Causal Analytics for Media Planning

Dick Dodson, Chief Research Officer, Telmar Group Inc.
and
Igor Mandel, Chief Statistician, Telmar Group Inc.

28st August 2015

Abstract

Audience Targeting is becoming increasingly important as the vast array of media options continues to grow and fragment. Traditional targeting methods no longer provide the optimal solution and over recent years ideas such as 'Involvement' and 'Engagement' have been introduced by our clients to try and improve targeting by adding the relationship between the vehicle and the user into the mix. While investigating how we might quantify such metrics we realised that this media requirement has similarities with the ongoing statistical pursuit of causal relationships within the study of human behaviour rather than correlational relationships which simply measure the degree that variables move in a predictable direction.

With most media targeting methodologies based on correlation and common movement of variables, Telmar undertook development of analytics that were closer to causality than just correlation. Telmar's first statistical development for Causality was to create an algorithm that removes random behaviour from correlation analytics. Once the random components were eliminated, algorithms for calculating Intrinsic Casual Probabilities were created. Both these developments have been realized and the desired algorithms tested. Throughout this paper the words, Causality, Intrinsic Causal Probabilities and Intrinsic Probabilities are used interchangeably.

Our paper describes the development work undertaken and shows some examples of the results achieved, based on GfK/MRI magazine readership data.

Key words: media targeting, causality, intrinsic probability

Introduction

Several developments in media and market targeting have been widely observed, but efforts to relate them to better understand their impact on media, audience, or target selection have been non-productive. These developments include:

1. The fragmentation of the media as the number of media alternatives expand with the growth of digital in all its forms, compounded by the ever-growing number of devices with which media can be consumed, while the number of potential customers to be reached has increased due to the social networks and the mass involvement of people with electronic world. This fragmentation leads to ever-increasing granularity in our ability to target audience, even down to the specific individual.

2. More data is available for targeting than ever before, including marketing and audience data. Concomitant with more data is the underlying belief that more data means better decisions.

3. C**ausality**, the statistical approach for the measurement of the extent that one variable causes a change in another variable, rather than simply being correlated with it, is a burgeoning field with the regular publication of monographs (see references in Section 1), articles, and a dedicated journal (Journal of Causal Inference). Taken at face value, causality approaches indicate a new tool for the evaluation of alternatives, including audience and marketing alternatives.

Taken together, one might think we are closer to better audience targeting, however the reality does not bear out the promise. More media means more alternatives to consider, while traditional evaluation tools have not evolved well to incorporate the new media landscape. More data often leads to less knowledge because the scale and speed of the data surpasses our ability to analyse it, and Causality has generated much debate about the underlying approaches used to produce a measurement.

1. Causal models in statistics – a summary

The principal approaches used in causality development are: Simultaneous Structural Equations originally conceived by S. Wright (Wright 1921; see references in Kline, 2010); Potential Outcomes (PO) proposed by J. Neyman (Neyman, 1923), and advanced by D. Rubin and others (Rubin, 2006), and the concept of Do-Operators and associated with them the Acyclic Graphs Structural Approach developed by J. Pearl and others since end of 1980s (Pearl, 2009). There are many other authors and proposals combining and modifying these ideas, although according to J. Pearl, almost all of these approaches incorporate the same things, using different terms and stressing different aspects of the problem. One common element in most of this work is the consideration of many interlinked variables and that the goal of the causal analysis is disentangling these influences and the subsequent evaluation of the individual impact of each cause on the effect. For instance, in the influential book Causality,

Models, Reasoning and Inference by J. Pearl (2009), all analyses are for many outcomes, but it is not clear what to do if there is only one outcome and many potential causes. This is the case in many applications.

In media planning often the issue is to find which vehicle will deliver a message with highest efficiency and there is usually just one measure of performance (*e.g.*, projected sales) and many inputs (vehicles). The previous exclusion of this case from causal analytics relegates us to non-causal solutions such as classical regression and other similar techniques which work with "simpler" models.

The Structural Graphs and Potential Outcome approaches are based on counterfactual variables. This approach assumes that the causal effect of the variable on the measure of performance could be derived from "interfering" – *i.e.* changing the value of the variable – and observing the change in the result, leaving all other variables intact.

The basic assumption of the counterfactual approach is that something will (or will not) happen, if something else didn't (or did) happen. A tightly defined goal combined with a large number of variables renders this approach ineffective. For example, one could ask if World War One would have happened if the Archduke Ferdinand had not been shot. This is an unanswerable question because of the many other variables at play and the specific, single, measurement goal. The solution proposed by the Counterfactual Approach is to forego prediction of the specific goal in favour of predicting the average goal across a population, however in audience targeting, we are dealing with a single specific goal. For more details about Causality approaches, see Lipovetsky and Mandel, 2015 (2).

The purpose of the causal analysis is to evaluate the intrinsic probabilities, or the parameters of the outcome Y=1 generated by different causes, including the random ones, with the observed data. Our proposed solution introduces the concept of probabilities to Casual analytics by allowing for a Causal variable to have a probability of having an effect as opposed to an on/off effect. We suggest that these probabilities are a way of quantifying the Involvement/Engagement values that have recently been promoted in the media planning.

2. Causality and Intrinsic Probability

2.1.Model description

(See Lipovetsky and Mandel 2015).

Consider a model of the direct impact of multiple causes onto the binary outcome Y with Y=1 and Y=0 meaning that the effect of interest has occurred or has not, respectively. A typical situation for media planning is that Y is the act of buying the product (or being aware of advertising, or having intention to buy, etc.), and influencing variables represent different media vehicles, or demographic groups. Consider a case of K attributes A1, A2,...,AK (where Ak=1 and Ak=0 denote the presence and the absence of a k-th attribute, with k=1,2,…,K). The attributes are represented by the variables which may be binary, ordinal, or nominal variables. A vector of the realized values of such attributes can be denoted as a=(a1,a2,..,aK), and this may represent levels of the same and/or different categorical variables, e.g., A1=1 means male, A2=1 means female, A3=1 means kids, A4=1 means teenagers, etc. – in other words, only values of the variable equal to 1 can produce the causal effect, while 0 – cannot (and data should be organised in such a way).

Let us assume that the attribute Ak creates the causal effect Y=1 with probability pk.

NB: The term "causal" here needs some comment. It means that one variable "creates" a certain probability effect – in other words that the combination of circumstances in this part of the world is such that event Y=1 appears with this probability when Ak appears. In this sense, the intrinsic probability to buy as a function of belonging to this subgroup is a substitute for the real causes which are never directly measurable. It would be better to use term "pseudo-causal," but for simplicity sake we leave "causal" keeping this substitutional nature in mind.

In the simplest case k=2, the probability that Y=1 would follow the rule of the union of the independent events:

$$S=p1+p2-p1.p2$$

In essence, it reflects that the coincidence of two causes does not produce anything more than one effect. Respectively, the probability of not having the causal effect would be presented as

$$1-S = (1-p1)(1-p2).$$

For any K, the causal effect of outcome Y=1 is defined as an intrinsic (latent) probability $S_{causal}(a)$ where (a) is a vector of the realized set of attributes, so that the probability of not-occurring of the event is:

$$1-S_{causal}(a)=\prod_{\{k:\,a_k=1\}}(1-p_k), \qquad (1)$$

where p1, p2,.., pK are parameters which represent the causal strength associated with the presence of each attribute Ak. Note that the absence of an attribute often implies the presence of the opposite attribute (e.g., the absence of the "male" attribute A1 contributes to the presence of the "female" attribute, A2). In other situations it could vary: for instance, a traffic accident may happen due to fog (A1), reckless driving (A2), ice conditions (A3), and other non-mutually exclusive causes. But what is important is each variable contains only one attribute (i.e. "Gender" would be presented as two variables with values Ak=1 in each).

