.0000	51.8859	4.026	20.4834	83.2884	11.1141
3	5	66.0000	54.5565	3.728	23.3158	85.7971	11.4435
4	2	17.0000	38.5332	5.979	5.7963	71.2702	-21.5332
5	8.5	94.0000	73.2503	3.545	42.1028	104.4	20.7497
6	6	79.0000	59.8976	3.296	28.8704	90.9247	19.1024
7	13.5	93.0000	99.9557	7.471	65.9024	134.0	-6.9557
8	8	66.0000	70.5797	3.342	39.5312	101.6	-4.5797
9	12.5	94.0000	94.6146	6.551	61.4013	127.8	-0.6146
10	7.5	82.0000	67.9092	3.208	36.9223	98.8961	14.0908
11	6.5	78.0000	62.5681	3.185	31.5919	93.5443	15.4319
12	8	65.0000	70.5797	3.342	39.5312	101.6	-5.5797
13	7.5	77.0000	67.9092	3.208	36.9223	98.8961	9.0908
14	8	75.0000	70.5797	3.342	39.5312	101.6	4.4203
15	7.5	62.0000	67.9092	3.208	36.9223	98.8961	-5.9092
16	12	85.0000	91.9441	6.105	59.1057	124.8	-6.9441
17	6	43.0000	59.8976	3.296	28.8704	90.9247	-16.8976
18	2.5	57.0000	41.2038	5.548	8.8010	73.6065	15.7962
19	5	33.0000	54.5565	3.728	23.3158	85.7971	-21.5565
20	7.5	65.0000	67.9092	3.208	36.9223	98.8961	-2.9092
21	6	33.0000	59.8976	3.296	28.8704	90.9247	-26.8976

Sum of Residuals	0
Sum of Squared Residuals	3968.9088
Predicted Resid SS (Press)	4728.5664

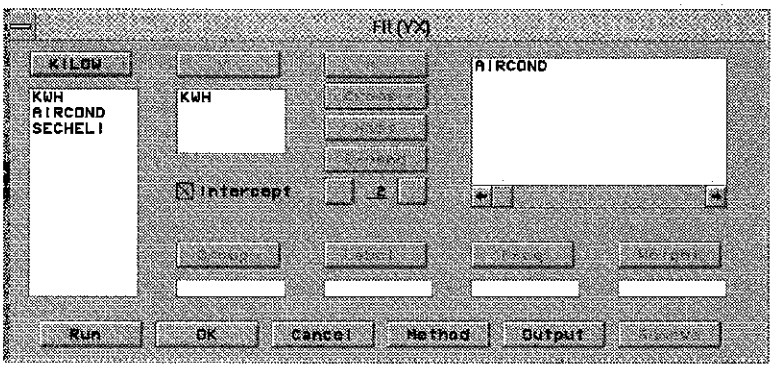
Obs	AIRCOND	Dep Var KWH	Predict Value	Std Err Predict	Lower95% Predict	Upper95% Predict	Residual
1	1.5	35.0000	35.8627	6.423	22.4197	49.3057	-0.8627
2	4.5	63.0000	51.8859	4.026	43.4585	60.3134	11.1141
3	5	66.0000	54.5565	3.728	46.7536	62.3593	11.4435
4	2	17.0000	38.5332	5.979	26.0186	51.0478	-21.5332
5	8.5	94.0000	73.2503	3.545	65.8295	80.6710	20.7497
6	6	79.0000	59.8976	3.296	52.9991	66.7960	19.1024
7	13.5	93.0000	99.9557	7.471	84.3181	115.6	-6.9557
8	8	66.0000	70.5797	3.342	63.5856	77.5739	-4.5797
9	12.5	94.0000	94.6146	6.551	80.9023	108.3	-0.6146
10	7.5	82.0000	67.9092	3.208	61.1939	74.6245	14.0908
11	6.5	78.0000	62.5681	3.185	55.9025	69.2337	15.4319
12	8	65.0000	70.5797	3.342	63.5856	77.5739	-5.5797
13	7.5	77.0000	67.9092	3.208	61.1939	74.6245	9.0908
14	8	75.0000	70.5797	3.342	63.5856	77.5739	4.4203
15	7.5	62.0000	67.9092	3.208	61.1939	74.6245	-5.9092
16	12	85.0000	91.9441	6.105	79.1666	104.7	-6.9441
17	6	43.0000	59.8976	3.296	52.9991	66.7960	-16.8976
18	2.5	57.0000	41.2038	5.548	29.5916	52.8160	15.7962
19	5	33.0000	54.5565	3.728	46.7536	62.3593	-21.5565
20	7.5	65.0000	67.9092	3.208	61.1939	74.6245	-2.9092
21	6	33.0000	59.8976	3.296	52.9991	66.7960	-26.8976

