Gilles Santini
IMS, France

8.2 VALIDATION OF DATA FUSION TECHNIQUES: WHAT CAN STATISTICAL THEORY DO FOR US?

INTRODUCTION

Most of the work which has been done in the past few years on fusion has focused on how actually to perform such an operation and what is its underlying logic.

Although a lot of practical experience has been accumulated, beside the general debate on the motivation for fusion, very little work is available on how to validate a given fusion.

Some rare papers have been given so far, trying to demonstrate either the pros or the cons, but they seem to lack sound statistical grounds and, through controversy, they have brought more confusion than safeguards to the users.

My intent here is not to discuss the usefulness of fusion nor the legitimacy of the method, nor to argue about the best way to operate. My sole concern is to go back to basics, restate what fusion attempts to do, and elaborate on what could be considered as possible statistical procedures to test the goodness of fit of a given fusion.

FUNDAMENTALS

Taking for granted that every one here has a minimal knowledge of fusion terminology I will summarise only some basic points on the foundations of fusion:

– The objective of fusion is to add to a recipients' sample simulated data available in the different sample ie the donors' sample.

– The operation should be done at individual level in such a manner that the resulting data file containing both original and transferred data is appropriate to draw legitimate results by way of cross tabulation of any two variables.

– Neglecting third order interactions, one will consider that the file obtained through fusion is 'good' if probabilitic assumptions drawn on the various two-way distributions reflected by the corresponding two-way tables do not differ from the ones drawn from the actual collected samples.

We will see that an alternative way to the above is to transform the problem into checking whether two samples of multi-dimensional data points come from the same population.

NAIVE STEPS

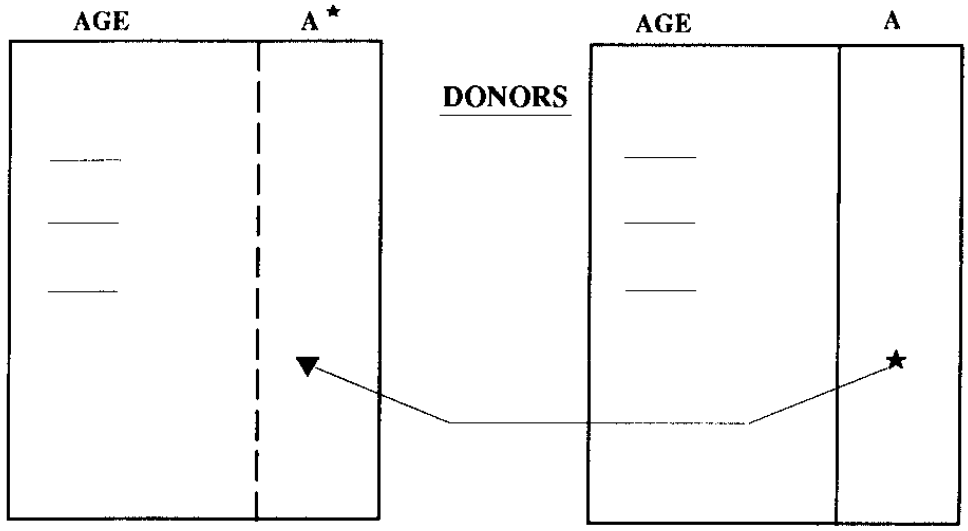
Case 1

Let us assume that the only available common variable between donors and recipients is age and that the only variable to be transferred is readership of magazine A (Chart 1).

We may for example observe the following distribution in the donors sample:

	Read A	Do not Read A	
< 35 Years	5	20	25
35 - 50 Years	10	30	40
> 50 Years	14	21	35

Chart 1



In this case, it is easy to stimulate readership of magazine A in the recipient sample by random drawing of a Bernouilli variable (read/do not read) with probability 5/20 in the less than 35 years group, 10/30 in the 35-50 years group and 14/21 in the over-50 years group.

