FUSION QUALITY ASSESSMENT

Dina Raimondi, Ipsos
Gilles Santini, G-LINE Invest

1- INTRODUCTION

During the past two decades, fusion techniques have been presented and experimented in most European countries, encountering more or less acceptability from the market. Attitudes in front of data resulting from a fusion ranged from indifference or unawareness of users which made no difference in the origin of these data and in the evaluation of their quality or reliability, to a reluctance or even a deny of these techniques. Specially on statisticians side, fusions were not considered as part of the statistical establishment, mainly for validation reasons, as we did not have, for instance, any capability to calculate the confidence intervals around estimates produced by a fusion.

Nevertheless, it is a fact that fusion is very efficient to recover information in many cases, and that it helps to answer many questions, as sub-sampling, non-response, length of questionnaire, budget issues, etc… some of these experiences have been presented in past Symposia. This is probably the reason why, a few years ago, fusion started to raise curiosity on the other side of the Atlantic, the program of the forthcoming ARF workshops which dedicates a significant amount of contributions to these techniques is a proof of the interest for fusions among practitioners of market and media research.

The questions remain : how can the accuracy of a fusion be evaluated, how can a user appreciate the reliability of the results, do we have the same standard error when a decision is taken on the basis of data resulting of a fusion as when it is taken on the basis of a unique specific sample ?

An experiment has been designed to answer such questions : the same fusion was replayed a large number of times on random sub-samples and the variability of the results was estimated. Based on a 20,000 respondents print audience file, considering socio-demo variables, habits of reading, and additional variables of socio-cultural type.

A three steps procedure :

- we extract a sub-sample of 8 000 respondents, which is split in 2 halves,
- we use these two sub-samples (1 as a donor, 2 as a receiver) to produce by fusion a third sample,
- we calculate audiences, ranking and schedule evaluations on sample 2 and on sample 3 produced by the fusion.

This is run 100 times.

Then we compare the results issued from the fusion datasets (samples 3) and the ones coming from the observed datasets (samples 2) in order to see if there are bias in audience results and in schedules performances by target.

The conclusions presented in this paper are based on a large amount of results in order to give them the suitable significance. The fusion algorithm involved is the latest available, including the use of a new distance to evaluate the similarity between donors and recipients.

2- PROCUSTREAN FUSION ALGORITHM

2.1- Context

From here on we will consider that a set of *common variables* C is collected for two comparables samples I and J of respondents and that a set of *additional variables* A is collected for sample I only.

Sample I is called the donor sample and sample J is called the recipient sample. The aim of the fusion process is to produce ascribed data on all additional variables for each respondent in the recipient sample.

2.2- Principles

The Procustrean Fusion Algorithm (PFA) obeys five principles :

1- Each recipient should receive data from a single donor.
2- The data collected for a donor is transferred as a whole to the linked recipients.
3- Any donor already linked should be highly discouraged to produce further links

4- The cross-distributions between common and additional variables should be preserved unchanged by the ascription process.
5- The similarity between two respondents should be evaluated globally.

The rational underlying the first two principles is simple : the aim is to avoid breaks of the inter-correlations between the additional variables during the ascription process.

The third principle protects against a decrease of the effective sample size.

The fourth principle refers to one of the basic requirements for a good fusion..

The last principle is somewhat more subtle. The idea behind is that if one considers that two donors are close (i.e. similar) they need to be so, not only on the basis of the common variables but also on the basis of the additional variables, otherwise the fusion process could distort the dependencies existing among these variables

Altogether the five previous principles are useful to protect the fused database against distortion of the relationships existing between the variables

2.3- Distances

Let $d_C^2(i',i'')$ be a distance defined between two donors (i',i'') on the basis of the common variables and likewise $d_A^2(i',i'')$ be a distance defined on the basis of the additional variables.
In order to fulfill the requirement expressed by the fourth principle the idea is to built a weighted distance on the common variables and to chose the weights applied on each of them in such a way that it will statistically be true that :

$$\boxed{d_C^2(i',i'') < \varepsilon_1 \Rightarrow d_A^2(i',i'') < \varepsilon_2}$$

