PRINT OPTIMIZATION: A PRACTICAL APPROACH

Shinichiro Okauchi, Dentsu Inc.

1. Backgrounds and Data Fusion

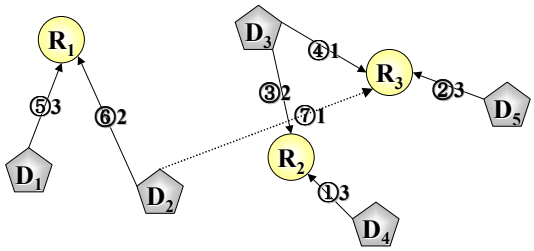
"Optimization of target reach according to media contact data" is one of the most fundamental subjects in the media research area for which we do not have enough measures yet.

Among various factors which make solutions difficult, it can be pointed out that media contact data itself is not sufficient – ideally speaking, single source data that has both large sample size and enough research items should be available, though it won't be achieved owing to economical or physical reason. These days while solutions using data fusion are proposed, the measures for the following issues are being improved.

- Ordinary communication target is different from the target audience which is set out in media planning – for example, in the case of the sleeping improving drug, those who are actually suffering from poor sleeping should be the original communication target. However, in utilizing media contact data it is expressed as one of demographic targets such as "male & female individuals from 35 to 49 years old". As a result, almost media plans deviates from the most optimum situation.
- Since media contact data which are used as currency by each medium are different one another, and thus planning based on overlapping contact to multiple media is not so easy.

As regards data fusion, since the proposal on applications to media planning of constrained statistical matching method adapting "Transportation problem" put forth by Soong and de Montigny (2001) was made, it has been discussed from the practical viewpoint actively in Japan. On the other hand, Okauchi and Mitsuda (2002) has developed a new method, distance advantage constrained statistical matching method. This enables to reduce calculation time for matching drastically at the same time maintaining advantages of the date rate on the donor. Also, it does not lower its quality of summation result.

Distance advantage constrained statistical matching method matches donor and receipient from the shortest distance combination, taking the sample weight into consideration. At the final phase of matching, though combinations of farthest apart are left, their influence on the summation result is negligible if a large-scale data fusion is tried out.



We produced and validated multimedia database by fusing ACR data for media contacts and consumption behavior in May 2001 and individual TV contact data (PM data) in May 2001.(***refer to Appendix***)

Distance advantage constrained statistical matching method completed matching in 60 seconds though the ideal constrained statistical matching utilizing "Transportation problem" needed about 40 hours (the personal computer used had "CPU AthlonMP1800 with 1 GB memory").

With this matching method which does not require much time, quick fused data analysis has become available by items which are utilized as Hook variables (demographic items and TV contact condition) in carrying out matching through extracting targets freely. Another benefit of this method lies in keeping data rate on the Donor side at the extracted target level.

Along with the trial of problem solution on this media contact data itself, as regards the method of optimal calculation (search), swift measures should be taken - even if media contact data are improved and if such data are not dealt with due care, the effect of its improvement will be less meaningful.

The need for most precision optimal logic will come out when the number of candidate vehicles are large, because verifications of all of the conceivable ad placement plans are not possible. For example, in case of conceiving once or less ad placement, combinations of about 270 magazines on ACR data in Japan can amount to $2^{270} = 1.9\times10^{81}$. In case of conceiving four times or

less multiple ad placements on the same vehicle, there are about $5^{270} = 5.3\times10^{188}$ the number of which is astronomically gigantic, and evaluation of reach or cost calculations for all the plans would be rather impractical with the current computer environment

$$2^{270} = 1.9\times10^{81} = 1900000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000$$

With data fusion, multimedia optimizer utilizing virtual single source data will be logically possible where larger scale optimum problem solutions will be necessary (on condition of feasible system environment and practical calculation time).

2. Utilization of Metaheuristics

As regards an optimal method based on media contact data, it is pointed out that conventional Hill Climbing ("HC") is apt to fall into locally optimal solution where improvement of HC or utilization of Metaheuristics is put forth. In order to deal with large scale single source data by data fusion effectively, its importance is getting heightened. As typical methods included in Metaheuristics, there are "Random Multi-start Local Search", "Genetic Algorithm (GA)", "Simulated Annealing", "Tabu Search and so on".

As regards the possibility of Metaheuristics (or the limitation of HC) on "target reach optimization based on media contact data", the following verifications with respect to small scale problems would give better understanding.

