Thomas Puliyel, Sankara Pillai, Ashutosh Sinha
Indian Market Research Bureau
Bombay, India

MAKING DATA SETS COMPATIBLE FOR DATA FUSION : RESULTS OF AN INDIAN EXPERIMENT

MAKING DATA SETS COMPATIBLE FOR FUSION: RESULTS OF AN INDIAN EXPERIMENT.

1. Introduction

Data fusion has been suggested as a method for resolving the difficulty of creating a single-source database that can handle any queries posed by a marketer or media planner. Fusion is an attractive idea, since it enables comparisons across different types of studies (eg. combining data from continuous panels and one-shot surveys), and helps reduce the demands made on an individual respondent.

When fused data bases first became available, there were many who had doubts about their validity. Till now, we have been accustomed to the idea of using information on many traits to predict behaviour on one aspect. Fusion is a case of information on a few traits being used to predict behaviour on many different variables. The question asked was: Suppose we collect real data on, say, media habits and then replaced this with fused data from a donor sample. Will we then get the same results ?

A number of experiments have since been carried out with this in mind, including one by the present authors (Thadani & Sinha, 1988), and the fusion of TGI data on BARB in Britain (O'Brien, 1990).

In this paper, we have set out an account of a fusion experiment carried out in India. We begin with a summary of our previous work and its implications. We then go to the present experiment; observed versus fused data using alternative sets of bridge variables. Lastly we have shown the results for a media plan using real as well as fused data.

2. The previous experiment

Indian Market Research Bureau (IMRB) runs a diary based panel to monitor television programme viewership, the Television Rating Points (TRP) System. The TRP is currently operational in nine cities, with a total sample size of 3100.

In our previous experiment, we collected data on readership, as well as a number of potential bridge variables from TRP panel members in four cities. The TRP panel was then split into two equal sub-samples, each half serving as donor and recipient for both readership and television viewership data.

By hindsight, some of the 'lessons' from this early experiment now seem rather obvious. Nevertheless they are valid and worth recalling.

(1) Fusion is more accurate in predicting the type of publications read (or television programmes watched) than in predicting the audience for a particular publication or television programme.

(2) Television viewership data tends to be less strongly correlated with product use or demographics. If a multimedia database is to be created, it makes better sense to transfer television viewing from a donor sample to a recipient readership database. In television viewing there are fewer 'skews' that have to be reproduced, making the fusion task a lot easier.

3. The present experiment

In this experiment, too, we have used the TRP panel. However, this time we have not attempted to fuse readership data. Instead, we have concentrated our efforts on the trasfer of television viewing behaviour onto recipient observed readership and product use data.

First, a word about the viewership data that has been transferred. Fusion has not been carried out for the full mass of ratings on individual programmes telecast. Drawing on our previous work, we thought it would be better to borrow data on frequency of viewing for programme types rather than particular telecasts.

3.1 Validitity of viewing frequency data.

As in the case with most TV audience measurement systems, the TRP data is too voluminous to be used effectively. Thus, our initial efforts were focussed on meaningful groupings of programmes based on their content, format, time slot and other factors influencing viewing behaviour. Once this grouping was done, the observed programme viewing data was converted into frequency of viewing for the programme type. With this step, two major advantages were obtained. Firstly, the database itself was reduced to a more manageable size. Second, and more significantly, the data became compatible with available readership data and many of the existing press schedule evaluation models could easily be used.

However, the first test was to ascertain whether the data reduction process, using frequency data would provide less reliable reach and exposure estimates. The estimated reach and OTS distributions of over 200 media plans were computed using the frequency data as well as the observed data. The findings as presented in Table 1 support the point of view that the lessons learnt from readership research can be successfully transplanted onto television. We believe that the use of frequency of viewing data would be even more worthwhile in countries where the TV viewership data tends to be voluminous due to the longer telecasting hours and multiplicity of channels.

