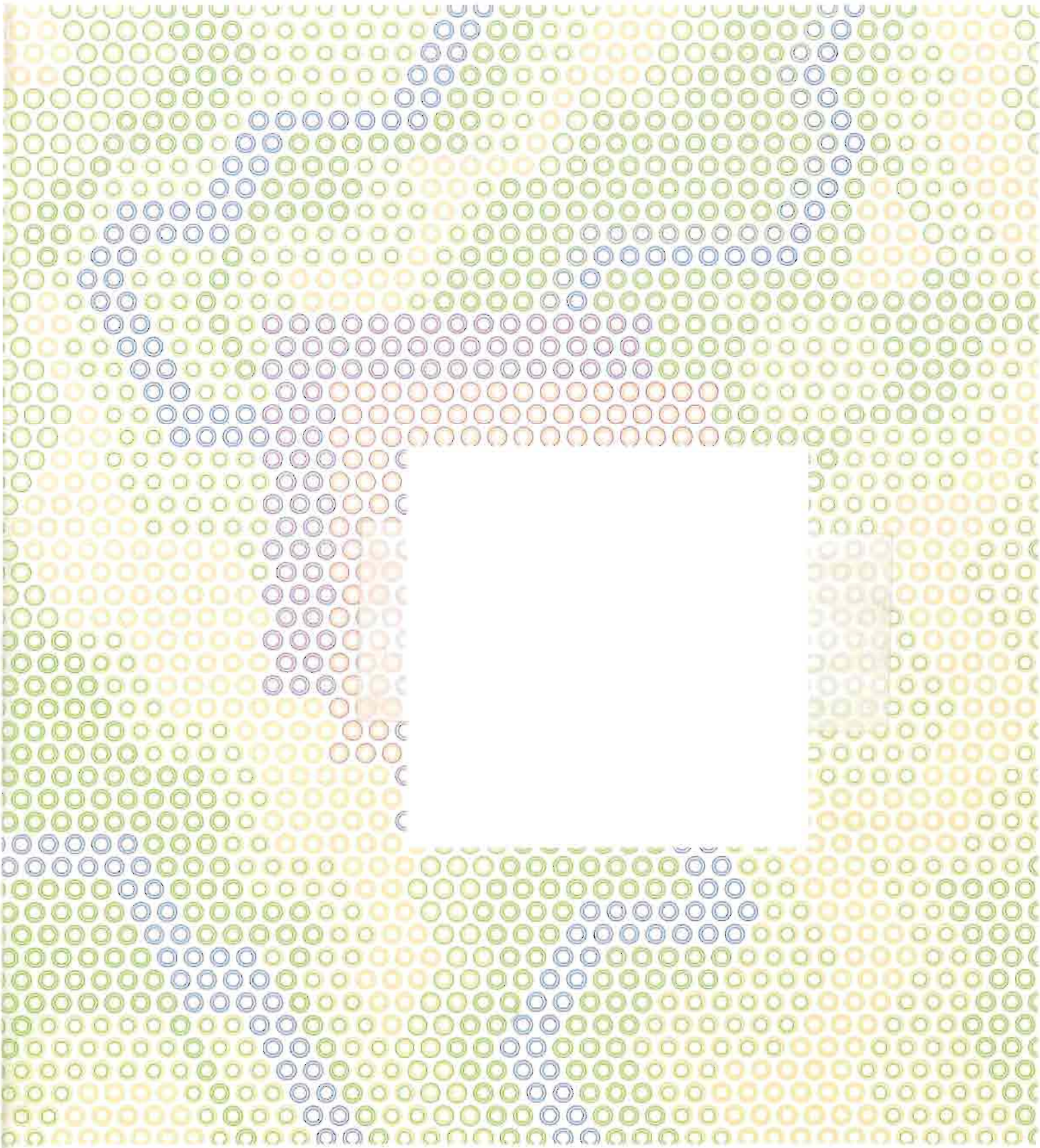
Philippe Waniez

RECLUS

Analyse exploratoire des données



collection reclus modes d'emploi n° 17

Collection Reclus Modes d'Emploi n°17

Analyse exploratoire des données

Philippe WANIEZ
ORSTOM, GIP RECLUS

GIP RECLUS

MAISON DE LA GÉOGRAPHIE, MONTPELLIER

Analyse exploratoire des données / WANIEZ Philippe
Montpellier: G.I.P. RECLUS, 1991.— 160 p., 187 fig., 30 cm.—
(Collection Reclus Modes d'Emploi n°17)
ISBN 2-86912-035-8
ISSN 0298-9689

Ce fascicule a été rédigé par Philippe Waniez, chargé de recherches à l'ORSTOM,
Maison de la Géographie

Secrétariat de publication: Régine Vanduick

Direction: Hervé Théry

G.I.P. RECLUS, Maison de la Géographie, 17 rue Abbé de l'Épée,
34000 Montpellier
Tél. 67 72 46 10; Fax 67 72 64 04

Analyse exploratoire des données

Philippe WANIEZ
ORSTOM, GIP RECLUS

Du même auteur

Au GIP RECLUS

• *Les données et le Territoire: initiation à la numérisation pour la cartographie statistique.* (en collaboration avec Violette Cabos).

• *Pratique de l'analyse statistique, SAS sur PC/PS, mini et gros systèmes.* Col. RECLUS modes d'emploi, n°15. (en collaboration avec Micheline Cosinschi).

Au GIP RECLUS et à l'ORSTOM

• *Les données et le Territoire: initiation au traitement informatique des données spatialisées.* (en collaboration avec Gérard Dandoy et Violette Cabos).

A l'ORSTOM

• *Les données et le Territoire: initiation à l'analyse en surfaces de tendances.* (en collaboration avec Yann Le Gauffey).

Aux Editions Eyrolles.

• *Cartographie sur Macintosh.*

• *Système d'information géographique: initiation pratique sur Macintosh.*

Remerciements

L'auteur tient à remercier, pour leur contribution proche ou moins proche à la genèse de cet ouvrage:

• Jean-Pascal Grevet, Directeur de Publication et Rédacteur en Chef de la revue *Icônes*.

• Micheline Cosinschi, Maître Assistant au Département de Géographie de l'Université de Lausanne.

• Pierre Gondard et Joël Bonnemaison, Directeurs Scientifiques au Département Société, Urbanisation, Développement de l'ORSTOM. Louis Arréghini, Allocataire de Recherche au Centre ORSTOM de Nouméa.

• Roger Brunet et Hervé Théry, Directeurs successifs du GIP RECLUS.

• Violette Brustlein, Régine Vanduick et Joël Charre (Maison de la Géographie de Montpellier), Françoise Pelletier et Christian Mullon (Laboratoire d'Informatique Appliquée de l'ORSTOM à Bondy), Christiane Weber (Université Louis Pasteur de Strasbourg, CNRS URA902).

Table

INTRODUCTION

L'analyse exploratoire des données, en anglais *Exploratory Data Analysis (EDA)*, imaginée par le statisticien J.W. Tukey (de l'Université de Princeton et des Laboratoires AT&T Bell) rencontre aujourd'hui, en Europe, un vif intérêt. Peu normative, l'analyse exploratoire insiste sur l'inadaptation des hypothèses sous-jacentes à la statistique classique, hypothèses souvent trop fortes au regard de la complexité des univers analysés. Elle cherche, de plus, à prendre en compte les anomalies ou les cas extrêmes, trop souvent considérés comme aberrants, car s'ajustant mal aux «lois» statistiques.

Enquêter comme un détective

Au lieu de rechercher à tout prix l'adéquation à un test statistique, et de prendre, de manière quasi rituelle, une décision de type probabiliste, l'analyse exploratoire s'intègre dans un processus de recherche combinant les deux méthodes. L'approche exploratoire conduit à «radiographier les données», à chercher ce qui se passe dans les chiffres, sans *a priori*. J. Tukey propose d'adopter les règles d'un détective qui examine la scène d'un crime: garder l'esprit ouvert, chercher, un indice après l'autre, les vérités enfouies sous la masse des données et isoler un problème qui, par la suite, se traitera par des méthodes moins intuitives.

L'analyse exploratoire ne doit donc pas être comprise comme une alternative à la statistique classique. En utilisant de manière extensive les représentations graphiques, en associant souvent plusieurs modes de représentation des données, en effectuant un retour constant aux tableaux d'origine, en regardant les données selon des perspectives variées, on maîtrise mieux les relations entre les variables et l'on reconnaît plus facilement les groupes d'individus. Ce retour à la vieille méthode «crayon-papier» aboutit à une réelle intimité de l'analyste avec ses données et permet d'éviter les conclusions prises à la hâte, de manière trop mécanique, conduisant parfois à nier

l'évidence ou mieux, à imposer des conclusions absurdes sous la foi d'un test mal adapté aux questions auxquelles il est censé apporter une réponse.

Il ne faudrait cependant pas croire que l'analyse exploratoire constitue un simple ensemble de règles visant à rendre plus prudente la démarche de l'analyste de données. Ce serait faire là une grave injustice à l'encontre des promoteurs de cette approche et, surtout, ignorer les outils d'analyse très originaux qu'ils nous ont légués. Aujourd'hui, l'informatique permet d'en exploiter toute la subtilité.

Analyse exploratoire et univers Macintosh

Cet ouvrage vise plusieurs buts. En premier lieu, il s'agit de combler un vide regrettable dans la littérature en langue française sur l'analyse statistique exploratoire. Bien entendu, on ne trouvera pas ici un nouveau traité alors que la traduction des œuvres de Tukey ou de ses successeurs reste à faire. Plus modestement, on a cherché à montrer l'originalité de cette autre approche des données numériques, tout en la replaçant dans les chapitres usuels de statistique: analyse univariée, bivariée et multivariée.

En second lieu, la présentation des différentes méthodes de l'analyse exploratoire ne peut plus être envisagée indépendamment des logiciels qui permettent, aujourd'hui, d'en tirer un meilleur parti. Ces logiciels apparaissent comme de véritables chefs-d'œuvres d'imagination et d'intelligence, tant sur le plan des interfaces logiciel/utilisateur, que sur celui de la mise en œuvre des méthodes exploratoires. On a donc tenté de présenter les principales différences entre ces logiciels originaux afin d'en faciliter le choix pour une application particulière.

Pour faciliter la présentation des méthodes et des logiciels, un exemple unique court tout au long de l'ouvrage. Il s'agit d'un ensemble de données démographiques et économiques tirées du recensement de la population de Nouvelle-Calédonie du 15 avril 1983 (figure n° 0.1). Elles ont été choisies en raison non seulement de l'intérêt des thèmes auxquels elles se rapportent, mais aussi pour la complexité de leurs distributions statistiques (valeurs exceptionnelles, distributions non symétriques, etc.). La figure n° 0.2 représente la carte des limites des communes de Nouvelle-Calédonie auxquelles se rapportent ces données.

Analyse exploratoire et ordinateur

Publié en 1977, l'ouvrage de référence de J. Tukey, Exploratory Data Analysis, est rédigé alors que plusieurs logiciels d'analyse statistique (que nous nommerons désormais «statisticiels») comme SPSS ou BMDP assurent l'essentiel des tâches de traitement des données des gros centres informatiques où règne encore la carte perforée. La micro informatique n'en est qu'à ses premiers balbutiements: le Personal

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
NOC01	N	BELEP	4.7	9.9	41.8	49.3	99.9	2.6	50.0	76.1	3.5	1805.3	0.0
NOC02	S	BOULOUPARI	7.8	23.1	35.5	45.3	84.5	8.2	23.5	78.0	3.7	405.3	63.6
NOC03	S	BOURAIL	23.5	8.3	32.4	45.7	77.8	13.2	37.7	84.8	3.9	337.0	79.5
NOC04	N	CANALA	26.4	-1.1	42.2	47.3	95.7	16.9	23.5	78.8	4.9	524.1	27.9
NOC05	S	DUMBEA	38.1	32.1	36.6	48.6	66.0	0.8	26.7	82.1	4.1	289.0	97.1
NOC06	S	FARINO	1.7	30.4	31.6	45.8	92.5	8.5	25.4	84.6	3.5	428.8	93.1
NOC07	N	HIENGHENE	11.9	-10.5	41.1	48.6	97.6	29.9	9.9	78.7	4.2	251.3	9.2
NOC08	N	HOUAILOU	27.5	3.7	41.5	49.1	93.3	17.0	17.2	77.4	4.8	421.0	45.6
NOC09	S	ILES-DES-PINS	8.9	17.5	42.7	45.8	95.1	20.8	30.5	79.4	6.5	569.5	21.2
NOC10	N	KAALA-GOMEN	8.5	-13.6	37.0	48.6	96.3	4.8	32.7	72.1	4.0	732.7	42.2
NOC11	N	KONE	20.1	17.7	36.8	48.5	94.2	8.9	35.5	79.1	5.0	539.6	53.1
NOC12	N	KOUMAC	9.7	•	36.2	50.2	90.8	2.3	38.0	88.8	4.6	468.3	76.6
NOC13	S	LA-FOA	14.4	5.1	35.3	49.8	83.0	13.1	33.3	82.6	4.0	371.3	66.3
NOC14	I	LIFOU	55.9	7.2	44.5	50.1	97.8	43.2	14.2	72.1	5.1	333.8	10.4
NOC15	I	MARE	31.7	10.9	46.1	49.8	99.2	30.1	6.8	73.1	5.7	232.4	4.7
NOC16	S	MOINDOU	2.6	-2.3	33.9	46.8	95.2	21.7	33.0	87.3	4.1	356.6	65.6
NOC17	S	MONT-DORE	100.5	37.1	37.6	48.8	67.7	0.6	21.4	84.5	4.6	350.7	97.1
NOC18	S	NOUMEA	413.5	7.2	31.4	49.8	63.3	0.1	33.2	86.0	3.6	284.1	99.0
NOC19	N	OUEGOA	10.1	-3.0	39.1	45.8	98.0	38.2	24.0	84.9	4.9	549.8	19.5
NOC20	I	OUVEA	19.1	-0.2	42.6	48.8	98.2	25.4	4.7	80.4	5.5	193.6	2.2
NOC21	S	PAITA	33.3	41.9	38.8	46.9	68.7	2.6	27.8	78.2	4.8	415.3	82.4
NOC22	N	POINDIMIE	25.1	21.1	38.3	47.8	93.8	47.2	13.3	77.9	3.8	210.4	29.2
NOC23	N	PONERIHOUEN	13.3	-6.4	39.4	46.7	97.8	17.8	34.1	80.4	4.8	731.8	19.3
NOC24	N	POUEBO	10.3	-15.7	43.4	46.6	99.7	52.7	17.4	80.6	5.5	671.0	5.5
NOC25	N	POUEMBOUT	4.8	-5.7	34.0	47.8	93.4	21.3	36.2	85.1	4.2	397.7	81.2
NOC26	N	POUM	5.6	•	38.0	46.4	97.4	0.0	52.4	81.9	4.3	1295.2	48.7
NOC27	N	POYA	13.5	-32.7	40.9	46.3	92.7	23.2	32.4	83.4	4.3	583.6	52.5
NOC28	S	SARRAMEA	3.3	35.3	42.0	47.8	96.9	25.3	22.2	82.4	4.8	487.9	82.0
NOC29	S	THIO	20.8	4.3	43.1	48.9	85.0	7.7	16.4	81.5	4.9	408.0	84.9
NOC30	N	TOUHO	13.1	14.0	38.5	47.4	97.2	33.5	14.7	73.8	4.7	287.2	21.3
NOC31	N	VOH	10.9	-14.6	38.6	48.6	96.5	11.7	36.1	82.6	4.4	881.1	55.5
NOC32	S	YATE	9.5	1.6	44.1	46.2	94.5	0.0	35.0	83.1	5.4	783.6	36.3

Identification des variables (le nom des variables est entre parenthèses)
V1 : Codes des communes (CODE)
V2 : Province d'appartenance, N=Nord, S=Sud, I=Iles Loyauté (PROVINCE)
V3 : Noms des communes (NOM)
V4 : Part de la populotion Néo-Calédonienne pour 1000 habitants (POPNC)
V5 : Evolution de la population communale 1976-1983 % (VAR76-84)
V6 : Part des 0-14 ans dans la population communale % (0-14ANS)
V7 : Part des femmes dans la population communale % (FEMMES)
V8 : Part des personnes nées en Nouvelle-Calédonie dans la papulation communale % (NENC)
V9 : Part des agriculteurs dans la population communale en activité % (AGRIC)
V10 : Part des salariés du secteur public dans la population communale en activité % (SALPUBLIC)
V11 : Part des personnes sachant écrire le français dans la population communale en activité % (ECRIT)
V12 : Nombre de personnes par résidence principale (POP/RESID)
V13 : Nombre de personnes pour 100 personnes en activité (DEPEN)
V14 : Part des résidences principales équipées de l'eau % (EAU)

• donnée manquante

figure n°0.1. Le tableau de données démographiques et économiques sur les communes de Nouvelle-Calédonie.

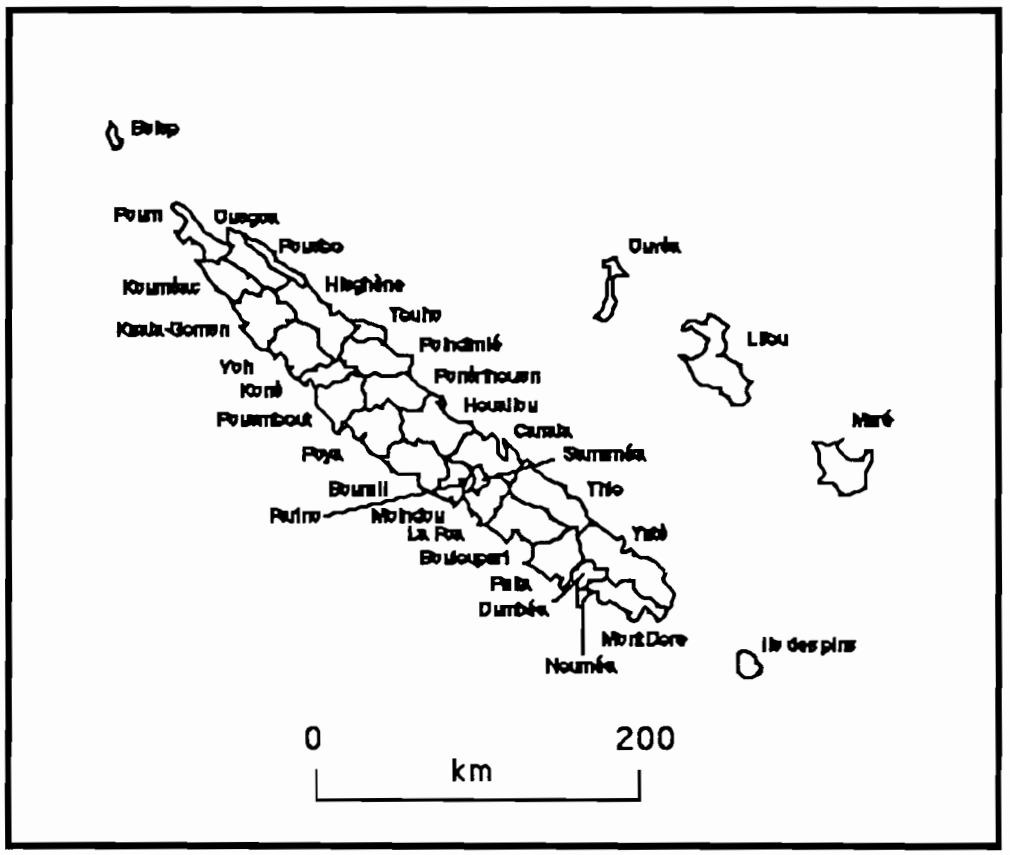


figure n°0.2. Les communes de Nouvelle-Calédonie.

Computer d'IBM ne verra le jour que cinq ans plus tard et le Macintosh d'Apple est un projet d'avant-garde. Convivialité et interactivité ne font pas encore partie du vocabulaire usuel de l'informatique.

Dans un tel environnement, et malgré une très grande puissance de calcul, l'utilisation de ces logiciels reste affaire de spécialiste et conduit le statisticien, ou plus simplement le détenteur de données, à se transformer en «pseudo-informaticien». Les enseignements de J. Tukey peuvent donc être compris comme une forme de réaction à l'usage bien souvent abusif des boîtes noires: en préconisant l'observation préalable et directe des données, on échappe à la tentation de soumettre un *job au batch* ou mieux, de se soumettre inconsidérément aux sorties de l'*output queue* !

Les ordinateurs graphiques comme le Macintosh ont sans aucun doute donné à l'analyse exploratoire l'environnement informatique qui lui faisait défaut pour atteindre un large public (malgré les efforts de l'équipe de Tukey sur IBM 360 et VAX DEC). Les concepts de bureau, d'icône et d'interaction ont permis le

développement de statisticiels très originaux qui n'éliminent pas la statistique courante, celle des modèles et des tests d'hypothèse, mais incite à ne plus considérer l'ordinateur sous le seul angle du calculateur.

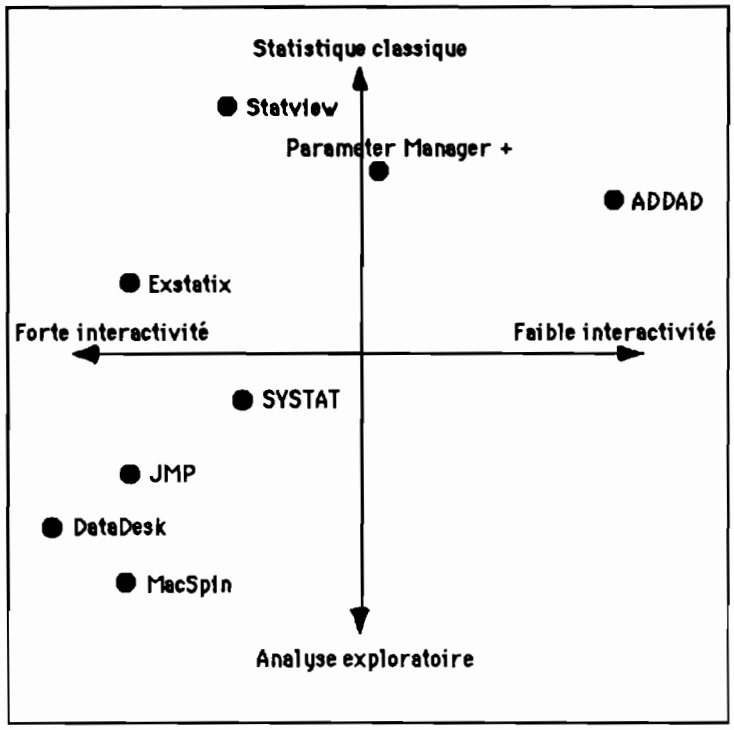


figure n° 0.3. Le positionnement relatif des statisticiels.

Statisticiels pour l'analyse exploratoire

Pour justifier le qualificatif statisticiel, un logiciel d'analyse des données doit assurer un nombre minimum de traitements usuels comme l'analyse de la variance, la régression multiple ou bien encore l'analyse factorielle. Par ailleurs, un grand nombre de méthodes statistiques considèrent que le tableau de données à analyser n'est en fait qu'un échantillon extrait d'une population plus large; les paramètres calculés doivent donc être assortis de tests de significativité qui donnent un seuil de confiance au-delà duquel les valeurs calculées ne peuvent être dues à des fluctuations aléatoires d'échantillonnage.

Ainsi, les tableurs du genre Excel ou les grapheurs comme CricketGraph ne doivent pas être considérés comme des logiciels d'analyse statistique, même s'ils offrent quelques possibilités limitées dans ce domaine comme le tracé d'histogrammes ou de diagrammes bivariés.

Si l'on écarte ces «croqueurs de nombres» (les *numbers crunchers* américains), les logiciels d'analyse statistique pour Macintosh demeurent assez nombreux. Il faut distinguer les logiciels généraux, ceux qui couvrent un large spectre d'applications, d'autres, plus spécialisés, qui répondent à des besoins particuliers, comme ceux des économètres, des laboratoires ou des études de marché. Les prix sont en général assez élevés, (plus de 5000 francs en moyenne).

Une étude publiée par le revue *Icônes* au début de l'année 1990 montre qu'il existe de notables différences entre les statisticiels pour Macintosh. A partir des analyses figurant dans ce dossier, il est possible d'exprimer le positionnement relatif de ces logiciels sur un graphique bivarié. L'abscisse représente le degré d'interactivité qui rend compte

de la plus ou moins bonne adaptation à l'interface du Macintosh. L'ordonnée traduit l'orientation vers l'exploration statistique par la présence des méthodes qui lui sont spécifiques (figure n° 0.3).

De notables différences apparaissent. Si le degré d'interactivité semble bon en général (le logiciel d'analyse des données de l'ADDAD est en fait l'adaptation d'une bibliothèque de programmes provenant d'un gros système, ce qui explique sa très faible convivialité), la distinction entre statistique classique et analyse exploratoire apparaît bien plus marquée.

Les logiciels qui nous intéressent ici sont ceux qui occupent le quadrant Sud-Ouest du graphique (en bas à gauche de la figure n° 0.3). On y trouve, proches les uns des autres, **SYSTAT, DataDesk, JMP** et **MacSpin**. Les trois premiers représentent de véritables encyclopédies statistiques: on y trouve une très grande variété de méthodes classiques remarquablement bien adaptées à l'interface du Macintosh. De son côté, **MacSpin** est entièrement dédié à l'exploration statistique. Les figures n° 0.4.A, 0.4.B et 0.4.C donnent les caractéristiques principales de ces quatre logiciels dans leur version disponible à la fin de l'année 1990.

NOM DU LOGICIEL	*Data Desk 3.0*	*JMP 1.0*	*Systat 5.0*	*MacSpin 2.0*
STATISTIQUE DESCRIPTIVE				
moyenne, écart-type, etc.	x	x	x	
fréquences	x	x	x	
TESTS PARAMETRIQUES				
Khi-deux	x	x	x	
t de Student	x	x	x	
TESTS NON-PARAMETRIQUES				
Kolmogorov-Smirnov	x		x	
Wilcoxon	x		x	
Kruskal-Wallis	x		x	
ANALYSE DE LA VARIANCE				
ANOVA	x	x	x	
ANCOVA		x	x	

figure n° 0.4.A. Les caractéristiques des statisticiels orientés vers l'analyse exploratoire: statistique descriptive et analyse de la variance.

NOM DU LOGICIEL	*Data Desk 3.0*	*JMP 1.0*	*Systat 5.0*	*MacSpin 2.0*
CORRELATION et REGRESSION				
Pearson R	x	x	x	
Spearman	x		x	
Kendall	x		x	
régression	x	x	x	
régression. polynomiale	x	x	x	
régression pas à pas	x	x	x	
régression non-linéaire	x		x	
ANALYSE DES DONNEES				
analyse en composantes principales	x	x	x	
analyse discriminante	x		x	
classification ascendantehiérarchique	x		x	
SERIES CHRONOLOGIQUES				
lissages			x	
autocorrélation			x	
ARIMA			x	
transformées de Fourier			x	
METHODES EXPLORATOIRES				
diagramme en tige et feuilles			x	
diagramme en boîte et moustaches	x	x	x	
graphiques bivariés interactifs	x	x	x	
matrice de graphiques bivariés	x	x	x	
brossage d'un nuage de points	x	x	x	x
tranchage d'un nuage de points	x			
toupie	x	x	x	x
animation 3D par masquage				x

figure n° 0.4.B. Les caractéristiques des statisticiels orientés vers l'analyse exploratoire: corrélation, analyse des données, séries chronologiques et méthodes exploratoires.

NOM DU LOGICIEL	*Data Desk 3.0*	*JMP 1.0*	*Systat 5.0*	*MacSpin 2.0*
GRAPHIQUES				
histogramme	x	x	x	
diagramme en boîte (box plot)	x	x	x	
diagramme en tige et feuilles	x		x	
graphiques bivariés (x,y)	x	x	x	x
droite de régression	x	x	x	
courbes de niveaux			x	
diagramme triangulaire			x	
graphiques trivariés	x	x	x	x
histogramme en 3 dimensions			x	
surface en 3 dimensions	x		x	
PREPARATION DES TABLEAUX				
sélection d'individus	x	x	x	x
calcul de nouvelles variables	x	x	x	x
pondération des observations		x	x	
recodages	x	x	x	x
GESTION DES DONNEES				
saisie à l'écran	x	x	x	x
valeurs manquantes	x	x	x	x
lecture/écriture de fichiers ASCII	x	x	x	x
nombre maximum de variables	mem	?	200	mem
nombre maximum d'individus	mem	?	disque	mem
PRIX INDICATIF EN FF	2 900	8 500	4 600	3 000

figure n° 0.4.C. Les caractéristiques des statisticiels orientés vers l'analyse exploratoire: graphiques, préparation des tableaux, gestion des données et prix (mem=mémoire).

LOGICIELS POUR L'EXPLORATION STATISTIQUE

Les logiciels **SYSTAT**, **DataDesk**, **JMP** et **MacSpin** présentents à leurs utilisateurs des visages bien différents. Ceci est dû, bien entendu, à l'image que leurs concepteurs ont voulu leur attribuer, en quelque sorte leur «personnalité» dans le monde de l'informatique. Cet argument lié au «marketing» n'est pas suffisant pour expliquer toutes leurs spécificités: ces logiciels sont, malgré le respect des normes propres au Macintosh, de conceptions générales (de *system design*) extrêmement diverses. L'interface utilisateur reflète ces particularités qu'il est utile de connaître car elles impliquent pour l'usager final des possibilités, mais aussi des limitations spécifiques.

Ainsi, selon la manière dont l'interface a été conçue, il est plus ou moins aisé de «creuser» une exploration, possible, ou impossible, d'interroger simultanément un graphique et le tableau de données auquel il est lié, etc. Autant que la variété des méthodes disponibles, la capacité d'un logiciel à interagir avec son utilisateur devient, en analyse exploratoire, un critère de choix essentiel.

1.1. SYSTAT

Considéré depuis longtemps comme le statisticiel de référence sur PC, **SYSTAT**, véritable encyclopédie de la statistique, doit combler les statisticiens les plus exigeants.

Edité par la société du même nom, **SYSTAT** est l'archétype du logiciel de statistique gouverné par un langage de commande proche du Basic. La version 5.0, disponible depuis le printemps 1990, dispose d'une interface plus «Macintosh» que précédemment puisque toutes les commandes sont accessibles à partir d'articles de menus déroulants. Systat s'adresse plus particulièrement aux statisticiens désirant convertir une partie de leurs applications, les plus légères, d'un gros ordinateur central vers un Macintosh. Notons qu'une version de ce logiciel existe aussi pour PC/PS sous MS-DOS, ce qui rend

possible le dialogue entre utilisateurs de ces principaux standards de la micro-informatique.

L'interface utilisateur se compose de trois fenêtres et d'une barre de menus. Après un biclic sur l'icône du logiciel apparaît la fenêtre nommée *SYSTAT Command* surmontée d'une barre composée de 7 menus (figure n° 1.1).

FILE réalise toutes les opérations nécessaires à l'ouverture, la sauvegarde ou la fermeture d'un fichier de données.

EDIT comprend les habituelles fonctions de copier/couper/coller.

DATA provoque le chargement du module de même nom nécessaire aux opérations de lecture des données dans un fichier texte, de calcul de nouvelles variables, etc.

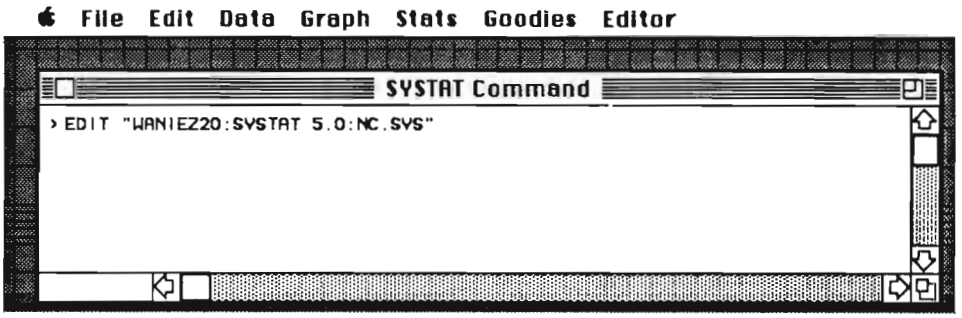
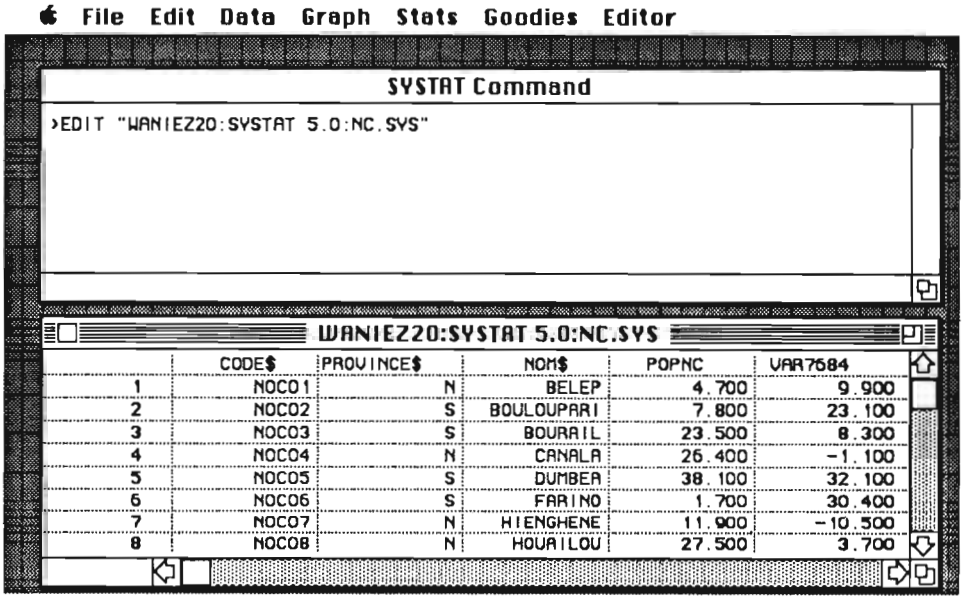


figure n° 1.1. SYSTAT: la fenêtre Systat Command.



	CODE$	PROVINCE$	NOM$	POPNC	VAR7684
1	NOCO1	N	BELEP	4.700	9.900
2	NOCO2	S	BOULOUPARI	7.800	23.100
3	NOCO3	S	BOURAIL	23.500	8.300
4	NOCO4	N	CANALA	26.400	-1.100
5	NOCO5	S	DUMBEA	38.100	32.100
6	NOCO6	S	FARINO	1.700	30.400
7	NOCO7	N	HIENGHENE	11.900	-10.500
8	NOCO8	N	HOUAILOU	27.500	3.700

figure n° 1.2. SYSTAT: la fenêtre Systat Command et celle d'édition du fichier "WANIEZ20:SYSTAT 5.0:NC.SYS".

GRAPH donne accès à toutes les fonctions graphiques du module SYGRAPH.

STAT donne accès à toutes les techniques d'analyse statistique disponibles.

GOODIES permet d'effacer ou de faire apparaître l'une ou l'autre des fenêtres, d'afficher la boîte à outils de dessin, d'afficher les options en cours, etc.

EDITOR charge l'éditeur qui n'est qu'un genre de tableur aux fonctions limitées à la saisie et à la correction des données.

La fenêtre *Systat command* reste en permanence en attente d'une commande. Une commande se présente sous la forme d'une chaîne de caractères qui ordonne au système de réaliser une opération donnée. Les commandes doivent être entrées derrière le signe >. Par exemple, si l'on souhaite éditer le fichier NC.SYS du dossier SYSTAT 5.0 du disque WANIEZ20, il faut commander (derrière le signe >) :

```
>EDIT "WANIEZ20:SYSTAT 5.0:NC.SYS"
```

Notons que la commande s'appelle EDIT et que ses paramètres sont compris entre les doubles apostrophes ("). Les paramètres disent à la commande sur quoi elle doit porter. Un retour-chariot provoque l'exécution de cette commande et l'affichage de la fenêtre de l'éditeur (figure n° 1.2).

Enfin, la troisième fenêtre, SYSTAT View, assure la visualisation des résultats d'un traitement, qu'ils soient numériques ou graphiques. Par exemple, en faisant les commandes GRAPH, puis PLOT FEMMES*DEPEN, le graphique bivarié s'affiche dans la fenêtre SYSTAT View (figure n° 1.3). En fonction du type de traitement, divers outils sont proposés à l'utilisateur. Ici, les outils pointeur, sélecteur et lasso permettent respectivement de désigner ou de sélectionner certains points du graphique qui sont alors marqués dans la fenêtre de l'éditeur par le symbole •.

La relation logique entre ces fenêtres facilite l'exploration des données: on découvre quelles observations forment un groupe dans le nuage de points et quelles sont celles qui sont isolées. C'est là le début de l'analyse exploratoire.

Depuis l'apparition de la version 5.0, il n'est plus obligatoire d'entrer directement des commandes souvent difficiles à mémoriser. On aboutit à un résultat semblable en activant d'abord le menu **GRAPH**, ce qui a pour effet d'inscrire la commande GRAPH dans la fenêtre *SYSTAT Command* (figure n° 1.4).

Puis, en sélectionnant le sous-menu **PLOT** (ce qui a pour effet d'inscrire la commande PLOT dans la fenêtre *SYSTAT Command*), et en choisissant les deux variables (figure n° 1.5) pour lesquelles on souhaite obtenir un graphique bivarié (ce qui a pour effet d'inscrire FEMMES*DEPEN derrière la commande PLOT dans la fenêtre *SYSTAT Command*), le graphique s'affiche dans la fenêtre *SYSTAT View*.

L'utilisation d'un langage de commande est assez inhabituel pour un logiciel fonctionnant sur Macintosh (mais la plupart des tableurs proposent

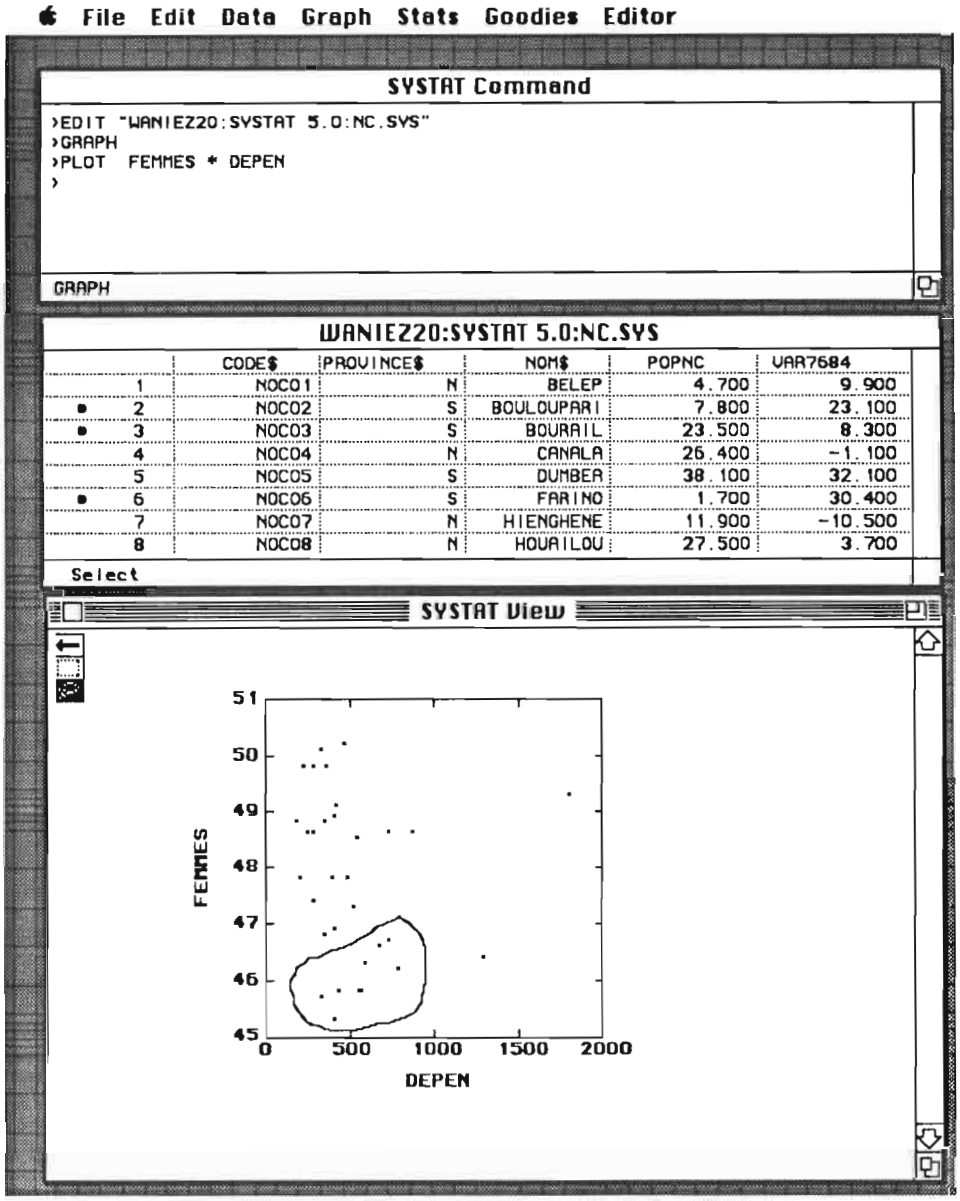


figure n° 1.3. SYSTAT: La fenêtre Systat Command, les fenêtres d'édition du fichier "WANIEZ20:SYSTAT 5.0:NC.SYS" et Systat View.

maintenant un langage de construction de macro-commandes regroupant plusieurs opérations). Le principal intérêt d'un tel langage réside dans la rédaction de programmes formés de commandes qui s'exécutent les unes après les autres: il suffit de changer le nom du fichier de données pour réitérer un traitement sur un autre ensemble de statistiques. Pour l'analyse exploratoire, cette option ne semble pas indispensable compte tenu du caractère inter-

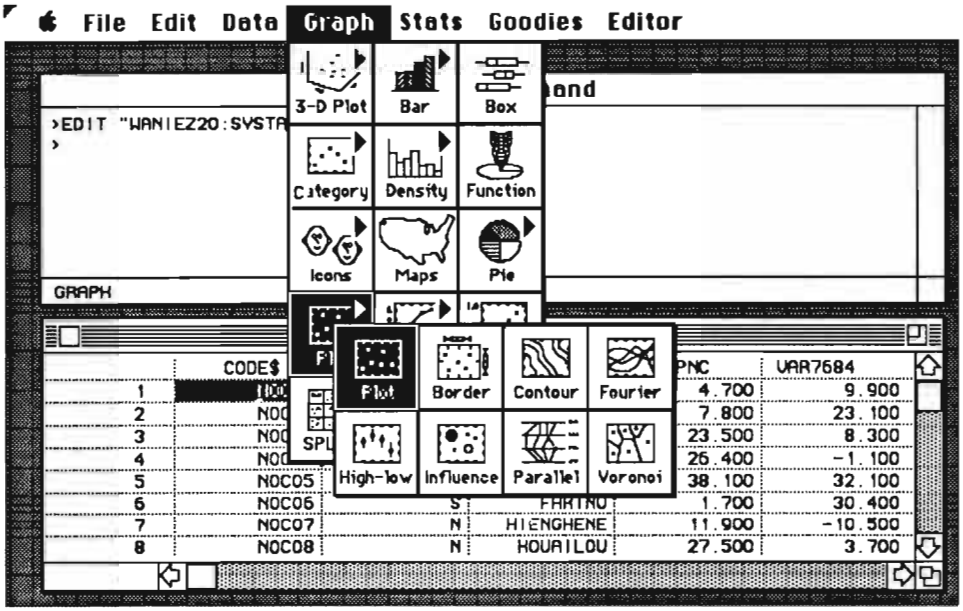


figure n°1.4. SYSTAT: l'utilisation des menus déroulants à la place des commandes.

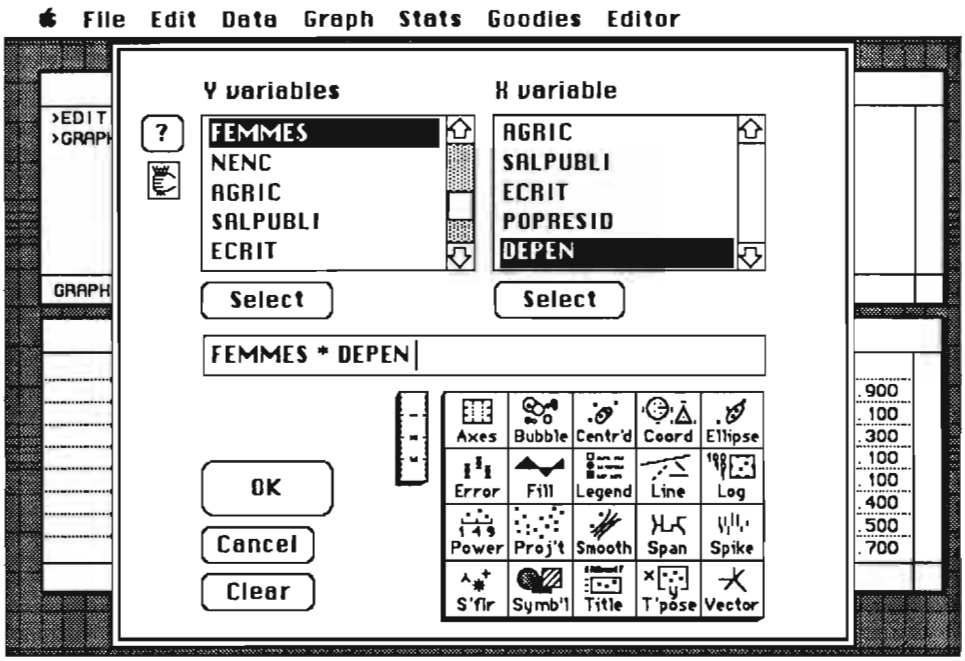


figure n°1.5. SYSTAT: la sélection des variables du graphique bivarié. Noter la présence d'un ensemble considérables d'options.

actif de cette démarche qui s'oppose à toute approche figée par un programme pré-défini. On notera enfin que **SYSTAT** dispose désormais d'une interface de type Macintosh qui le rend comparable à d'autres statisticiels disponibles pour cet ordinateur.

1.2. DataDesk

Comme l'indiquent son icône et son nom, **DataDesk** s'appuie sur le concept de bureau, si fréquent dans le monde du Macintosh, pour offrir à l'utilisateur une interface résolument «orientée objet».

Dès l'ouverture d'un fichier de données, plusieurs ensembles d'objets font leur apparition sur le bureau qui porte la marque **DataDesk 3.0**. On trouve tout d'abord la fenêtre *DATA* portant le nom du fichier ouvert (NC.DATA) et contenant une série d'icônes portant chacune le nom de la variable qu'elles figurent (figure n° 1.6): chaque variable est donc un «objet» variable qui va pouvoir être traité comme n'importe quel objet du bureau (sélection, déplacement, copier, couper et coller, etc.).

Sur la partie droite, on trouve, comme il se doit, pour «coller» aux concepts du Macintosh, l'icône du fichier de données et une corbeille (*Trash*) qui recevra les objets devenus inutiles.

Dans la partie supérieure de l'écran, 8 menus sont proposés:

FILE réalise toutes les opérations nécessaires à l'ouverture, la sauvegarde

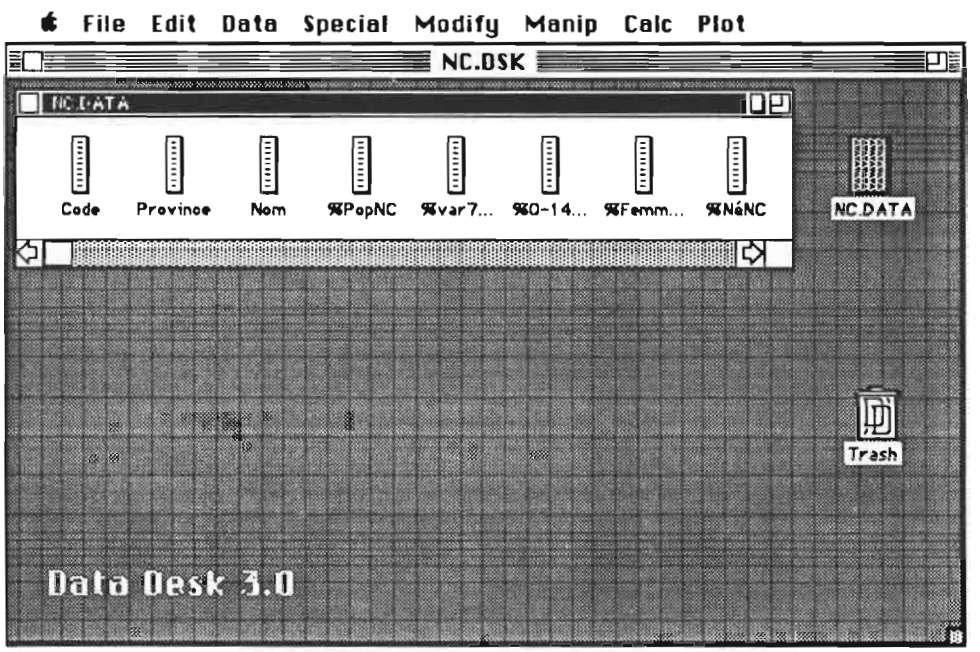


figure n° 1.6. DataDesk: le bureau avec ses icônes.

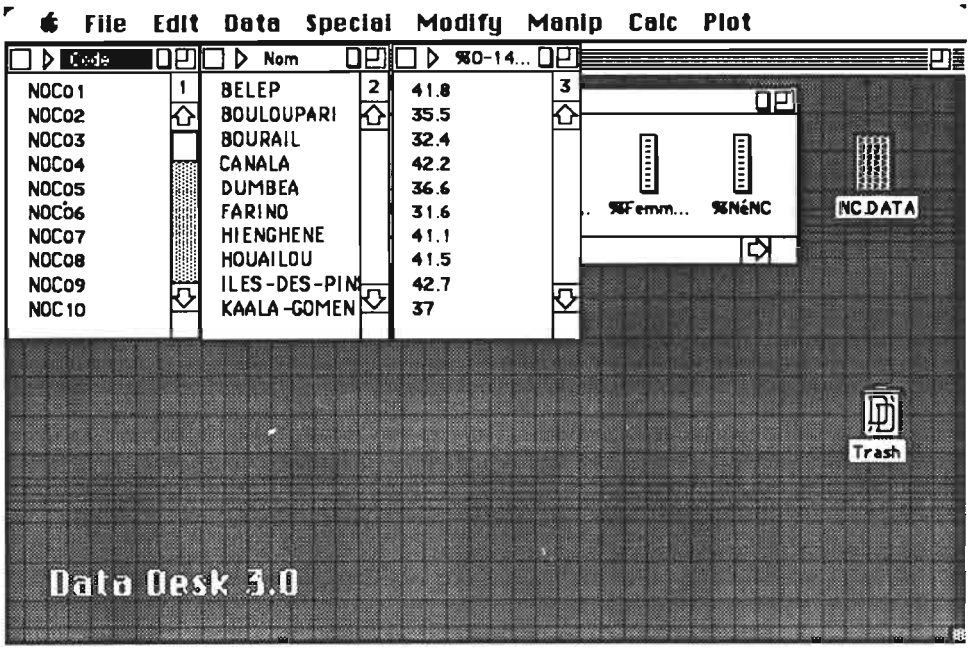


figure n° 1.7. DataDesk: Le bureau avec 3 variables ouvertes.

ou la fermeture d'un fichier de données.

EDIT comprend les habituelles fonctions de copier/couper/coller.

DATA donne accès aux fonction de création des dossiers et des fichiers de données ainsi que de leur contenu en proposant différents types d'icônes.

SPECIAL comprend diverses fonctions comme vider la corbeille ou recherche une icône particulière (lorsqu'il y en a trop pour la surface de l'écran).

MODIFY permet de modifier les graphiques et offre une large panoplie d'outils pour approfondir l'étude des graphiques.

MANIP donne accès à un ensemble de fonctions sur les variables comme, par exemple, le rangement des observations en ordre croissant ou décroissant.

CALC déclenche l'exécution des analyses statistiques.

PLOT réalise une grande variété de graphiques.

Contrairement à la majorité des logiciels traitant des données numériques, il n'y a pas dans **DataDesk** de tableau de données proprement dit. Plus précisément, il existe un tableau «virtuel» composé des variables que l'utilisateur a souhaité visualiser. Ceci se fait par un biclic sur l'icône de chaque variable désirée ou par la sélection multiple d'un ensemble d'icônes (un clic sur la première variable, puis un clic sur chaque variable supplémentaire avec la touche majuscule enfoncée, et choix de l'article **OPEN** du menu **DATA**). Les valeurs des variables ouvertes apparaissent alors dans des fenêtres différentes (figure n° 1.7).

Chacune des fenêtres des variables est indépendante des autres: on peut la déplacer, l'agrandir, la fermer, etc. Mais les lignes sont liées entre elles: un déplacement de l'ascenseur dans une des fenêtres se traduit par un déplacement simultané dans toutes les fenêtres ouvertes. Dans chaque fenêtre, les données peuvent être modifiées, sélectionnées, copiées et collées, etc.

Pour réaliser un traitement quelconque, un graphique bivarié, par exemple, il faut avoir préalablement sélectionné les icônes des variables entrant dans ce traitement. En choisissant l'article **SCATTERPLOT** du menu **PLOT**, une nouvelle fenêtre, contenant le graphique, fait son apparition sur le bureau qui commence à devenir très encombré. Heureusement, ces fenêtres peuvent être déplacées et redimensionnées (figure n° 1.8).

Le menu **MODIFY** donne accès à un ensemble d'articles permettant de modifier le graphique ou d'en poursuivre l'exploration plus avant. L'article TOOLS permet, en particulier, d'interagir avec le graphique, de manière semblable à ce que propose **SYSTAT** (figure n° 1.9).

Lorsque qu'un point est désigné par l'outil *ad hoc*, l'observation correspondante apparaît en fond inversé dans toutes les fenêtres de variables ouvertes (ici la variable Nom, figure n° 1.10). Cette liaison n'est pas limitée à deux fenêtres et peut être utilisée aussi bien pour des résultats numériques que pour des graphiques.

De plus, ce procédé est réflexif: une observation désignée par un biclic sur sa valeur dans une des fenêtres de variables ouvertes provoque sur le

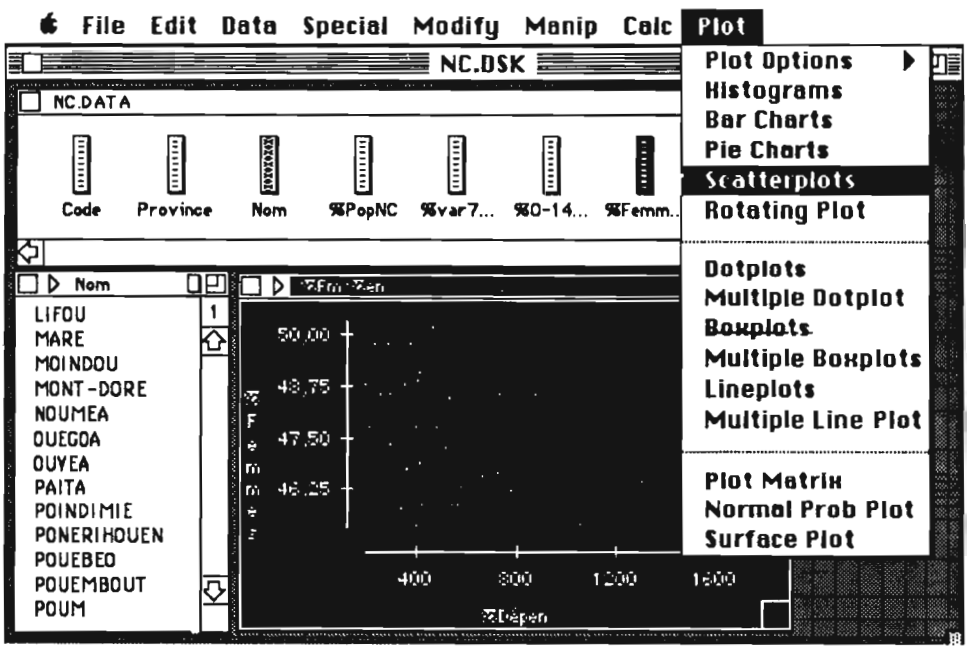


figure n° 1.8. DataDesk: un graphique bivarié sur le bureau.

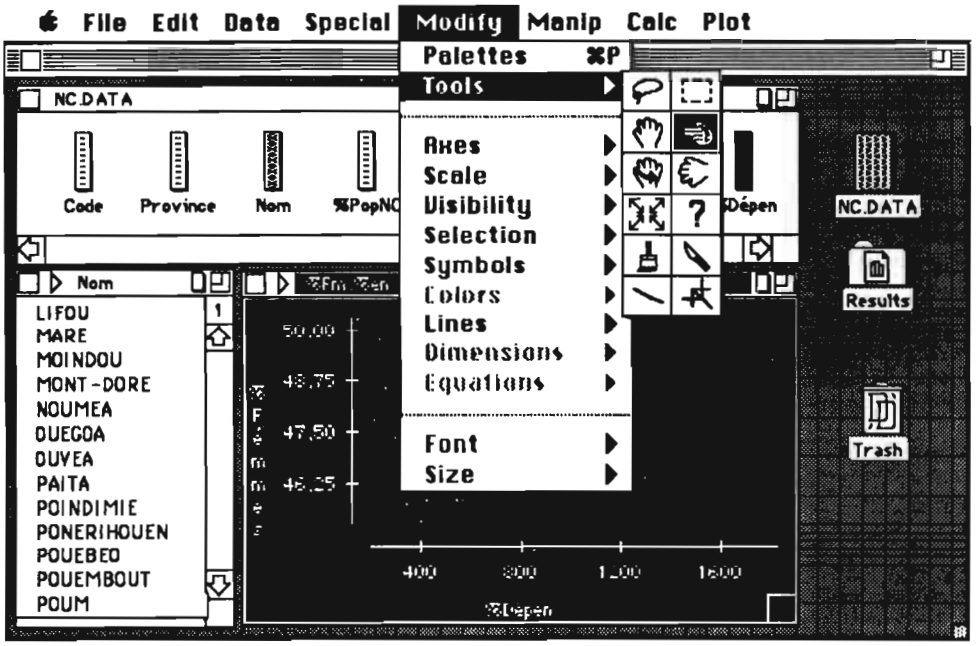


figure n° 1.9. DataDesk: le choix d'un outil dans le menu MODIFY.

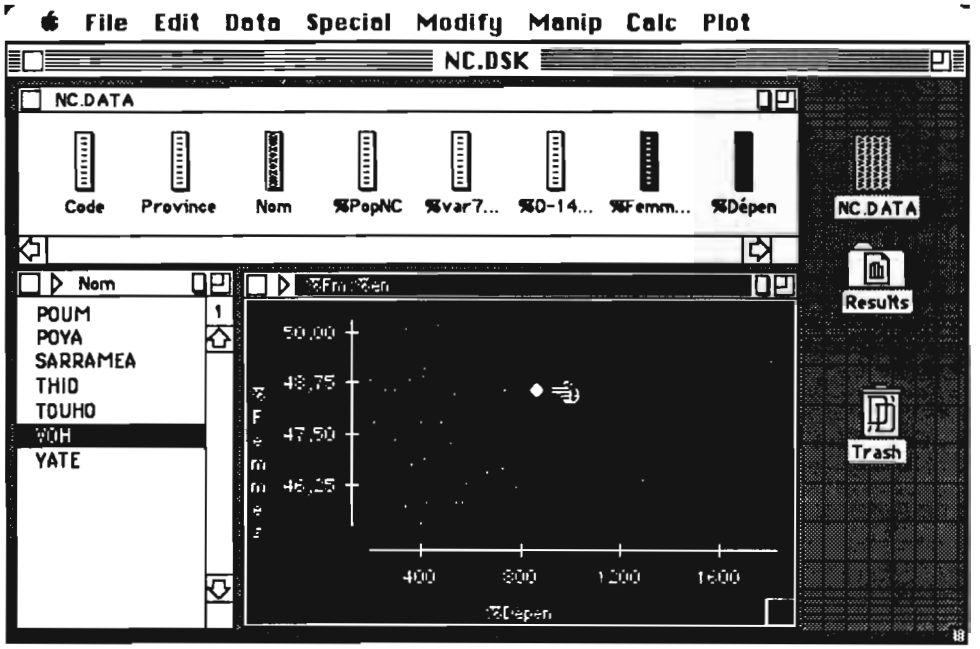


figure n° 1.10. DataDesk: la désignation d'une observation sur le graphique.

graphique une surintensité du point représentant cette observation sur le graphique (ici Nouméa, figure n° 1.11).

Enfin, sur le plan du dialogue entre l'utilisateur et le logiciel, **DataDesk** présente une originalité très intéressante: dès qu'un traitement est réalisé, son résultat est engrangé dans un dossier nommé **RESULTS**. Des icônes rappellent chaque type d'analyse. Un simple clic sur l'une d'elles conduit à un nouvel affichage de ce résultat (figure n° 1.12).

Par sa conception même, **DataDesk** apparaît bien adapté à l'approche exploratoire: les liaisons logiques entre les fenêtres et le stockage des résultats obtenus au cours d'une session de travail rendent l'interaction entre l'utilisateur et l'ordinateur vraiment efficace. A l'usage, néanmoins, on peut être quelque peu gêné par l'empilement des fenêtres sur le bureau. L'utilisation d'un écran de grand format (A4, ou mieux A3) apporte un confort d'utilisation qui permet de profiter encore plus de la convivialité de ce remarquable logiciel.

1.3. JMP

Avec **JMP**, prononcer *jump*, l'entrée (ou, comme le suggère l'icône, le saut) de SAS Institute dans le monde du Macintosh constitue un événement. Disons tout de suite, à l'intention des spécialistes, que **JMP** n'est pas une nouvelle version du célèbre *Statistical Analysis System* pour Macintosh, ni même l'un de ses nombreux modules. Il s'agit en fait d'un «prototype de ce que

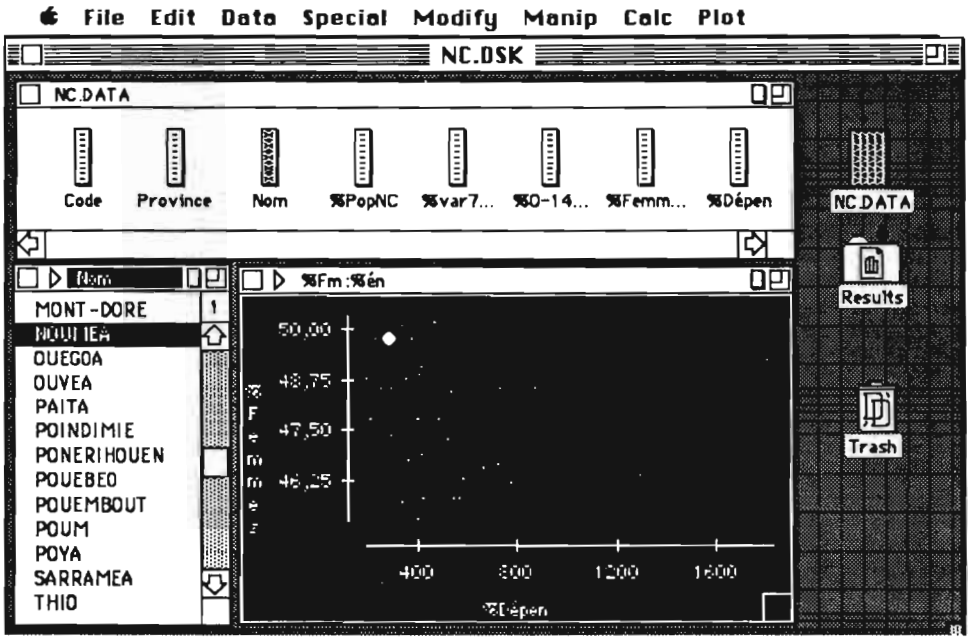


figure n°1.11. DataDesk: la désignation d'une observation sur la variable Nom et sa localisation sur le graphique.

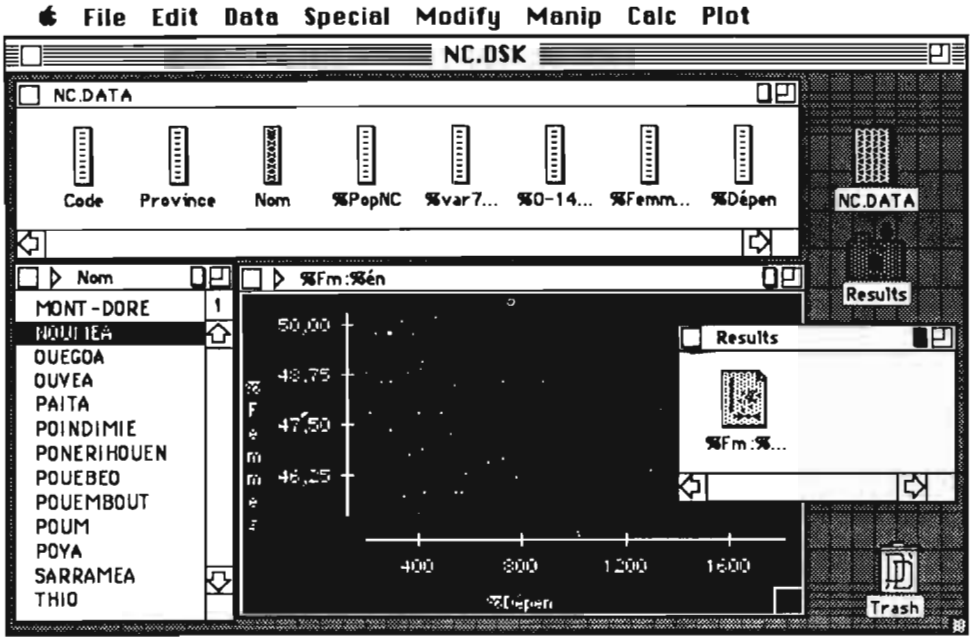


figure n° 1.12. DataDesk: l'enregistrement des résultats dans le dossier RESULTS.

sera prochainement le système SAS en matière de statistiques et de représentations interactives des données».

L'ouverture d'un fichier de données se traduit par l'affichage d'un tableau qui, sous un air sans doute familier aux utilisateurs de tableurs, cache une extraordinaire concentration d'intelligence. En effet, chaque variable possède deux cases qui sont en fait des menus déroulants permettant de définir leur rôle dans les traitements et leur type sur le plan statistique (figure n° 1.13).

Le type statistique correspond en fait aux trois échelles de mesure courantes:

- L'échelle d'intervalle, ainsi nommée car elle permet d'exprimer la distance d'une observation par rapport à une origine (la valeur zéro) choisie arbitrairement et grâce à une unité de mesure constante sur toute l'étendue des valeurs possibles (comme, par exemple, des températures). Notons que **JMP** considère toutes les données quantitatives comme relevant de l'échelle d'intervalle, ce qui est un abus de langage sans grande conséquence sur les traitements auxquels ces données seront soumises.
- L'échelle ordinale appliquée lorsqu'un rang peut être affecté à chaque observation en fonction d'une mise en ordre ou d'un classement pour un critère donné (comme les modalités «un peu», «beaucoup», «passionnément»).
- L'échelle nominale dont les valeurs sont des modalités exprimant des qualités ou des situations sans qu'aucun ordre particulier puisse être identifié (comme des noms de famille ou des couleurs).