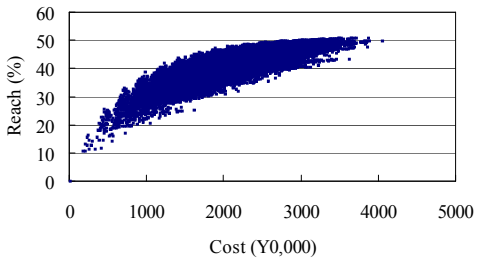
1) Conditions of Validation

- Media contact data used: ACR (Video Research Ltd.)
- Validation area: Tokyo (within a radius of 30km)
- Target: F1 (female individuals from 20 to 34 years old age)
- Ad placement budget: 20 Million Yen (About US$170,000)
- Candidate vehicle (magazines): 10 magazines of Josei Jishin, Josei Seven, an-an, non-no, More, With, ViVi, CanCam, JJ and Ray
- Each vehicle's upper limit ad placement: two times
- Evaluation index: "condition 1" = reach, "condition 2" = more than twice reach, "condition 3" = more than 3 times reach
- Optimization method: GA (100 individuals × 100 generations, mutation rate 2%, rank selection, 1 point crossing 50%) and HC
- Validation method: Absolute evaluation by all search

2) Validation Result

- Condition 1
  - Both GA and HC researched optimal solutions

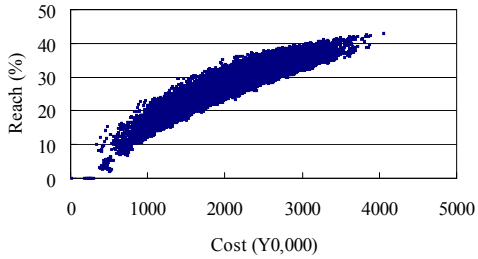
Item	Reach 1+ (%)	Cost (Y0,000)	Selected medium									
			Josei Jishin	Josei seven	an・an	non・no	MORE	With	ViVi	CanCam	JJ	Ray
Optimal solution	47.2	1875	1	1	1	1	1	1	1		1	
HC	47.2	1875	1	1	1	1	1	1	1		1	
GA	47.2	1875	1	1	1	1	1	1	1		1	



◆ Condition 2

- Only GA's optimal solution is searched
- 100 individuals ×100 generations cannot find optimal solutions (Average reach 2+ is 33.0%)
- 200 indivduals × 100 generations searched 100% optimal solutions

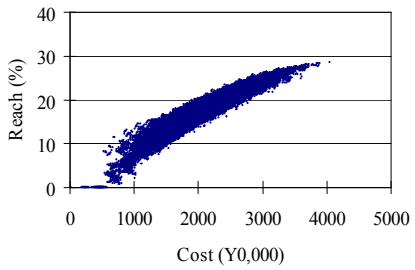
Item	Reach 2+ (%)	Cost (Y0,000)	Selected medium									
			Josei Jishin	Josei seven	an・an	non・no	MORE	With	ViVi	CanCam	JJ	Ray
Optimal solution	35.33	1950		2		2	2				2	
HC	28.6	1960		2			1		2	2	2	
GA	35.33	1950		2		2	2				2	



◆ Condition 3

- Only GA's optimal solutions are serached.
- With 100 individuals × 100 generations, optimal solutions are search at the rate of 12% probability (Average reach 3+ is 20.1%)
- GA1~3 show selected plans.

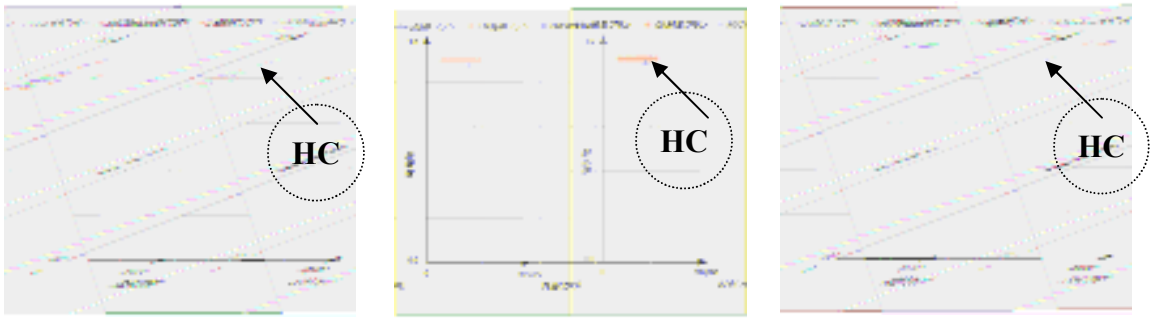
Item	Reach 3+ (%)	Cost (Y0,000)	Selected medium									
			Josei Jishin	Josei seven	an・an	non・no	MORE	With	ViVi	CanCam	JJ	Ray
Optimal solution	20.46	1980	2	2	1				1	1	2	
HC	18.8	1855		1			2	1	1	1	2	
GA_1	20.4	1985		1	2	1			1	2	2	
GA_2	20.46	1980	2	2	1				1	1	2	
GA_3	20.3	1930	1	2	2				1	2	1	