Table 1

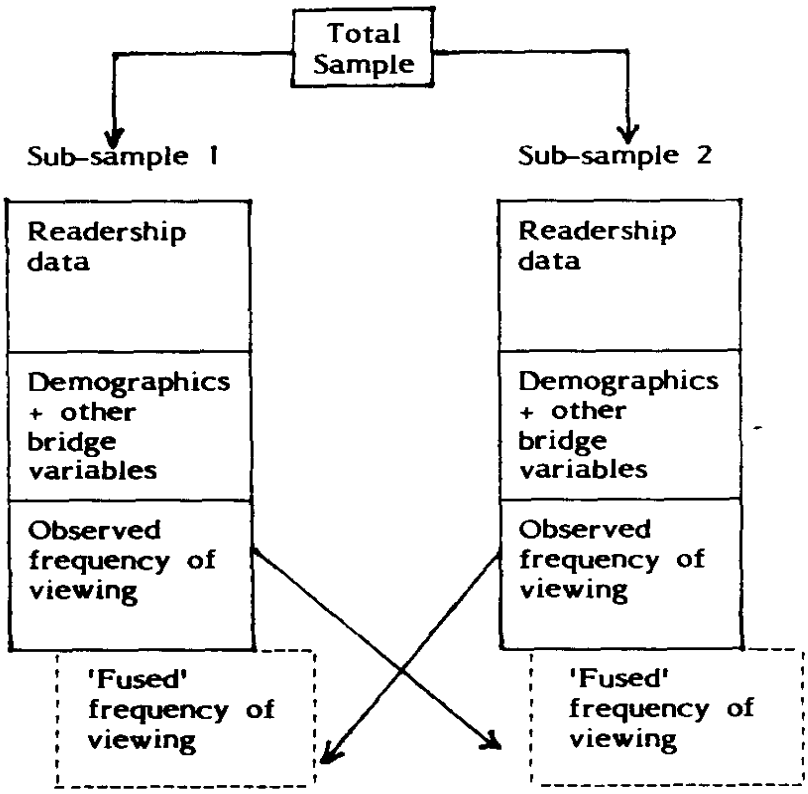
Reach and Exposure : Comparison of estimates based on frequency model with observed data

	% of media evaluations
Variation in reach	
0 - 1%	34
1 - 2%	36
2 - 4%	18
4 - 6%	10
6% +	2
Variation in gross exposures	
0 - 2%	33
2 - 4%	47
4 - 6%	14
6% +	6

3.2 Data preparation for fusion

We administered a questionnaire that covered readership, demographics and usership data (for a few products). Information was also collected on a number of potential bridge variables other than demographics; such as claimed heaviness of television viewing, languages understood, and the claimed frequency of viewing of a few types of programmes (eg. "programmes based on Hindi film songs like Chhayageet and Chitrahaar"). The experiment covered two TRP centres, namely, Bombay and Madras. The total sample in this experiment was 744.

The sample was then split randomly into two halves. One half of this sample was used as a donor and the second half as recipient. The process was repeated once again, this time with the second half serving as the donor. In this way, we have a fairly large sample available for our tests of validity. Diagrammatically, this can be expressed as follows :



The series of steps for isolating bridge variables and for carrying out the experimental fusion is described below.

3.3 The search for explanatory variables

A factor analysis was carried out on the TRP frequency of viewing data to identify the co-ordinates of television viewing. The following key programme types were identified :

In Bombay	In Madras
Marathi language viewing	News, topical and sports
English language viewing	Tamil cinema based
Morning transmission	Morning transmission
News bulletins (across languages)	Hindi language viewing
Hindi serials	Tamil language/second channel viewing
Hindi cinema music	

Similarly, claimed frequency of viewing data was also factor analysed to identify 'traits' (as far as response to the question). Thereafter, a number of analysis including analysis of variance and regression were carried out with the viewing 'constructs' described above as the dependent variables.

Some of the results are worth reporting. It is found that claimed frequency of viewing is often a superior predictor of viewing behaviour as compared with demographic variables. Two examples of this (drawn from Bombay) are presented below.

	Percent of all viewing accounted for by the		
English language programmes	**Top 25%**	**Top 33%**	**Top 50%**
Education (7 levels)	40	51	71
Claimed frequency - English news (4 levels)	66	73	82
Hindi cinema music programmes			
Age x sex (6 levels)	25	33	50
Claimed frequency - Hindi film music (4 levels)	31	41	60

Seeing how promising the 'claimed frequency' variables were, it was decided to carry out and compare two sets of marriages in Bombay - one based entirely on demographics and the other based on some demographic and some claimed frequency variables. The variables employed in the two sets of marriages, Demographics and Demographic Plus are listed below :

	Demographics	Demographic Plus
For cell creation	Television ownership Mother tongue Gender	Television ownership Mother tongue Gender
For computing distances	Age Education	Age Education Hours watch TV on average weekday Claimed frequency - Hindi serials Claimed frequency - English programmes

Marriages were carried out within the cells described above (eg. male, Marathi-speaking, television owner). The final matching was based on Euclidean distances after normalising the remaining variables.