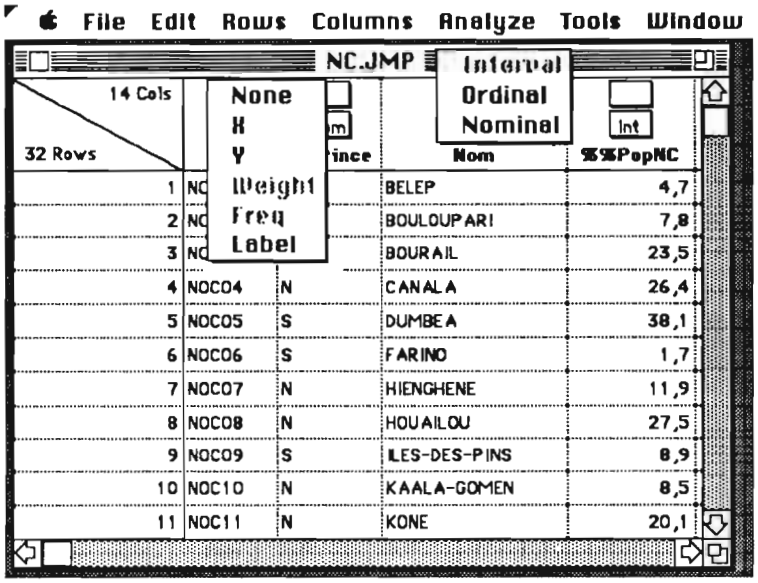


figure n°1.13. JMP: le tableau de données avec ses menus déroulants destinés à définir le type et le rôle de chaque variable.

Cette énumération n'aurait pas sa place ici si elle n'était l'une des bases de la conception du logiciel. En effet, la sélection d'un type de variable définit le jeu de traitements qui pourra lui être appliqué. Par exemple, on ne peut pas faire une régression avec les codes des provinces! Ceci constitue un garde-fou important lorsque le logiciel est utilisé à des fins didactiques: on ne peut faire n'importe quoi, n'importe comment. Mais l'«intelligence» de **JMP** va bien au-delà de la prudence pédagogique: en fonction de la plate-forme statistique choisie par l'utilisateur, c'est telle ou telle autre technique statistique qui est automatiquement sélectionnée.

On accède à l'une des six plates-formes statistiques par le menu **ANALYZE**. Une plate-forme est une fenêtre interactive qui permet d'analyser les données, d'explorer ies graphiques et d'enregistrer les résultats obtenus. Pour réaliser une analyse, il faut choisir l'échelle de mesure et le rôle de chaque variable. Les variables Y sont considérées comme «exogènes» (dites aussi «variables dépendantes» ou «variables à expliquer», etc.): elles représentent le phénomène objet de l'étude. Les variables X sont des variables «endogènes» (dites aussi «variables indépendantes» ou «variables explicatives», etc.): elles sont censées être en relation avec les valeurs des variables Y.

- **Distributions of Y's**: décrit la distribution de chaque variable Y à l'aide d'histogrammes ainsi que d'autres graphiques et paramètres statistiques.
- **Fit Y by X**: ajuste une variable Y

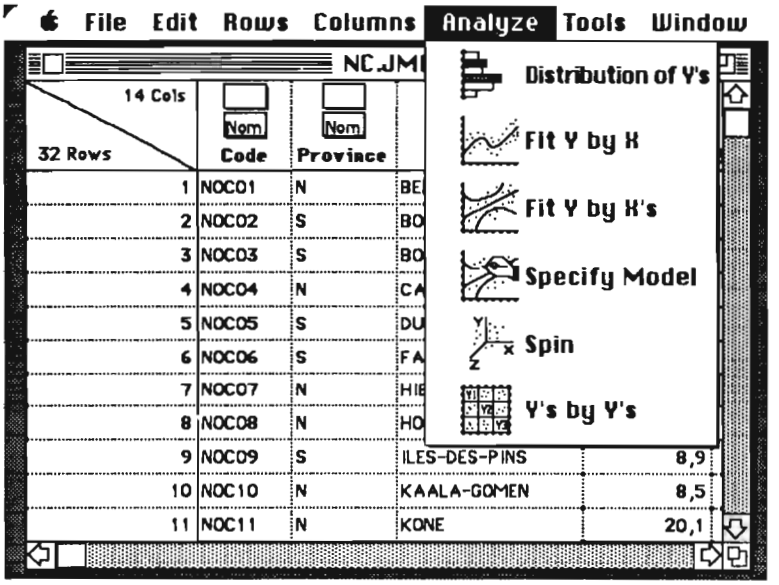


figure n°1.14. JMP: le tableau de données et les plates-formes statistiques du menu ANALYZE.

par une variable X, et cela conformément aux échelles de mesure adoptées. Selon le cas, il s'agit de corrélation, de régression simple, polynomiale ou logistique, de tableaux croisés ou de tests de comparaison de moyennes entre groupes.

• **Fit Y by X's**: ajuste une variable Y par toutes les variables X, et cela conformément aux échelles de mesure adoptées. Selon le cas, il s'agit de régression multiple ou d'analyse de la variance ou de la covariance.

• **Specify Model**: permet de réaliser des analyses de covariance avec un grand nombre d'options.

• **SPIN**: donne accès à la méthode de la toupie.

• **Y's by Y's**: calcule les corrélations entre plusieurs variables Y et affiche le *Scatterplot Matrix*, l'un des graphiques centraux de l'analyse exploratoire.

Tous les traitements proposés par le menu **ANALYZE** se caractérisent par une interactivité poussée à l'extrême. Par exemple, après avoir donné le rôle X à la variable %0-14ans, et le rôle Y à Pop/Resid, et activé l'article **Fit Y by X** le logiciel trace un graphique bivarié (figure n° 1.15). Un clic sur un point a pour effet de le souligner dans le tableau de données (comme le faisaient les deux précédents logiciels).

L'étude de cette relation pourrait très bien s'arrêter là. Mais le bouton **Fitting** réalise l'ajustement d'une courbe, ici une droite, qui montre que les communes les plus jeunes sont aussi celles où les résidences principales sont le plus peuplées, ce que confirme le coefficient de détermination (Rsquare=0.51).

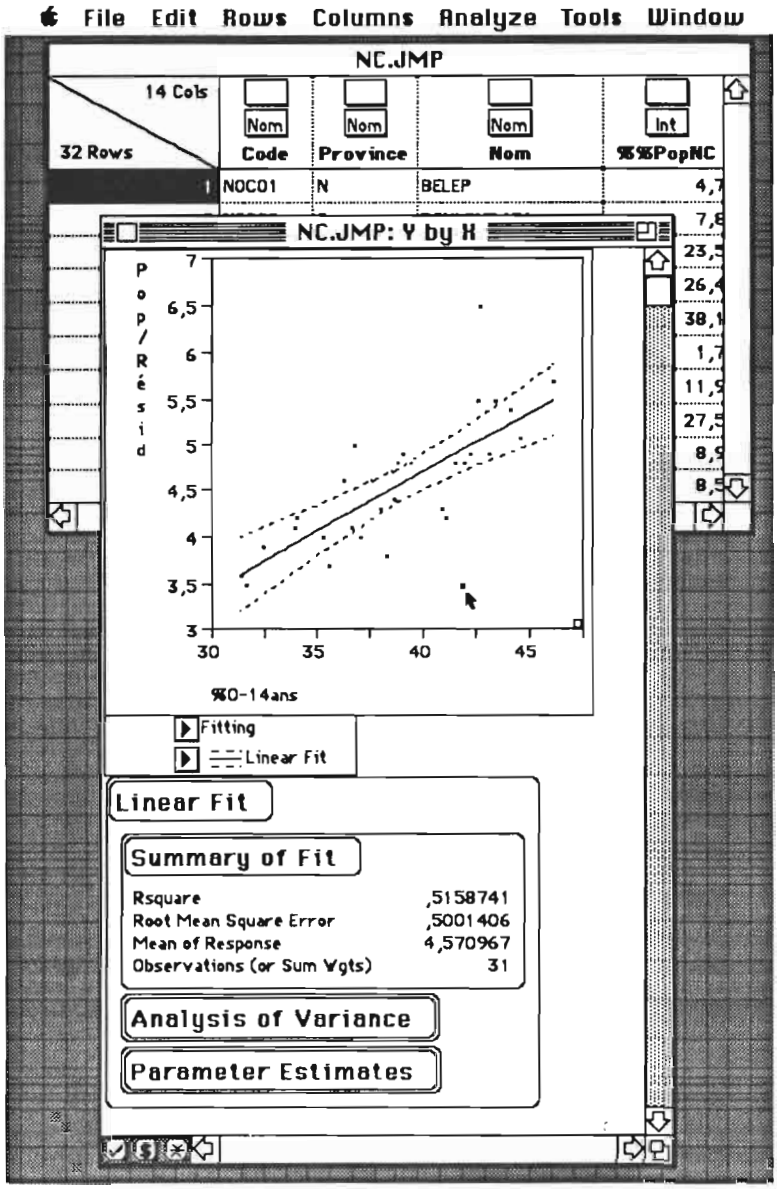


figure n°1.15. JMP: Un graphique bivarié accompagné d'un ajustement linéaire.

L'étude de cette relation pourrait encore s'arrêter là. Mais **JMP** propose les deux boutons *Analysis of Variance* qui donne le tableau d'analyse de la variance, et *Parameter Estimate* qui affiche les coefficients de la droite de régression. Pour poursuivre l'analyse, il suffit de cliquer sur l'un ou l'autre de ces boutons pour faire apparaître les résultats des traitements.

Les autres menus sont beaucoup moins importants sur le plan conceptuel. Mais les articles auxquels ils donnent accès démontrent, s'il en est encore besoin, l'imagination que les

programmeurs de SAS Institute ont déployée pour faire de **JMP** un produit extrêmement convivial.

FILE réalise toutes les opérations nécessaires à l'ouverture, la sauvegarde ou la fermeture d'un fichier de données. De plus, l'article **TRANSFORM** donne accès à toute une panoplie d'options très utiles pour produire de nouveaux tableaux de données à partir de tableaux existant déjà: association de tableaux, tris sur plusieurs clés, sous-tableaux par sélection sur plusieurs critères, etc.

EDIT comprend les habituelles fonctions de copier/couper/coller. L'article **JOURNAL** permet de conserver dans un fichier texte le contenu des fenêtres successivement actives ce qui peut être très pratique pour la recherche d'éventuelles erreurs de manipulation.

ROWS assure un ensemble de fonctions portant sur les lignes du tableau de données. En particulier, grâce à l'article EXCLUDE/INCLUDE, il devient possible d'exclure ou d'inclure à nouveau toute observation d'une série de traitements, sans avoir à l'effacer du fichier de données.

COLUMNS permet d'agir sur les variables définissant leurs rôles dans les analyses (variables X ou Y, concurremment aux menus déroulants présents dans chaque colonne). Ce menu autorise aussi l'ajout de nouvelles colonnes, ou le déplacement à l'intérieur du tableau de colonnes existant déjà ainsi que leur suppression réelle ou virtuelle (on ne voit plus ces colonnes, mais elles existent encore dans le fichier).

TOOLS comprend les articles d'une boîte à outils qui définissent l'action de la souris à un moment donné. Dans une fenêtre contenant des résultats, ces outils permettent de couper, de déplacer ou d'interroger JMP qui en précise alors la signification statistique.

WINDOW simplifie l'accès aux multiples fenêtres qui s'ouvrent pour chaque plate-forme statistique. Les noms de ces fenêtres en sont les principaux articles: une fois sélectionnées, ces fenêtres deviennent actives et accessibles, à la lecture par de classiques ascenseurs.

On retiendra de cette présentation que **JMP**, tout comme **DataDesk**, conduit l'utilisateur à analyser ses données pas à pas, afin de lui éviter de se noyer dans les chiffres, tout en gagnant sur le temps de traitement, ce qui, avec un micro-ordinateur demeure aujourd'hui encore un atout très important. Grâce à **JMP**, on peut, à tout moment, approfondir une relation particulière ou explorer une nouvelle voie ce qui correspond bien au précepte central de l'Analyse Exploratoire.

1.4. MacSpin

MacSpin n'est pas à proprement parler un logiciel d'analyse statistique mais, comme l'indique l'éditeur lui-même, un logiciel d'analyse graphique des données. Il s'agit néanmoins d'un système très original d'étude des données statistiques qui rendra bien des services à tous ceux qui ne veulent pas (ou ne peuvent pas) suivre les lois contraignantes de la statistique classique. D'ailleurs, comme l'indique intelligemment la documentation, on aura

intérêt à utiliser **MacSpin** conjointement avec un véritable statisticiel afin de préciser les structures découvertes par l'estimation plus précise de leurs paramètres statistiques.

MacSpin est entièrement dédié à l'une des méthodes propres à l'analyse exploratoire, la toupie (*spiner*). Son fonctionnement sera présenté dans la troisième partie de cet ouvrage.

La plus grande partie du bureau de **MacSpin** est occupée par le nuage de points blancs sur fond noir dont les coordonnées ne sont que les valeurs sur les variables X, Y et Z sélectionnées dans la fenêtre supérieure droite de l'écran nommée Variables. Le système d'axes permet de savoir sous quel angle le nuage de points est observé dans cette galaxie (figure n° 1.16).

Sur la gauche, on trouve une boîte à outils qui assure les fonctions d'identification et de sélection des points et, surtout, la rotation du nuage de points autour de l'un des trois axes, d'où le nom de toupie donnée à cette méthode.

Comme les logiciels précédents, **MacSpin** simplifie l'exploration du nuage de points. Par exemple, lorsqu'on clique sur un point avec l'outil d'identification (le petit cercle en haut et à gauche de la boîte à outils), son nom apparaît en regard. La fenêtre Observations réalise, à l'envers, la même opération: un clic sur un nom ou un groupe de noms provoque le soulignement des points correspondants du nuage. Mais, dans tous les cas, les lignes des observations sélectionnées sont également soulignées dans le tableau de données auquel on accède par le menu

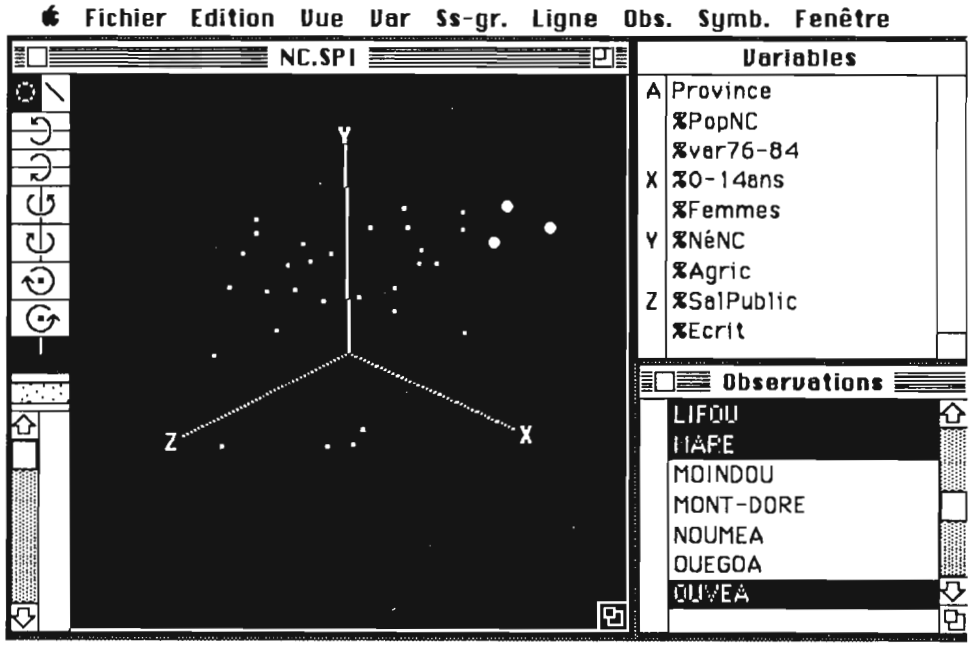


figure n° 1.16. MACSPIN: le bureau.

Fichier Edition Vue Var Ss-gr. Ligne Obs. Symb. Fenêtre

NC.SPI

	Province	%PopNC	%var76-84	%0-14ans	%Femmes
BELEP	1	4,7	9,9	41,8	49,3
BOULOUPARI	2	7,8	23,1	35,5	45,3
BOURAIL	2	23,5	8,3	32,4	45,7
CANALA	1	26,4	-1,1	42,2	47,3
DUMBEA	2	38,1	32,1	36,6	48,6
FARINO	2	1,7	30,4	31,6	45,8
HIENGHENE	1	11,9	-10,5	41,1	48,6
HOUAILOU	1	27,5	3,7	41,5	49,1
ILES-DES-PI...	2	8,9	17,5	42,7	45,8
KAALA-GO...	1	8,5	-13,6	37	48,6
KONE	1	20,1	17,7	36,8	48,5
KOUMAC	1	9,7		36,2	50,2
LA-FOA	2	14,4	5,1	35,3	49,8
LIFOU	3	55,3	7,2	44,5	50,1
MARE	3	31,7	10,9	46,1	49,8
MOINDOU	2	2,6	-2,3	33,9	46,8
MONT-DORE	2	100,5	37,1	37,6	48,8
NOUMEA	2	413,5	7,2	31,4	49,8
OUEGOA	1	10,1	-3	39,1	45,8
OUVEA	3	19,1	-0,2	42,6	48,8
PAITA	2	33,3	41,9	38,8	46,9

figure n°1.17. MACSPIN: le tableau de données sur lequel on a défini un groupe d'observations.

EDITION. Il existe donc un lien logique entre toutes ces fenêtres.

La barre supérieure de l'écran propose un ensemble de menus. Mis à part les classiques **FICHIER** et **EDITION**, les autres menus assurent soit le contrôle de l'environnement de travail, soit la définition et l'étude des observations statistiques.

VUE contrôle le graphique en offrant le choix entre plusieurs positions d'origine pour les axes et en permettant l'affichage des points en noir sur fond blanc ou en blanc sur fond noir.

VAR comprend tous les articles nécessaires au recodage des variables existant déjà dans le fichier en cours d'analyse, ou à la création de nouvelles variables comme, par exemple, des pourcentages ou des rapports.

SS-GR. permet de réunir plusieurs sous-groupes et d'en extraire les individus qui composent leur intersection.

LIGNE est un menu qu'il faut utiliser conjointement avec l'outil ligne de la boîte à outils (en haut à droite). En joignant un ensemble de points avec cet outil, on obtient une ligne brisée reliant des points aux caractéristiques proches. On peut tracer et enregistrer plusieurs lignes sur le même graphique.

OBS. isole ou exclut un groupe de points du graphique et permet de rechercher un point donné parmi tous les points du graphique.

SYMB. donne une palette de symboles simplifiant l'identification d'individus ou de groupes d'individus particuliers.

FENETRE permet de choisir les fenêtres devant apparaître sur l'écran ainsi que leur mode d'affichage (superposé, etc.).

D'autres fenêtres composent l'interface utilisateur de MacSpin comme GROUPES et LIGNES qui contiennent respectivement la définition des lignes tracées par l'utilisateur et des groupes d'observations qu'il a pu constituer.

Pour conclure ce chapitre sur les interfaces utilisateur, tentons de dégager les caractéristiques communes aux logiciels d'analyse statistique orientés vers l'analyse exploratoire.

En premier lieu, tous ces logiciels se caractérisent par une très grande interactivité, comme il se doit sur Macintosh. Seul **SYSTAT** est gouverné par un langage de commande qui, pour l'Analyse Exploratoire demeure de peu d'utilité. Cette interactivité se traduit par une avalanche de menus déroulants souvent placés hors de la traditionnelle barre de menus, par une foule de boutons de toutes formes, par le déplacement d'objets variés (points, icônes, etc.), en bref, par un véritable réseau d'actions dans lequel l'utilisateur doit véritablement naviguer au plus proche. Ainsi conçue, l'interactivité constitue une véritable richesse dont l'acquisition, loin d'être facile et immédiate, passe par un véritable apprentissage.

Le deuxième trait commun réside dans les relations que ces logiciels établissent entre le tableau de données d'une part et les résultats de traitement d'autre part, qu'ils soient graphiques ou numériques. Ceci simplifie énormément le travail de lecture, de compréhension et de vérification des traitements. Grâce aux liens dynamiques entre fenêtres (toute modification sur un objet, une valeur, ou un graphique figurant dans une fenêtre est immédiatement répercutée dans les autres fenêtres contenant cet objet), il n'est plus nécessaire d'imprimer les sorties avant la fin de l'analyse.

Véritables chefs-d'œuvres de conception informatique **SYSTAT**, **DataDesk**, **JMP** et **MacSpin** proposent une très grande variété de méthodes exploratoires univariées, bivariées et multivariées présentées dans les chapitres qui suivent.

EXPLORATIONS UNIVARIÉES

De même qu'en statistique classique, l'analyse exploratoire commence par un examen attentif des distributions des variables. Elles expriment différentes propriétés mesurées sur chaque individu et doivent être examinées de trois façons complémentaires: la localisation, l'étendue et la forme.

La localisation d'une distribution s'exprime par une valeur caractéristique de l'ensemble des valeurs prises par une variable. C'est elle qui résume le mieux l'ensemble des valeurs. Habituellement, la localisation s'exprime à l'aide de paramètres comme la moyenne arithmétique, le mode ou la médiane.

L'étendue d'une distribution exprime la dispersion des individus. Comme précédemment, on résume souvent l'étendue par une valeur unique, l'écart-type par exemple.

La forme d'une distribution est une caractéristique un peu plus difficile à appréhender comme en témoignent les adjectifs suivants: normale, en cloche, symétrique, uni-modale ou pluri-modale, etc.

La description de la localisation, de l'étendue et de la forme des variables permet d'accéder à une bonne connaissance élémentaire des données. L'acquisition de cette familiarité passe par l'emploi de techniques de description qui prennent en général la forme de résumés numériques (moyenne et écart-type) ou graphiques (histogrammes).

L'analyse exploratoire considère que l'analyse des formes des distributions statistiques est au moins aussi importante que l'étude des localisations ou des étendues. Bien entendu, ces formes peuvent s'exprimer par leur équation caractéristique; mais qui peut aisément identifier la forme d'une distribution à à l'aide de son équation, mis à part les statisticiens professionnels? En privilégiant la représentation des formes au moyen de graphiques, on se met en situation non seulement de mieux en percevoir les aspects les plus subtils, mais aussi d'en communiquer aisément les principales caractéristiques.

Un autre parti-pris de l'analyse exploratoire, corollaire du premier, revient à ne considérer les résumés

numériques que pour ce qu'ils sont: des raccourcis. Ce n'est qu'en fonction des enseignements apportés par l'examen des distributions qu'on peut ensuite choisir les paramètres numériques les mieux à même de résumer les caractéristiques des données.

L'analyse doit commencer par les données elles-mêmes, et non par leur résumé. En effet, les données étant souvent difficiles à obtenir, il faut éviter, au moins dans un premier temps, de les réduire trop brutalement à une information schématique.

Il faut tout d'abord rappeler qu'il suffit bien souvent de procéder à une simple mise en ordre des données pour les faire parler; d'où l'omniprésence des techniques de tri dans les logiciels traitant des données numériques.

La seconde méthode exploratoire abordée ici, le diagramme en tige et feuilles, facilite la perception rapide de la forme d'une distribution relative à une série de données.

La description de cette allure générale doit néanmoins être complétée par des paramètres plus précis. En analyse exploratoire, les résumés numériques résistants, peu sensibles aux valeurs exceptionnelles, sont préférés aux paramètres plus classiques comme la moyenne et l'écart-type.

Enfin, le diagramme en boîte et moustaches permet de visualiser de nombreuses caractéristiques comme, par exemple, les queues des distributions.

2.1. Mettre de l'ordre dans les données

La technique la plus facile pour examiner les valeurs d'une variable consiste à classer les individus dans l'ordre croissant ou décroissant des valeurs. Une lecture même superficielle d'un tel tableau ordonné (figure n° 2.1) permet de remarquer que:

- les taux de variation de la population sont plus souvent positifs que négatifs, et qu'il existe deux valeurs manquantes figurées par un point (•).
- la part des agriculteurs représente plus de la moitié de la population active dans une seule commune (NOC24, Pouébo), et qu'elle reste en général inférieure à 30%.
- 30 communes sur 32 renferment moins de 10% de la population de Nouvelle Calédonie, Nouméa (NOC18) représentant à elle seule 41.3% de la population totale du Territoire.

Avec ces observations élémentaires, on dispose déjà d'une information quelque peu élaborée, véhiculée par un commentaire qui apparaît soit très généraliste (taux de variation de la population négatif ou positif), soit très exceptionnaliste (poids de Nouméa). Il reste néanmoins assez difficile de se faire une bonne idée de la forme des distributions.

2.1.1. SYSTAT

Pour réaliser un tri avec **SYSTAT**, il faut introduire le tableau de données dans l'éditeur à l'aide de la commande EDIT ou bien en cochant le bouton EDIT

dans la fenêtre de dialogue affichée consécutivement à la sélection de l'article **OPEN** du menu **FILE** (figure n° 2.2). Le tableau de données se surimpose alors à la fenêtre de commande (figure n° 2.3).

Le choix de l'article **SORT** du menu **DATA** conduit à un dialogue permettant de choisir la ou les variables sur lesquelles le tri doit être réalisé. Cette sélection se fait de manière très simple: on choisit une variable en cliquant sur son nom et en appuyant sur le bouton SELECT (figure n° 2.4). Ici, une seule variable est retenue; il s'agit du pourcentage d'agriculteurs dans la population active (AGRIC).

Code	VAR76-84	Code	AGRIC	Code	POPNC
NOC27	-32,7	NOC26	0	NOC06	1,7
NOC24	-15,7	NOC32	0	NOC16	2,6
NOC31	-14,6	NOC18	0,1	NOC28	3,3
NOC10	-13,6	NOC17	0,6	NOC01	4,7
NOC07	-10,5	NOC05	0,8	NOC25	4,8
NOC23	-6,4	NOC12	2,3	NOC26	5,6
NOC25	-5,7	NOC01	2,6	NOC02	7,8
NOC19	-3,0	NOC21	2,6	NOC10	8,5
NOC16	-2,3	NOC10	4,8	NOC09	8,9
NOC04	-1,1	NOC29	7,7	NOC32	9,5
NOC20	-0,2	NOC02	8,2	NOC12	9,7
NOC32	1,6	NOC06	8,5	NOC19	10,1
NOC08	3,7	NOC11	8,9	NOC24	10,3
NOC29	4,3	NOC31	11,7	NOC31	10,9
NOC13	5,1	NOC13	13,1	NOC07	11,9
NOC14	7,2	NOC03	13,2	NOC30	13,1
NOC18	7,2	NOC04	16,9	NOC23	13,3
NOC03	8,3	NOC08	17,0	NOC27	13,5
NOC01	9,9	NOC23	17,8	NOC13	14,4
NOC15	10,9	NOC09	20,8	NOC20	19,1
NOC30	14,0	NOC25	21,3	NOC11	20,1
NOC09	17,5	NOC16	21,7	NOC29	20,8
NOC11	17,7	NOC27	23,2	NOC03	23,5
NOC22	21,1	NOC28	25,3	NOC22	25,1
NOC02	23,1	NOC20	25,4	NOC04	26,4
NOC06	30,4	NOC07	29,9	NOC08	27,5
NOC05	32,1	NOC15	30,1	NOC15	31,7
NOC28	35,3	NOC30	33,5	NOC21	33,3
NOC17	37,1	NOC19	38,2	NOC05	38,1
NOC21	41,9	NOC14	43,2	NOC14	55,9
NOC12	•	NOC22	47,2	NOC17	100,5
NOC26	•	NOC24	52,7	NOC18	413,5

figure n° 2.1. Les trois variables triées dans l'ordre décroissant.

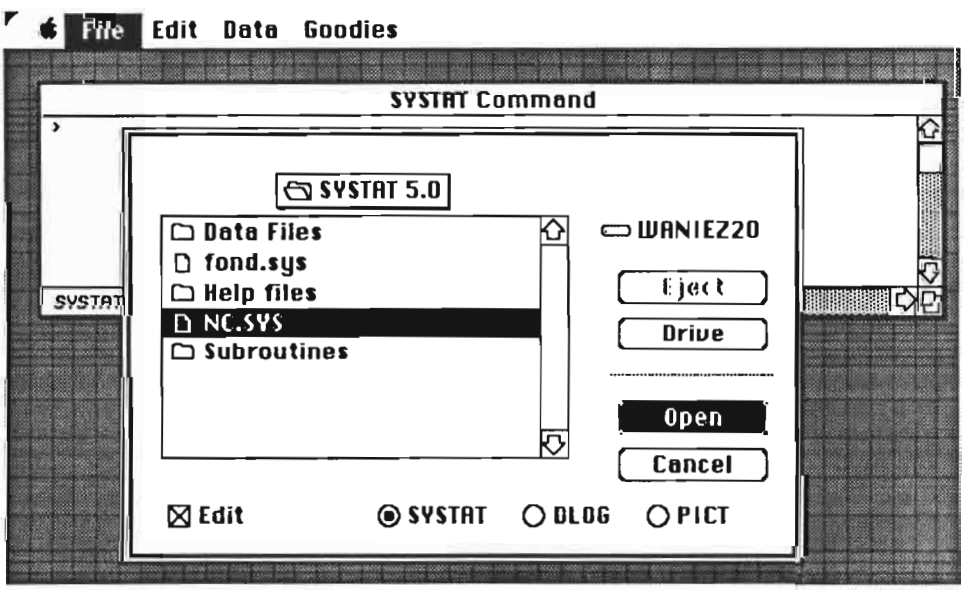


figure n° 2.2. SYSTAT: l'ouverture d'un tableau de données en vue de son édition (bouton EDIT coché).

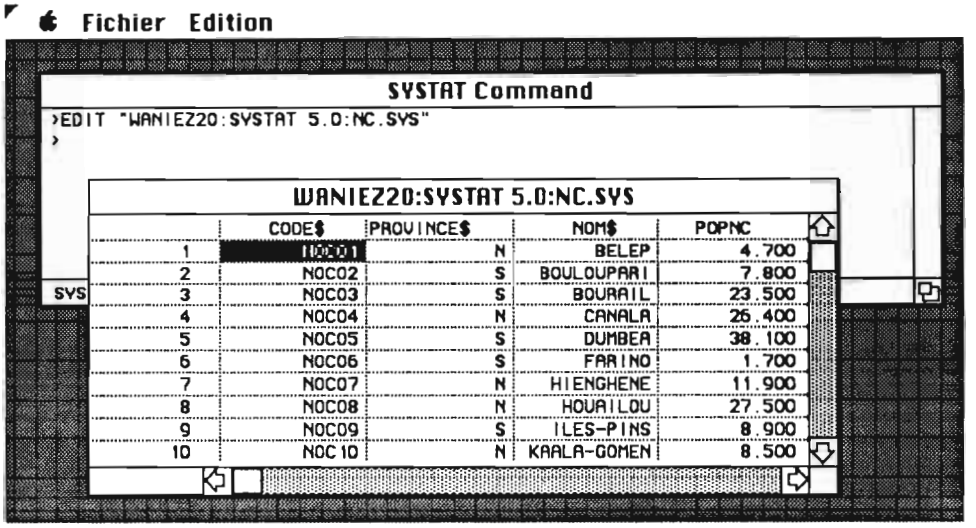


figure n° 2.3. SYSTAT: l'affichage du tableau dans l'éditeur.

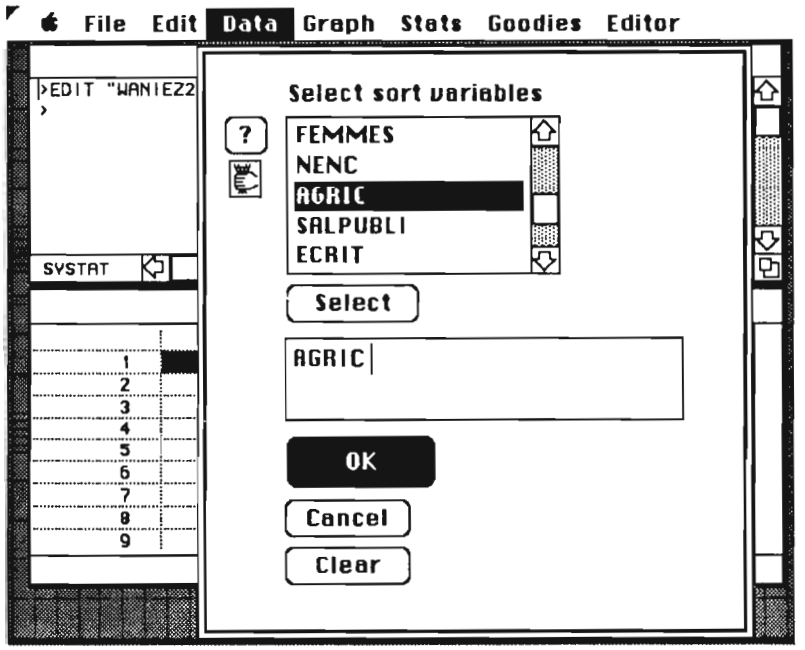


figure n° 2.4. SYSTAT: la sélection de la variable de tri du tableau.

Un clic sur le bouton OK provoque l'affichage d'un nouveau dialogue: afin de ne pas écraser le tableau initial par le tableau trié, le logiciel demande le nom du fichier contenant le résultat du tri. **SYSTAT** propose automatiquement le nom d'origine suivi du mot *sorted* (trié en anglais), mais rien n'interdit de choisir un nom différent (figure n° 2.5).

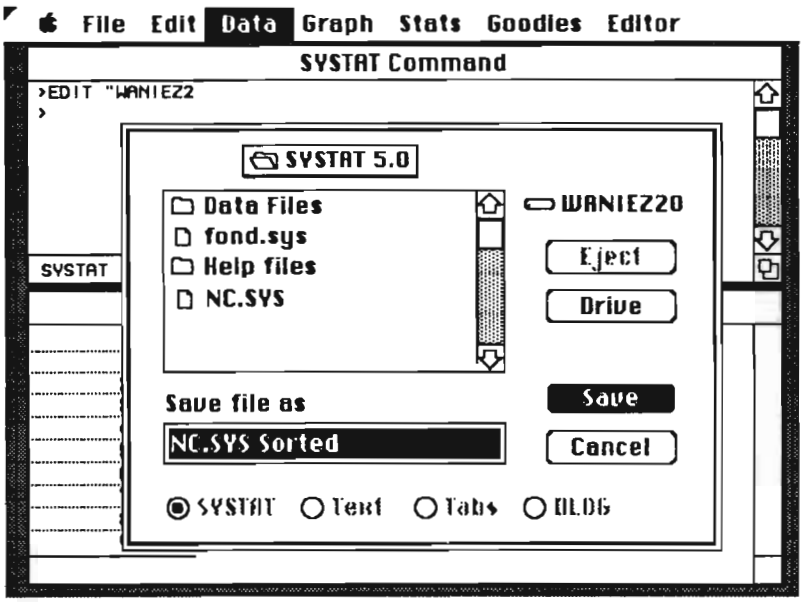
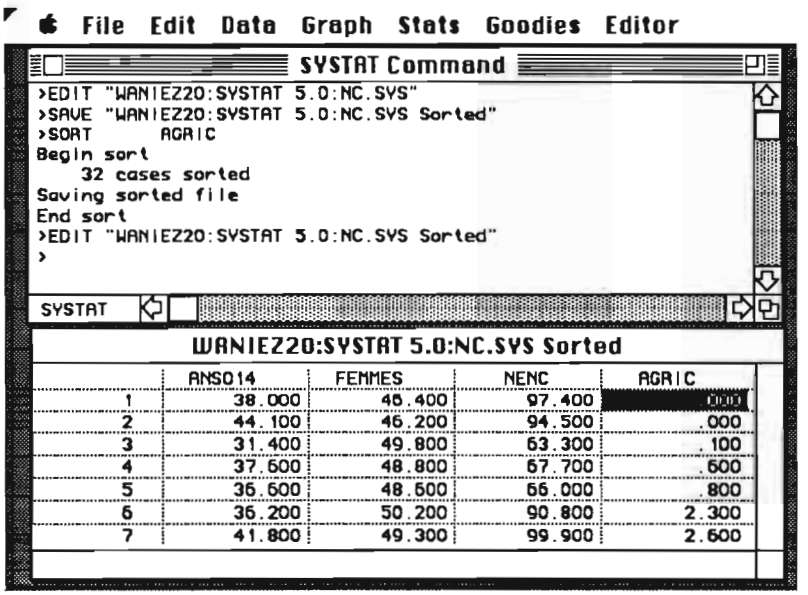


figure n° 2.5. SYSTAT: le choix du nom du fichier contenant le tableau trié.



	ANSO14	FEMMES	NENC	AGRIC
1	38.000	46.400	97.400	.000
2	44.100	46.200	94.500	.000
3	31.400	49.800	63.300	.100
4	37.600	48.800	67.700	.600
5	36.600	48.600	66.000	.800
6	36.200	50.200	90.800	2.300
7	41.800	49.300	99.900	2.600

figure n° 2.6. SYSTAT: l'affichage du tableau trié.

Un dernier clic sur le bouton SAVE déclenche l'exécution du tri. Un nouvel **OPEN** avec l'option EDIT permet de vérifier si tout s'est bien passé: les lignes du tableau apparaissent réarrangées selon les valeurs croissantes de la

variable AGRIC (figure n° 2.6). La documentation ne signale pas s'il existe une possibilité de tri à partir de la plus grande valeur jusqu'à la plus petite.

Au cours des choix successifs de l'utilisateur, **SYSTAT** traduit chaque requête par une commande dans son propre langage de commande. Par exemple, toutes les opérations décrites ci-dessus se résument aux 4 commandes suivantes:

```
>EDIT "WANIEZ20:SYSTAT 5.0:NC.SYS"
>SAVE "WANIEZ20:SYSTAT 5.0:NC.SYS Sorted"
>SORT      AGRIC
>EDIT "WANIEZ20:SYSTAT 5.0:NC.SYS Sorted"
```

En changeant le nom de la variable dans l'instruction SORT, on peut réaliser un nouveau tri, sans devoir parcourir à nouveau l'ensemble des menus: une possibilité qui peut faire gagner beaucoup de temps...

2.1.2. DataDesk

Avec **DataDesk**, l'ouverture du fichier contenant le tableau à trier se fait dès l'entrée dans le logiciel (figure n° 2.7). Diverses options sont proposées: l'ouverture d'un fichier de type **DataDesk**, déjà enregistré sur disque par ce logiciel, l'importation à partir d'un fichier texte, le collage de données en provenance du presse-papier, ou bien encore, la saisie directe à partir du clavier.

Un dialogue de sélection apparaît à l'écran, si le choix porte sur un fichier à ouvrir, grâce auquel on indique son nom.

Le bureau spécifique à **DataDesk** se remplit alors des icônes représentant les variables du fichier ouvert. Pour

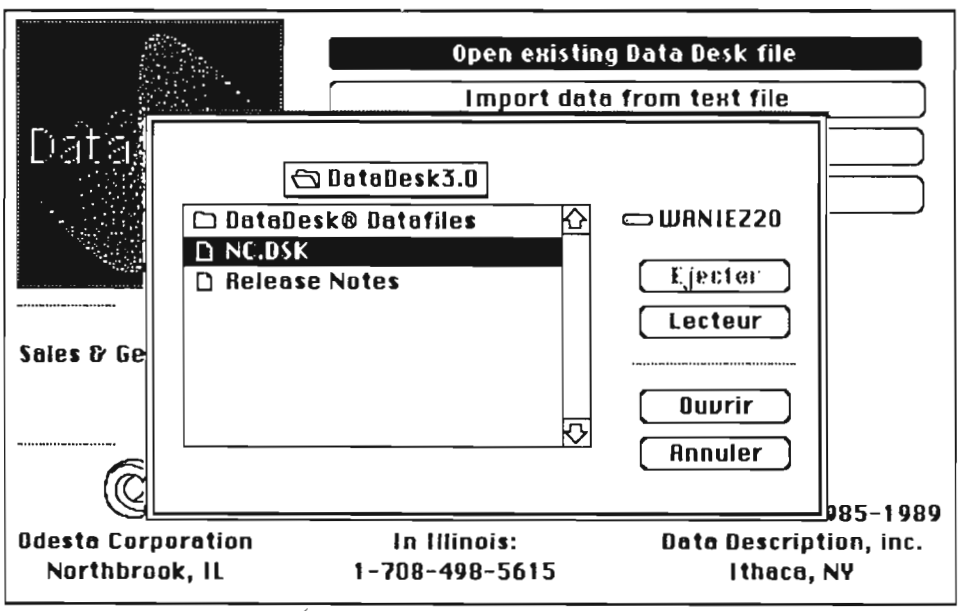


figure n° 2.7. DataDesk: la sélection d'un fichier.

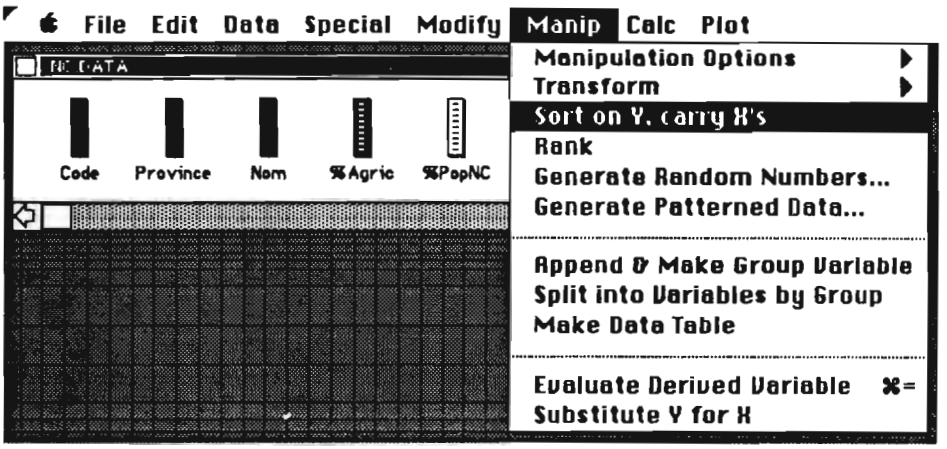


figure n° 2.8. DataDesk: la définition des conditions d'un tri (variables Y et X).

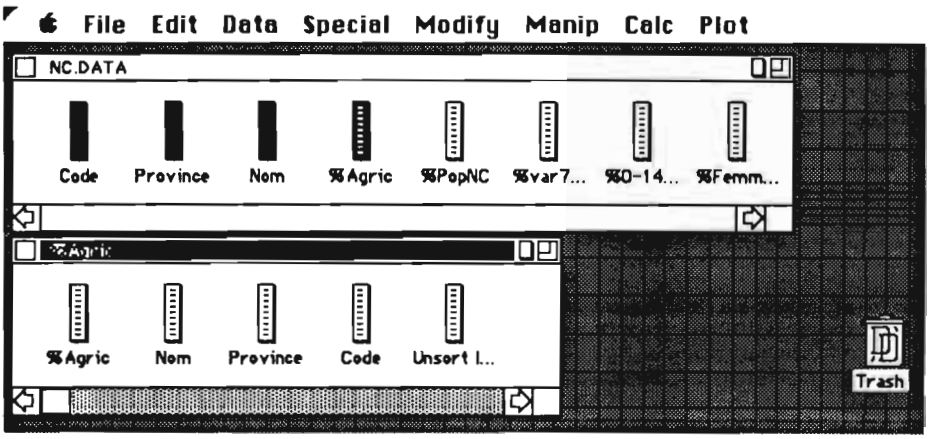


figure n° 2.9. DataDesk: le tableau d'origine et tableau trié.

procéder au tri, il faut procéder en trois temps (figure n° 2.8):

• choisir la variable de tri par un clic sur son icône, ici %Agric. On désigne symboliquement cette variable par la lettre Y.

• choisir les variables qui accompagneront Y dans le tableau trié par un clic sur leurs icônes, touche majuscule enfoncée, ici Code, Province et Nom). On désigne symboliquement l'ensemble de ces variables par la lettre X. Le tableau de données résultant du tri ne contiendra donc que la variable Y et les variables X, ces dernières étant en général un sous-ensemble des variables du tableau d'origine.

• sélectionner, dans le menu MANIP, l'article **SORT ON Y, CARRY X'S**.

Une nouvelle fenêtre s'affiche portant le nom de la variable de tri (%Agric). Outre cette variable, elle

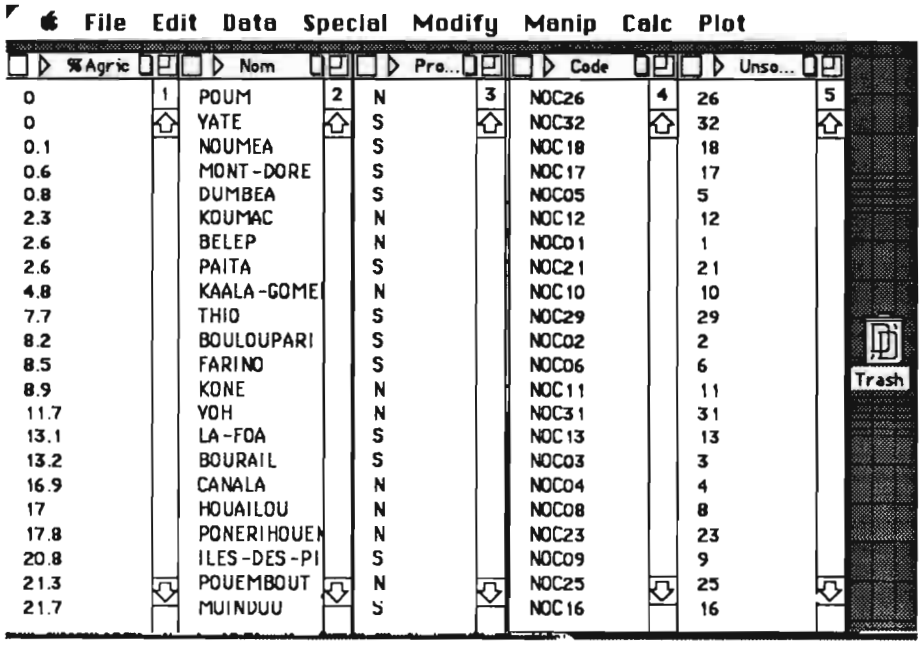


figure n° 2.10. DataDesk: l'ouverture des icônes des variables du tableau trié.

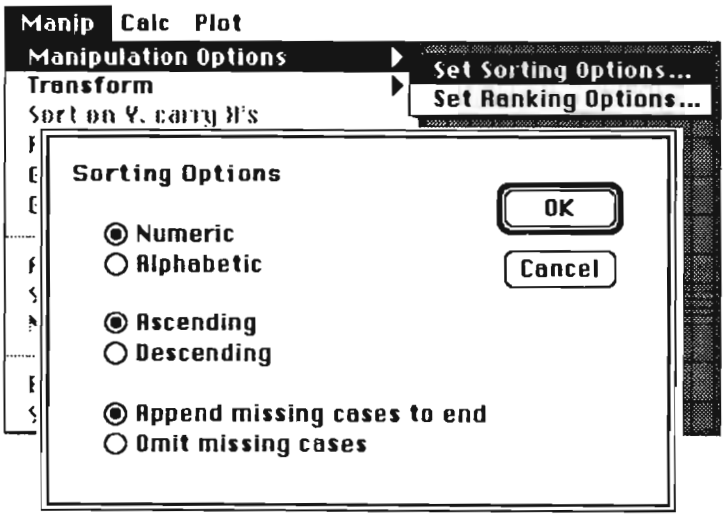


figure n° 2.11. DataDesk: les options de tri.

contient, comme prévu les icônes des variables X et une icône supplémentaire nommée Unsort Indice (figure n° 2.9).

Unsort Indice est une variable numérique contenant le rang des observations avant le tri. On vérifie que le tri s'est bien déroulé en ouvrant les icônes du tableau %Agric (figure n° 2.10).

DataDesk ne peut procéder directement à des tris de profondeur multiple, sur plusieurs variables simultanément, comme **SYSTAT** ou même JMP (voir ci-après). Mais, vis-à-vis de SYSTAT, il offre une bien plus grande richesse d'options de tri. On y accède en choisissant l'article **SET SORTING OPTIONS** du sous-menu **MANIPULATION OPTIONS** du menu **MANIP** (figure n° 2.11):

• tri numérique ou alphabétique: un tri alphabétique sur une variable numérique considère les valeurs comme une chaîne de caractères. L'ordre final est donc constitué de toutes la valeurs

commençant par 0, puis 1, 2, etc., même si ces valeurs varient entre 1 et 1 000 000.

• tri dans l'ordre croissant ou décroissant des valeurs.

• valeurs manquantes rangées à la fin du tri ou supprimées d'emblée de celui-ci.

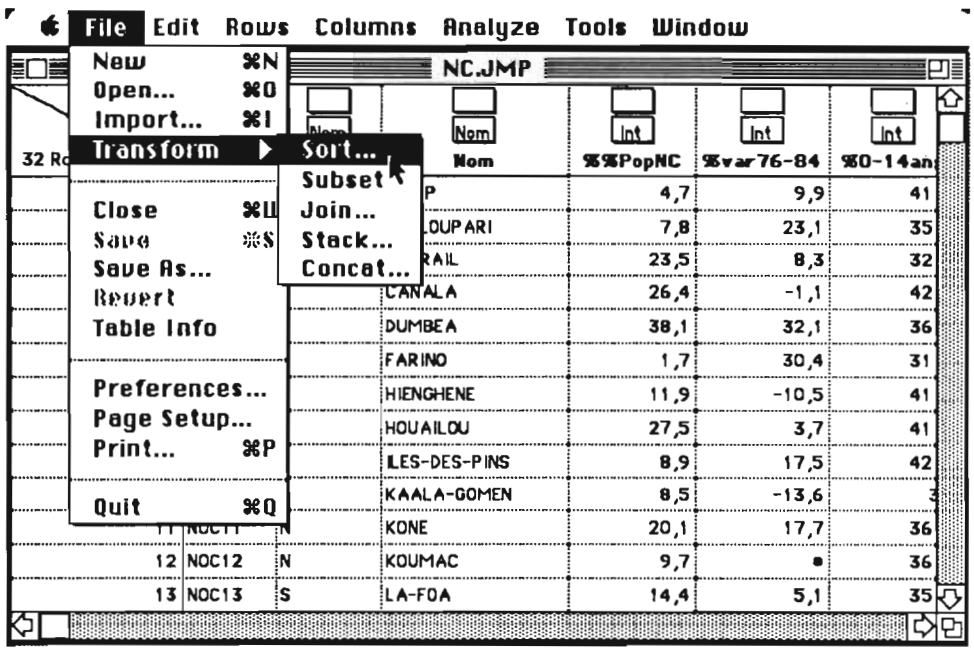


figure n°2.12. JMP: les menus pour la sélection du tri.

2.1.3. JMP

Pour trier un tableau de données avec **JMP**, il faut l'avoir préalablement ouvert à l'aide de l'article **Open** du menu **File**. Lorsque les données sont affichées à l'écran, on sélectionne l'article **Sort** du sous-menu de l'article **Transform** du même menu **File** (figure n° 2.12). Une boîte de dialogue s'affiche alors à l'écran (figure n° 2.13).

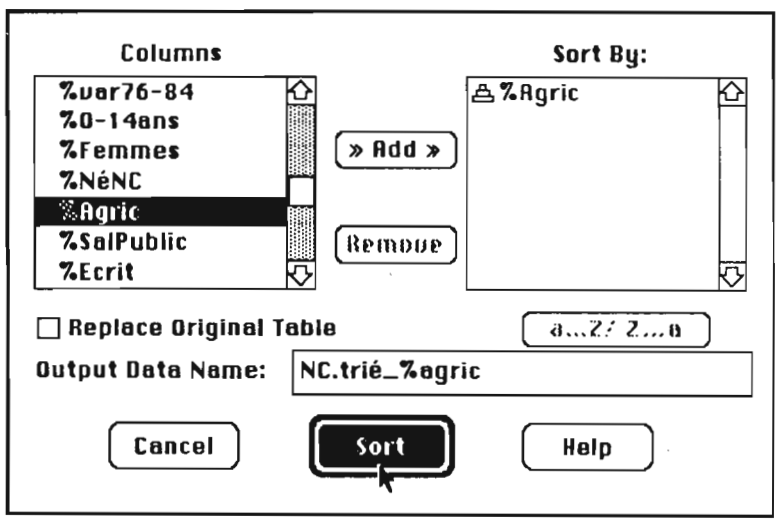


figure n°2.13. JMP: la boîte de dialogue pour le choix des options d'un tri.

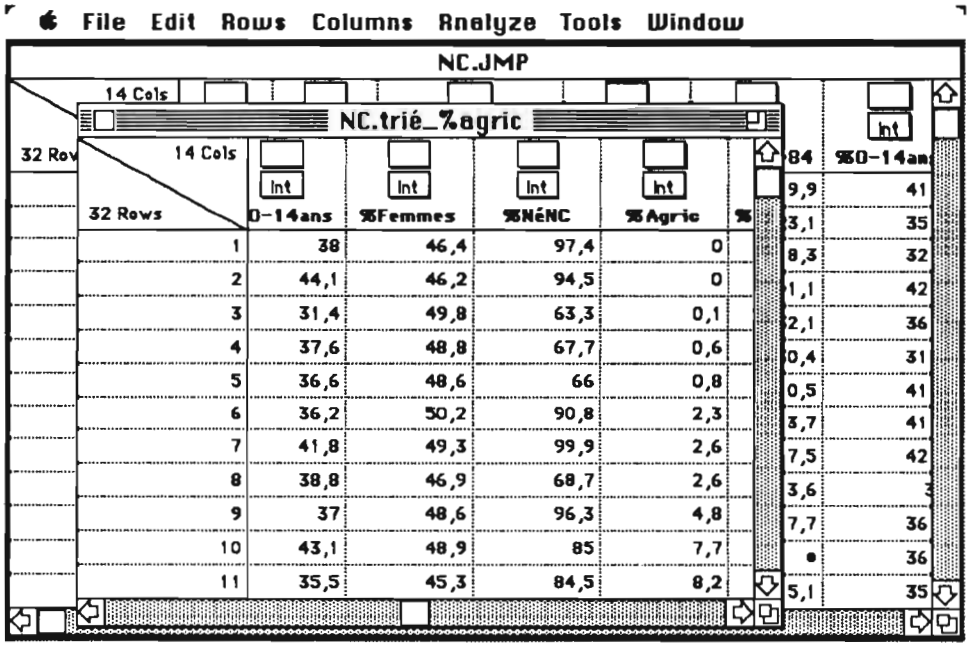


figure n° 2.14. JMP: l'affichage du tableau trié.

On peut ainsi sélectionner les variables de tri du tableau (ici la seule variable %AGRIC) et fixer l'ordre du classement; par défaut, cet ordre est ascendant comme l'indique l'icône figurant en face du nom de la variable retenue (sa base est plus large que sa partie supérieure).

Un simple clic sur le bouton a...Z/Z...a permet de renverser cet ordre (l'icône correspondante est une pyramide inversée), c'est-à-dire d'ordonner les individus de la plus grande à la plus petite valeur.

Enfin, cette boîte de dialogue permet de choisir entre le remplacement du tableau affiché par le tableau trié, ou bien l'enregistrement suivi de l'affichage d'un nouveau tableau: dans ce cas, le nom du nouveau tableau est saisi dans la zone d'édition prévue à cet effet. Le tableau trié s'affiche ensuite à l'écran (figure n° 2.14).

L'ordre des variables dans le tableau reste néanmoins inchangé, ce qui rend difficile l'identification des individus lorsque l'écran ne peut contenir toutes les variables, ce qui est ici le cas.

JMP offre donc la possibilité de rapprocher la variable ayant servi au tri du tableau de la ou des variables permettant d'identifier les individus. Après avoir sélectionné la variable à déplacer (%Agric), l'article **Move to First** du menu **Columns** (figure n° 2.15) la place à côté des variables Code, Province et Nom. On peut alors réellement associer un nom ou une appartenance provinciale à chaque valeur (figure n° 2.16).

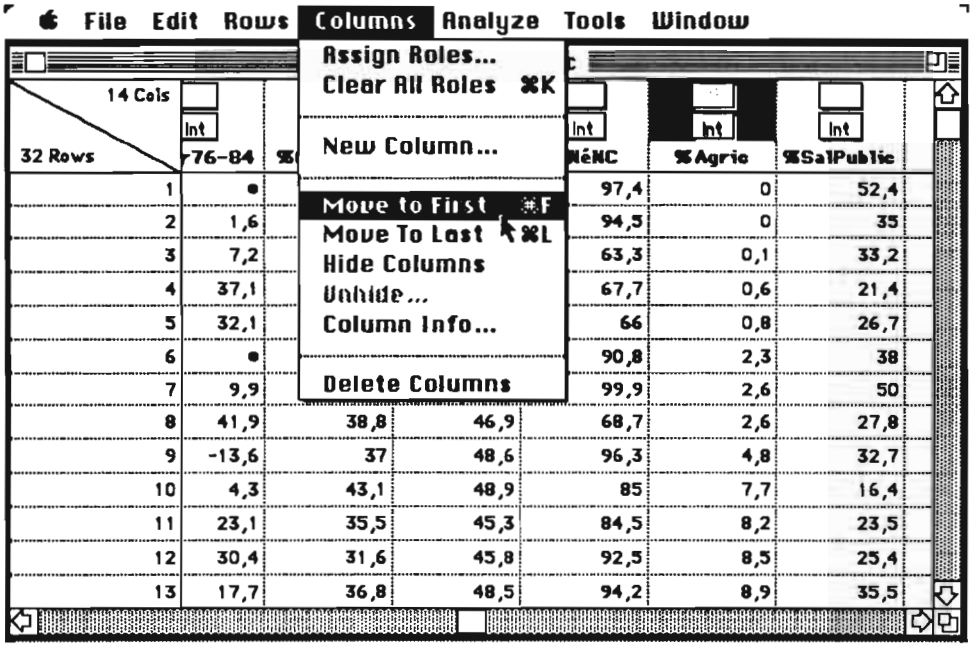
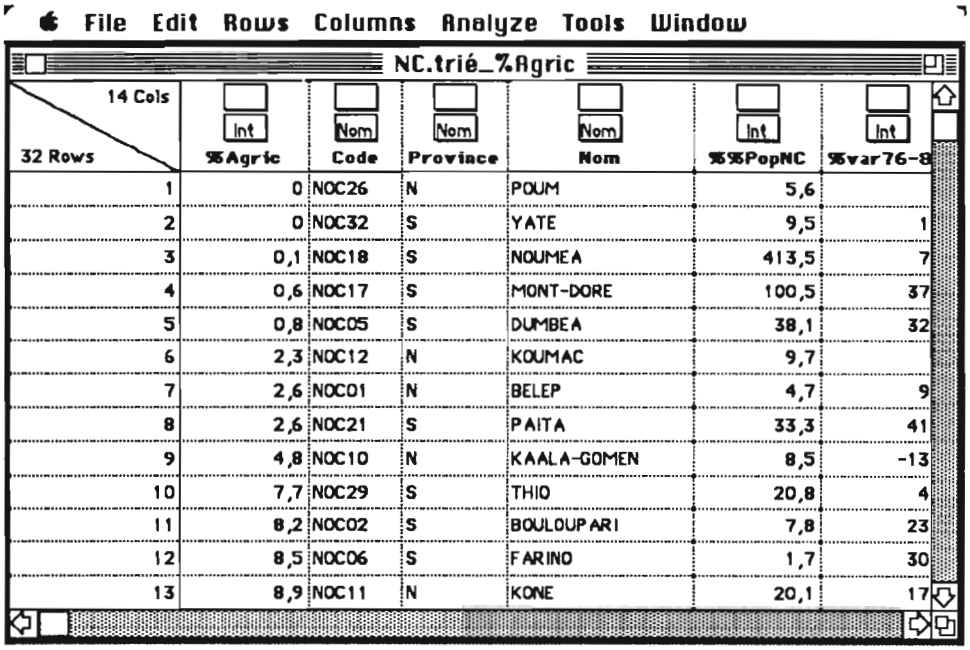


figure n° 2.15. JMP: Déplacement d'une colonne du tableau de données.



	%Agric	Code	Province	Nom	%PopNC	%var76-8
1	0	NOC26	N	POUM	5,6	
2	0	NOC32	S	YATE	9,5	1
3	0,1	NOC18	S	NOUMEA	413,5	7
4	0,6	NOC17	S	MONT-DORE	100,5	37
5	0,8	NOC05	S	DUMBEA	38,1	32
6	2,3	NOC12	N	KOUMAC	9,7	
7	2,6	NOC01	N	BELEP	4,7	9
8	2,6	NOC21	S	PAITA	33,3	41
9	4,8	NOC10	N	KAALA-GOMEN	8,5	-13
10	7,7	NOC29	S	THIO	20,8	4
11	8,2	NOC02	S	BOULOUPARI	7,8	23
12	8,5	NOC06	S	FARINO	1,7	30
13	8,9	NOC11	N	KONE	20,1	17

figure n° 2.16. JMP: la variable %Agric se retrouve à côté des variables d'identification des individus.

2.2. Le diagramme en tige et feuilles

Dans son ouvrage *Exploratory Data Analysis*, John W. Tukey propose une technique simple pour visualiser des données classées dans l'ordre ascendant ou descendant des valeurs. Le diagramme tige et feuilles (*stem an leaf plot*) permet d'apprécier la forme de leur distribution. Ce type de diagramme fait appel aux nombres entiers, ce qui simplifie l'approche initiale du tableau de données.

Considérons, par exemple, la part des agriculteurs exploitants dans la population active. La valeur maximum s'élève à 52.7%, 52% en arrondissant à l'entier inférieur. On construit la tige en traçant une colonne de 6 chiffres correspondant aux dizaines, soit les valeurs de 0 dizaine à 5 dizaines (0 à 50%). Les feuilles s'attachent à la tige d'après les valeurs des individus: pour 52%, le chiffre 2 apparaît derrière la valeur 5 de la tige; lorsque plusieurs individus entrent dans la même dizaine,

20, 21,21, 23, 25 25 et 29

on empile leurs valeurs de la manière suivante:

20 1 1 3 5 5 9

Cette technique permet d'aboutir facilement à la visualisation des formes de distributions (figure n° 2.17). Ainsi, il apparaît très clairement que:

- la distribution des taux de variation de la population présente une forme à peu près en cloche, très différente des deux autres variables.
- la proportion d'agriculteurs dans la population totale et la contribution des communes à l'ensemble des habitants du territoire présentent des formes semblables, avec une très forte représentation des faibles valeurs (sans qu'on sache pourtant si ce sont les même communes dans les deux cas).

Par rapport aux traditionnels histogrammes, le principal apport du diagramme en tige et feuilles réside dans sa capacité à conserver les valeurs. Mais, comme pour l'histogramme, la principale difficulté de construction du diagramme en tige et feuilles revient au choix de l'intervalle de classe choisi. Cet intervalle conditionne directement

VAR76-84		AGRIC		POPNC	
-3	2	0	0000022247888	0	12344578899
-2		1	133677	1	000133349
-1	5430	2	0113559	2	003567
-0	653210	3	038	3	138
+0	13457789	4	37	4	
+1	0477	5	2	5	5
+2	13			***	hors limites
+3	0257			10	0
+4	1			41	3

figure n°2.17. Diagrammes en tige et feuille des variables de la figure n°2.1.

l'allure générale de la distribution, et par conséquent, une part importante de ce qui sera dit sur chaque variable ainsi représentée. Disons tout de suite qu'il n'existe pas de méthode idéale pour choisir l'unité de construction de la tige. Le principe général stipule qu'il ne faut pas avoir une trop grande proportion de l'ensemble des observations se retrouvant dans la même classe.

Pour la contribution des communes à la population du Territoire, on aurait pu choisir un intervalle de 100 (soit une tige composée des valeurs 0, 1 2, 3 4. Dans ce cas, 30 communes sur 32 se seraient retrouvées face à la valeur 0, et rien n'aurait pu être remarqué mis à part la présence de deux communes exceptionnelles. Dans tous les cas, l'intervalle est constant, sauf pour les valeurs «hors limites».

L'attention du lecteur est donc attirée sur les risques encourus lors de la comparaison des formes de distributions visualisées à l'aide de diagrammes en tige et feuilles dont la tige n'aurait pas été construite de la même manière pour chaque graphique. De telles comparaisons peuvent aboutir à des conclusions absurdes: dans ce cas, il est préférable d'utiliser d'autres techniques de description, plus complexes que le graphique en tige et feuilles, mais donnant aussi des résultats dépendant moins directement de la technique employée.

2.2.1. SYSTAT

Seul logiciel, parmi les quatre retenus ici, à proposer le tracé de diagrammes en tige et feuilles, **SYSTAT** démontre son «encyclopédisme statistique».

Après avoir ouvert le fichier contenant les données à l'aide de l'article **OPEN** du menu **FILE**, l'article **STEM** du menu **GRAPH** permet de choisir le type de diagramme en tige et feuilles (figure n° 2.18). Une boîte de dialogue s'ouvre alors (figure n° 2.19): elle remplit deux fonctions principales. Dans la partie supérieure, l'utilisateur choisit la ou les variables à représenter;

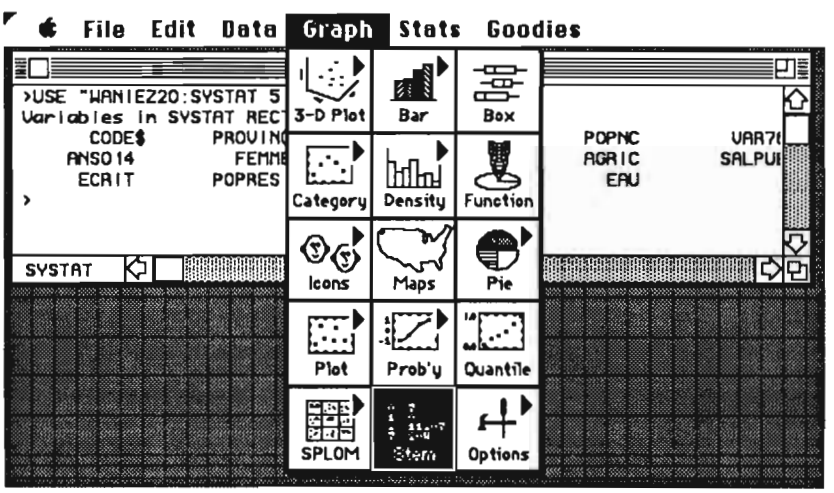


figure n° 2.18. SYSTAT: le choix du type de graphique STEM.