Sum of Residuals	0
Sum of Squared Residuals	3968.9088
Predicted Resid SS (Press)	4728.5664

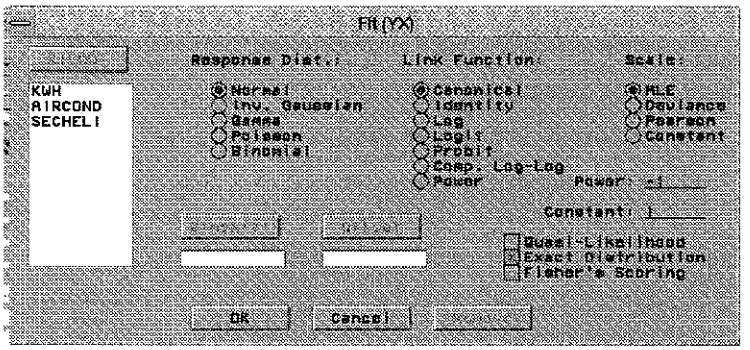

KWH
AIRCOND
140
120
100
80
60
40
20
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14

En Sas insight, d'autres actions sont nécessaires :

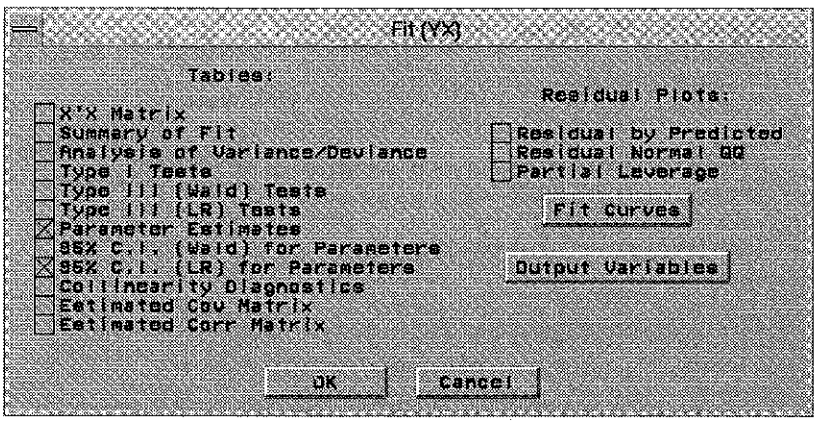
- appel du menu déroulant Analyze
- choix de la rubrique *fit YX*



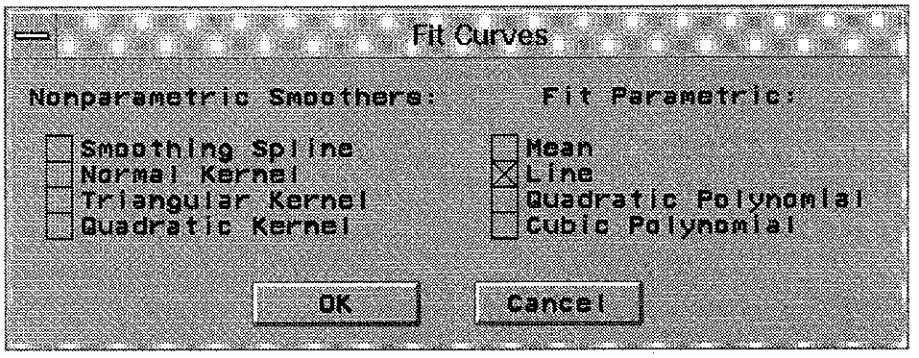
- choix de la méthode grâce au bouton METHOD



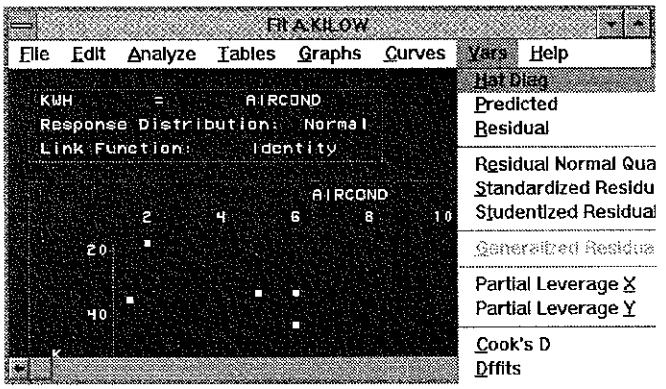
- sélection des sorties souhaitées grâce au bouton OUTPUT