In this experiment, we have used a donor just once as far as possible. In some situations, however, donors were used more than once, because of unequal cell sizes across the two sub-samples.

3.4 Testing for validity

Early tests of fusion validity placed a great deal of emphasis on the successful replication of marginal totals. However, we now feel that a better and more realistic test would be one of selectivity as suggested by Rothman (1988) and others. Essentially, we would like to see the relationships that exist in the original data to be resurrected, to the extent possible, in the fused data. For example, if younger people have a less-than-average heaviness of viewing, the phenomenon should not be lost in the fusion process. Nor would we want a selectivity to appear where none really exists.

In these two-way tabulations, the variables of interest would be :

(1) Demographic variables which would almost invariably form part of the target group definition.

(2) Product use which may be specified by itself, or together with demographic variables.

(3) Readership data to examine the relationships between programme viewership and readership. We do find a nexus between the two that cannot be explained by demographic variables alone. For example, the reader of a news magazine has a substantially higher viewership of news and topical television programmes.

Observed and predicted selectivity was measured for 11 television programme types in Bombay and 12 programme types in Madras. Further, we have also compared the performance of media evaluation runs using observed and fused frequency of viewing data.

3.5 Selectivity : Observed versus fused

In India, the television scene is different from that in Europe and other parts of the world. Since channel options are limited to two, there is very little audience fragmentation. Popular programmes reach upto 75% of the audience with just one telecast. As a result, the selectivity in frequency of programme viewership is, in general, lower than what would be expected in other countries.

Thus, the real test of fused data would be in ensuring that whatever selectivity exists should not be regressed to the mean and thus lost. We analysed the selectivity indices for the 11 programmes in Bombay across eight target groups defined on demographics, reading habits and product usage. Out of the 88 indices created, the selectivity that existed in the original data was retained or marginally better emphasised in 63% of the cases while using the 'Demographic' fusion. This improves by another 5 percent in the case of the 'Demographic Plus' fusion involving additional bridge variables of viewing frequency. In Madras too, a similar level of success was obtained through the 'Demographic Plus' marriages.

Table 3

Changes in selectivity : Segmentwise analysis in Bombay

Segments	'Demographic' Fusion		'Demographic Plus' Fusion	
	selectivity worse than observed	selectivity same or better than observed	selectivity worse than observed	selectivity same or better than observed
Men	5	6	5	6
Youth	4	7	4	7
Matriculates	3	8	3	8
Upper income	4	7	4	7
Reader of any Marathi publication	4	7	3	8
Reader of any English magazine	4	7	2	9
Toothpowder users	3	8	3	8
Cartridge shaving system users	6	5	4	7
No. of cases	**33**	**55**	**28**	**60**
Percentage of cases	**37%**	**63%**	**32%**	**68%**

Table 4

Changes in selectivity : Segmentwise analysis in Madras

Segments	**No. of programmes in 'Demographic Plus' where selectivity was ....**	
	worse than observed	**same or better than observed**
Men	-	12
Youth	3	9
Matriculates	4	8
Upper Income	4	8
Reader of any Tamil film magazine	4	8
Reader of any English magazine	7	5
Toothpowder users	5	7
Cartridge shaving system users	4	8
No. of cases	**31**	**65**
Percentage of cases	**32%**	**68%**

3.6 Cumulative reach and Schedule Evaluation : Fused versus observed data

One other test that was carried out to check the validity of the fused data was its usage for estimating cumulative reach of programme types.

Using the TRP frequency data the expected reach of eight programmes were estimated employing the observed frequencies as well as the frequencies obtained through both approaches to fusion.

The eight programmes selected for this exercise included programmes of varying subjects, content and viewership levels. The expected reach figures were computed for a single episode, ten episodes and the 'plateau levels'.