Recognizing the fact that most of the collected data is categorical, it is a common statistical practice to define the distance between two respondents as the Euclidean distance of their representations in the multidimensional space associated with the correspondence analysis of the data. Such a choice is justified by the fact that the resulting measure is not affected by random noise and is scale-effect free .
One will writes with usual notations:

$$\boxed{d_C^2(i',i'') = r_C^{-1}\sum_{k=1}^{r_C}\left(\psi_C^k(i') - \psi_C^k(i'')\right)^2}$$

where r_C is the number of retained factorial axis .

and likewise :

$$\boxed{d_A^2(i',i'') = r_A^{-1}\sum_{k=1}^{r_A}\left(\psi_A^k(i') - \psi_A^k(i'')\right)^2}$$

One can also define a derived weighted distance :

$$\boxed{\begin{array}{l}\tilde{d}_A^2(i',i'') = \sum_{k=1}^{r_C}\varpi_k^2\left(\psi_C^k(i') - \psi_C^k(i'')\right)^2 \\ \text{with } \sum_{k=1}^{r_C}\varpi_k^2 = 1\end{array}}$$

where the ϖ_k^2 weighs are estimated in such a way that :

$$\boxed{\tilde{d}_A^2(i',i'') \approx d_A^2(i',i'')}$$

This can be done by spherical regression using all the pairs of points or a sub-sample of them .

Finally one will use as an indicator of the distance between two donors (i',i'') the quantity :

$$\boxed{\delta^2(i',i'') = Max\left\{d_C^2(i',i''),\overset{\%}{d}_A^2(i',i'')\right\}}$$

The usefulness of this approach comes from the fact that it is possible to compute the same distance indicator for a pair of the recipients (j',j'') since the only quantities involved in the formula are based on the common variables :

$$\boxed{\delta^2(j',j'') = Max\left\{d_C^2(j',j''),\overset{\%}{d}_A^2(j',j'')\right\}}$$

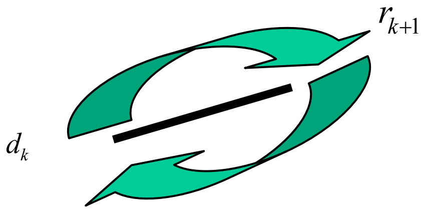
As a reminder of the method which basically consists in stretching or cutting the distance components one will note that Procustrus in the ancient times was doing the same to fit his visitors in the beds of his hostel by stretching or cutting their legs From this much dreaded inn keeper comes the name of the algorithm .

2.4- Marriage Algorithm

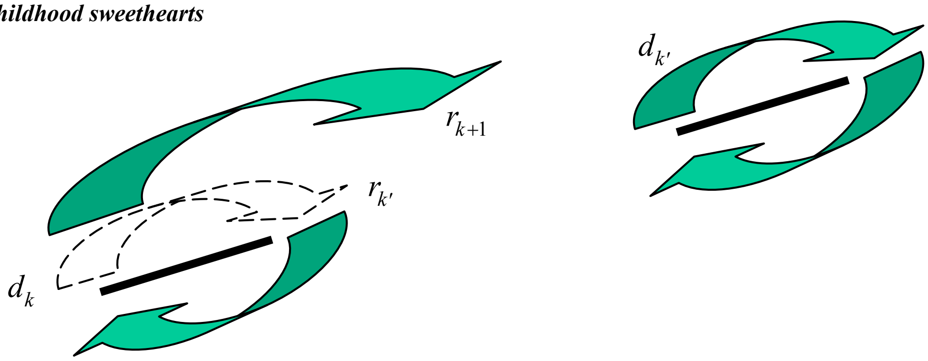
Once the method used to measure how close (i.e. similar) two respondents are it is necessary to produce links between donors and recipients.
The PFA method is quite classic in this way and search for pairs that exhibit a form of reciprocal affinity.