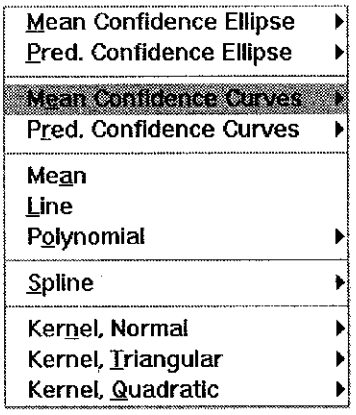
- enrichissement complémentaire de tracés grâce au bouton FIT CURVES



- le résultat est susceptible d'être enrichi immédiatement



nous l'avons enrichi par les courbes d'intervalles de confiance de la moyenne à 95 %



- la récupération du résultat chiffré hors de Sas se fera par appel au menu déroulant File à la rubrique *Tables*

KWH = AIRCOND
Response Distribution: Normal
Link Function: Identity

Parametric Regression Fit

Type	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat
Line	1.0000	5609.6626	19.0000	208.8899	0.5856	26.8546

Parametric Regression Fit

Prob > F
0.0001

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
Model	1.0000	5609.6626	5609.6626	26.8546	0.0001
Error	19.0000	3968.9088	208.8899	.	.
C Total	20.0000	9578.5714	.	.	.

Parameter Estimates

Variable	DF	Estimate	Std Error	T Stat	Prob >|T|
INTERCEPT	1.0000	27.8511	7.8065	3.5677	0.0021
AIRCOND	1.0000	5.3411	1.0307	5.1821	0.0001

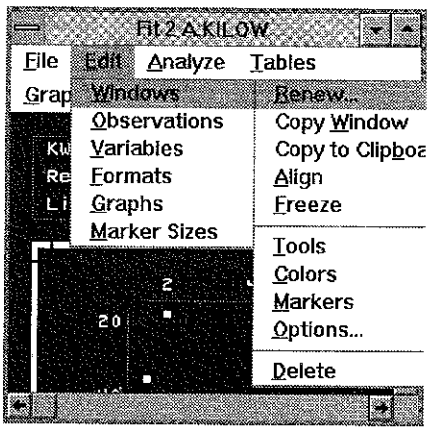
Parameter Estimates

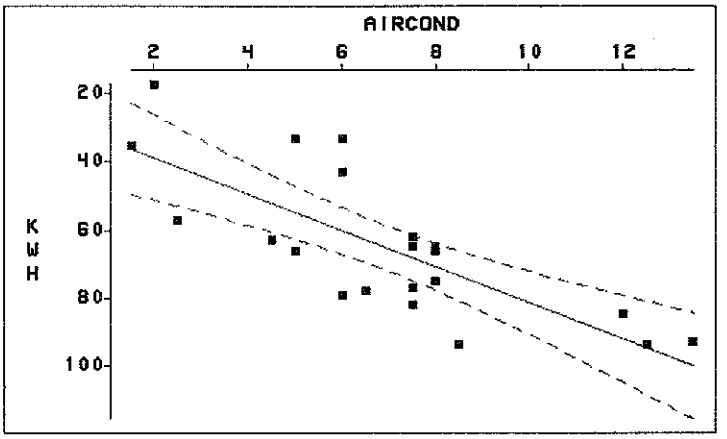
Tolerance	Var Inflation
.	0.0000
1.0000	1.0000

95% C.I. for Parameters

Variable	Estimate	Lower	Upper
INTERCEPT	27.8511	11.5118	44.1903
AIRCOND	5.3411	3.1839	7.4983

- la récupération du résultat graphique hors de Sas se fera après la sélection du graphique par appel au menu déroulant Edit à la rubrique *Copy to clipboard*