This validation indicates necessity of Metaheuristics for complex subjects.

When dealing with large scale multimedia contact data created by data fusion, appropriate methods or searching magnitudes differ in each case. This is why we have been seeking for a way to reduce calculation time and to stabilize the search result to make this system practical.

Part of planning systems within Dentsu adopt hybrid GA which takes HC results in initial plans. Through the combination with other approximate algorithms for better GA's performance, with relatively small scale search, good results (at least improved results compared to HC results) can be obtained.



("Darwin", newspaper area distribution optimizer: In accordance with searching, GA results are being improved by exceeding initial solutions with HC.)

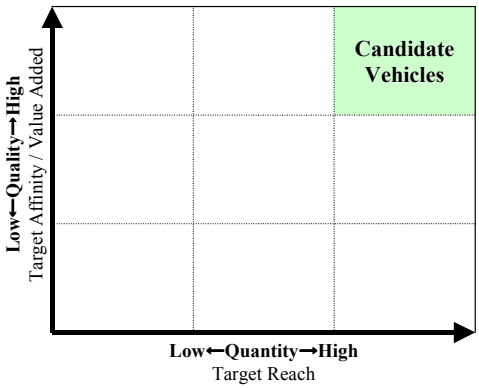
3. Utilization of All Search

Though All Search is quite a clear-cut optimal method, if the number of candidate vehicles which exceed certain amounts are set, its practice becomes impossible. On the other hand, in the case of the limited number of candidate vehicles, All Search will be possible within practical time. It can be utilized effectively especially in the case of a simple subject like "target reach optimization in one area for one medium".

In media planning for practical print media, all of the vehicles in media contact data are not necessarily suitable for candidate vehicles.

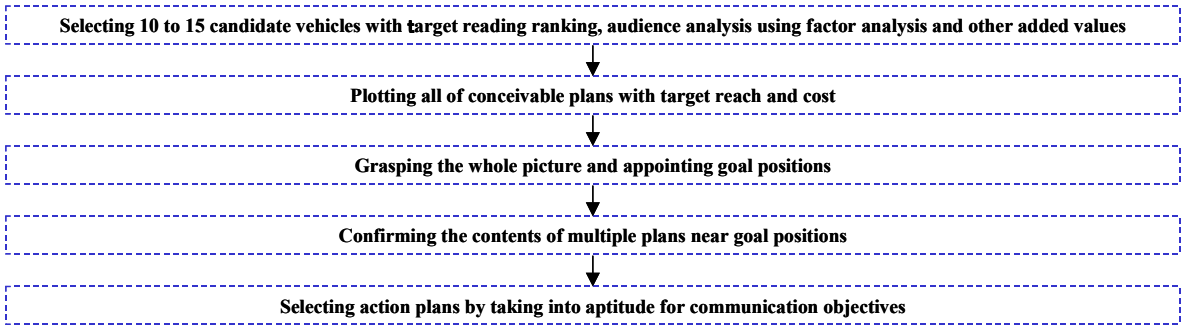
- Vehicles which are less compatible with communication subjects (products or messages) are not suitable as candidate vehicles.
- Vehicles which are in quite low target reach will be looked into separately from the framework of media plans
- Selection is performed through looking into not only direct effects on Audience by place advertisements but also promotion development momentum, influence on media companies or the government administration, etc (obviously there are selections made based on added values like these rather than direct effect on audience by ad).

Judging from the above, candidate vehicles are often narrowed down before "Target reach optimization". In these cases, we should select All Search, not HC or GA.



1) Optimization Flow by All Search

Proposed workflow here will be as follows.



