Media:Time Integrated Planning Database

Technical Description

A report prepared for:

Media:Time

April 2017

RSMB Limited, Savoy Hill House, 7-10 Savoy Hill, London WC2R 0BU
Tel: + 44 (0)20 7808 3600 Email: contact@rsmb.co.uk Web: www.rsmb.co.uk
Registered office as above. Registered in England and Wales No. 2173860

Contents

1. INTRODUCTION

This document contains a technical description of the data fusions/integration conducted by RSMB on behalf of MediaTime (Netherlands) for an integrated planning database (IPD). The objective was to create a database which reflected the functionality and audience levels provided by the main media trading currencies in the context of multi-media reach and frequency. To assist in creating this database, MediaTime undertook a Media Survey which was based on a questionnaire and one week 'time use' diary that collected media habits.

2. THE PRINCIPLE OF DATA FUSION

The principle of the fusion is to attach an individual's donor responses to each recipient respondent by finding the donor respondent whose characteristics are closest to that of the recipient respondent. These characteristics are called linking variables or hooks and have different levels of importance depending on how well they discriminate between the key outputs on each media currency.

The success of a fusion is dependent on many factors including:

- the number of hooks available
- the consistency of definitions (how similar the questions were on the surveys and how well the profiles matched)
- how well the hooks can predict the key output variables
- how sample design structures are accounted for
- donor/recipient ratios

3. TECHNICAL DESCRIPTION

3.1 Background

Normally, when fusing two industry media databases the importance of the linking variables can only be assessed separately with respect to each media. However, the MediaTime survey provided a single source database that can help to understand relationships between different media and meant therefore that the importance of linking variables could be assessed with respect to media interactions.

The methodology applied was fuse the MediaTime dataset onto the Press (NOM) survey to produce the hub survey onto which the industry samples for Individuals 13+ could each

in turn could then be fused to. The NOM (Nationaal Onderzoek Multimedia), the national readership survey in the Netherlands, was deemed suitable for this purpose. As the media response variables are required for the other industry fusions the MediaTime/Press (NOM) fusion is undertaken first.

3.2 Datasets

The following datasets were used for these fusions in order to tie in with the main fieldwork period for the MediaTime survey:

Survey	Data Period	Survey Sample Size
MediaTime Survey	14 Sept to 18 Oct 2015	2,904 individuals
Press (NOM)	2015-II/2016-I	17,151 individuals
Television (SKO)	31 Aug – 22 Nov 2016	3,003 individuals
Radio (NLO)	1 Sep – 31 Oct 2015	7,535 individuals
Internet and Apps (DDMM)	2 May 2016 – 5 June 2016	8,158 individuals
Outdoor	March 2017	N/A

For the daily reporting surveys the samples used in their respective fusion were reduced so that only respondents who reported sufficiently were used in the fusion. This decision on selecting the sample cut-off criteria is subjective and a trade-off between using a larger sample size and having as much 'full' respondent media data as possible. Below are the sample sizes used in the fusions and the criteria used for creating this sample:

Survey	Fusion Sample Size Individuals	Fusion Sample Size Households	Reporting Criteria
MediaTime Survey	2,818	2,818	Reported on all seven days of the diary.
Television (SKO)	2,286	1,237	Individuals 13+ and reported for at least 63 days out of 84
Radio (NLO)	7,939	7,939	Individuals 13+
Internet (DDMM)	7,026	5,812	Individuals 13+ and responded for 28 out of 35 days
Press (NOM)	17,151	17,151	-
Press (NOM – for TV)	16,533	16,533	Respondents that watch TV

Press (NOM – Internet)	16,674	16,674	Home Internet Users
Outdoor (NOM)	16,188	16,188	Individuals Aged 13-75

3.3 Linking variables

A key aspect in determining the success of any data fusion is the number and compatibility of the linking variables. Linking variables need to have similar profiles otherwise the information carried across can become skewed.

For the fusions Sex and the 12 Provinces were set as critical cells. This meant an individual was always matched to someone in the same province and of the same gender. The reason Sex and Province were selected as critical cells is that Sex is the main discriminator in media behaviour and geography is used so that regional media behaviour is carried across from the donors.

The other linking variables related to demographics and media behaviour, see section 3.5; the latter would assist in helping to preserve the cross media relationships.