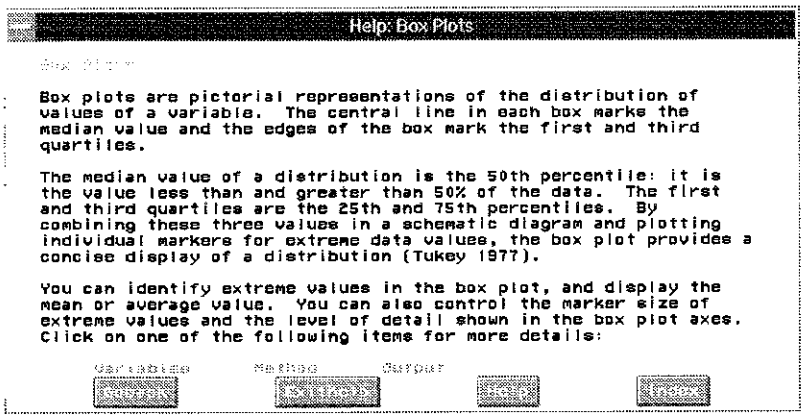


Dans la prochaine version du logiciel, les problèmes de fond noir auront disparu.

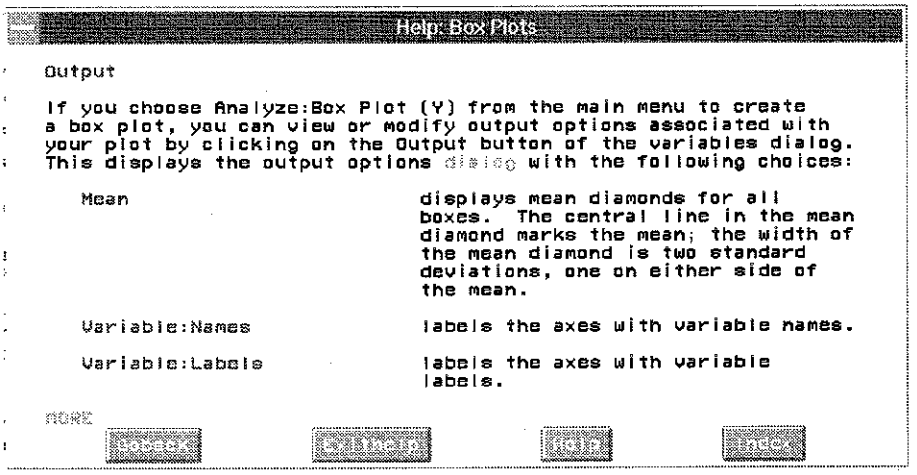
Exemple 3 : Une aide en ligne pour consolider votre exploration

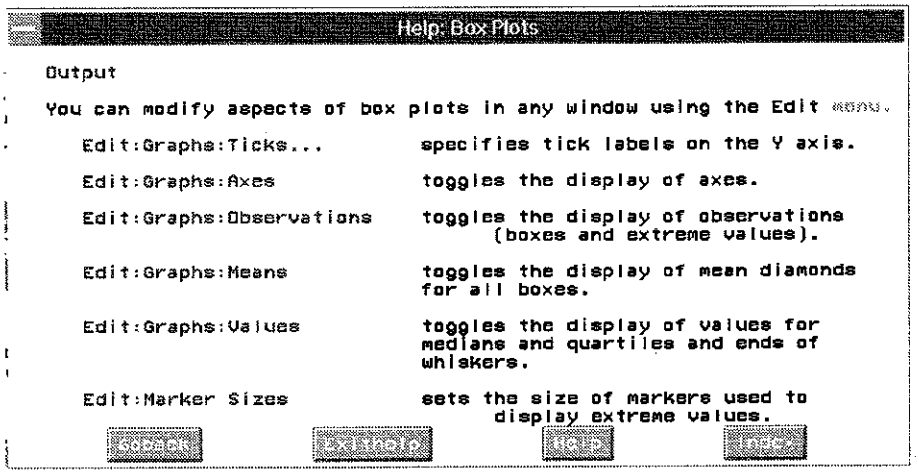
Le ministère du tourisme a rassemblé des informations sur les campings. Je souhaite me faire une petite idée à ce sujet :

- ouverture de la table a.fc
- appel des *boites à moustaches* sur les emplacements disponibles (rubrique *box plot (Y)* du menu analyze)
- grâce à la rubrique *Reference* du menu déroulant Help, nous avons pu rafraîchir nos connaissances sur la signification des moustaches