There is also an irreducible latent probabilistic "random cause" that represents other factors that are not explicitly accounted for by the set of attributes. It is assumed that this random cause is: a) independent of other attributes; b) its effect (denoted as r for random) is constant across all configurations of attributes that may be present for a particular individual. These assumptions yield the expected probability at the population level as the union of the causal and random sources,

$$S(a) = S_{causal}(a) + S_{random} - S_{causal}(a) \cdot S_{random}$$

or in the explicit form:

$$S(a) = S_{causal}(a) + r - r \cdot S_{causal}(a) = 1 - (1-r)\left(1 - S_{causal}(a)\right) = 1 - (1-r) \prod_{\{k:\, a_k = 1\}} (1 - p_k) \tag{2}$$

The aim of the proposed model is the estimation of K+1 parameters, p1,..,pK, and r, on the basis of the sample of the realized outcomes Yi(a)={1,0} and the associated attribute vectors.

Concerning the motivation for the model, we can see the following arguments. Our setup acknowledges the asymmetric nature of causality, and the model (1)-(2) for intrinsic probability assumes that a single cause is sufficient for an event to happen, whereas for an event not to occur, all potential causes should be ineffectual. It contrasts with a common binary logistic regression, where all the attributes contribute additively to the probability of the event occurring or not occurring. Also, the model assumes that a random cause is irreducible and is present within the sample probabilities S(a). Finally, in the considered model, the main role is played by the presence of attributes (one value of the variable), rather than by the changing levels of the variables in classical models based on the concept of regression, Potential Outcomes, and other models.

Thus, we can say that each cause works as an independent entity and is associated not with the whole variable (like a binary "gender"), but with the separated levels (grades) of the variable (like two variables of "males" and "females"). It is different from the traditional statistical way of making models: one should look at these "grade-related yields" rather than at the coefficients of general association (*i.e.* regression), linking the whole "gender" to the outcome. Each level of the potentially causal variable produces an outcome with its own intrinsic probability. And if there are some causes, which cannot be associated with any measured variables but still produce the outcome, we relate them to the random cause. A typical example of such random causes is when customers buy a product regardless of advertising, or promotions (a sales "baseline" which is hard to estimate).

As an explicit example how formula (2) works, consider data with three variables (X), so in total there are eight cells of all combinations of their values, and in each cell we find the proportions Si of the outcome of variable S(a), so the proportion of Y=1 is in the base size of each cell. The cells and corresponding proportions Si are presented in Table 1. Of course, in a particular real data set, some cells could be empty. The variables in Table 1 are orthogonal, so they are statistically independent.

Table 1. Example of data set with three binary variables.

	x1	x2	x3	Si
	0	0	0	0.09141
	1	0	0	0.73409
	0	1	0	0.25630
	1	1	0	0.80300
	0	0	1	0.57608
	1	0	1	0.86570
	0	1	1	0.63409
	1	1	1	0.89563

Let us consider how to estimate parameters of the model (2) by data like those given it Table 1. Regrouping and taking the logarithm of the equation (3), and using notations

$$y_i = \ln(1 - S_i), \quad b_0 = \ln(1-r), \quad b_k = \ln(1 - p_k), \tag{3}$$

we represent (3) in the linearized form:

$$y_i = b_0 + b_1 x_{i1} + ... + b_K x_{iK} . \quad (4)$$

Thus, the problem of estimation of the parameters bk is reduced to the ordinary least squares (OLS) linear regression, with the known solution

$$b = (X'X)^{-1} X'y , \quad (5)$$

where y (4) is a vector of Nth order, X is the design matrix of xik values (completed by the additional column of all 1s, which corresponds to the intercept b0 in the model), b is the vector of all K+1 parameters in (5). If there are not enough observations, the matrix of the second moments X'X in (5) could be close to singular, and its inversion is impossible, or it yields too inflated coefficients. In such a case, we can use a regularization imposed onto the parameters which produces the ridge-regression and other techniques used to struggle with multicollinearity.

By the estimated coefficients b (4)-(5), each original parameter of probability can be obtained from the relations (3) by the transformation:

$$r = 1 - \exp(b_0), \quad p_k = 1 - \exp(b_k) . \quad (6)$$

The relations (6) show that the parameters b should be negative which can be achieved by their special parameterization if necessary (or just forced to be zero otherwise).

To illustrate this approach, let us return to Table 1, take yi=ln(1-Si) as the dependent variable (3), and construct the model (4). Its coefficients are presented in the first column of Table 2. These coefficients are transformed by (6) to the probabilities r of the random impact and pi of the causes, which are given in the second numerical column in Table 2. In the next column, Table 2 also presents the original values of cause probabilities used in this simulated data. Comparison of the estimated and the original values shows a good quality of the estimation with the relative errors of several percent or less shown in the last column of Table 2. The coefficient of multiple determination in this model (4) equals 0.998, and its value adjusted by degrees of freedom equals 0.995, so the quality of the model is very high.

Table 2. Regression coefficients and probability estimates.

Coefficients of regression		Estimates of causal probability		Original values used in simulation	Relative error, % to original values
0	-0.102722		0.09762	0.10	-2.38
1	-1.240261	1	0.71069	0.70	1.53
2	-0.224870	2	0.20138	0.20	0.69
3	-0.697456	3	0.50215	0.50	0.43

It is important to note that a design matrix like Table 1 is orthogonal, so the x-variables have zero correlations. In such a situation, coefficients of multiple linear regression equal the coefficients in the pair regression of y on each x separately, which makes calculations even simpler. If a cell of certain variables' combination is empty, the number of rows in the design table can be reduced. But even in such a case, it is possible to hold the whole design matrix substituting zero with a small proportion value, say, S=0.005.

In applications, researchers could often be interested in estimating an additive share of influence of each cause in the effect, or *decompose the effect* into mutually excluded components. In order to achieve this, we can use the following formula (see derivation in (Lipovetsky and Mandel, 2015)):

$$S_{ik} = S_i \frac{\ln(1 - p_k)^{x_{ik}}}{\ln(1 - S_i)} \quad (7)$$

where the total of the causally induced effects (including the random one corresponding to the index k=0) in each cell equals the predicted proportion (these Si divided by total observed effect could be calculated also as a percent):

$$S_i = \sum_{k=0}^{K} S_{ik} \quad . \tag{8}$$

The decomposition is a very convenient feature of the model – it answers a different question about the nature of the calculations. If intrinsic probability shows how intensively each variable generates the outcome, decomposition of the result (7)-(8) shows, what is the final effect of this variable in combination with all others. The conclusions from these two statistics are also very different. High p(k) may happen, for example, in a small magazine (where frequency Ak=1 is very low). This magazine may be good for advertising, but its share in a total market S_{ik} will be low exactly because of its small size. Or, it may happen that pk is still high, but this magazine had such a wide duplication with other magazines, that just a small fraction of cases Y=1 generated by this magazine contributed (they moved in concert with other variables). In another environment its share could be much higher. Or, pk could be high, but the random component could also be extremely high and the random component masks the effect of this variable.