3.4 Media Imperatives

As well as containing demographic and geographical linking variables, after fusing the MediaTime to Press (NOM) the dataset then contained useful media contact data for television, radio and internet each of which can be utilised as linking variables in order to enhance the fusion process.

For each of the major media industry surveys, consistent media imperatives were required for that media for recipients on the MediaTime survey and also donors on the industry respondent dataset that would describe each individual's behaviour.

In order to compensate for survey measurement differences, the industry response variables were calibrated to the MediaTime average for that measure. This was done for contacts for that measure on each survey.

To create respective media linking variables to be used for each fusion, a principal components analysis (PCA) was performed for the appropriate media response variables for respondents in the MediaTime survey. To use all the given media's response variables as links would be impractical as they would be fragmented and highly correlated. The principal components analysis is a statistical technique that creates successive linear combinations of the data, with the first principal component explaining the most amount of variation (i.e. predicting power) in the response variables and each successive principal component explaining a diminishing amount. This technique allows a smaller number of

factors to be used than the original dataset as the first handful of principal components account for most of the variation in the individual's viewing behaviour.

The 15 principal components used in the Hub/Press fusion explained 53% of the Hub publication media imperatives. The 8 principal components used in the TV/Press fusion explained 73% of the Hub TV media imperatives whilst the 8 principal components used in the Radio/TV fusion explained 75% of the Hub Radio media imperatives.

3.5 Importance Weights

The MediaTime survey has the advantage of containing a wealth of media consumption data so for each fusion the importance of a linking variable can be assessed with respect to an individual's cross media consumption.

The variables used were from the diary are in Appendix 1 A1 and included the following:

- Television viewing by station by week and day part
- Radio listening by station by week and day part
- Reading by title by week part
- Internet by genre by week part

Collating these variables together, a factor analysis of consumption was then performed. A factor analysis is a statistical multivariate technique that creates independent variables from a multivariate dataset, which can then in turn be analysed as correlations are accounted for. This factored dataset effectively contained variables that described each individual's patterns of media consumption.

Using Analysis of Variable techniques (ANOVA), each linking variable was assessed to examine how much of the variability of these patterns could be explained solely by that linking variable. In this way a hierarchal set of importance weights could be created reflecting the importance of the linking variable in discriminating patterns of media consumption. These weights were then used in the fusion matching algorithm to ensure the factors that were most important were highly matched.

3.6 Fusion Process and Matching

The fusion process attempts to find exact matched for recipients for the donor pool in terms of all the linking variables. In practice exact matches cannot be found across all linking variables and the next best match is found using the concept of statistical distance. The distance measurement used is Mahalanobis' distance. This distance is used as it accounts for the following:

- Scaling of hooks

- Correlations between the hooks

In addition this distance is weighted by the aforementioned importance weights to recognise the predictive power of each linking variable with respect towards the data to be fused.

To ensure that there is no overuse of particular donors, constraints are built into the matching process. This is important as due to sample sizes for some of the fusions there was already an inevitable high average use of donors.

If each recipient was matched with its best donor then some donors may be used an unacceptably large number of times, reducing the effective sample size and causing these donor's responses to dominate any subsequent analyses. To prevent the use of a donor being used too frequently a multiplicative penalty weight is applied each time a donor is used so that the distance between respondents is artificially increased. This helps to reduce the frequency of a donor being used therefore giving a better effective sample size but at the expense of the possible best distance solution. To obtain the optimal fusion solution these penalty weights have to be adjusted iteratively to give an acceptable 'distance/effective sample' size solution.

3.7 Media probabilities

The objective of the process is a final fusion file that contains contact rates (likelihood of exposure) to all media.

For Print, this survey is effectively the hub survey so probabilities for reach and frequency can be calculated as they would in the standalone media currency.

For TV, the probability of viewing to a spot has been calculated for each of the donors based on their typical observed behaviour. After the fusion these have been carried across and then calibrated to published data within a demographic matrix (as far as sample sizes will allow) for each channel time segment itemised. If samples are small this segmentation has been reduced. An algorithm employed in the calibration process ensures probabilities are not greater than 1.