If the distribution of age groups is similar within the donor and the recipient samples, the frequency of readership in each cell will be similar within the donors' and the recipients' files. If the distribution of age groups is different within the donor and the recipient samples, the frequency of readership in each cell will still be similar within the donors' and the recipients' files.

Case 2

Let us assume now that we want to transfer two correlated variables, ie readership of magazine A and magazine B. The previous Monte Carlo method will not be adequate since in the merged file the frequency of reading A and the frequency of reading B will be independent. In order to avoid breaking the correlation of A and B, one can think of matching, one to one, donors and recipients within the same age group, and transfer all readership from a donor to its matched recipient. If the numbers of donors and recipients per cell are equal, this matching can be done randomly with conservation of both marginal and crossed distributions of A and B within cells (Chart 2).

If the numbers of donors and recipients per cell are unequal, random sampling with replacement from a donor cell (age group), will produce, through matching into the corresponding recipient cell, simulated readership with frequencies similar to the original ones.

It can readily be seen that the Case 1 mechanism is similar to the latter mechanism although the one to one matching is of a different nature.

Case 3

Let us assume now that we have more than one common variable. The previous method can readily be generalised by considering cells which are the combination of the common variables.

Chart 2

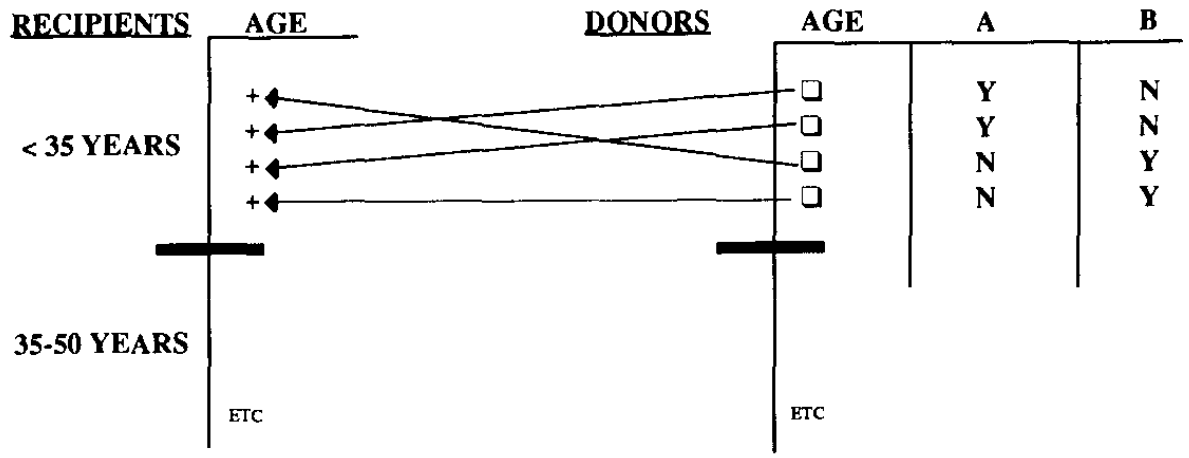
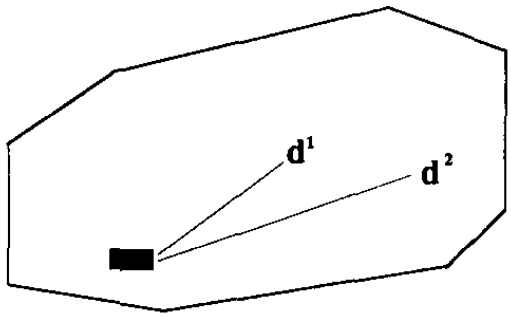


