

BUILDING A PANEL FROM A RECENT READING DATABASE THROUGH FUSION

Flávio Ferrari, IBOPE Pesquisa de Mídia
Alexandre Crivellaro, IBOPE Group

The use of the “recall” methodology for Press audience measurement is widely adopted for the good quality of its results and for good cost x benefit relationship. Nevertheless, although very effective for the evaluation of a Newspaper’s performance, it presents important limitations for planning advertising for this medium, since it does not allow the simulation of the campaign results (reach, frequency and frequency distribution). The research using “panel” methodology is an effective alternative, but with high costs for implementation and maintenance, mainly considering the necessary sample size to represent a very fragmented behavior. The present work will delve into a creative alternative for the use of fusion algorithms to build a virtual panel from a group of individual respondents using a “recall” survey. This proposal allows the joining of the advantages of the recall research (practicality and low cost) with the benefits of the “panel” survey (possibility of campaign simulations) preserving the quality of the results.

1. Introduction:

Data fusion has proven to be an excellent tool for media applications, because it allows to obtain more information for an individual through the crossing of variables obtained from two or more distinctive sources. Some data fusion techniques, also known as statistical matching, were presented and classified as unconstrained [1] and constrained [6] and afterwards assessed [2]. A very similar study has been presented recently [3], where the Radio audience data was fused using a similar process, also based on recall information and a simple optimization algorithm.

In this study we are considering another possible practical application for these methods in order to find the solution of the limitations imposed by the “recall” methodology in the Press research.

Although the “recall” research is easy to implement and has a much lower cost than the panels, it is not the best solution for reach and frequency studies, as they register the behavior of each individual for a very limited period of time (usually 24 to 48 hours). Studies have shown that the rate of memorization [4,5] tends to decay over time, therefore creating a barrier in order to obtain longer periods of cooperation.

On the other hand, the fact that the research is conducted by an interviewer helps to obtain a good response rate for a few days.

Individual probabilities based on frequency of reading is another alternative, already adopted in some markets for media planning purposes, but with important limitations for small penetration newspapers.

So, the alternative of building a virtual panel from the “recall” research emerges as an attractive solution, which deserves to be assessed.

2. Methodology of Recall Survey

The “recall” survey, which is regularly performed by IBOPE in São Paulo and other 8 cities, was used as a basis for this work, makes use of three questions for the identification of the audience behavior: which Newspaper did the respondent read yesterday, the day before yesterday and sporadically. For the first two questions, we accept five options (limit) as an answer; and for the third question, we accept a maximum of three additional options.

Individual demographic characteristics are identified in this questionnaire: social class, sex, age, position in the family, level of education, geographic region, etc. The respondents necessarily have to be residents in the surveyed area, and have to be 10+ years old. Today, 80% of these interviews are conducted face to face and 20% are conducted over the phone.

2.1 Sample

The sample is selected in two stages. In the first stage, the clusters are probabilistically selected through the PPS method (Probability Proportional Size), which takes the population as a basis for the selection and, in the second stage, the selection of interviewees within the cluster, which is done through proportional quotas of the following variables: sex, age and economic status.

This study is carried out daily, with approximately 220 interviews per day, and the result is published on a monthly basis, on an average of ninety days (three months), with approximately 19,800 individual interviews (for the São Paulo metropolitan area). In this paper, we are going to use a sample of 19.768 people (close to three months of field work), obtained through this study.

This is not a magic figure and it has an important motives for its adoption: the main reason is that it is divisible by 14, aiming at not losing any individual in the fusion.

The sample weighting at the end of the process corrects the possible distortions of proportion and it is performed using the rim-weight method, with about 70 control variables. Differently from the so-called constrained statistical matching method [2], recently presented by R. Soong [8], the original weights are not used in the fusion process, as a new weighting solution will be adopted for the fused data.

3. Fusing Recall Data

3.1 The Process:

The fusion goal is to obtain a sample with size = $\frac{n}{14}$ of “virtual individuals”, with 14 continuous days of audience behavior, by fusing the “n” individuals from the same survey but for whom we have information for only 24 hours.