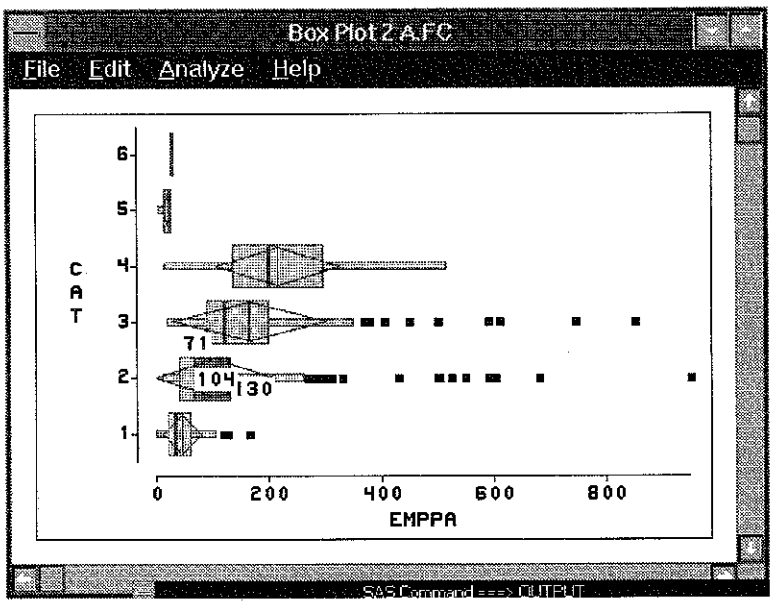


Aussi en ce qui concerne les options de sortie.





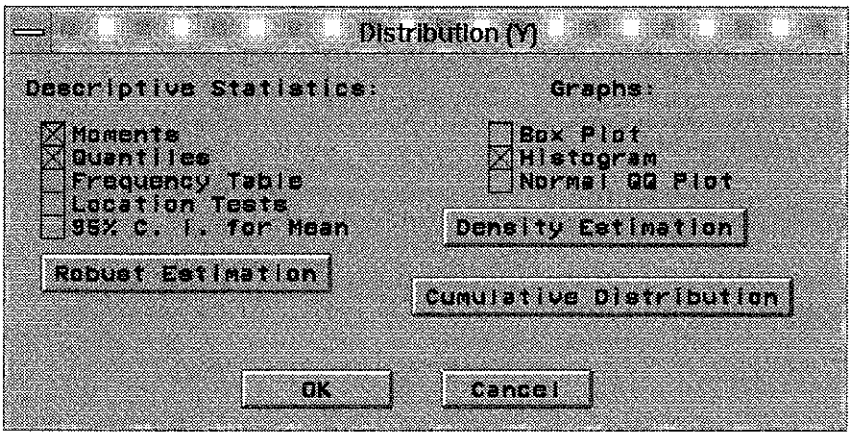
pour obtenir une compréhension immédiate de la sortie.



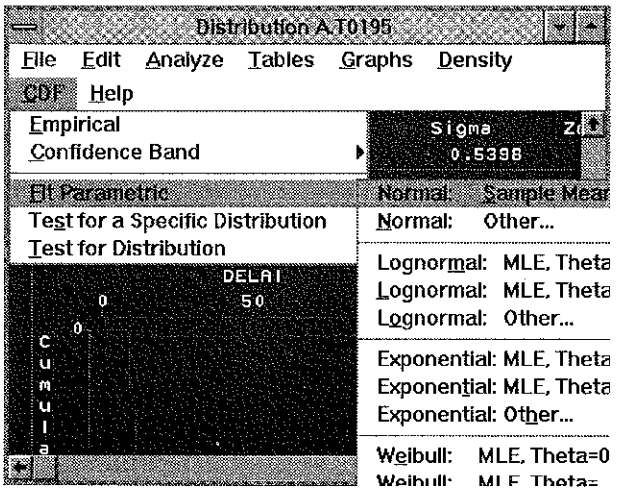
Exemple 4 : Une mise au point progressive et ouverte sans *a priori*

Une enquête est effectuée auprès d'entreprises. On dispose d'une statistique sur le délai de réponse au questionnaire :

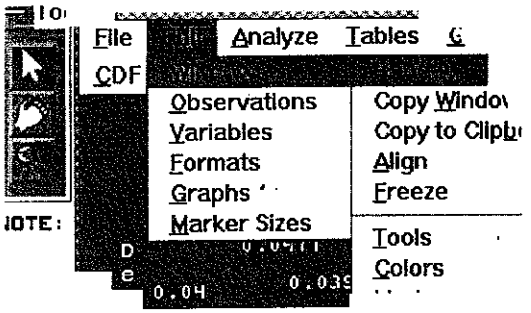
- ouverture de la table
- un bar chart permet de se faire une première idée de la distribution (rubrique *bar chart (Y)* du menu Analyze)
- on poursuit l'analyse de la distribution en demandant un histogramme (rubrique distribution (Y) du menu Analyze)