Chart 3



d^2 IS LESS SIMILAR THAN d^1 TO r

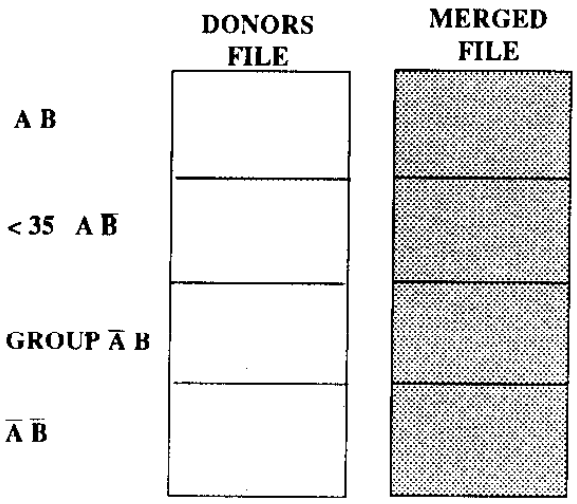
However, we quickly hit a limit since the produced cells are very small, resulting in the impossibility of performing the random drawing correctly. The idea then is to work not within a given cell but within a neighbourhood. To construct such a neighbourhood, the cases are mapped into an Euclidean space, the topology of which is meant to reflect similarity of the cases based on the common attributes as expressed by the common variables (Chart 3).

Around a recipient, the donors within its neighbourhood do not play the same role since the further away they are from the recipient the more they are dissimilar. According to this it is not possible simply to work by random drawing within the neighbourhood. In order to overcome such difficulty, fusion experts have designed various algorithms which attempt both to match donors and recipients according to their closeness and to produce data distributions similar to the original ones.

PARTIAL TESTING

Considering the above, a simple way to check how good a job the fusion algorithm has performed is to compare the distribution of the merged variables in the merged file and in the donor's file. Since the donor and the recipient

Chart 4



samples may be different, it is necessary to do so by variables breaks.

A simple statistical test to use is the chi-squared test. However, since such test is sensitive to empty cells, one might consider using zero-cell correction in some cases.

A particular case is concerned with duplications. For such purpose, the test may be set up as in Chart 4.

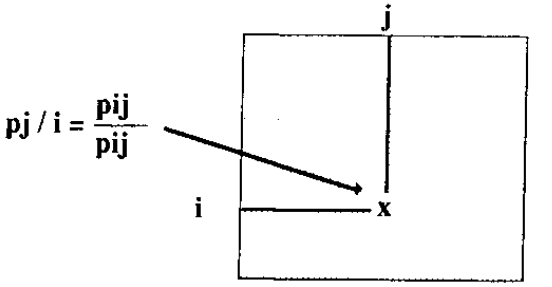
GLOBAL TESTING

Partial testing cannot produce a global indication of how well a fusion has performed. Indeed, it gives us valuable information on what needs to be worried about and what is satisfactory. However, we need some kind of global criteria.

Let us note X the common variables and A the transferred ones. Neglecting third order actions one can think of comparing the two examples in Chart 5 where everything is crossed by everything.

Such tables are known as Burt tables. We want to compare such tables but we have to face the fact that the donors' and the recipients' file may be different according to the X's.

In fact what we really want to compare are condition probabilities, as follows (for each elementary combination the marginal value is set to 100%):



With an appropriate metric such a job can be done by a type of multivariate categorical analysis known as multivariate analysis of correspondences*. Since such a method is only concerned with conditional probabilities, it will eliminate all scale effects on the X's. We will neglect all other type of differences between the two samples since that can only induce us to be more pessimistic about fusion that is justified.