The segmentation matrix is as follows:

AB1	*	13-34	*	Male
B2CD		35-54		Female
		55+		

For Radio, the process is the same as TV but the fusion is not calibrated to the published data but to the published data with a commercial listening time index applied. The commercial listening time indices are the advertising listening rates relative to all listening in the same timeslot based on 2013 data sourced from the NLO meter panel and GfK online database.

For Internet the contact rate is based on pageviews. This has been calibrated for each website required in the itemisation.

The probability of listening to a spot has been calculated for each of the donors based on their typical observed behaviour.

4. FUSIONS

The following sections detail the processes for the fusions that were undertaken.

4.1 MediaTime/Press (NOM) Fusion (Hub Survey)

The initial fusion is the expansion of the MediaTime survey onto the Press survey.

RSMB conducted a comprehensive assessment of all the potential linking variables. The final list for the MediaTime/Press fusion is:

- Sex (Critical Cell – Male, Female)
- Province (Critical Cell – 12 Provinces)
- Age (Actual Age)
- Urbanity (Rural/Urban)
- Household Size (1,2,3,4,5+)
- Head of Household (Males only – Yes, No)
- Highest Level of Education Completed (Lower, Middle, Higher)
- Working/Not Working
- Hours Worked per Week by CIE (30+, 8-29, 1-8, 0)
- Income Class (Low, Middle, High)
- TV Viewing: High/Medium/Low
- Television Viewing Frequency (0, 1, 2, 3, 4, 5, 6+ Days)
- Radio Listening: High/Medium/Low
- Radio Listening Frequency (0, 1, 2, 3, 4, 5, 6+ Days)
- Internet Usage: High/Medium/Low
- PC Device used for Internet (Yes, No)
- Laptop Device used for Internet (Yes, No)
- Mobile Device used for Internet (Yes, No)
- Tablet Device used for Internet (Yes, No)
- Games Console Device used for Internet (Yes, No)
- TV (Internet) Device used for Internet (Yes, No)
- Wealth (W1-W5)
- Social Class (A, B1,B2, C, D)
- Newspaper Readership: High/Medium/Low
- Magazine Readership: High/Medium/Low
- Weeklies Readership: High/Medium/Low
- Monthlies Readership: High/Medium/Low
- 15 Press Media Imperatives

Following the discriminant analysis the variables given the highest importance in the matching were Radio Listening, Radio Frequency, and Age. The full listing is in Appendix 1 B1.

The Press survey had 17,151 respondents and MediaTime had 2,818 respondents making the average number of times a donor was used around 6. The full fusion frequency distribution is in Appendix 1 D1. The improvement on random and matching is in Appendix 1 C1.

4.2. TV (SKO)/Press (NOM) Fusion

Following the expansion of MediaTime onto the Press to create the Expanded Hub dataset the other media were then added.

RSMB conducted a comprehensive assessment of all the potential linking variables for the TV/Press fusion. The final list for the TV/Press fusion is:

- Sex (Critical Cell – Male, Female)
- Province (Critical Cell – 12 Provinces)
- Age (Actual Age)
- Household Size (1,2,3,4,5+)
- Presence of Children age 0 to 5 (Yes, No)
- Presence of Children age 6 to 12 (Yes, No)
- Presence of Children age 13 to 17 (Yes, No)
- Chief Income Earner Working/Not Working
- Urbanity (Rural/Urban)
- Head of Household Highest Level of Education (Lower, Middle, Higher)
- Head of Household (Males Only – Yes, No)
- Working/Not Working
- Highest Level of Education Completed (Lower, Middle, Higher)
- Radio Listening: High/Medium/Low
- TV Viewing: High/Medium/Low
- Frequency of using Twitter (High, Medium, Low)
- Internet at Home (Yes, No)
- Wealth (W1-W5)
- 8 TV Media Imperatives

The 8 media imperatives were derived as detailed in Section 3.5, in order to create hooks so that individuals could be matched on their TV viewing behaviour. For each individual, response variables were calculated from TV respondent level data and Press data.

Following the discriminant analysis the variables given the highest importance in the matching were Radio Listening, Age, and TV Viewing, along with some of the media imperatives. The full listing is in Appendix 1 B2.