In short, if the intrinsic probability is a pure effect of the variable in isolation from randomness and from influence of other variables, the decomposed share of the variable in total effect is the complex product of all these factors in combination.

2.2.Numerical Simulations and algorithm modifications

To test the validity of the proposed procedure, a series of experiments on generated data were performed. Generation of Y=1 as a function of the different X was done by applying uniform distribution with a given frequency. If, say, P1=0.1 and frequency of X1=0.2, in a dataset, 20% of all observations have the value X1=1. Out of these 10% of randomly selected observations would have C(1)=1, where C(1) is unobserved effect generated by X1 variable. Then for this observation if the sum of C(1)+C(2)+…C(r) >0, Y=1, otherwise Y=0. Having just observed Y and X, one should find all Pk.

We used two methods of ***data generation***: (1) ***Random***, when on each step with a given set of parameters everything was updated and (2) ***Random with pasting,*** when the data set was generated randomly just once and then copied and pasted. Respectively, two methods of running the ***simulation*** were applied: (1) ***random*** (when for each set generated in random mode all calculations were performed) and (2) ***bootstrap***, when for each dataset generated in random and pasted mode, we applied the bootstrap procedure with a given number of runs (in which the same number of observations was randomly selected each time from the data). In both scenarios the average across all runs was calculated together with first and last quintiles.

The following parameters were alternated in the experimental design:

- ***Number of cases n*** in a data set (100 - 1,000,000);
- ***Number of causal variables*** (from 1 to 100);
- ***Correlations between X variables*** (-0.9 - +0.9);
- ***Level of randomness r*** (0.1 - 0.8);
- ***Intrinsic probabilities for X variables*** (Equal or Different);
- ***Fraction of X=1 in input variables*** (5%-90%);
- ***Number of model's runs for estimation*** (1-100).

After the modelling, the estimated probabilities in (4)-(6) (after averaging among the runs) were compared with the values used in data generation using primarily two types of statistics: ***correlations*** between values used in generation and estimated values; and ***relative error***, calculated as absolute difference between estimated and generated values, divided by original value used in generation (if, for example, if Pk=0.1 was used in generation and the average estimation was 0.12, the relative error is abs(0.12-0.1)/0.1=20% (see also table 2).

Out of many settings let's show just some to illustrate the level of quality of the estimation. In one experiment with eight variables, the original coefficients could take any random values (they are not controlled, as in a more typical setting), i.e. in each dataset everything was absolutely random and for each set just one estimate was made. The results of 40 simulations are shown in Table 3, where the average correlation of original Y with Xs is 0.05, and the maximum correlation equals 0.15.

Table 3. Quality of the parameters estimation for all Pk random, 40 runs (n=10,000)

	p1	p2	p3	p4	p5	p6	p7	p8	r
Average correlations between original and estimated values among 40 runs	0.69	0.87	0.78	0.81	0.83	0.86	0.74	0.80	0.64
Median error, % to original value	35	20	23	32	27	33	21	21	40

The first row in Table 3 shows that correlations between the original and estimated values are quite large, so the procedure definitely captures the main features of the data. This is especially important because the original datasets have practically no correlations among Y and X variables, so in this situation, the traditional statistical methods completely fail. The second row in Table 3 shows that median error is about 20-30% of the original values. Of course, it is not an ideal but a good enough result in a situation where original data are uncorrelated and estimates each time are made just once. Other experiments showed that the estimations only slightly depend on the level of the mutual correlations between X variables, so the problem of multicollinearity is much less troubling in this approach as in common regression modelling.

Many experiments showed that direct estimation of the parameters based on (4) – (6) does not always produce the best values possible, especially when the number of variables increases. The reason for this is the duplication of many variables and the random component attribute, which leads to the random coefficient becoming inflated (while the coefficients of the variables are often estimated pretty well). To overcome this problem, the following idea has been implemented: let's find some subset of the data, where just part of the variables may have the causal influence on the result, while others do not, and estimate the values of the coefficients only on these subsets. For example – if one finds a subset where all X=0, the all Y=1 the effect should logically be attributed only to the random component R; or on another subset X1=1, but all other X = 0 – all cases with Y=1 are to be explained only by random and X1 values and so on.

This ***Subsets algorithm*** turned out to be very effective. Coefficients for each variable could be estimated, in principle, several times: X1 paired only with R, X1 paired only with R and X2 and so on. R, in turn, could be estimated as many times as any other estimations take place and then averaged. Experiments showed though that the best way to make subset estimation is to use single isolated X variables and isolated R (in which case we do not need regression, just calculate a fraction of Y=1 cases for subset when all X=0).

One set of results supporting this statement is provided in table 4. Each column represents the specific method used to estimate the same values of Pk, shown under Variable title. R took two levels – 0.1 and 0.7 (10% and 70%). Each value in a table is an average among 30 bootstrap runs (for example, 0.21 in the upper left corner means, that after 30 runs the average estimate of the coefficient was 0.21). Two statistics of quality – correlation between generated and average estimated values and median error for all IP's are at the bottom of the table. Single subset means that just the particular variable was isolated. Pair subset means that all pairs of this variable with all others were isolated. The values of the coefficient for this variable were estimated and then averaged to be shown in the table. The table leads to some interesting conclusions.

Table 4. Different models applied to different data settings (bootstrap, 30 runs, n=10,000)

		Types of models used for estimation						
		Direct IP model	Direct IP model	Regression	Singles subsets	Singles subsets	Pairs subsets	Pairs subsets
	Random R generated, %	10%	70%	10%	10%	70%	10%	70%
Variables	IP generated	1	2	3	4	5	6	7
x1	0.1	0.21	0.04	0.03	0.12	0.29	0.11	0.52
x2	0.2	0.43	0.32	0.04	0.20	0.57	0.15	0.35
x3	0.3	0.69	0.18	0.08	0.38	0.28	0.33	0.44
x4	0.4	0.51	0.22	0.04	0.35	0.57	0.71	0.49
x5	0.5	0.68	0.33	0.07	0.54	0.76	0.52	0.80
x6	0.6	0.71	0.40	0.09	0.70	0.49	0.53	0.55
x7	0.7	0.75	0.33	0.10	0.78	0.61	0.74	0.73
x8	0.1	0.08	0.05	0.02	0.12	0.05	0.08	0.07
x9	0.2	0.38	0.26	0.04	0.25	0.16	0.32	0.41
x10	0.3	0.71	0.38	0.08	0.31	0.33	0.23	0.53
x11	0.4	0.38	0.22	0.04	0.37	0.33	0.24	0.18
x12	0.5	0.49	0.27	0.05	0.50	0.53	0.43	0.32
R		1.00	1.00	0.58	0.10	0.70	0.16	0.60
Correlation		0.39	0.68	(0.24)	0.98	0.75	0.85	0.57
Median error		31%	45%	85%	12%	19%	15%	52%

1. Growth of **random causes** for direct models (columns 1 and 2) from 10 to 70% yielded significant growth in error, from 31 to 45%, yet the correlations increased, which is somewhat inconclusive and counterintuitive. But the same shift for single and pairs subsets (models 4, 5 and 6, 7) show consistent detriment of the quality: the correlations declined (0.98 to 0.75 and from 0.85 to 0.57, respectively), and the errors increased (12% – 19%; 15% - 52%). Amazingly though, for the single subset model a huge increase in randomness still produces very acceptable results, for example, a 19% of error is much better than 45% for the direct model.