The fusion process used here can be described as an interactive process where, for a specific individual on a specific day, we search for 13 other individuals with less distance (see section 3.2) to complete a panel of 14 days, consequently, for each interaction, 14 individuals are withdrawn from the list and the process continues until the moment where there are no individuals in the list of donors. In the beginning of the process, we randomized the sample to avoid biases toward a specific demographic variable.

Each individual, whom we have the pattern of 48-hours (yesterday and the day before) is split into two individuals with information covering 24-hours, maintaining the same demographic variables. Even though this may be contrary to the fusion process, where we try to “fuse” different individuals into a single one instead of “splitting” one into two, this procedure strongly enhances the algorithm implementation. Both alternatives were tested, and the best results were obtained by splitting the individual into two donors.

For each individual that we seek, we use a score concept, where we attribute points according to the distance of characteristics between the donor and the receptor, the greater the distance, the lower the score obtained. Further ahead in this paper, the variables used in the process will be shown in detail as well as their respective weights. We also applied scores due to some characteristics of each respondent’s reading patterns. In other words, we also consider the behavioral analysis of that person.

When we have many candidates with the same distance for the same receptor, we always choose the first option in the randomic list, however we know that this point can be improved in the future through an optimization process, considering not only the best one-to-one match but also the best overall matching scores.

The “fused” panel is re-weighted afterwards, because we build, in practice, a sub-sample of the original sample, where this sub-sample, for each day, has to represent the same universe. Consequently, we were not able to reproduce exactly the same reach of the “recall” research for the period, but even so, we managed to generate very close results, considering the targets and days of the week. We will present the comparative results in the next sessions.

The first part of the process is to come up with a linked-list [7] with all individuals imported from the recall research, the donor’s list. In this study, we are working with 19.768 cases, which will yield 39.546 donor individuals, seeing that we are considering that yesterday’s response is one individual and the day before yesterday’s response is another. This list is randomly organized, in order to avoid a predetermined sequence of any given variable. This randomized organization was built in such a way that we can generate the same individual combination at any given moment.

The fusion algorithm is prepared to create a panel of x days. Our exercise was performed with $x= 14$ that, respecting medium characteristics and original sample size, offers the best relation.

The second step is to create 14 empty lists, one for each day of the panel, where we will store individuals from the fused panel, and we calculate the maximum reach indexes using the recall information, for each Press media we consider in our database, as the endpoint of the reach curve we want to accomplish.

For each newspaper we create a cluster of readers using the previous index calculated by the recall survey software analysis, and use them to create a sub-panel through the fusion algorithm. This process is then repeated for every newspaper considering the superpositions among them and the *non readers* at all. This process will yield a very close reach index to that generated by the recall study itself.

Eg.: In the fusion process for a specific newspaper, suppose we have a 20% reach, with no superposition, so we have a sub-panel of approx. 7.907 individuals that is obtained by $(39.536 * 0.2)$. We attribute the first individual to one of the 14 lists, depending on his cooperation day, for instance, if it is a Monday, we attribute it to the first list, if it is a Tuesday, to the second list and so forth. This individual now also becomes a receptor.

Among the remaining donors, now totaling 7.906, we look for the shortest distance (according to the process described in detail in the next item) for the next available day. We repeat the process until 14 days are completed.

As we have to consider all the superpositions among these newspaper, a specific algorithm based on optimization techniques was adapted to solve this problem.

Table 1
FUSED DATABASE AND RECALL DATABASE DURING THE FUSION PROCESS

Fused Database					Recall Database		
Id	Day 1	Day 2	...	Day 14	Id	Day	Id Fused
1	16580	23929		853	22348	7	2
2	18374	3824		22348	16580	1	1
3	23114	3331		18616	18374	1	2
...					22770	6	
(m)					23929	2	1
					853	7	1
					1825	5	
					23114	1	3
					3824	2	2
					18616	7	3
					3331	2	3
					...		
					(n)		

In this case we can notice that for individual 1, from the fused panel, we used individuals 16580, 23929, ..., and 853 from the recall panel to fill out days 1, 2, ..., and 14 respectively. These individuals are discarded from the donor individual list. This process advances until there isn't any individual available in the recall list.

The efficiency of the algorithm (execution speed) was not a concern factor, because tests demonstrated a very comfortable performance. This fusion process we are describing in this paper took a little bit more than 50 seconds to be performed, using a Pentium IV PC with 1.8 Mhz of speed.