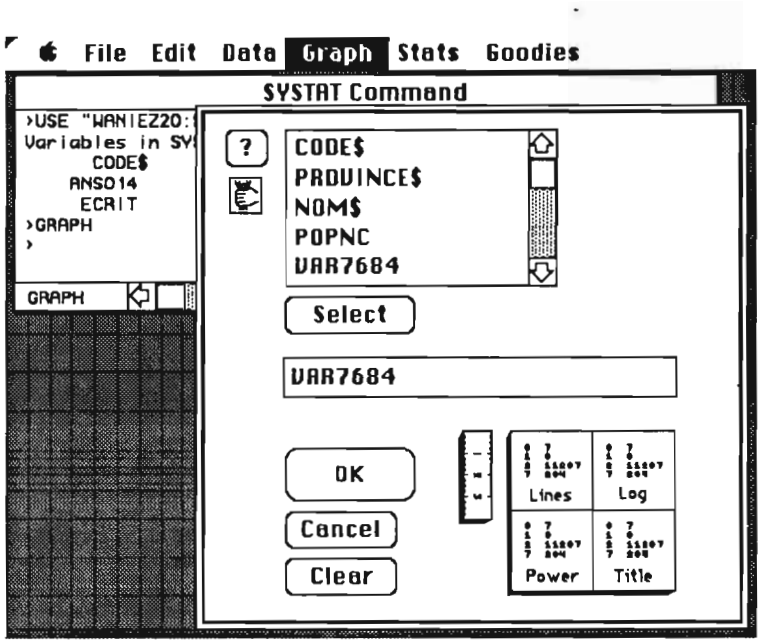


figure n° 2.19. SYSTAT: la fenêtre de sélection des variables à représenter et des options du graphique STEM.

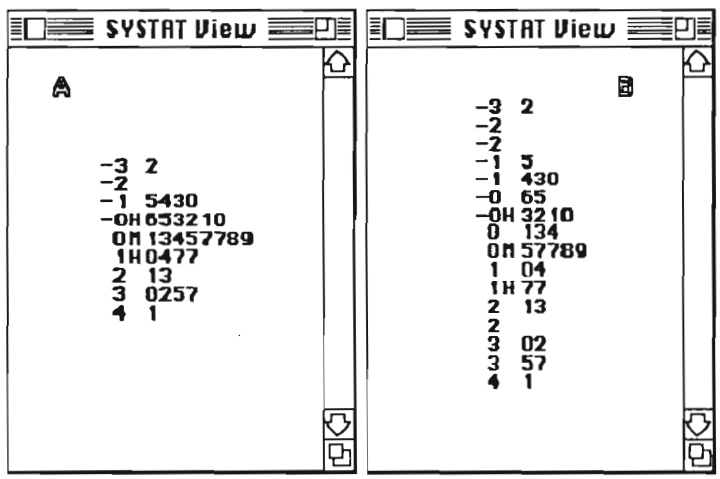


figure n° 2.20. SYSTAT: l'effet de l'option LINES. A: sans option, B: avec LINES=16.

leurs noms sont copiés dans la fenêtre d'édition prévue à cet effet, au-dessous du bouton SELECT.

La partie inférieure est dédiée à des options. La règle donne la possibilité de choisir les dimensions du graphique.

LOG et POWER transforment les variables sélectionnées en logarithmes ou en puissance. Ces options sont indispensables pour analyser des effectifs. De telles transformations agissent sur les données soit en «resserrant» l'étendue des valeurs (transformation LOG), soit, au contraire, en la «dilatant» (transformation POWER). Ceci a pour effet direct une modification de la forme du diagramme. Ces options sont donc d'un usage délicat et ne doivent être employées qu'après avoir vérifié que le graphique, sans transformation, n'est pas exploitable. Notons que LOG et POWER n'affectent pas les données du fichier, qui resteront donc inchangées, mais uniquement les valeurs de la variable au moment du tracé du diagramme.

TITLE donne accès à une zone d'édition où l'utilisateur saisit le titre devant figurer sur le graphique.

Enfin, LINES est sans doute l'option la plus indispensable à qui veut contrôler le tracé du diagramme. Grâce à elle, on peut en effet choisir le nombre de lignes du diagramme. Sur une même variable, le résultat obtenu apparaît bien différent selon qu'on laisse faire le

logiciel (figure n° 2.20.A) ou qu'on lui indique le nombre de lignes qu'il doit tracer (figure n° 2.20.B).

Comme pour les tris, tous les choix successifs de l'utilisateur dans les menus ou les boîtes de dialogue sont traduits par **SYSTAT** dans son langage de commande. Par exemple, les opérations décrites ci-dessus se résument aux 4 commandes suivantes:

```
>USE "WANIEZ20:SYSTAT 5.0:NC.SYS"
Variables in SYSTAT RECT file are:
      CODE$        PROVINCE$          NOM$          POPNC
    ANSO14          FEMMES          NENC          AGRIC
     ECRIT        POPRESID         DEPEN            EAU
>GRAPH
>STEM    VAR7684
>STEM    VAR7684/LINES=16
```

Il suffit donc de changer le nom de la variable dans l'instruction STEM ou d'ajouter des options derrière le nom de cette variable (ici /LINES=16) pour réaliser un nouveau diagramme, sans avoir à parcourir l'ensemble des menus.

Les possibilités offertes par **SYSTAT** pour le tracé de diagrammes en tige et feuilles sont donc assez variées et, comme il s'agit du seul logiciel à proposer ce genre de graphique, ces qualités doivent être soulignées. On peut néanmoins regretter l'absence de lien dynamique entre la fenêtre graphique et l'éditeur de données, qui aurait permis d'identifier les observations par un clic sur le diagramme.

2.3. Les résumés numériques résistants

Les utilisateurs des méthodes statistiques classiques connaissent bien la sensibilité de certains paramètres aux valeurs exceptionnelles. Or, c'est sur ces paramètres que se fonde l'essentiel de leurs analyses. Comment ne pas s'interroger sur le bien-fondé de l'information véhiculée par la moyenne arithmétique et l'écart-type, deux paramètres parmi les plus fréquemment utilisés en statistique. En moyenne, la population de chaque commune de Nouvelle Calédonie représente 72.3‰ de la population du totale du Territoire. Sans Nouméa et sa «banlieue», Mont Dore, ce chiffre tombe à 16.2‰ seulement (les différences sont encore plus importante pour l'écart-type: 72.3 contre 12.2). Deux observations exceptionnelles peuvent donc modifier profondément la localisation et l'étendue d'une distribution.

Pour palier ces inconvénients, on fait appel à des résumés numériques résistants, ainsi nommés car ils ne sont pas trop sensibles aux valeurs exceptionnelles. Parmi l'ensemble de ces techniques, celles dites «ordinales», parce que basées sur les rangs des individus, sont sans doute les plus aisées à mettre en application. Ainsi, à la moyenne arithmétique comme indicateur de la localisation d'une variable, on préfère la médiane: lorsque les individus sont classés dans

N		32
maximum	100.0%	413,50
	99.5%	413,50
	97.5%	413,50
	90.0%	50,56
quartile	75.0%	26,07
mediane	50.0%	13,20
quartile	25.0%	8,60
	10.0%	3,72
	2.5%	1,70
	0.5%	1,70
minimum	0.0%	1,70

avec Nouméa et Mont Dore

N		30
maximum	100.0%	55,900
	99.5%	55,900
	97.5%	55,900
	90.0%	33,140
quartile	75.0%	23,900
mediane	50.0%	12,500
quartile	25.0%	8,325
	10.0%	3,440
	2.5%	1,700
	0.5%	1,700
minimum	0.0%	1,700

sans Nouméa et Mont Dore

figure n° 2.21. Les résumés numériques basés sur la médiane et les quartiles, avec Nouméa et Mont Dore et sans ces communes.

(*hinge*). Le pivot inférieur (premier quartile) correspond à la valeur en dessous de laquelle on trouve un quart de l'effectif total; de même, le pivot supérieur correspond à la valeur au-dessus de laquelle on trouve un quart de l'effectif total. L'intervalle compris entre le pivot inférieur et le pivot supérieur (soit la moitié de l'effectif total) porte le nom d'intervalle interquartile; les spécialistes de l'analyse exploratoire préfèrent le terme d'étendue du centre (*midspread*) ce qui semble bien plus imagé.

l'ordre croissant ou décroissant, la valeur médiane partage les individus en deux ensembles d'effectifs égaux.

Pour apprécier l'étendue, on fait appel aux premiers et troisièmes quartiles qui, dans le jargon de l'analyse exploratoire, prennent le nom de pivots

Le caractère résistant de la médiane et de l'étendue du centre apparaît très nettement: avec Nouméa et Mont Dore, la médiane de la part de chaque commune dans la population totale de Nouvelle-Calédonie atteint 13.2‰ alors que sans ces deux communes, sa valeur est de 12.5‰ soit seulement 0.7‰ de

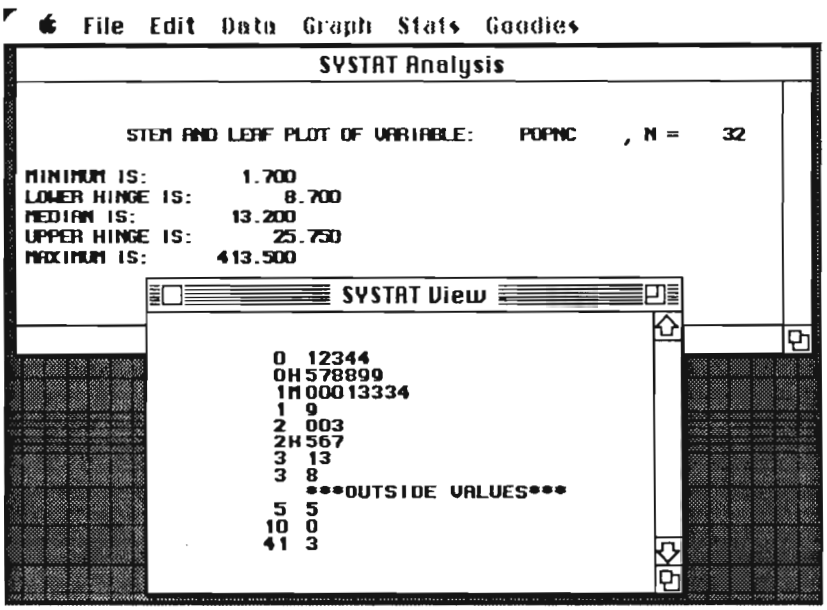


figure n° 2.22. SYSTAT: le résumé numérique résistant.

différence. Cette résistance s'affirme aussi sur les quartiles: 8.6 contre 8.3 pour le premier, 26.1 contre 23.9 pour le troisième.

2.3.1. SYSTAT

Avec **SYSTAT**, les résumés résistants accompagnent les diagrammes en tige et feuilles. Il n'y a donc rien à faire de plus. Le tableau numérique s'affiche dans une fenêtre standard nommée SYSTAT ANALYSIS (figure n° 2.22). On y trouve le nombre d'observations entrant réellement dans le calcul (à l'exclusion des valeurs manquantes), le minimum et le maximum, la médiane et les quartiles n° 1 et n° 3 (*lower hinge* et *upper hinge*).

2.3.2. DataDesk

Le contenu des résumés numériques résistants proposés par **DataDesk** dépend de ce que souhaite obtenir l'utilisateur. En effet, l'article **SELECT SUMMARY STATISTICS...** du sous menu **CALCULATION OPTIONS** du menu **CALC** (figure n° 2.23) donne accès à une boîte de dialogue grâce à laquelle on peut sélectionner les indicateurs que l'on souhaite calculer (figure n° 2.24).

Une partie seulement des indicateurs fournis par **DataDesk** répondent au critère de résistance défini plus haut. Il suffit d'aller cocher les cases nécessaires. Outre les classiques médiane, minimum, maximum, quartiles, on trouve dans ce tableau l'étendue interquartile (Interquartile Range), la valeur centrale entre deux pourcentiles (*Mid k %*), et la différence entre deux pourcentiles (*k-th %ile Diff*), forme généralisée à tout pourcentile de l'intervalle interquartile. Pour toutes les valeurs relatives aux pourcentiles, le logiciel propose une zone d'édition: ici, 25 fait référence aux quartiles, mais 10 ferait de la même manière, référence aux déciles et 100 aux centiles.

Lorsque tous les paramètres à calculer ont été choisis, il suffit de sélectionner l'article **SUMMARY REPORTS** du menu **CALC** pour effectuer le traitement proprement dit. Il dure de quelques secondes à quelques minutes, en fonction de la quantité d'information à intégrer au calcul. Puis, une fenêtre contenant tous les résultats s'ouvre (figure n° 2.25).

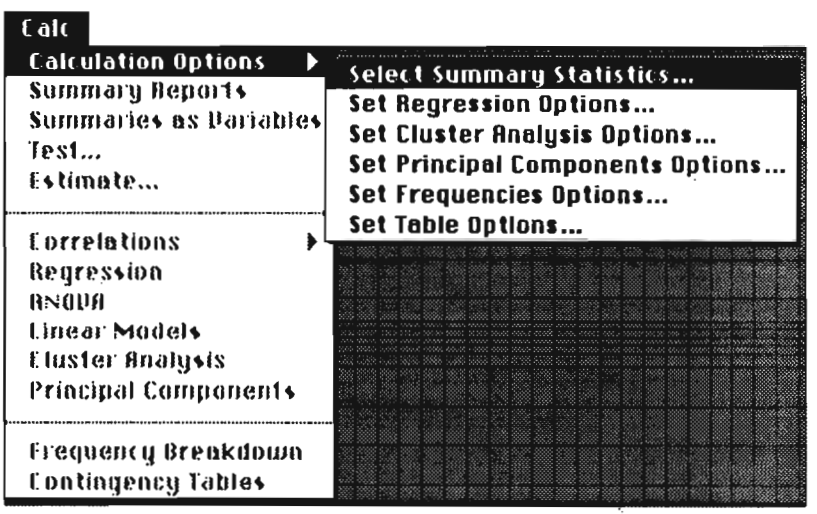


figure n° 2.23. DataDesk: Pour définir le contenu du résumé statistique...

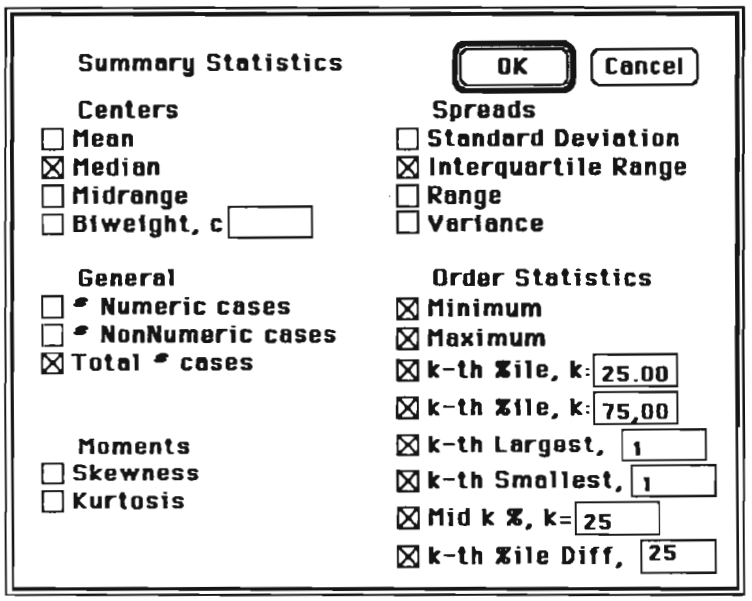


figure n° 2.24. ...il suffit de sélectionner les indicateurs proposés par DataDesk.

Si le calcul des résumés statistiques résistant à l'aide de **DataDesk** semble extrêmement simple, il ne faut pas en conclure que ce logiciel s'en tienne à cette tâche, somme toute assez banale. En effet la fenêtre des résultats possède des vertus (un peu) cachées. Tout d'abord, elle possède un lien dynamique avec la fenêtre des données qui permet de procéder à un nouveau calcul, simplement en déplaçant l'icône d'une autre variable (figure n° 2.26). Après quelques instants, les nouveaux paramètres s'affichent dans la fenêtre des résultats (figure n° 2.27).

Loin d'être un gadget, cette possibilité correspond bien aux principes de l'analyse exploratoire. En effet, rien n'impose la réalisation de calculs massifs dès le premier contact avec les données. En procédant pas à pas, avec intuition et méthode, sans se laisser dépasser par un amas de chiffres insurmontable, le chercheur se met en position de trouver plus facilement le résultat qu'il pressent.

Mais la puissance de **DataDesk** ne

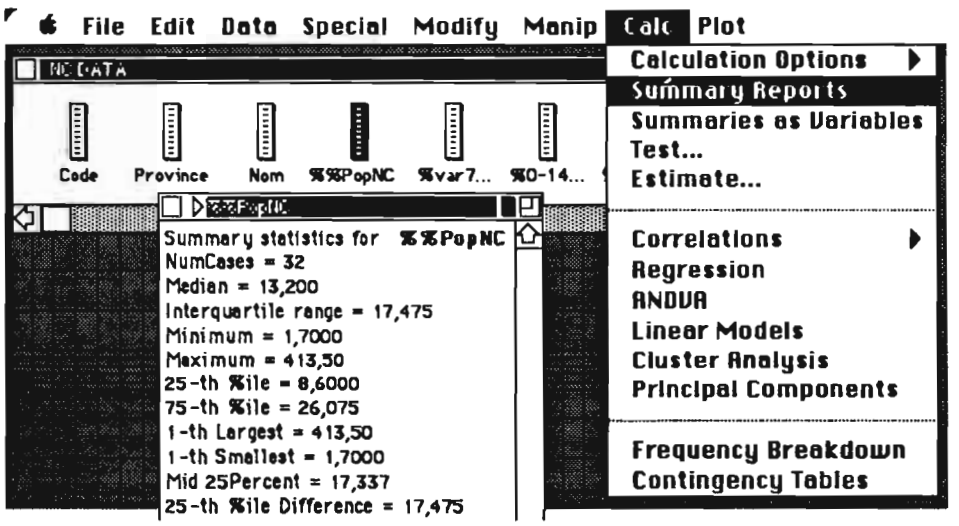


figure n° 2.25. DataDesk: un résumé statistique résistant de la variable %%POPNC.

s'arrête pas là. Un clic sur le nom d'une des variables donne accès à un nouveau menu dans lequel on peut choisir un traitement complémentaire (ce type de menu porte le nom anglais de *hyperview menu*). Par exemple, un histogramme peut compléter utilement la série des paramètres numériques (figure n° 2.28 et 2.29).

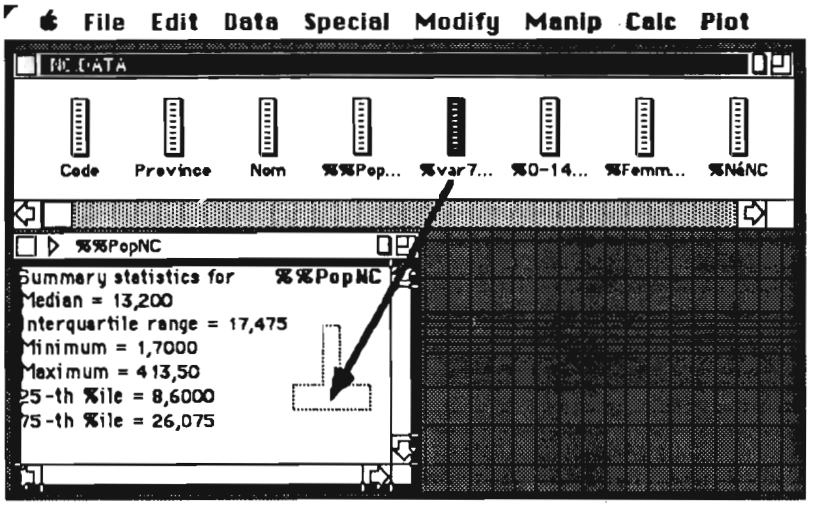


figure n° 2.26. DataDesk: le déplacement de l'icône d'une nouvelle variable...

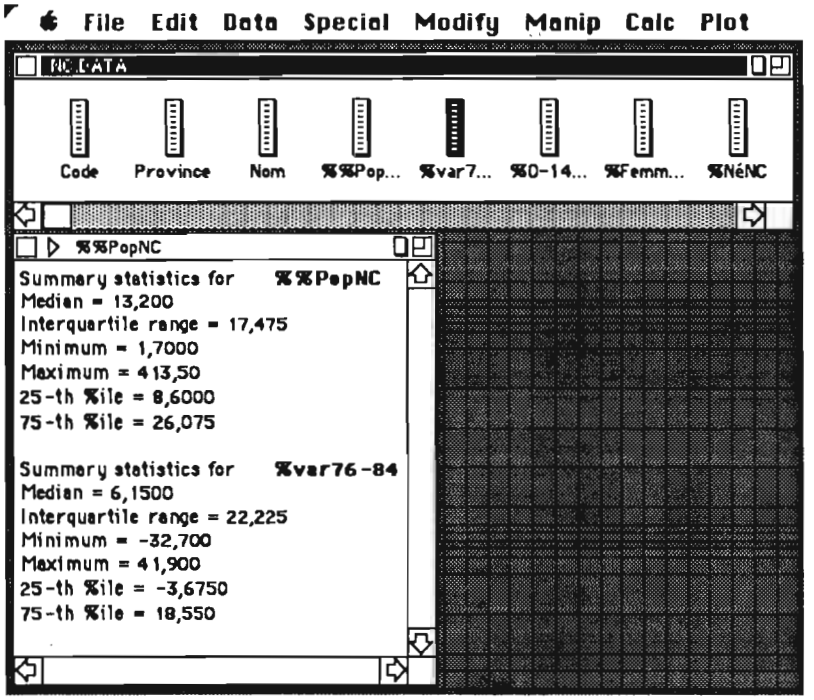


figure n° 2.27. ... et la fenêtre des résultats est immédiatement mise à jour.

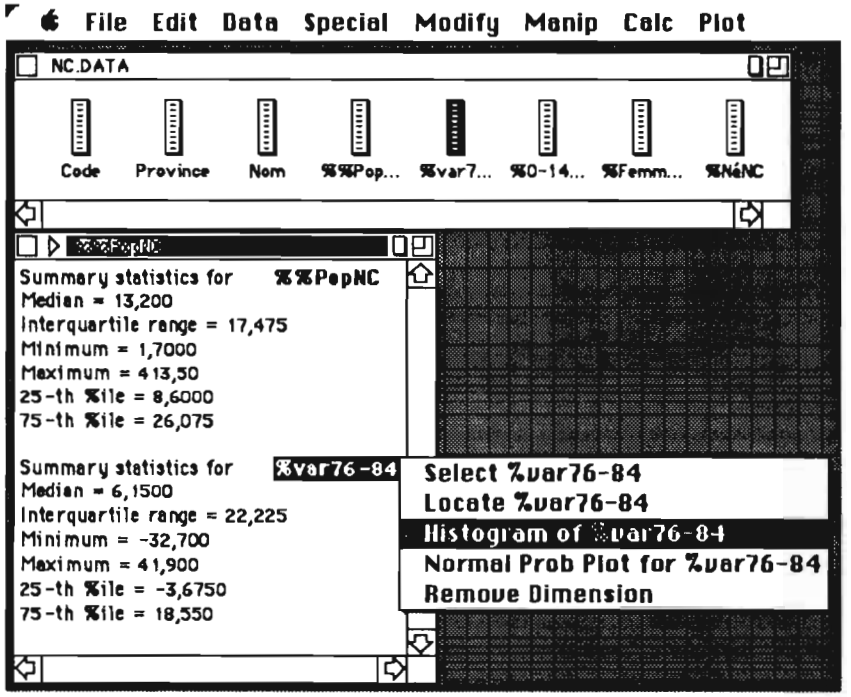


figure n° 2.28. DataDesk: le choix du tracé d'un histogramme dans le menu hyperview.

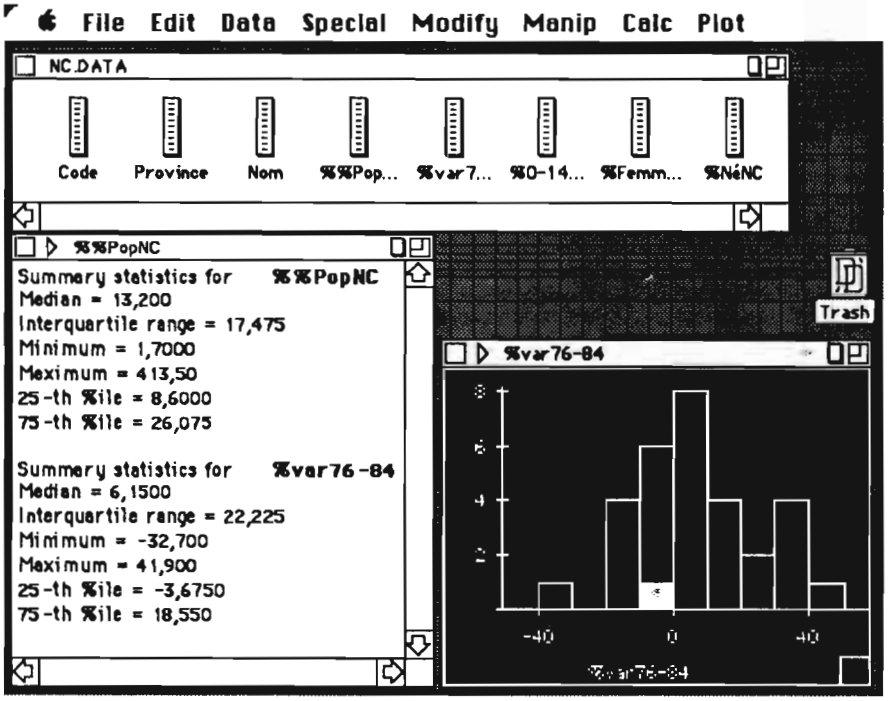


figure n° 2.29.DataDesk: l'affichage de l'histogramme de la variable %var76-84.

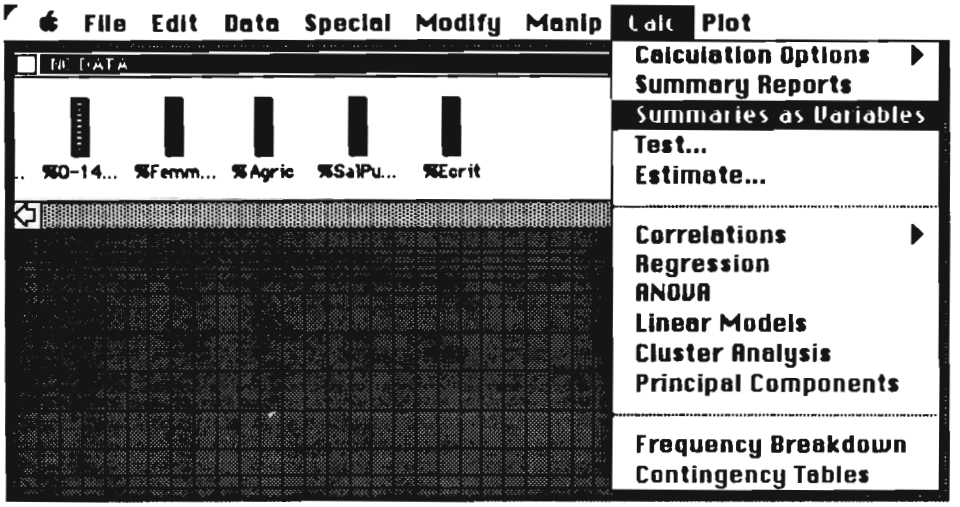


figure n° 2.30. DataDesk: l'enregistrement des paramètres statistiques.

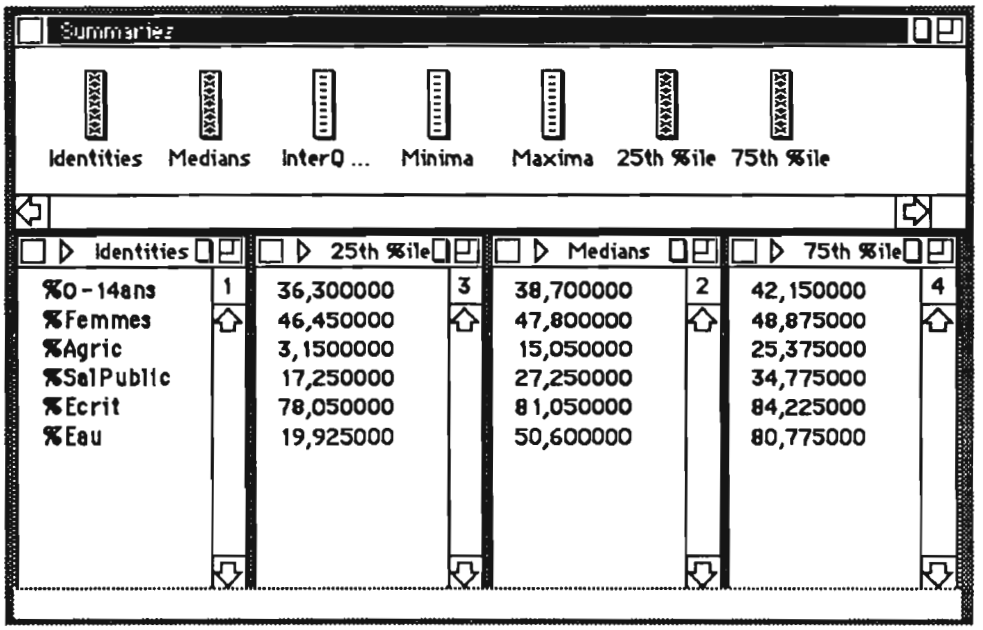


figure n° 2.31. DataDesk: la fenêtre SUMMARIES et les icônes des variables contenant les différents paramètres calculés.

Enfin, **DataDesk** offre une intéressante possibilité de stockage des paramètres statistiques calculés dans un nouveau tableau de données. Cela peut être très pratique si, par exemple, on souhaite comparer les médianes des différents pourcentages qui composent les variables figurant dans le tableau de données. On y accède par l'article **SUMMARIES AS VARIABLES**, littéralement résumés comme variables, du menu **CALC**, après avoir sélectionné la

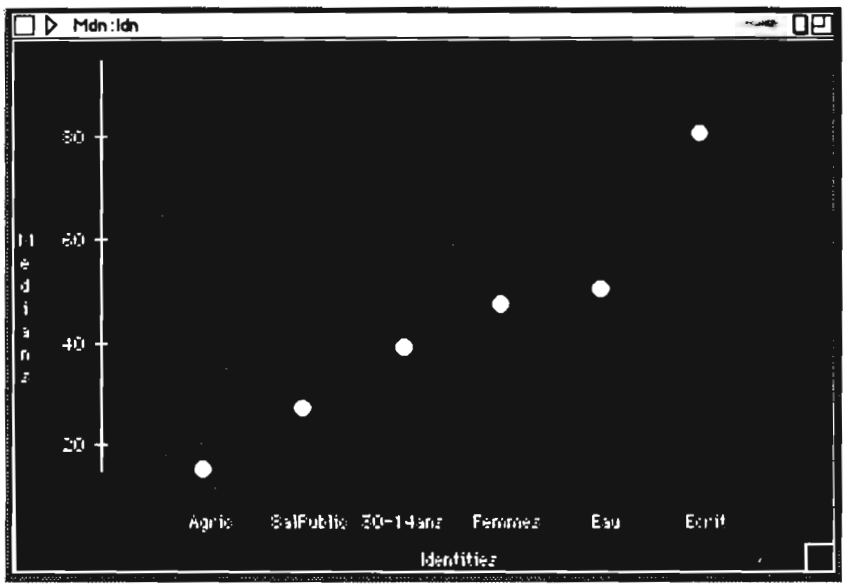


figure n° 2.32. DataDesk: un graphique de comparaison des médianes.

ou les variables devant entrer dans le calcul. (figure n° 2.30).

DataDesk ouvre une nouvelle fenêtre intitulée SUMMARIES (figure n° 2.31) contenant autant d'icônes de variables que de paramètres calculés (le choix des paramètres s'effectuant comme auparavant avec l'article **SELECT SUMMARY STATISTICS...** du sous menu **CALCULATION OPTIONS** du menu **CALC**.

La variable IDENTITIES renferme les noms des variables sur lesquelles le calcul a été effectué. En ouvrant les fenêtres d'édition de quelques-unes de ces variables, par exemple celles des noms, du premier quartile, de la médiane et du dernier quartile, on s'aperçoit qu'il s'agit d'un véritable nouveau tableau de données. Comme tel, il peut à son tour faire l'objet de calculs ou de représentations graphiques. La comparaison des valeurs des différentes médianes ne pose alors plus aucun problème (figure n° 2.32).

DataDesk présente, pour le calcul de résumés numériques résistants une large panoplie d'options qui en font un logiciel très bien dessiné pour un usage exploratoire. On ne peut manquer d'être étonné par les possibilités d'enchaînement des traitements qu'il permet.

2.3.3. JMP

Avec **JMP**, les résumés numériques résistants constituent une partie des sorties proposées en standard par la plate-forme **DISTRIBUTION OF Y'S** du menu **ANALYZE**. Il faut donc, préalablement à cette analyse, fixer aux variables pour lesquelles on souhaite obtenir un résumé le rôle Y. Cela se fait soit à partir des menus déroulants situés dans la partie supérieure de chaque colonne du tableau de données, soit à l'aide de l'article **ASSIGNING ROLES** du menu **COLUMNS** (figure n° 2.33).

JMP ouvre alors une fenêtre nommée DISTRIBUTION qui contient une description complète de chacune des variables retenues: histogramme et diagramme en boîte et moustaches (voir § 2.4, ci-après), quantiles, moments

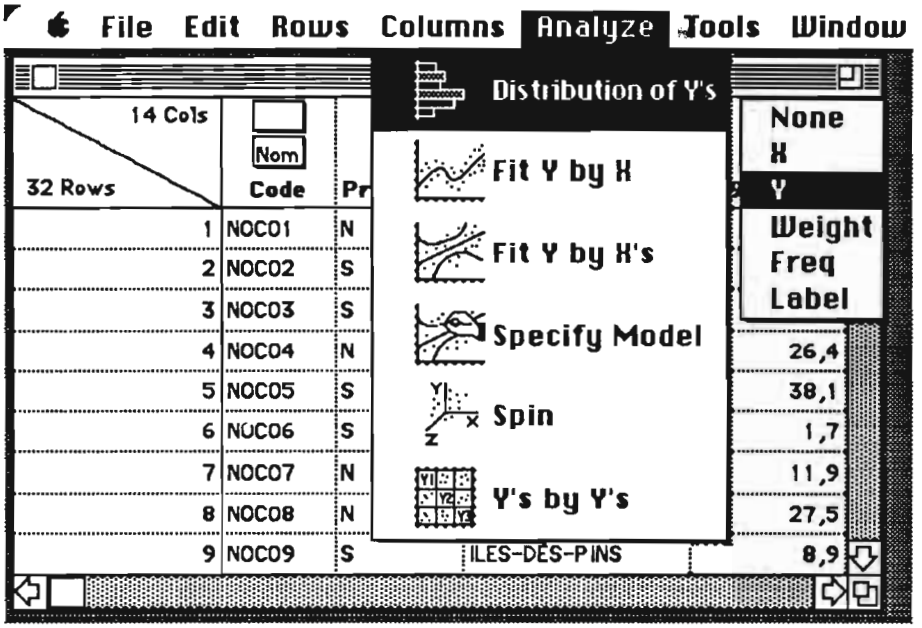


figure n° 2.33. JMP: Fixer le rôle Y à la variable... et sélectionner l'article DISTRIBUTION OF Y'S.

(moyenne, écart-type, etc.) ainsi que le t-test de Student. Dans le paragraphe consacré aux quantiles (figure n° 2.34), on trouve bien entendu le minimum, le maximum, les quartiles Q1 et Q3 et la médiane; cette liste est complétée par les déciles 10% et 90% et les pourcentiles 0.5%, 2.5%, 97.5% et 99.5%.: difficile d'en demander plus!

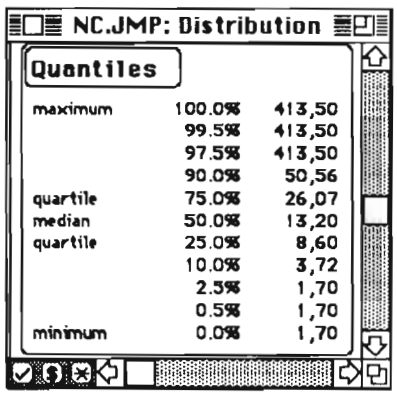


figure n° 2.34. JMP: un résumé statistique résistant.

Dans le coin inférieur gauche de la fenêtre des résultats figurent trois boutons utiles pour une exploration complémentaire des données:

- options d'affichage complémentaires et de sélection des tableaux en sortie.
- options d'enregistrement des résultats.
- aide en ligne et affichage de toutes les sorties possibles.

Par exemple, il peut s'avérer intéressant de remplacer les valeurs d'origine d'une variable par les rangs des observations sur cette même variable. De telles transformations conduisent à l'analyse statistique non-paramétrique, plus résistant au sens donné ici à ce mot, que les techniques faisant appel à la variance. Pour récupérer ces rangs dans le tableau de données, il suffit de

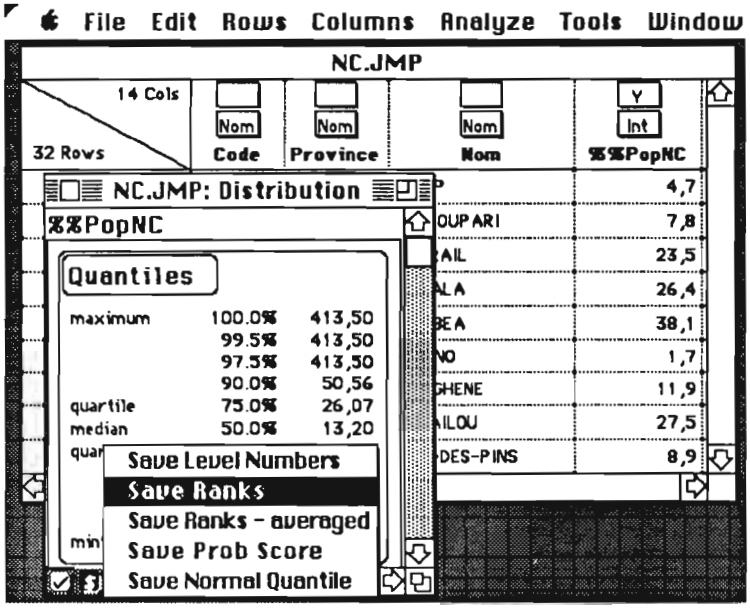


figure n° 2.35. JMP: l'enregistrement des rangs sur la variable %%POPNC.

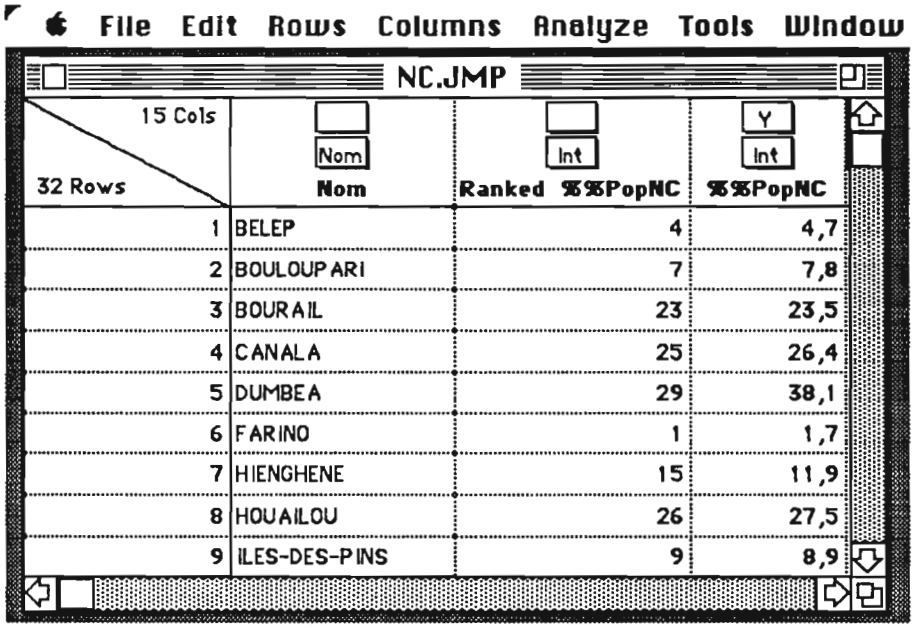


figure n° 2.36. JMP: les rangs sur la variable %%POPNC enregistrés dans la nouvelle variable Ranked %%POPNC.

choisir l'article SAVE RANKS du menu $ auquel on accède par un clic sur le bouton $ (figure n° 2.35). JMP crée alors une nouvelle variable nommée *Ranked* suivi du nom de la variable d'origine (figure n° 2.36).

2.4. Le diagramme en boîte et moustaches

Si les caractéristiques des distributions décrites précédemment permettent de mémoriser les ordres de grandeur, elles doivent être complétées par un graphique pour rendre compte de la forme de ces distributions. Le diagramme en boîte et moustaches, lui aussi développé par J. Tukey, représente non seulement les pivots, mais visualise d'autres propriétés bien utiles, comme par exemple les queues des distributions.

Dessiner un diagramme en boîte et moustaches comprend les étapes suivantes (figure n° 2.37):

- A - on trace une échelle de longueur égale ou supérieure à l'étendue des valeurs et comprenant la valeur minimale et maximale. L'unité de graduation de l'échelle est celle de l'unité de mesure dans laquelle les valeurs sont exprimées.
- B - la boîte est d'abord tracée; sa largeur est sans importance, mais sa longueur représente la distance entre les deux pivots; on appelle P_1 le premier pivot et P_3 le troisième.
- C - dans la boîte, la médiane, M est représentée par un trait dans la largeur.
- D - on repère l'individu présentant une valeur égale ou supérieure à P_1 - (1.5 x (P_3-P_1)); on nomme I_1 cet individu et V_1 sa valeur. De manière symétrique, on repère l'individu présentant une valeur égale ou inférieure à P_3 + (1.5 x (P_3-P_1)); on nomme I_3 cet individu et V_3 sa valeur. Les deux moustaches sont alors tracées en joignant par une ligne I_1 et P_1 et I_3 et P_3.

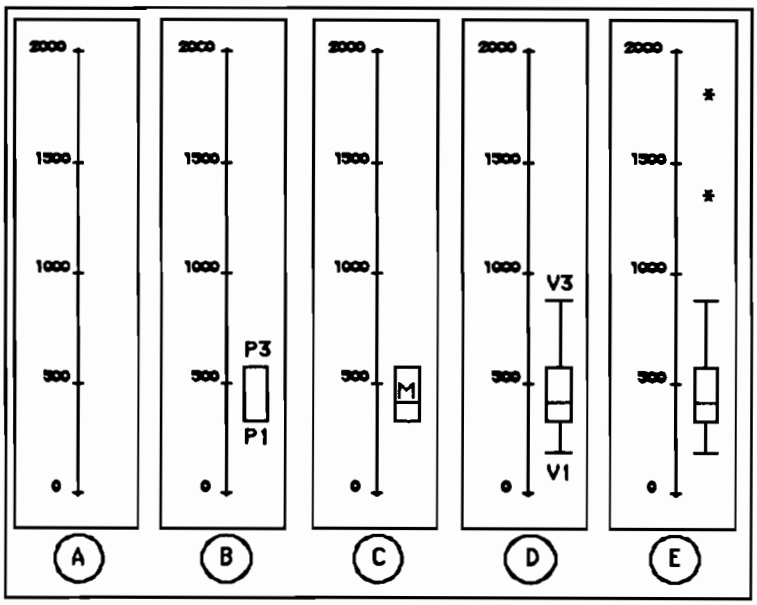


figure n° 2.37. Les étapes de la construction d'un diagramme en boîte et moustaches. La variables représentée ici est le taux de dépendance (‰).

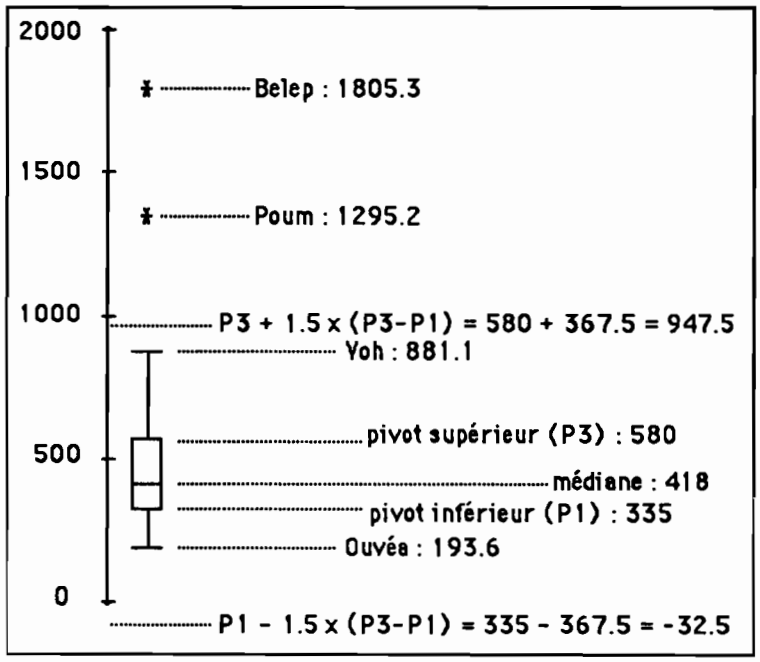


figure n° 2.38. Le diagramme en boîte et moustaches du taux de dépendance (‰).

• E - les individus situés à l'extérieur de la boîte ou des moustaches, ceux dont la valeur est supérieure à V_3 ou inférieure à V_1, sont marqués par un symbole (un point ou une astérisque).

Cette technique de construction du graphique en boîte et moustaches appelle quelques explications sur le tracé des moustaches. Les limites supérieures et inférieures de celles-ci sont définies de manière à repérer facilement les observations exceptionnelles. Il n'y a pas à proprement parler de règle précise pour fixer les bornes inférieure et supérieure de chaque moustache. Tukey conseille d'utiliser un coefficient de 1.5 ou 3, mais d'autres auteurs conseillent 1 ou 1.5. L'important, semble-t-il, est d'adopter toujours la même règle lorsqu'on souhaite comparer des diagrammes.

Par rapport aux résumés numériques, le diagramme en boîte et moustaches facilite non seulement la perception des caractéristiques principales des distributions, mais simplifie le repérage des valeurs extrêmes en les localisant les unes par rapport aux autres (figure n° 2.38).

2.4.1. SYSTAT

Pour tracer des diagrammes en boîte et moustaches avec **SYSTAT**, il faut procéder initialement de la même manière que pour le diagramme en tige et feuilles (figure n° 2.39):

• ouvrir le fichier contenant les données à l'aide de l'article **OPEN** du menu **FILE**,
• choisir l'article **BOX** du menu **GRAPH**,
• sélectionner la ou les variables à représenter dans la boîte de dialogue conçue à cet effet.

En plus du tracé «standard» du diagramme en boîte et moustaches, **SYSTAT** propose quelques possibilités supplémentaires intéressantes. En premier lieu, il est possible, pour une variable donnée, de tracer autant de diagrammes qu'il y a de modalités dans une autre variable, discrète, définissant des groupes d'observations. Le tableau

de données sur la Nouvelle Calédonie contient ce type d'information: la variable PROVINCE présente trois modalités: I pour la Province des Iles, N pour la Province du Nord et S pour celle du SUD. Si l'on souhaite obtenir un diagramme du taux de dépendance pour chaque province, il suffit se sélectionner ces deux variables (figure n° 2.40)

Ce procédé permet de comparer les formes des distributions pour chacune des provinces. Des différences sensibles apparaissent très nettement (figure n° 2.41).

On observe nettement que les communes de la Province des Iles ont 3 personnes par actif en activité, avec une très faible dispersion autour de la médiane. Par contre, dans la Province Nord, ce nombre est plus élevé, plus dispersé, et comprend deux cas exceptionnels, les communes de Poum et de Poya qui dépassent 10 personnes. On peut sans doute expliquer ces «anomalies» par l'inadaptation du concept d'actif à la population des tribus indigènes et par les difficultés qu'a sans doute rencon-

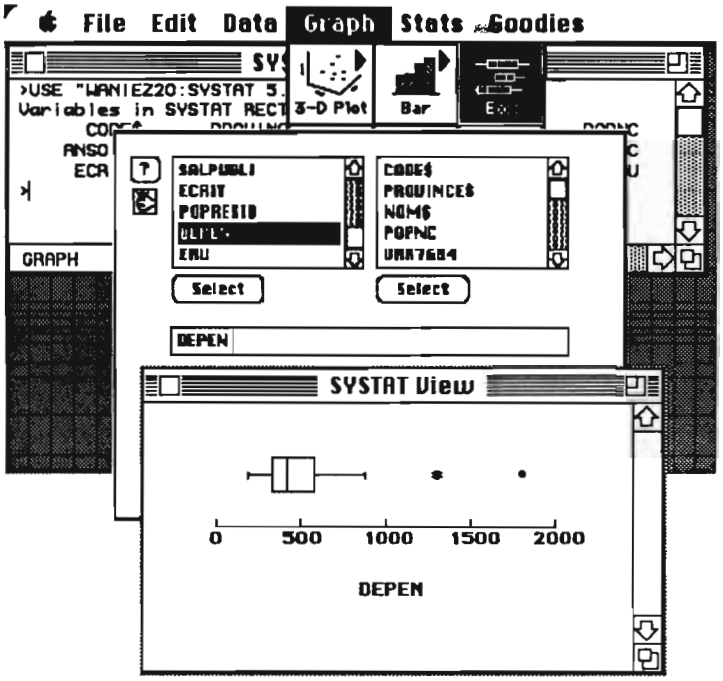


figure n° 2.39. SYSTAT: le tracé du diagramme en boîte et moustache du taux de dépendance.

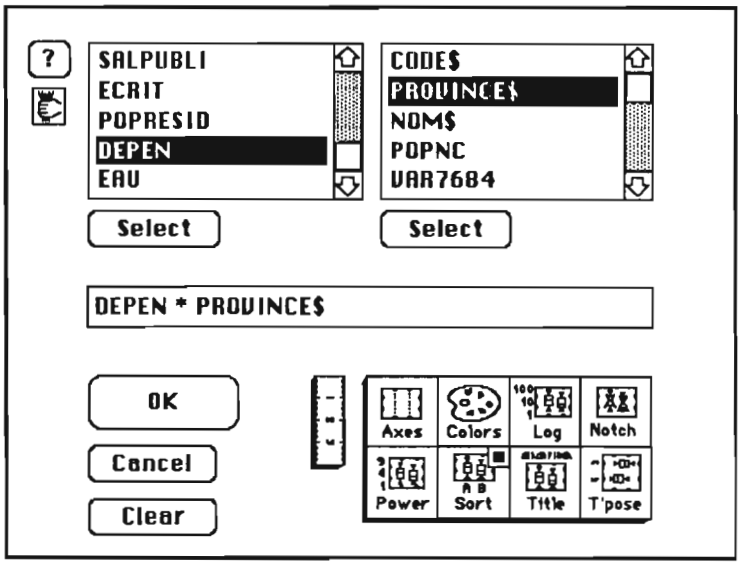


figure n° 2.40. SYSTAT: la sélection des variables pour tracer autant de diagrammes de la variable DEPEN qu'il y a de modalités dans la variable PROVINCE$.

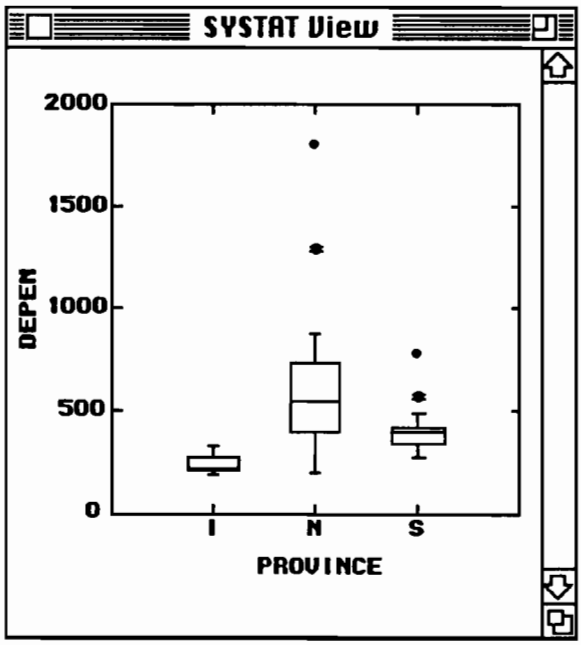


figure n° 2.41. SYSTAT: le diagramme en boîte et moustaches du taux de dépendance des provinces de Nouvelle Calédonie (I=Province des Iles, N=Province Nord, S=Province Sud).

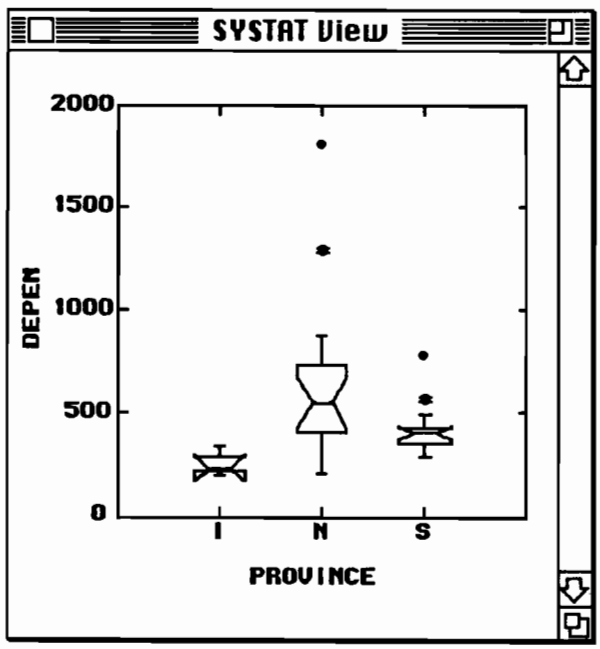


figure n° 2.42. SYSTAT: les diagrammes en boîte et moustaches entaillés du taux de dépendance des provinces de Nouvelle Calédonie.

trées l'INSEE pour effectuer le recensement de 1983. Toujours est-il que le diagramme signale de manière évidente ces cas particuliers qui rendraient peu interprétables les résumés non-résistants que sont les classiques moyenne et écart-type. Enfin, si les communes de la Province du Sud sont situées, en général, en dessous de 5 personnes par actif en activité, Yaté et l'Ile des Pins sont aussi deux cas particuliers; mais la distribution apparaît bien plus resserrée que celle de la Province du Nord, ce qui traduit une certaine homogénéité du taux de dépendance dans les communes concernées.

L'autre possibilité supplémentaire offerte par **SYSTAT** dans ce chapitre réside dans les diagrammes en boîte et moustaches entaillés (*Notched Box Plots*). On y accède en sélectionnant l'option du même nom dans la boîte de sélection des variables. Avec cette extension, le diagramme initial, purement descriptif, se voit complété par un intervalle de confiance à 95% portant sur la médiane et ayant donc valeur probabiliste. L'entaille correspondant à cet intervalle part de la médiane (figure n° 2.42) et se poursuit jusqu'à une valeur au-delà de laquelle, si l'on pouvait prélever 100 échantillons au hasard, seulement 5 médianes présenteraient des valeurs supérieures ou inférieures. Ces encoches sont très utiles pour juger de la significativité des différences entre groupes.

Dans le cas du taux de dépendance, les trois médianes sont significativement différentes les unes des autres: il n'y a aucune correspondance entre les

encoches. Mais d'autres cas de figure peuvent se présenter sur d'autres variables. Par exemple, les diagrammes entaillés de la part des personnes âgées de 0 à 14 ans dans la population totale: s'ils permettent de conclure que la population des communes de la Province des Iles est plus jeune que celle des deux autres provinces, ces diagrammes n'autorisent pas à départager le Nord du Sud, même si les valeurs médianes sont différentes: d'un point de vue probabiliste, cette différence n'est pas significative, au seuil de 5% (figure n° 2.43).

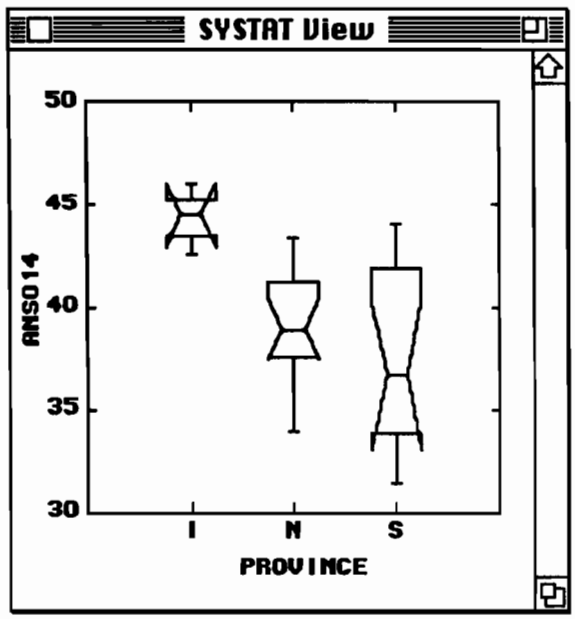


figure n° 2.43. SYSTAT: les diagrammes en boîte et moustaches entaillés de la part des personnes de 0 à 14 ans par rapport à la population totale, dans les provinces de Nouvelle Calédonie.

Cette approche graphique de la significativité des paramètres statistiques, ici la médiane, complète astucieusement les diagrammes en boîte et moustaches. On peut ainsi, de manière simple et immédiate, saisir les différences entre groupes. A titre d'exercice, le lecteur pourra décrire les différences entre les trois provinces calédoniennes, sur le plan de l'équipement en eau des résidences principales et sur celui de l'évolution de la population (figure n° 2.44).

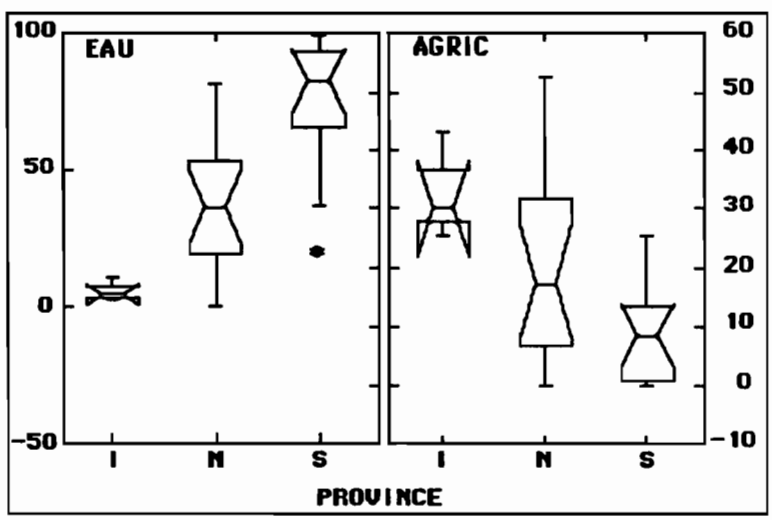


figure n° 2.44. SYSTAT: les Diagrammes en boîte et moustaches entaillés de la part des résidences principales disposant de l'eau courante et de la proportion d'agriculteurs dans la population active.

2.4.2. DataDesk

Pour réaliser un diagramme en boîte et moustaches avec **DataDesk**, il faut sélectionner l'icône d'une variable présente sur le bureau et choisir l'article **BOXPLOTS** du menu **PLOT**. Dans une nouvelle fenêtre intitulée BOXPLOT, le diagramme s'affiche en

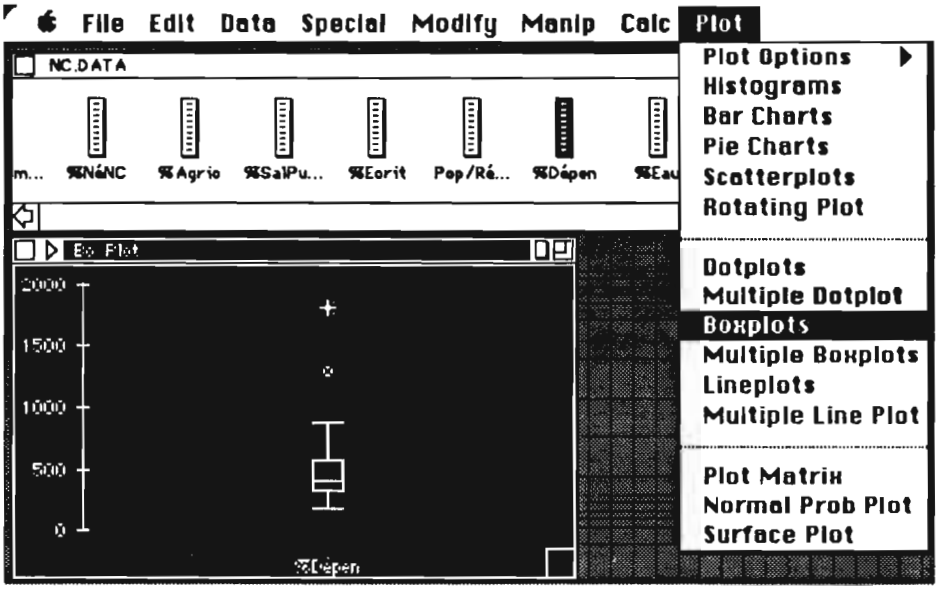


figure n° 2.45. DataDesk: les opérations nécessaires au tracé d'un diagramme en boîte et moustaches.

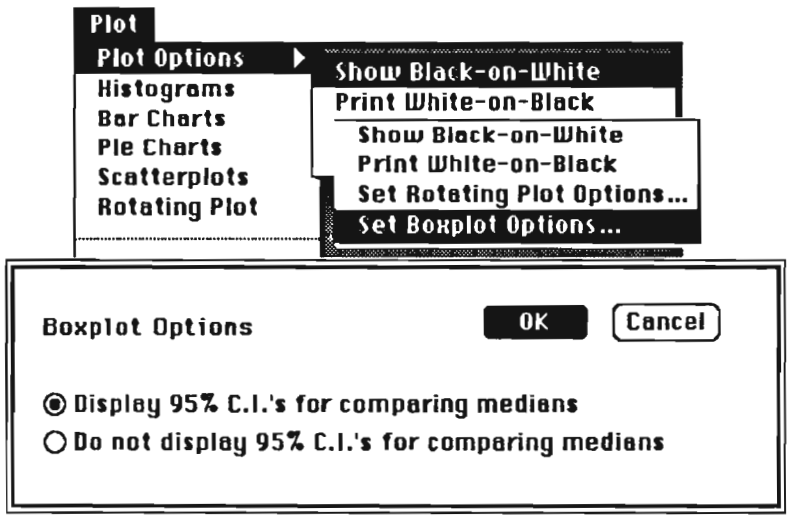


figure n° 2.46. DataDesk: Les options de tracé du diagramme en boîte et moustaches.

blanc sur fond noir, option standard d'affichage du logiciel (figure n° 2.45).

En utilisant certains des articles du sous-menu **PLOT OPTIONS** du menu **PLOT**, il est possible de modifier les caractéristiques de l'affichage (figure n° 2.46). D'une part, grâce à l'article **SHOW BLACK ON WHITE**, le graphique s'affiche en noir sur fond

blanc, ce qui peut s'avérer bien utile pour des documents devant être imprimés. D'autre part, l'article **SET BOXPLOT OPTIONS...** donne accès à une boîte de dialogue permettant d'adjoindre au diagramme un intervalle de confiance de la médiane.

Il ne s'agit pas à proprement parler d'entailles sur le diagramme, comme le fait **SYSTAT**, mais d'une zone ombrée qui se surimpose au diagramme (figure n° 2.47).

La fenêtre d'affichage du diagramme est dotée d'un menu *hyperview* avec

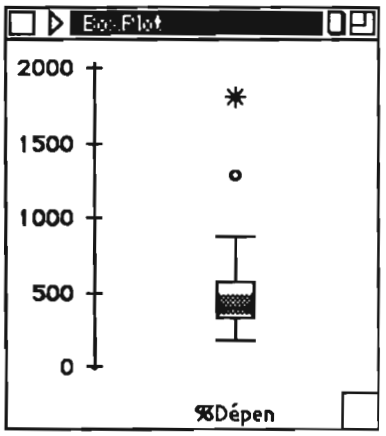


figure n° 2.47. DataDesk: le diagramme en boîte et moustaches du taux de dépendance muni de l'intervalle de confiance à 95% de la médiane.

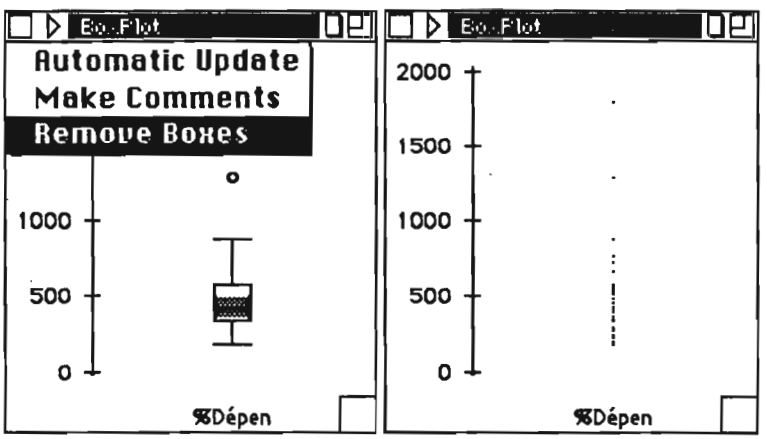


figure n° 2.48. DataDesk: la suppression de la boîte et des moustaches avec le menu hyperview de la fenêtre BOXPLOT.

lequel la boîte et les moustaches peuvent être retirées du graphique pour laisser place aux seuls points représentatifs des observations (figure n° 2.48). Cette option peut s'avérer très pratique car elle autorise, par un lien dynamique, le repérage des points par une sélection dans l'une des fenêtres d'édition des variables (figure n° 2.49).

Pour conclure, signalons que **DataDesk**, permet aussi de réaliser un diagramme en boîte et moustaches pour chaque modalité d'une variable discrète. Il faut d'abord sélectionner l'icône de la variable à représenter, puis celle de la variable contenant les catégories et, enfin, choisir l'article **MULTIPLE BOXPLOTS** du menu **PLOT** (figure n° 2.50).

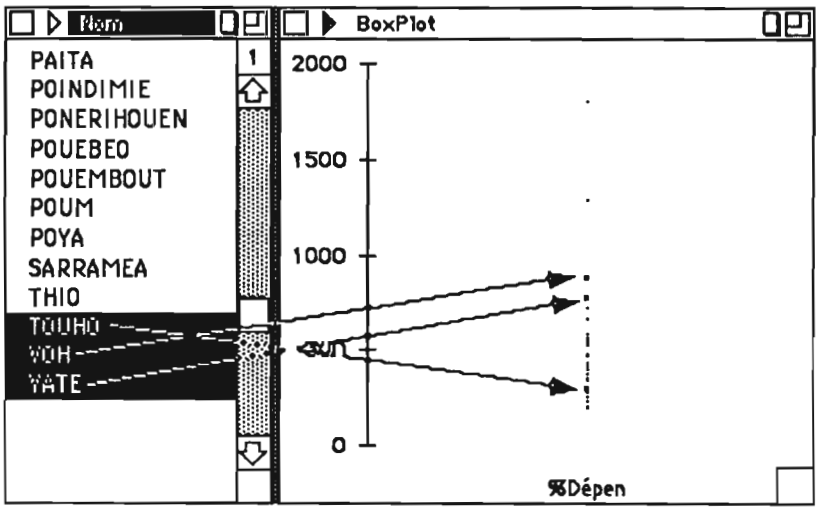


figure n° 2.49. DataDesk: le repérage des observations, sur le diagramme privé de boîte, par un clic dans la fenêtre d'édition de la variable NOM.