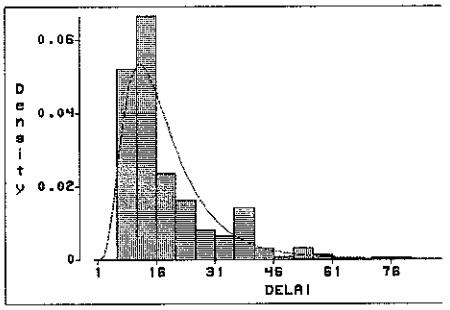
- on demande ensuite au vu de l'histogramme, la courbe de densité de la loi de la distribution étudiée : la log normale est la plus vraisemblable (rubriques du menu Density)



- il est nécessaire de rechercher aussi le meilleur histogramme pour déjouer les apparences trompeuses par l'appel aux outils (la petite main)



- je puis enfin retenir l'histogramme suivant qui me rappellera le second pic des retours du questionnaire



DELAI

Parametric Density Estimation

Distribution	Mean / Theta	Sigma	Zeta / C
Lognormal	0.0000	0.5398	2.7728

Moments

N	2642.0000	Sum Wgts	2642.0000
Mean	18.8861	Sum	49897.0000
Std Dev	12.7406	Variance	162.3236
Skewness	2.0093	Kurtosis	4.7298

Moments

USS	1371055.00	CSS	428696.707
CV	67.4605	Std Mean	0.2479

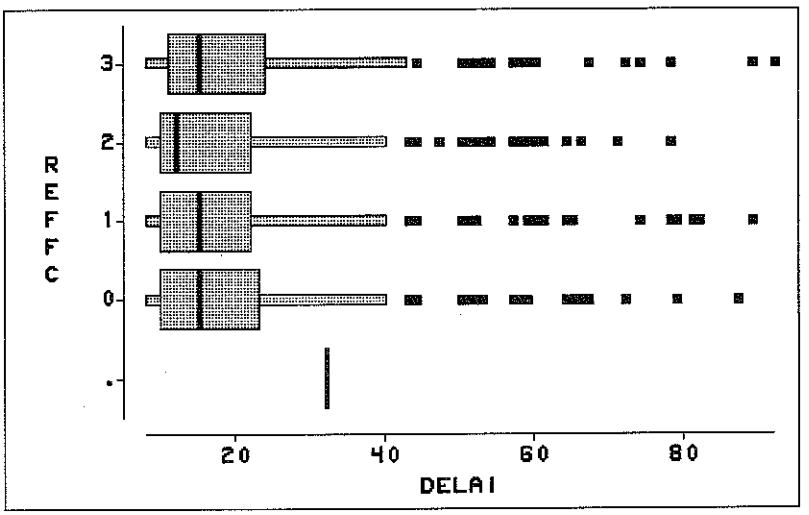
Quantiles

100% Max	92.0000	99%	66.0000	Range	84.0000
75% Q3	22.0000	95%	43.0000	Q3-Q1	12.0000
50% Med	15.0000	90%	37.0000	Mode	9.0000
25% Q1	10.0000	10%	9.0000		.
0% Min	8.0000	5%	9.0000		.
	.	1%	8.0000		.

- une petite excursion extérieure me permet de poursuivre l'analyse grâce à la soumission d'un Proc rank établissant 4 classes sur les variables d'effectif et du chiffre d'affaire.

```
   PROC RANK DATA=a.t0195 OUT=a GROUP=4;
4      VAR effc cac;
5      RANKS reffc rcac;
6     RUN;

NOTE: The data set WORK.A has 2642 observations and 5 variables.
NOTE: At least one W.D format was too small for the number to be printed.
   The
      decimal may be shifted by the "BEST" format.
NOTE: The PROCEDURE RANK used 2.97 seconds.
```

- retour à Insight pour poursuivre l'analyse visuelle enrichie : un box plot by reffc