Let us note $V_R(i)$ the closest recipient to donor i and $V_D(j)$ the closest donor to recipient j .
If $\{V_R(i)=j\}\wedge\{V_D(j)=i\}$ one will said that i and j are reciprocal neighbors. This strong relationship should result in a link between them. With famous nicknames the following figures exhibits the various types of configurations :

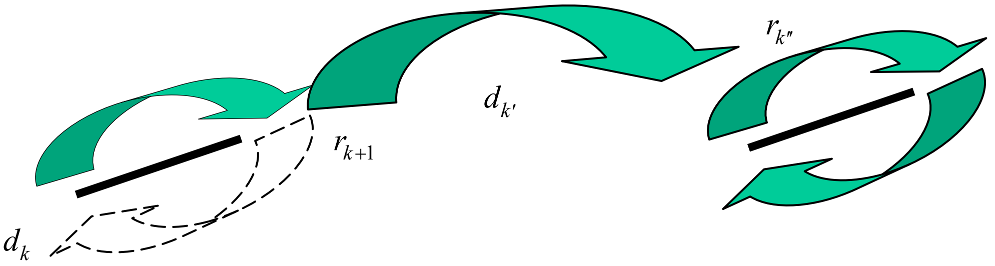
Love at first sight



Childhood sweethearts



Assiduity



The other cases are called *marriage of convenience*.

Closest neighbor
Next neighbor
Marriage

Arrows lengths are not significant

2.5- Optimization

Once all the links are established the PFA method looks for an optimization of the marriages configuration in a way that satisfies as much as possible the fourth principle. This is done by breaking wisely some of the convenience marriages and replacing them by better ones.

The objective set to this optimization is to keep after transfer the existing statistical relationships between the common and the additional. variables

To do so PFA looks at the set of binary variables $\{c_k\}$ associated with each category of each common variable and at the set of binary variables $\{a_l\}$ associated with each category of each additional variable.

For each couple (c_k, a_l) the following (2x2) frequency table is built:

$c_k \otimes a_l$	Donors	Recipients
$c_k \wedge a_l$	f_d	f_r
$\bar{c}_k \vee \bar{a}_l$	$1-f_d$	$1-f_r$

The search is aimed at minimizing the number of significant χ^2 differences between the two rightmost columns.

In practice the main interest of this phase is to produce datasets that show little distortion in terms of targets groups. It is particularly useful when the type of distances used produces many cross tabs bearing significant differences. With the use of the Procustrean distance such discrepancies are much less numerous (a small %) so optimization may often be skipped which save a large part of the computing burden.

3- EXPERIMENT

3.1- Objective

The objective of the experimentation is to assess whether one is likely to take similar media-decision from a fused data base to the ones taken from a direct random sample.

Since random samples incorporate a certain degree of variability decision taken from two independent comparable random samples may vary.

The question is :

"Is the variability similar when one uses fused databases rather than distinct random samples ?"

A statistically valid answer to this important question can be provided by comparison for bias and difference in spread between of the outcomes of several fused databases coming from an original dataset with the same outcomes of several independent random

Since costs prohibit real life experiences we have set up a large scale Monte Carlo experiment which can be replicated if one wishes to assess the variability of fusion in a specific case.

3.2- Method

We start with a typical audience study from the AEPM which provides us with a well balanced sample *S* of roughly 20000 respondents for which we have considered answers to 9 socio-demo variables, 44 habits of reading , and 48 additional variables about practices , uses and habits.

We consider *K*=100 runs.