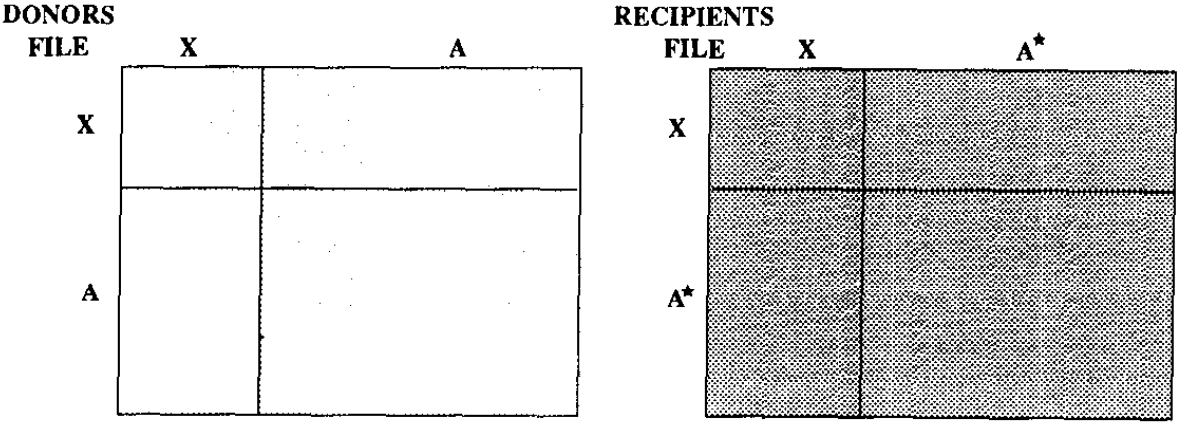
This being considered, analysis of correspondence maps all cases within the donors' and the recipients' samples into a common multidimensional space where each data point is positioned according to a set of continuous coordinates. Let F_d^k and F_r^k respectively be such coordinates for the donors and the recipients.

* This method, developed by JP Benzecri (1973), is somehow related to canonical analysis as shown in Greenacre (1984).

VALIDATION OF DATA FUSION TECHNIQUES: WHAT CAN STATISTICAL THEORY DO FOR US?

Chart 5

Burt tables



The question is are these two sets of points coming from the same distribution of multi-dimensional variables of which $\{F_d^k \mid d \varepsilon D\}$ and $\{F_r^k \mid \varepsilon R\}$ are observations?

Since no distribution theory is known on multi-dimensional variables, we have to rely on non parametric testing theory. Work due to two US statisticians, J Friedman and L Rafsky, has greatly helped us in designing the following method.

THE MULTIVARIATE TWO-SAMPLE PROBLEM

At that point we need to go back to some statistical methodology. The multivariate two-sample problem can be stated as follows: let us consider two samples of size m and n from distributions $F\chi$ and $F\gamma$ where x and y are multidimensional random variables defined in $R\Upsilon$. We wish to test whether the two samples come from the same distribution ie in statistical terms, we want to test the null hypothesis $F\chi = F\gamma$ against the general alternative $F\chi \neq F\gamma$. This is a multivariate analogue of a well known univariate problem for which nonparametric tests are available and frequently used in practitice.

Among them, let us focus on what is known as the Wald-Wolfowitz Runs Test. This test works as follows.

Let us arrange the two samples (termed x and y) into one combined ordered sample:

$$X < X < Y < X < - - - < Y < X$$

$$a_1 \; a_2 \; \beta_1 \; a_3 \qquad\qquad \beta_m \; a_n$$

We count the number of runs of like elements in the above sequence. Under the null hypothesis such a count is asymptotically distributed as a normal variable with mean and variance which depend only on the size of the two samples and are easily derived by way of a combinatorial argument. The idea behind all this is that we expect samples from the same distribution to have scattered values and samples from distinct distribution to have clustered values.

Chart 6

The MST of a set of points is a connected graph that spans all the points and has minimal length