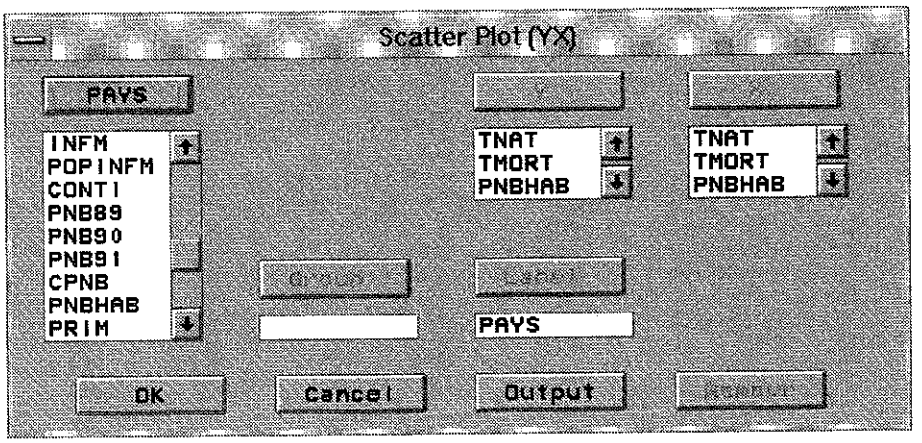


Exemple 5 : Aussi des outils exploratoires multidimensionnels

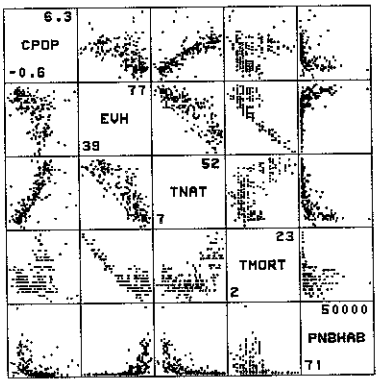
Le dernier exemple portera sur un tableau récapitulatif de la revue Population de l'Ined.

Il s'agit d'une table Pays contenant un certain nombre d'informations sur tous les pays du globe :

- ouverture de la table
- notre souci est d'essayer d'appréhender visuellement les liens entre différentes variables démographiques quantitatives et une mesure de la richesse par le biais dans un premier temps d'un ensemble de croisements variable par variable : rubrique *Scatter plot (YX)* du menu Analyze



- un léger apprentissage est nécessaire pour s'habituer à ce genre de présentation (comment lire les axes ?)



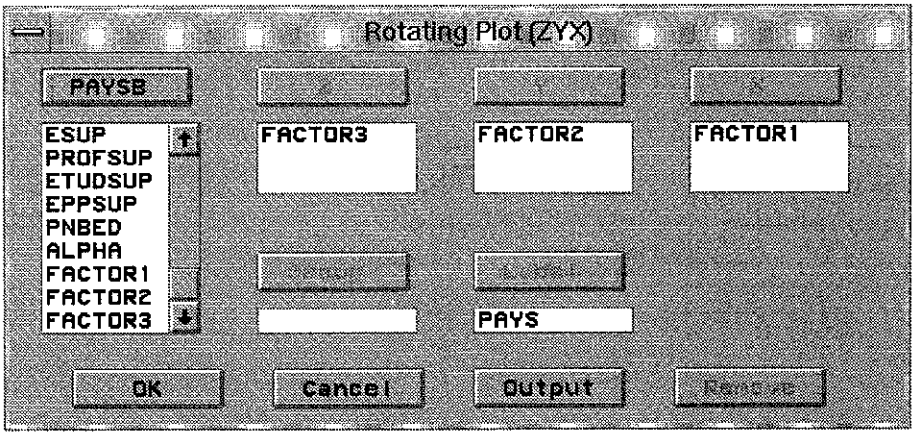
- possibilités de faire plusieurs opérations sur cette synthèse :
  la sélection d'un point
  l'exclusion d'un point
  la coloration
  la loupe (voir Tools)
  le brushing
- dans un second temps, appel d'une procédure extérieure sur ces données : Proc factor récupérant les coordonnées des projections des profils-lignes sur les 3 premiers axes d'une analyse en composantes principales

```
PROC FACTOR DATA=a.pays OUT=a.paysb NFACTORS=3;
 VAR  cpop evh tnat tmort  pnbhab;
RUN;
```

