Gilles Santini
IMS France
Paris, France

8.5 Clustering techniques

METHODOLOGY: THE STATE OF THE ART

Very many different kinds of clustering techniques have flourished in the past. Most of them were designed to deal with well behaved quantitative data such as those enjoyed by lucky statisticians who work in the fields of physics or biology. Doomed to work with media-related data we have had to leave the shores of statistical paradise some years ago. Since then we have learnt quite a few things about the difference between theory and practice.

(1) Clustering methods *are* useful: this technique provides the researcher with a powerful tool for constructing analytic grids. These are useful when investigating large amounts of multidimensional data or tracking qualitative trends over time.

(2) But clustering algorithms can produce almost anything, including nonsense, from the data: clustering methods cannot be restricted to a black box algorithm. They require user-controlled step by step analysis of the data and some experience in mastering the various statistical methods.

(3) Results should be robust against three statistical disasters: redundancy (dependence) or the variables; scale effects of the phenomenon measured; outliers amongst the observations.

(4) The clusters produced (types) should obey two principles: parsimony — we want the right number of types; stability — we cannot accept that the constructed types be sensitive to small parts of the data.

To meet such requirements, in the past couple of years we have put together and used a clustering method called TRF which proved to have some value. It has been used on data sets containing over 300 variables and 15,000 cases.

Phase one

(1) First the method constructs a factorial base using Correspondence Analysis, a technique that wipes out random noise and weakens outliers. This technique is especially interesting for our purpose since it is highly insensitive to redundancy within the set of the variables and to differences in frequency among the attributes of the cases measured.

(2) The rank of this base (generally around 10) is fixed according to the speed of decrease of the amount of inter-points variance drawn by each additional axis. All cases are projected in the resulting factorial base. From then on a case is described by its co-ordinates in the factorial base (ie. it is related to the original data by the value of several linear contrasts built on the original variables — as many linear combinations as axes in the base).

Phase two

(1) In the space of the factorial base the data can be represented as a constellation of points (each point is a particular case). We like to call it the 'typological soup'. Our first aim is to find where this soup is thickest (in other words we look for the lumps). This can be done using DIDAY's statistical algorithm. It first creates clusters by simple allocations to centroids. This creates a first partition of the set of cases and then produces several such partitions by changing the original random seeds of the centroids. The process continues by looking at all cells of the combined partition and putting together cases that are in the same cells. The number of granules produced is decided by rejecting all those whose size is less than a few % of the total number of cases. Rejected cases are put aside until the end of the clustering process, since they are in a low density region of the factorial base and could disturb the next aggregation process.

(2) The centre of each granule is computed and placed in the factorial base. The user should check that granules are spread more or less over the entire space.

(3) A hierarchical tree of possible partition of the granules is computed. Cutting the tree at various levels produces different numbers of sets of granules — we call these sets primary types. Looking at the gain of a statistical measure can help the user to decide what is the right number to choose. Once such a decision is taken all cases belonging in the same primary type are merged and the centre of each primary type is computed.

Phase three

(1) Each case is allocated to one of the primary types on the base of the closest-distance criterion. New types (secondary types) are thus constructed. The amount of migration from primary types is checked. If too much migration is observed allocation is iterated until a more stable solution is reached.

(2) Isolated cases are allocated to the resulting secondary types. The result produce the final types.

Phase four

(1) The original variables are tabulated for each of the final types. Vertical and horizontal percentage are drawn. Measures of association (Phi-square) between variables and types can help in interpreting the nature of the final types.
(2) Complementary variables can be used to describe the final types. Graphs are a must for such a description.

APPLICATIONS

Readership approach: "Tell me what you read and I will tell you who you are."

Several recent cases have proved to be efficient in the analysis of media communication using typologies of respondents built out of reading habits alone. Such attempts offer an alternative to attitude and behavioural items, the limitations of which have raised some controversy in France.

A French government agency, the C.E.O (Centre for Opinion Studies), is in charge of measuring the level of audience and satisfaction of TV viewers. To explore the general public observed reactions, each show is analysed using types previously constructed on the basis of magazine readership, the underlying idea being that viewers filter out aspects of the TV shows which are not within the scope of their cultural background. People with similar cultural background are believed to be in the same type.

MEDIAS, a professional magazine asked us recently to cluster press readers in homogeneous groups. Such a study has been conducted in collaboration with Gilles Boisson from Intermarco Conseil. The main results are outlined in **Tables 1** and **2**

Multi information data bases: Some results from the CESP 'Media-Product' study.

Clustering techniques have been used by the CESP to merge samples and deliver homogeneous data base to their subscribers. Furthermore the mass of printed output was so huge (11 volumes) that most media and agencies were interested mostly in general or sectorial syntheses. For example, we recently performed a typology on all women in the data base (7,906) using 308 consumption variables. The study was done for a women's magazine who used it to develop a promotional brochure. But we can additionally identify the levels of readership within types of their direct competitors, and point out how discriminating the various clusters can be.