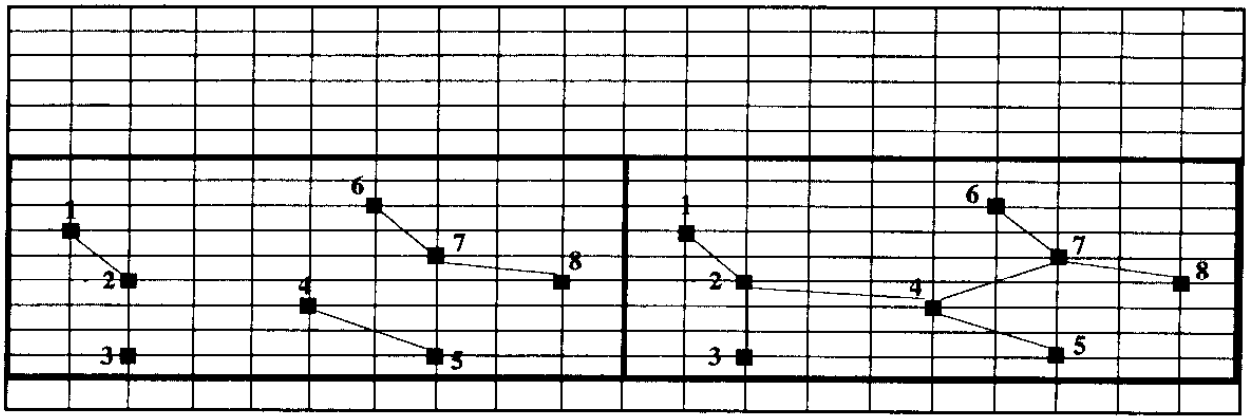
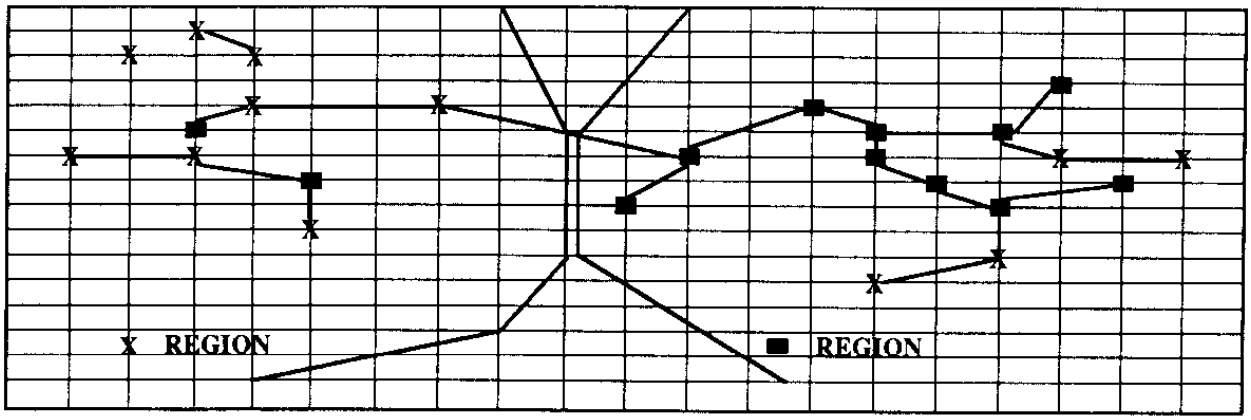


Chart 7



The same idea applies in Υ dimensions, the problem being to traverse all points in a somehow ordered way. Following Hartigan (1975) who pointed out that a Minimal Spanning Tree (MST) provides an excellent description of point sets, Friedman and Rafsky (1979) have suggested the use of such structure for our purpose.

A MST can readily be constructed first by connecting each point to its closest neighbour and then connecting each fragment so created to another one by the smallest possible link (Chart 6).

In one dimension the MST of a set of points reduces to a simple linkage of the sample points

ordered by value as required for the Wald-Wolfowitz Runs Test.

In Υ dimension , we tend to believe that the two samples come from the same distribution if the MST of the pooled sample does not show regions dominated by one sample only (Chart 7).

Following such an approach, Friedman and Rafsky have generalised the Wald-Wolfowitz Runs Test to the multivariate two-sample problem, leading to the following procedure (Charts 8 and 9).

Under the null hypothesis and limited *a priori* validity conditions the number R of dispoint fragments is distributed as a normal random variable with mean and variances respectively equal to:

$$E(R) = \frac{2mn}{N} + 1$$

$$V(R) = \frac{2mn}{N(n-1)} \left\{ \frac{2mn-N}{N} + \frac{C-N+2}{(N-2)(N-3)} \left[(N(N-1)-4mn+2 \right] \right\}$$

with C = the number of edge pairs that share a common node

$$C = \frac{1}{2} \sum_{i=1}^{N} d1(di-1)$$ where di is the degree of linkage of the ith node

Here we have 12 nodes of degree 1
23 " 2
8 " 3
1 " 4

with some arithmetic, this leads to C = 53 and
E(R') = 22.8
V(R) = 10.5

Since $|R-E(R)| < 1.96\, V^{\frac{1}{2}}(R)$, we accept at level 5% the null hypothesis ie both donors and recipients behave as samples from the same distribution.