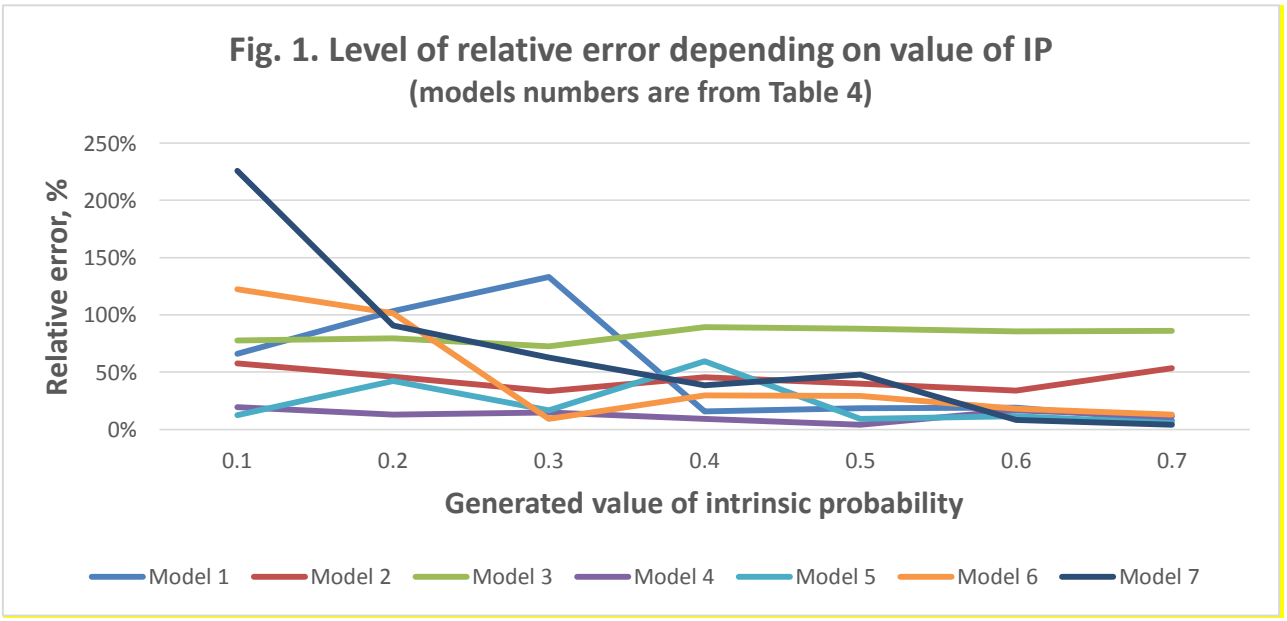
2. Direct models work much worse than **single subset** models and, to a lesser extent, the pairs subset model. The Single subset model gives in fact a near perfect result for R=0.1: correlation 0.98, error 12%.

3. Using **pairs** in the subset algorithm does not help. The results are significantly worse than for singles. However, this option cannot be discarded from consideration because in practice single subsets may not exist (or are very small in size), while pairs provide yet better estimates than the whole data set.

4. A serious problem with direct estimates is the inflation of **R**, regardless of the value used in generation (estimates of R are 1 both for 0.1 and 0.7 generated values), while individual coefficients are still estimated well. This is in sharp contrast with subsets, which estimates R perfectly with singles (0.1 and 0.7).

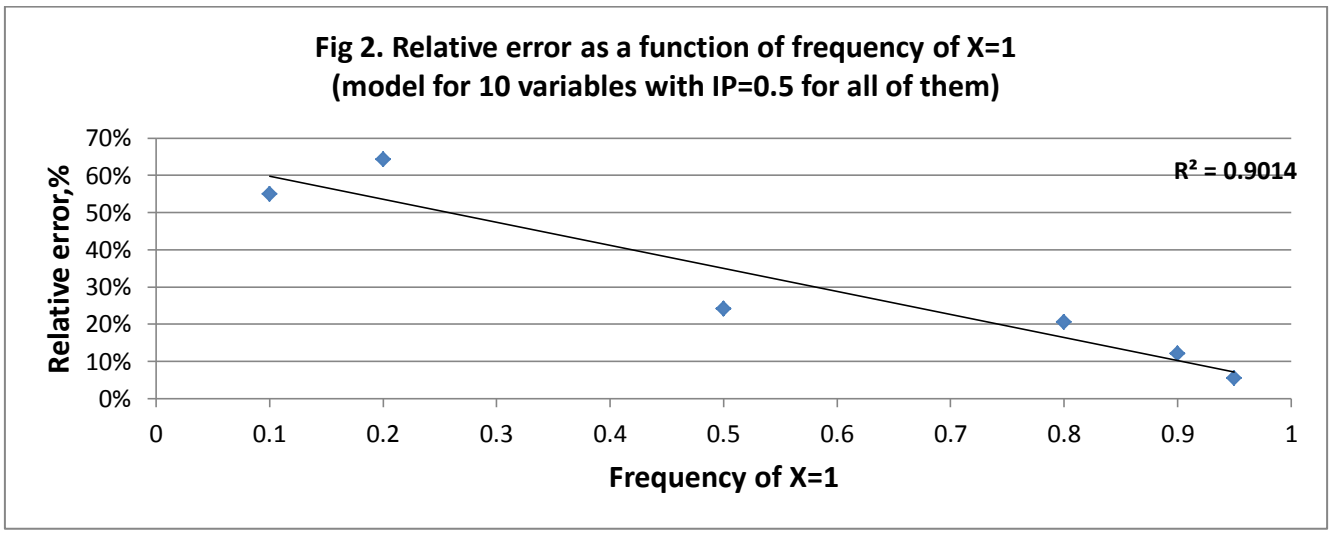
5. The **regression** results are the worst of all, with negative correlation between the real values and their estimates, and an error of 85% (although the random component was just 10%). This is remarkable because it demonstrates that traditional methods trying to capture the relationship where two effects --randomness and individual causal outcomes--from each variable are intermixed, fail. The IP estimates go deeper into the data – even though the correlation between X and Y variables are zero.

6. It is important to understand how the quality of estimation depends on the **level of intrinsic probability**. Fig. 1 (based on table 4, where values of coefficients with identical IPs, like x1 and x8, x2 and x9 were averaged) gives the answer. One can see, generally, a decline of the errors when IP increases. This is explained but just in part, by the design because the IP value is in the denominator of the error. Some values are still going up with a large IP. It is, generally, logical as it is easier to capture the "most intensive signal" than a weak one.



Of special interest is the extent that the quality of the estimates depends on the frequency of the input variables. As a general tendency, the higher the frequency of X=1, the better estimate, as shown on Fig.2 after applying the single subset procedure to the data with the same IP (0.5) for all ten variables, but this regularity was not firmly confirmed for many experiments and needs additional investigation.

NB: Often variables with low frequencies are also estimated very well.



All these and many other experiments show that the IP methodology can estimate the hidden drivers of the observed effect and reveals structural relations that would be invisible otherwise.

3. Modelling of the efficiency of print media

To demonstrate the application of this methodology to a practical media planning problem, we made many models with real (survey) data.