If the size of All Search is around 500,000 plans, for upper limit ad placement 1, candidate vehicles should be narrowed down to less than 19, and for upper limit ad placement 2, candidate vehicles should be narrowed down to less than 12.

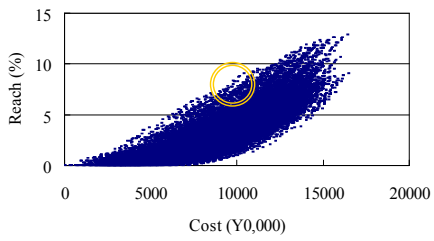
Number of upper limit ad placement	Number of candidate vehicle	Numer of candidate plan	Number of upper limit ad placement	Number of candidate vehicle	Numer of candidate plan
1	10	1,024	2	10	59,049
1	12	4,096	2	11	177,147
1	14	16,384	2	12	531,441
1	16	65,536	2	13	1,594,323
1	18	262,144	2	14	4,782,969
1	19	524,288	2	15	14,348,907

2) Examples of Optimization

First of all, for narrowed down 10 newspaper vehicles, plot on the graphs of all plans possible under the following conditions.

- Media contact data used: ACR (Video Research Ltd.)
- Validation Area: Tokyo (within a radius of 30km)
- Target: M1 (male individuals from 20 to 34 years old age)
- Ad placement budget: \100,000,000 (approx. US$830,000)
- Candidate vehicles (Newspapers): 10 papers such as Asahi, Mainichi, Yomiuri, Nikkei, Sankei, Nikkan sports, Sports nippon, Sankei sports, Sports houchi and Seikyo.
- Upper limit ad placement per vehicle: 2 times
- Evaluation index: more than 3 times reach

Item	Reach 3+ (%)	Cost (Y0,000)	Selected medium									
			Asahi	Yomiuri	Nikkei	Mainichi	Sankei	Nikkan sports	Sports houchi	Sports nippon	Sankei sports	Seikyo
Optimal solution 1	8.3	9727.5	1	1	2			2	2	2		
Optimal solution 2	8.3	9904.5	2		1		2	2	2	2	2	
Optimal solution 3	8.3	9772.5	2	1	2			2	1	2	1	



As the next step, multiple plans which are superior existing around optimal points with 8.3% of "R3+" in 100,000,000 yen ad placement budget will be picked up. For example, optimal solutions 1 to 3 in the chart are the plans which have almost even cost efficiency of evaluation indexes. At this point, "final practicing plans which include 'Sankei' twice which enables merchandise sampling through distributors will be selected", and this kind of judgment process will also enhance the better understanding of practice plans for advertisers.

4. Conclusion

There are many other subjects left for "Optimization of target reach according to media contact data".

As a measure solution for insufficient media contact data itself, data fusion is tried out. Its validation results show possibility of approach from the multimedia viewpoint, acquiring actual communication targets firmly.

As regards data fusion, its utilization environment is available as can be seen in making fused data easily which can retain donor side data collecting rate with regard to randomly extracted targets by using distance advantage constrained statistical matching etc., with lesser calculation load.

In dealing with such large scale media contact data, Metaheuristics should be indispensable, and thus its technical improvement will have to be promoted. For example, part of Dentsu's planning systems adopt hybrid GA containing HC results in the initial plan where with relatively small scale search, favorable results (at least it shows improved results compared to HC) are achieved.

As for All Search, though it is primitive, it will be the superlative optimization method if it is possible to utilize. When narrowing down of candidate vehicles in accordance with communication subjects is possible, with All Search clear solutions can be provided. In case of Metaheuristics, if search size is smaller, the probability of selecting favorable plans will be higher where narrowing down of candidate vehicles becomes important.

Appendix: New Matching Method and Its Performance
(Production and validation of multimedia database by fusing ACR data and individual TV contact data)

1) Conditions Of Validation

- Data used:
  - ACR data (May of 2001, May/7~, Kanto Area valid 2623 samples, Video Research Ltd.)
  - Individual TV contact data (PM data, May of 2001, May/7~Jun/13`Kanto area, valid 1270 samples, Video Research Ltd.)