For each run :

- we draw randomly without replacement a sub-ample S^k of 8000 respondents.
- his sub-sample is randomly split into two halves S_1^k and S_2^k of 4000 each.
- then we consider for S_2^k the observed common and additional variables : $S_2^k = \left[C_2^k \mid A_2^k\right]$
- next we use $S_1^k = \left[C_1^k \mid A_1^k\right]$ and $S_2^k = \left[C_2^k \mid \varnothing\right]$ to produce by fusion a third sample $S_3^k = \left[C_2^k \mid A_3^k\left(A_1^k; C_1^k, C_2^k\right)\right]$.
- we perform some work (audience calculations, ranking and schedule evaluations) for a large number of targets on the dataset $S_2^k = \left[C_2^k \mid A_2^k\right]$ producing a set of results R_O^k
- we perform the same work on dataset $S_3^k = \left[C_2^k \mid A_3^k\right]$ producing a set of results R_F^k

Once all runs are performed we compare the two set of results R_O^k and R_F^k . To do so for each couple of resulting quantities $\left(x_O^k \in R_O^k, x_F^k \in R_F^k\right)$ we calculate the mean and variance across the runs :

$$\boxed{\begin{aligned} \overline{x}_O &= \tfrac{1}{100}\sum_{k=1}^{100} x_O^k \\ \sigma_{x_O}^2 &= \tfrac{1}{100}\sum_{k=1}^{100}\left(x_O^k\right)^2 - \left(\overline{x}_O\right)^2 \end{aligned}} \qquad \boxed{\begin{aligned} \overline{x}_F &= \tfrac{1}{100}\sum_{k=1}^{100} x_F^k \\ \sigma_{x_F}^2 &= \tfrac{1}{100}\sum_{k=1}^{100}\left(x_F^k\right)^2 - \left(\overline{x}_F\right)^2 \end{aligned}}$$

The question is how far apart are those results?

A first and quite natural idea could be to measure how many *standard deviations* one mean stands from the other. This however is not entirely adequate since a difference in mean may be very small compared to the standard deviation and yet be very significant.

To properly assess the significance of a difference in means one has to look for the number of *standard errors* that they are apart. Standard errors measures the accuracy with which the sample mean estimates the *true* value of this quantity. In the case of means they turn out to be equal to the standard error divided by the square root of the sample size. Furthermore a classic statistical test known as the *Student's t-test* enables us to decide whether a difference in means is statistically significant or not[1].

This being said we should also check for the numerical importance of a statistically significant difference since a large number of observations may generate significance diagnostics for differences which are numerically negligible.

[1] Since most results are proportions p ,we have applied this test on transformed variables using $z = \log\left(\frac{p}{1-p}\right)$ and have taken in account in the calculation of the t statistic possible unequal variances.

Last but not least we may decide to ignore differences that are both statistically significant and numerically non negligible because their impact will be null in practice.

In the present case we will use the following compound decision scheme :

1. Discard targets which are too small (less than 25)
2. Discard results with level less than 1% in the target group
3. Neglect differences which are less than 1% point or less than one standard deviation in magnitude
4. Test for statistical significance at level 0.05.

Any elementary difference in results that pass this compound test will be considered as a discrepancy.

The fusion method will shows a systematic bias if more than a few % of the results show such discrepancies positively or negatively

Also if the results coming from the fusion datasets doesn't exhibit more fluctuations that those computed from the observed ones the ratio of the precision factors should be close to one:

$$\boxed{\delta = \frac{\sigma_{x_F} / \bar{x}_F}{\sigma_{x_O} / \bar{x}_O} \cong 1}$$

This can be verified by looking at each the results in the two datasets and testing statistical significance for the difference in variance (F-test).

3.3- Calculations

In order to have at hand results covering a wide variety of cases we have considered 630 targets built by combination of 30 subpopulations defined from the additional variables which are ascribed by the fusion process with 20 subpopulation built from the socio-demo's plus the total sample.

On each of these 630 targets we have calculated the audience of 44 media vehicles covering the whole spectrum of type of magazines as well as the performances of 25 random schedules varying from one target to the next.

Altogether the number of different media rankings computed during the experiment amounts to a total of : $100 \times 2 \times 630 = 126\ 000$ and the number of schedule evaluations to $100 \times 2 \times 630 \times 25 = 3\ 150\ 000$!