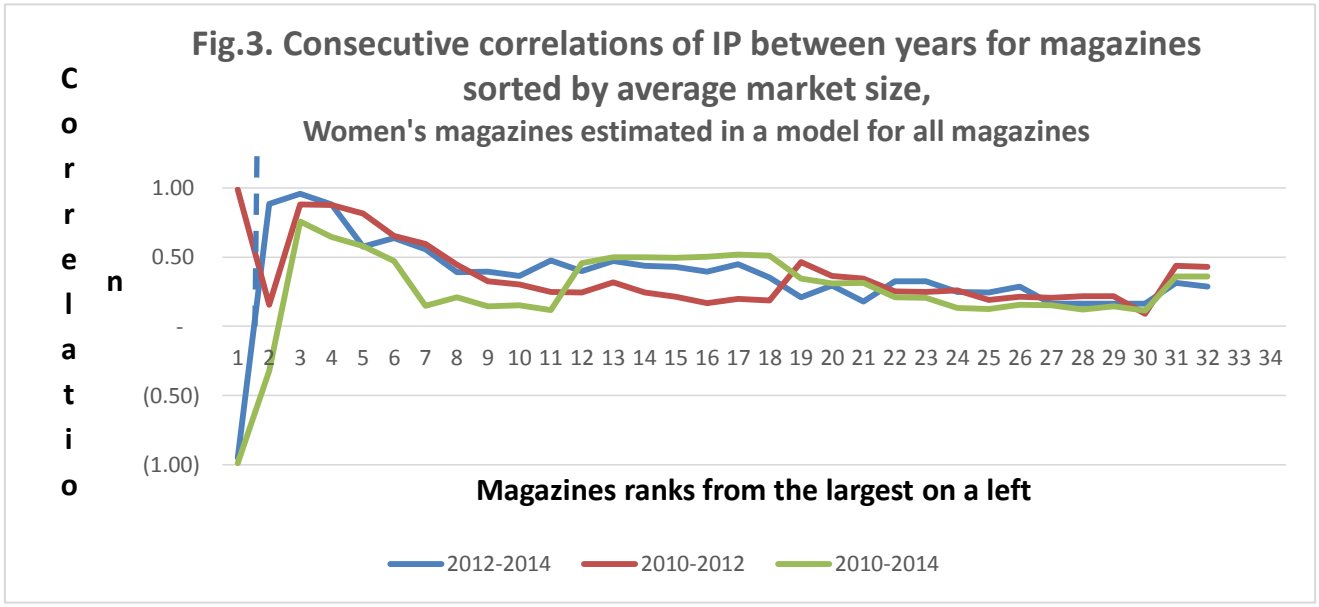
The purpose was to understand how the causality model would affect the selection of the best media vehicles. The target brand was **Woman's Big Ticket Apparel**; the target audience was **Women 18+**; the pre-selected media included 34 woman's magazines and 39 men's magazines which are also read by a relatively small number of women. We used the MRI 2010, 2012, and 2014 Doublebase® studies. See Appendix 1 for details about the magazines. Details of the analysis are:

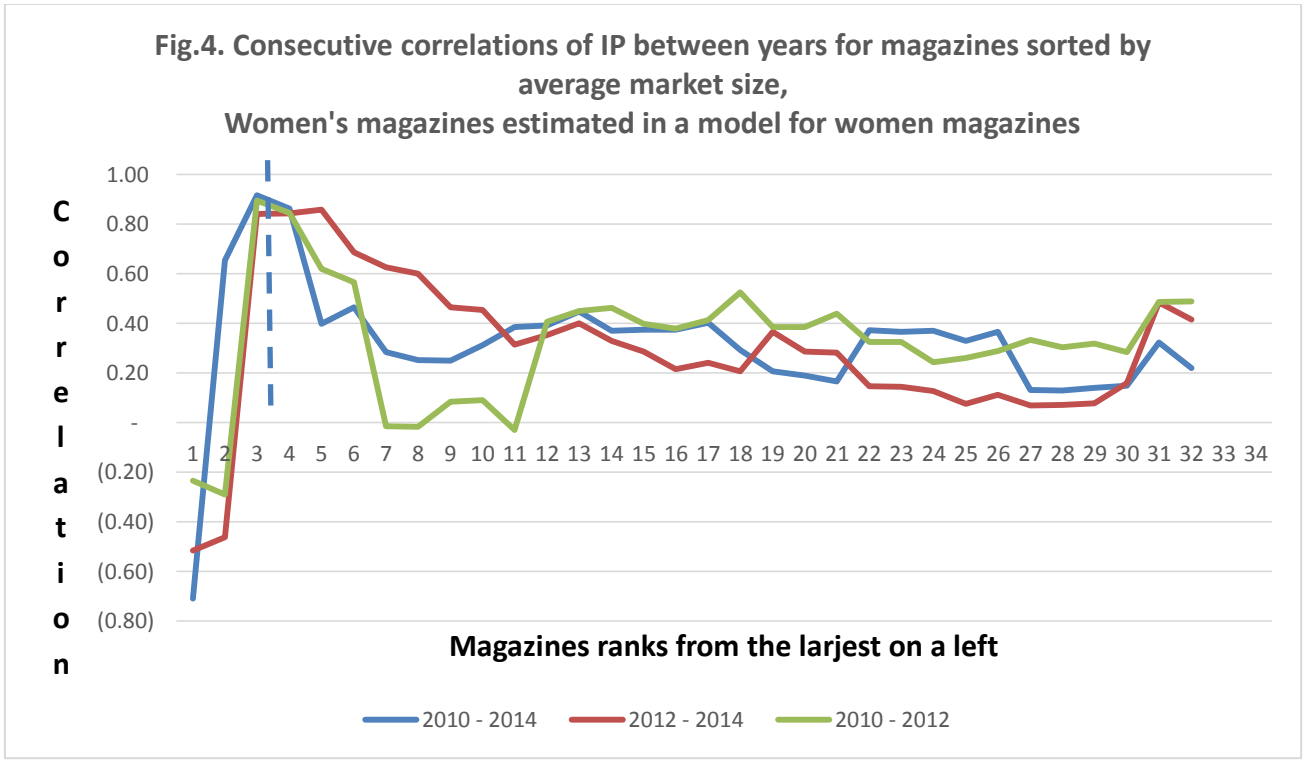
1. Models were built for each database for the following combinations of the parameters:
   a) Women – All magazines
   b) Women – Women's Magazines
   c) Women – Men's Magazines
   d) Men – Women's Magazines

We will focus on models a) and b) and touch on the models c) and d) just in passing. All models were made using the single subset method, as it seems most reliable (see 2.2).

2. We looked at the stability of the IP values over time by correlating the values for each year. This produced a consistent result in part due to economic fluctuations between 2010 and 2014 and the impact of these fluctuations on high-end apparel purchases. We applied the following approach:

We expected (in general this was confirmed by data) that any tendencies are more pronounced on large magazines We calculated the average market size for three years for each magazine and sorted them in descending order. Then moving correlations were calculated between the first and consecutive values of IP for these magazines - i.e., the correlation between 2010 and 2012 years for first three (the largest) magazines in a list (the starting value), then for the first four and so on till the end. Two charts of these correlations for models types a) and b) are shown in Fig.3 and Fig. 4.