Although the Wald-Wolfowitz Runs Test is not very powerful* in the univariate situation, Friedman and Rafsky have given evidence by way of numerical simulation that their multivariate analogue is more powerful than other generalised non parametric tests such as the Smirnov test whenever the dimensionality is large.

Such is the case in the fusion context. This, added to the fact that the topological nature of the MST is similar to the one of fusion, makes us believe that the above statistical method should be considered for performing global tests on fusions.

TESTING PROCEDURE

As we said at the beginning, the object of this paper is not to present actual data but to raise methodological issues. We will conclude then with what in our opinion should be the right way to test whether fusion leads to reliable data or not.

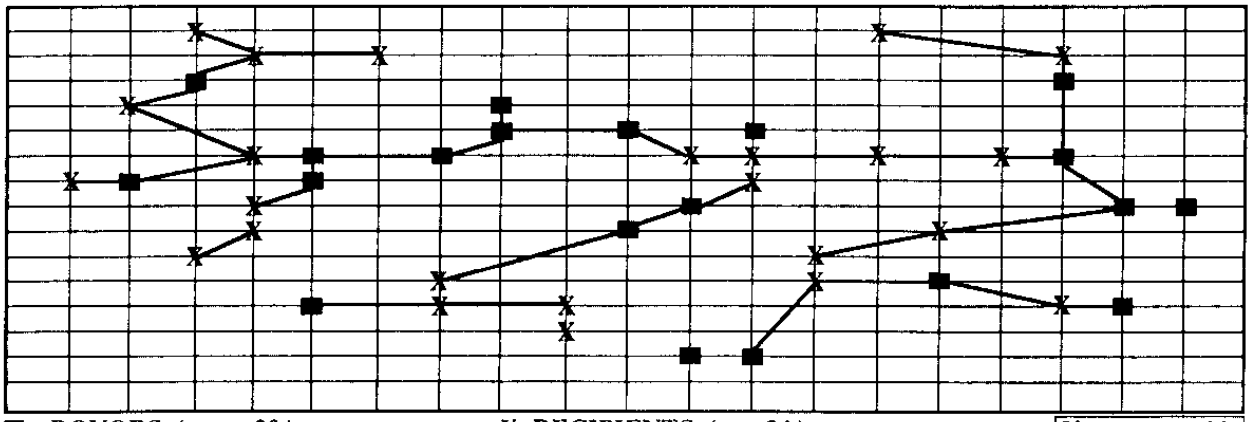
(1) Compare the distribution of the common variables in the donor and the recipient samples: the chi-square test or the COCHRAN T test are good candidates for that. If significative differences are detected we should ask ourselves whether this is correct and if not, if we are ready to ignore it (considering then that the recipient sample is the right one) or eventually improve it by dropping or reconsidering some of the common variables.

(2) Check whether the fusion algorithm has performed 'smoothly' ie whether some abnor-

* The power of a test is the probability of rejecting a false null hypothesis.