The Press survey had 16,533 respondents (those who watched TV) and the TV sample had 2,286 respondents making the average number of times a donor was used around 7. The TV sample was defined by selecting respondents who reported sufficiently over the 12 week period in order to be able to define typical behaviour whilst not be too punitive so as to restrict the sample available. The full fusion frequency distribution is in Appendix 1 D2. The improvement on random and matching is in Appendix 1 C2.

4.3 Radio (NLO)/Press (NOM) Fusion

RSMB conducted a comprehensive assessment of all the potential linking variables for the Radio/Press fusion. The final list for the Radio/Press fusion is:

- Sex (Critical Cell – Male, Female)
- Province (Critical Cell – 12 Provinces)
- Age (Actual Age)
- Age of Head of Household (Actual Age)
- Urbanity (Rural/Urban)
- Household Size (1,2,3,4,5+)
- Household Composition (Single, Couple, With Children)
- Presence of Children age 0 to 5 (Yes, No)
- Presence of Children age 6 to 12 (Yes, No)
- Presence of Children age 13 to 17 (Yes, No)
- Highest Level of Education Completed (Lower, Middle, Higher)
- Respondent Hours Worked per Week (30+, 8-29, 8<)
- Head of Household (Males Only - Yes, No)
- Head of Household Highest Level of Education Completed (Lower, Middle, Higher)
- Head of Household Hours Worked per Week (30+, 8-29, 8<)
- Radio Listening (High/Medium/Low)
- 8 Radio Media Imperatives

The 8 media imperatives were derived so that individuals could be matched on their Radio listening behaviour. For each individual, response variables were calculated from Radio respondent level data and Press data.

Following the discriminant analysis the variables given the highest importance in the matching were Radio Listening and Age, along with some of these media imperatives. The full listing is in Appendix 1 B3.

The Press survey had 17,151 respondents and the Radio sample had 7,939 respondents giving an average number of times a donor was used at around 2. As with TV, the Radio sample was defined by selecting respondents who reported sufficiently over the 12 week period in order to be able to define typical behaviour whilst not be too punitive so as to restrict the sample available. The full fusion frequency distribution is in Appendix 1 D3. The improvement on random and matching is in Appendix 1 C3.

4.4 Internet and Apps (DDMM)/Press (NOM) FUSION

RSMB conducted a comprehensive assessment of all the potential linking variables for the Internet/Press fusion. The final list for the Internet/Press fusion is:

- Sex (Critical Cell – Male, Female)
- Province (Critical Cell – 12 Provinces)
- Age (Actual Age)
- Social Class (A, B1, B2, C, D)
- Urbanity (Rural/Urban)
- Highest Level of Education Completed (Lower, Middle, Higher)
- Chief Income Earner (Males Only – Yes, No)
- Working Status (Full Time, Part Time, Unemployed)
- Presence of Children age 0 to 5 (Yes, No)
- Presence of Children age 6 to 12 (Yes, No)
- Presence of Children age 13 to 17 (Yes, No)
- Number of Children (0,1,2,3+)
- Household Size (1,2,3,4,5+)
- Internet Weight of Usage (High, Medium, Low, None)

Following the discriminant analysis the variables given the highest importance in the matching was Age. The full listing is in Appendix 1 B4.

The Press Survey had 16,674 respondents as it was restricted to those who had used the internet in the last month. The Internet sample had 6,047 respondents giving an average number of times a donor was used at around 3. As with TV and Radio, the Internet sample was defined by selecting respondents who reported sufficiently over the 5 week period in order to be able to define typical behaviour whilst not be too punitive so as to restrict the sample available. The full fusion frequency distribution is in Appendix 1 D4. The improvement on random and matching is in Appendix 1 C4.

5. OUTDOOR

For Outdoor the complexities of the Outdoor methodology mean that integrating this into the planning database warrants a different approach from fusion.

For the industry for the Outdoor reach and frequency model the methodology uses (among other things):

Travel Survey
Traffic counts
Probability models
Extrapolation
Visibility adjustment factors
Computer simulations

It was clear that this would be too difficult to encapsulate all of these into the integrated planning database such that it would be able to deliver credible results for any campaign the user wished to run.

The solution instead is to replicate the reach and frequency results for a number of key typical schedules that could be selected by the user to assist in their tactical planning using this media.

The process would rely on identifying key determinants that discriminate contacts across general or specific campaigns. In discussion with the currency provided a key segmentation was designed based on Sex, Age Group, Education level, and Region.