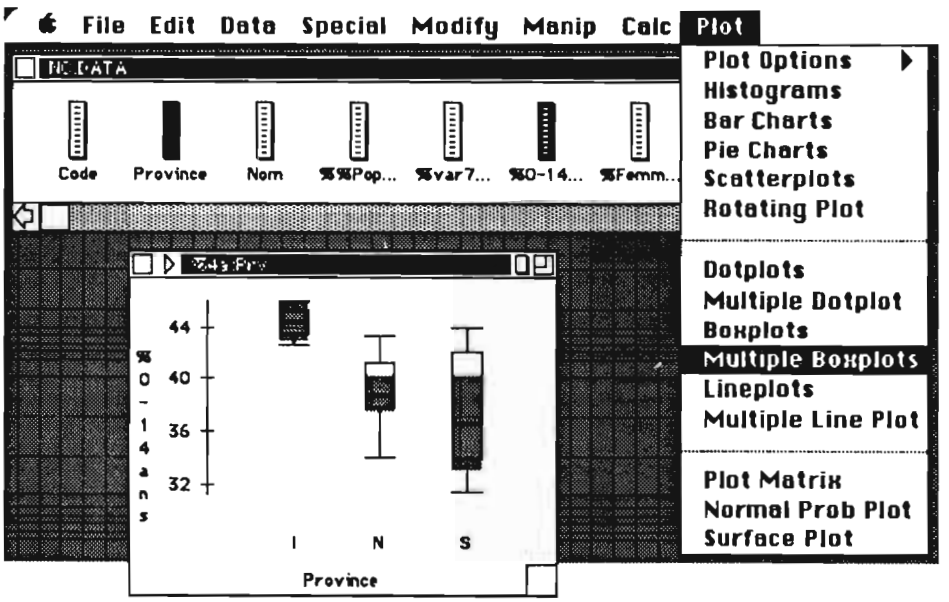


figure n° 2.50. DataDesk: les diagrammes en boîtes et moustaches de la part des 0-14 ans dans la population totale des communes des trois provinces calédoniennes.

Le résultat semble plus lisible que celui obtenu avec SYSTAT, en particulier pour comparer les intervalles de confiance de la médiane.

2.4.3. JMP

JMP réalise les diagrammes en boîte et moustaches en même temps que les résumés numériques résistants. Il faut donc, comme précédemment, utiliser la plate-forme **DISTRIBUTION OF Y'S** du menu **ANALYZE**, après avoir fixé au préalable le rôle Y aux variables pour lesquelles on souhaite obtenir un diagramme.

La fenêtre nommée DISTRIBUTION affiche alors un histogramme de chaque variable retenue, accompagné d'un diagramme en boîte et moustaches ainsi qu'un résumé numérique résistant (figure n° 2.51). Le tracé du diagramme est quelque peu différent de ceux réalisés par SYSTAT ou par **DataDesk**. En effet, les moustaches joignent directement le minimum et le maximum de la variable. Les valeurs exceptionnelles ne sont donc pas soulignées par un symbole particulier. De plus, un losange (appelé *diamond* dans la documentation) figure l'intervalle de confiance à 95% de la moyenne; mais celui de la médiane n'y est pas. Par contre, la proximité de l'histogramme et du résumé numérique constitue une présentation très pratique des caractéristiques de la variable étudiée, et cela d'autant plus que chacun de ces éléments peut être temporairement supprimé de la fenêtre d'affichage par un simple clic sur l'un des boutons prévus à cet effet.

Entre la fenêtre DISTRIBUTION et la fenêtre contenant les tableaux des données, il existe un lien dynamique permettant de repérer dans le tableau de données les observations qui appartiennent à l'une ou l'autre des classes de l'histogramme. Pour cela, il suffit de cliquer sur la classe souhaitée pour voir s'inverser les lignes concernées (figure n° 2.52). Par contre, un clic directement sur le diagramme en boîte et moustaches ne produit aucun effet.

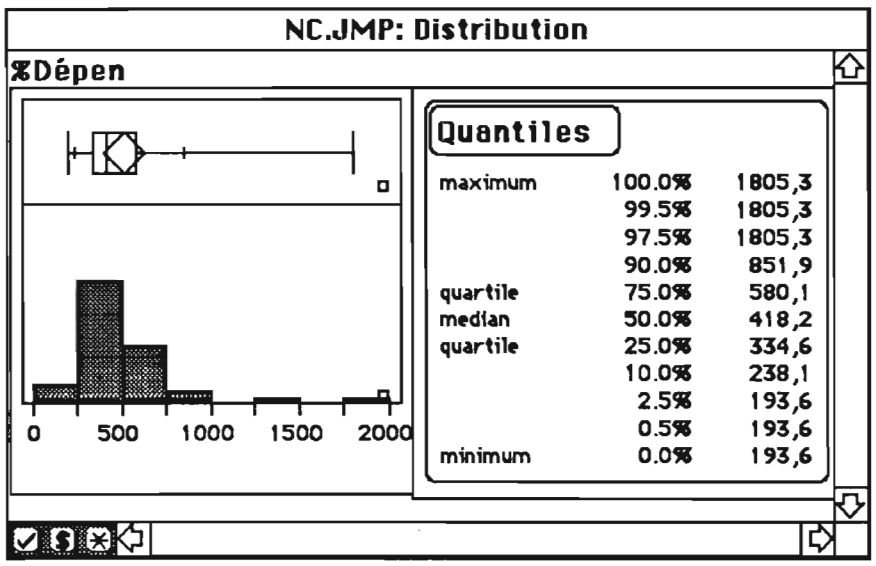


figure n° 2.51. JMP: un diagramme en boîte et moustache accompagné d'un histogramme et d'un résumé numérique résistant.

Dans la logique de **JMP**, la réalisation de diagrammes par groupes, pour chacune des régions par exemple, revient déjà à étudier la relation entre deux variables. Pour ce faire, il faut donc sélectionner l'article **FIT Y BY X** du menu **ANALYZE** après avoir fixé le rôle X à la variable PROVINCE et Y la variable %DEPEN (figure n° 2.53).

La fenêtre Y BY X présente alors pour chaque province, un graphique semblable à celui offert par **DataDesk** lorsqu'on lui demande de supprimer la boîte et les moustaches

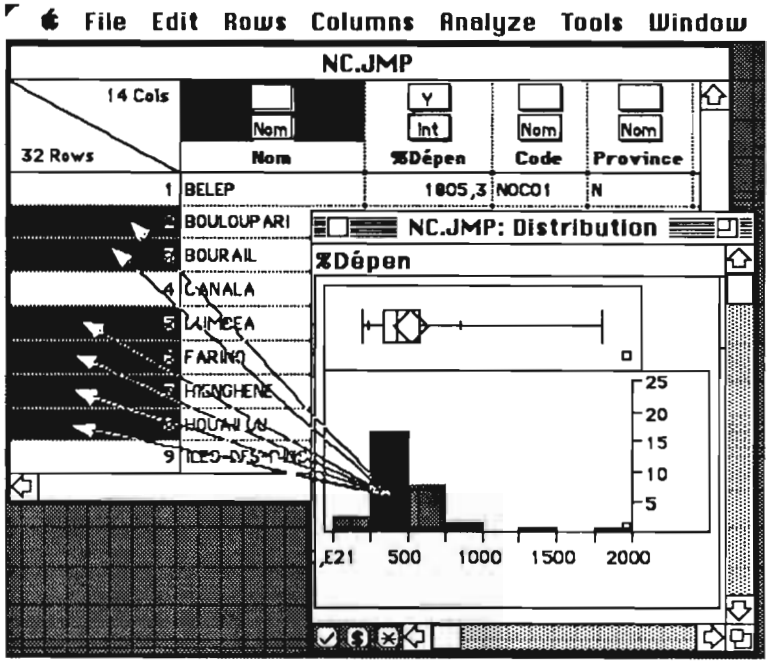


figure n° 2.52. JMP: le fonctionnement du lien dynamique entre l'histogramme et le tableau de données (les flèches ont été ajoutées par la suite).

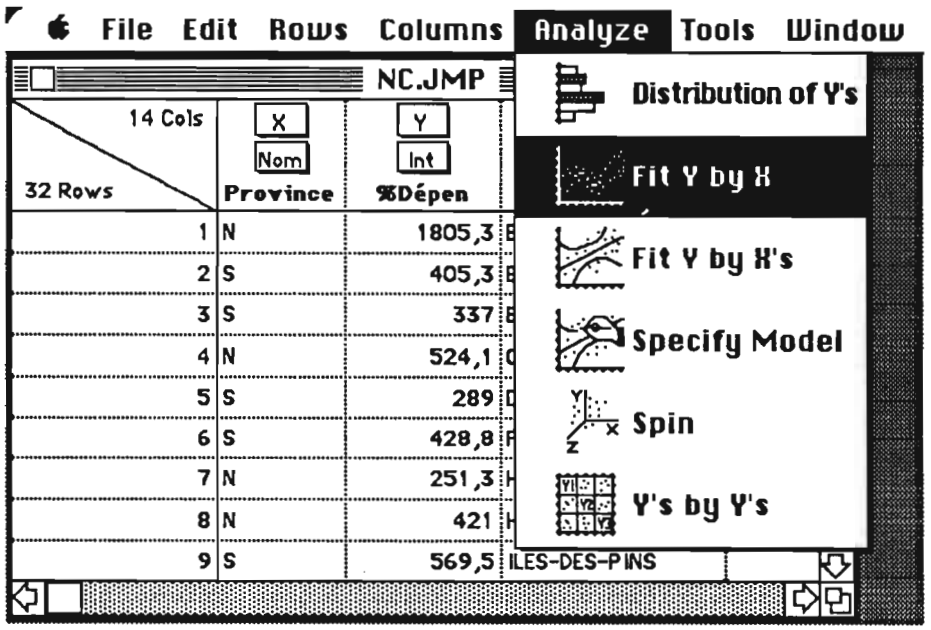


figure n° 2.53. JMP: le choix de l'article FIT Y BY X pour réaliser des diagrammes en boîte et moustaches par groupes.

(figure n° 2.48). Heureusement, un menu hyperview propose l'article **FIT QUANTILES** qui assure le tracé des boîtes (figure n° 2.54). Cela a pour effet de tracer les boîtes ainsi que les premiers et derniers déciles, sans les moustaches (figure n° 2.55).

La largeur des boîtes diffère en fonction du nombre d'observations dans chaque groupe. Ici, la Province des Iles présente donc, avec 3 communes seulement, une taille plus fine que la Province Nord qui compte 16 communes. Cette relativisation, immédiate à la première lecture, permet d'éviter de conclure trop rapidement sur de petits groupes.

Entre la fenêtre Y BY X et la fenêtre contenant les tableau des données, il existe un lien dynamique grâce auquel un clic sur le diagramme provoque le soulignement de l'observation dans le tableau (figure n° 2.56). On notera également sur cette figure que s'affiche automatiquement le numéro de l'observation désignée par l'utilisateur (ici 1).

Par contre, les intervalles de confiance de la médiane, si pratiques pour évaluer le degré de significativité de

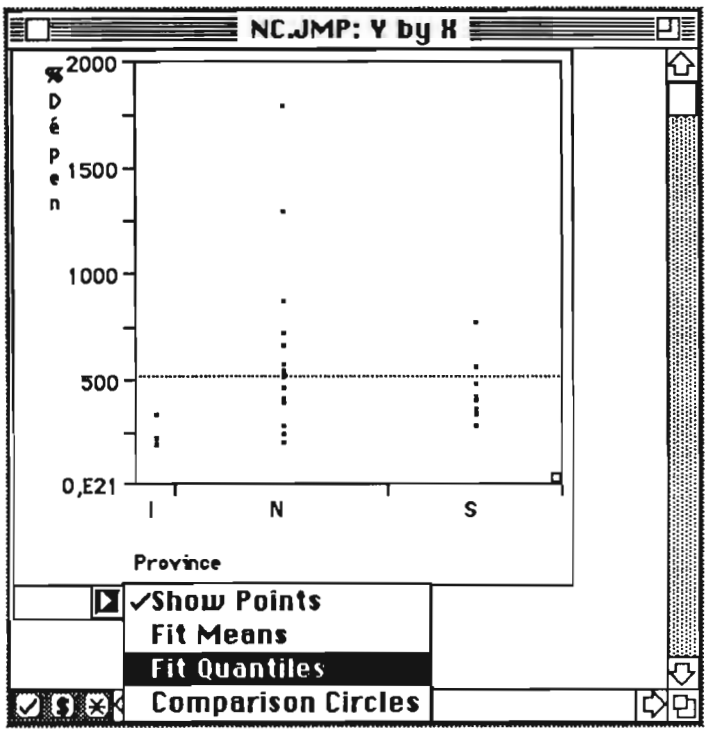


figure n° 2.54. JMP: le graphique des points représentant les valeurs du taux de dépendance dans chaque province.

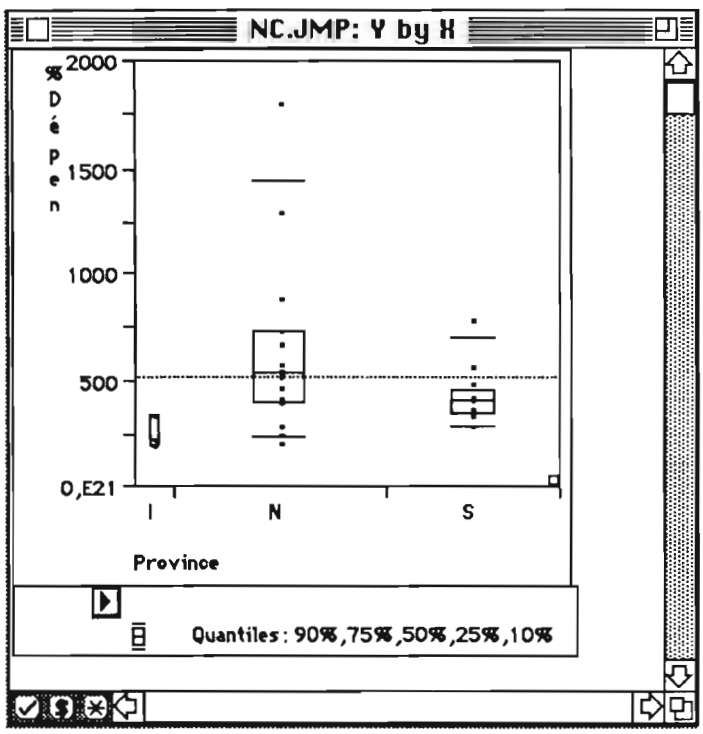


figure n° 2.55. JMP: le tracé des boîtes et des déciles extrêmes.

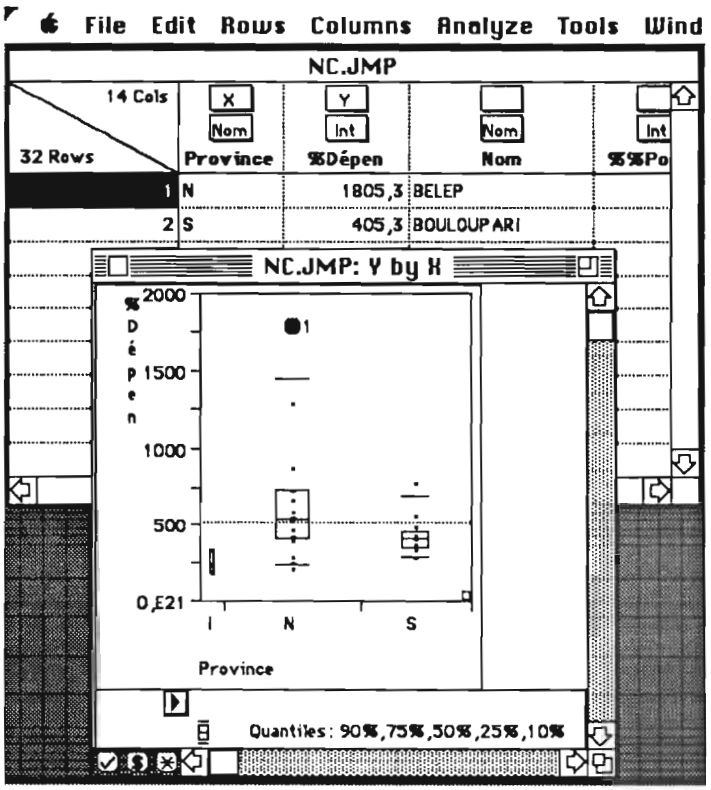


figure n° 2.56. JMP: l'interactivité entre la fenêtre Y BY X et le tableau des données.

différences entre les groupes ne sont pas calculés. D'autres outils astucieux constituent néanmoins des solutions originales pour les moyennes (cercles de comparaison des moyennes par exemple).

En conclusion, dans le domaine des diagrammes en boîte et moustaches, JMP propose des fonctions proches de celles ses concurrents, même si l'approche résistante, si caractéristique de la méthode exploratoire, n'est pas toujours mise en avant.

EXPLORATIONS BIVARIÉES

S'il est parfois suffisant d'examiner quelques variables indépendamment les unes des autres, il faut, dans la majorité des cas, analyser les relations qu'elles entretiennent entre elles. Ceci apparaît d'autant plus nécessaire que l'objet d'étude ne se laisse pas appréhender par une seule série de mesures. L'exploration bivariée nécessite l'examen de trois caractéristiques: l'intensité de la relation, sa direction et sa forme.

L'intensité d'une relation peut être comprise comme le degré de correspondance des valeurs sur une variable par rapport à une valeur sur l'autre variable. Par extension, l'intensité exprime le degré de prédiction des valeurs d'une variable par celles d'une autre variable. Avec la terminologie particulière à l'analyse exploratoire, l'intensité a trait à l'importance relative du lisse et du rugueux (*smooth* et *rough*). Plus la relation est lisse, plus son intensité est forte; en d'autres termes, plus le rugueux subsiste après le lissage des données, plus faible apparaît l'intensité de la relation entre les deux variables. L'une des étapes les plus importantes de l'exploration consiste donc à adopter un lissage des données de manière à exprimer une relation à la fois la plus simple et la plus intense possible.

La direction d'une relation exprime le fait qu'aux valeurs fortes de l'une des deux variables correspondent systématiquement les valeurs fortes de l'autre (et qu'aux valeurs faibles correspondent les valeurs faibles), ou bien encore, qu'aux valeurs fortes de l'une des deux variables correspondent systématiquement les valeurs faibles de l'autre (et qu'aux valeurs faibles correspondent les valeurs fortes). On désigne le premier cas de figure par l'expression «relation positive» et le second par «relation négative». Bien entendu, rien n'empêche l'apparition d'une relation présentant plusieurs directions l'une après l'autre. C'est le cas, par exemple, de toutes les relations en forme de U, d'abord négative puis positive.

Enfin, la forme, troisième caractéristique importante d'une relation, traduit

son modelé, l'allure générale que présente le lissage. Elle peut prendre l'aspect d'une ligne droite, dans le cas où les différences d'une observation à l'autre sont proportionnelles sur les deux variables, ou bien d'une courbe dans les autres cas.

Pour être complète, une exploration bivariée doit permettre d'identifier à la fois la forme, la direction et l'intensité d'une relation. Plusieurs outils d'analyse sont disponibles:

- le graphique bivarié permettant d'apprécier la forme du nuage de points-observations,
- l'ajustement d'une courbe correspondant à une fonction mathématique simple, et respectant certains critères fixés à l'avance
- le lissage des données de proche en proche.

3.1. Les graphiques bivariés

L'examen d'une relation entre deux variables doit toujours commencer par le tracé d'un graphique bivarié (en anglais *scatterplot*) sur lequel chaque axe figure l'une ou l'autre des variables. Les observations sont représentées par un point ayant pour coordonnées leurs valeurs sur les variables. Ainsi, l'observateur se trouve face à un nuage de points sur lequel la direction et dans une moindre mesure, la forme et l'intensité sont immédiatement perceptibles. Pour s'en convaincre, il suffit d'observer quatre exemples très différents de relations (figure n° 3.1).

- Entre la part des salariés du secteur public dans la population active (%SALPUBLIC) et le nombre de personnes pour 100 actifs ayant un emploi (%DEPEN), il existe une relation nettement positive (figure n° 3.1.a). Le nuage de point présente une plus grande dispersion vers les fortes valeurs, et deux communes se détachent avec des chiffres exceptionnellement élevés.

- On observe aussi une relation positive entre la proportion des jeunes de 0 à 14 ans (%0-14ANS) dans la population totale et le nombre de personnes par résidence principale (POP/RESID). Cependant, cette relation semble d'une intensité moins forte car le nuage de points apparaît plus dispersé (figure n° 3.1.b).

- A l'inverse, entre le pourcentage de résidences principales équipées de l'eau courante (%EAU) et le nombre de personnes par résidence principale, la relation est négative, avec une intensité sans doute proche de la précédente (figure n° 3.1.c).

- Enfin, il ne semble pas y avoir de relation entre la part des femmes dans la population totale (%FEMMES) et le nombre de personnes par résidence principale. En effet, le nuage de points apparaît très dispersé sans aucune direction privilégiée (figure n° 3.1.d).

Ces quatre exemples montrent clairement qu'un simple examen des graphiques bivariés renseigne déjà beaucoup sur l'existence de relations entre variables. Bien entendu, ce lien apparent doit pouvoir faire l'objet d'une

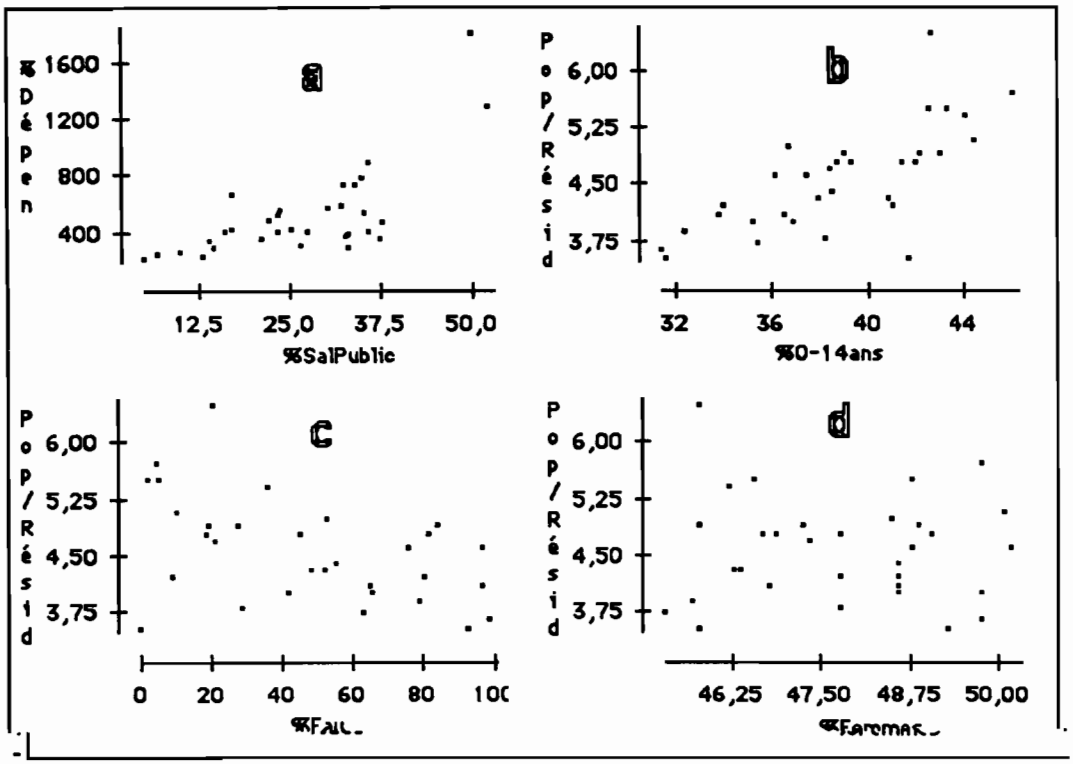


figure n°3.1. Quatre graphiques bivariés présentant des formes, des directions et des intensités différentes.

interprétation en rapport avec le sujet étudié avant même d'être précisé par un quelconque ajustement ou lissage.

3.1.1. SYSTAT

Pour réaliser un graphique bivarié avec **SYSTAT**, il faut, après avoir ouvert le fichier contenant les données, choisir l'article **PLOT** du sous-menu **PLOT** du menu **GRAPH**. Notons que le sous-menu **PLOT** propose 8 types de graphiques bivariés différents (figure n° 3.2). Les types PLOT (graphique bivarié simple) et BORDER PLOT (graphique bivarié dont les axes sont bordés par des diagrammes en boîte et moustaches) sont les plus fréquemment utilisés en analyse exploratoire. Le logiciel ouvre alors un dialogue permettant de sélectionner les variables en ordonnée (Y) et en abscisse (X), ainsi qu'une vingtaine d'options de tracé (figure n° 3.3).

Un clic sur le bouton OK provoque l'affichage du graphique bivarié dans la fenêtre SYSTAT VIEW (figure n° 3.4).

L'article **BORDER PLOT** du sous-menu **PLOT** offre au praticien de l'analyse exploratoire une option très intéressante: au graphique bivarié de base, on ajoute les diagrammes en boîte et moustaches de chacune des deux variables (figure n° 3.5.A).

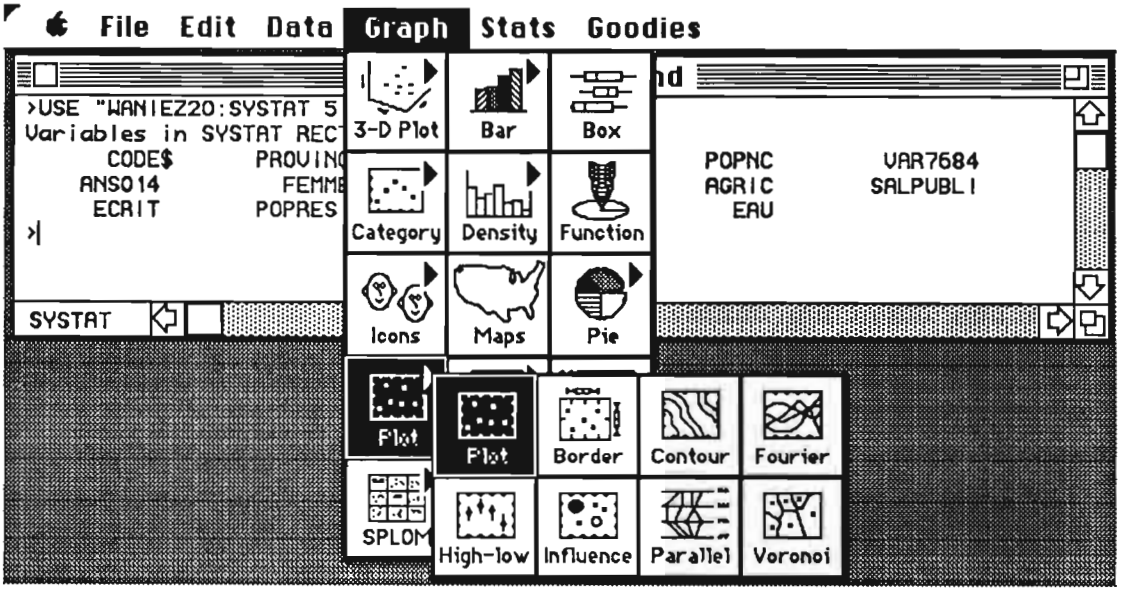


figure n° 3.2. SYSTAT: la sélection d'un type de graphique bivarié.

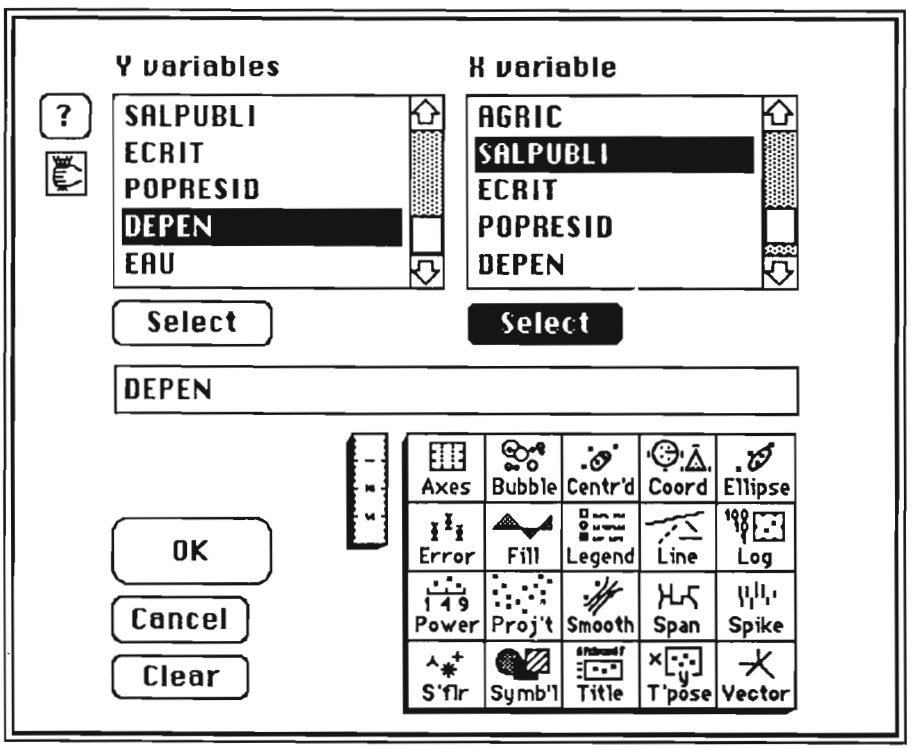


figure n° 3.3. SYSTAT: la sélection des variables Y (en ordonnée) et X (en abscisse).

Une autre manière de rendre compte de la densité de points consiste à adopter l'option STRIPE: à chaque point correspond une ligne perpendiculaire à chaque axe (figure n° 3.5.B). Sur ce second graphique, on observe ainsi, assez nettement, que les valeurs extrêmes EAU sont plus «resserrées» que celles de la variable POPRESID.

Compte tenu du caractère «encyclopédique» de **SYSTAT**, il est impossible de présenter ici tous les types de graphiques bivariés que propose ce logiciel. Disons cependant un mot du type INFLUENCE PLOT auquel on accède par l'article **INFLUENCE** du sous menu **PLOT** du menu **GRAPH**. Sur un tel graphique, les points sont placés comme sur un graphique bivarié simple, mais ils sont représentés par des cercles qui traduisent

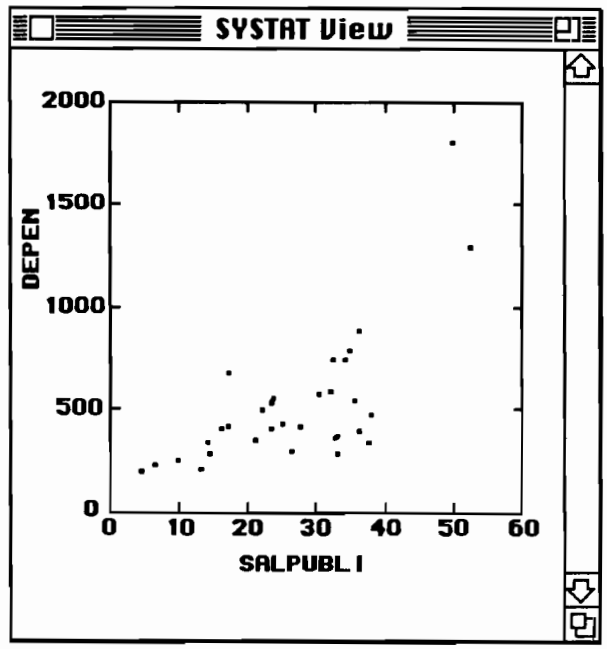


figure n° 3.4. SYSTAT: le graphique bivarié de la part des salariés du secteur public dans la population active (SALPUBLI) et du taux de dépendance (DEPEN)

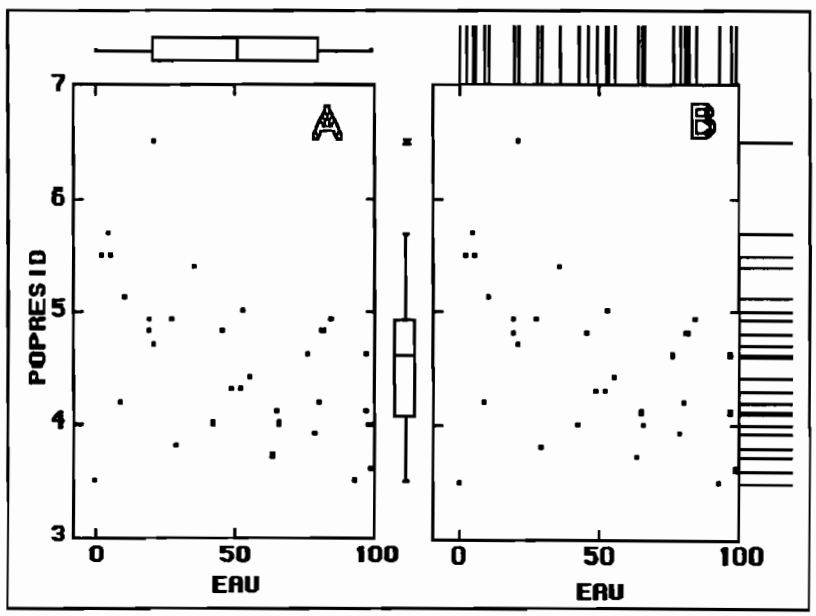


figure n° 3.5. SYSTAT: le tracé du type de graphique bivarié avec l'option BORDERPLOT (A) et STRIPE (B).

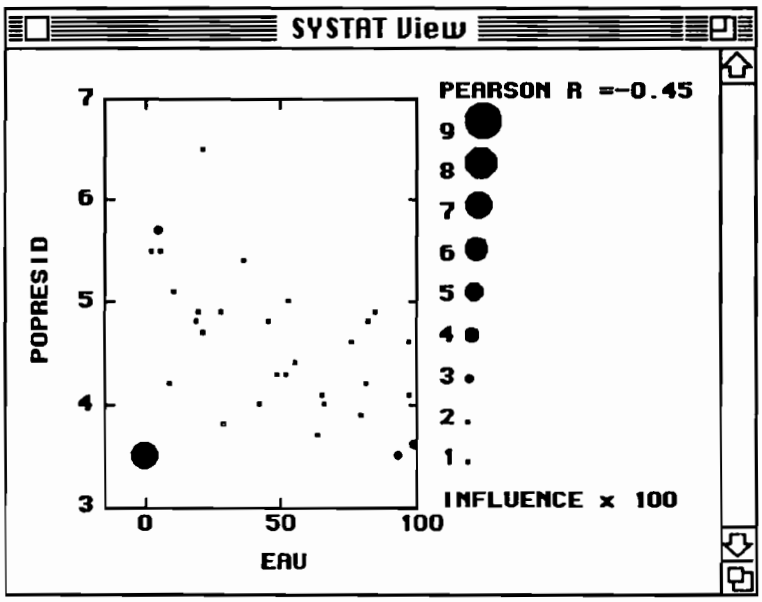


figure n°3.6. SYSTAT: un graphique bivarié avec l'option INFLUENCE.

leur influence sur le coefficient de corrélation linéaire de Pearson.

Par exemple, la relation linéaire entre le nombre de personnes par résidence principale (POPRESID) et la proportion des résidences principales équipées de l'eau apparaît très (trop) influencée (figure n° 3.6) par un point exceptionnel (il s'agit de la commune de Belep). De cette manière, le graphique bivarié d'influence complète utilement l'évaluation de l'intensité d'une relation.

Au total, **SYSTAT** propose de nombreux types de graphiques bivariés dotés d'un grand nombre d'options, parfois introuvables ailleurs.

3.1.2. DataDesk

Lorsque le fichier de données a été ouvert et que les icônes de ses variables sont présentes sur le bureau de **DataDesk**, le tracé d'un diagramme bivarié nécessite d'abord la sélection l'icône de la variable figurant l'axe des ordonnées, puis celle des abscisses. En choisissant l'article **SCATTERPLOT** du menu **PLOT**, la fenêtre graphique s'ouvre et le diagramme est dessiné (figure n° 3.7).

Rappelons que cette fenêtre graphique possède des liens dynamiques avec les fenêtre d'édition des variables qui s'ouvrent lorsqu'on clique sur les icônes correspondantes. Ceci facilite beaucoup le repérage des observations par un simple clic sur les points du diagramme bivarié.

Lorsqu'on souhaite examiner les distributions de chacune des variables composant le graphique bivarié, il suffit de cliquer sur le graphique, à l'emplacement du nom de cette variable. Apparaît alors un menu *hyperview* donnant accès à diverses fonctions comme, par exemple le tracé d'un histogramme qui s'affiche alors immédiatement (figure n° 3.8).

En plus des liens dynamiques entre fenêtres, **DataDesk** renferme une boîte à outils destinés à l'étude des

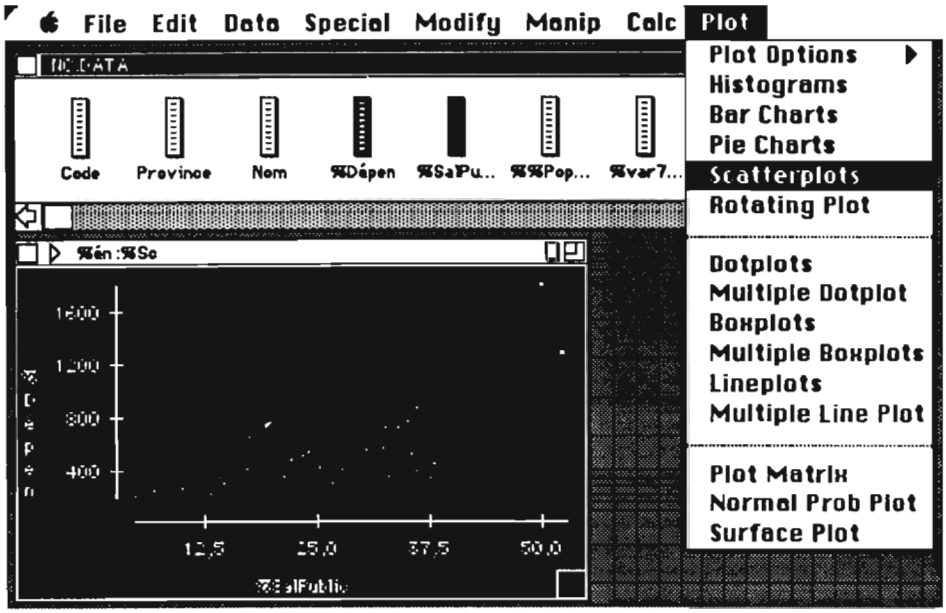


figure n° 3.7. DataDesk: le tracé d'un graphique bivarié.

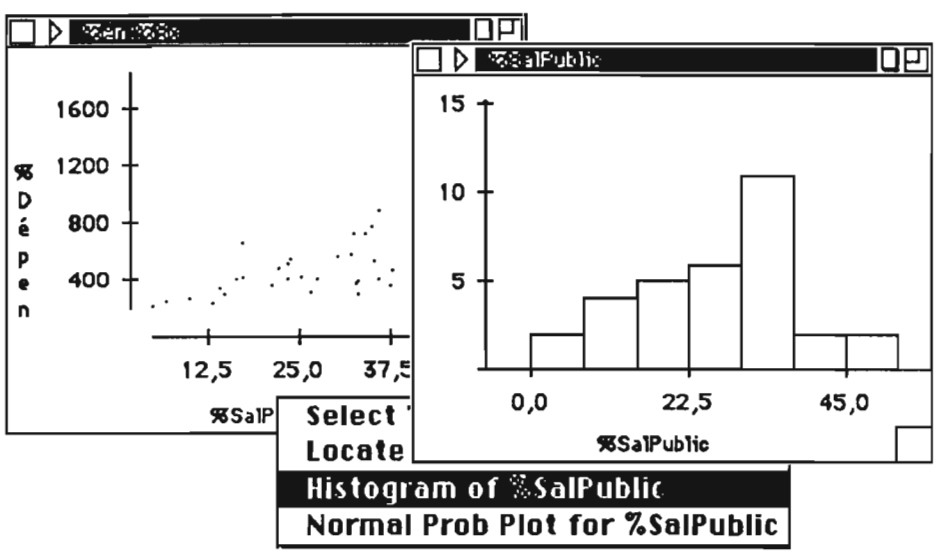


figure n° 3.8 DataDesk: l'histogramme avec l'article HISTOGRAM OF ... du menu hyperview associé au nom de la variable.

graphiques. On y accède par l'article **TOOLS** du menu **MODIFY** (figure n° 3.9). Ces outils graphiques permettent:

- de sélectionner des observations à l'aide du lasso, du rectangle ou du doigt.
- de visualiser les liens entre plusieurs fenêtres. En sélectionnant une partie du graphique, les observations

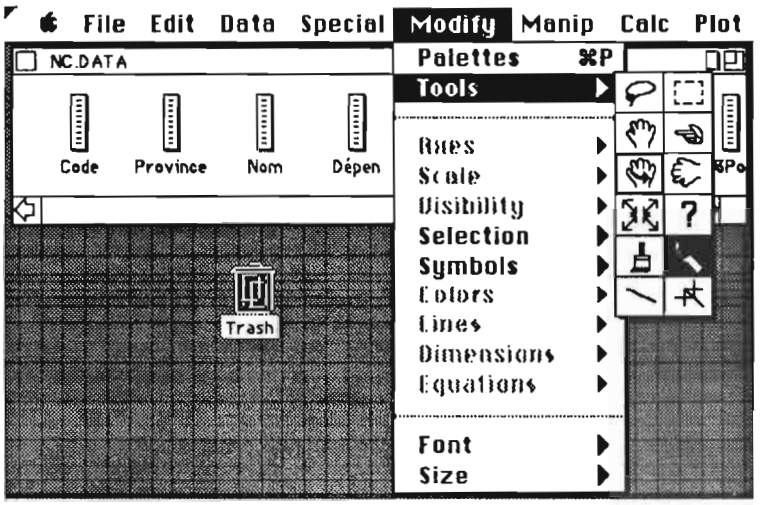


figure n°3.9. DataDesk: les outils d'étude d'un graphique.

qui s'y trouvent sont soulignées dans les autres fenêtres ouvertes, qu'il s'agisse des fenêtres d'édition ou d'autres fenêtre graphiques.

• de réaliser, à l'aide de la brosse et du couteau, le «brossage» ou le «tranchage» (*brushing et slicing*) du nuage de points pour étudier simultanément plusieurs graphiques bivariés, en mettant en évidence des «paquets» de points, ou en découpant des tranches, verticales ou horizontales, pour examiner le comportement de ces tranches par rapport à d'autres variables.

Par exemple, les communes appartenant à la tranche centrée sur 25% de salariés du public dans la population active (figure n° 3.10.A) présentent, en général, un fort accroissement de la population entre 1976 et 1984, sans que la proportion de jeunes dans la population totale soit au maximum (figure n° 3.10.B).

Ces outils présentent donc une très grande nouveauté qui constitue, à notre avis, une progression spectaculaire par rapport à l'analyse exploratoire telle qu'elle est décrite par Tukey. Avec son ingénieuse boîte à outils, les liens dynamiques entre fenêtres, ses menus *hyperview*, les graphiques bivariés de

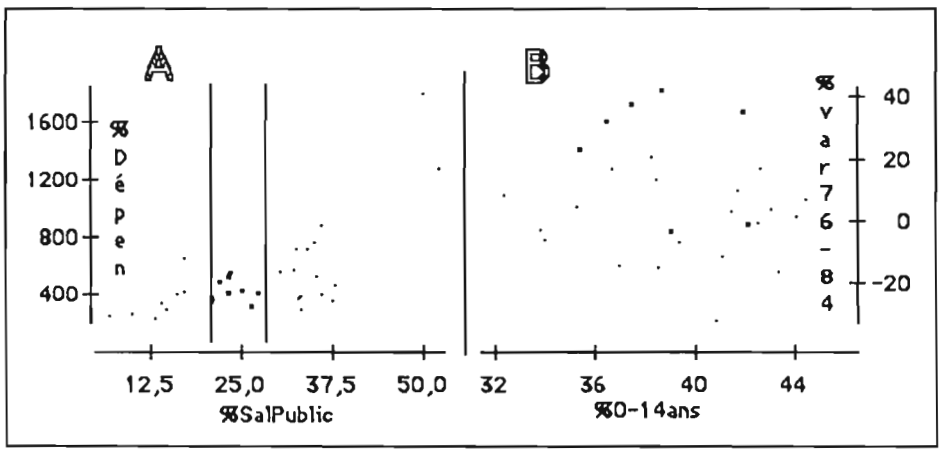


figure n°3.10. DataDesk: le tranchage d'un graphique.

DataDesk forment de réels outils d'exploration interactive des relations entre variables.

3.1.3. JMP

C'est grâce à la plate-forme **FIT Y BY X** du menu **ANALYZE** que **JMP** réalise des graphiques bivariés. L'utilisation de cette plate-forme a déjà fait l'objet d'une présentation au chapitre 2.4, à propos du tracé de diagrammes en boîtes et moustaches par province. Dans ce cas, il s'agissait d'étudier une variable continue en fonction des modalités d'une autre variable, discrète cette fois-là.

Ici, le problème posé est quelque peu différent puisqu'on cherche à percevoir, à l'aide d'un graphique bivarié, la relation pouvant exister entre deux variables continues.

Après avoir fixé les rôles X et Y aux variables concernées (figure n° 3.11), l'activation de l'article FIT X BY Y provoque l'affichage du graphique bivarié (figure n° 3.12) dans une fenêtre intitulée Y by X.

Comme **DataDesk**, **JMP** propose une boîte à outils auxquels on accède en activant les articles du menu **TOOLS** (figure n° 3.13).

On y trouve en particulier la brosse avec des fonctions semblables à celles décrites plus haut: en «brossant» une partie du nuage de points dans un graphique, les points représentant les mêmes individus sur les autres graphiques, ou les lignes du tableau de données sont sélectionnés (figure n° 3.14). Par contre, la documentation de **JMP** ne fait pas état d'un couteau pour le «tranchage» du nuage perpen-

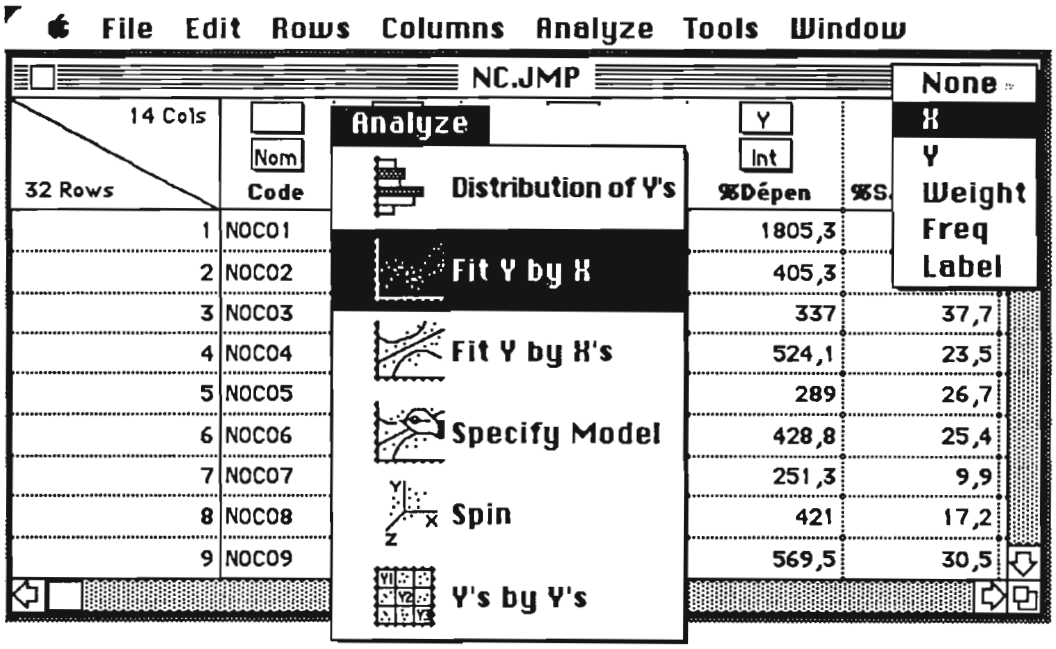


figure n° 3.11. JMP: l'affectation des rôles X et Y aux variables.

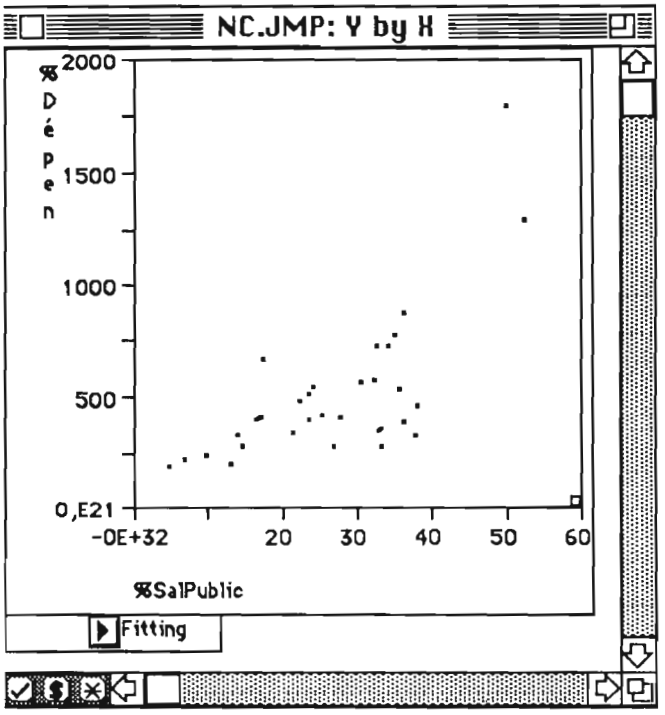


figure n°3.12. JMP: le graphique bivarié tracé par JMP.

diculairement à l'un des deux axes.

JMP présente donc des fonctions proches de celles de **DataDesk** pour le tracé et l'étude des graphiques bivariés. Par contre, **SYSTAT** occupe une place un peu différente: l'interactivité y est plus faible que chez ses concurrents, mais la variété des graphiques disponibles compense en partie cette limitation.

3.2. Résumer une relation

Si les graphiques bivariés facilitent la prise de contact avec une relation statistique, ces graphiques demeurent insuffisants pour

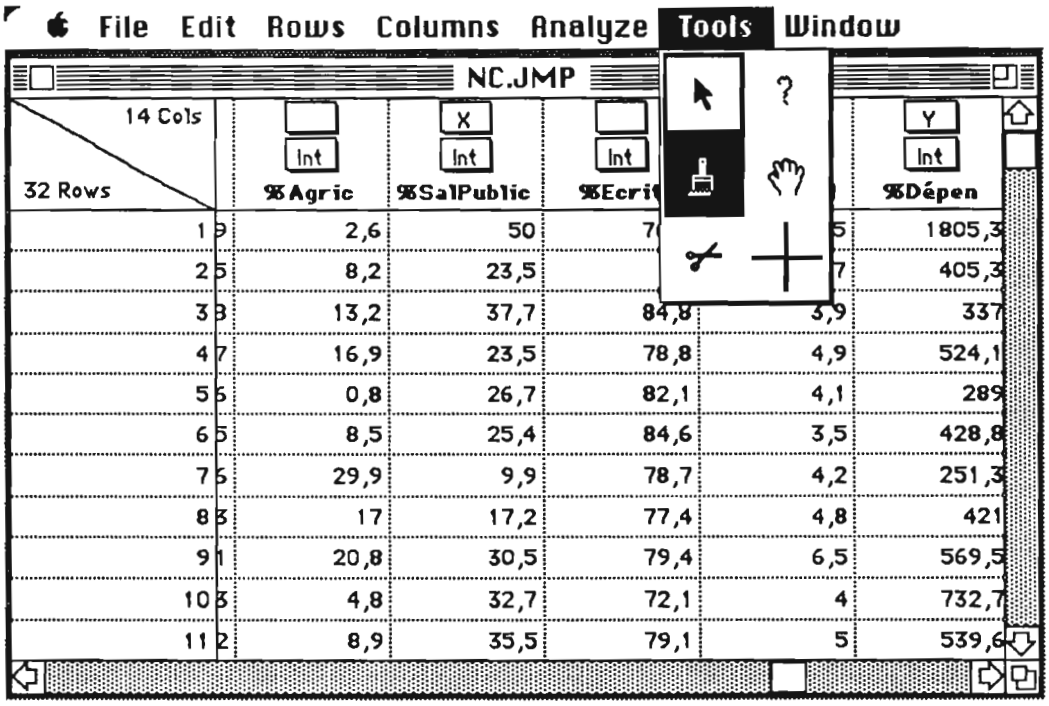


figure n°3.13. JMP: la boîte à outils pour l'étude approfondie des graphiques.

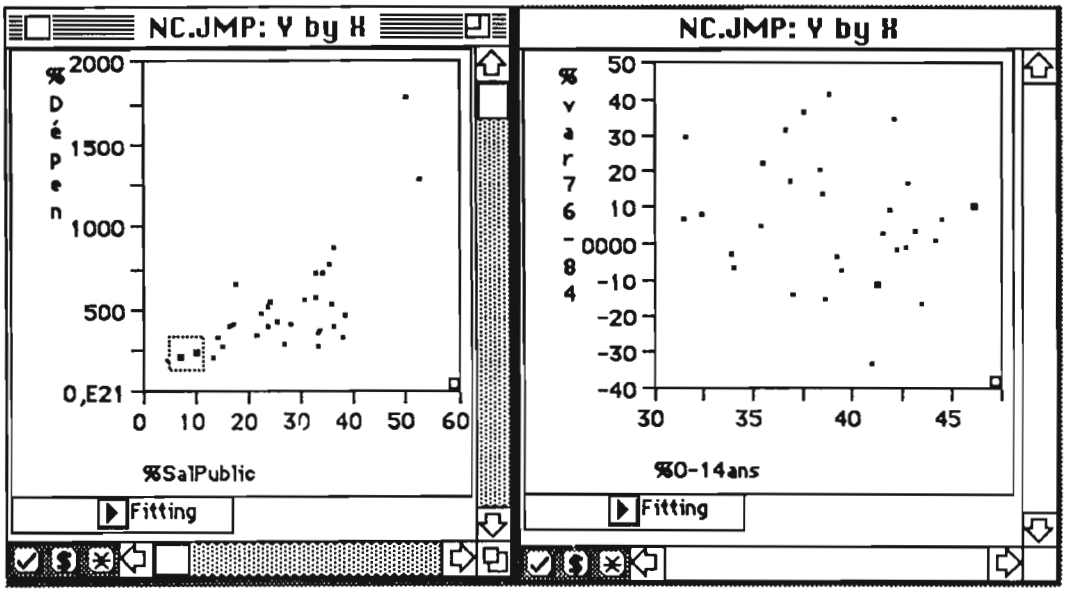


figure n°3.14. JMP: le «brossage» du nuage de points.

exprimer les parts respectives du lisse et du rugueux. Or, il existe de nombreuses techniques, bien plus systématiques, d'évaluation de ce rapport qui ont pour intérêt principal de ne pas dépendre de la perception, souvent subjective, de l'analyste. Elles sont toutes fondées sur le tracé d'une courbe ajustant «au mieux» le nuage de points. Toute la discussion sur le choix de l'une ou l'autre de ces solutions réside, bien entendu, dans ce qu'on considère comme étant «le mieux» pour exprimer une relation. S'agit-il de la forme la plus simple? S'agit-il de rendre compte le plus fidèlement possible des accidents ou ruptures? Tenter de résumer une relation statistique dépend largement, en définitive, des critères retenus pour ajuster la courbe qui, en exprimant le lisse permet d'évaluer le rugueux.

3.2.1. La régression linéaire dans l'environnement exploratoire

S'il est une méthode statistique dont le succès ne se dément pas, c'est bien celle des moindres carrés, avec sa technique la plus connue, la régression linéaire. Rares sont les manuels de statistique qui ne lui consacrent pas un chapitre. Comme il ne s'agit pas à proprement parler d'une méthode exploratoire, le lecteur désirant en savoir plus sur la régression linéaire se reportera aux ouvrages cités en bibliographie. Ici, nous nous limiterons à une présentation «littéraire» de la méthode des moindres carrés visant à montrer qu'elle n'est qu'une technique d'ajustement parmi d'autres.

Rappelons simplement ici que, lorsque le nuage de points présente une forme allongée et inclinée, on peut faire

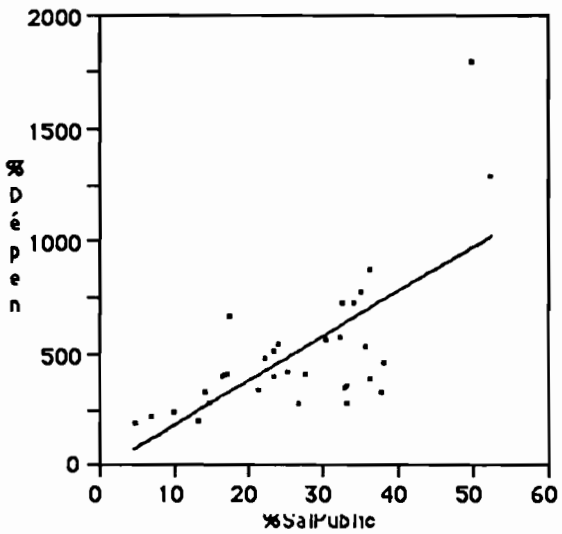


figure n° 3.15. La droite des moindres carrés du taux de dépendance par rapport à la part des salariés du secteur public dans la population active.

l'hypothèse qu'une ligne droite peut rendre compte de cette relation. En quelque sorte, on tente d'ajuster le modèle simple de la ligne droite aux données observées. Par exemple, le graphique bivarié de la part des salariés du secteur public et du taux de dépendance (figure n° 3.1) montre qu'il existe sans doute une relation entre ces deux variables.

Le modèle de la ligne droite a pour équation:

$$Y = aX + b$$

La variable Y s'appelle variable exogène (ou bien encore dépendante, ou à expliquer). La variable X s'appelle variable endogène (ou bien indépendante, ou explicative): elle est supposée influencer les valeurs de la variable Y. Si l'on cherche à estimer le taux de dépendance par la proportion de salariés du secteur public, Y est le taux de dépendance et X les salariés du secteur public.

Le choix d'une variable exogène dépend donc directement de la question qu'on se pose. Puisqu'on connaît les valeurs de Y et de X, il suffit d'évaluer les paramètres a (la pente de la droite) et b (la valeur sur Y lorsque X vaut 0).

Le critère des moindres carrés s'énonce de la manière suivante: la somme des carrés des écarts entre les valeurs observées et les valeurs estimées par l'intermédiaire de la droite de régression doit être la plus petite possible lorsque ces écarts sont mesurés parallèlement à l'axe des Y.

Pour la régression entre les salariés du secteur public (%SALPUBLIC, X) et le taux de dépendance (%DEPEN, Y), l'équation de régression est la suivante (figure n° 3.15):

%DEPEN = 19.9(%SALPUBLIC) - 15.4

Reste à évaluer l'intensité de cette relation. Elle est exprimée par le coefficient R^2 (dit aussi coefficient de détermination) qui varie entre 0 et 1. Il s'agit du rapport entre la somme des carrés des écarts des observations à la droite de régression (dite variation résiduelle) et la somme des carrés des écarts des observations à la moyenne arithmétique de la variable exogène (dite aussi variation totale). On a donc:

$$1-R^2 = \frac{\text{Variation résiduelle}}{\text{Variation totale}}$$

Lorsque R^2 vaut 1, toute la variation de la variable exogène est en relation

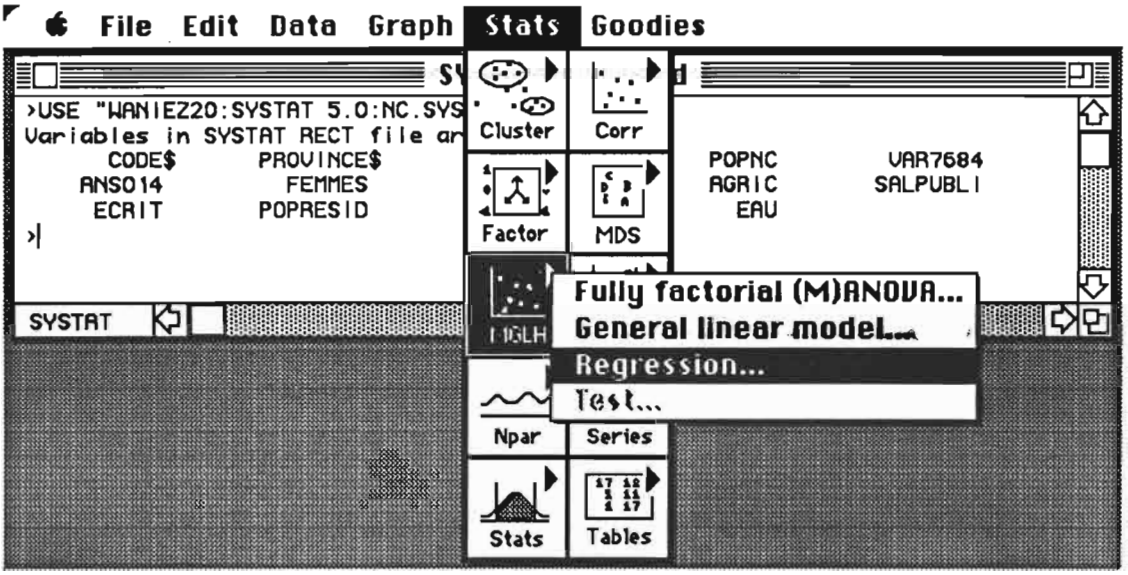


figure n° 3.16. SYSTAT: le choix du menu REGRESSION...

avec la variation de la variable endogène: tous les points sur le graphique bivarié sont alignés. Entre 0 et 1, la valeur de R^2 exprime une variété de situations allant de l'absence de relation, à une forte relation non-fonctionnelle (ne correspondant pas strictement aux valeurs estimées par la droite de régression).

Dans le cas de la relation entre le taux de dépendance et les salariés du secteur public, R^2 vaut 0.48. Ainsi, moins de 48% de la variation du taux de dépendance correspond à la variation des salariés du secteur public. La relation existe, mais elle est ténue.

3.2.1.1. SYSTAT

Après avoir ouvert le fichier de données, **SYSTAT** réalise une régression lorsqu'on sélectionne l'article **REGRESSION** du sous menu **MGLH** du menu **STATS** (figure n° 3.16).

MGLH signifie Multivariate General Linear Hypothesis, ou, en français Modèle Linéaire Multivarié Généralisé. En effet, la régression n'est qu'un cas particulier d'ajustement dans lequel la variable exogène et la (ou les) variable(s) endogène(s) sont continues. Mais on peut très bien réaliser d'autres types d'ajustements sur des variables discrètes, tout en recourant à la méthode des moindres carrés.

SYSTAT ouvre alors un dialogue destiné à la sélection des diverses variables (figure n° 3.17). Il contient, dans sa partie supérieure, une fenêtre dotée d'un ascenseur renfermant la liste des variables du fichier. A sa droite figure une zone d'édition de plusieurs lignes où s'écrivent les noms des variables endogènes (INDEPENDENT). Enfin, dans la zone d'édition intitulée DEPENDENT doit apparaître le nom de la variable exogène. Dans ces deux

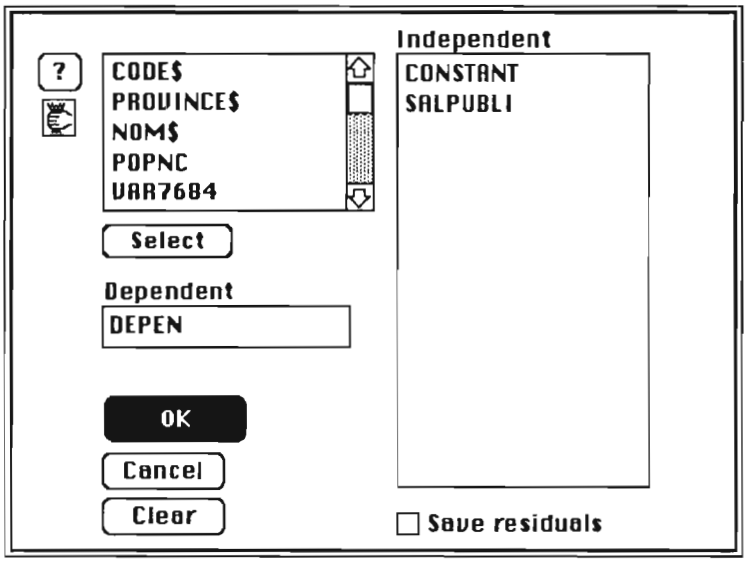


figure n° 3.17. SYSTAT: le dialogue de sélection de la variable exogène et des variables endogènes.

boîtes d'édition, les noms peuvent être saisis à partir du clavier, ou bien être transférés, par un simple clic, à partir de la fenêtre listant les noms des variables. Notons que, pour une régression avec un terme constant (cas le plus courant), la première variable endogène doit s'appeler CONSTANT.

Lorsque les calculs sont achevés, une nouvelle fenêtre s'ouvre: elle présente les résultats des calculs en trois parties (figure n° 3.18):

- un résumé numérique donnant, en particulier la valeur du R^2 (SQUARED MULTIPLE R).
- le tableau des coefficients de régression
- une analyse de la variance.

Le tracé de la droite de régression est indépendant des calculs. Il s'agit en fait de l'option SMOOTH du tracé des graphiques bivariés. SYSTAT propose une très grande variété de lissages (figure n° 3.19). En choisissant l'option LINEAR, le logiciel trace sur le graphique bivarié la droite de régression.

En conclusion, SYSTAT réalise correctement le calcul et le tracé d'une

SYSTAT Analysis

DEP VAR: DEPEN N: 32 MULTIPLE R: .695 SQUARED MULTIPLE R: .483
ADJUSTED SQUARED MULTIPLE R: .466 STANDARD ERROR OF ESTIMATE: 238.997

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	-15.362	109.404	0.000	.	-0.140	0.889
SALPUBLI	19.889	3.759	0.695	1.000	5.292	0.000

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	1599437.882	1	1599437.882	28.002	0.000
RESIDUAL	1713581.997	30	57119.400		

figure n° 3.18. SYSTAT: le tableau des résultats d'une régression.