3.2 Variables and Distance Metrics:

This fusion process considers two types of variables for the analysis, demographic (the individual characteristics) and behavioral (relative to the consumption of the measured medium). The demographic variables are as follows: socio-economic class, sex, age, day of the week in which the cooperation occurs, and geographic region. These variables are the same as the variables used in the traditional recall survey. The behavioral variables are obtained during the process of the raw data import. For this study, we use two types of behavioral variables; the recent reading declaration and total number of different newspaper read.

For all demographic variables we use the concept of weighted Euclidian distance, which may change according to the importance of each variable, as shown in the equation (1), where (k) represents the number of variables (demographic and behavioral), (i) is the recipient individual and (j) is the donor individual.

Equation 1
DISTANCE EQUATION

$$d_{ik(j)} = \sqrt{\sum_{k=1}^p (X_{ik(i)} - X_{jk})^2 W_k}$$

For the last behavioral variable we measure the distance between the donor and receptor comparing the number of newspaper read and the titles. We noticed that this is a very important variable in the process, since an alteration in its weight value may substantially affect the total reach resulting from the panel. As we control the sub-panel of specific titles we have to consider also the null reading in a way to produce the expected reach.

**Tables 2 and 3:
DEMOGRAPHICS AND BEHAVIOR VARIABLES IN DETAIL, AND THEIR RESPECTIVE WEIGHTS.**

Demographics					Behavior	
Class	Sex	Age	Region	Day of Week	Qty Titles	Same Titles
w=2		w=1	w=1	w=4	w=1	w=5
AB	M	10-14	Center	Monday	1 to 5	1 to 99
C	F	15-19	Nort	Tuesday		
DE		20-24	East	Wednesday		
		25-29	South	Thursday		
		30-34	West	Friday		
		35-39	N+L Suburbs	Saturday		
		40-49	S Suburbs	Sunday		
		50-59	W Suburbs			
		60+				

3.3 Matching Success Rate:

Statistical matching is about finding similar respondents considering a list of variables. It may turn out inadequate if there is no perfect match, so respondents are brought together with some differences in the corresponding variables.

The fusion algorithm performance may be monitored and assessed in regards to the ability to join similar individuals, given the elected variables for the fusion.

Obviously, this doesn't guarantee that the final result of the fusion process is suitable for the proposed objective (to join different individuals who, from a press consumption standpoint, may be considered as if they were the same person).

Therefore, apart from controlling the success rate of individual matching, it is necessary to evaluate the quality of the resulting information.

In the following table we can analyse the successful matching rate by each variable in this fusion process. We can observe that the success rate varies according to the weight given to a specific variable. In this study, the greatest concern was maintaining the good ratings for the socioeconomic level, sex and day of the week

**Table 4
LIST OF VARIABLES USED IN THE FUSION PROCESS AND THEIR MATCHING STATUS.**

Variable	% Successful matching	Weight
Class	88,50%	3
Sex	93,14%	2
Age	72,27%	2
Region	79,60%	2
Day of Week	82,65%	2
Behavioral	Not measured	5

The weights for the different variables were attributed according to their relevance for the preservation of the quality of the end product.

Social class and gender are determinant for the behavior and are relevant for the subsequent utilization of information for the media planning. Age and geographic region are relevant, however, they allow more flexibility.

Variable "number of different titles" was included in the list of matching variables to reduce a possible reach bias, by the attribution of a greater titles behavioral diversity for each individual.

"Titles" code demanded a more careful study, as in the fusion process, when joining different people, can "create" new individuals of "hybrid" behavior. By favoring the individual matching of those individuals who stated to have read to similar Titles in the period actually covered by the "recall", offers a guarantee for minimizing this bias.

It's important to mention that, according to the results of regular research, the great majority of individuals consume a reduced number of Titles in a very systematic manner.

Consequently, the development of a more intelligent algorithm was the key for handling different titles and its consumption combinations due to not produce a different pattern of individual reading, that would produce different results of a recall survey and consequently not being a good tool for planning different titles.

Next table shows matching quality evolution along days of panel. As expected, quality decreases day by day, due to the reduction of available donors, but is still at an acceptable level in a 14 days panel.