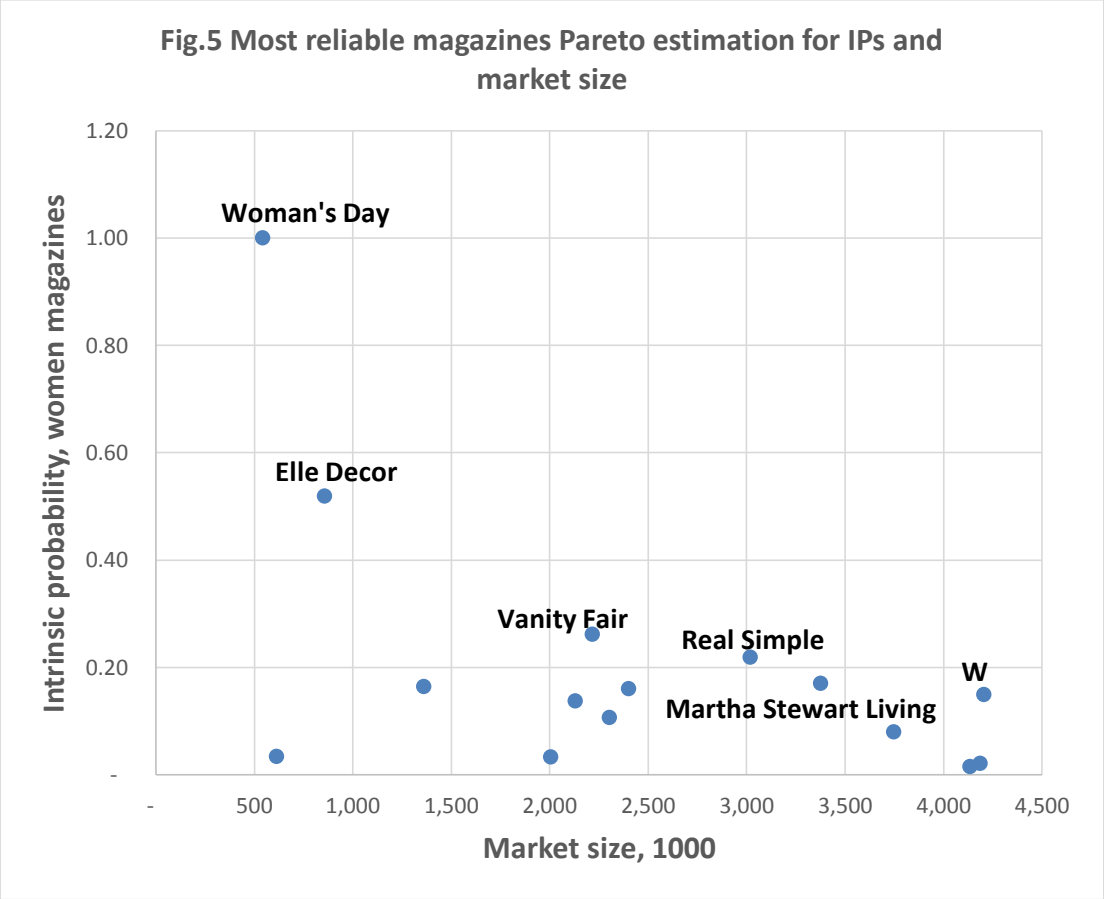
In both cases there is a subgroup of large magazines, where IPs are very highly correlated (determined by a broken vertical line), and these groups for the two estimates are identical in that they include the following five magazines: *Better Homes & Gardens*, *Good Housekeeping*, *Woman's World*, *Family Circle*, and *Cosmopolitan*. The correlation between years for these magazines is larger than 0.85, i.e. IPs are very consistent. It should not be considered as a very definitive indicator of consistency, for example, one can add the next magazine, *Southern Living*, which insignificantly reduces the level of correlation. If one is interested in finding a small group of magazines which demonstrate very similar performance any time, these six would be good candidates. It's interesting to note that among women's magazines estimated by the woman's model (Fig.4) two years – 2010 and 2012 – demonstrate extremely high correlations for any subsets of magazines. But usually people are interested in the most recent data, so our next analysis will be done based on data of 2014.

3. Summary information about all the magazines for 2014 is provided in Appendix 1. The columns 6 and 7 contain two sets of IP – estimated for all magazines (73) and estimated for women's magazines (34) only. They are correlated on a level 0f 0.97 (without the extremely high values in Woman's day – 0.92), which is very high, what definitely gives a good indication of the directionality of the estimates. . However, even greater confidence in the results can be had if the IPs are estimated on the same absolute level by two different models. We selected the magazines, which have difference between IPs in two models no more than 20% (to average estimate of the two) and put information about them into table 5. It turned out that 15 out of 34 (44%) have very similar values of the IP, what adds a lot to the confidence level. In the first row are correlation coefficients between IP derived from women's magazines and other indicators. The correlation with a traditional Index is 0.73, *i.e.* about 54% of the information about these IP could be derived from the Index. The relationship between indices and IP is non-linear, but, generally, positive. Yet IP provides additional information, which the index cannot do (see example below on Fig.6). With this information, one can make a meaningful selection of the magazines, based on business priorities.

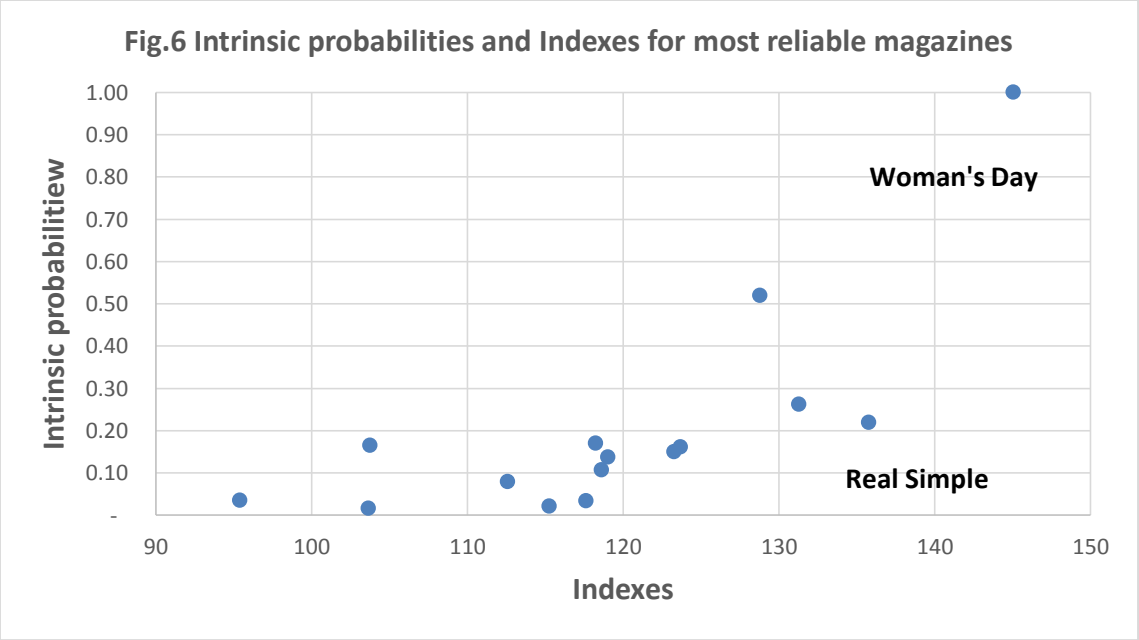
Table 5. Most reliable magazines with stable IP in two classes of models