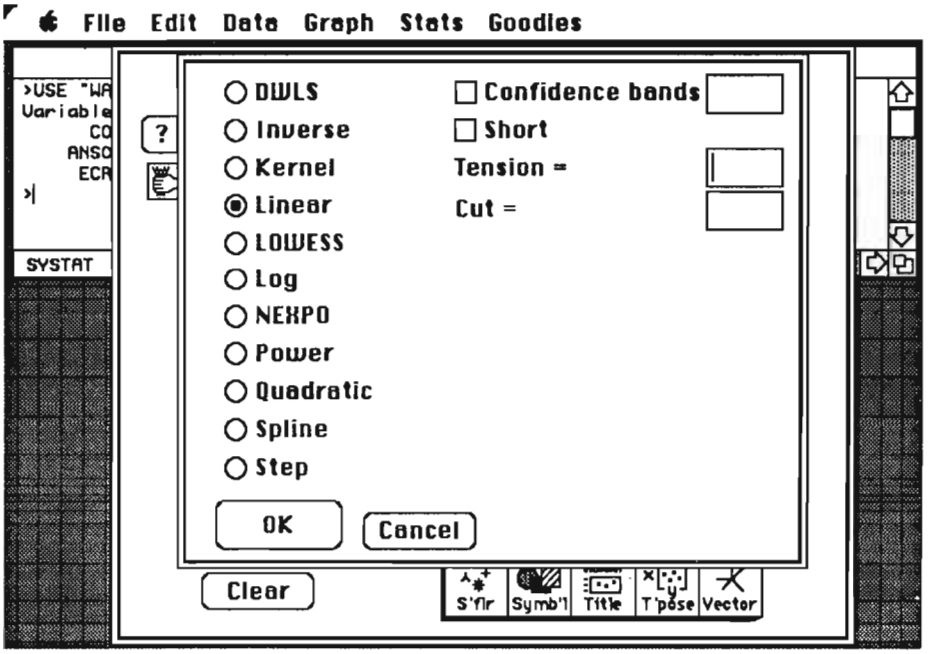


figure n°3.19. SYSTAT: les options de lissage proposées par SYSTAT.

droite de régression, mais de manière très «classique», avec peu d'interactivité.

3.2.1.2. DataDesk

Une fois fixées la variable Y et la ou les variables X (par un clic sur leurs icônes), **DataDesk**, comme **SYSTAT**,

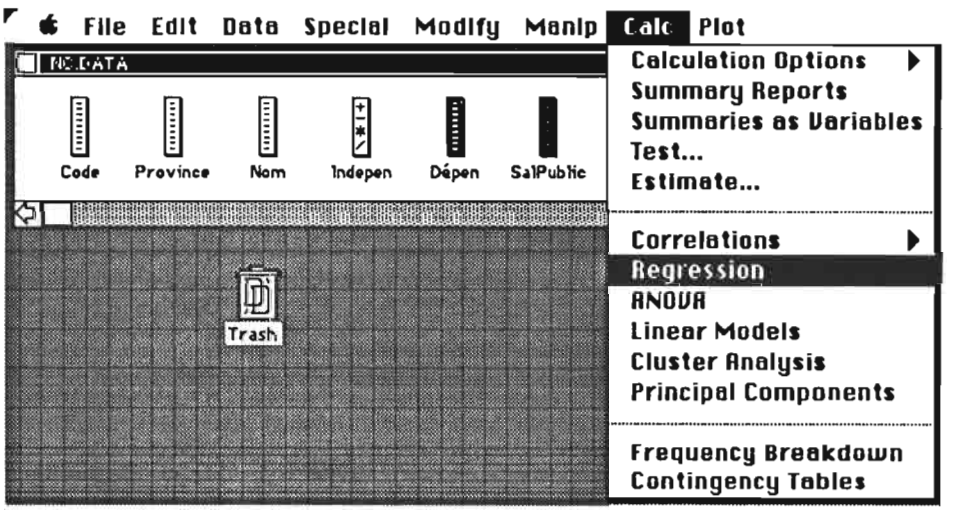


figure n°3.20. DataDesk: la sélection de la régression.

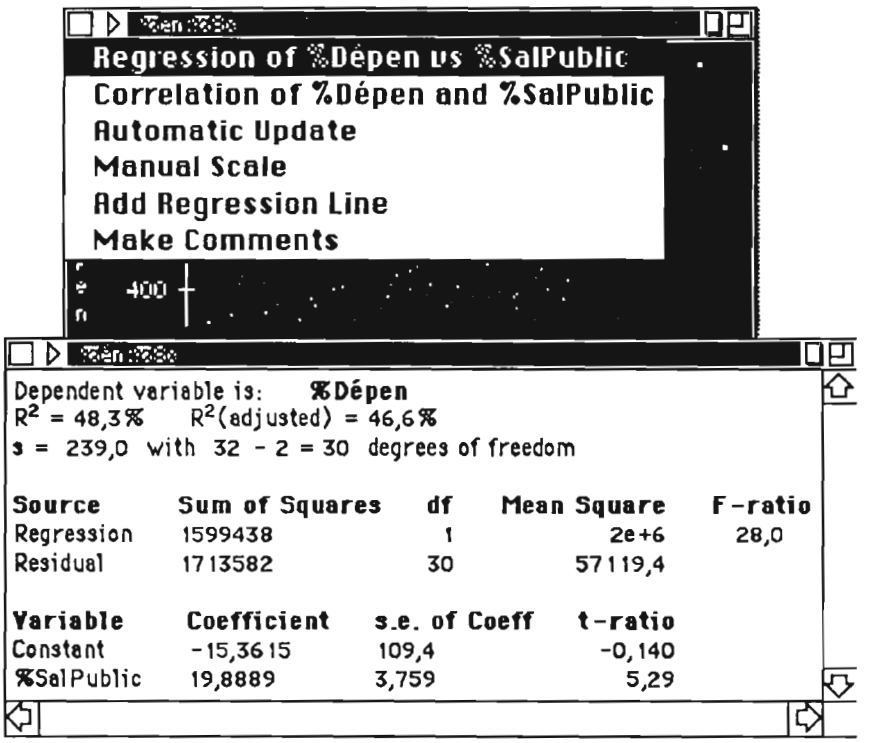


figure n° 3.21. DataDesk: les calculs de régression effectués par le menu hyperview du graphique bivarié.

réalise une régression à l'aide de l'article **REGRESSION** du menu **CALC** (figure n° 3.20).

Cette voie «classique» se double d'un chemin de traverse très bien adapté à l'approche exploratoire. **DataDesk** propose en effet, sur la fenêtre d'un graphique bivarié, un menu *hyperview* donnant accès à la régression linéaire, et au tracé de la droite de régression. En choisissant dans le menu *hyperview* l'article **REGRESSION OF ...**, les résultats des calculs s'affichent à l'intérieur d'une nouvelle fenêtre. De manière très classique, on y trouve le coefficient de détermination, une analyse de la variance et les coefficients de régression (figure n° 3.21)

Toujours dans le menu *hyperview* du graphique bivarié, l'article **ADD A REGRESSION LINE** ajuste une droite des moindres carrés au nuage de points tracé précédemment (figure n° 3.22).

Les outils de brossage et de tranchage décrits à propos des graphiques bivariés demeurent accessibles après le tracé de la droite de régression. **DataDesk** offre ainsi un environnement d'analyse dans lequel la régression apparaît elle-même comme un objet d'exploration. De plus, les liens dynamiques entre toutes les fenêtres ouvertes facilitent l'identification des observations sur le graphique, et donc la réflexion sur l'ajustement obtenu.

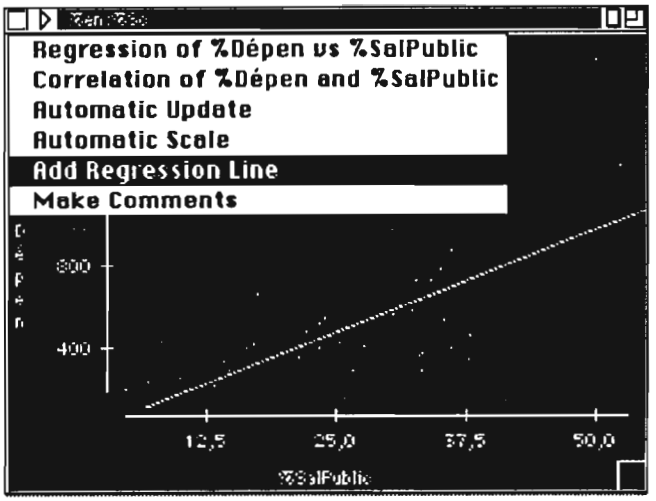


figure n°3.22. DataDesk: le tracé d'une droite de régression par le menu hyperview du graphique bivarié.

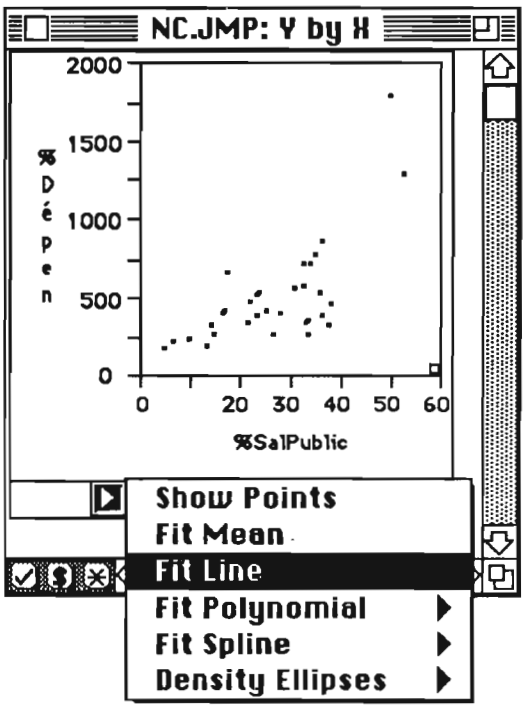


figure n°3.23. JMP: la réalisation de l'ajustement linéaire par l'article FIT LINE du menu pop-up FITTING.

3.2.1.3. JMP

Comme pour les graphiques bivariés, c'est par la plate-forme **FIT Y BY X** du menu **ANALYZE** qu'on réalise la régression avec **JMP**. Lorsque le graphique bivarié a été tracé à l'écran, de la manière décrite en 3.1, il suffit de choisir le menu *pop-up* **FITTING** et de sélectionner l'article **FIT LINE** (figure n° 3.23). On notera ici que **JMP** offre d'autres possibilités d'ajustement, polynomial ou par splines; elles seront examinées plus loin dans ce chapitre.

Le logiciel trace alors la droite de régression sur le graphique bivarié. Un nouveau menu *pop-up*, intitulé LINEAR FIT fait sont apparition. Il permet de choisir les couleurs du graphique, et surtout, d'enregistrer les valeurs estimées et les résidus (figure n° 3.24).

Un clic sur le bouton LINEAR FIT provoque l'affichage du tableau de régression qui comprend un résumé

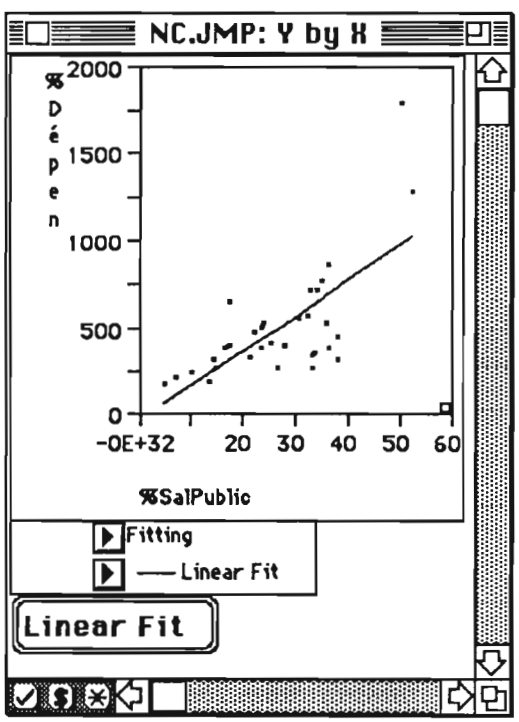


figure n°3.24. JMP: le tracé d'une droite de régression.

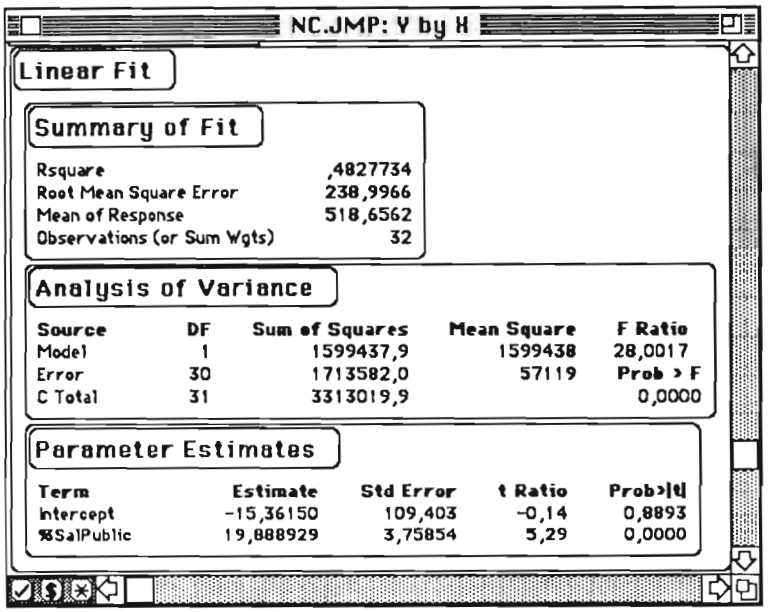


figure n°3.25. JMP: le tableau numérique de la régression.

numérique contenant le coefficient de détermination, une analyse de la variance et les coefficients de régression (figure n° 3.25).

Un autre clic sur le bouton LINEAR FIT supprime l'affichage du tableau: l'utilisateur peut ainsi organiser sa lecture sans encombrer le bureau du Macintosh.

Dans le domaine de la régression linéaire, il existe donc une très grande ressemblance entre **JMP** et **DataDesk**: l'analyse est menée directement sur le graphique bivarié, ce qui permet d'éviter les ajustements abusifs. Il apparaît donc clairement que l'approche exploratoire n'exclut pas l'usage approprié des techniques de la statistique paramétrique comme la régression linéaire. Mais, en règle générale, on évitera bien des erreurs d'interprétation en faisant précéder ces calculs d'une exploration des nuages de points à l'aide des outils adéquats.

3.2.2. Examiner les résidus dans l'environnement exploratoire

L'une des difficultés d'utilisation de la droite des moindres carrés réside dans sa sensibilité au fortes valeurs exceptionnelles. Cela n'est pas étonnant puisque ce sont les carrés des écarts qui entrent dans le calcul des paramètres de la droite de régression. Dans leur ouvrage sur l'analyse exploratoire, F. Hartwig et B.E. Dearing citent le cas d'une relation qui, bien qu'apparaissant nettement positive sur le graphique bivarié, présente, avec la droite des moindres carrés, une direction négative. Ce basculement est dû à quelques observations particulières.

La validation d'un ajustement apparaît donc nécessaire avant d'en tirer des conclusions (et peut-être même des décisions). Le graphique bivarié permet, dans bien des cas, de vérifier l'existence d'une relation. Mais il est souvent nécessaire d'aller plus loin:

- de vérifier que les écarts entre les valeurs observées de la variable exogène et celles calculées par la droite de régression, nommés résidus, d'une part ne varient pas systématiquement, et d'autre part, ne présentent pas de valeurs exceptionnellement élevées.
- de choisir un autre type d'ajustement que la droite des moindres carrés, soit linéaire, comme la droite de Tukey, soit non-linéaire, comme les ajustements polynomiaux.

On peut toujours faire passer une droite dans un nuage de points; cela ne signifie par pour autant que cette droite a une quelconque signification. Pour qu'une droite de régression soit correcte, il faut que les résidus ne présentent pas de configuration particulière. Un graphique bivarié dont l'axe des abscisses figure la variable endogène, et celui des ordonnées les résidus permet souvent de détecter un mauvais ajustement.

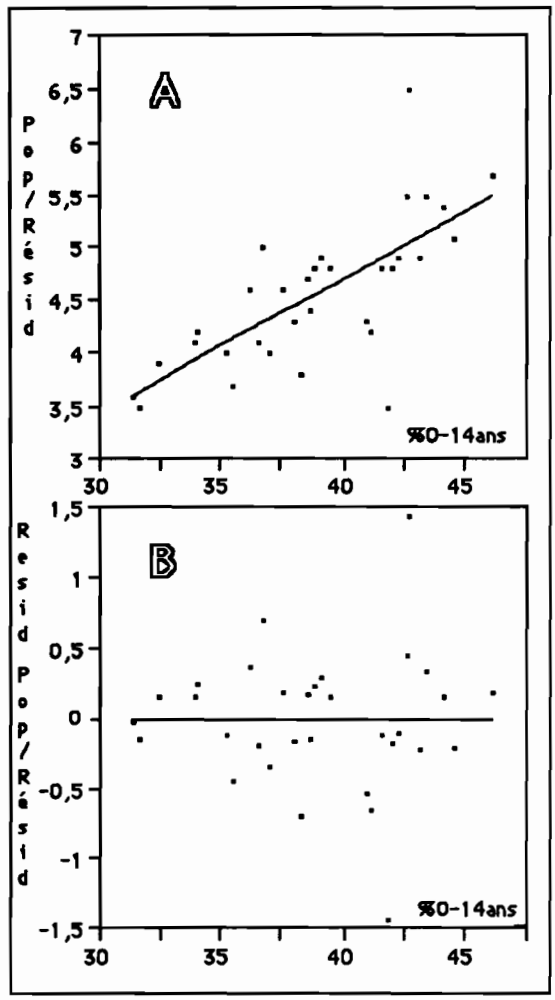


figure n° 3.26: Régression du nombre de personnes par résidence principale par rapport aux 0-14 ans dans la population totale.
A: droite de régression, B: graphique des résidus.

Le cas de figure idéal se présente lorsque les résidus sont disposés de part et d'autre d'une ligne parallèle à l'axe des abscisses ayant pour origine la valeur 0 sur l'axe des ordonnées (elle représente donc la droite de régression). On observe un phénomène de ce type sur la régression du nombre de personnes par résidence principale (Y) par rapport à la proportion des personnes âgées de 0 à 14 ans dans la population totale (X) (figure n° 3.26.A). La direction de la droite est positive: dans les communes présentant une grande proportion de jeunes, les résidences principales abritent un plus grand nombre de personnes. Les résidus ne présentent pas de forme particulière de part et d'autre de la droite de régression (figure n° 3.26.B).

Mais d'autres cas, bien moins favorables, peuvent survenir. On peut distin-

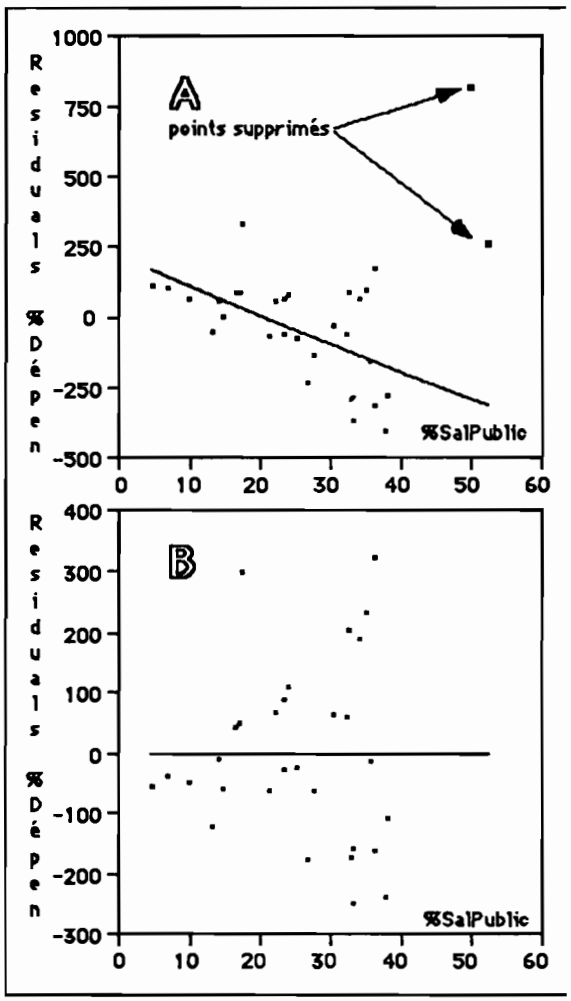


figure n° 3.27. Résidus de la régression du taux de dépendance par rapport à la part des salariés du secteur public dans la population active. A: avec Belep et Poum, B: sans.

guer deux cas de figure bien différents. En premier lieu, il arrive que les résidus présentent une nette direction sur leur graphique (figure n° 3.27.A). C'est le cas avec la régression du taux de dépendance (Y) par rapport à la part des salariés du secteur public dans la population active (X). A l'exception de deux communes «atypiques», Belep et Poum, les résidus décroissent très nettement: La droite est sans doute mal ajustée à cause des deux communes «bizarres». Ceci justifie qu'elles soient exclues, et que l'on recommence tous les calculs.

En second lieu, la régression ne peut être valide lorsque les résidus ont tendance à augmenter ou à diminuer avec les valeurs de la variable endogène. Le nuage de points prend dans ce cas la forme d'un triangle dont la droite de régression suit à peu près l'une des médianes

Un phénomène de ce genre fait son apparition après que Belep et Poum aient été écartées de la précédente régression: les résidus s'accroissent alors que le pourcentage de salariés du public augmente. Là encore, la régression n'est pas correcte (figure n° 3.27.B).

La suppression des observations considérées comme «aberrantes» ne conduit donc pas obligatoirement à un bon ajustement. Cette solution, pratique à première vue, s'avère souvent inopérante et de plus, a pour conséquence une perte d'information difficile à contrôler.

3.2.2.1. SYSTAT

L'examen des résidus avec **SYSTAT** nécessite leur stockage préalable dans un fichier. Il suffit, pour cela, de cocher l'option SAVE RESIDUALS du dialogue destiné à la sélection des diverses variables (figure n° 3.17). Le nouveau fichier, dont le nom par défaut est celui du fichier contenant les variables de la régression suivi du suffixe WORK, peut contenir divers résultats de calculs (figure n° 3.28).

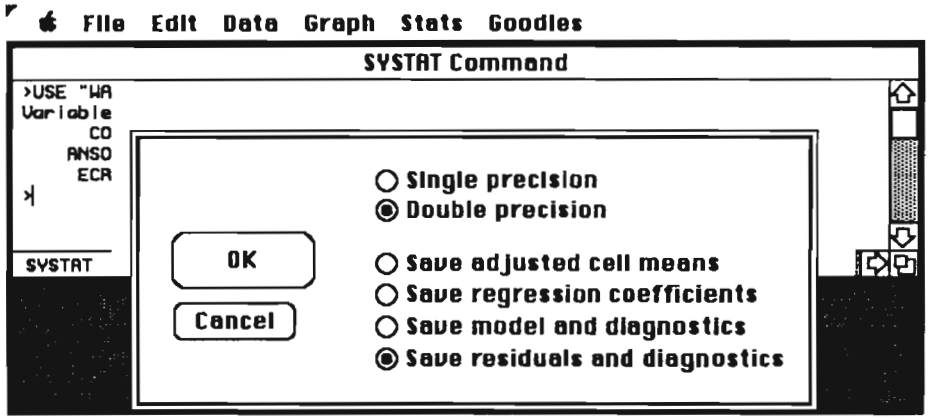


figure n°3.28. SYSTAT: l'enregistrement des résidus dans un fichier.

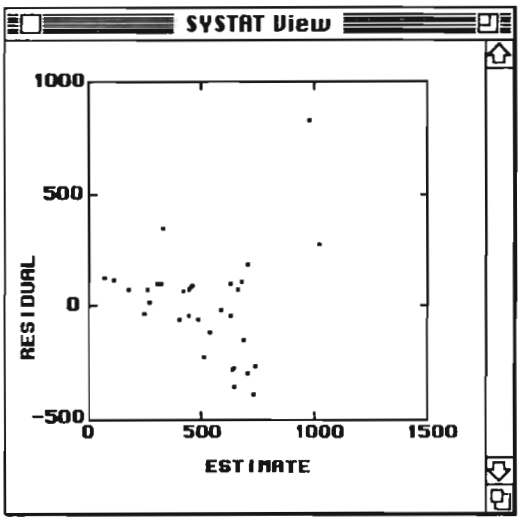


figure n°3.29. SYSTAT: le graphique des résidus par rapport aux estimations.

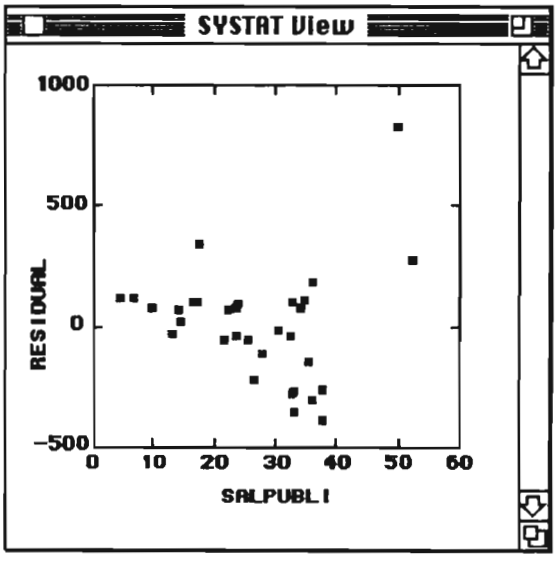


figure n°3.30. SYSTAT: le graphique des résidus par rapport à la variable endogène.

Après avoir ouvert le fichier au suffixe WORK, on peut tracer un graphique bivarié des estimations (ESTIMATE) et des résidus (RESIDUAL) (figure n° 3.29):

Mais l'obtention d'un graphique bivarié des résidus par rapport à la variable endogène est bien plus compliquée. En effet, ces données se trouvent dans deux fichiers différents. Il faut donc les associer dans le module DATA de **SYSTAT**, ce qui nécessite la rédaction d'un petit programme. Ce n'est qu'après l'exécution de ce programme que le logiciel peut tracer le graphique bivarié des résidus (RESIDUAL) par rapport à la variable endogène (SALPUBLI) (figure n° 3.30) .

On notera que la forme des nuages de points des figures n° 3.29 et 3.30 sont identiques. En effet, lorsqu'il s'agit d'une régression simple (avec une seule variable endogène), tracer le graphique bivarié des résidus par rapport aux estimations, ou celui des résidus par rapport à la variable endogène revient au même. De ce fait, la lourdeur d'utilisation de **SYSTAT** apparaît moins gênante.

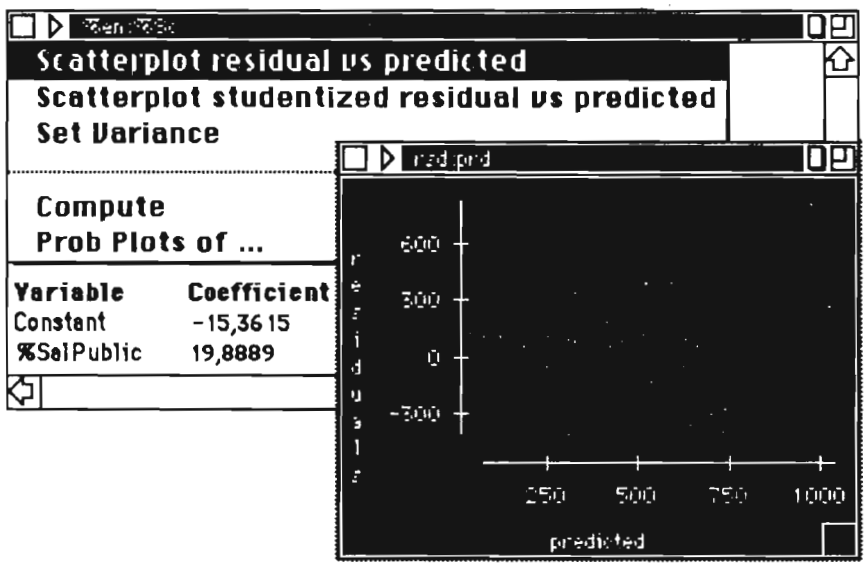


figure n°3.31. DataDesk: Graphique des résidus par rapport aux estimations.

3.2.2.2. DataDesk

Revenons à la fenêtre contenant les résultats des calculs de régression réalisés par **DataDesk** (figure n° 3.21). Elle est aussi dotée d'un menu *hyperview* grâce auquel l'analyse peut-être approfondie.

Par exemple, l'article **SCATTERPLOT RESIDUAL VS PREDICTED** (littéralement graphique bivarié des résidus par rapport aux estimations) permet de rechercher une éventuelle corrélation entre ces deux variables. Le graphique montre, comme avec **SYSTAT** (figure n° 3.31) une tendance des résidus à s'éloigner de la valeur 0 en fonction des estimations.

Le menu *hyperview* de la fenêtre contenant les résultats des calculs de régression propose un second article très pratique. Pour conserver les résidus de la régression, il suffit de cocher l'article **RESIDUAL** pour obtenir, dans le dossier des résultats (RESULTS), l'icône d'une nouvelle variable nommée RESIDUALS. Naturellement, cette nouvelle icône de variable peut être transférée dans le dossier des données et faire l'objet de nouveaux traitements comme n'importe quelle autre variable du fichier de données (figure n° 3.32 et n° 3.33).

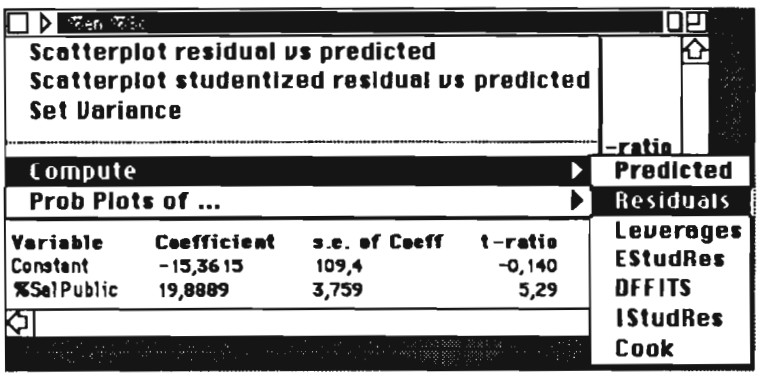


figure n° 3.32. DataDesk: le calcul des résidus.

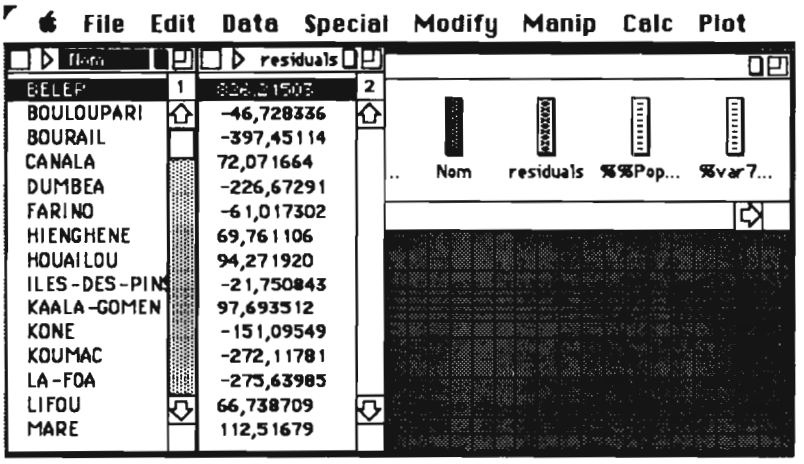


figure n° 3.33. DataDesk: l'enregistrement des résidus dans le tableau de données et l'ouverture de la fenêtre d'édition des variables RESIDUALS et NOMS.

3.2.2.3. JMP

JMP, comme **DataDesk** permet d'enregistrer les estimations et les résidus dans le tableau de données. Le logiciel ajoute ainsi une ou deux nouvelles variables qui portent le nom de la variable exogène précédé des mentions PREDICTED (estimations) et RESIDUALS (résidus) (figure n° 3.34)

A partir de là, il suffit d'utiliser à nouveau la plate-forme **FIT Y BY X** du menu **ANALYZE** pour réaliser le graphique bivarié des résidus avec la variable exogène, ou avec les estimations. Lorsque les trois fenêtres sont affichées à l'écran (celle du tableau de données, celles du graphique bivarié de la variable exogène et de la variable endogène, celle du graphique bivarié de la variable endogène avec les résidus), il est préférable de les disposer de manière à pouvoir les examiner simultanément (figure n° 3.35).

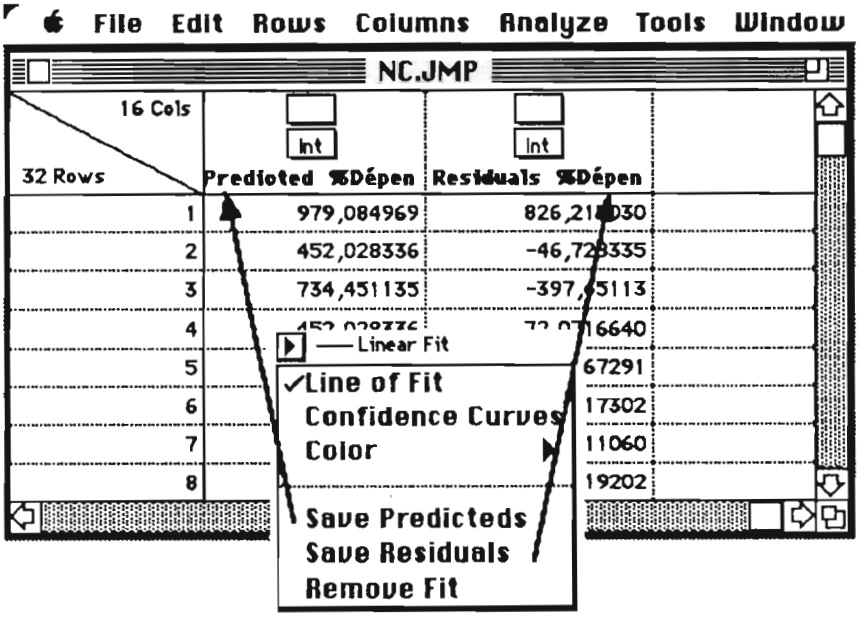


figure n°3.34. JMP: l'enregistrement estimations et des résidus dans le tableau de données.

En sélectionnant l'outil brosse, et en le déplaçant sur le graphique de la régression, on observe que les résidus les plus négatifs, ceux qui «tirent» le graphique bivarié des résidus vers le bas, correspondent aux principaux centres de la côte Ouest de Nouvelle Calédonie, au Nord de Nouméa. Le nombre de personnes pour 100 actifs ayant un emploi (%DEPEN) y est anormalement faible, compte tenu de la proportion de salariés du secteur public dans la population active (%SALPUBLIC).

L'exploration des résidus facilite donc la détection d'un «effet régional» qu'une carte des résidus confirmerait, sans doute à des degrés divers, pour la majorité des communes de la côte Ouest. L'ouvrage de L. Sanders et F. Durand-Dastès, *L'Effet régional* montre tout le parti que l'on peut tirer d'une telle exploration.

3.2.3. Une autre technique d'ajustement linéaire: la droite de Tukey

Du fait qu'elle recourt à des médianes, et non pas à la moyenne et à la variance, la droite de Tukey présente une plus grande résistance que celle des moindres carrés, c'est-à-dire une moins grande sensibilité aux valeurs exceptionnelles.

3.2.3.1. La construction graphique de la droite de Tukey

Voici la description des étapes nécessaires à la construction graphique de la droite de Tukey.

- répartir les observations en trois groupes d'effectifs à peu près égaux, en fonction des valeurs de la variable endogène (X).
- déterminer les médianes du premier groupe sur la variable endogène (X) et sur la variable exogène (Y), et les nommer respectivement MEX1 et MEY1.
- déterminer les médianes du troisième groupe sur la variable endogène (X) et sur la variable exogène (Y), et les nommer respectivement MEX3 et MEY3.
- tracer, sur le graphique bivarié, la ligne droite entre les points de coordonnées (MEX1,MEY1) et (MEX3,MEY3).
- Déplacer la ligne droite, parallèlement à l'axe Y, de manière à ce que la

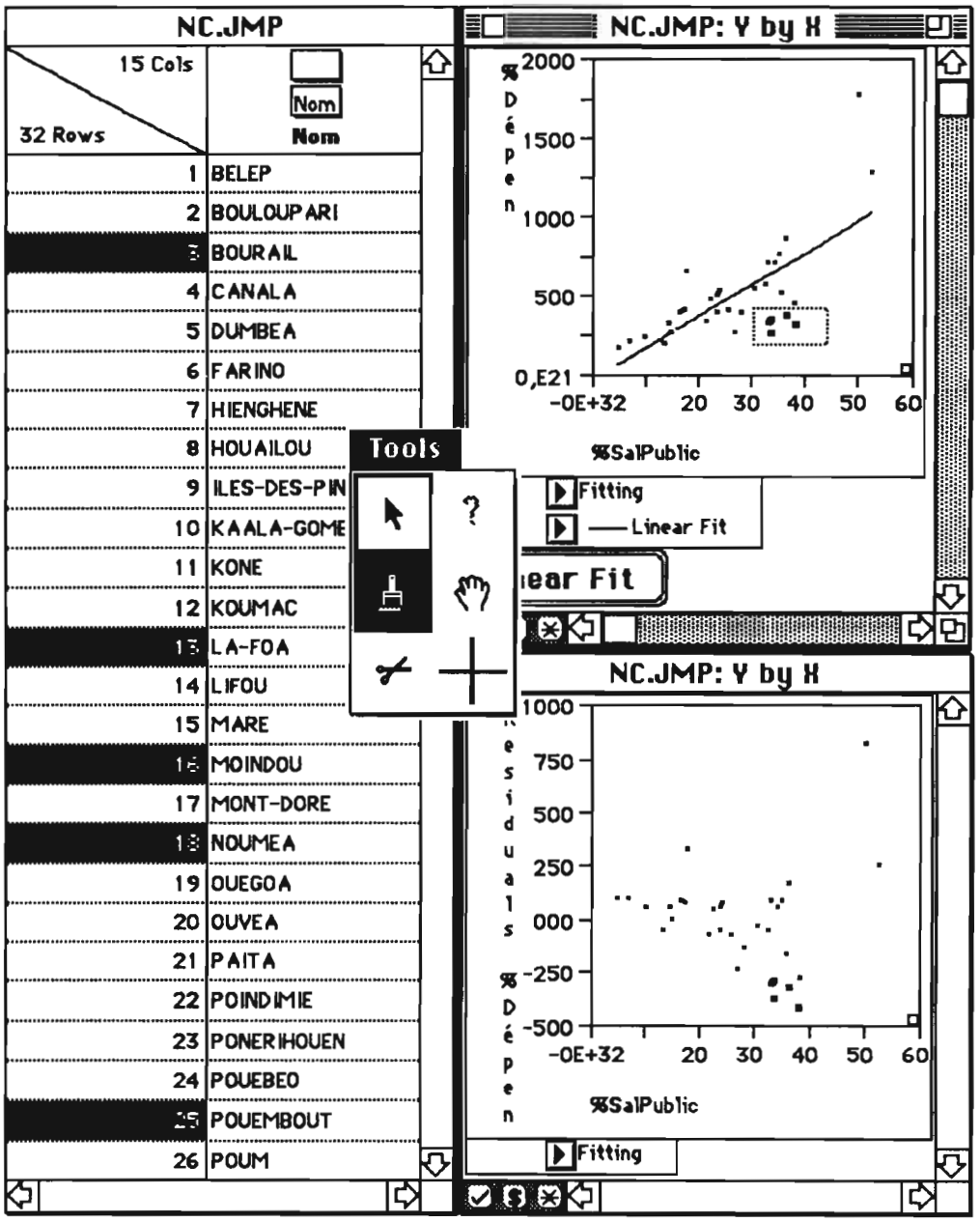


figure n° 3.35. JMP: l'examen simultané de la régression, des résidus et du tableau de données, à l'aide de l'outil brosse.

moitié des observations se trouve au-dessus de la ligne, et la moitié au-dessous: la ligne de Tukey est tracée.

Cette méthode de construction de la droite de Tukey est assez lourde et ne figure en «standard» dans aucun article des logiciels analysés ici. On se limitera donc à montrer comment il faut s'y prendre avec DATADESK pour aboutir au résultat attendu, sachant que les étapes à franchir sont du même type dans les autres logiciels.

3.2.3.1.1. DataDesk

Le tracé d'une droite de Tukey à l'aide de **DataDesk** comprend cinq étapes.

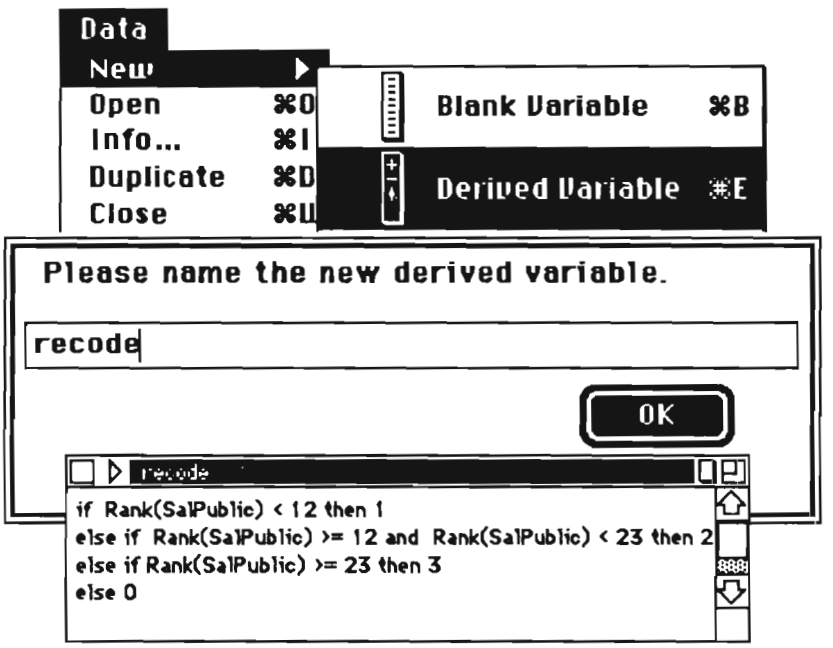


figure n°3.36. DataDesk: la création d'une nouvelle variable RECODE et la rédaction de l'expression de ses valeurs.

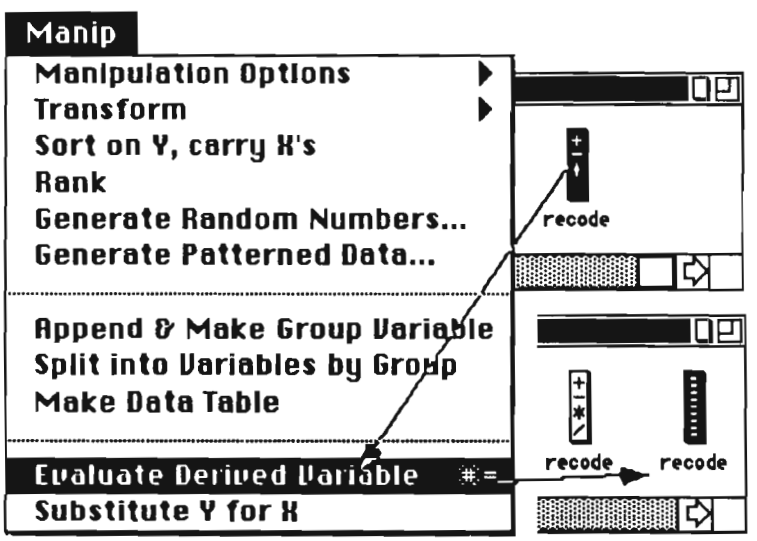


figure n°3.37. DataDesk: l'évaluation de la nouvelle variable RECODE et son rangement sur le bureau de DataDesk.

• Pour répartir les observations en trois groupes d'effectifs à peu près égaux, en fonction des valeurs de la variable endogène (X), on crée une nouvelle variable, nommée RECODE, à l'aide de l'article **DERIVED VARIABLE** du sous-menu **NEW** du menu **DATA**. Pour cette nouvelle variable, le logiciel affiche une fenêtre qui contient la définition (un *script*) du recodage (figure n° 3.36). En voici la traduction en clair:

• si lorsque la variable SALPUBLIC est rangée de la plus petite à la plus grande valeur, le rang d'une observation est inférieur à 12 (32 observations ventilées en 3 groupes donnent à peu près 11 observations par groupe), alors la nouvelle variable RECODE vaut 1 (les observations appartiennent au premier groupe).

• sinon, si le rang d'une observation est compris entre 12 et 23, alors la nouvelle variable RECODE vaut 2 (les observations appartiennent au second groupe).

• sinon, si le rang d'une observation est supérieur ou à 23, alors

la nouvelle variable RECODE vaut 3 (les observations appartiennent au troisième groupe).

• sinon, pour toute autre valeur (donnée manquante, par exemple), la nouvelle variable RECODE vaut 0.

Ainsi définie, la variable RECODE n'a pas vraiment d'existence: son icône, faite de signes de calcul arithmétique (+-*/) se distingue des autres icônes de variables sur le bureau. Pour créer la variable RECODE, il faut activer l'article **EVALUATE DERIVED VARIABLE** du menu **MANIP** (figure n° 3.37). L'icône de la nouvelle variable apparaît alors sur le bureau.

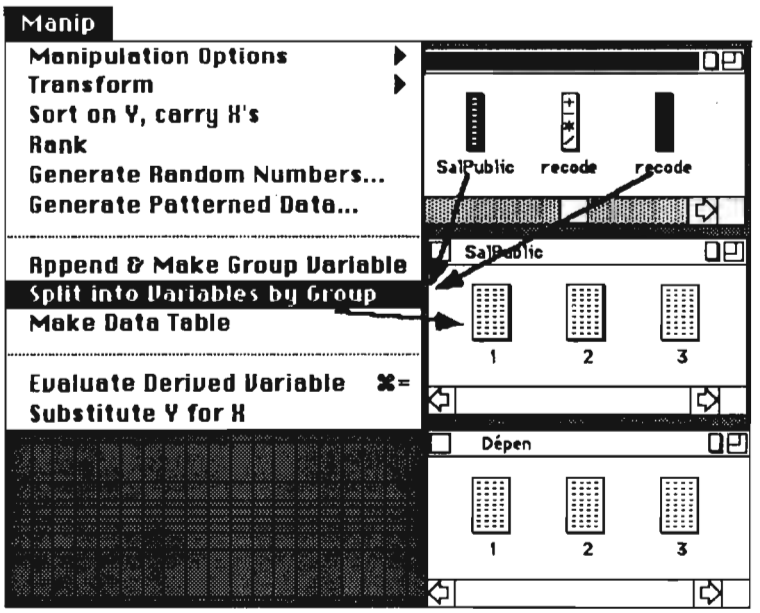


figure n°3.38. DataDesk: l'«éclatement de la variable SALPUBLIC en trois nouveaux tableaux selon les modalités de la variable RECODE.

Pour sélectionner les observations de chaque groupe, **DataDesk** fait «éclater» chaque variable en trois tableaux différents (numérotés 1, 2 et 3, en fonction du numéro de groupe de la variable RECODE) par activation de l'article **SPLIT INTO VARIABLES BY GROUP** du menu **MANIP**. Ces tableaux sont rangés dans un dossier partant le nom de la variable endogène (figure n° 3.38). **DataDesk** fait de même pour la variable exogène (DEPEN).

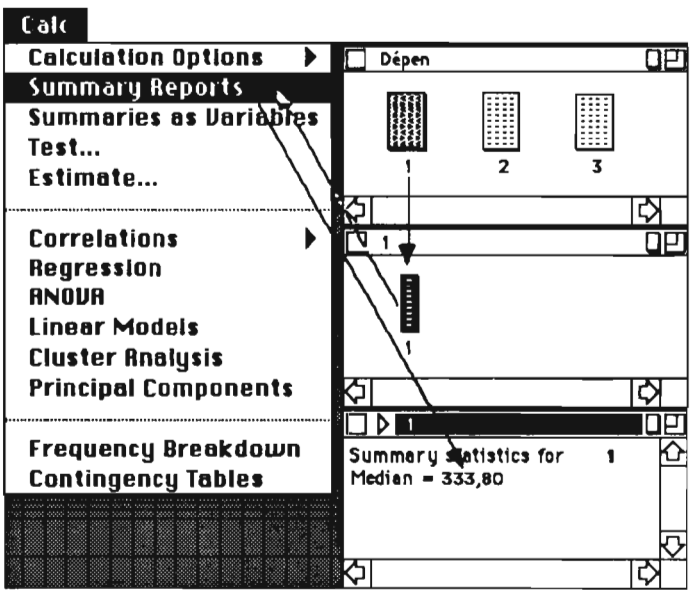


figure n°3.39. DataDesk: le calcul de la médiane du premier groupe de la variable DEPEN.

Dans chaque nouveau tableau (1, 2, 3), il y a une variable qui porte le même numéro que celui du tableau (figure n° 3.39): il s'agit des valeurs de la variable endogène (ou exogène, selon que le dossier ouvert s'appelle SALPUBLIC ou DEPEN) pour les seules observations du groupe dont la variable porte le numéro. En choisissant l'article **SUMMARY REPORTS** du menu **CALC**, la médiane du groupe s'affiche dans une nouvelle fenêtre (celle du premier groupe pour la variable DEPEN sur la figure n° 3.39 soit MEY1 dans la terminologie adoptée ci-dessus).

En réitérant cette opération sur le troisième groupe pour la variable DEPEN, et sur les groupes 1 et 3 pour la variable SALPUBLIC, on obtient les médianes désirées, soit:

- médiane du premier tiers des observations sur SALPUBLIC: **MEX1 = 14.7**
- médiane du troisième tiers des observations sur SALPUBLIC: **MEX3 = 36.15**
- médiane du premier tiers des observations sur DEPEN: **MEY1 = 333.8**
- médiane du troisième tiers des observations sur DEPEN: **MEY3 = 635.7**

Pour faire apparaître ces deux points sur le graphique bivarié, il suffit de saisir les valeurs ci-dessus après avoir ouvert les fenêtres d'édition de SALPUBLIC et de DEPEN. Puis, en choisissant l'article **SCATTERPLOT** du menu **PLOT**, le graphique bivarié s'affiche (figure n° 3.40). Enfin, la droite de Tukey est tracée sur le graphique bivarié, préalablement importé dans un logiciel de dessin. On fait d'abord

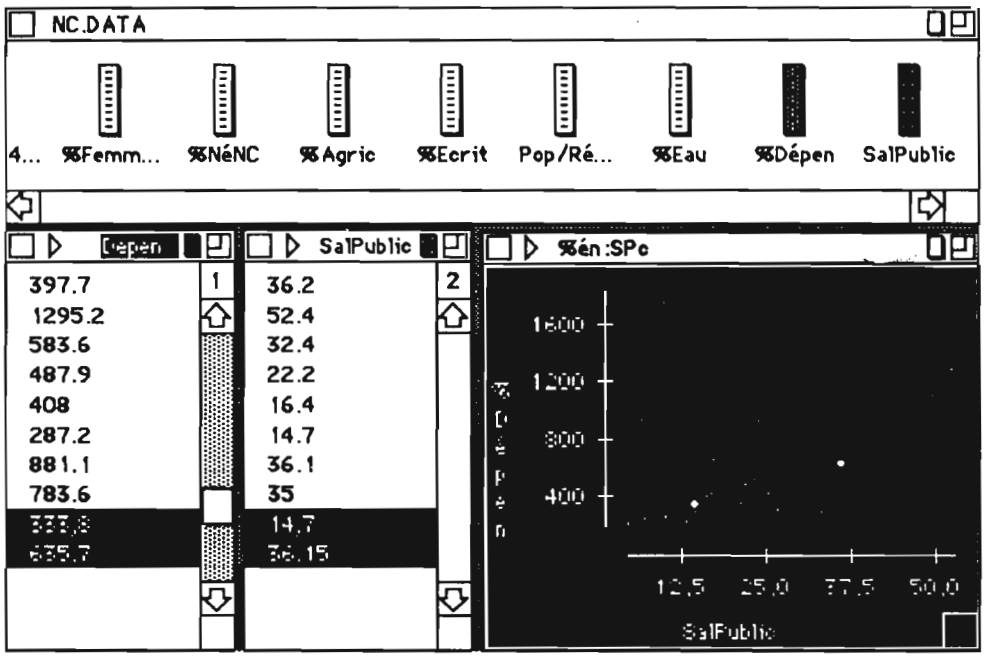


figure n° 3.40. DataDesk: le tracé du graphique bivarié avec deux point supplémentaires correspondant aux médianes des groupes.

passer la droite par les deux points correspondant aux médianes, puis, on la déplace de manière à avoir la moitié des observations au-dessus et au-dessous de la ligne (figure n° 3.41, deux points sont invisibles car situés sur la ligne).

3.2.3.2. La construction arithmétique de la droite de Tukey

Pour gagner du temps et de la précision, il apparaît préférable de recourir à l'estimation des paramètres a et b de la droite, dans l'équation:

Le paramètre a (la pente de la droite) a pour expression:

$$a = \frac{(MEY3-MEY1)}{(MEX3-MEX1)}$$

Le paramètre b (l'ordonnée à l'origine) est la médiane d'une nouvelle variable D dont les valeurs correspondent à l'expression suivante (l'indice i désigne chaque individu):

$$D_i = Y_i - bX_i$$

La construction arithmétique de la droite de Tukey nécessite, dans tous les cas, le calcul des médianes MEX1, MEX3, MEY1 et MEY3. On procédera donc de la même manière que celle exposée ci-dessus pour obtenir ces valeurs. Il faut ensuite évaluer les paramètres a et b. Pour a, le calcul est très simple:

$$a = \frac{(635.7 - 333.8)}{(36.15 - 14.7)}$$

$$a = 14.07$$

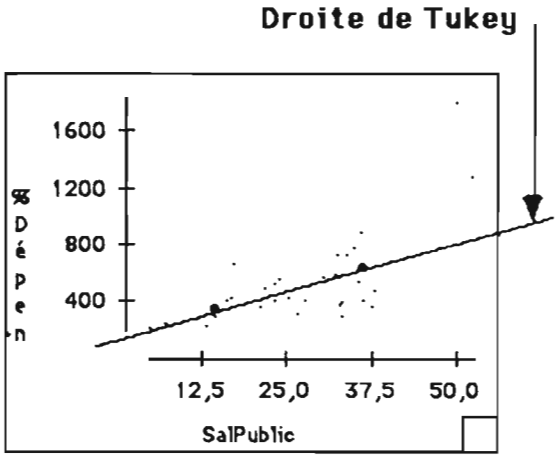


figure n° 3.41. le tracé de la droite de Tukey.

Pour b, il faut obligatoirement passer par le calcul de la nouvelle variable D. Après avoir défini la formule de calcul (figure n° 3.42), les valeurs de D sont obtenues en activant l'article **EVALUATE DERIVED VARIABLE** du menu **MANIP**. Puis, le paramètre b de l'équation de la droite

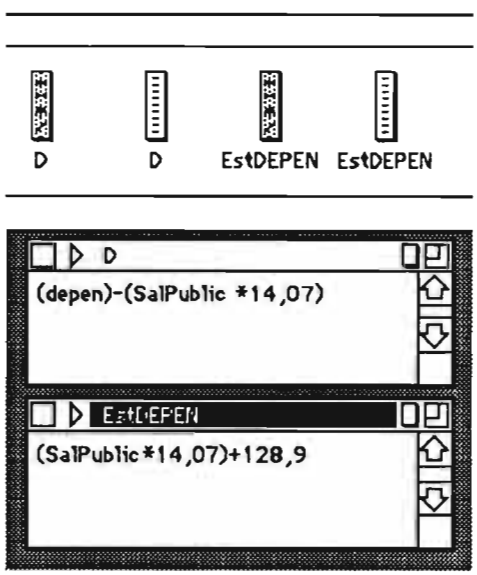


figure n° 3.42. DataDesk: le calcul du paramètre et des estimations du taux de dépendance.

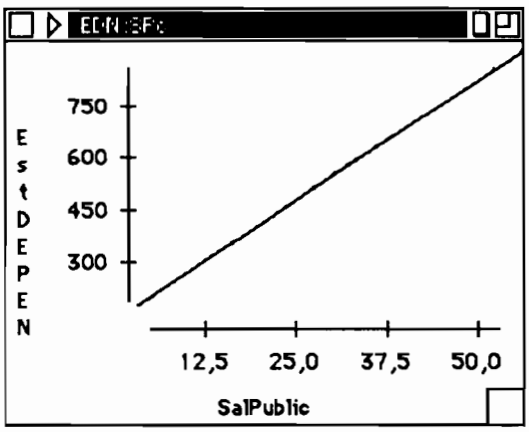


figure n° 3.43. DataDesk: le tracé de la droite de Tukey.

est calculé en choisissant l'article **SUMMARY REPORTS** du menu **CALC**: il s'agit de la médiane de cette variable. On trouve: b = 128.9

Enfin, le calcul des estimations du taux de dépendance passe aussi par la création d'une nouvelle variable (ESTDEPEN) dont la formule est:

ESTDEPEN = (SALPUBLIC * 14.07) + 128.9

La droite de Tukey est obtenue en traçant le graphique bivarié de la variable ESTDEPEN (Y) avec la variable endogène SALPUBLIC (figure n° 3.43, la pente apparaît différente de celle de la figure n° 3.41 car la graduation de l'axe Y n'est pas la même).

De même que l'équation de la droite des moindres carrés, l'équation de la droite de Tukey permet de calculer les résidus par création d'une nouvelle variable (RESDEPEN) en appliquant la formule:

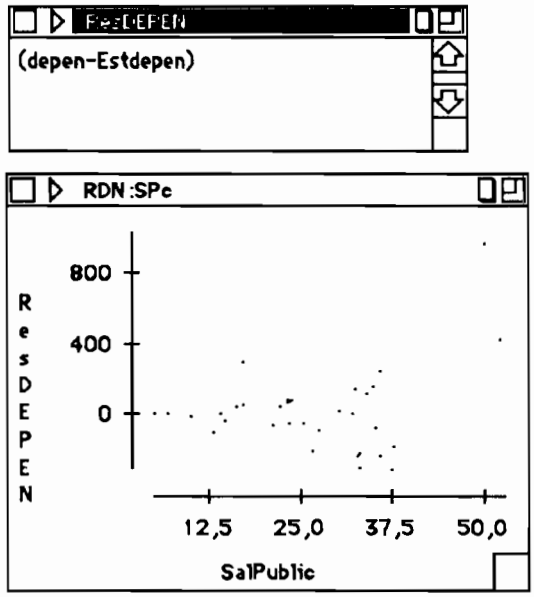


figure n° 3.44. DataDesk: le graphique bivarié de la variable endogène et des résidus obtenus à l'aide de l'équation de la droite de Tukey.

RESDEPEN = DEPEN - ESTDEPEN

On peut ainsi visualiser les résidus (RESDEPEN) par rapport à la variable endogène (SALPUBLIC) en traçant leur graphique bivarié (figure n° 3.44). On remarque que la pente négative des résidus est beaucoup moins accentuée qu'avec la droite des moindres carrés (figure n° 3.27.a).

La droite de Tukey montre qu'un autre critère que celui des moindres carrés peut être choisi pour tracer une droite dans un nuage de points sur un graphique bivarié. Mais, si la ligne droite demeure pratique à utiliser, en raison de la simplicité de son équation, il ne s'agit pas, beaucoup s'en faut, de la seule technique d'ajustement. Beaucoup de relations ne peuvent être bien résumées par une droite. Lorsque le graphique bivarié, ou une régression

non validée conduisent à la conclusion qu'une simple ligne droite ne peut pas rendre compte de manière satisfaisante de la relation entre deux variables, il demeure possible de tenter un ajustement curviligne

3.2.4. Ajustements non-linéaires

A partir du moment où l'on cherche à ajuster autre chose qu'une droite, il existe un très grand nombre de techniques permises par **SYSTAT, DataDesk** et **JMP**.

3.2.4.1. La droite des moindres carrés après transformation des variables.

Lorsqu'on fait appel à une transformation des variables d'une régression linéaire, on réalise un ajustement «intrinsèquement linéaire» car, une fois que les variables X ou Y, ou X et Y ont

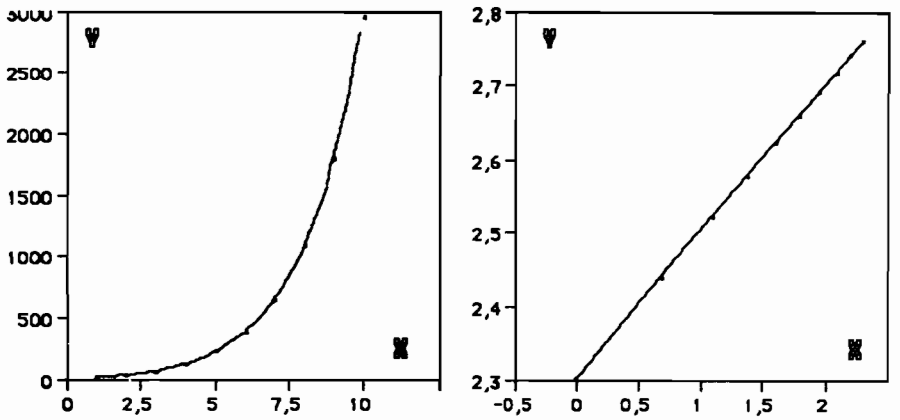


figure n°3.45. La courbe représentative de la fonction Y = e0.5X+3 et la droite représentative de la fonction lnY = 0.5X+3.

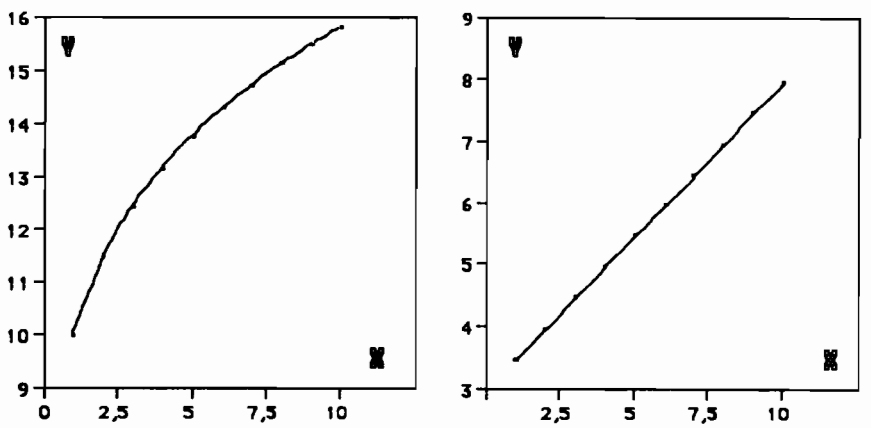


figure n°3.46. La courbe représentative de la fonction Y = 10X0.2 et la droite représentative de la fonction lnY = 0.2ln X+2.3.

été transformées, les calculs sont identiques à ceux d'un ajustement linéaire par la droite des moindres carrés. On peut ainsi ajuster des courbes dont la fonction peut être:

• exponentielle:
$Y = e^{aX+b}$

On applique aux valeurs de la variable exogène (Y) une transformation en logarithmes naturels (de base e). L'équation de régression devient (figure n° 3.45):

$\ln Y = aX+b$

• puissance:
$Y = bX^a$

On applique aux valeurs de la variable exogène (Y) et à celles de la variable endogène une transformation en logarithmes naturels (de base e). L'équation de régression devient:

$\ln Y = a\ln X+\ln b$

On peut imaginer d'autres transformations fonctionnelles, mais les transformations exponentielle et puissance donnent accès à des familles de courbes assez variées pour répondre à de nombreux besoins. Elles ont toutes en commun d'être «linéarisables» et donc «monotones» (de pente positive ou négative sur toute l'étendue des valeurs). Tous les logiciels de statistique étudiés ici proposent une large gamme de fonctions de transformations des données. Leur utilisation est toujours basée sur le même principe: après la création d'une nouvelle variable dans le tableau de données, des valeurs lui sont affectées par l'intermédiaire d'une expression mathématique composée d'un nom de fonction précédant le nom de la variable sur laquelle elle s'applique. Pour réaliser ce traitement, les chemins sont différents dans les trois logiciels.

```
>USE "WANIEZ20:SYSTAT 5.0:NC.SYS"
Variables in SYSTAT RECT file are:
      CODE$       PROVINCE$          NOM$          POPNC         VAR7684
    ANSO14         FEMMES           NENC          AGRIC         SALPUBLI
     ECRIT       POPRESID          DEPEN            EAU
>data
FILE IN USE IS WANIEZ20:SYSTAT 5.0:NC.SYS

>let lndepen=log(depen)
>let lnsalpub=log(salpubli)
>save "WANIEZ20:SYSTAT 5.0:NCTRANS.SYS"
>run

   32 CASES AND   16 VARIABLES PROCESSED.
SYSTAT FILE CREATED.

>systat
>USE "WANIEZ20:SYSTAT 5.0:NCTRANS.SYS"
Variables in SYSTAT RECT file are:
      CODE$       PROVINCE$          NOM$          POPNC         VAR7684
    ANSO14         FEMMES           NENC          AGRIC         SALPUBLI
     ECRIT       POPRESID          DEPEN            EAU         LNDEPEN
  LNSALPUB
>GRAPH
>PLOT  LNDEPEN * LNSALPUB/SMOOTH=LINEAR
```

3.2.4.1.1. SYSTAT

Avec **SYSTAT**, il faut opérer les transformations dans le module DATA. Le programme ci-après donne la liste des instructions nécessaires. Après avoir ouvert le fichier de données, l'instruction DATA donne accès à ce module. Deux instructions LET permettent de réaliser les transformations; leur syntaxe est la suivante:

LET nouvelle variable=**fonction** (ancienne variable)

Ensuite, l'instruction SAVE enregistre les données dans un nouveau fichier qui peut être ouvert et traité par le menu **GRAPH** comme n'importe quel autre fichier. Tout cela n'est pas bien compliqué, mais sans doute un peu lourd et bien peu convivial.

(voir copie d'écran page ci-contre)

3.2.4.1.2. DataDesk

Pour réaliser des transformations avec **DataDesk**, il faut créer de nouvelles variables à l'aide de l'article **DERIVED VARIABLE** du sous-menu **NEW** du menu **DATA**. Pour cette nouvelle variable, le logiciel affiche une fenêtre contenant la définition (un *script*) de la transformation (figure n° 3.47).

DataDesk demande ensuite le nom de la nouvelle variable et ouvre la fenêtre destinée à recevoir son *script*. Puis, en activant l'article **EVALUATE DERIVED VARIABLE** du menu **MANIP**, l'icône de la nouvelle variable fait son apparition sur le bureau. Elle peut être traitée comme n'importe quelle autre variable du fichier.