Table 5
MATCHING SUCCESS ALONG THE PANEL DURATION

Days	% Successful matching
1	90,1
2	89,3
3	88,9
4	88,5
5	88,3
6	87,5
7	87,3
8	84,8
9	84,0
10	83,6
11	83,3
12	83,4
13	82,6
14	82,5

4. Comparison

4.1 Recall Survey Versus Fused Panel Survey

The most important figures to compare between the recall research and the fused panel are Reach and Average Reading per Issue for each title.

Table 6 shows that the created panel preserves the characteristics of the recall survey.

Reach figures are lower for fused panel because we are comparing a 14 days reach period (fused) with “habitual” reading (recall – use to read). Differences are greater for those titles with lower level of subscription (more eventual reading).

Average Reading per Issue numbers is the same, as expected due the building process.

Table 6
COMPARING REACH AND READING PER ISSUE FIGURES FOR MAIN TITLES

TITLES	REACH		AVERAGE READ	
	RECALL	FUSED	RECALL	FUSED
TOTAL NEWSPAPER	61,4	64,8	15,66	15,64
FOLHA DE SAO PAULO	21,4	20,8	4,67	4,66
O ESTADO DE SAO PAULO	16,3	15,3	3,59	3,59
DIARIO DE SAO PAULO	14,3	12,4	2,17	2,16
AGORA SAO PAULO	9,7	9,0	2,15	2,14
LANCE!	5,9	4,5	1,08	1,08
DIARIO DO GRANDE ABC	4,8	4,1	0,75	0,74
JORNAL DA TARDE	4,2	4,1	1,03	1,01
JORNAIS DE BAIRRO	4,0	2,5	0,30	0,25
METRO NEWS	3,1	2,4	0,39	0,38
PRIMEIRAMAIO	1,9	1,1	0,13	0,12
GAZETA MERCANTIL	1,3	1,2	0,25	0,25
FOLHA GUARULHOS	1,0	0,6	0,13	0,13
O DIARIO DE MOGI	0,7	0,7	0,18	0,17

4.2 Fused Panel Versus Single Source Study

It is also interesting to compare some results from the fused panel obtained through this method with recent reading figures. For the same region and similar time frame, we have the possibility to look at TGI (Target Group Index) press results.

For this comparison we've choose two main titles: "Folha de São Paulo" (FSP) and "O Estado de São Paulo" (OESP). Due differences of methodology, total reach figures and reading per issue are higher in TGI than in recall study for both titles, as expected in recent reading studies.

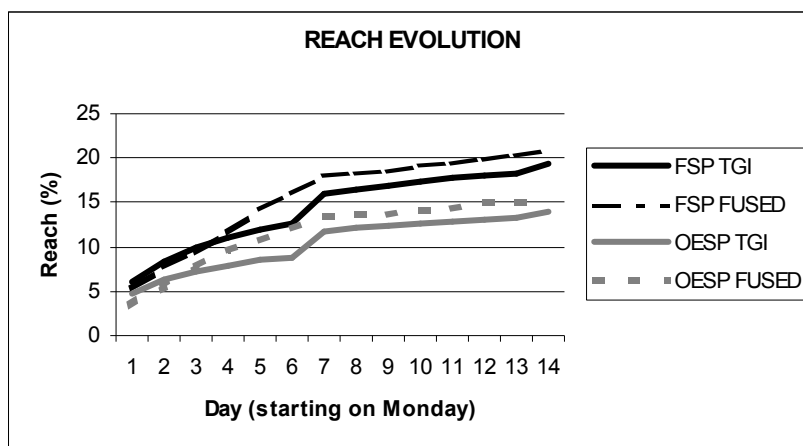
Table 7
COMPARING TGI AND RECALL FOR TWO MAIN TITLES

Main Figures (Age 10+)	TGI	Recall
FSP Mo/Sa - 1 edition	6,04	4,65
FSP Sun - 1 edition	9,86	4,74
FSP Tot Reach	29,70	21,37
OESP Mo/Sa - 1 edition	4,70	3,56
OESP Sun - 1 edition	7,69	3,74
OESP Tot Reach	22,00	16,33

But it is interesting to notice that reach evolution curve (14 days period) is similar, with some advantage for Fused Panel. TGI reach curves, using a model based on frequency of reading, are more smooth and continuous, with a strong jump on first Sunday.