The results which have been considered are :

For each audience analysis :

- Audience levels in %
- Ranks within the 44 magazines

For each schedule evaluation :

- Total GRP
- Reach
- Coverage 2+
- Coverage 3+
- Coverage 4+
- Coverage 5+

3.4- Results

A first check to perform was to verify that the target sizes were close when estimated from the fused datasets compared to the observed ones.
It turns out that the target sizes relative discrepancy :

$$\boxed{\Delta_N = \left| \frac{N_F - N_O}{0.5 \times (N_F + N_O)} \right|}$$

shows an average across the 630 studied targets equal to 0.033 with a standard deviation of 0.042 which is quite good considering the large variation existing in the target sizes.

The others major results are summarized in the following tables :

MEDIA VEHICLES	% of results with positive bias (average bias)	% of results with negative bias (average bias)	% of results with difference in variance (average delta)
Audience levels	0.7 % (0.02)	2.0 % (0.02)	10.6% (1.06)
Rankings	2.5 % (4.31)	3.2 % (5.23)	24.5 % (1.12)

SCHEDULES	% of results with positive bias (average bias)	% of results with negative bias (average bias)	% of results with difference in variance (average delta)
GRP'S	0.2 % (0.05)	7.7 % (0.09)	5.4 % (1.06)
Reach	0.2 % (0.02)	11.8 % (0.03)	6.7 % (1.08)
Coverage 2+	0.3 % (0.02)	8.5 % (0.03)	7.1 % (1.07)
Coverage 3+	0.3 % (0.02)	5.7 % (0.02)	7.0 % (1.06)
Coverage 4+	0.2 % (0.02)	3.5 % (0.02)	7.5 % (1.06)
Coverage 5+	0.2 % (0.02)	2.0 % (0.02)	8.5 % (1.05)

Although most results are quite satisfactory, one can see that the fused datasets results exhibits however a clear but negligible positive bias which seems to affect mainly the reach in 12% of the cases but at only an average level of 3 percentage points whenever it occurs.

One will also observe this important fact that the stability of the results is very high.

To better judge the magnitude of the figures exhibited by the above table one would also like to compare it to random matching without replacement.
Here are the results following strictly the same experiment protocol :

MEDIA VEHICLES	% of results with positive bias (average bias)	% of results with negative bias (average bias)	% of results with difference in variance (average delta)
Audience levels	3.6 % (0.02)	20.9 % (0.02)	34.0% (1.16)
Rankings	15.5 % (6.33)	14.8 % (7.26)	49.0 % (1.26)

SCHEDULES	% of results with positive bias (average bias)	% of results with negative bias (average bias)	% of results with difference in variance (average delta)
GRP'S	3.8 % (0.10)	69.3 % (0.20)	20.1 % (1.16)
Reach	3.5 % (0.03)	68.4 % (0.05)	21.9 % (1.18)
Coverage 2+	2.8 % (0.03)	61.7 % (0.05)	27.5 % (1.18)
Coverage 3+	2.4 % (0.02)	54.9 % (0.04)	32.8 % (1.19)
Coverage 4+	2.0 % (0.02)	48.0 % (0.03)	36.2 % (1.19)
Coverage 5+	1.4 % (0.02)	39.8 % (0.03)	37.6 % (1.19)

Fusion clearly does much better!

4- CONCLUSION

On the basis of the performed experiment that should be replicated with other studies, one should conclude that :

- **there is no major or discouraging bias linked to the fusion process when compared to results issued from independent samples for all common media-planning applications (ranking and schedules)**,
- **there is no strong variability from one fusion to another, which make the fusion process robust.**

In addition this experiment shows that the formulas which are applied to independent samples can be applied to data issued from a fusion process in order to have a good idea of results precision.

It demonstrates that fusion methods may enter the category of classical statistical methods and do not induce any unreasonable risk for the users, in the context of professional usage of media data, that is to say, not on the base of individual level comparison which makes no sense in view of the effective use of such data , but on the base of media performance estimated for a defined marketing or media target.