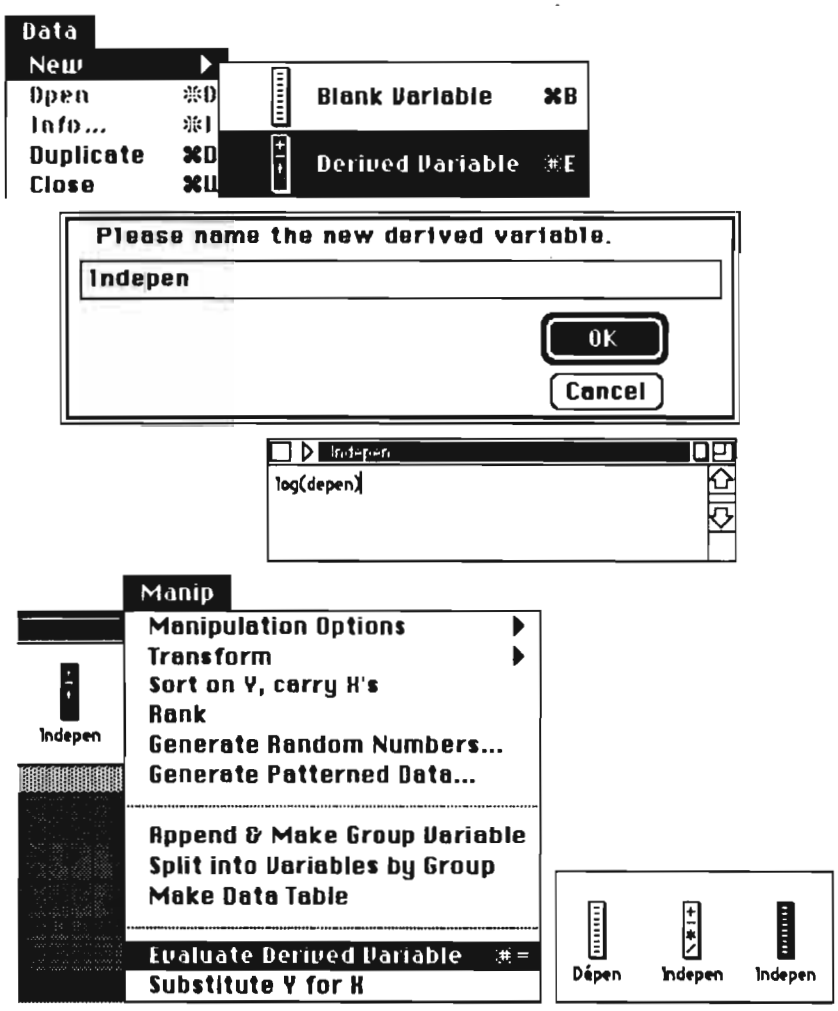


figure n° 3.47. DataDesk: les étapes nécessaires à la transformation de la variable DEPEN en logarithmes.

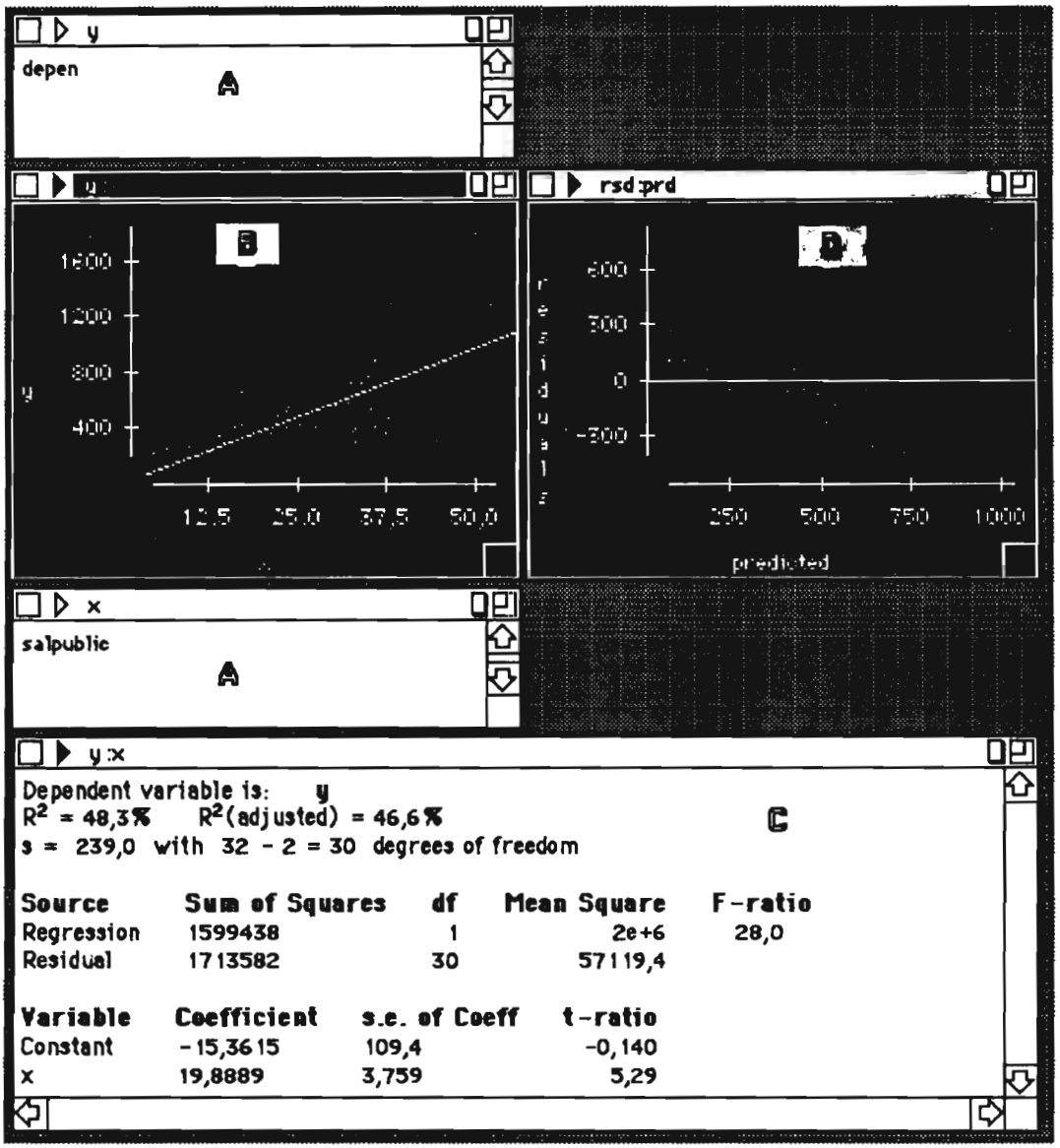


figure n° 3.48. DataDesk: le dispositif de régression pour l'analyse exploratoire des effets des transformations des variables.

Mais **DataDesk** ne s'arrête pas là. Il est possible de mettre en place sur le bureau un dispositif complet de transformation des données, avec visualisation du graphique bivarié des variables exogène et endogène ainsi que de celui des résidus et des estimations et affichage du tableau de régression (figure n° 3.48). Ce dispositif comprend:

- les fenêtres ouvertes de deux variables dérivées (A) X et Y, contenant respectivement les noms des variables endogène (ici SALPUBLIC) et exogène (ici DEPEN);

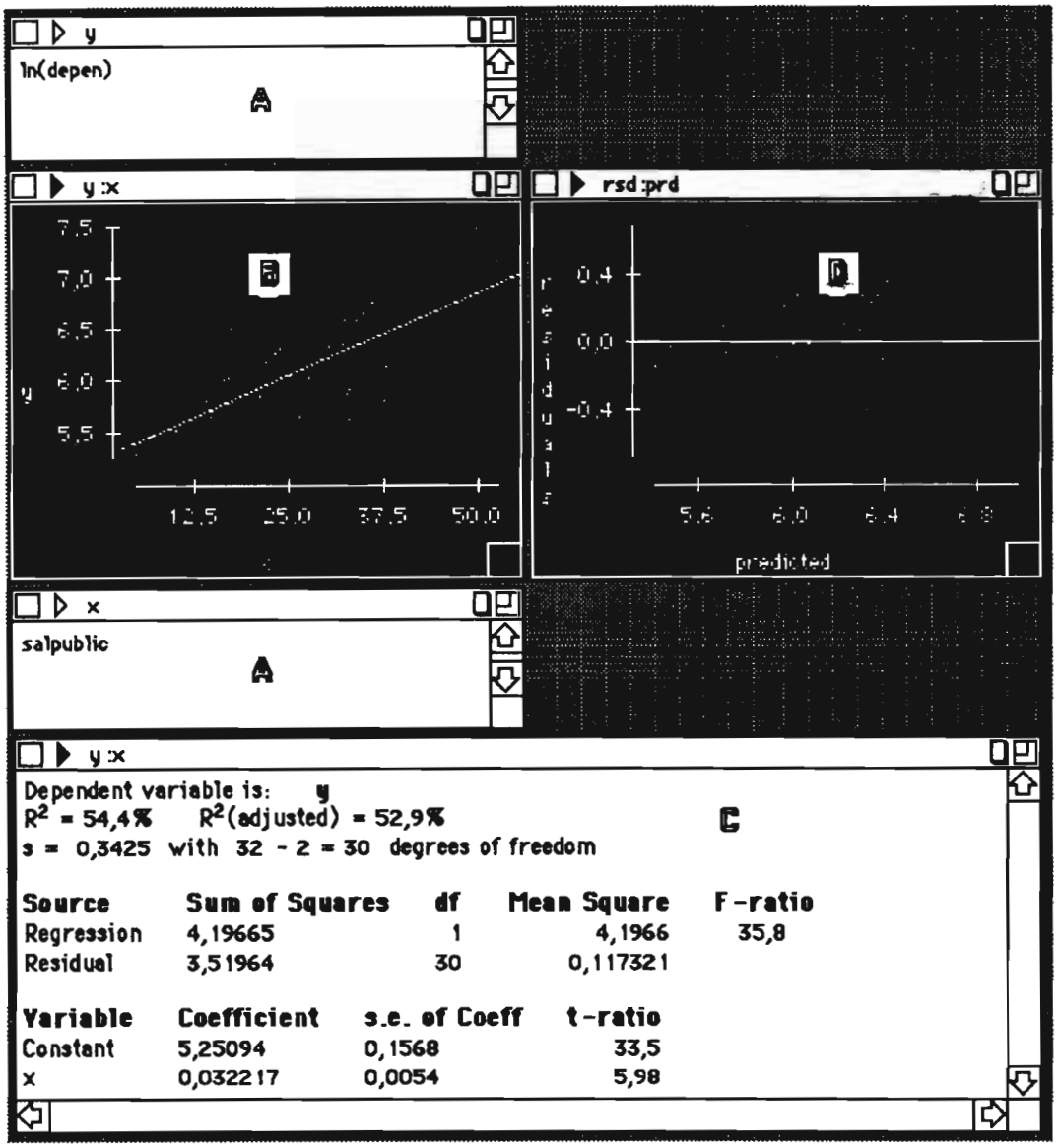


figure n° 3.49. DataDesk: le dispositif de régression pour l'analyse exploratoire après transformation en logarithmes de la variable DEPEN.

• la fenêtre du graphique bivarié des variables Y et X (B);

• le tableau de la régression de Y par rapport à X (C);

• le graphique des résidus par rapport aux estimations (D);

Lorsque toutes les fenêtres B, C et D sont affichées à l'écran, il faut activer l'article **AUTOMATIC UPDATE** (mise à jour automatique) de leurs menus *hyperview* respectifs. Comme toutes les fenêtres sont liées entre elles par des liens dynamiques, toute modification dans l'une ou l'autre des fenêtres de définition des variables X ou Y (A) sera

automatiquement répercutée dans les autres fenêtres: le graphique bivarié de X et Y (B) verra son allure générale changer, la régression sera recalculée (C) et, par voie de conséquence, les résidus se disposeront différemment sur leur graphique.

Par exemple, le simple ajout de la fonction ln devant le nom de la variable Y (ici DEPEN) dans la fenêtre contenant son *script* (A) provoque la modification du contenu de toutes les autres fenêtres B, C et D. Toutes ces opérations ne demandent que quelques secondes pour s'exécuter. Les résultats (figure n° 3.49) montrent que si le coefficient de détermination s'accroît légèrement, passant de 0.47 à 0.53, soit 6% d'explication en plus, les résidus présentent maintenant une forme en U par rapport à la droite de régression: cet ajustement, bien meilleur que le précédent n'est pas encore totalement satisfaisant!

Il apparaît donc indéniable que l'environnement exploratoire de la régression proposé par **DataDesk** constitue un progrès remarquable par rapport aux «habitudes» dans ce domaine. L'examen des effets d'une transformation par rapport à une autre devient un simple jeu d'écriture dont les conséquences, immédiatement perceptibles, doivent être prises en compte pour, le cas échéant, reformuler le modèle.

3.2.4.1.3. JMP

La transformation d'une variable

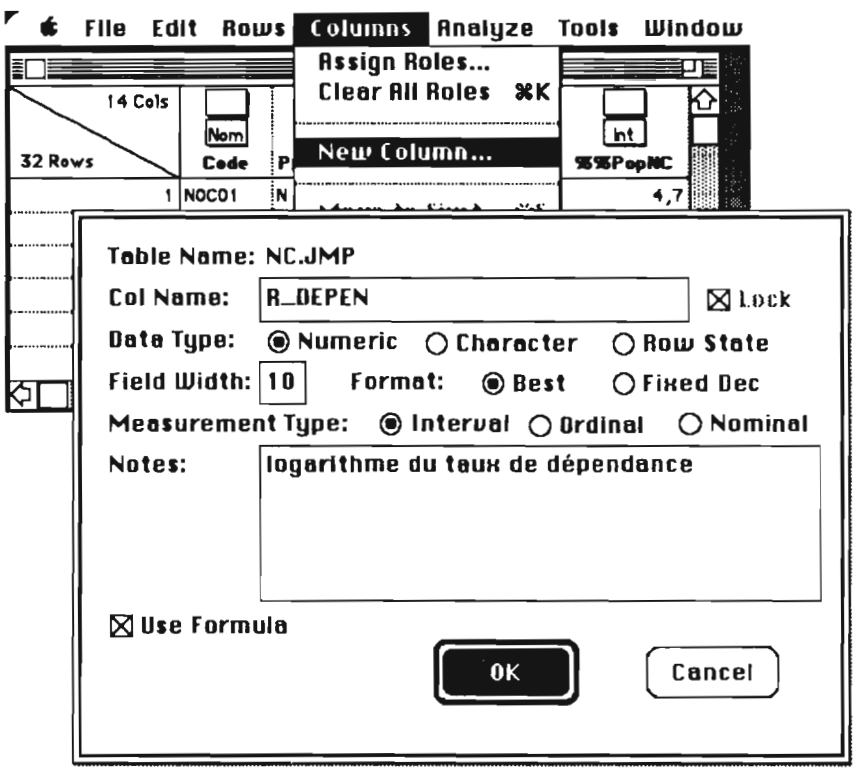


figure n°3.50. JMP: la création d'une nouvelle colonne.

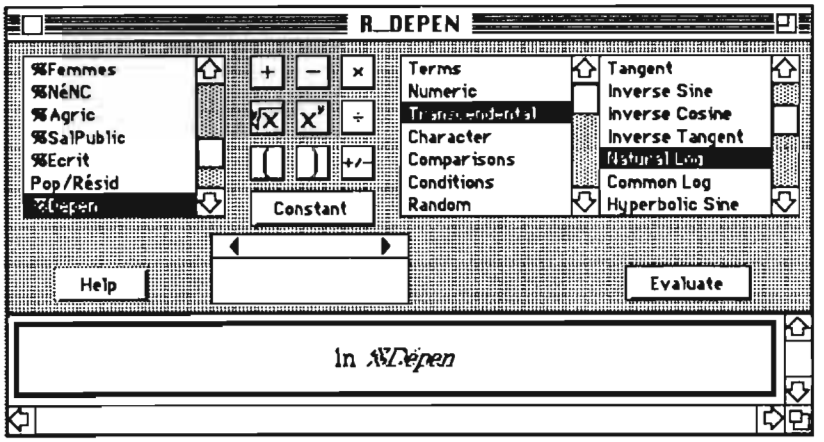


figure n° 3.51. JMP: le Calculator.

avec **JMP** s'effectue en créant une nouvelle variable à l'aide de l'article **NEW COLUMN** du menu **COLUMNS**. Un dialogue (figure n° 3.50) demande de fixer le nom (*Col Name*) et le type (*Data Type*, *Field Width* et *Measurement Type*) de la nouvelle variable, et éventuellement de noter sa signification (*Notes*). Si l'on coche le bouton Use Formula, JMP affiche le *Calculator*, genre de calculatrice très sophistiquée réalisant les transformations souhaitées.

Le *Calculator* comprend trois fenêtres dotées de menus déroulants (figure n° 3.51). A gauche, la fenêtre des variables permet de sélectionner celles qui entrent dans la composition de la formule de transformation. Au centre, la fenêtre des familles de fonctions les rassemble par groupes, de manière à simplifier la sélection d'une des fonctions disponibles; par exemple, les fonctions logarithmes appartiennent à la famille *Transcendental*. Par un clic sur une famille, les noms de fonctions apparaissent dans la fenêtre de droite.

La partie inférieure du *Calculator*, la fenêtre d'édition de formule, sert à rédiger la formule qui s'appliquera à la nouvelle variable. Par exemple, pour calculer le logarithme de la variable %DEPEN, il faut parcourir les étapes suivantes:

- sélectionner, dans la fenêtre centrale, la famille de fonction *Transcendental*;
- choisir, dans la fenêtre de droite, la fonction *Natural Log*. Le mot ln fait son apparition dans la fenêtre d'édition de formule;
- sélectionner la variable %DEPEN dans la fenêtre de gauche: son nom est écrit derrière le mot ln;
- déclencher le calcul par un clic sur le bouton EVALUATE;

Une nouvelle variable, portant le nom donné dans la fenêtre de définition est rangée dans le tableau de données. Elle peut être traitée comme n'importe quelle autre variable de ce tableau, en définissant son rôle et en choisissant une plate-forme d'analyse.

3.2.4.2. Autres types d'ajustements

Les logiciels analysés ici réalisent d'autres types d'ajustements non-linéaires soit par ajustement d'une fonction complexe et souvent non-monotone (de pente positive sur une partie de l'intervalle de variation, puis négative, puis à nouveau positive, etc...), soit par le lissage de données de proche en proche. Dans tous les cas, on cherche à mieux rendre compte de la forme de la relation observée sur le graphique bivarié, par rapport à ce que peut faire une courbe d'équation simple. Mais en rendant le modèle plus complexe, on augmente sans doute la difficulté de son interprétation.

3.2.4.2.1. SYSTAT

Dans **SYSTAT**, le tracé de la droite de régression n'est qu'une possibilité de lissage parmi les autres. On accède aux 10 autres types de lissage (figure n° 3.19), en sélectionnant, comme pour la régression, l'option SMOOTH attachée à l'article **PLOT** du sous-menu **PLOT** du menu **GRAPH**. Les lissages apparais-

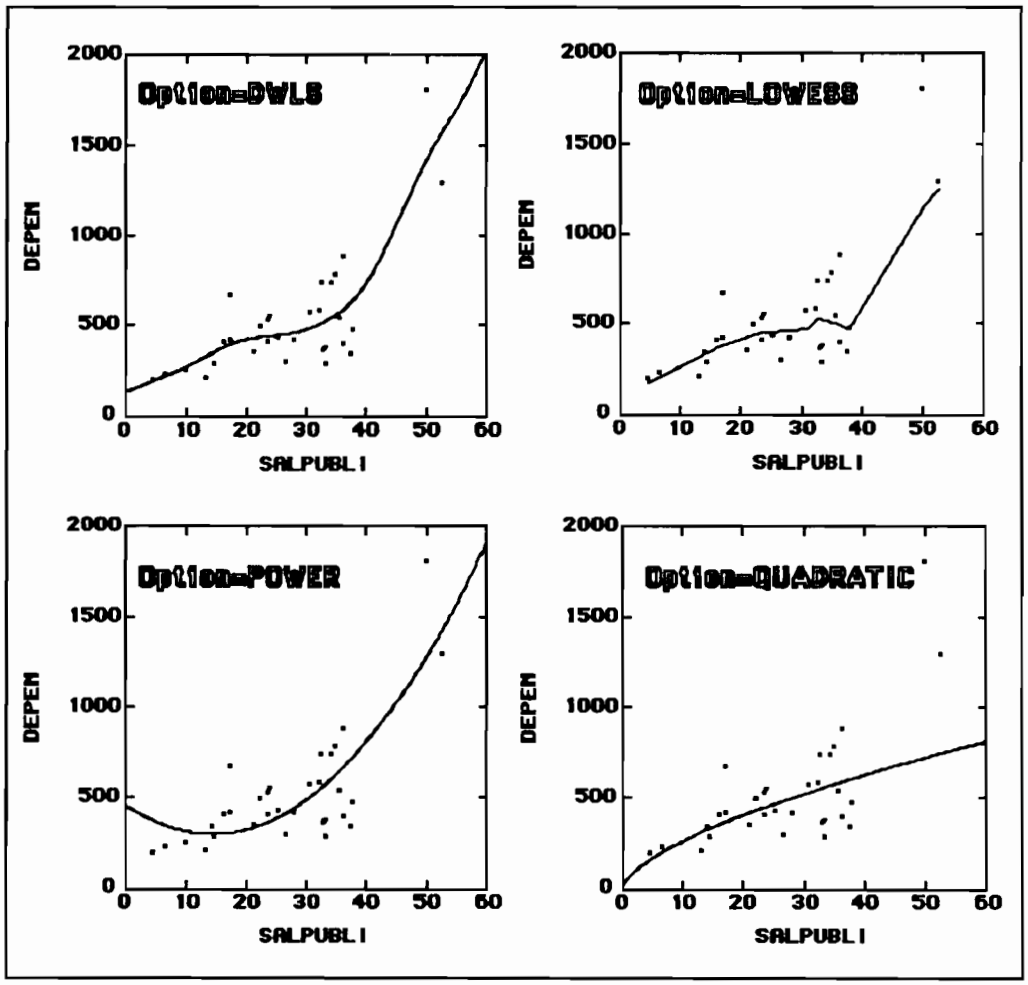


figure n° 3.52. SYSTAT: Quatre exemples de lissages.

sent très différents les uns des autres sur le plan de leur forme régulière ou plus «tourmentée», de leur caractère monotone ou non, du temps nécessaire à leur calcul. Sans doute faut-il procéder à plusieurs essais avant de choisir le meilleur ajustement, ou en tout cas, celui sur lequel on pourra faire un commentaire argumenté (figure n° 3.52).

Pour visualiser plusieurs lissages de type différent sur le même graphique bivarié, il faut, à chaque fois, parcourir l'ensemble des menus. Le langage de programmation permet néanmoins de s'affranchir de cette lourdeur puisqu'il suffit, comme dans le programme ci-après, de modifier le nom de l'option de lissage dans l'instruction PLOT, pour tracer une nouvelle courbe.

```
>USE "WANIEZ20:SYSTAT 5.0:NC.SYS"
Variables in SYSTAT RECT file are:
      CODE$       PROVINCE$          NOM$          POPNC        VAR7684
    ANSO14          FEMMES           NENC          AGRIC       SALPUBLI
     ECRIT        POPRESID          DEPEN            EAU
>GRAPH
>PLOT   DEPEN * SALPUBLI/SMOOTH=DWLS
>PLOT   DEPEN * SALPUBLI/SMOOTH=LOWESS
>PLOT   DEPEN * SALPUBLI/SMOOTH=POWER
>PLOT   DEPEN * SALPUBLI/SMOOTH=QUAD
```

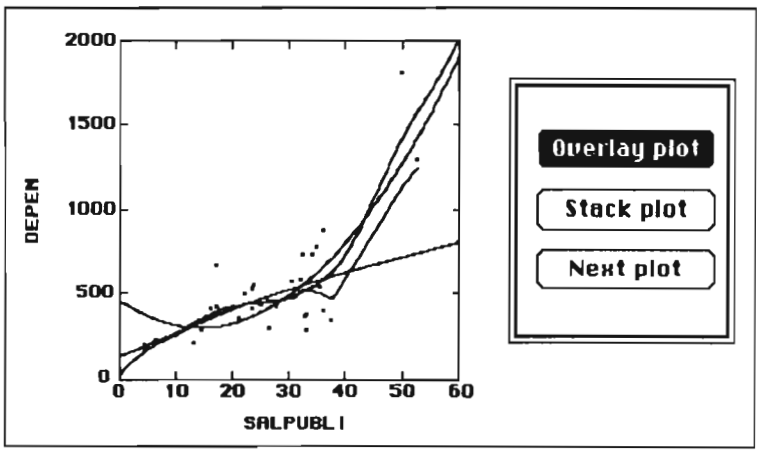


figure n° 3.53. SYSTAT: l'affichage de lissages superposés.

La représentation a lieu sur le même graphique (les courbes sont donc superposées) si l'on choisit le bouton OVERLAY PLOT (figure n° 3.53). On peut regretter que, dans ce cas, chaque courbe ne soit pas étiquetée automatiquement par le nom du lissage qu'elles représente.

3.2.4.2.2. JMP

En plus de la régression linéaire, avec ou sans transformation, **JMP** propose deux autres méthodes: l'ajustement de polynômes d'ordre 2 à 6 et l'ajustement de courbes *Splines*.

L'ajustement polynomial, dit aussi régression polynomiale, est obtenu en appliquant la méthode des moindres carrés à une fonction du type:

$$Y = a_1X + a_2X^2 + ... + a_pX^p + b$$

La plus grande valeur de l'exposant (p) désigne l'ordre du polynôme; l'ordre 1 correspond à la régression linéaire telle qu'elle est décrite plus haut.

Comme pour la régression, c'est par la plateforme **FIT Y BY X** du menu **ANALYZE** que **JMP** réalise la régression polynomiale. Lorsque le graphique bivarié a été tracé à l'écran, il suffit de choisir le menu *pop-up* **FITTING**, de sélec-

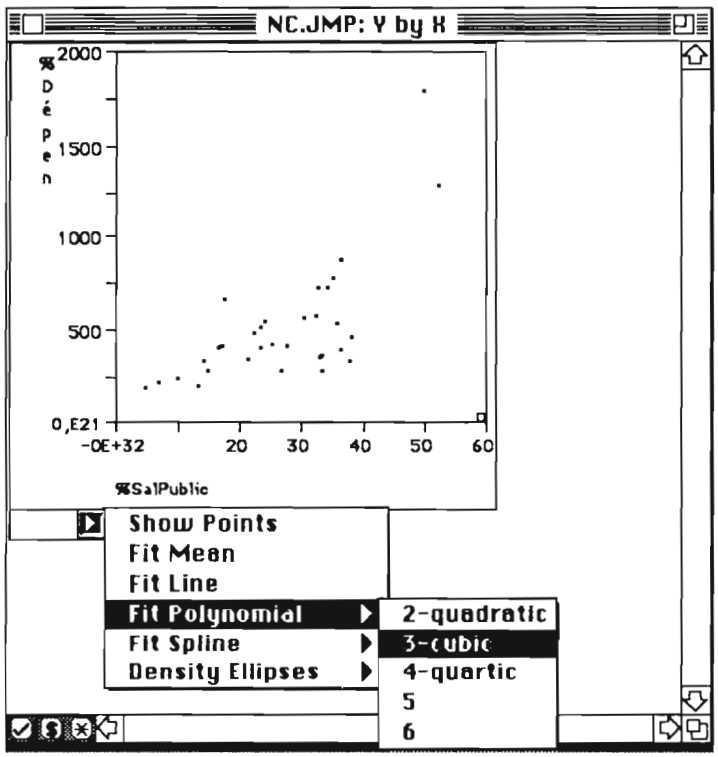


figure n° 3.54. JMP: la sélection d'un ajustement polynomial de degré 3.

tionner l'article **FIT Polynomial** (figure n° 3. 54), et de choisir le degré du polynôme.

Si la courbe obtenue est bien plus complexe que la droite des moindres carrés, la sortie des résultats se présente de la même manière (figure n° 3.55). Le tableau *Summary of Fit* donne le R^2, ici 0.71, soit beaucoup plus que celui obtenu par la droite de régression, 0.48 (figure n° 3.25). Le tableau Parameter Estimates donne les coefficients de l'équation du troisième degré:

%DEPEN= +62.15 %SALPUBLIC
-2.6 %SALPUBLIC^2
+0.038 % SALPUBLIC^3
-94.76

L'ajustement d'un polynôme à la place d'une droite ne dispense pas de l'examen des résidus. Comme précédemment, ceux-ci ne doivent pas présenter de forme particulière pour que la régression puisse être validée. Pour enregistrer les résidus, il suffit de sélectionner l'article **SAVE RESIDUALS**.

JMP propose aussi l'ajustement de courbes *Splines*, dont la forme varie en fonction d'un paramètre Lambda (figure n° 3.56). Plus la valeur de ce paramètre est petite, plus la courbe suit la rugosité des points; plus ce paramètre est grand, plus la courbe ressemble à une ligne droite.

L'approche exploratoire permet de faire varier les valeurs de Lambda pour

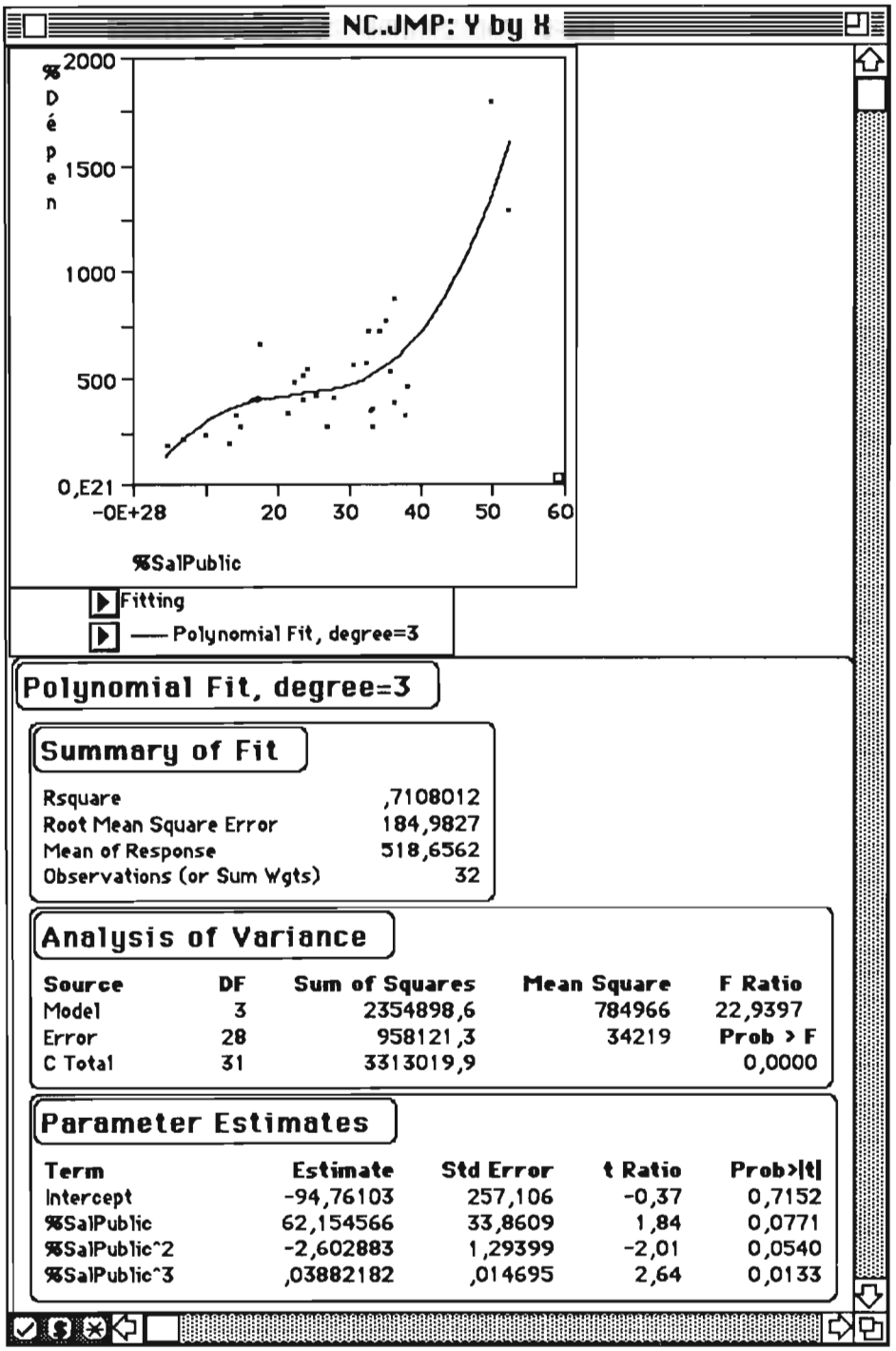


figure n° 3.55. JMP: l'ajustement, par un polynôme d'ordre 3, du taux de dépendance (%DEPEN) par rapport à la part des salariés du secteur public dans la population totale (%SALPUBLIC).

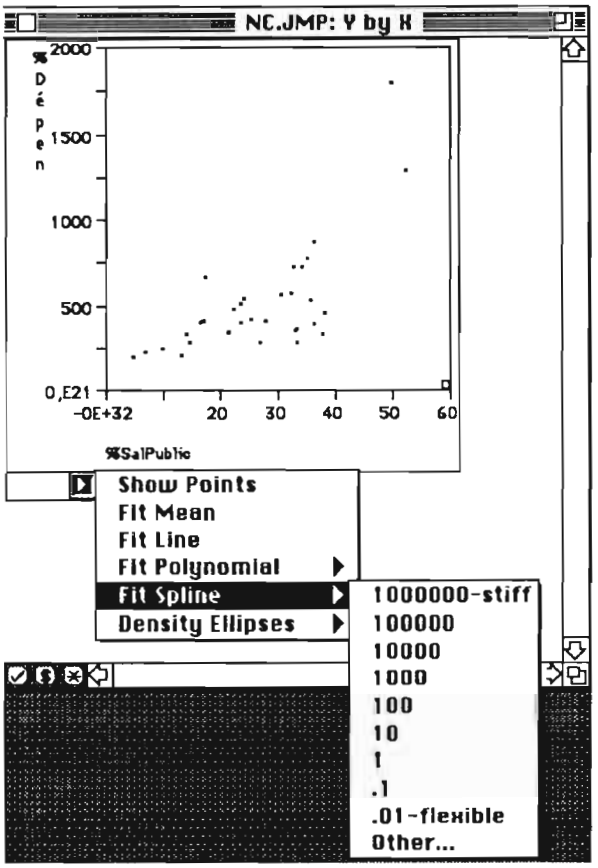


figure n° 3.56. JMP: les différentes valeurs du paramètre Lambda des courbes Splines.

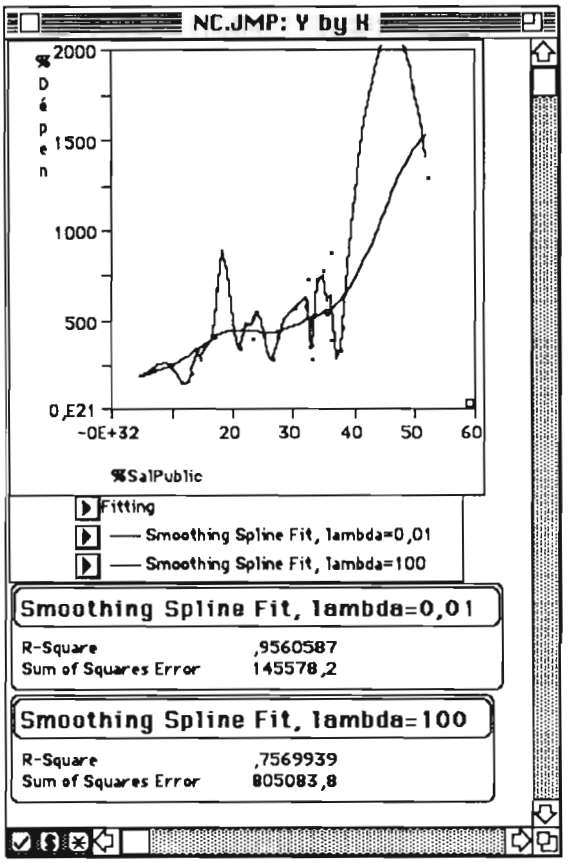


figure n° 3.57. JMP: l'ajustement de deux courbes Splines.

en examiner les effets (figure n° 3.57). On observe qu'un Lambda très faible donne un R^2 proche de 1, mais que la courbe ne correspond à rien pour les valeurs exceptionnelles. Avec une valeur plus élevée, l'ajustement ressemble à celui d'un polynôme d'ordre 3, avec un R^2 légèrement supérieur, 0.75. L'examen des résidus demeure possible après les avoir enregistrés avec l'article **SAVE RESIDUALS**.

EXPLORATIONS MULTIVARIÉES

L'analyse statistique multivariée s'applique à des ensembles de données dont les individus ont été mesurés sous divers angles complémentaires. Ainsi, le nombre de variables à prendre en compte simultanément dépasse souvent plusieurs dizaines, et les méthodes exposées au chapitre 3 s'avèrent insuffisantes. Il faut donc disposer de techniques capables d'exprimer l'information véhiculée par ces multiples variables en termes de liaisons, de ressemblances ou de dissemblances.

L'une des branches de la statistique, connue sous le nom d'«Analyse des données» s'attache à développer un ensemble de techniques descriptives multivariées dont le principe général revient à «plonger individus et variables dans des espaces géométriques tout en faisant la plus grande économie d'hypothèses, et à transformer les données pour les visualiser dans un plan, ou les classer en groupes homogènes, et ceci tout en perdant le minimum d'information» (J.M. Bouroche et G. Saporta).

L'Analyse des données se compose de deux ensembles de techniques complémentaires où chaque observation représente un point dans un espace géométrique dont les variables sont les dimensions: avec une variable, on a affaire à une ligne, avec 2 variables, un plan, avec 3 variables, un volume, 4 variables et plus, un hyper-volume.

Les méthodes factorielles visent à produire un résumé de l'information par projection du nuage de points multidimensionnel sur un sous-espace formé par les axes principaux d'allongement de ce nuage. Ces axes, ou facteurs, rendent compte des associations entre variables et de ce fait, leur nombre apparaît bien plus réduit que celui des variables d'origine. Selon le critère choisi pour calculer les distances destinées à exprimer l'allongement du nuage, on procède à une Analyse en Composantes Principales (ACP), ou bien à une Analyse Factorielle des Correspondances (AFC).

Les méthodes de classification automatique servent à constituer des

groupes plus ou moins homogènes. Elles rassemblent dans une même classe les observations proches les unes des autres, dans le nuage de points multidimensionnel. Les diverses techniques de classification automatique diffèrent d'après le critère retenu pour apprécier les ressemblances entre observations et, surtout, en fonction du mode de progression de l'algorithme de constitution des classes. La progression descendante conduit à la segmentation successive de l'ensemble des individus, ou des classes d'individus, en fonction de différentes combinaisons de propriétés. Avec la progression ascendante, les classes sont formées par agrégations successives d'individus ou des classes d'individus, jusqu'à la classe ultime, la plus hétérogène, celle qui rassemble tous les individus en une seule classe. Selon le cas, on a affaire à une hiérarchie de classes emboîtées (Classification Ascendante Hiérarchique, CAH), ou à des classes disjointes deux à deux (Nuées Dynamiques).

L'Analyse des Données connaît un très grand succès, depuis une vingtaine d'années, dans tous les domaines où l'observation de phénomènes complexes impose l'étude de grands tableaux de données. Avec la montée en puissance des ordinateurs et la diffusion, de plus en plus large, de logiciels faciles à employer, les diverses méthodes, ou variantes de méthodes de cette statistique descriptive multidimensionnelle a sans doute contribué de manière décisive, au progrès des sciences naturelles et biologiques, et même des sciences humaines.

L'exploration statistique ne doit pas représenter, ici, pour le multivarié, comme ailleurs avec l'uni- ou le bivarié, une alternative à l'Analyse des Données. L'exploration multivariée propose des outils permettant de prendre contact avec des données complexes en les regardant sous divers angles. Cette approche *ante* Analyse des Données peut se compléter d'une approche *post* Analyse des Données permettant une amélioration de la lecture des résultats (facteurs ou classes) provenant de cette dernière, et, ainsi, un approfondissement incitant à dépasser la simple mais stérile description des sorties des programmes informatiques. Il n'y a donc pas «concurrence déloyale» entre l'approche exploratoire, qui ne demande pas de connaissance mathématique particulière, et l'Analyse des Données, plus «mathématisée». Sans renoncer à la reproductibilité des techniques «automatiques», l'approche exploratoire invite à refuser de se laisser enfermer dans des «boîtes noires».

Comme les méthodes d'Analyse des Données ont fait l'objet d'un très grand nombre d'ouvrages, on se limitera ici à l'exposé des méthodes exploratoires.

4.1. Les principes de l'exploration multivariée

En mettant au point leur logiciel PRIM-9, en 1972, à l'Université de Standford, M.A. Fishkeller, J.H. Friedman et J.W. Tukey ont mis en pratique les principes de l'exploration multivariée telle qu'ils la proposaient. En effet, PRIM est formé par les initiales

des 4 opérations de base grâce auxquelles l'exploration d'un nuage de points multidimensionnel devient une réalité.

4.1.1. Quatre principes pour une méthode

• P pour Projection.

Dans le monde réel, les objets sont observés en perspective: un même objet apparaît d'autant plus petit qu'il est éloigné de l'observateur. De plus, la combinaison par le cerveau des images transmises par les deux yeux permet de rendre aux objets leur relief. Malheureusement, les nuages de points multidimensionnels auxquels font appel les statisticiens pour analyser leurs données n'ont pas d'existence matérielle. Il faut donc recourir, comme le font les différentes méthodes d'analyse factorielle, à la projection des points de l'espace multidimensionnel sur un plan.

• R pour Rotation.

La rotation permet de créer l'illusion de la troisième dimension. En regardant le nuage de points sous divers angles on cherche à identifier des organisations particulières. Cette reconnaissance des formes du nuage de points ouvre la voie de l'interprétation des données statistiques.

• I pour Isoler.

Isoler un ensemble de points pour mieux les observer revient à s'interroger sur l'existence de groupes présentant des caractéristiques particulières. L'isolement consiste, d'une part, à étudier le groupe pour lui-même, en définissant un sous-ensemble d'observations devant être analysé à part, et d'autre part, à examiner ce groupe par rapport aux autres observations (ou aux autres groupes), en les marquant par un signe ou une couleur particulière.

• M pour Masquer.

En masquant certaines parties du nuage de points, en fonction de critères qui n'ont pas contribué directement à sa construction, on cherche à discriminer les observations *a priori*. Ainsi, il est possible de faire des hypothèses sur le rôle joué par telle ou telle autre caractéristique.

4.1.2. La reconnaissance des formes

Les auteurs du logiciel PRIM-9 ont découvert, de manière empirique, que les projections des nuages de points présentaient des formes récurrentes qu'il faut s'efforcer de reconnaître. En adoptant une échelle d'intérêt de ces formes, et en les classant de la moins intéressante à la plus intéressante, on peut distinguer:

• les nuages de points en forme de disque ou d'ellipse peu allongée correspondant à une distribution normale (figure n° 4.1.A). Très importantes en statistique inférentielle, car elles correspondent à certaines conditions d'échantillonnage devant être respectées pour que la généralisation de l'échantillon à toute la population soit valide, les distributions normales sont les moins intéressantes en analyse exploratoire.

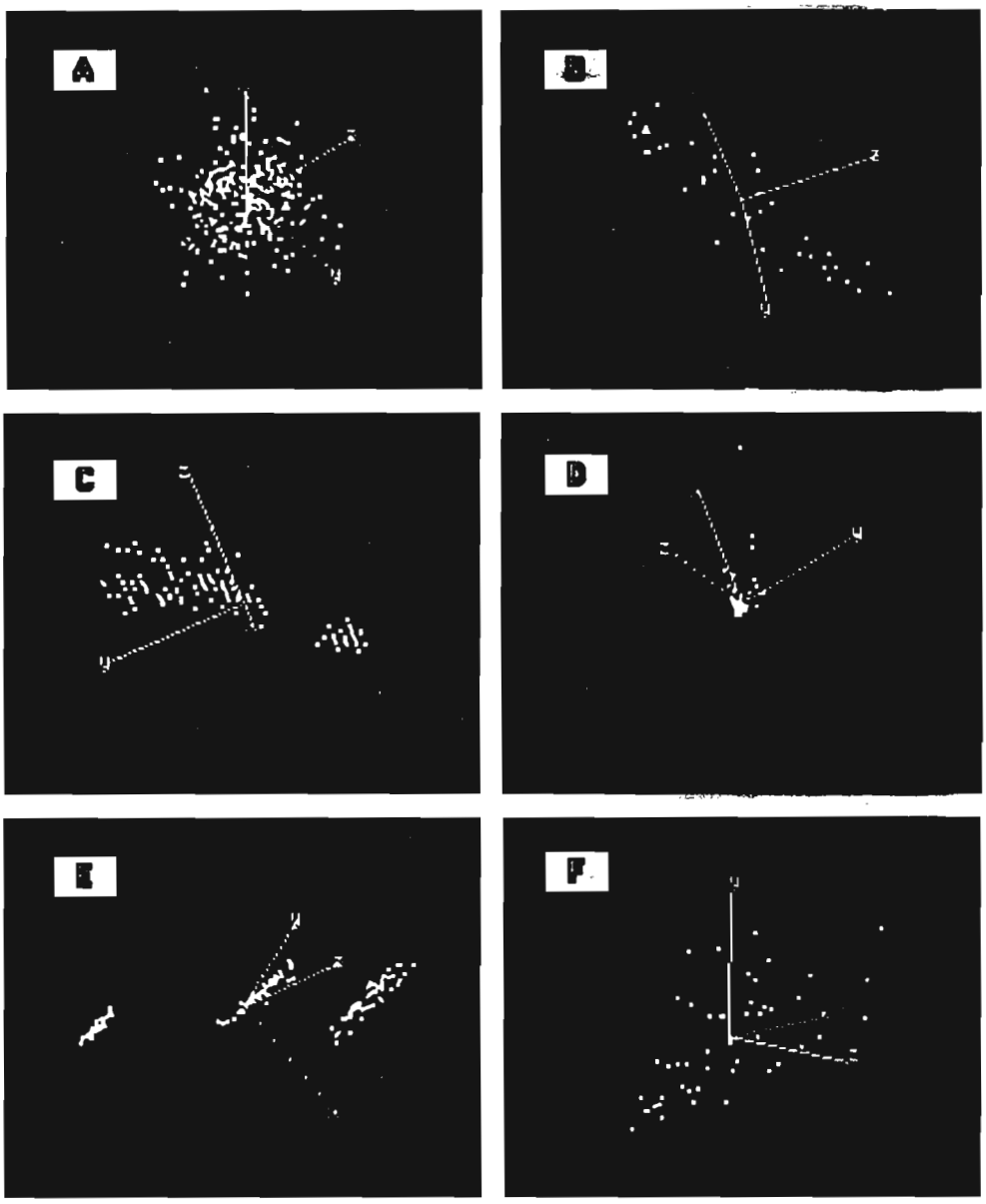


figure n°4.1. Six formes «significatives» de nuages de points.

• les alignements de points sont d'un plus grand intérêt (figure n° 4.1.B). En effet, ils expriment l'existence de tendances, de relations entre les variables, comme il en a été question au précédent chapitre; la différence vient ici de la caractéristique multidimensionnelle du nuage.

• les groupes de points séparés nettement les uns des autres traduisent l'existence de populations différentes au sein du même tableau de données (figure n° 4.1.C). Dans un tel cas de figure, il apparaît souvent préférable d'isoler chaque groupe d'individus pour les examiner séparément, ce qui

facilite l'examen postérieur des différences entre les différents groupes.

• les surfaces minces traduisent l'existence de combinaisons de variables qui interagissent sur une autre variable. Cette configuration correspond à la régression multiple de la statistique «classique».

• les observations exceptionnelles, qui n'entrent pas dans les formes décrites ci-dessus doivent toujours faire l'objet d'un examen particulier (figure n° 4.1.D). Il peut s'agir d'erreurs de saisie, mais si cela n'est pas le cas, on doit s'interroger sur ces exceptions, les masquer, pour éventuellement les réintroduire plus tard dans l'analyse.

• enfin, d'autres formes plus complexes apparaissent quelquefois. Il s'agit du bâton (figure n° 4.1.E), de l'«aile d'oiseau» (figure n° 4.1.F) ou du «lapin à oreille molle».

4.2. Du bivarié au multivarié: les matrices de graphiques bivariés

La première étape d'une analyse factorielle consiste à calculer les distances entre les variables. Celles-ci sont enregistrées dans un tableau carré (matrice de coefficients de corrélation, matrice de distances du χ^2) ayant autant de lignes et de colonnes qu'il y a de variables (figure n° 4.2).

Dans chaque case située à l'intersection d'une ligne et d'une colonne de cette matrice, on trouve une mesure de la distance, ou de la ressemblance entre

	V1	V2	V3
V1	d(V1,V1)	d(V1,V2)	d(V1,V3)
V2	d(V2,V1)	d(V2,V2)	d(V2,V3)
V3	d(V3,V1)	d(V3,V2)	d(V3,V3)

figure n° 4.2. La matrice carrée des distances entre 3 variables.

les variables correspondantes: la valeur d(V1,V2), de la case de coordonnées (V1,V2), représente la distance entre les variables V1 et V2. Le plus souvent, les valeurs de la matrice sont symétriques par rapport à la diagonale, ce qui s'exprime par d(V2,V3)=d(V3,V2). Grâce à cette propriété, la matrice peut être allégée en ne représentant que la diagonale et sa partie supérieure ou inférieure (figure n° 4.3).

Lorsque la matrice représente des distances, les cases situées dans la diagonale prennent une valeur exprimant une distance nulle. Par contre, s'il s'agit d'une matrice de ressemblance, les cases situées dans la diagonale prennent une valeur traduisant la parfaite similitude.

Il existe une certaine analogie entre la matrice de distances de l'Analyse des Données, et la matrice de graphiques

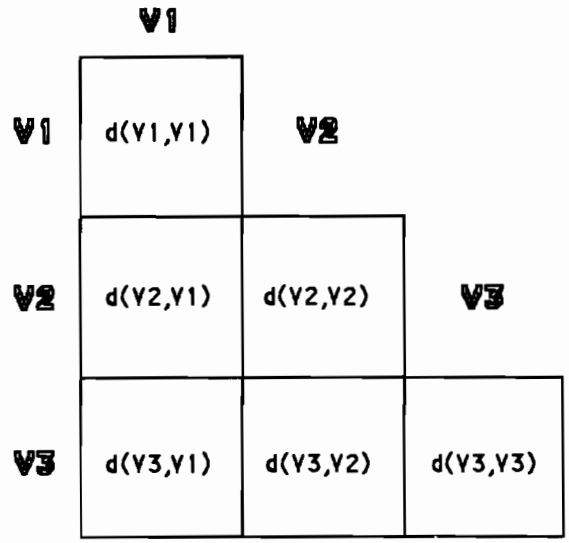


figure n° 4.3. La partie inférieure et la diagonale de la matrice carrée des distances entre 3 variables.

bivariés (*Scatterplot Matrix*) de l'Analyse Exploratoire. Une seule différence sensible: au lieu d'apprécier les distances par des valeurs numériques, on se base sur les formes des graphiques bivariés.

Plus l'ensemble des points figurant dans chaque case de la matrice de graphiques bivariés apparaît allongé (dans la terminologie exploratoire, plus il est lisse), plus les variables auxquelles se rapportent ces graphiques sont en relation.

La figure n° 4.4 représente les relations entre la part des salariés du secteur public dans l'ensemble des actifs (%SALPUBLIC), le nombre de personnes par résidence principale (POP/RESID) et le nombre de personnes pour 1 000 actifs ayant un emploi (%DEPEN).

On observe qu'il existe une relation positive assez forte entre la première variable (%SALPUBLIC) et la troisième (%DEPEN), et négative entre la première (%SALPUBLIC) et la seconde (POP/RESID), cette autre relation étant visiblement beaucoup moins intense. Par contre, entre les seconde et troisième variables, aucune relation ne semble exister.

Ces observations sont confirmées par la matrice des coefficients de corrélation linéaire, qui ne sont autres que la racine carrée du R^2 décrit au chapitre précédent, doté d'un signe précisant le sens de la relation. On notera que cette matrice contient des 1 dans la diagonale signifiant la parfaite corrélation linéaire d'une variable avec elle-même.

Une telle matrice présente les inconvénients de son seul avantage: elle fournit un résumé numérique des relations très pratique, mais est incapable de suggérer l'existence d'une relation non-linéaire (ce qui est quand même le cas de %SALPUBLIC avec %DEPEN). De plus, elle ne rend pas compte des valeurs exceptionnelles alors même que le coefficient de corrélation linéaire est extrêmement sensible à de telles valeurs.

En faisant une lecture conjointe de la matrice des graphiques bivariés et de celle des coefficients de corrélation linéaire, on évitera très certainement des erreurs courantes, et encore trop fréquentes, dans l'utilisation des matrices de corrélation. Par voie de conséquence, l'interprétation des résultats de l'Analyse en Composante

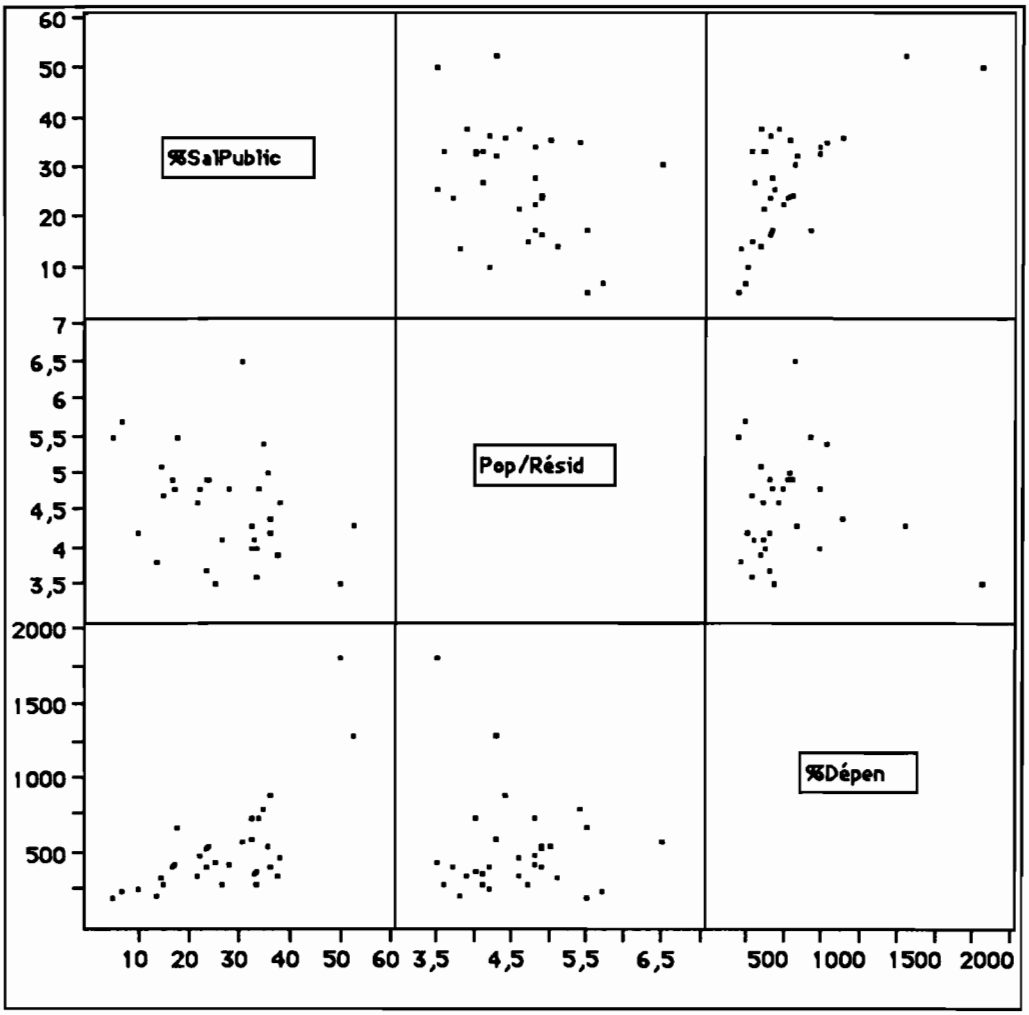


figure n°4.4. La matrice de graphiques bivariés de la part des salariés du secteur public dans l'ensemble des actifs (%SALPUBLIC), du nombre de personnes par résidence principale (POP/RESID) et du nombre de personnes pour 1000 actifs ayant un emploi (%DEPEN).

Variable	%SalPublic	Pop/Résid	%Dépen
%SalPublic	1,0000	-0,3645	0,6948
Pop/Résid	-0,3645	1,0000	-0,1297
%Dépen	0,6948	-0,1297	1,0000

figure n°4.5. La matrice des coefficients de corrélation linéaire des variables de la figure n°4.4 ci-dessus.

Principales (ACP), qui utilise la matrice des coefficients de corrélation linéaire, apparaîtra plus aisée et plus sûre.

4.2.1. SYSTAT

Pour réaliser une matrice de graphiques bivariés avec **SYSTAT**, il faut ouvrir le fichier de données puis sélectionner l'article **SPLOM** du sous-menu **SPLOM** du menu **PLOT**.(figure n° 4.6).

Le logiciel ouvre ensuite un dialogue permettant de sélectionner les variables à représenter (figure n° 4.7). Notons la présence, dans la partie inférieure droite de ce dialogue, de la

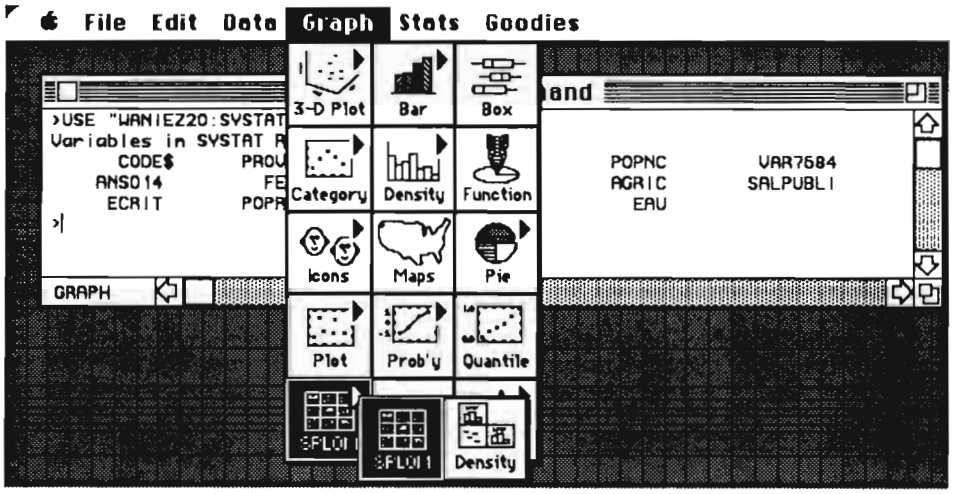


figure n°4.6. SYSTAT: la sélection de l'article SPLOM du sous-menu SPLOM du menu PLOT.

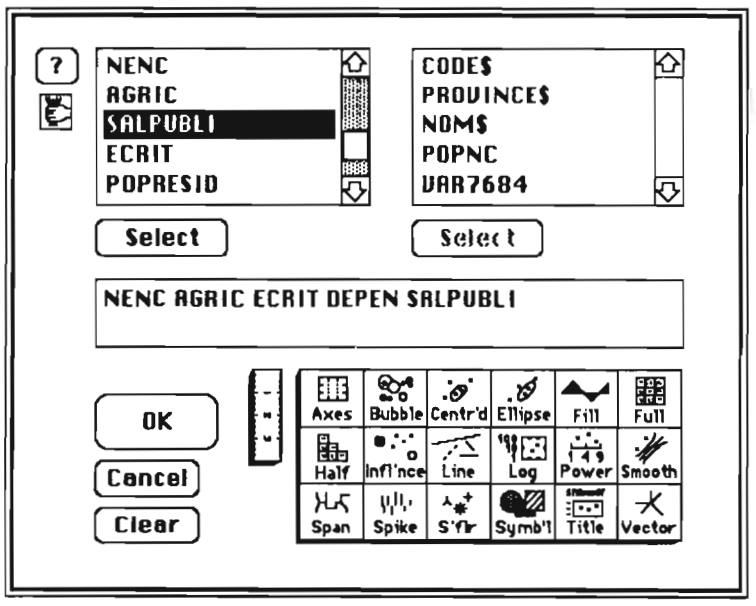


figure n°4.7. SYSTAT: le choix des variables constituant la matrice de graphiques bivariés.

plupart des options déjà proposées pour les graphiques bivariés. De plus, l'option HALF permet de ne tracer que la partie inférieure de la matrice de graphiques bivariés.

Si l'on ne retient aucune de ces options, le graphique obtenu représente les nuages de points des variables prises deux à deux, avec leur nom dans la diagonale (figure n° 4.8).

Mais il apparaît souvent utile de compléter ce genre de graphique par une information relative à la distribution de chaque variable (figure n° 4.6).

Ceci est possible en sélectionnant, non pas l'article **SPLOM**, mais l'article **DENSITY** du sous-menu **SPLOM** du menu **PLOT**. Trois boutons permettent de choisir le contenu des cases de la diagonale (ici, l'option STRIPE a été retenue). En combinant ces options avec celles relatives au contenu des graphiques bivariés (ici INFLUENCE), on obtient des graphiques assez complets et parfois même complexes.

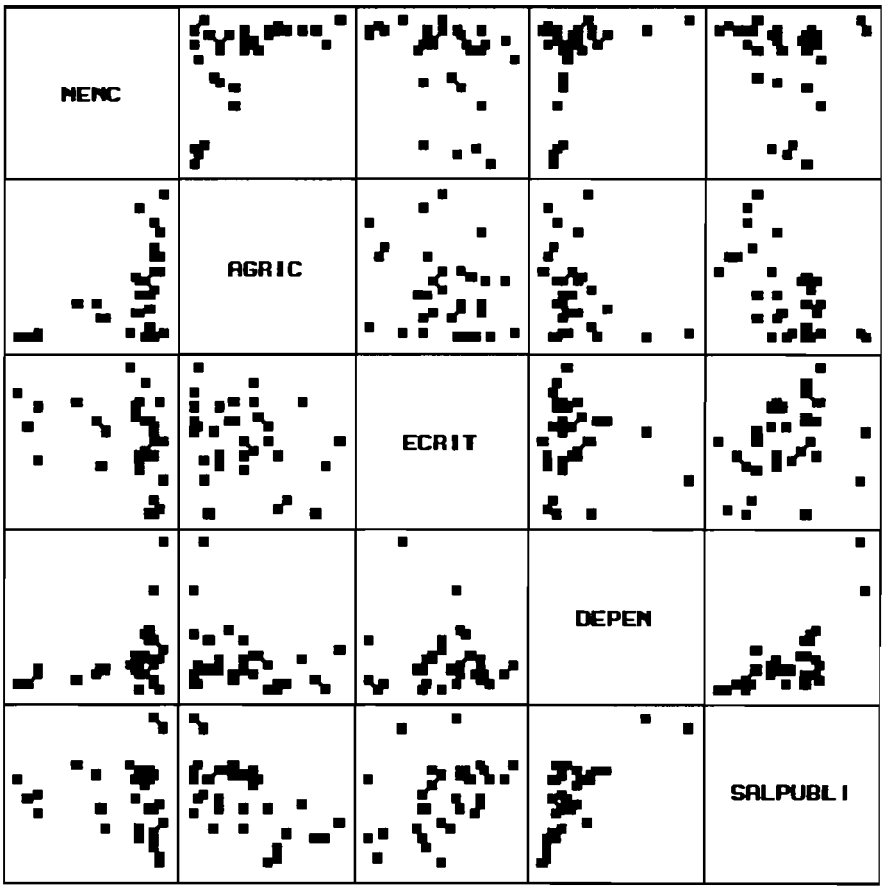


figure n° 4.8. SYSTAT: la matrice de graphiques bivariés de 5 variables du fichier sur les communes de Nouvelle-Calédonie.

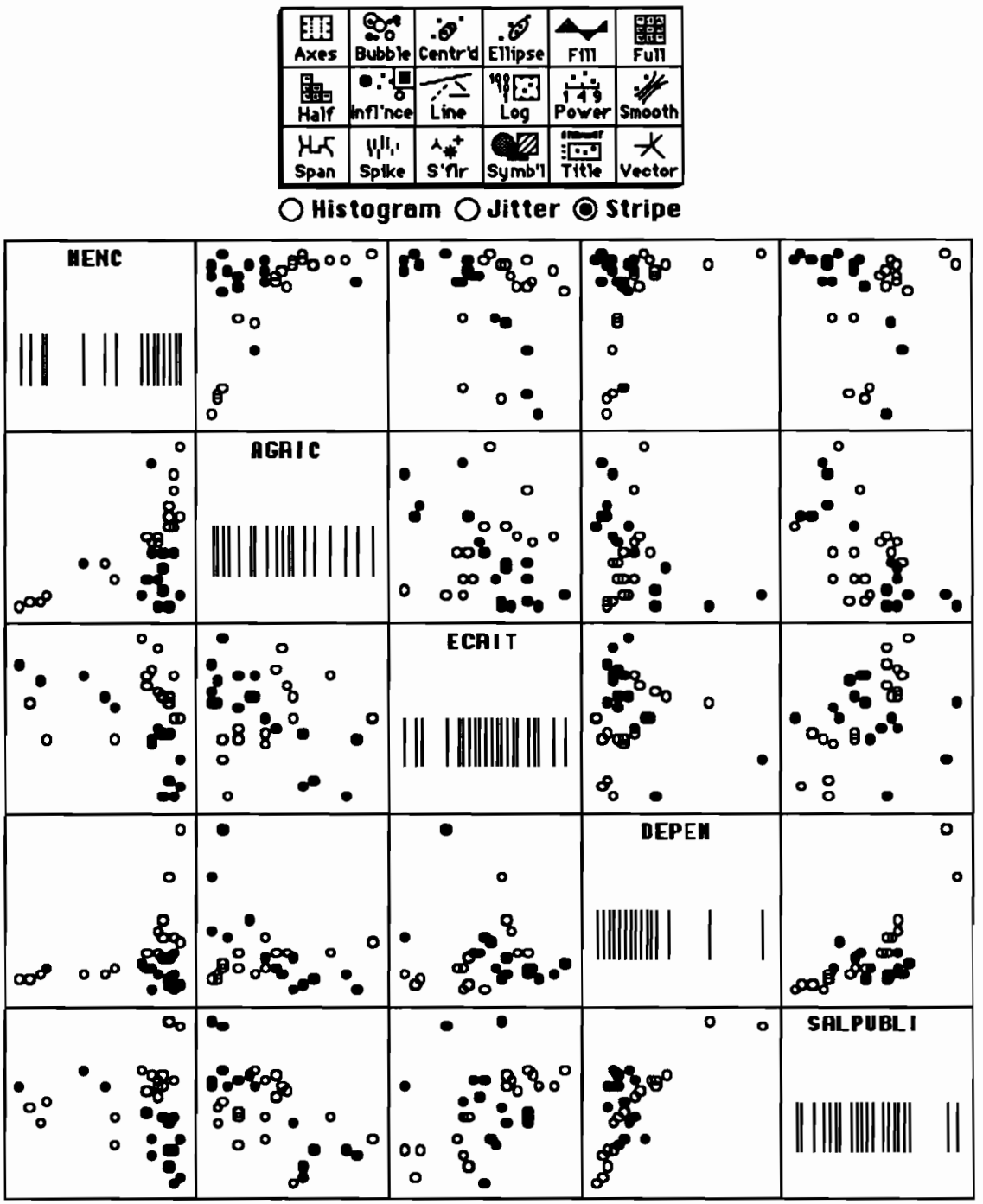


figure n°4.9. SYSTAT: la matrice de graphiques bivariés de 5 variables du fichier sur les communes de Nouvelle-Calédonie, avec les options STRIPE et INFLUENCE.