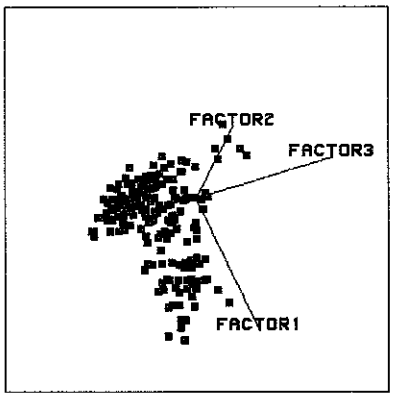
Avec ses résultats papier et fichier :

```
Initial Factor Method: Principal Components

                    Prior Communality Estimates: ONE

        Eigenvalues of the Correlation Matrix:  Total = 5  Average = 1

                        1          2          3          4          5
Eigenvalue         3.2372     1.0074     0.5824     0.1105     0.0625
Difference         2.2298     0.4250     0.4719     0.0480
Proportion         0.6474     0.2015     0.1165     0.0221     0.0125
Cumulative         0.6474     0.8489     0.9654     0.9875     1.0000

           3 factors will be retained by the NFACTOR criterion.

                          Factor Pattern

              FACTOR1    FACTOR2    FACTOR3

CPOP          0.71406    0.55146    0.39482    Croissance population
EVH          -0.93486    0.29223   -0.00881    Esp,rance de vie (homme)
TNAT          0.94229    0.17218    0.11743    Taux brut de natalit, (/1000)
TMORT         0.67604   -0.71393    0.07947    Taux brut de mortalit, (/1000)
PNBHAB       -0.71304   -0.28024    0.63745    PNB par habitant

                    Variance explained by each factor

                       FACTOR1    FACTOR2    FACTOR3
                      3.237207   1.007380   0.582404


                Final Communality Estimates: Total = 4.826991

                    CPOP        EVH        TNAT       TMORT      PNBHAB
                0.969862   0.959440   0.931343   0.973040   0.993306


                  Scoring Coefficients Estimated by Regression

            Squared Multiple Correlations of the Variables with each Factor

                       FACTOR1    FACTOR2    FACTOR3
                      1.000000   1.000000   1.000000

                  Standardized Scoring Coefficients

              FACTOR1    FACTOR2    FACTOR3

CPOP          0.22058    0.54742    0.67791    Croissance population
EVH          -0.28879    0.29009   -0.01512    Esp,rance de vie (homme)
TNAT          0.29108    0.17092    0.20163    Taux brut de natalit, (/1000)
TMORT         0.20883   -0.70870    0.13645    Taux brut de mortalit, (/1000)
PNBHAB       -0.22027   -0.27818    1.09451    PNB par habitant
```

- ouverture de la table enrichie
- représentation des coordonnées des points sur les 3 premiers axes des facteurs par l'appel de la rubrique *Rotating plot (Z Y X)* du menu Analyze



- pour obtenir une visualisation en 3 dimensions de la répartition de points, difficilement perceptible sur le papier : manipulation avec la petite main !



Conclusion

Sas Institute a donc fourni un outil convivial et véritablement interactif pour le statisticien chercheur qui doit explorer de nouvelles données.

Cet outil, par son caractère visuel, permet d'aller plus vite en appréhendant assez rapidement des tendances, les pièges, les exceptions.

Il contient une aide en ligne assez fournie pour soutenir un statisticien hésitant. Il faut cependant être bien formé à la statistique pour l'utiliser à bon escient.

Des allers et retours vers l'extérieur (Sas ou non Sas) sont possibles.

Cet outil n'est quand même pas surpuissant : pour l'utiliser face aux immenses tables France entière, etc ; il vous sera nécessaire d'en extraire des sous-ensembles vu la limite actuelle de la mémoire de l'ordinateur et du système d'exploitation.

La prochaine version d'Insight permettra de mémoriser et donc de rejouer des scénarios.

Espérances concernant la version 6.11 :

- il sera possible de saisir des données comme sur une feuille de tableur ;
- les relations entre une variable X et plusieurs Y se visualiseront sur un seul graphique avec plusieurs courbes (line plot) ;
- les boites à moustaches seront réservées aux variables numériques, vu l'introduction de "mosaïques" pour la visualisation de variables nominales ;
- apparition de manettes de réglages concernant l'ajustement des courbes de densité ;
- meilleure sauvegarde des graphiques sous forme de fichiers Gif, Tiff ou Postscript ;
- possibilité de double-cliquer sur une observation et de voir son contenu détaillé ;
- améliorations dans la modélisation ;
- augmentation des possibilités de transformation des variables ;
- "Petite main" automatique pour l'ajustement des axes et des marques d'échelles ;
- possibilités nouvelles pour la représentation en trois dimensions ;
- animation et mémorisation de scénarios.