The reach and frequency results of these schedules (0 to 10+) were be run using the industry software. In this way, all the complexities of the media measurement were accounted for.

Each NOM respondent has a segment based on Sex, Age Group, Education level, and Region. A respondent is assigned a probability of viewing a campaign 0 times, 1 time, 2 times, and so on up to 10 or more times. These probabilities are calculated from the proportion of the population in this segment that has viewed the campaign 0 times, 1 time, and so on in the campaign data file. The average contacts figure is also picked up from the campaign file.

In addition, due to small differences between the NOM and Outdoor segment population estimates a calibration routine is employed to ensure that these probabilities and the average contacts exactly match the original for All Individuals whilst preserving the discrimination of the segmentation.

For each respondent and for each campaign, there is a probability of viewing 0 times, 1 time, 2 times, and so on up to 10 or more times. There is also an average contacts figure.

Pointlogic

PointLogic required a separate (albeit weaker) solution that ran on a given rate. A Poisson model was fitted to the reach figure for each respondent. The rate constant for the respondent is therefore given by:

$$\lambda = -\ln(1 - reach)$$

Each respondent has a single rate for each campaign. By design this will match the reach figures produced by the standard solution. An index is applied for matching the contact frequencies.

6. OUTPUTS

Following the fusions, the next step is to create inputs which can then be used in a mixed media campaign.

Respondent level summary data will be held as personal probabilities in the fused database. These are based on the average behaviour for each of the donors in the relevant currency dataset.

These personal probabilities are equivalent to:

- Press – personal average issue readership
- TV – personal average GRP
- Radio All-Time and Commercial – personal average GRP
- Internet and Apps – rate of exposure

These probabilities or rates of exposure have been calibrated to currency levels for each of the breaks (channels, time of day etc.) in the fused database. The calibration is limited to a small matrix of demographics based on sex, age and social grade.

For Outdoor frequency distributions are pre-calculated for a range of typical schedules.

Due to issues in incorporating the above, for PointLogic a rate has been supplied representing average contacts.

Classifications from the NOM (e.g. demographics) can be used to create target groups.

The bureau documentation details the methodology to be used to create results for mixed media schedules from this database.

7. Validation

By definition there is no validation for mixed media campaigns using the fused database but we can assess the validity of the media probabilities in the database and methodology by comparing the results of single media campaigns using this against the published results when using currency data.

Reach and Frequency

Reach and frequency tests were completed for TV and radio using the supplied industry currency based campaigns. These tests are carried out to ensure that the industry currency has been preserved during the fusion process and the reach and frequency results are comparable to a 'typical' campaign.

TV

The TV fusion data was compared against 125 campaigns with total GRPs per campaign ranging from 13 to 1398. The sample campaigns were based on a variety of demographic targets – all adults, shoppers 13+, social grade A, men 20 to 49 and women 20 to 49. Details of the GRPs and reach % indices are below:

Measure	Average Index	Average Abs. Difference
GRPs	100	0
Reach (%)	104	2.75

See Appendix 2 for the full UAT

All key measures were well preserved during the TV fusion. As expected, GRPs were preserved exactly for all campaigns and reach % and frequency were preserved exceptionally well across all types of campaigns.

Radio

The radio fusion was compared with 8 radio currency based campaigns. This initial analysis was carried out on campaigns across 'All Adults' for 6 different time segments sets. The total GRPs per campaign ranged between 330 and 990. Details of the GRP and Reach % index are below:

	All Schedule Comparisons		Complete Schedule Comparisons	
Measure	**Average Index**	**Average Abs. Difference**	**Average Index**	**Average Abs. Difference**
GRPs	99.7	2.2	100	0
Reach (%)	86	10	89	6.7

See Appendix 3 for the full UAT

GRPs were preserved for 6 out the 8 campaigns. In schedules 6 and 7; there were radio stations missing from the fused data, so for the GRPs to match in these cases is unrealistic but the resulting GRP are different to the anticipated level, see Appendix.

Internet and Apps

There were no reach and frequency campaigns for Internet and Apps.

Outdoor

Outputs are designed to exactly match reach and frequency campaigns.

The Pointlogic solution will exactly match reach. Frequency of contacts and GRPs are corrected by using an index.