Ainsi, malgré une interactivité avec le graphique réduite, **SYSTAT** réalise d'intéressantes matrices de graphiques bivariés.

4.2.2. DataDesk

Le tracé d'une matrice de graphiques bivariés avec **DataDesk** suit le même mode opératoire que celui des graphiques bivariés proprement dit. Lorsque les icônes des variables sont sur le bureau du logiciel, il faut sélectionner les variables devant composer la matrice par un clic sur leurs icônes respectives. Ensuite, en choisissant l'article **PLOT MATRIX** du menu **PLOT** (figure n° 4.10), **DataDesk** ouvre une fenêtre par graphique, l'assemblage des fenêtres constituant, au fur et à mesure, la matrice de graphiques bivariés (figure n° 4.11).

Remarquons que **DataDesk** ne trace que le triangle supérieur de la matrice, ce qui a pour effet de réduire de manière importante le temps de traitement, sans occasionner de perte d'information.

Dans la diagonale de la matrice graphique, on trouve, pour chaque variable, un graphique nommé NORMAL PROBABILITY PLOT. Il s'agit d'une technique simple d'évaluation de la ressemblance des distributions à la loi Normale.

Il s'agit d'un graphique bivarié où l'on trouve, en ordonnée, les valeurs de la variable étudiée, et, en abscisse, les valeurs théoriques de la loi Normale estimée par la fonction NSCORES du logiciel. Pour que la loi Normale corresponde à la distribution observée, tous les points doivent être alignés sur une droite.

Bien entendu, tous les outils destinés à l'étude des graphiques bivariés, décrits au chapitre 3, demeurent disponibles avec une matrice de graphiques bivariés:

- menus *hyperview*, situés dans la partie gauche de chaque fenêtre permettant le calcul et le tracé d'une droite de régression.
- menus *hyperview* sur le nom de chaque variable permettant d'en obtenir une description univariée.
- outils de sélection, de tranchage et de brossage, destinés à l'examen de points, ou de groupes de points sur un

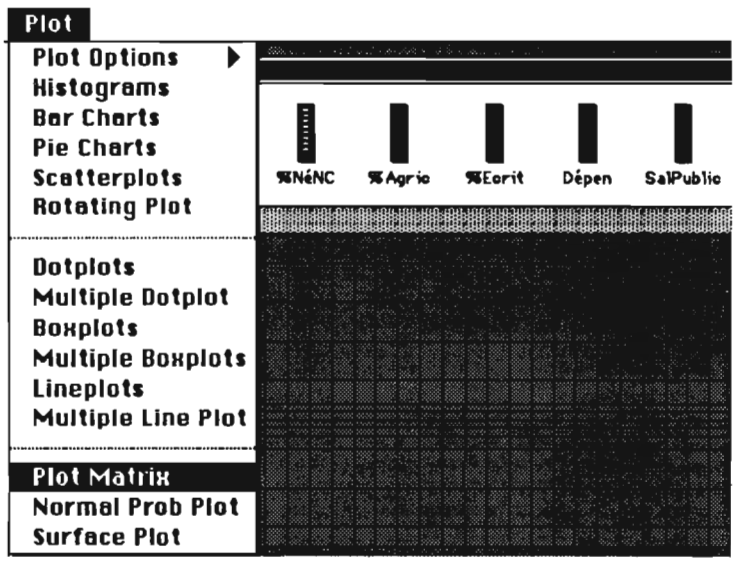


figure n° 4.10. DataDesk: le choix du menu PLOT MATRIX pour tracer une matrice de graphiques bivariés.

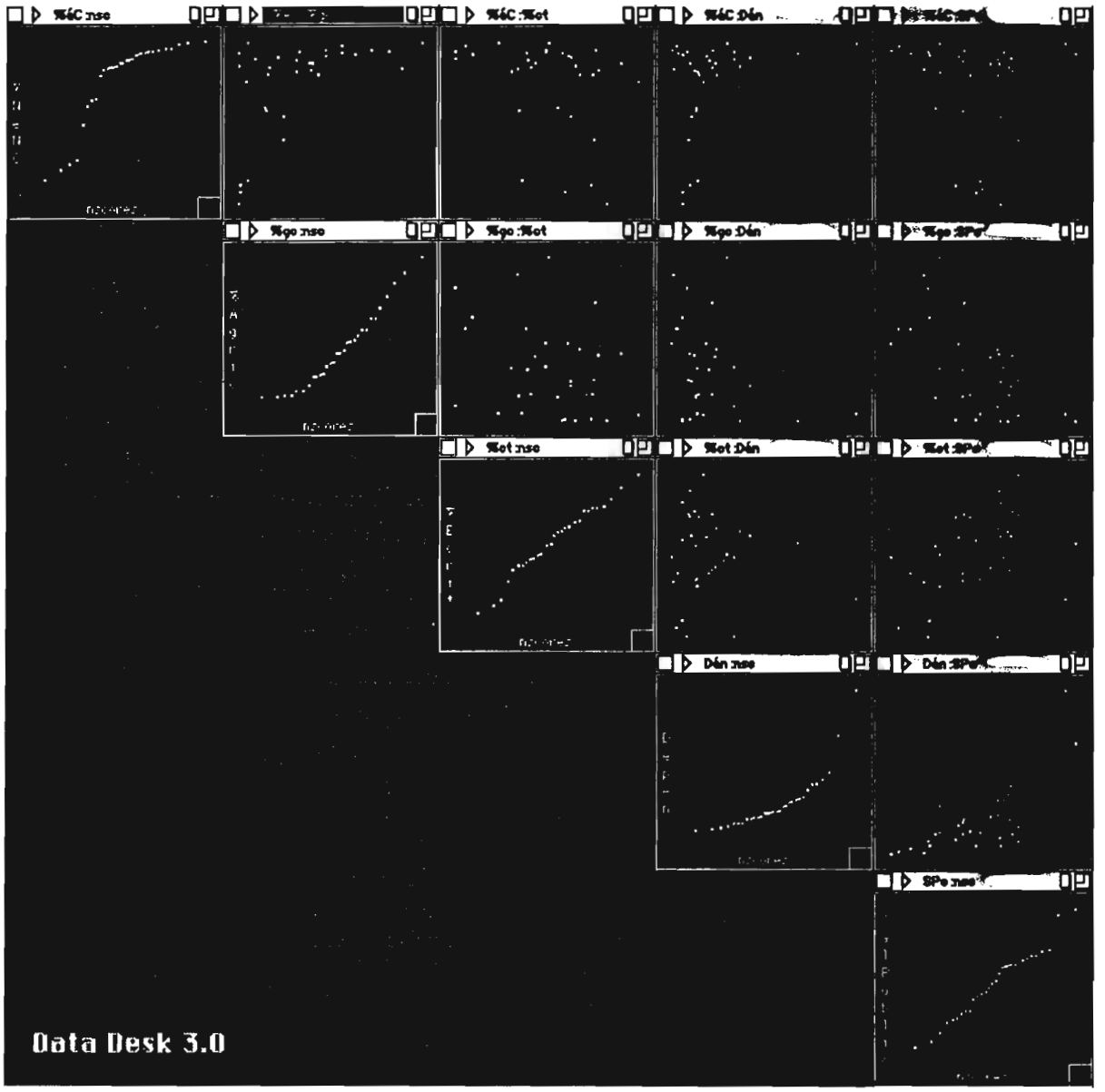


figure n° 4.11. DataDesk: la matrice de graphiques bivariés de 5 variables du fichier sur les communes de Nouvelle-Calédonie.

graphique, en relation avec tous les autres graphiques.

De plus, lorsque l'article **AUTOMATIC UPDATE** des menus *hyperview* attachés à chaque fenêtre graphique a été activé, **DataDesk** provoque, grâce aux liens dynamiques qu'il a établis, la répercussion de toute modification des données dans la matrice de graphiques bivariés.

4.2.3. JMP

Dans la terminologie propre à **JMP**, l'exploration multivariée correspond à la plate-forme **Y's BY Y's** à laquelle on accède par le menu **ANALYZE** (figure n° 4.12).

Les variables constituant la matrice de graphiques bivariés doivent donc jouer le rôle Y qui leur est attribué soit à l'aide des menu *pop-up* des en-têtes de colonnes, soit avec le dialogue destiné à cet effet, qui s'ouvre lorsqu'aucun rôle n'a encore été défini (figure n° 4.13). Dans ce cas, il suffit de cliquer sur les noms des variables retenues pour les sélectionner.

La fenêtre des résultats, intitulée MULTIVARIATE, comprend quatre parties. On y trouve tout d'abord la matrice des coefficients de corrélation linéaire précédemment décrite. Puis vient la matrice des graphiques bivariés proprement dite (figure n° 4.14). Les cases situées en dehors de la diagonale représentent les nuages de points; seuls les noms des variables apparaissent dans la diagonale.

Vient ensuite un tableau très original. Il s'agit de la matrice inverse des coefficients de corrélation linéaire (figure n° 4.15). Les valeurs situées dans la diagonale de cette matrice renseignent sur le degré de corrélation linéaire de chaque variable par rapport à toutes les autres. Par exemple, la variable %SALPUBLIC qui présente la valeur la plus élevée est aussi celle qui, sur l'ensemble des

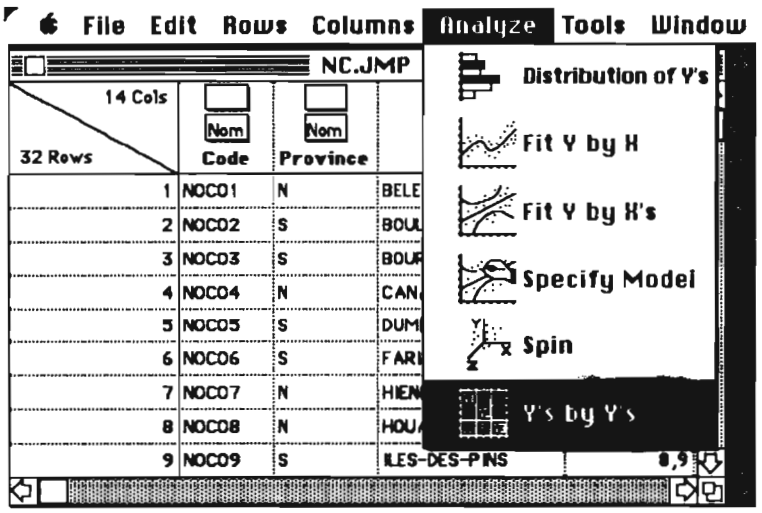


figure n° 4.12. JMP: la sélection de la plate-forme Y's BY Y's du menu ANALYZE.

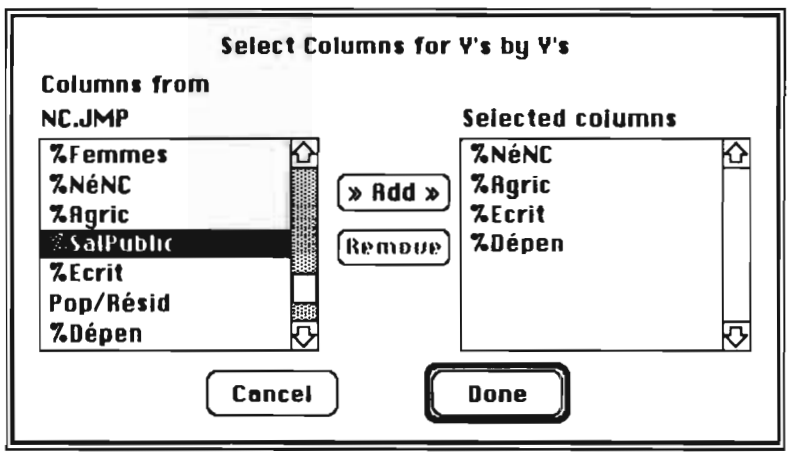


figure n° 4.13. JMP: la sélection des variables devant composer la matrice de graphiques bivariés.

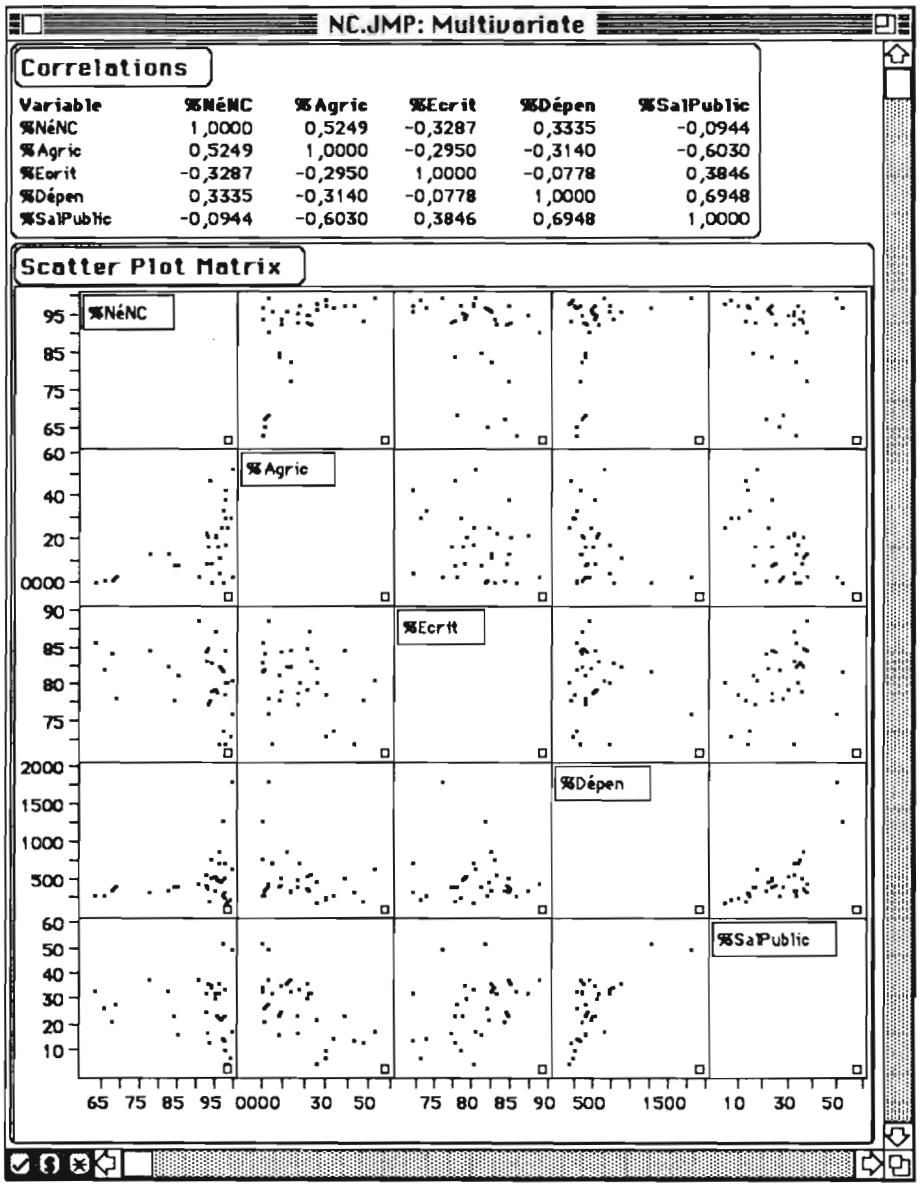


figure n°4.14. JMP: la matrice des coefficients de corrélation linéaire et celle des graphiques bivariés.

graphiques bivariés, présente, le plus souvent, un allongement révélateur de corrélations plus ou moins fortes.

Enfin, le graphique de la distance multivariée de Mahalanobis a pour fonction de simplifier le repérage des observations les plus exceptionnelles, compte tenu de l'ensemble des variables figurant dans la matrice des graphiques bivariés. Le premier point se détache nettement. En cliquant dessus, sa ligne dans le tableau de données est soulignée: il s'agit de la commune de Belep dont le comportement particulier a maintes fois été révélé. Notons que cette

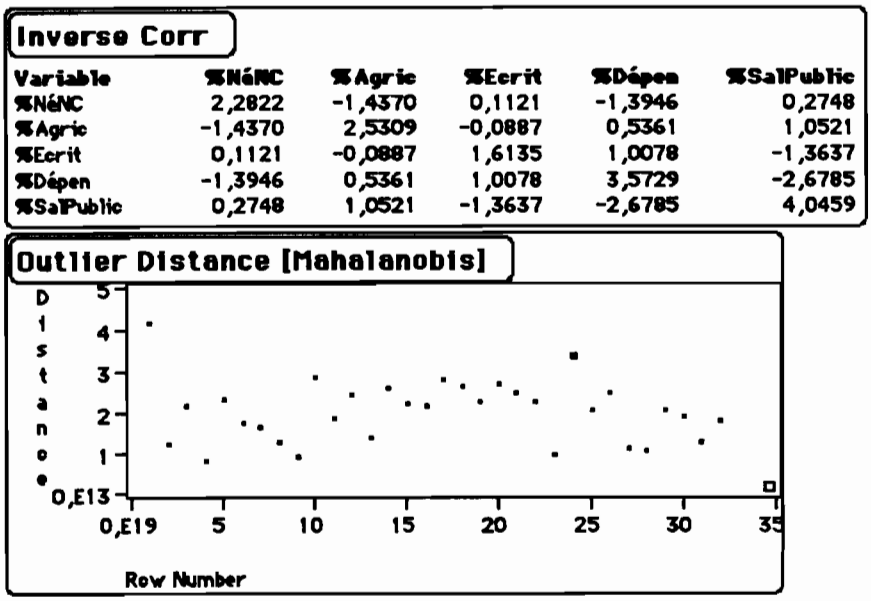


figure n°4.15. JMP: la matrice inverse des coefficients de corrélation linéaire et le graphique des distances de Mahalanobis.

distance multivariée de Mahalanobis peut être enregistrée dans le tableau de données à l'aide d'un menu *pop-up* **$**.

L'outil de brossage peut être directement utilisé sur la matrice de graphiques bivariés. Mais on n'a pas accès directement sur ce graphique au calcul, ou au tracé des droites de régression. Il faut, dans ce cas, revenir à la plate-forme **FIT Y BY X**. Sur ce plan, les concepteurs de **JMP** ont sans doute privilégié leur logique de plate-forme: **Y's BY Y's** n'est pas **FIT Y BY X**. Affaire de point de vue, sans doute!

En conclusion, on retiendra, à propos des matrices de graphiques bivariés la variété des options de **SYSTAT**, l'immense interactivité de **DataDesk**, la cohérence ainsi que les tableaux complémentaires de **JMP**.

4.3. L'exploration galactique: la toupie

L'une des méthodes les plus intéressantes et originales de l'analyse exploratoire multivariée repose sur un graphique trivarié, que l'on peut faire tourner autour de ses trois axes, afin d'observer le nuage de points sous divers angles. Cette figure porte différents noms dans la littérature anglo-saxonne: *3D plot*, *Spin*, ou bien encore *Rotating plot*. En français, l'expression «Graphique Rotationnel» apparaît parfois, mais il semble à la fois plus français et plus imagé de parler de Toupie, dont la traduction anglaise est *Spinning top*. En effet, d'après le Dictionnaire alphabétique et analogique de la langue française Robert, une toupie est «un jouet d'enfant, formé

d'une masse conique, sphéroïdale, etc., munie d'une pointe sur laquelle elle peut se maintenir en équilibre en tournant». De cette définition, on retient les idées de volume et de rotation qui apparaissent précisément comme les caractères les plus originaux de ce graphique. La métaphore peut s'étendre à la méthode d'analyse elle-même: d'une certaine manière, la toupie constitue un vaisseau d'exploration des galaxies. Chaque étoile représente une observation localisée dans l'espace multidimensionnel (ou multivarié) en fonction de ses valeurs sur les variables formant le système d'axes.

4.3.1. Construire une toupie

Les nuages de points analysés au cours du chapitre 3 ont été construits en localisant chaque observation en fonction de ses valeurs sur deux variables formant les axes orthogonaux d'un plan. En considérant une troisième variable, on introduit une troisième dimension représentée par un axe orthogonal aux deux autres: le nuage de points acquiert ainsi une épaisseur.

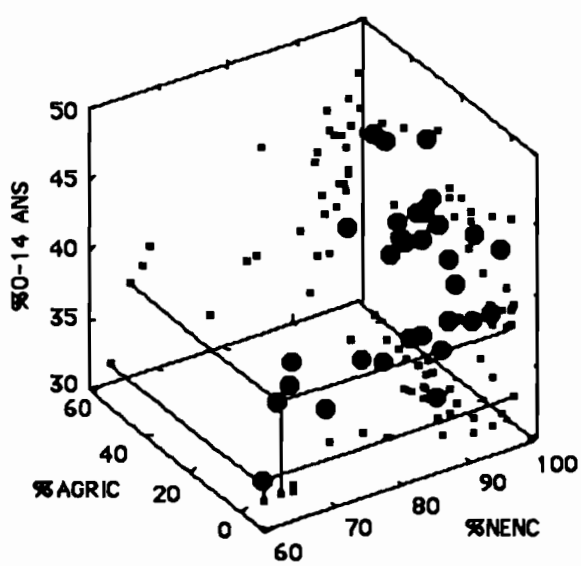


figure n°4 .16. Le nuage de points tridimensionnel formé par les valeurs de communes sur les variables %NENC, %AGRIC et %0-14 ANS.

On peut représenter un tel nuage en perspective. Par exemple, chaque commune de Nouvelle-Calédonie forme un point (gros et rond) sur le graphique construit en fonction du pourcentage de ses habitants nés en Nouvelle-Calédonie, de celui des agriculteurs par rapport à la population active et, enfin, de celui des personnes âgées de 0 à 14 ans par rapport à la population totale (figure n° 4.16).

En chaque plan formé par les variables prises deux à deux, on obtient une «boîte» qui renforce l'impression de volume. Les communes peuvent être projetées sur chaque face, l'ensemble de ces projection formant à son tour un nuage de points (petits et carrés) bivarié comme ceux du chapitre 3.

On notera que, si la lecture d'un seul nuage de points bivarié est aisée, il apparaît plus difficile de retenir et de mettre en relation trois graphiques bivariés simultanément (les outils de brossage et de tranchage sont là pour simplifier cette tâche). Par ailleurs, la perspective adopte un angle de vue qui n'est pas toujours le meilleur pour examiner chaque groupe de points. Le rôle de la toupie est précisément de faciliter l'examen du volume sous tous les angles.

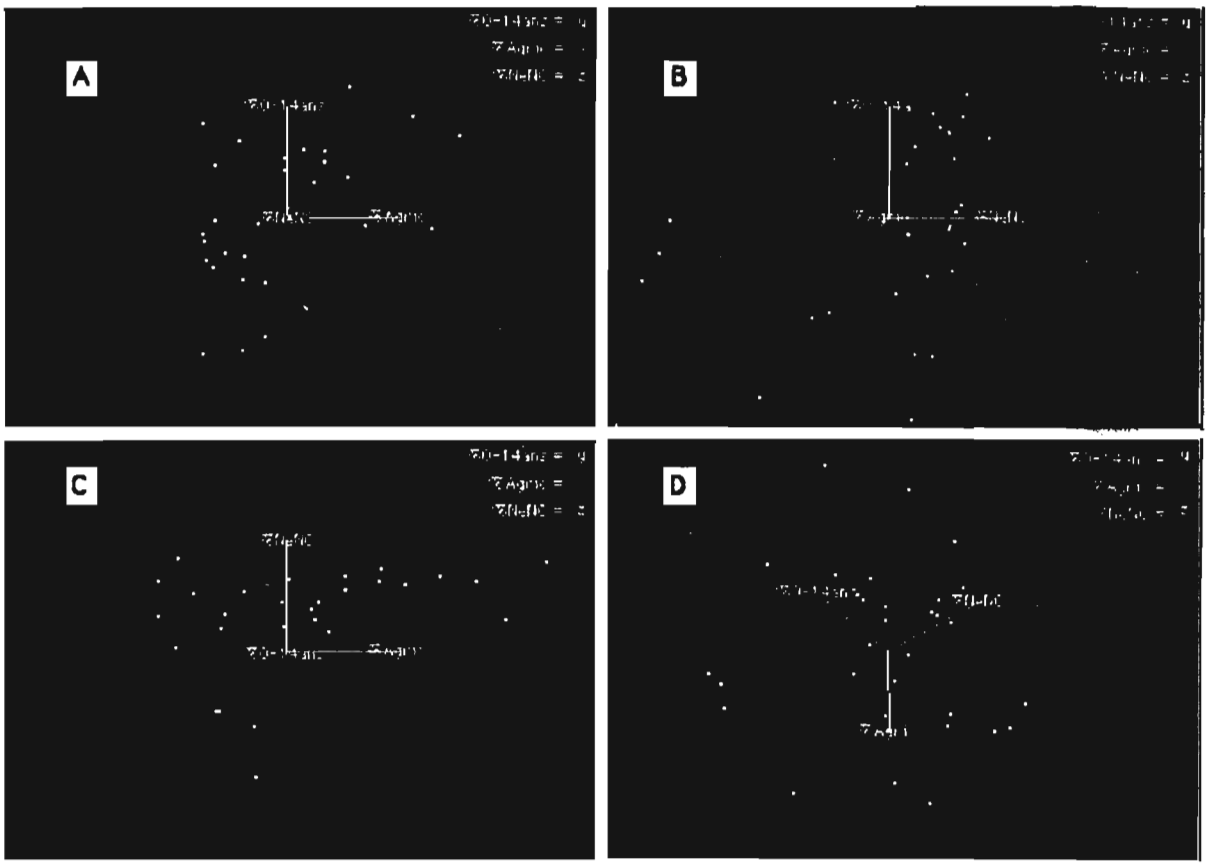


figure n° 4.17. La projection d'un nuage de points tridimensionnel sur des plans d'orientations différentes.

La construction d'une toupie correspond au premier principe de l'analyse exploratoire multivariée: P pour Projection. Au lieu de recourir à l'artifice de la perspective pour rendre compte du volume formé par le nuage de points tridimensionnel, on le projette sur un plan figuré par l'écran de l'ordinateur. Selon l'orientation de ce plan par rapport aux axes de référence, la projection du nuage de points révèle diverses configurations, diverses formes qu'il faut interpréter.

Pour faciliter l'observation, on place en général le système d'axes au centre du nuages, sur le point correspondant à la médiane de chacune des variables. Lorsque le plan de projection est parallèle à deux axes, le troisième disparaît, ou plus précisément, il est confondu avec l'origine du système d'axes. Dans tous les autres cas, lorsque le plan de projection forme un angle compris entre 0 et 90°, tous les axes demeurent visibles.

Pour illustrer ce propos, nommons X le pourcentage d'agriculteurs dans la population active (%AGRIC), Y, celui des 0-14 ans (%0-14) dans la population totale et Z, la proportion de la population totale née en Nouvelle-Calédonie (%NENC). Les parties A, B, et C de la

figure n° 4.17 représentent le nuage de points projeté sur des plans parallèles, respectivement à XY, YZ et XZ: seuls les axes concernés sont visibles. Par contre, sur la partie D, les 3 axes (ou, plus exactement, leurs projections) sont visibles, leur longueur dépendant de l'inclinaison par rapport à l'un ou l'autre des axes.

La construction d'une toupie revient donc à choisir les variables relatives aux trois dimensions, à placer les axes sur le nuage de points afin de pouvoir le faire tourner ensuite autour de l'un des axes.

4.3.1.1. SYSTAT

On accède à la toupie de **SYSTAT** lorsqu'après avoir ouvert le fichier de

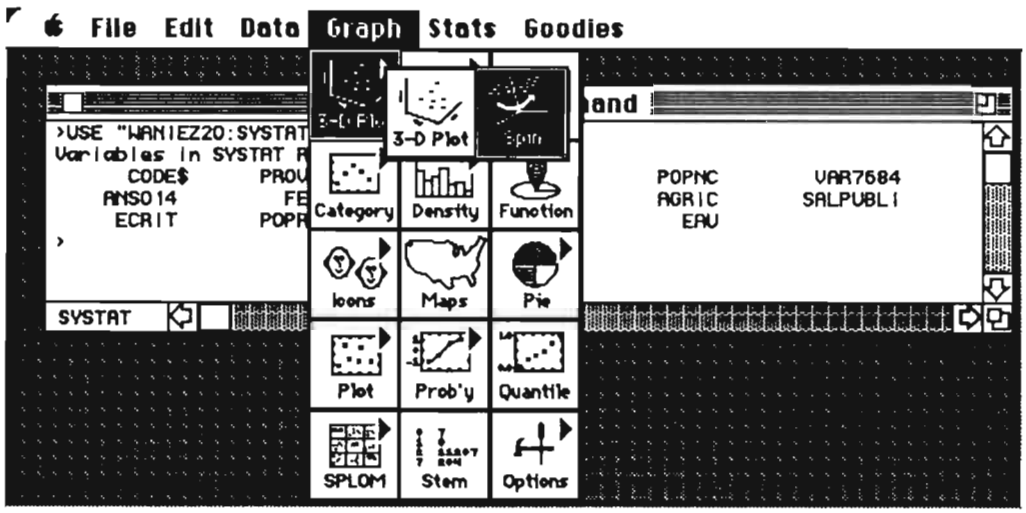


figure n° 4.18. SYSTAT: le menu d'accès à la toupie.

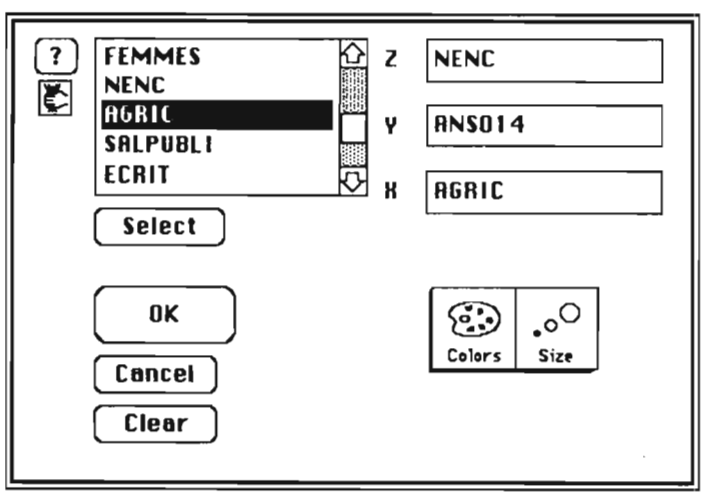


figure n° 4.19. SYSTAT: la sélection des variables Z, Y et X de la toupie.

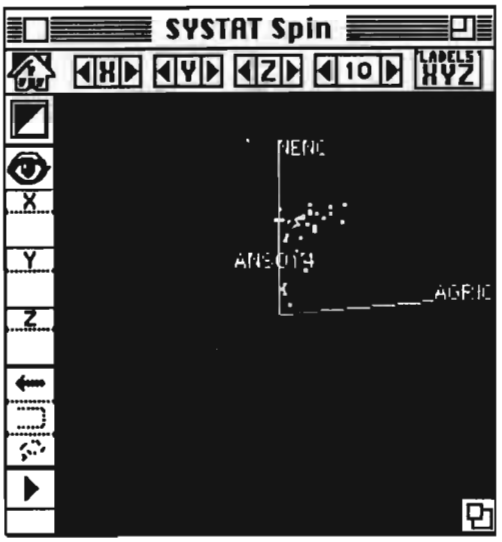


figure n° 4.20. SYSTAT: l'affichage de la toupie en position initiale, avant rotation.

données, on choisit l'article **SPIN** du sous-menu **3D PLOT** du menu **GRAPH**. (figure n° 4.18).

Le logiciel ouvre alors un dialogue destiné à sélectionner les trois variables du graphique (figure n° 4.19), puis affiche la toupie dans une fenêtre particulière (figure n° 4.20).

En position initiale, avant rotation, l'axe Z correspond à l'axe vertical, légèrement incliné vers l'avant; X forme l'axe horizontal de la largeur, après une légère rotation vers l'arrière autour de Z; Y constitue l'axe horizontal de la profondeur, après une légère rotation vers l'avant autour de Z. Notons que le système d'axes n'est pas au centre du nuage de points, mais localisé aux valeurs 0 sur les trois variables.

4.3.1.2. DataDesk

Lorsque les variables sont sur le bureau de **DataDesk**, il faut sélectionner trois d'entre elles par un clic sur leurs icônes (figure n° 4.21). L'ordre de la sélection a une importance, puisque c'est de lui que dépend l'identification des axes: la première variable correspond à Y, la seconde à X et la troisième à Z.

En choisissant l'article **ROTATING PLOT** du menu **PLOT**, la toupie fait son apparition dans une nouvelle fenêtre. En position initiale, avant rotation, la figure représente une projection du nuage de points sur un plan parallèle à XY. L'axe Z, de face, demeure invisible (figure n° 4.22). Notons que le système d'axes est au centre du nuage de points, mais que cette position peut être transférée aux coordonnées 0,0,0 ou bien encore au point moyen grâce à l'article **SET ROTATING PLOT OPTIONS** du sous-menu **PLOT OPTIONS** du menu **PLOT**.

Lorsque les variables sont sur son bureau, **DataDesk** propose un outil de contrôle de la dispersion du nuage. On y accède par un article du sous-menu **TOOLS** du menu **MODIFY**. En plaçant le curseur sur le nuage de points, on provoque par un clic une concentration par rapport aux axes (figure n° 4.23).

Au contraire, un clic à l'extérieur du nuage accentue sa dispersion par rapport aux axes. Cet outil s'avère très utile lorsque quelques points situés très loin de la majorité des autres provoquent une telle concentration de ces

derniers qu'il est impossible de les examiner.

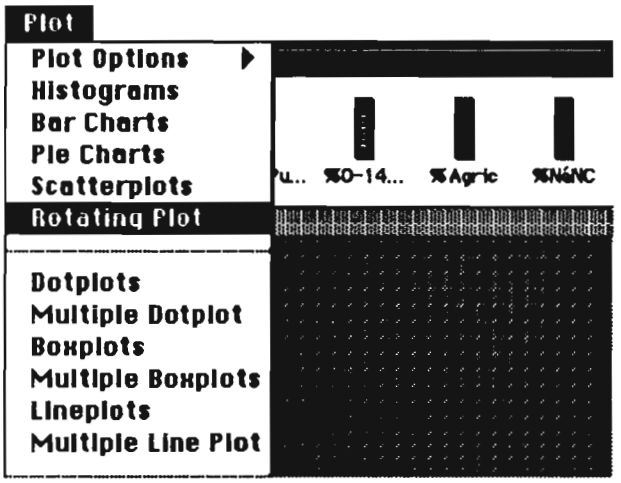


figure n° 4.21. DataDesk: la sélection des 3 variables destinées à la construction de la toupie.

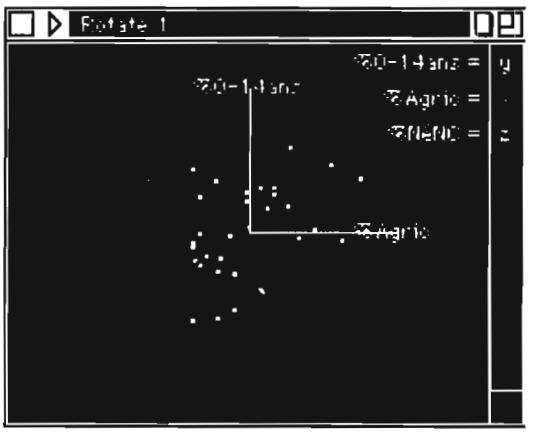


figure n° 4.22. DataDesk: l'affichage de la toupie en position initiale.

4.3.1.3. JMP

On accède à la toupie de **JMP** par la plate-forme **SPIN** du menu **ANALYZE** (figure n° 4.24). Si les rôles des variables n'ont pas encore été fixés, le logiciel

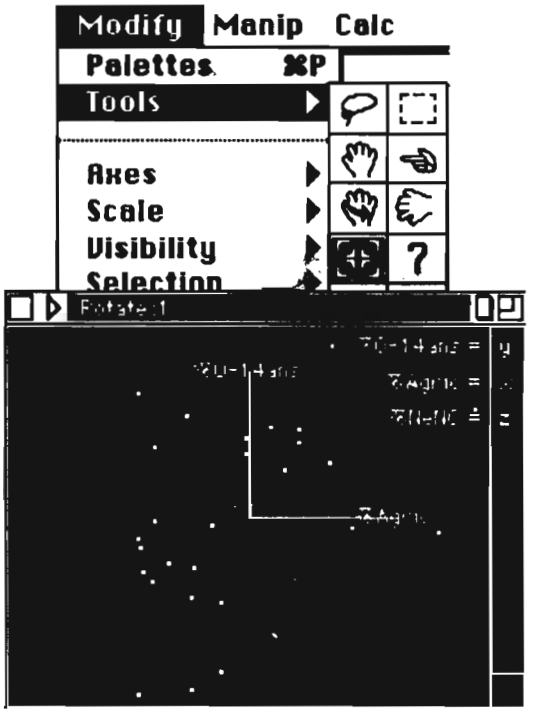


figure n° 4.23. DataDesk: le contrôle de la dispersion de l'affichage du nuage.

ouvre un dialogue comprenant deux fenêtres (figure n° 4.25): celle de droite donne la liste des variables présentes dans le fichier de données. Par un clic sur l'un de ces noms, la variable correspondante est sélectionnée, et son nom apparaît dans la fenêtre de gauche.

Après avoir sélectionné 3 variables, un clic sur le bouton DONE provoque l'affichage de la toupie dans une fenêtre nommée SPIN (figure n° 4.26). Les axes prennent la dénomination X, Y et Z en fonction de l'ordre de sélection des variables.

Notons que la position de départ coïncide avec un plan de projection parallèle au plan XY, Z demeurant invisible.

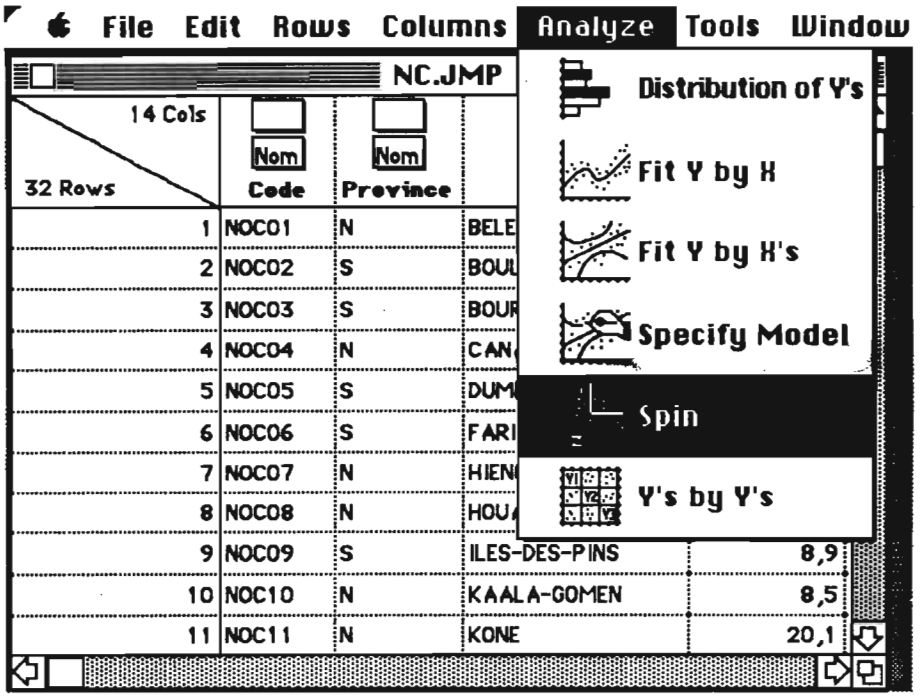


figure n° 4.24. JMP: le choix de la plate-forme SPIN.

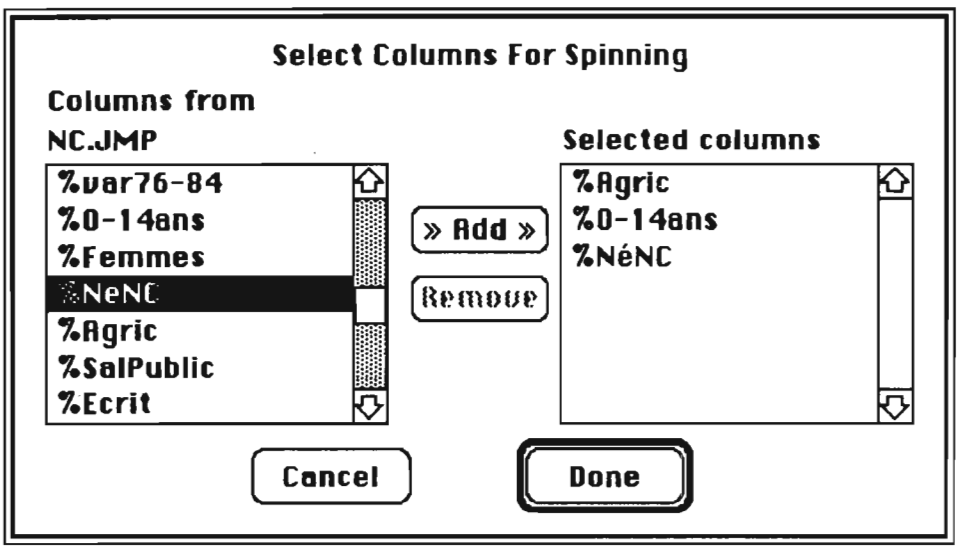


figure n° 4.25. JMP: la sélection de variables de la toupie.

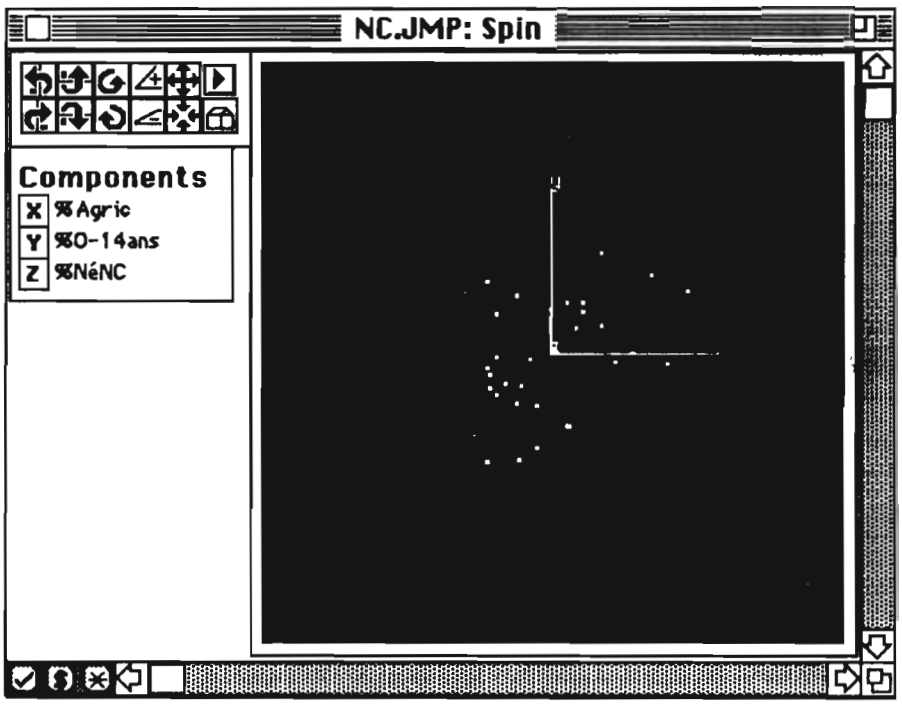


figure n°4.26. JMP: la toupie en position initiale.

Dans la boîte à outils,  permet, comme avec **DataDesk** de contrôler la dispersion du nuage de points par rapport aux axes. Un clic sur  accentue la dispersion, alors que  provoque un resserrement.

4.3.1.4. MacSpin

Seul logiciel entièrement dédié à la toupie, **MacSpin** affiche d'emblée un graphique correspondant aux 3 premières variables du fichier retenu.

La modification des variables définissant les axes se fait très simplement à l'aide de la fenêtre des Variables (figure n° 4.27). Il suffit de sélectionner l'une d'entre elles par un clic sur son nom, puis de faire glisser ce nom jusqu'à la flèche de l'un ou l'autre des axes X, Y ou Z. Le graphique est immédiatement modifié en fonction de son nouveau système d'axes.

En position initiale, les axes sont placés au centre du nuage de points, mais l'article **ORIGINE A ZERO** du menu **VUE** permet de les placer aux coordonnées 0,0,0 (figure n° 4.28). Le plan de projection initial est parallèle à la facette XY, Z étant invisible.

4.3.2. Faire tourner la toupie

La toupie exploratoire peut tourner autour de chacun de ses axes. La rotation permet de créer l'illusion de la troi-

sième dimension et se rapporte au second principe de l'analyse exploratoire: **R** pour Rotation. Cette rotation peut être réalisée de deux façons:

- en déplaçant un pointeur directement sur le graphique. Ce mode de fonctionnement porte le nom de rotation libre.
- en provoquant la rotation autour de X, Y ou Z grâce aux outils destinés à cet effet: un clic correspond à un angle de rotation prédéterminé. Ce mode de fonctionnement s'appelle la rotation contrôlée

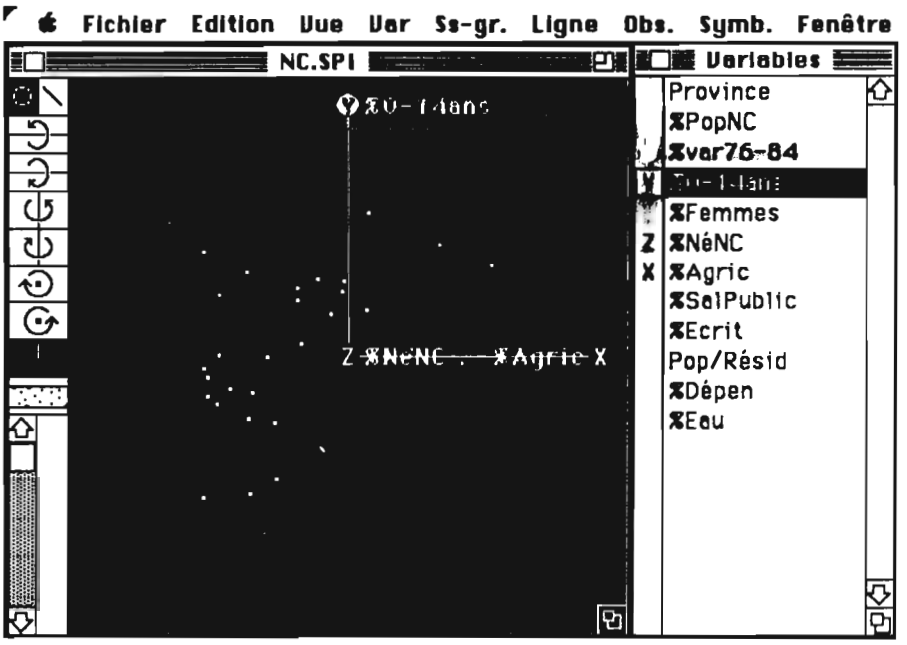


figure n°4.27. MacSpin: la toupie en position initiale et la sélection des variables définissant les axes.

Pour rechercher des formes significatives, deux méthodes peuvent être utilisées l'une après l'autre:

- animer le nuage de points d'un mouvement uniforme: le dessin se met alors à s'animer jusqu'à ce que l'utilisateur, en découvrant une vue intéressante, impose un arrêt sur image. Dans ce cas, on parle de rotation continue.
- par contre, lorsque l'on veut affiner une vue, il est préférable de faire pivoter la toupie pas à pas.

Lorsque la vue est jugée satisfaisante, il apparaît utile d'enregistrer la nouvelle projection du nuage de points par rapport au système de coordonnées d'origine.

Enfin, il faut pouvoir, dans tous les cas, revenir à la position d'origine.

Comme la rotation supposée n'est vraiment perceptible que lorsque le graphique se déplace sur l'écran,

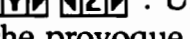
figure n°4.28. MacSpin: placer l'origine du graphique sur la valeur des 3 variables.

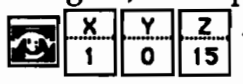
l'exposé se limitera, pour les quatre logiciels retenus, à l'analyse des cinq fonctions suivantes:

- rotation libre.
- rotation contrôlée.
- rotation continue.
- enregistrement d'une nouvelle projection.
- retour à la position d'origine.

4.3.2.1. SYSTAT

- rotation libre: elle s'effectue en saisissant un axe par un clic sur son extrémité, puis, le bouton de la souris enfoncé, en faisant faire à cet axe un mouvement circulaire.

- rotation contrôlée: chaque axe dispose d'un bouton de contrôle de sa rotation  . Un clic sur la flèche de gauche provoque une rotation vers la gauche; un clic sur la flèche de droite conduit à une rotation en sens inverse. L'angle de rotation produit par un seul clic (ici 2 degrés par clic) peut être fixé grâce à un bouton destiné à cet effet  .

- enregistrement d'une nouvelle projection: on obtient les coordonnées d'une nouvelle projection, par rapport au système d'axes d'origine, en cliquant sur l'icône de l'œil  .

Les valeurs sont indiquées en degrés par rapport à la valeur 0,0,0 (ici, 1° pour X, 0° pour Y et 15° pour Z).

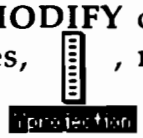
- retour à la position d'origine: en cliquant sur l'icône de la maison  (*home*), le graphique reprend sa position d'origine.

4.3.2.2. DataDesk

- rotation libre: deux outils du sous-menu TOOLS du menu MODIFY permettent la rotation libre du nuage de points. Avec la main horizontale , on décrit un mouvement circulaire après avoir saisi l'un des axes; le fonctionnement est donc le même qu'avec SYSTAT. Le second outil, main verticale fléchée , provoque une rotation perpendiculaire au sens de déplacement: si la main va de la gauche vers la droite, le nuage de points tourne autour de l'axe vertical du plan de projection, dans le sens inverse des aiguilles d'une montre; un déplacement de la droite vers la gauche provoque une rotation autour de l'axe vertical, dans le sens des aiguilles d'une montre. Par contre, si la main va du haut vers le bas, le nuage de points tourne autour de l'axe horizontal du plan de projection.

- rotation contrôlée: il n'y a pas de bouton de contrôle de la rotation, mais il est quand même possible de faire bouger le graphique pas à pas à l'aide des touches **I, J, K, L** du clavier. Les flèches du clavier sont également utilisables. Le pas de la rotation, nommé ici sensibilité (*sensitivity*), peut aussi être fixé, mais de manière intuitive, sans mesure précise de l'angle.

- rotation continue: lorsqu'une rotation contrôlée a commencé par utilisation des touches **I, J, K, L**, ou des flèches du clavier, le mouvement ainsi provoqué peut être poursuivi par un clic sur le graphique à l'aide de la main verticale fléchée  ; la rotation continue cesse dès qu'un nouveau clic intervient.

• enregistrement d'une nouvelle projection: en sélectionnant l'article **SHOW EQUATIONS** du sous-menu **EQUATIONS** du menu **MODIFY**, on obtient les coordonnées d'une nouvelle projection, par rapport au système d'axes d'origine, à chaque arrêt de la rotation. Dans le même menu, l'article **SHOW EQUATIONS DURING ROTATION** affiche les équations à chaque pas de rotation; du fait du temps de calcul nécessaire, la vitesse de rotation chute de manière considérable. De plus, il est possible d'enregistrer les coordonnées de chaque point sur le plan de projection: l'article **RECORD PROJECTION** du sous-menu **DIMENSION** du menu **MODIFY** crée trois nouvelles variables,  , nommées

YPROJECTION, XPROJECTION (pour le plan de projection), et ZPROJECTION (pour la profondeur) que l'on peut traiter ensuite comme n'importe quelle autre variable.

• retour à la position d'origine: en sélectionnant l'article **HOME** du menu *hyperview* attaché à la fenêtre du graphique, le système d'axes revient à sa position d'origine.

4.3.2.3. JMP

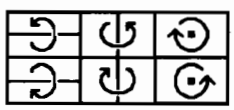
• rotation libre: main verticale  , provoque une rotation perpendiculaire au sens de déplacement, de manière semblable à la main verticale fléchée de **DataDesk**. L'outil main verticale  , est l'un des articles du menu **TOOLS**.

• rotation contrôlée: chaque axe est doté de deux boutons de contrôle de rotation  . Ceux de gauche correspondent à l'axe Y, ceux du centre à X, et ceux de droite à Z. Pour chaque axe, un clic sur le bouton supérieur provoque une rotation dans le sens des aiguilles d'une montre, et dans le sens inverse pour le bouton inférieur. L'angle de rotation produit par un seul clic peut être fixé grâce à deux boutons  : celui du haut permet d'ouvrir l'angle, alors que celui du bas le ferme; lors du clic sur l'un ou l'autre de ces deux boutons, la valeur de l'angle est affichée.

• rotation continue: elle est possible en rotation libre par l'outil main verticale  , mais avec la touche majuscule enfoncée. Elle cesse par un nouveau clic sur le graphique.

• retour à la position d'origine: en cliquant sur l'icône de la maison  (*home*), le graphique reprend sa position d'origine.

4.3.2.4. MacSpin

• rotation contrôlée: chaque axe est doté de deux boutons de contrôle de rotation  Pour chaque axe, un clic sur le bouton inférieur provoque une rotation dans le sens des aiguilles d'une montre, et dans le sens inverse pour le bouton inférieur.

• enregistrement d'une nouvelle projection: il est possible d'enregistrer les coordonnées de chaque point sur le

plan de projection: l'article **CREER** du menu **VAR** crée une nouvelle variable que l'on peut traiter ensuite comme n'importe quelle autre variable. Un dialogue permet d'en donner son nom (par défaut PROVISOIRE) et d'indiquer sur quel axe (X, Y ou Z) doit être effectué le calcul. Il est possible de conserver un état de la rotation grâce à l'article **ENREGISTRER** du menu **VUE**, ou de faire alterner deux vues successives par l'article **ECHANGER** du même menu.

• retour à la position d'origine: en sélectionnant l'article **R.A.Z.** du sous menu **ORIENTATION** du menu **VUE**, le graphique reprend sa position d'origine.

4.3.3. Former des groupes

Lorsqu'un groupe de points fait son apparition sur une vue particulière obtenue après rotation de la toupie, il s'avère souvent utile de les désigner par un symbole particulier. Ainsi, lorsque la rotation sera poursuivie, les nouvelles positions des points solidaires d'une groupe pourront être facilement examinées afin de savoir pourquoi ce groupe est visible dans certaines positions et invisible dans d'autres. Ce marquage répond au troisième principe de l'analyse exploratoire multivariée: I pour Isoler.

Illustrons ce propos par un exemple. Dans une position donnée de la toupie, on observe un regroupement de trois points. Afin de mieux les distinguer, ils sont successivement sélectionnés (figure 4.29.A) et marqués d'un symbole particulier: une croix (figure n° 4.29.B). Après une nouvelle rotation de la toupie, un nouveau groupe est identifié, sélectionné (figure n° 4.29.C) et, enfin, marqué par un symbole qui lui est propre: un bâton vertical. Durant cette dernière rotation, le premier groupe, marqué d'une croix, s'est lui aussi déplacé: on remarque ainsi que, s'il est homogène, peu dispersé sur le plan de projection formé par

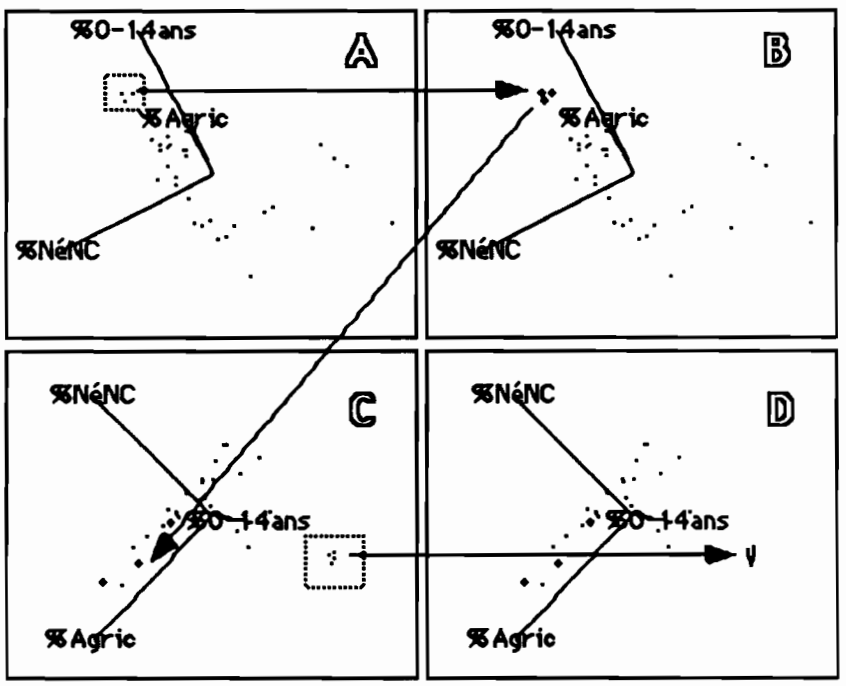


figure n° 4.29. Le marquage de groupes d'individus.

les variables %0-14ANS et %NENC (figure n° 4.29.A), il est beaucoup plus hétérogène lorsqu'on prend en compte la variable %AGRIC.

Les communes de Lifou, Maré et Pouebo qui forment ce groupe sont donc caractérisées par une forte population jeune, par une très grande proportion d'autochtones, et par un pourcentage variable des agriculteurs dans la population active, toujours au-dessus de la moyenne, mais plus fort pour Pouebo, 52.7%, que pour Lifou et Maré, respectivement 43.2% et 30.1%.

Au-delà du simple repérage graphique des groupes, on a souvent besoin de former des sous-ensembles d'observations pour les examiner séparément des autres, de manière à connaître leurs caractéristiques particulières. Dans ce cas, on cherche à définir une nouvelle variable où chaque observation porte le numéro du groupe auquel elle appartient. Ceci permet, par exemple, de calculer la moyenne des groupes pour les trois variables de la toupie. Dans l'exemple ci-dessus, on obtient:

• pour le groupe marqué d'une croix (communes de Lifou, Maré et Pouebo):

%0-14ANS	44.6%
%AGRIC	42.0%
%NENEC	98.9%

• pour le groupe marqué d'un bâton vertical (communes de Nouméa, Mont Dore, Dumbéa et Paita):

%0-14ANS	36.1%
%AGRIC	01.0%
%NENEC	66.4%

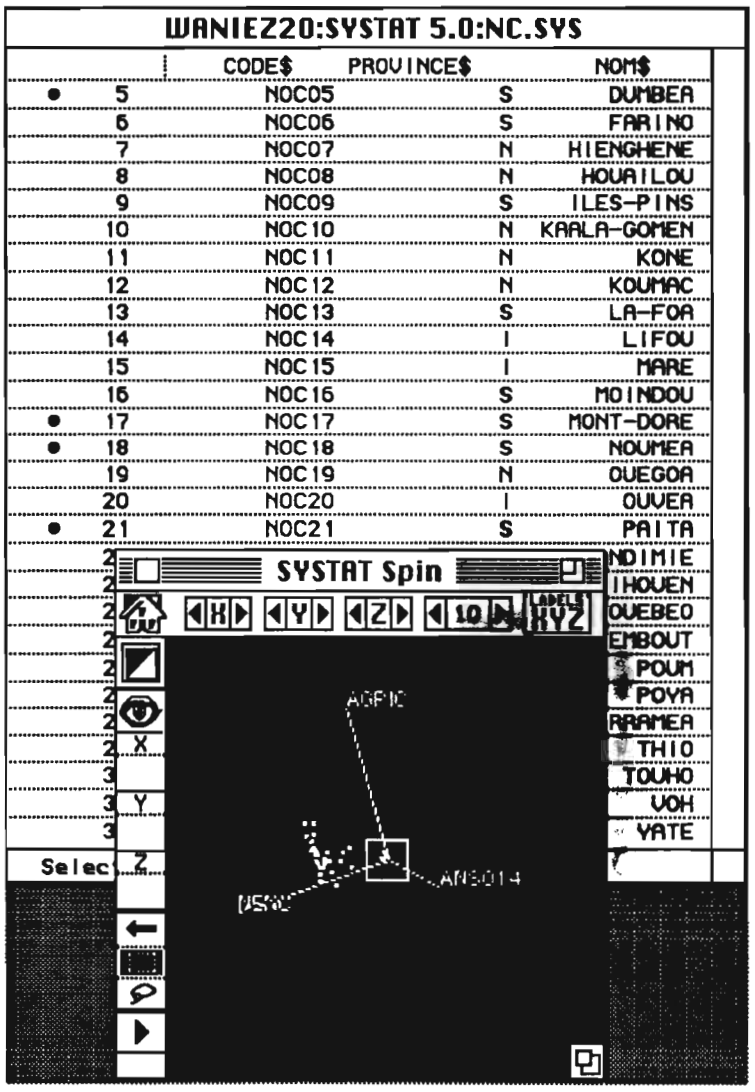


figure n°4.30. SYSTAT: la sélection d'un groupe d'observations et leur marquage dans la fenêtre d'édition.

Bien entendu, à partir du moment où l'on a pu enregistrer le groupe d'appartenance de chaque observation, tous les traitements pouvant être réalisés sur l'ensemble deviennent possibles sur chaque groupe.

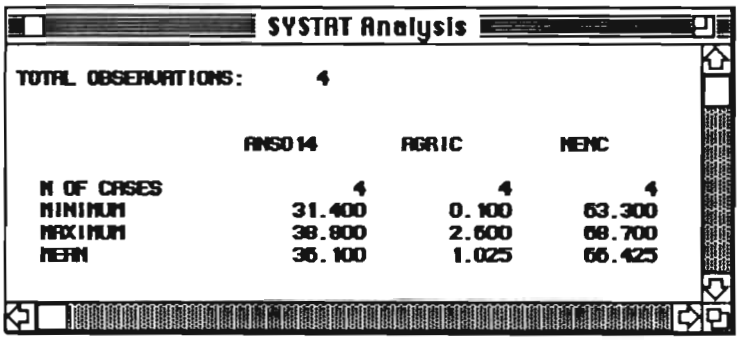


figure n° 4.31. SYSTAT: un résumé numérique calculé sur les 4 observations précédemment sélectionnées.

Ainsi, pour connaître les possibilités des logiciels, il apparaît indispensable de savoir comment:

- on désigne des groupes d'observations;
- on définit de nouvelles variables contenant les numéros de groupes.

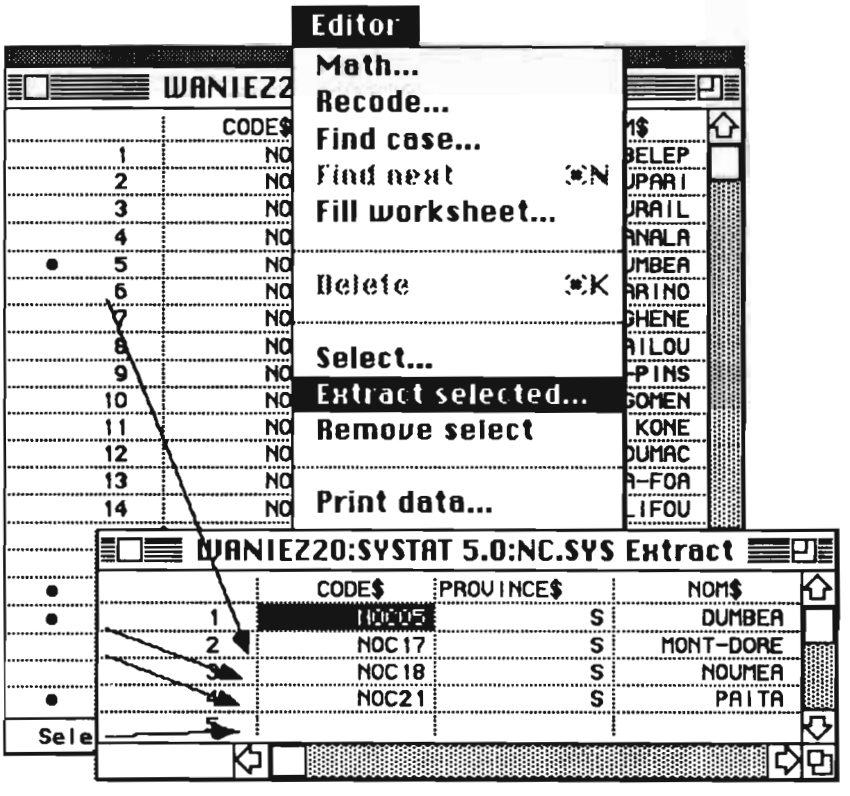


figure n° 4.32. SYSTAT: l'édition du nouveau fichier EXTRACT obtenu par sélection des observations marquées.

4.3.3.1. SYSTAT

Pour interagir avec une toupie réalisée par **SYSTAT**, il faut, comme pour les graphiques bivariés, avoir préalablement ouvert le fichier de données en cochant l'option EDIT. Lorsque le tableau apparaît à l'écran, la toupie peut être tracée de la manière indiquée en 4.3.1. L'option EDIT rend actifs les outils de sélections de points: flèche, carré ou lasso. En sélectionnant sur le graphique, à l'aide du carré, les points formant un groupe, l'éditeur leur affecte le symbole de sélection, • (figure n° 4.30).

A partir du moment où les observations sélectionnées sur le graphique sont marquées dans l'éditeur, tous les traitements à venir ne portent que sur ces seules observations. Par exemple, en choisissant l'article **STATISTICS** du menu **STATS**, on obtient un résumé

statistique qui ne porte que sur les 4 observations retenues (figure n° 4.31).

Mais SYSTAT peut faire plus encore: en activant l'article **EXTRACT SELECTED** du menu **EDITOR**, le logiciel constitue un nouveau fichier contenant toutes les variables du fichier initialement ouvert, mais uniquement les observations sélectionnées. On obtient ainsi un fichier dérivé pouvant être analysé séparément (figure n° 4.32).