The results reveal that both approaches to fusion give quite satisfactory estimates of the observed data. The variation from the observed data tended to be greater for very small viewership programmes, but even in such cases, the absolute variation was within acceptable limits. Also, the percentage variations tended to decrease further with a increase in the number of spots. Among the two methods, the 'demographic plus' method with viewing habits data gave marginally better results, and perhaps more significantly, consistently improved results. Out of the 24 estimates presented in Table 5, the 'Demographic Plus' method gives a smaller variation from the observed data in 22 cases.

Table 5

Cumulative Reach : Observed versus Fused data

(Fig. in 000s)	Observed Data	'Demographic' Fusion	'Demographic Plus' Fusion
Programme/Type:			
Reach of a single telecast			
9.00 pm serial	27.9	25.6	26.7
Rangoli	16.2	17.9	16.2
Chitrahaar	29.0	32.7	28.6
Marathi serial	5.6	8.1	5.8
Hindi News	15.2	13.3	13.4
World of Sport	2.8	4.0	2.0
Sunday Live Sports Telecast	3.2	3.8	3.2
The World This Week	10.5	12.4	8.9
Cumulative reach of ten telecast			
9.00 pm serial	44.7	44.9	44.0
Rangoli	32.9	31.7	32.4
Chitrahaar	42.3	46.3	42.6
Marathi serial	20.4	24.5	19.8
Hindi News	27.3	25.3	26.9
World of Sport	6.2	8.5	6.2
Sunday Live Sports Telecast	14.6	16.8	13.9
The World This week	25.3	25.5	21.0
Maximum cumulative reach (Plateau level)			
9.00 pm serial	48.1	50.3	47.9
Rangoli	38.9	38.4	40.0
Chitrahaar	47.2	49.8	45.8
Marathi serial	37.0	40.7	33.9
Hindi News	40.1	37.0	39.8
World of Sport	9.9	10.4	10.0
Sunday Live Sports Telecast	21.0	24.4	19.6
The World This Week	32.0	30.0	26.9

Similar results were also noticed when full media schedules were compared, with the 'Demographic Plus' approach giving consistently better results. The results hold good for product categories with narrowly focussed target group definitions as well as those with broad, mass market definitions.

One of the cases considered was the media plan for a washing machine. Washing machines, being luxury items in India, normally have their advertising focussed on the upper-upper income group. The target group was defined as women, in the 25-44 age group, with household incomes over Rs.2500 p.m. The plan that we considered had 15 spots in Bombay and 17 in Madras. In both cities, there were a mixture of national as well as regional programmes. While both approaches performed equally well on reach, the 'Demographic Plus' version was better able to predict exposure values.

The other case presented here is of a FMCG. The results point in the same direction as in the previous case, thus providing further evidence of the superiority 'demographic plus'.

Table 6

Summary results of the schedule evaluations

	Observed Data	'Demographic Fusion	'Demographic Plus' Fusion
Washing Machine :			
Reach (000s)	23.3	25.2	25.5
% reached	96.1	95.7	97.2
Gross Exposures (000s)	112	125	113
Avg. Exposure	4.45	4.97	4.42
Toothpaste:			
Reach (000s)	273.1	274.5	272.5
% reached	96.2	96.6	95.9
Gross Exposures (000s)	1176	1764	1774
Avg. Exposures	6.50	6.42	6.51

Conclusions:

To conclude, the Indian experience has been that the usage of frequency of programme viewing data provides very good estimates of reach and exposure, and in addition possesses the advantage of easier data base management. In effect, this is a case where concepts learnt in readership research have been applied to television.

The current experiment confirms the suitability of the frequency of programme viewing data for the fusion process. It was found that claimed frequency of viewing is often a superior predictor of viewing behaviour as compared with demographic variables.

The experiment also shows that using variables pertaining to broad viewing habits as bridge variables enhances the selectivity as well as provides reliable estimates of the observed viewership data. We believe that this experiment which uses frequency data as input for fusing databases will enable easier comparisons across readership and television audience measurement systems thus permitting multi-media comparisons and evaluations.

REFERENCES

O'Brien, S., (1990) "The Role of The Data Fusion in Actionable Media Targetting in the 1990's", The 43rd ESOMAR Marketing Research Congress, Monte Carlo.

Rothman, James, (1988) "Testing Data Fusion" ESOMAR Seminar on Media and Media Research, Madrid.

Thadani R P, and Sinha A, (1988) "A Study on the validity of Data Fusion: The Indian Experience". Readership Research : Theory & Practice Proceedings of the Fourth International Symposium, Barcelona.