Chart 8

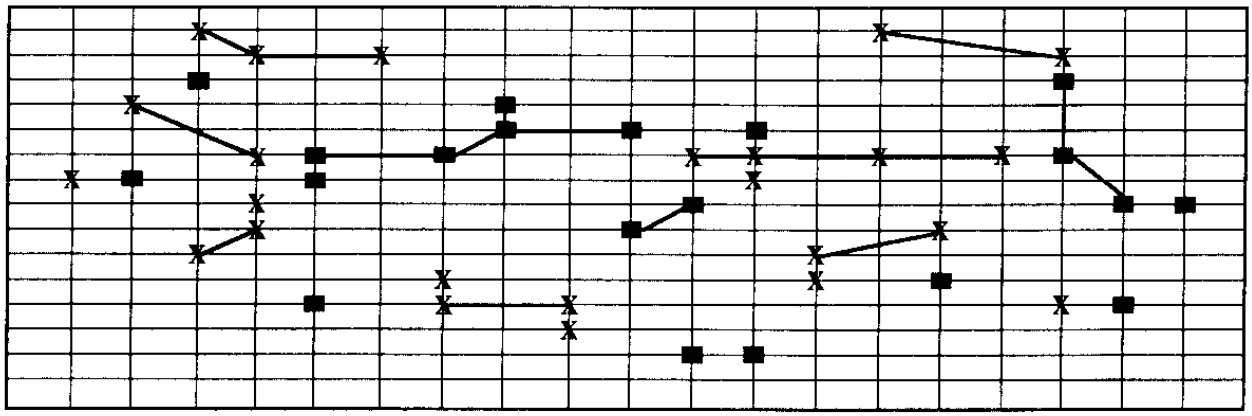
Construction of the MST



■ DONORS (m = 20) X RECIPIENTS (n = 24) N = m + n = 44

Chart 9

Deletion of the mixed edges



R = 19 DISJOINT FRAGMENTS

mal behaviour has been observed. This differs from one algorithm to another one. Indicators may be:

- distribution of the number of established links per donor.

- distribution of the types of marriages.

- average distance per type compared with the total average distance between married respondents.

- relative cardinality of the donors and the recipients in each cell compared with the overall relative cardinality.

- distribution of the types of marriages per cell.

- total average distance between married respondents per cell compared across cells.

These checks do not guarantee the reliability of the produced data but are important indications of what may have happened.

(3) Compare for the total population on some meaningful (rather larger) subpopulations the absolute levels of the transferred variables between the donors' file and the recipients' file. If the donor sample and the recipient sample are different, reweighting of the donor sample according to the recipient major common variables should be done in order to get corrected figures for comparison.

(4) Perform as many partial-tests as economically possible as suggested previously in this paper and produce an indication of how many are significant or not at the 5% level.

(5) Perform for each cell of the fusion our proposed global test of comparison of the donors versus the recipients. Examine cells showing significant differences*. Depending on the size and the nature of the cells accept or reconsider the operations.

If every thing is OK, one should accept the fusion in confidence. Otherwise, one should try to evaluate how important is the defect for practical use, warn the users of possible inadequacies, and eventually reject the fusion.

References

Antoine, Jacques & Santini, Gilles (1986). An experiment to validate fusioned files obtained by the referential factorial method. ESOMAR Seminar on 'New developments in media research'. Helsinki.

Antoine, Jacques & Santini, Gilles (1987). Fusion techniques alternative to single-source methods? *European Research*, August **15**, 3.

Benzecri, Jean-Paul (1973). *Analyse de donnees.* Tome 1 La Taxinomie. Dunod.

Friedman, Jerome H & Rafsky, Lawrence C (1978). Graphics for the multivariate two-sample problem. Stanford Linear Accelerator Center Publication **2193**. August.

Friedman, Jerome H & Rafsky, Lawrence C (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics* **7**, 4.

Greenacre, Michael J (1984). *Theory and applications of correspondence analysis.*

* Two way representations of the data points based on the MST structure can be used for such purposes: see Friedman & Rafsky (1978).

Hartigan, John (1975). *Clustering algorithms.* Wiley. New York.

Rothman, James (1988). Testing data fusion. ESOMAR Seminar on 'Media and media research: How far can we go?'. Madrid.

Santini, Gilles (1986). Méthodes de fusion: Nouvelles réflexions, nouvelles expériences, nouveaux enseignements. *IREP,* décembre.

Wendt, Friedrich (1983). The AG.MA model. Montreal Proceedings.

Wiegand, Jüngen (1986). The combining of two separately derived data-sets into an integrated intermedia planning system: the German model of partnership. ESOMAR Seminar on 'New developments in media research, Helsinki'.