	Magazines	Audience (000s)	Market size (000s)	Apparel consumption, % to Audience	Index	IP from all magazines model	IP from women magazines model
	Correlation of IP for women magazines	(0.58)	(0.55)	0.73	0.73	1.00	1.00
1	W	10,603	4,207	0.40	123	0.17	0.15
2	Glamour	11,290	4,188	0.37	115	0.02	0.02
3	Southern Living	12,401	4,136	0.33	104	0.02	0.02
4	O The Oprah Magazine	10,347	3,749	0.36	113	0.07	0.08
5	Martha Stewart Living	8,878	3,379	0.38	118	0.17	0.17
6	Real Simple	6,914	3,021	0.44	136	0.22	0.22
7	Allure	6,039	2,404	0.40	124	0.15	0.16
8	Ebony	6,037	2,304	0.38	119	0.11	0.11
9	Vanity Fair	5,251	2,219	0.42	131	0.29	0.26
10	Essence	5,567	2,133	0.38	119	0.12	0.14
11	House Beautiful	5,301	2,007	0.38	118	0.03	0.03
12	First For Women	4,081	1,363	0.33	104	0.19	0.16
13	Elle Decor	2,073	859	0.41	129	0.50	0.52
14	Latina	2,002	615	0.31	95	0.03	0.03
15	Woman's Day	1,169	546	0.47	145	1.00	1.00

4. In particular, one can place the values of IP and market size into one chart (Fig. 5) and select the best magazine, having both the high IP and a large market penetration. The six magazines (which have names on the chart) belong to the Pareto first optimum layer, which means, that they cannot be overcome by both criteria by any other magazine. After considering the first layer, one can eliminate it and detect the second layer of the best candidates and so on. There are other ways to evaluate this information. Each of the obtained causal coefficients (intrinsic probabilities) should be interpreted as an internal feature of the media vehicle, which gives it additional value *vs* typical random, which is equal to the "baseline" level and therefore can be utilized.



5. To compare the results to a traditional Index, we put the same magazines on another scatterplot (Fig.6). As may be seen, the composition of the Pareto layers changed (which is not surprising, because Indexes and IPs are positively, not negatively correlated). Just two magazines from the fig. 5 remain in a first layer.



6. It interesting to look at the behaviour of women in men's magazines. The volume of the market for woman's big-ticket apparel in these is of course much smaller than in women's magazines, but there is a strong presence of women regardless. A summary data for the men's magazine are presented in Table 6.

Table 6. Men's magazines in relation with women apparel

	Average Index for women in men's magazines	Coefficient of variation for index	Magazines with index less than 100, %	Correlations between IP for men and women in men's magazines	Magazines with Men's IP higher than Women's IP,%	Average index for women in women's magazines
	1	2	3	4	5	6
2010	61	0.35	97%	0.62	15%	108
2012	69	0.38	82%	(0.04)	18%	110
2014	104	0.25	31%	0.62	8%	116

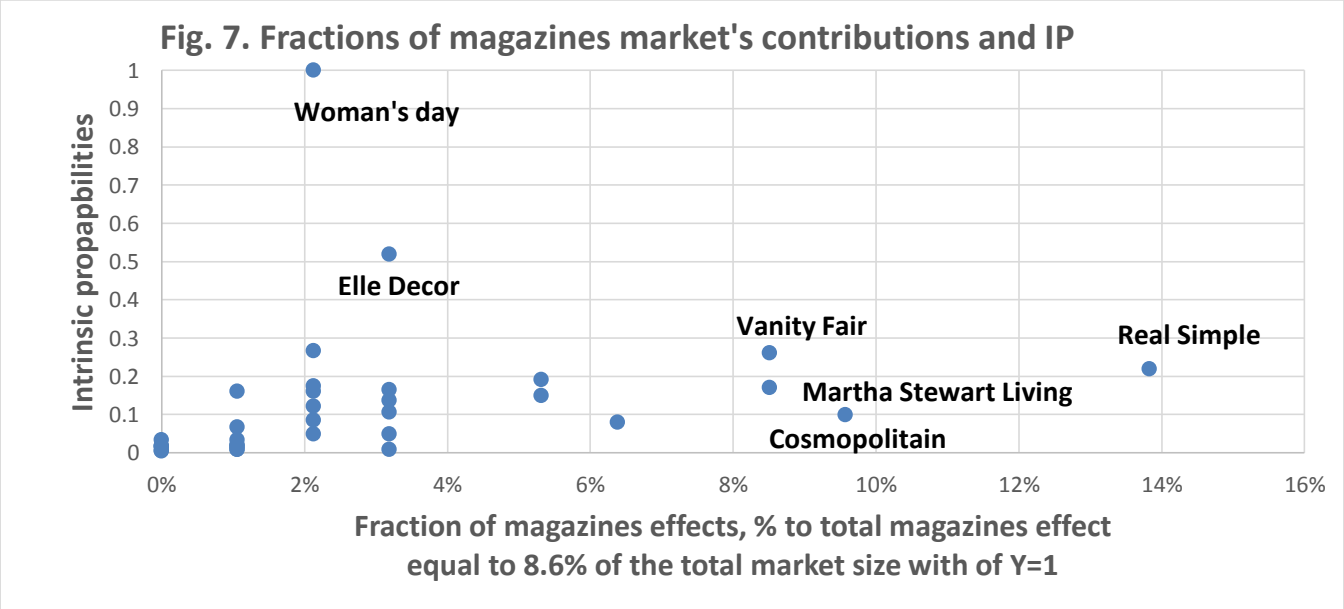
One can see that a change in the market occurred over time. The indexes grow for these 5 years and even exceeded the average of 100 in 2014 (column 1), in quite a consistent fashion (comparatively small variation, column 2). The number of attractive magazines with index higher than 100 dramatically changed in 2014 (column 3), from 3 to 69%; some of the magazines became quite effective (like *Popular Science* with index 136; *Guns & Ammo*, 127; *Motorcyclist*, 126, etc.).

NB: Remember the target is high-end women's apparel, including a leather jacket. Many motorcyclists wear leather jackets

For comparison, column 6 contains average indexes in women's magazines. As expected, they are much higher, but the gap is narrowing.

Correlations between IP for men and for women in men's magazines (column 4), shows a dramatic change in 2012 and then forms a "new normal". The level of correlation 0,62 shows that similar sets of the magazines are attractive to both genders, but this question begs more investigation. What is surprising is on first glance that women's efficiency in men's magazines (for women apparel) is much higher than men's, and it is growing, but not as fast as indexes (column 5).

7. Decomposition of the magazines effects, performed by formula (7) provides different information, than intrinsic probabilities themselves. For comparison, let's look at these two values for Women in Women's magazines as shown on fig. 7 for 2014 data (see the values of contributions in Appendix)



First, the share of the baseline (or "random", typical buying) is overwhelming – it is 91.4% of all purchases and not show here. It is, generally, expected – advertising seldom plays the key role in purchasing of the established good (and in our case it is not even advertising, but just reading). For that reasons fractions of effects were recalculated, taking the remaining part (belonging to specifics of the magazines readership, 8.6%, as equal to 100%) – and these values are presented on a chart.