  * Recipient by ACR data (having affluent amount of information), donor by PM data
- Daypart pattern to be used:

Daypart Pattern #1
83 Daypats/Station

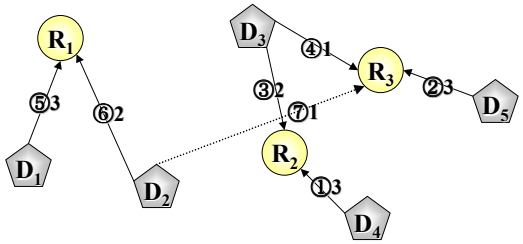
	Mon	Tue	Wed	Thu	Fri	Sat	Sun
5							
6	1	14	27	40	53	66	75
7	2	15	28	41	54		
8							
9	3	16	29	42	55	67	76
10							
11	4	17	30	43	56		
12	5	18	31	44	57	68	77
13							
14	6	19	32	45	58		
15						69	78
16							
17	7	20	33	46	59	70	79
18	8	21	34	47	60	71	80
19	9	22	35	48	61		
20							
21	10	23	36	49	62	72	81
22							
23	11	24	37	50	63		82
24							
25	12	25	38	51	64	73	
26							
27							
28	13	26	39	52	65	74	83

Daypart Pattern #4
26 Daypats/Station

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
5							
6			1			11	19
7			2				
8							
9						12	20
10							
11			3				
12			4			13	21
13							
14							
15							
16							
17			5			14	22
18			6			15	23
19							
20							
21			7			16	24
22							
23			8				25
24							
25			9			17	
26							
27							
28			10			18	26

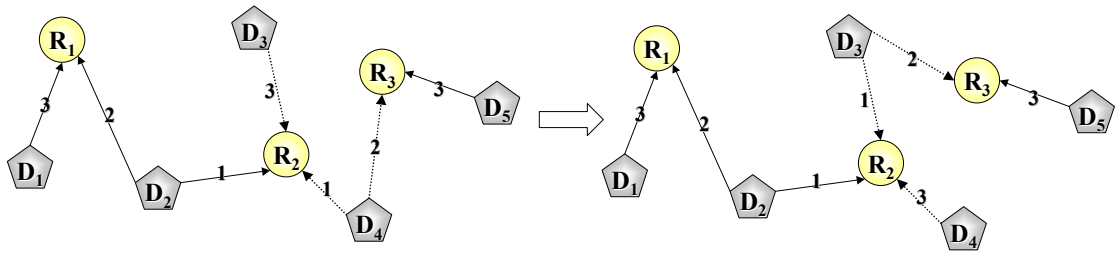
- Critical Hook Variable:
  - Genders
- Hook Variable:
  - Age (from 12 to 69 years old age)
  - Occupation
  - Married or not
  - Role at home
  - Structure of family
  - TV Viewing conditions (Daypart No.4)

  *Hook variable weight will be decided through optimization by simple ”Unconstrained statistical matching” and GA.
- Distance Calculation: Distance calculation of ACR and PM respondents will be carried out by ”Weighted nearest neighbor method” within the same gender group.
- Matching Method:
  - Distance advantage constrained statistical matching method matches Donor and Recipient from the shortest distance combination (from the largest sum of variable weight), taking the sample weight into consideration.



  - Existing Method : (”Constrained statistical matching adopting “Transportation problem”)

  *Making the initial solution with the ”North-west corner method”, and optimizing with ”Stepping stone method”
  *Take considerable amount of time in optimization



- Summation of Matching Results:
  - TV contact rate per 83 items of "Daypart No.1" are summed up per item as follows

    *Demographic attributions (gender, age, occupation, married or not, roles at home)

    *Consumption behavior (soft drinks, beers, detergents, shampoos, cosmetics, automobiles, PCs, wrist watches, cigarettes, snacks, mobile phones)

    *Other media contacts (newspapers, magazines)
- Remarks:
  - Based on random sampling, sample weight within "Recipients" (or "Donors") will be dealt evenly.

2-1) Hook Variable Setting For Distance Calculation

We set the point for each 6 variables to become in the scale of 0 to 1 according to how similar the recipient and donor are. For example, when they are identical, its value is 1. However, as regards TV contact conditions, owing to variances of absolute values between ACR data and PM data, each respondent is classified into no contact, low contact and high contact to set each point.

- Age ([W_1:1~10]×[0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1])
  - Relations on age variances and [0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1] are as follows.