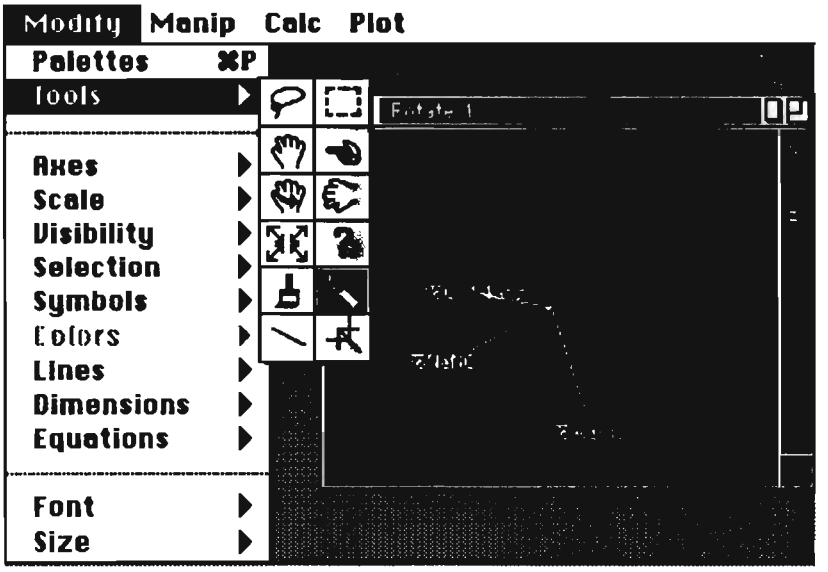


figure n° 4.33. DATADESK: le choix d'un outil de sélection.

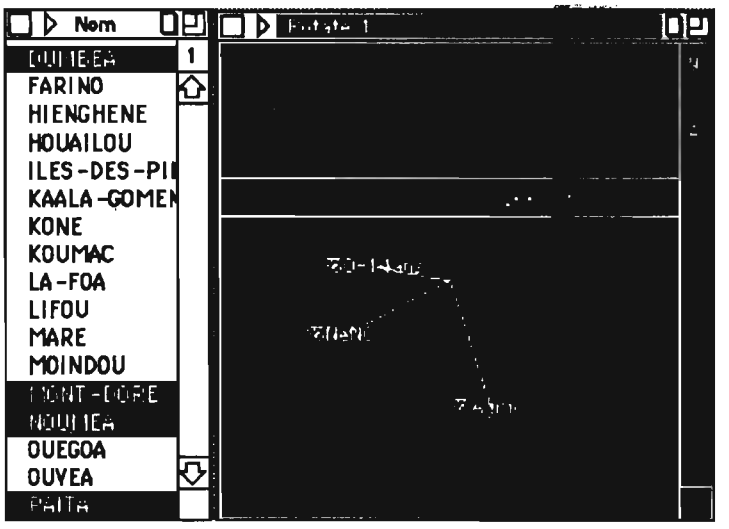


figure n° 4.34. DataDesk: le tranchage de la toupie.

4.3.3.2. DataDesk

Lorsque la toupie est présente sur le bureau de **DataDesk**, la sélection d'une partie du nuage de points se réalise en choisissant l'un des outils de sélection dans le sous-menu **TOOLS** du menu **MODIFY**. Les divers outils utilisables sont le lasso, le carré, le doigt, la brosse et les couteaux (figure n° 4.33).

Ils fonctionnent de la même manière qu'avec les graphiques bivariés. Par exemple, le tranchage du nuage dans le sens vertical a pour effet de sélectionner toutes les observations incluses dans le couloir défini entre le début de l'appui sur le bouton de la souris et son relâchement. Les points des observations sélectionnées apparaissent en double intensité et, grâce aux liens dynamiques entre fenêtres, il est facile de les repérer dans celle d'édition de la variable NOM (figure n° 4.34).

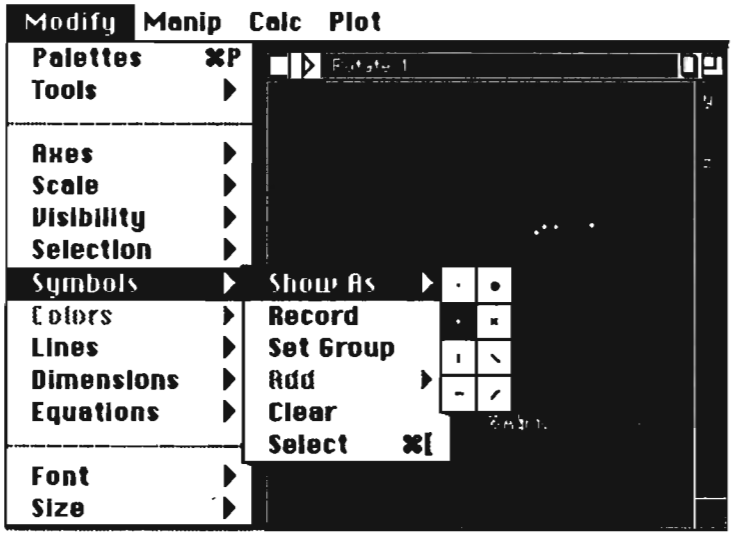
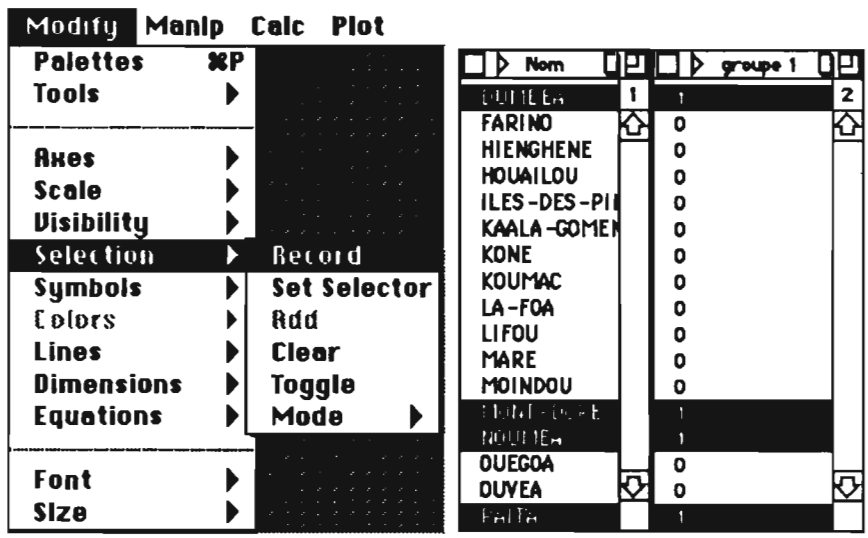


figure n° 4.35. DataDesk: l'affectation d'un symbole particulier à un groupe d'observations préalablement sélectionnées.

Les points sélectionnés peuvent subir divers types d'opérations. En premier lieu, l'article **SHOW AS** du sous-menu **SYMBOLS** du menu **MODIFY** permet d'affecter un symbole au groupe, choisi parmi les huit symboles proposés (figure n° 4.35). Sur un écran couleur, les symboles colorés sont disponibles. Ce marquage particulier facilite le repérage du groupe pendant les rotations ultérieures.

En second lieu, toute sélection de points peut donner lieu au calcul d'une nouvelle variable binaire. Lorsqu'une observation appartient à la sélection, elle prend la valeur 1, sinon, on lui affecte la valeur 0.

Cette opération a lieu lorsqu'on active l'article **RECORD** du sous-menu **SELECTION** du menu **MODIFY**. **DataDesk** demande le nom de la nouvelle variable (ici GROUPE 1) et la place sur le bureau. En ouvrant la fenêtre d'édition par un clic sur l'icône de la nouvelle variable, on peut observer que les observations sélectionnées sur la toupie, dont le nom est souligné dans la fenêtre NOM, prennent la valeur 1, et 0 dans le cas contraire (figure n° 4.36).



=figure n° 4.36. DataDesk: le calcul d'une variable binaire d'appartenance à un groupe.

En utilisant l'article **SPLIT INTO VARIABLES BY GROUP** du menu **MANIP**, **DataDesk** génère un ensemble de dossiers (un dossier par variable à étudier), comprenant chacun deux tableaux, un pour les observations prenant la valeur 0, et

n'appartenant donc pas au groupe, et un autre pour celles du groupe, prenant la valeur 1. Dans chaque tableau ne figure qu'une variable nommée respectivement 0 ou 1, sur laquelle il est possible de poursuivre l'analyse en calculant, mais ce n'est qu'un exemple, un résumé statistique (figure n° 4.37). En renommant les variables, et en regroupant les icônes dans le même tableau, on obtient un nouveau fichier semblable à celui obtenu avec **SYSTAT**.

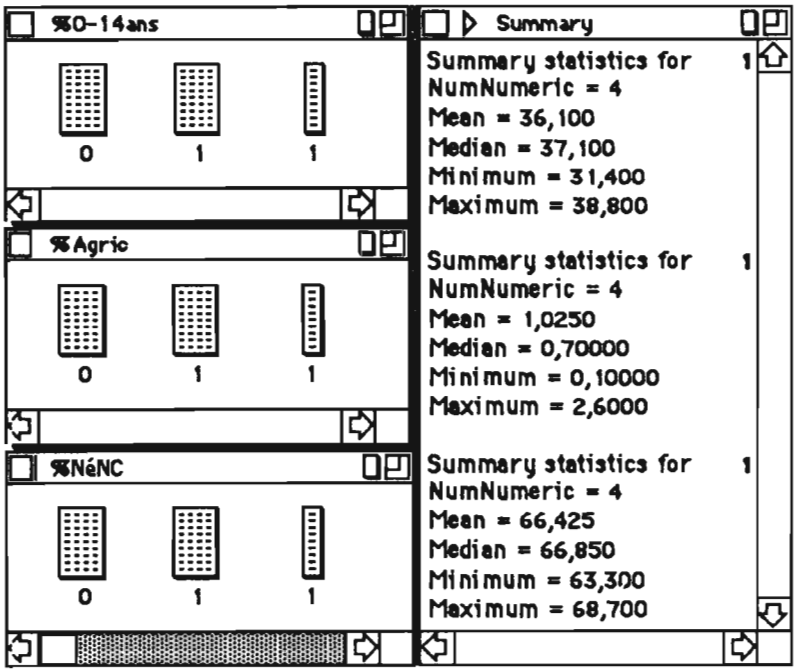


figure n°4.37. DataDesk: le calcul d'un résumé statistique pour les observations du groupe 1, sur les 3 variables de la toupie.

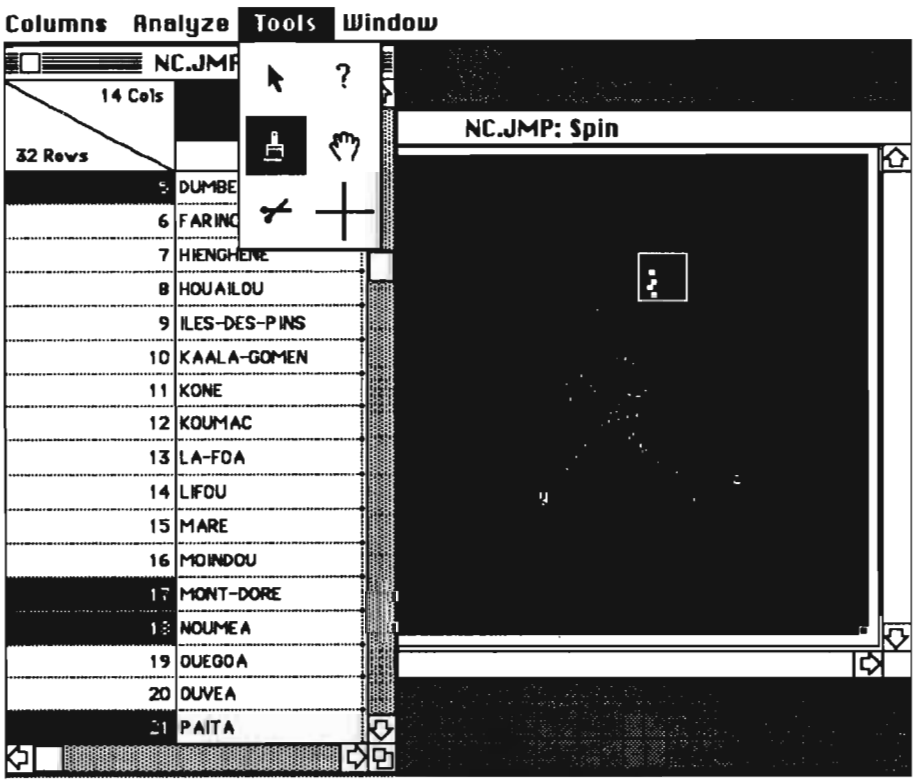


figure n°4.38. JMP: le brossage de la toupie.

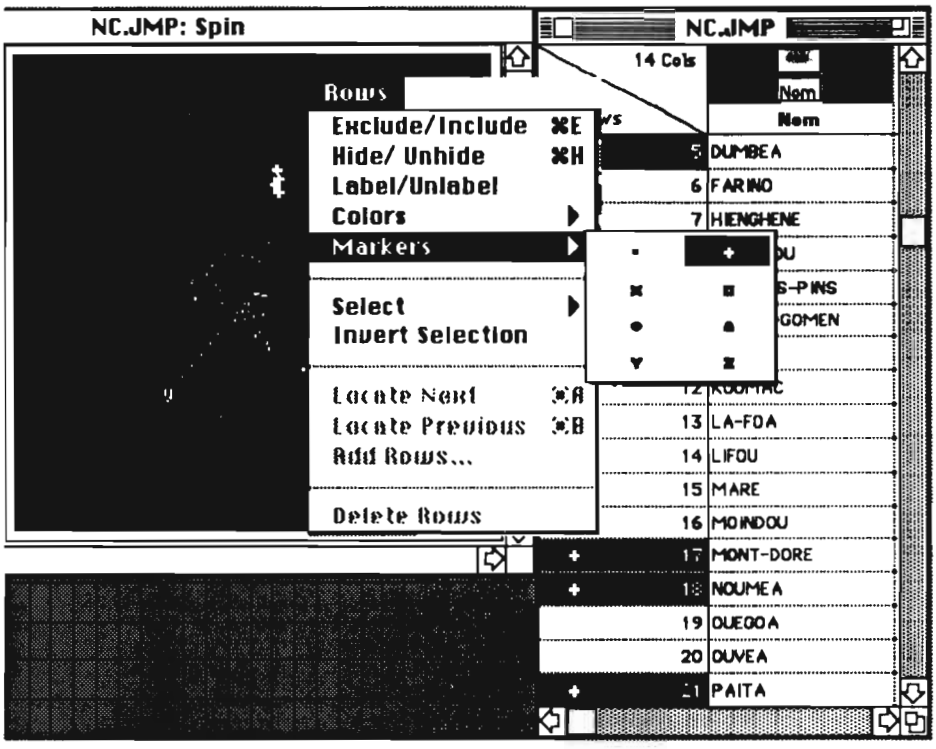


figure n°4.39. JMP: le marquage d'un groupe par un symbole.

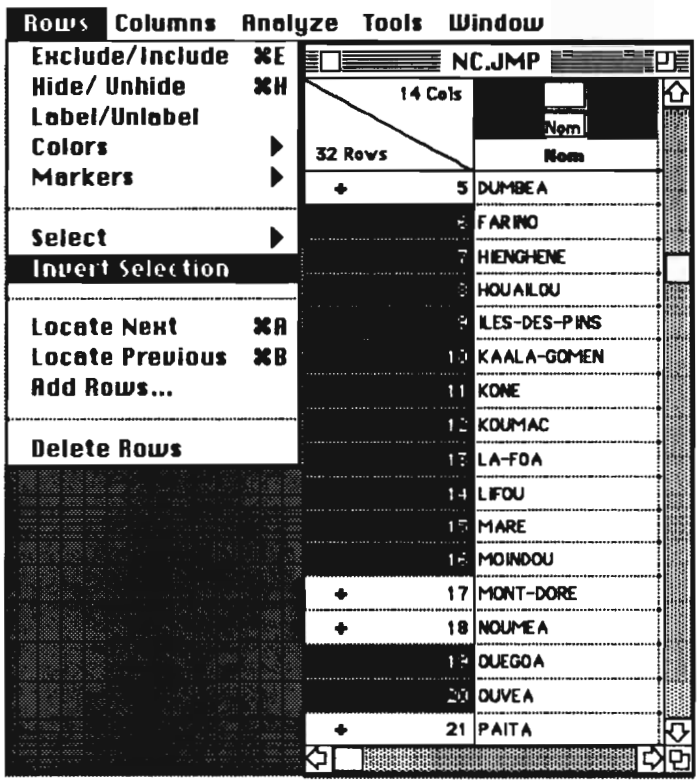


figure n°4.40. JMP: inverser la sélection.

4.3.3.3. JMP

Pour sélectionner un ensemble de points sur la toupie de **JMP**, les outils flèche ou brosse sont utilisables. La sélection provoque une surintensité des points sélectionnées et leur soulignement dans le tableau de données (figure n° 4.38). Dès lors, les observations retenues peuvent faire l'objet de diverses manipulations.

Tout d'abord, il est possible de modifier leur couleur et leur symbole en activant respectivement les articles des sous-menus **COLORS** et **MARKERS** (figure n° 4.39). Le symbole choisi pour le groupe s'affiche alors dans la colonne d'identification des

lignes de l'ensemble du groupe, en même temps que leur soulignement.

Ensuite, il est possible d'exclure soit les individus sélectionnés, soit ceux qui restent. Ceux qui sont exclus ne disparaissent pas du tableau de données, mais sont marqués d'un signe particulier, et seront ignorés de toutes les analyses suivantes, jusqu'à ce qu'ils soient inclus à nouveau. Dans le cas où l'on souhaite travailler uniquement sur le groupe formant la sélection, il faut exclure toutes les observations ne lui appartenant pas. Cela se fait simplement en choisissant successivement les articles **INVERT SELECTION** puis **EXCLUDE** du menu **ROWS** (figures n° 4.40 et 4.41).

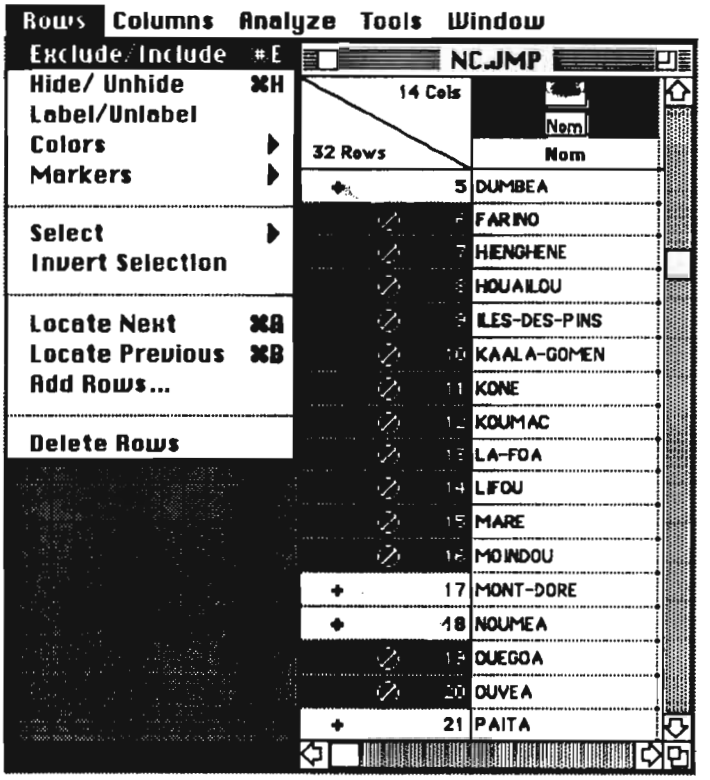


figure n° 4.41. JMP: exclure la sélection.

0-14ans

Quantiles		
maximum	100.0%	38,800
	99.5%	38,800
	97.5%	38,800
	90.0%	38,800
quartile	75.0%	38,500
median	50.0%	37,100
quartile	25.0%	32,700
	10.0%	31,400
	2.5%	31,400
	0.5%	31,400
minimum	0.0%	31,400

Agric

Quantiles		
maximum	100.0%	2,6000
	99.5%	2,6000
	97.5%	2,6000
	90.0%	2,6000
quartile	75.0%	2,1500
median	50.0%	0,7000
quartile	25.0%	0,2250
	10.0%	0,1000
	2.5%	0,1000
	0.5%	0,1000
minimum	0.0%	0,1000

NéNC

Quantiles		
maximum	100.0%	68,700
	99.5%	68,700
	97.5%	68,700
	90.0%	68,700
quartile	75.0%	68,450
median	50.0%	66,850
quartile	25.0%	63,975
	10.0%	63,300
	2.5%	63,300
	0.5%	63,300
minimum	0.0%	63,300

figure n° 4.42. JMP: un résumé numérique résistant calculé sur les trois variables formant la toupie.

Ainsi, seules les observations marquées du signe + demeurent actives pour les calculs ultérieurs, par exemple, un résumé numérique résistant obtenu par la plate-forme **DISTRIBUTIONS OF Y** du menu **ANALYZE** (figure n° 4.42).

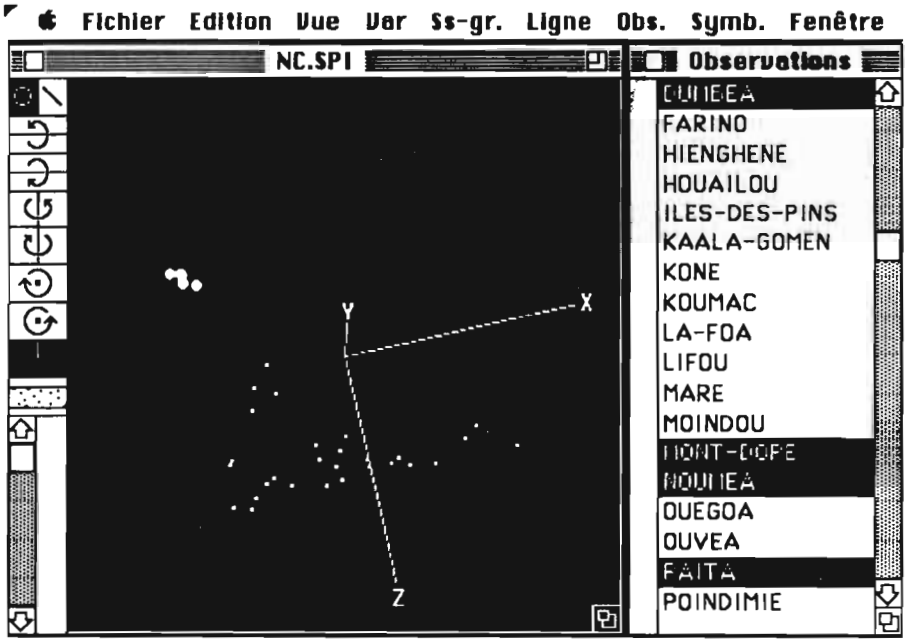


figure n° 4.43. MacSpin: la sélection des observations d'un groupe.

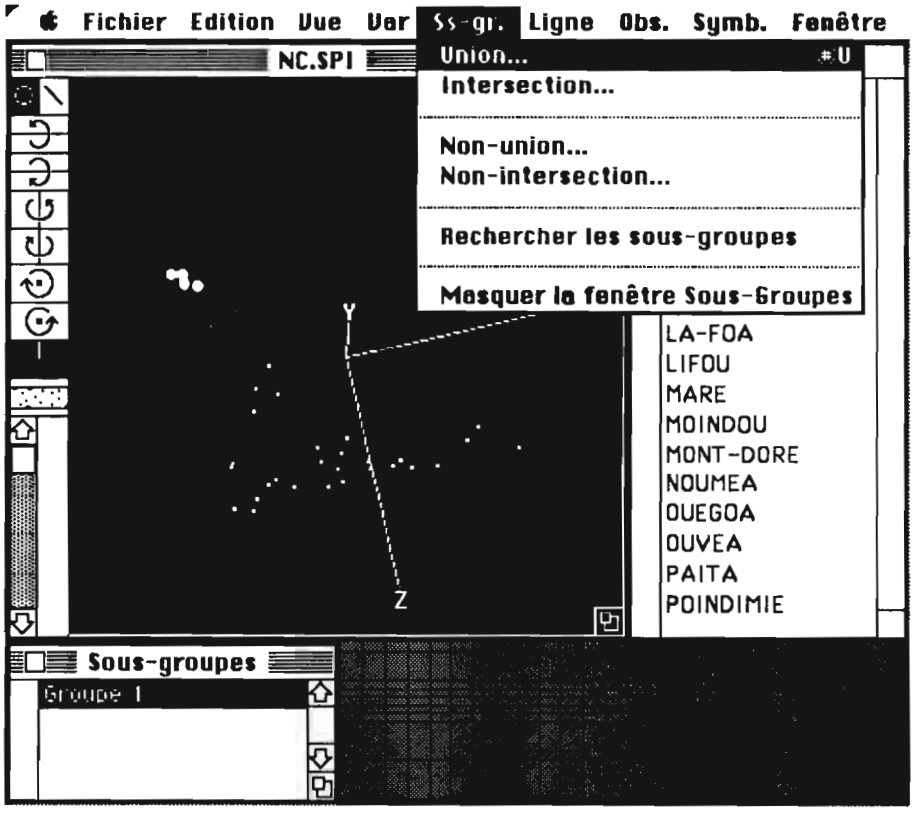


figure n° 4.44. MacSpin: l'identification d'un groupe.

4.3.3.4. MacSpin

Il existe trois méthodes de sélection d'un ensemble de points sur la toupie de **MacSpin**. En cliquant sur chaque observation à retenir, la sélection est soulignée dans la fenêtre OBSERVATIONS (figure n° 4.43). Réciproquement, un clic dans cette fenêtre sur les noms des observations devant composer un groupe le fait apparaître en surintensité sur le graphique. Enfin, par un cliquer-glisser, tous les points du rectangle sont sélectionnés.

Ce n'est pas parce que des points sont sélectionnés qu'ils forment effectivement un groupe. Un groupe existe à partir du moment où les points sélectionnés sont effectivement désignés comme devant former un groupe. L'article **UNION** du menu **SS-GR** réalise cette opération (figure n° 4.44): en l'activant, le logiciel ouvre un dialogue destiné à recevoir le nom du groupe.

A partir du moment où il existe un ou plusieurs groupes de points rassemblés sous le même nom, il apparaît intéressant de réaliser des opérations logiques sur un ou plusieurs de ces groupes, préalablement sélectionnés dans la fenêtre SOUS-GROUPES. **MacSpin** propose 4 opérations dans son menu **SS-GR**:

- l'union sélectionne tous les points

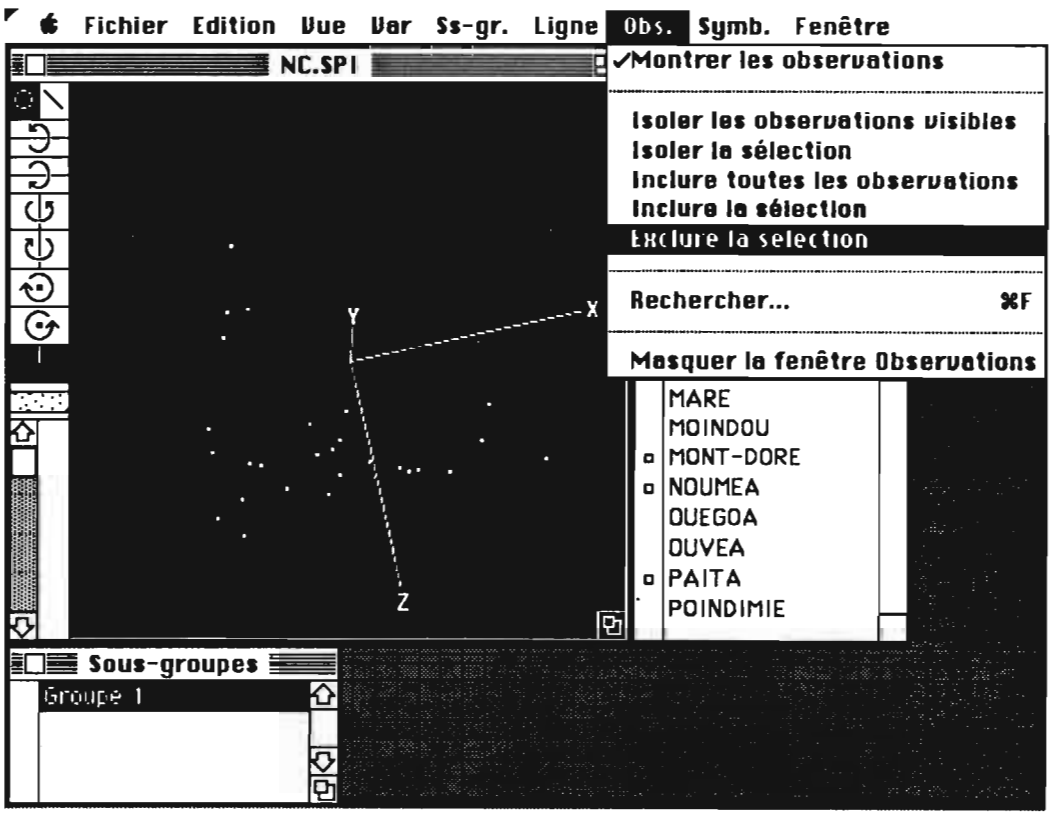


figure n° 4.45. MacSpin: le traitement des observations sélectionnées.

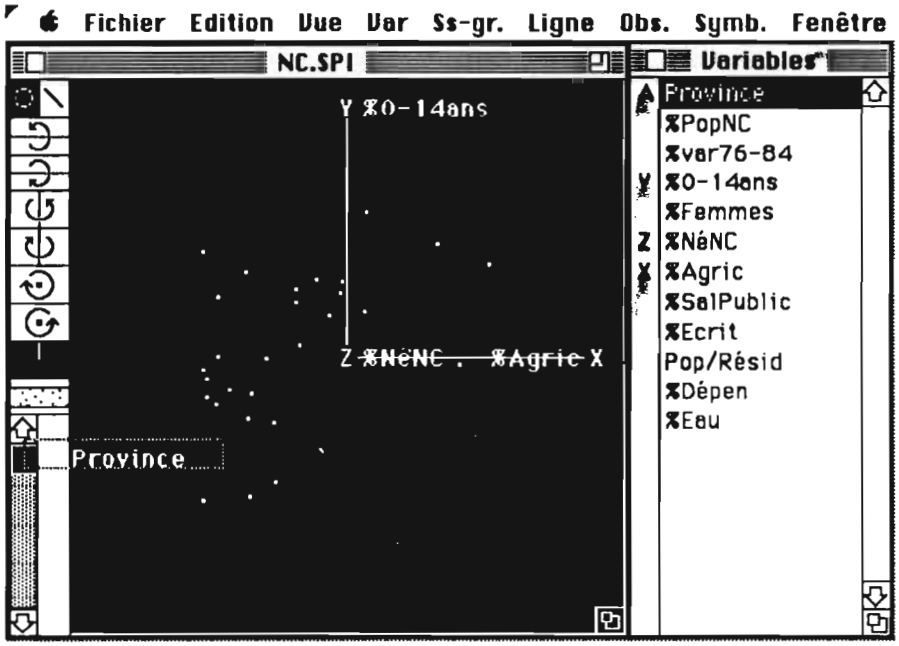


figure n° 4.46. MacSpin: la sélection d'une variable d'animation.

appartenant au moins à l'un des groupes retenus;

• l'intersection sélectionne tous les points communs aux groupes retenus;

• la non-union sélectionne tous les points qui n'appartiennent à aucun des groupes retenus;

• la non-intersection sélectionne tous les points qui n'appartiennent pas, en même temps, à tous les groupes retenus.

Notons qu'une même observation peut figurer dans plusieurs groupes en même temps, ce qui ouvre la porte à l'étude de sous-ensembles flous.

Définir des groupes, leur intersection, leur union, sert avant tout à marquer l'univers sur lequel on souhaite poursuivre l'analyse. Le menu **OBS** a pour fonction de définir ce qui doit advenir des observations sélectionnées (figure n° 4.44). Il est possible de les:

• isoler: seule la sélection figure sur la toupie;

• exclure: seules les observations non-sélectionnées sont représentées sur la toupie.

Notons que les points invisibles à un moment donné, peuvent être réinclus dans la toupie lorsqu'on le désire.

Sur le plan du traitement des groupes, **MacSpin** occupe donc une position originale: l'accent y est mis sur des éléments d'algèbre booléenne. Ceci correspond bien au parti-pris de l'analyse exploratoire: les groupes se font et se défont au gré des hypothèses. Les outils propres à ce logiciel invitent à adopter une démarche pratiquement expérimentale.

4.3.4. Vers l'exploration multivariée

Pour prendre en compte simultanément plus de trois variables dans le cadre de la toupie, il existe deux méthodes différentes: le masquage et la recherche des composantes principales.

4.3.4.1. Le masquage

La technique du masquage correspond au quatrième principe de l'analyse exploratoire: **M** pour **M**asquer. On cherche ainsi à examiner s'il existe une structuration particulière du nuage de points tridimensionnel, en fonction d'une quatrième variable. En quelque sorte, on procède à une exploration quadrivariée, où la quatrième dimension est le temps: les masquages successifs, par l'impression visuelle qu'ils laissent à l'observateur des parties du nuage de points, donnent une profondeur supplémentaire à la lecture de l'information.

4.3.4.2. MacSpin

La technique du masquage est particulièrement bien mise en valeur par **MacSpin**. L'option d'animation, qu'il ne faut pas confondre avec la rotation de la toupie autorise un affichage séquentiel du nuage de points dans l'ordre croissant ou décroissant des valeurs d'une quatrième variable.

Pour mieux exposer cette technique cinématique par essence même, il apparaît préférable d'examiner un exemple. On cherche à savoir si la toupie construite à l'aide du pourcentage d'agriculteurs dans la population active (X), de celui des 0-14 ans dans la population totale (Y) et de la proportion de la population totale née en Nouvelle-Calédonie (Z) présente une structuration particulière en fonction de la province d'appartenance des communes. Le tableau de données contenant une variable PROVINCE, celle-ci peut être utilisée pour l'animation. Il suffit de la sélectionner par un clic sur son nom dans la fenêtre VARIABLES, puis de la faire glisser jusqu'à l'ascenseur situé dans le coin gauche de l'écran (figure n° 4.46).

Le contrôle de l'animation se fait par action sur cet ascenseur: en le faisant descendre jusqu'au plancher, on vide la toupie de l'ensemble de ses points (figure n° 4.47.A); seul le système d'axes demeure. Par un clic sur la flèche supérieure, l'ascenseur remonte d'un étage, c'est-à-dire d'une modalité. La première dans l'ordre retenu pour le codage correspond à la Province Nord. Pour mieux le voir, ce groupe a été marqué du symbole + (figure n° 4.47.B). On observe, de manière très nette, que, dans une position particulière de la toupie, les communes de la Province Nord forme une bande d'orientation gauche-droite. Un nouveau clic sur la flèche supérieure fait apparaître les communes de la Province Sud (figure n° 4.47.C), marquées du symbole •.

De toute évidence, l'orientation haut-bas montre que les communes des deux provinces n'ont pas du tout le même comportement vis-à-vis des 3 variables analysées, même s'il existe une partie commune à leurs deux

nuages. Enfin, un nouveau clic sur la flèche supérieure fait apparaître les communes de la Province des Iles (figure n° 4.47.D), marquées par le symbole □ . Celles-ci sont situées dans la partie supérieure gauche du nuage formé par les communes de la Province Nord et apparaissent comme le prolongement de ces dernières.

En animant la toupie d'un mouvement lié à une quatrième variable, on visualise des sections particulières du nuage de points qui peuvent révéler des éléments de la structuration de l'ensemble. Le masquage accroît donc de manière sensible l'intimité de l'analyste avec son sujet.

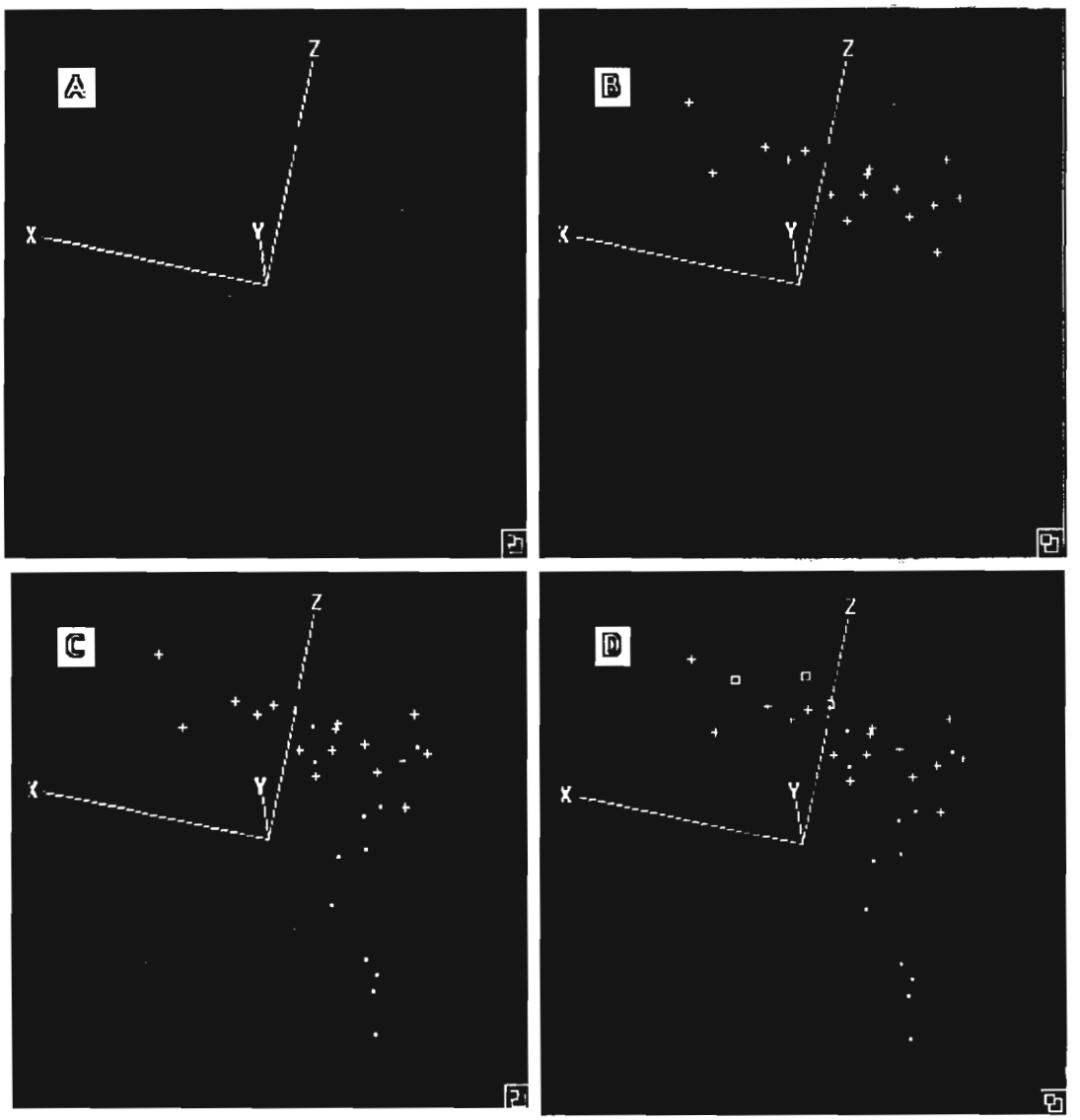


figure n° 4.47. MacSpin: l'animation en fonction dela variable PROVINCE.

4.3.4.3. La toupie et les composantes principales

Retrouver ici l'une des principales techniques de l'Analyse des Données ne doit pas apparaître comme un paradoxe. Rappelons, une fois encore, que l'approche exploratoire ne rejette aucune des techniques classiques d'analyse statistique. Elle les situe dans un environnement permettant de mieux en tirer parti, comme cela a déjà été le cas pour la régression, au chapitre 3.

Le lecteur trouvera dans la bibliographie plusieurs ouvrages traitant de l'analyse en composantes principales (ACP). Il ne semble donc pas nécessaire de se livrer à une paraphrase de ces excellents livres. Rappelons simplement que l'ACP est une méthode de réduction du nombre de variables, méthode factorielle car cette réduction ne se fait pas par simple élimination de variables considérées *a priori* comme peu signifiantes, mais par la construction de nouvelles variables obtenues par combinaison linéaire des variables d'origine. Si ces variables ne sont pas indépendantes linéairement, il est possible de les remplacer par un nombre plus petit de composantes principales.

D'un point de vue géométrique, l'ACP revient à rechercher les axes principaux d'inertie du nuage de points multidimensionnel formé par les observations sur lesquelles on a mesuré un grand nombre de variables. Le premier axe passe par le centre de gravité du nuage de points et le traverse dans la direction de son plus grand allongement: il rend compte de la plus grande dispersion possible. Le second axe, construit perpendiculairement au premier, passe également par le centre de gravité et rend compte, lui aussi, de la plus grande dispersion possible, mais compte tenu de celle dont le premier axe a déjà rendu compte.

Pour connaître l'importance relative de chaque composante, on recourt au pourcentage de la dispersion totale dont chacune d'elle rend compte. Si toutes les P variables étaient linéairement indépendantes, chaque composante représenterait 1/P*100% de la dispersion totale. Dans le cas du fichier de données portant sur les communes de Nouvelle-Calédonie, chaque composante devrait rendre compte de 1/11*100=9.1%. Or, le tableau des pourcentages (figure n° 4.48) montre que la première composante principale représente 37% de la dispersion, la seconde 18.9%, la troisième 13.2%, etc.

EigenValue:	4,0775	2,0786	1,4521	1,1210
Percent:	37,0685	18,8962	13,2010	10,1911
CumPercent:	37,0685	55,9647	69,1657	79,3568

figure n° 4.48. les taux d'inertie des 4 premières composantes principales calculées sur 11 variables relatives aux communes de Nouvelle-Calédonie.

En analysant le nuage de points multidimensionnel (ne pouvant donc pas être matérialisé) dans sa projection sur le plan formé par les deux premiers axes, c'est près de 56% de sa dispersion qui devient visible. L'approche classique consiste à examiner les plans de projection deux à deux. Mais, construisant une toupie avec les 3 premiers axes, la visibilité du nuage de points atteint 69%: l'environnement exploratoire apporte à l'ACP un moyen supplémentaire de lecture de la forme du nuage de points, en permettant de prendre en compte simultanément trois composantes principales, et même quatre si nécessaire, grâce à la technique du masquage.

4.3.4.2.1. JMP

Le fonctionnement de **JMP** donne une bonne idée du parti que l'on peut tirer de l'ACP associée à la toupie.

Avant même de réaliser l'ACP, il faut faire appel à la plate-forme **SPIN** du menu **ANALYZE**. Si les rôles des variables n'ont pas encore été définis, il suffit de sélectionner toutes les variables devant faire l'objet de l'analyse, et non pas seulement trois variables comme précédemment. **JMP** présente alors une toupie construite autour des trois premières variables sélectionnées. Notons qu'il est possible de visualiser les observations en fonction des autres variables sélectionnées au départ: il suffit de sélectionner la lettre X, Y ou Z dans la colonne des variables et de la déplacer par un glissé de souris jusqu'à la case de la nouvelle variable (figure n° 4.49).

L'analyse en composantes principales est une option de la toupie à laquelle on accède par le premier menu *pop-up* situé dans la partie inférieure gauche de la fenêtre SPIN (figure n° 4.50).

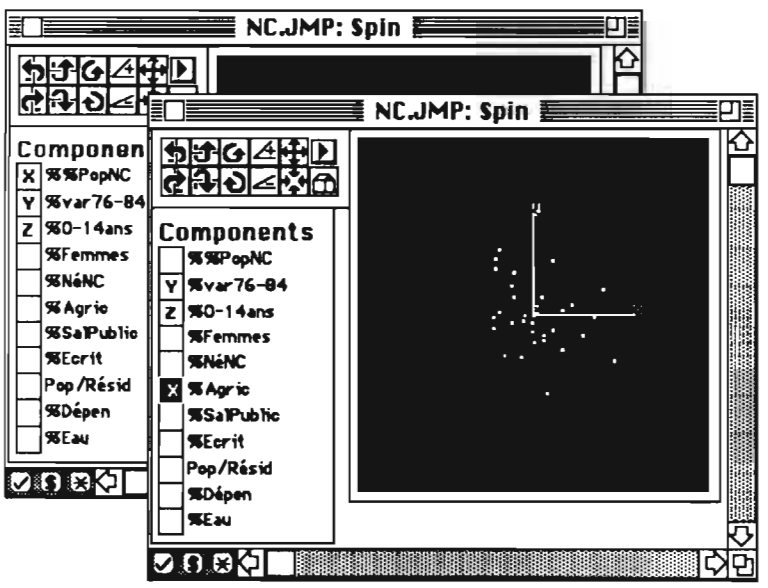


figure n° 4.49. JMP: la sélection de trois variables dans une liste.

JMP réalise l'ACP sur l'ensemble des variables sélectionnées après l'activation de la plate-forme SPIN, ici, sur 11 variables. Le calcul dure quelques secondes et les axes factoriels sont tracés dans le système d'axes formé par les trois variables constituant la toupie avant l'utilisation de l'article **PRINCIPAL COMPONENTS**. En faisant glisser la lettre X jusqu'à la case PRIN COMP 1, la lettre Y jusqu'à PRIN COMP 2 et Z jusqu'à PRIN COMP 3, on construit une toupie dont les axes sont les

composantes principales avec des points localisés en fonction des coordonnées des observations sur ces axes factoriels (figure n° 4.51).

Dans l'espace factoriel, JMP trace les rayons formés par les traces des axes d'origine que sont les variables de l'analyse. La longueur apparente d'un rayon rend compte de la contribution de la variable qu'il représente à l'espace factoriel. Autrement dit, plus le rayon est long, plus la variable d'origine joue un rôle important.

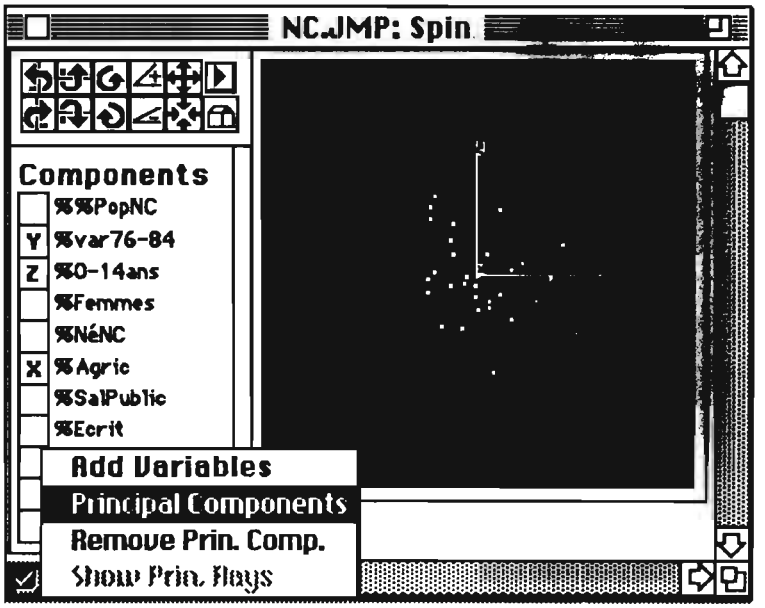


figure n° 4.50. JMP: l'activation de l'article PRINCIPAL COMPONENTS.

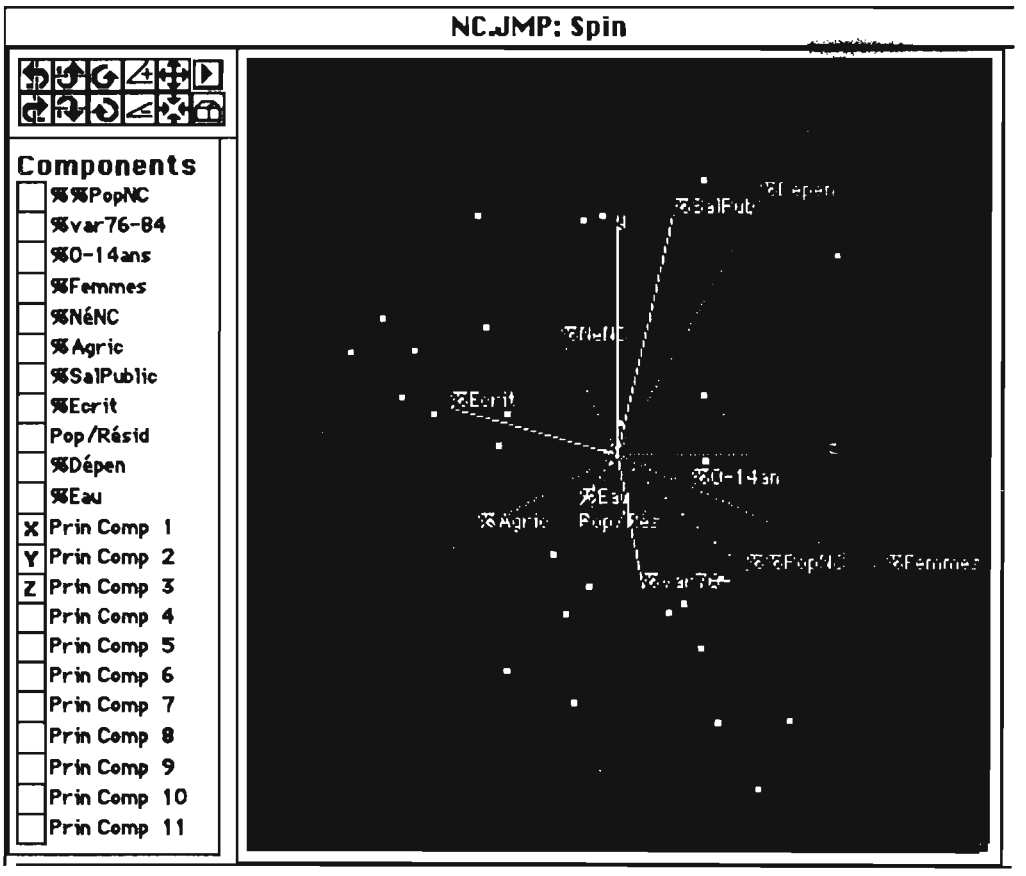


figure n° 4.51. JMP: le plan des axes n°2 et n°3.

Les angles formés par les rayons et les axes factoriels montrent le degré de concordance entre ces facteurs et les variables d'origine. Plus la valeur de l'angle tend vers 0° ou de 180°, plus l'axe correspond à la direction de la variable d'origine; inversement, plus l'angle est droit, moins l'axe factoriel représente la variable d'origine. Ainsi, plus les rayons s'alignent sur les axes, mieux ces derniers rendent compte des variables. La rotation de la toupie doit donc être guidée par le souci d'obtenir des projections sur lesquelles les axes factoriels tendent vers l'alignement avec les rayons des variables d'origine. Les liens dynamiques entre les fenêtres, les outils de brossage et de tranchage permettent ensuite d'examiner les positions des observations dans l'espace trifactoriel et d'avancer des explications sur la structuration du nuage de points multidimensionnel.

CONCLUSION

L'approche exploratoire pour l'analyse des données statistiques se compose de deux volets complémentaires formés, d'une part, de techniques spécifiques et d'autre part, d'un environnement informatique particulier. Les techniques spécifiques ont été développées par J.W. Tukey et ses élèves. Elles mettent l'accent sur la visibilité de l'information statistique par des graphiques originaux, diagramme en tige et feuilles, diagramme en boîte et moustaches, toupie, et par des résumés statistiques résistants basés sur la médiane et les quartiles, plus que sur la moyenne et l'écart-type. A sa manière, l'approche exploratoire réhabilite les graphiques des «ingénieurs» d'autrefois, en faisant de l'écran de l'ordinateur le moyen d'une «hyper-vision» nécessaire aux décideurs d'aujourd'hui.

Mais l'approche exploratoire, qui ne se limite pas aux représentations graphiques, peut être étendue à toute la statistique «classique» ainsi qu'à l'Analyse des Données. Au lieu d'utiliser ces techniques comme des «boîtes noires» ou des «moulins à statistiques», comme cela est encore très souvent le cas, le point de vue exploratoire invite à regarder les données et les résultats, pas à pas, afin de vérifier en permanence si les analyses sont effectuées dans les conditions d'application optimales. Cette manière de procéder garantit l'analyste contre les problèmes qui ne peuvent manquer de surgir lorsqu'on traite une information avec une méthode inadaptée.

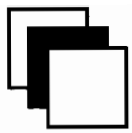
Ces deux volets de l'approche exploratoire présentent un commun dénominateur: l'informatique. Ainsi, le bon usage d'un logiciel d'analyse statistique, d'un statisticiel, semble désormais aussi important que la connaissance des procédés de calcul. On peut même avancer sans risque que l'ordinateur s'avère être un excellent moyen d'approche expérimentale de la statistique auprès des non-mathématiciens. Il semble donc utile, pour conclure, de souligner les points forts de chacun des logiciels, d'exprimer, en quelque sorte, des critères de choix.

• **SYSTAT** est sans doute le plus complet. Son caractère «encyclopédique» en fera le fidèle compagnon des statisticiens professionnels. Grâce à son langage de programmation, il comblera tout ceux qui souhaitent réaliser des traitements répétitifs sur des ensembles de données différents. L'approche exploratoire y apparaît comme surimposée: ceci se traduit par une interactivité insuffisante entre les données, les tableaux numériques et les graphiques. Notons enfin que la vitesse de chargement et de réaction sur Macintosh SE n'est pas grande au regard des 4 mégaoctets obligatoires pour faire fonctionner ce logiciel.

• **DataDesk** apparaît en parfaite adéquation avec les techniques et l'approche exploratoires. Odesta Corporation qui le fabrique, a mis de son côté un atout décisif en consultant P.F. Velleman dont l'expérience en matière d'analyse exploratoire fait autorité aux Etats-Unis. La principale originalité de ce logiciel réside sans doute, dans les liens dynamiques généralisés entre les fenêtres complétés par les menus *hyperview*. De ce fait, l'interactivité autorise une véritable navigation dans les structures numériques pour en découvrir toutes les facettes, même les moins accessibles. On ne peut manquer d'être impressionné par un tel chef-d'œuvre de génie logiciel. Mais la multiplication des fenêtres ouvertes sur le bureau s'avère, à l'usage, un peu encombrante sur le petit écran des Macintosh Plus, SE, SE/30 et Classic. Pour exploiter toute la souplesse (*versatility*) du logiciel, un écran de grande taille (A4 ou A3) et doté de la couleur (comme celui de la gamme des Macintosh II) apparaît souhaitable.

• **JMP**, dernier né des logiciels analysés ici présente de nombreux atouts au premier rang duquel il faut placer l'intelligence. Non pas que les autres en soient dénués, mais l'intelligence dont il est fait état ici a trait aux plate-formes d'analyse, véritables «postes d'aiguillages» et garde-fous contre une utilisation abusive des techniques statistiques. Obliger l'utilisateur à définir le type de chaque variable et le rôle qu'elle tient dans l'analyse réduit de manière considérable les risques d'erreur de choix d'une méthode. Une telle qualité conduit à recommander ce logiciel pour l'enseignement de la statistique, et cela d'autant plus qu'il existe une version limitée à 500 valeurs, nommée JMP-IN. Très complet sur le plan des méthodes statistiques (mais la classification fait néanmoins défaut), ce statisticiel bénéficie des liens dynamiques entre fenêtres, des menus *pop-up*, et d'une excellente gestion des données incluant leur portabilité vers SAS.

• **MacSpin** s'adresse à tous ceux qui souhaitent expérimenter la toupie, sans avoir accès aux autres techniques d'analyse, mais en profitant de quelques spécificités de ce logiciel comme, par exemple, l'étude booléenne des groupes.



BIBLIOGRAPHIE

1. Analyse exploratoire

Tukey J.W. (1977), *Exploratory Data Analysis.* Reading, MA.: Addison-Wesley Publishing Company, 688 p.

Il s'agit du traité fondamental de l'analyse exploratoire, par son propre inventeur. On y trouve l'exposé des méthodes de représentations graphiques, de résumés numériques résistants, de lissages, etc. Il n'y est fait aucune mention des moyens informatiques utiles.

Hartwig F., Dearing B. (1982), *Exploratory Data Analysis.* Sage University Paper, Series: Quantitative Applications in the Social Sciences, n°16. Beverly Hills, CA, Sage Publications, 83 p.

Ce petit livre d'initiation explique comment utiliser les principales techniques de l'analyse exploratoire dans le domaine de la science politique. Un glossaire facilite la familiarisation du lecteur à la terminologie propre à l'EDA.

Jambu M. (1989), *Exploration Statistique et informatique des données.* Paris: Dunod, Col. Dunod informatique, 506 p.

Issu d'un enseignement d'analyse des données à l'Institut National des Télécommunications, cet ouvrage demande, pour être intégralement compris, un niveau supérieur en mathématiques. Cependant, l'exposé de certaines méthodes exploratoires ne nécessite pas une telle spécialisation. Les aspects informatiques n'y sont que peu étudiés. Il s'agit, par ailleurs, d'un excellent traité d'analyse des données, par l'un des grands spécialistes français de la question.

Velleman P.F. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis.* Boston, MA: Duxbury Press.

L'analyse exploratoire exposée par l'un des créateurs de DataDesk.

Charre J., Dumolard P. (1989), *Initiation aux pratiques informatiques en géographie.* Paris, Masson, Col. Géographie, 199 p.

Destiné aux étudiants de géographie des premier et second cycles universitaires, cet ouvrage montre comment analyser des données statistiques relatives à un espace géographique. Tout en insistant sur les conditions d'application

et sur les limites de chaque technique, les auteurs y développent une méthode pré-exploratoire fondée sur un logiciel de leur fabrication nommé INFOGEO.

Bertrand R., Valiquette C. (1986), *Pratique de l'analyse statistique des données.* Québec: Presses de l'Université de Québec, 379 p.

Cet excellent ouvrage d'initiation à l'analyse des données statistiques en sciences humaines est l'un des rares à reprendre, en langue française, les exposés fondamentaux de Tukey. Les auteurs y expliquent, dans un langage accessible aux non-mathématiciens, les différences fondamentales entre analyse exploratoire et analyse confirmatoire. L'informatique n'y est malheureusement qu'évoquée, surtout pour insister sur les risques qu'elle est supposée faire courir à ses utilisateurs.

2. Statistique et Analyse des Données

Bouroche J.M., Saporta G. (1980), *L'Analyse des Données.* Paris: Presses Universitaires de France, Col. Que sais-je? n°1854, 125p.

Cosinschi M., Waniez P. (1990), *Les Statisticiels.* Icônes, n°21, Février/ Mars 1990, pp. 59-83

Cosinschi M., Waniez P. (1990), *Pratique de l'analyse statistique, SAS sur PC/PS, mini et gros systèmes.* Montpellier: GIP RECLUS, Col. RECLUS modes d'emploi, n°15, 175p.

Fenelon J.P. (1981), *Qu'est-ce que l'Analyse des Données.* Paris: Editions Lefonen (26 rue des Cordelières, 7501), 311p.

Sanders L., Durand-Dastès. (1985), *L'effet régional: les composantes explicatives dans l'analyse spatiale.* Montpellier: GIP RECLUS, Col. RECLUS Modes d'emploi, n°4, 47 p.

Sanders L., (1990), *L'analyse des données statistiques en géographie.* Montpellier: GIP RECLUS, Col. Alidade, 250 p.

Tomassone R., Lesquoy E., Miller C. (1983), *La régression, nouveaux regards sur une ancienne méthode statistique.* Paris, Masson, 180 p.

Adresses utiles

SYSTAT est une marque déposée de SYSTAT Inc.. Importateur en France: Statilogie, 40 rue du Colonel Pierre Avia, 75015 Paris.

DataDesk est une marque déposée de Data Description Inc.. Hyperview est une marque déposée de Data Description. ce logiciel est diffusé aux Etats Unis par Odesta Corporation. Importateur en France: Alpha Systèmes Diffusion, Miniparc ZIRST Grenoble Meylan, 43 chemin du Vieux Chêne, 38240 Meylan.

JMP est une marque déposée de SAS Institute Inc.. Importateur en France: SAS Institute, 50 avenue Daumesnil, 75012 Paris.

MacSpin est une marque déposée de D^2 Software. Importateur en France: Bruno Rives & Associés, 6 avenue Franklin Roosevelt, 75008 Paris.

GLOSSAIRE

Boîte et moustaches (*Box and whiskers*): résumé graphique d'une distribution statistique la médiane, les pivots ainsi que les valeurs exceptionnelles.

Brossage (*Brushing*): opération de définition de groupes disjoints d'observations par regroupements sur un graphique bivarié représentant une relation.

Diagramme en tige et feuille (*Stem and leaf plot*): figure ressemblant au diagramme à bâtons de la statistique classique. Ici, les bâtons sont formés par un empilement de chiffres correspondant au dernier chiffre de la valeur chaque individu sur la variable dont le diagramme représente la distribution.

Direction (*direction*): dans une relation, la direction est dite positive lorsque qu'aux fortes valeurs de l'une des deux variables correspondent les fortes valeurs de l'autre. Respectivement, la direction est dite négative lorsque qu'aux fortes valeurs de l'une des deux variables correspondent les faibles valeurs de l'autre.

Droite de Tukey (*Tukey line*): méthode d'ajustement linéaire résistante. La ligne droite, joignant d'abord les valeurs médianes des premier troisième tiers des observations, est ensuite déplacée sur le graphique bivarié de manière à ce que la moitié des observations apparaissent au-dessus de la ligne, et l'autre moitié au-dessous.

Etendue (Spread): mesure de la dispersion des individus sur une variable. En statistique classique, la dispersion est mesurée par la variance; on préfère, en analyse exploratoire, l'intervalle interquartile, plus robuste. Ne pas confondre avec l'étendue (*Range*) qui, en statistique classique correspond à la différence entre la valeur maximale et la valeur minimale.

Exception (*Outlier*): observations situées au-dessous ou au-dessus des pourcentiles 10% et 90%. En analyse exploratoire, les exceptions font l'objet d'une attention particulière due au scepticisme de l'analyste.

Forme (*Shape*): dans une relation, ligne droite ou courbe figurant l'allure générale du lisse.

Graphique bivarié (*Scatterplot*): figure sur laquelle chaque observation est représentée par un point sur un plan dont les axes sont deux variables. Ce type de graphique est très pratique pour détecter toutes sortes de relations, linéaires ou non.

Graphique des résidus (*Residual plot*): graphique bivarié présentant le rugueux d'une relation. L'axe des abscisse figure la variable endogène et celui des ordonnées les résidus de l'ajustement.

Intensité (*Intensity*): dans une relation, degré de correspondance du lisse avec les valeurs d'origine.

Lisse (*Smooth*): structure simple et sous-jacente d'une relation. Elle présente en particulier la forme et la direction de cette relation.

Localisation (*Location*): valeur qui résume le mieux l'ensemble des données. A la moyenne arithmétique de la statistique classique, on préfère, en analyse exploratoire, la médiane.

Ouverture (*Openness*): le premier des deux grands principes de l'analyse exploratoire. L'analyste doit aborder les données sans modèle a priori, et en ne cherchant de relation entre les variables qu'à partir des variables elles-mêmes.

Pivot inférieur ou supérieur (*Lower ou upper hinge*): Valeur en-dessous de laquelle (respectivement au-dessus de laquelle) sont situées un quart des observations. En statistique classique, ces pivots prennent les noms de premier et troisième quartile.

Relation linéaire (*Linear relationship*): lorsqu'une relation peut être résumée par une ligne droite, elle est dite linéaire. En statistique classique, l'ajustement correspond à une droite de régression; on préfère, en analyse exploratoire, la droite de Tukey, plus robuste.

Relation monotone (*Monotonic relationship*): relation dont la direction est toujours soit positive, soit négative.

Relation non-linéaire (*Non-linear relationship*): lorsqu'une relation ne peut être résumée par une droite, elle est dite non-linéaire. Une relation est intrinsèquement non-linéaire lorsqu'après une série de transformations (logarithmes, etc.), il demeure impossible d'en rendre compte par une droite. L'analyse exploratoire met l'accent sur la détection des relations non-linéaires.

Résidus (*Residuals*): différences entre les valeurs observées et les valeurs estimées par une relation. Ils expriment la partie rugueuse d'une relation.

Résistance (*Resistance*): qualité de certaines techniques de mesure de la localisation, de l'étendue ou de l'intensité d'une relation qui s'exprime par une relativement faible influence des exceptions.

Rugueux (*Rough*): dans une relation, déviation par rapport au lisse, plus communément appelé résidu.

Scepticisme (*Skepticism*): le second des deux grands principes de l'analyse exploratoire. L'analyste doit toujours s'interroger sur la valeur des indicateurs qu'il retient et considérer les situations «exceptionnelles» comme «significatives».

Toupie (*Spining top*): graphique trivarié. En faisant pivoter ce graphique autour de l'un de ses axes, on observe des configurations particulières du nuage de points formé par les observations. Certaines formes significatives peuvent ainsi êrte mises en évidence.

Tranchage (*Slicing*): opération de définition de groupes disjoints d'observations par définition d'intervalles sur l'un des deux axes du graphique bivarié représentant une relation.

Transformations (*Reexpressions*): s'applique aux variables numériques pour résoudre, au moins partiellement, les problèmes liés à la non-normalité des distributions ou la non-linéarité des relations. En générale, une transformation revient à appliquer une fonction mathématique (log, etc.) aux valeurs d'origine.

INDEX

COLLECTION RECLUS MODES D'EMPLOI

1. Observatoire de la dynamique des localisations. Création de l'information. Manuel pour l'emploi du bordereau d'enregistrement des données (avril 1985).

2. Pour la Géographie Universelle, charte de la rédaction (juillet 1985).

3. Thérèse SAINT-JULIEN, La diffusion spatiale des innovations (septembre 1985).

4. Lena SANDERS, François DURAND-DASTES, L'effet régional: les composantes explicatives dans l'analyse spatiale (novembre 1985).

5. Robert FERRAS, L'Espagne, écritures de géographie régionale (décembre 1985).

6. Colette CAUVIN, Henri REYMOND, Nouvelles méthodes en cartographie (janvier 1986).

7. Jean-Paul VOLLE, Bulgarie: les Systèmes de peuplement (février 1986).

8. André DAUPHINÉ, Jean-Yves OTTAVI, Atlas structurel des climats de la France (mai 1986).

9. Colette CAUVIN, Henri REYMOND, Abdelaziz SERRADJ, Discrétisation des données et représentation cartographique (mars 1987).

10. Anne-Marie LAKOTA, Christian MILELLI (coord.), Emplois, entreprises et équipements en Ile-de-France (avril 1987).

11. Fernand JOLY, Carte géomorphologique de la France au 1:1 000 000, quart Nord-Ouest (juin 1987).

12. André DAUPHINÉ, Christine VOIRON-CANICIO, Variogrammes et structures spatiales (février 1988).

13. Fernand JOLY, Carte géomorphologique de la France au 1:1 000 000, quart Nord-Est (novembre 1988).

14. Robert FERRAS, Les Géographies Universelles et le monde de leur temps (mai 1989).

15. Micheline Cosinschi, Philippe Waniez, Pratique de l'analyse statistique SAS sur PC/PS, mini et gros systèmes (juillet 1989).

16. Anne-Marie LAKOTA, Christian MILELLI (coord.), L'Ile-de-France en mouvement, colloque de novembre 1989 (mai 1990).

17. Philippe WANIEZ, L'analyse exploratoire des données (avril 1991)

Prix: 120 F