Second, majority of magazines having comparatively big contributions (like *Real Simple, Marta Stewart Living*, etc.) have been already included in Pareto optimal sets (see fig. 5) . The only significant contributor is *Cosmopolitan* (but it was not preliminary

selected in table 5 – so, did not appear earlier). It means, that decisions are, in general, to be consistent to each other, yet contributions reflect, of course, different aspect of reality, as explained in 2.2.

Third, the correlation between these two metrics is positive (0.22), but not that high (after exclusion of extreme cases like Women's day it grows to 0.44 or higher).The IP shows the internal intensity, while contribution – the particular play of the given media in its combination with all others. The higher contribution – the better it worked, but in given combination. If one changes the environment – the results could be different.

In general, the conclusions one would draw from a traditional analyses using indices and market penetration support the more extended conclusions drawn by the application of our causality and decomposition approach.

Conclusion

We have considered for the first time the implementation of a new approach to causal modelling based on the direct accounting of the internal relationship between the causal impacts and the outcome effect in media planning. The proposed model is a significant departure from the regular regression, or statistical learning models in general, as well as from the traditional models of causal analysis. In the suggested model, each causal variable effects the outcome individually, not cumulatively with others, which contrasts with the traditional statistics where the outcome accumulates the combined effect of all the variables of influence, and adding variables improves the goodness of fit. Also, unlike in the traditional methods, the random cause is not considered as something to be "minimized", but rather as a reflection of all causes which were not captured by the introduced variables. This approach to the analysis and estimation of causal relations demonstrates several important features:

a) It offers a way to estimate the causal relationships, when many possible causes generate one effect – a situation very typical for media planning applications;
b) It allows one to estimate the intensity of the causal relationships in the data, even if there is no correlation between Y and X variables, or when causal variables are highly correlated among themselves, or when coefficients of variables are equal to each other, or when the random component in the data is very high. All these features make it very different from the traditional statistical and causal approaches (primarily based still on regression).
c) It works with just frequency tables (providing they exist for all or many combinations of the predictors) so there is no need for the original observational data sets on the level of respondents.
d) Parameter estimation is technically very simple.
e) It works with data of high dimensionality, since the orthogonal design matrix allows us to reduce the estimation to paired regressions or data allow for finding a special subsets with beneficial features.

We have demonstrated how this approach could be used for media planning, where the pure effect of media, free of randomness, should be more useful than the traditional approaches which are inherently distorted by random noise. In particular we would contend that the decomposed % of IP as described above gives a method of quantifying the 'engagement' or 'involvement' of a media audience with the product. The table below shows 5 women's magazine which have varying readership levels (5m – 14m) but all have similar Index levels for product purchase (119 – 131). However the decomposed IP %s range from 2% to 10%. We would contend that this reflects the varying 'engagement' or 'involvement'.

	A new Measure of Engagement?				
	Vanity Fair	Cosmopolitan	W	Self	Elle
Readership	5,251	14,426	10,603	5,014	5,176
Index	131	119	123	124	124
IP	0.26	0.10	0.15	0.19	0.17
Decomposed %	**9%**	**10%**	**5%**	**5%**	**2%**

Acknowledgements

We thank V. Jain and V. Grover for the development of software for this approach.

GfK/MRI for the data used in the magazine examples.

References

Kline R.B. Principles and Practice of Structural Equation Modelling, Guilford Press, 2010.

Mandel I., Fusion and causal analysis in big marketing data sets. *Proceedings of JSM 2013, ASA, 2013.*

Neyman J., On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9, Translated from 1923 Polish original article by D. M. Dabrowska and T. P. Speed in: Statistical Science, 1990, 5 (4): 465–472.

Pearl J., Causality: Models, Reasoning, and Inference. *Cambridge University Press, 2009.*

Rubin D., Matched Sampling for Causal Effects. *Cambridge University Press, 2006.*

Wright S., Correlation and causation, *J. of Agricultural Research, 1921, 20, 557–585.*

Lipovetsky S. and Mandel I., Modelling Probability of Causal and Random Impacts, *Journal of Modern Applied Statistical Methods, 2015, forthcoming*

S. Lipovetsky and I. Mandel. Book Review: Handbook of Causal Analysis in Social Research, by Stephen L. Morgan, *Technometrics, 2015, forthcoming*

Appendix 1. Characteristics of the women's magazines, MRI 2014

		Audience, 1000	Market size, 1000	Apparel consumption, % to Audience	Index	IP from all magazines model	IP from women magazines model	Contributions, % to total of all magazines
	1	2	3	4	5	6	7	8
	Women	121,781	39,190	0.32	100			
	Woman's Apparel	39,190						
1	Better Homes & Gardens	30,344	10,352	0.34	106	0.03	0.01	3%
2	Good Housekeeping	16,961	5,591	0.33	102	0.01	0.00	0%
3	Cosmopolitan	14,426	5,522	0.38	119	0.15	0.10	10%
4	Woman's World	17,290	5,432	0.31	98	0.01	0.01	1%
5	Family Circle	16,296	5,216	0.32	99	0.02	0.01	1%
6	W	10,603	4,207	0.4	123	0.17	0.15	5%
7	Glamour	11,290	4,188	0.37	115	0.02	0.02	1%
8	Southern Living	12,401	4,136	0.33	104	0.02	0.02	1%
9	The Oprah Magazine	10,347	3,749	0.36	113	0.07	0.08	6%
10	Working Mother	9,741	3,525	0.36	112	-	0.05	3%
11	Martha Stewart Living	8,878	3,379	0.38	118	0.17	0.17	9%
12	Real Simple	6,914	3,021	0.44	136	0.22	0.22	14%
13	Seventeen	7,745	2,691	0.35	108	0.08	0.05	2%
14	Allure	6,039	2,404	0.4	124	0.15	0.16	2%
15	Redbook	6,664	2,373	0.36	111	0.07	0.12	2%
16	Ebony	6,037	2,304	0.38	119	0.11	0.11	3%
17	Vanity Fair	5,251	2,219	0.42	131	0.29	0.26	9%
18	Essence	5,567	2,133	0.38	119	0.12	0.14	3%
19	Elle	5,176	2,060	0.4	124	0.13	0.17	2%
20	House Beautiful	5,301	2,007	0.38	118	0.03	0.03	1%
21	Self	5,014	1,994	0.4	124	0.11	0.19	5%
22	Shape	5,062	1,986	0.39	122	0.05	0.02	0%
23	Women's Health	6,262	1,963	0.31	97	-	0.01	0%
24	Brides	4,957	1,817	0.37	114	0.12	0.09	2%
25	Every Day with Rachael Ray	5,003	1,801	0.36	112	0.03	0.02	1%
26	Marie Claire	3,831	1,580	0.41	128	0.04	0.02	0%
27	First For Women	4,081	1,363	0.33	104	0.19	0.16	3%
28	Teen Vogue	3,274	1,302	0.4	124	0.10	0.16	1%
29	Harper's Bazaar	3,064	1,197	0.39	121	0.45	0.27	2%
30	Soap Opera Digest	3,020	982	0.33	101	0.03	0.02	0%
31	Elle Decor	2,073	859	0.41	129	0.50	0.52	3%
32	Latina	2,002	615	0.31	95	0.03	0.03	0%
33	Woman's Day	1,169	546	0.47	145	1.00	1.00	2%
34	Vogue	853	338	0.4	123	0.13	0.07	1%
	Random (% to total market)					0.29	0.15	91%