Same	1	2	3	4	5	6	7	8	9	10 and greater
1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0

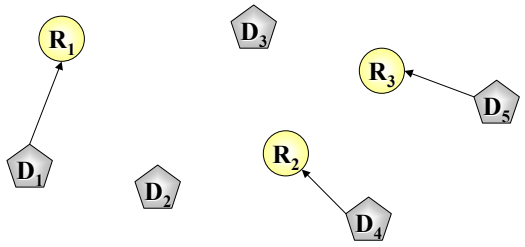
- Occupation ([W_2:1~10]×[0,1])
- Married or not, and roles at home ([$W_{3,4}$:1~10]×[0,1])
- Family structure (10 items except "others", each [$W_{5\text{-}1\sim10}$:1~10]×[0,0.1])
- TV Contact Conditions (156 items, each [W_6:1~10]×[0,0.5,1]×1/156)
  - Daypart (No.4)'s 26 items×6 channels
  - 156 items are dealt evenly.
  - 3 layers of [0 (viewing time: 0), L (viewing time: little), H (viewing time: a lot)] are classified based on viewing time per 26 items.
  - L and H divide [Total minus no contact] into two where it is an odd number, middle positioned sample is L
  - Relations between [0,L,H]and [0,0.5,1] are as per the following table.

H	0	0.5	1
L	0.5	1	0.5
0	1	0.5	0
	0	L	H

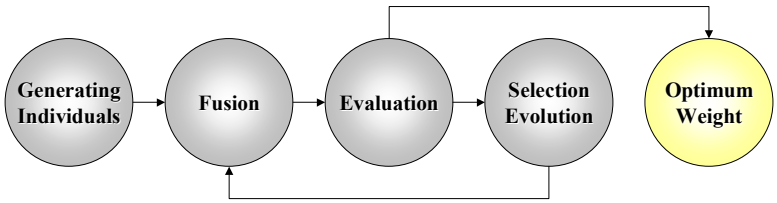
2-2) Optimization of Hook Variable Weight for Distance Calculation

Variable weight which sets the relative importance of 6 variables has been decided by GA's optimization. Practically, each variable weight was optimized at 11 levels ranging from 0 to 10 so that the square value deviances of TV contact rates of original data and fused data can become minimal.

- Matching method:
  - The most simple "Unconstrained statistical matching"



- Optimization Method:
  - Calculate optimal values of Weight (W_1~W_6) set by GA



- Evaluation Function:
  - Comparison between original data and fused data

*Variances of individual TV contact data
*Daypart (No.1)'s 93 items × 6 channels × 2 (male and female) × 6 (age, ten years old each)

3) Summary of Matching Results

Validation was made through comparison between values of fused data and values of original ACR data or PM data (Summing up TV contact rates divided into 498 Daypart; 83 weeks-Time Zones × 6 Channels). As summing up objects, each variable for demographic, consumption behavior and other media contacts was used. The following table shows correlation coefficient average and rate of correlation value of 0.7 and greater. In 90 items corresponding to 81% of the all 111 items, the strong correlation of over 0.7 is seen, and 0.4 and greater was confirmed for all the items. Even in comparison of existing methods, no inferior matching is obtained.

With existing methods, though matching with the current personal computers will take about 40 hours, the distance advantage constrained statistical matching method can complete matching in about 60 seconds (the personal computer used had "CPU AthlonMP1800 with 1 GB memory").

[Daypart (No.1)'s TV Contact Rate Comparison (total %)]

Summation variable	Comparison Data	Nunber of items	Proposed method (Distance Advantage)				Existing method			
			Correlation coefficient average		Rate of correlation 0.7 and greater		Correlation coefficient average		Rate of correlation 0.7 and greater	
			Correlation coefficient	Rank correlation	Correlation coefficient	Rank correlation	Correlation coefficient	Rank correlation	Correlation coefficient	Rank correlation
Demographics	PM	24	0.958	0.955	95.8%	95.8%	0.958	0.955	95.8%	100.0%
Demographics	ACR	24	0.848	0.835	91.7%	91.7%	0.848	0.835	91.7%	91.7%
Consumption behavior	ACR	48	0.753	0.740	68.8%	75.0%	0.756	0.737	75.0%	70.8%
Newspaper readership	ACR	5	0.843	0.855	100.0%	80.0%	0.819	0.853	80.0%	100.0%
Magazine readership	ACR	10	0.736	0.697	70.0%	70.0%	0.755	0.727	70.0%	60.0%
Total		111	0.820	0.808	81.1%	82.9%	0.822	0.810	82.9%	82.0%

[TV contact rate (498 Daypart) by Coca Cola takers with Proposed Method (distance advantage)]