Fused Panel curves are more "real shape", starting at a reasonable lower point but passing over TGI curves after the fourth day, for both titles.

Table 8 (and graphic)
COMPARING REACH EVOLUTION FOR TWO MAIN TITLES



Reach Evolution (Age 10+)	FSP		OESP	
	TGI	Fused	TGI	Fused
Monday	6,0	5,3	4,7	3,5
Tuesday	8,4	7,6	6,3	5,7
Wednesday	9,9	9,5	7,3	7,9
Thursday	11,0	11,8	7,9	9,7
Friday	11,8	14,2	8,5	10,9
Saturday	12,5	16,1	8,9	12,2
Sunday	16,0	18,1	11,8	13,4
Monday	16,5	18,3	12,1	13,7
Tuesday	17,0	18,5	12,4	13,8
Wednesday	17,4	19,1	12,6	14,2
Thursday	17,7	19,3	12,8	14,5
Friday	18,0	19,8	13,0	15,0
Saturday	18,3	20,2	13,2	15,1
Sunday	19,3	20,8	14,0	15,3

4.3 Campaign Evaluation (TGI x Fuse Panel)

We've also compared, just for a short example, the result of an evaluation of a typical newspaper campaign on both studies, using same titles, for two target groups: Total Population and High Level 18+.

The fact that results are very similar is not a coincidence. With this number of insertions per titles (six) and including Sundays, fused panel and TGI tend to offer similar results. Excluding Sundays or increasing insertions would lead to more "favorable" results within fused panel.

Table 9
COMPARING A TYPICAL CAMPAIGN – TGI x FUSED PANEL

Campaign Evaluation - Reach	Age 10+		AB 18+	
	TGI	Fused	TGI	Fused
14 days, 4 ins/title Mo/Sa and 2 ins/title Sundays (OESP and FSP)	25,0	24,9	51,6	50,3

5. Conclusions

Our experience conducting "recall" studies and "recent reading" (TGI) studies in Brazil leads us to the conclusion that the recent reading methodology normally offers greater "total reach" figures, but market recognizes this as an overestimation.

On the other hand, building a fused panel from recall study may be an interesting contribution for campaigns evaluations (reach and frequency), giving its "real shape" results.

To try to obtain a result that is closest to reality is always the goal of a fusion system. It's also important to find generalized solutions, in other words, systematically applicable in all situations.

The solution we have introduced in this paper serves this need in a way that we consider these challenges acceptable. Although it requires improvements, it is indicative that there is a possible alternative for the creation of Press virtual panels based on "recall" research.

We have noticed that the use of behavioral variables are of great importance for the process, and we intend to continue improving the use of these variables, even though some demographics determine the profile of a consumer.

Good results can be obtained through choosing an ideal donor and this fact, undoubtedly, will lead to the development of more advanced fused methods, match up to the method used by Soong [8] in his audience and TGI data fusion study for Mexico City.

References:

- [1] Radner, D.B., Allen, R., Gonzales, M. E., Jabine, T.B. and Muller, H.J. (1980) Report on Exact and Statistical Matching Techniques, *Statistical Policy Working Paper 5*, US Dept. of Commerce, Washington DC, US Government Printing Office.
- [2] Willard L. Rodgers (1984), An evaluation of statistical matching, *Journal of Business and Economic Statistics*, 2-1, 91-102.
- [3] Crivellaro A., Ferreira A., (2002), Building a Radio Panel Using Fusion and Recall Surveys, *Esomar, Week of Audience Measurement*.
- [4] Sudman, S., Bradburn, N.M., (1973), Effects of Time and Memory Factors on Response Surveys, *Journal of American Statistical Association*, V 68, 344, 805-815
- [5] Mathiowetz, N.A., The Effect of Length of Recall on the Quality of Survey Data, *Joint Program in Survey Methodology*, University of Maryland
- [6] Barr, R.S. and Turner, J.S. (1980), Merging the 1977 Statistics Income and the March 1978 Current Population Survey. Washington DC, Dept. of the Treasury.
- [7] Sedgewick, R., (1998), Algorithms in C, 90-109, Addison-Wesley.
- [8] Soong, R. Montigny, M. (2001), The Anatomy of Data Fusion, *Worldwide Readership Research Symposium*, 87-109.