TABLE 1
What the different groups read

THE STRONG-WILLED 8%	
42% read	Le Monde
33%	Le Nouvel Observateur
29%	L' Officiel des Spectacles
27%	L' Express
24%	Télérama
22%	Le Point
20%	Pariscope Télé 7 Jours
19%	Photo Sciences et Vie
16%	L' Expansion
15%	L' Humanité Le Matin
13%	Pilote Le Figaro
12%	L'Equipe du Lundi
11%	Cent Idées
10%	Télé Poche France Soir
9%	Le Journal du Dimanche Historia Paris-Match
8%	Spectacle du Monde
6%	Jours de France Le Parisien

THE FASHION-CONSCIOUS 31%	
40% read	Modes & Travaux
33%	Télé 7 Jours
28%	Art et Décoration
26%	Marie-Claire Marie-France
21%	Sélection Femmes d'Aujourd'hui
20%	Jours de France
19%	Femme Pratique Paris-Match
17%	Parents Modes de Paris
16%	Elle Votre Beauté
12%	Maison de Marie-Claire L'Express Cent Idées Mon Jardin/ Ma Maison Votre Maison
10%	Maison et Jardin Science et Vie L'Echo de la Mode
9%	Ma Maison/ Mon ouvrage Historia
8%	Le Point
7%	Le Nouvel Observateur

ENTHUSIASTS 19%	
62% read	a regional daily
43%	Télé 7 Jours
30%	Télé Poche
27%	L'Equipe
22%	Sélection
20%	Le Chasseur Français Onze
18%	L'Automobile Sport Auto
16%	L'Action Automobile
15%	France Football
14%	La Prévention Routière
12%	Science & Vie Système D Paris-Match
11%	Spécial Dernière
10%	L'Express
9%	Pilote La Pêche et les Poissons Weekend Historia France Dimanche
8%	Modes et Travaux Parents L'Huma Dimanche Ici Paris
7%	Photo Le Point
6%	Jours de France
5%	Journal du Dimanche France Soir

TABLE 1
What the different groups read (Cont.)

THE ROMANTIC 16%

45% read	Télé Poche
36%	Nous Deux
31%	Télé 7 Jours
27%	Intimité
20%	Modes et Travaux Podium
19%	Hit
18%	Ici Paris
17%	France Dimanche
16%	O.K. ou Bonheur
13%	Chez Nous Télé Star
12%	Modes de Paris
11%	Confidences Spécial Dernière Parents
8%	Sélection
6%	Femmes d'Aujourd'hui Onze Weekend
7%	Bonne Soirée Jours de France

THE MORAL-MINDED 7%

68% read	a regional daily
44%	Le Pélerin
35%	La Vie
27%	Notre temps
23%	Modes & Travaux
20%	Clair Foyer
19%	Echo de notre Temps
15%	Sélection
13%	Femmes d'Aujourd'hui
9%	Modes de Paris
8%	Télérama
7%	Echo de la Mode Bonne Soirée Bonheur
6%	Le Chasseur Francais Paris Match
4%	La Croix L'Express Science et Vie Rustica
3%	Point de vue Nous Deux Jours de France

THE SELF-ISOLATED 15%

93% read	a regional daily
11%	Télé 7 Jours
3%	Télé Poche
2%	Chez Nous Femmes d'Aujourd'hui Modes de Paris France Dimanche Paris-Match
1%	Bonne Soirée Rustica Spécial Dernière Jours de France Weekend

TABLE 2
Profiles of the different groups

		Total	The fashion-conscious	The romantic	The moral-minded	The self-isolated	The enthusiasts	The strong-willed
% readers								
Sex:	Men	47	27	37	34	54	80	63
	Women	53	73	63	66	46	20	37
	Housewives	42	59	44	52	39	15	26
	Working women	21	34	25	18	12	8	14
Average age		*43*	*43*	*39*	*51*	*54*	*38*	*36*
Aged:	15-24	20	17	29	13	7	27	32
	25-34	20	22	21	10	9	24	27
	35-49	22	24	20	19	21	23	20
	50-64	19	21	15	20	26	18	11
	65 and over	19	16	15	38	37	8	10
With children aged:								
	2 or less	7	8	10	5	3	7	8
	7 or less	18	19	24	11	9	20	17
	15 or less	38	38	52	29	26	42	34
Education:								
	Higher	11	14	1	9	3	10	47
	Secondary	40	49	48	38	26	56	44
	Primary	49	37	51	53	71	34	9
Occupation:								
	Higher grades	5	5	0	4	2	5	18
	Middle grades	19	25	12	13	9	23	30
	Small businessman or farmer	7	7	5	9	12	7	1
	Foreman, workman	22	16	33	11	21	33	10
	Student	11	9	14	7	3	15	22
	Non-working housewife	17	22	20	23	17	6	8
	Retired	19	16	16	33	36	11	11
Dwelling:								
	Rural	28	25	26	38	37	23	7
	Towns of less than 100.000	28	28	31	31	33	32	10
	Towns of more than 100.000	28	30	27	24	30	32	20
	Paris conurbation	16	`17	16	7	0	13	63
Owns:								
	Colour TV	36	39	35	34	32	41	37
	Dishwasher	16	20	8	14	8	17	26
	Takes wintersports holidays	12	16	6	6	4	15	25
	Been to cinema in last 7 days	8	8	6	4	3	10	25