PROBABLE IMPRESSIONS:
A HYBRID APPROACH TO WEB PLANNING IN A MULTIMEDIA SETTING

Harm Hartman, Ipsos

Abstract

It is standard practice in the planning of digital media to measure success in terms of page impressions. This was a challenge that EMS, the single source multimedia currency of the affluent European across 21 markets, set out to tackle in 2011. This paper describes the design, production and usage of the EMS / comScore web planner, which was built by Ipsos.

It is a hybrid model, dividing Page Impressions into two components: the probability of visiting a certain site on an average day, and the average number of pages visited per day. For all international websites the average page views per day were derived from comScore data across the different European markets. Linking the comScore distribution of average page views to the EMS universe of main income earners from the top 20% household income was done by Data Fusion, based on Chaid Trees. To allow Reach &Frequency multimedia modelling to take place across Print, TV and web, the full binominal model needed modification to deliver average page views and total page impressions as gross reach and as an input to web planning.
Experience from planners in the last two years makes clear that using this hybrid model will change the way we plan the web. The EMS / comScore Webplanner allows user to do combined campaign planning, delivering the total reach and frequency of contacts for Print, TV or Websites individually or in combination.

Keywords

Digital, Media Planning, Probabilities, Page Views, Data Fusion, Audience Research, EMS, comScore

Probable Impressions
A hybrid approach to web planning in a multimedia setting

1. Introduction

The reason page views are ruling the digital world is obvious: they can easily be counted on each site and are therefore very suitable for planning and selling advertisements. Capping, running a digital campaign until a certain targeted numbers of clicks or page impressions are reached is daily business, but quite different from the way campaigns are run with printed ads. The requirements for audience research in the digital space are therefore also higher; to plan the web page impressions are a minimal level of required detail.

EMS, the single source media currency for the Affluent European population, offers multi media planning with probabilities based on a recall methodology. In the past, websites were measured on a patronage level, based on the probability of visiting a site on an average day in the week. The granularity of page views was absent, because respondent are incapable to recall the number of pages visited for each site. To allow for real multimedia planning EMS had to adapt page views into the framework of the R&F model. Three issues had to be tackled: reliable page view data, fusion into the EMS databases and a proper algorithm.

The solution for combining page views into the multi-media R&F is splitting page views into two components: the probability of visiting a site on a certain day and the average number of pages viewed on this day. Average pages per day per site for all European markets are delivered by comScore. Data fusion, using Chaid trees, was used to connect the average pages viewed to the right respondent. The full binomial, used within the R&F model, needed a slight adjustment to make it suitable for planning on page views.

After testing this solution in 2011, the EMS / comScore Webplanner has been in the market for two years. The experience of planning and the rapid changes in the digital space introduce new issues to be solved over the following couple of years.

2. Background

2.1 EMS

EMS, officially the European Media & Marketing Survey, has been the currency for TV viewership and Print readership among the European elite for more than 15 years. Since its start in 1995 EMS has expanded to measure the top 20% of the population in 21 countries in Western and Central Europe annually. Since 2006, EMS has expanded to the Middle East, Africa (2009) and Latin America (2013). Since 2003, websites have been part of the main questionnaire and since 2012 also apps on mobiles or tablets are measured as part of the currency.

The leading principle has always been single source measurement based on a solid CATI methodology with probability sampling, using data fusion techniques (ascription) to solve missing data issues with follow-up surveys. The benefit of single source information is the strength of the relationships between audience information, target groups and profiling information, which allows for multimedia planning. Single source information is however heavily dependent on the cooperation of respondents and their willingness to provide all this information in a sometimes long and repetitive interview or follow-up questionnaire. To solve this problem, in 2010 the methodology was changed into a mixed mode design (CATI in combination with access panels), and a Random Block Design to cut the interview length by asking only 50% of the titles, channels and sites, and using data fusion, based on Chaid trees, to complete the missing media information[1].

2.2 Multi Media R&F

Most media planning tools belong to one of two categories: probability or model based. Probabilistic tools calculate, from audience surveys, individual probabilities of contact between respondents and media, and aggregate respondent level models of exposures, for various target groups. Model based tools use, for each target group, a series of aggregated formulae (adjusted to audience results), which calculate reach and frequency distribution, from duplication matrices measures from the data[2].

From the start EMS has followed the probability framework, modeling probabilities from the combination of recent behaviour (last time read/viewed/visited) and general behaviour (normal frequency of reading/viewing/visiting)[3]. The main reasons EMS provides these probabilities for each title, channel or site are consistency and transparency:

- All software suppliers produce the same R&F results in their media planning tools, and therefore the whole industry is working with the same numbers.
- Media plans are consistently independent of the target group used, because the same individual probabilities are fuelled in the algorithms
- Transparency in algorithms used: no black box is used
- Using plain frequency distributions from the data is more a reflection of the past, using probabilities based on a combination of recent and general behaviour gives more a picture of the future.

Within a typical EMS multimedia plan probabilities of Print, TV and Web can be easily combined, although in essence we have to deal with quite different metrics:

- Print titles have a probability to read an Average Issue
- TV channels have (within EMS) a probability to see a spot within a certain day part of a weekend or week day
- Websites have a probability to visit a site on an average day (or week).

Adding Page Impressions to this model doesn't change the nature of multimedia planning: we still combine different forms of Opportunity to See (OTS), but instead of an OTS for the site, it is now an OTS for a page, carrying a form of advertisement[4]. Adding page views in the R&F model are necessary as a specific carrier of information, necessary to create credibility and alignment with the way the web is sold. You could say that Page Impressions are the GRPs for the web.

[1] Less Questions, More Data: Revitalizing the European currency in single source affluent audience measurement, Harm Hartman, Print & Digital Research Forum, San Francisco 2011
[2] Stop being Discrete, François Charton & Antoine Taconet, WRR Symposium Valencia 2009
[3] Probabilities are modeled in a way that frequent 'users' of specific media get high and in-frequent low probabilities, with corresponding recency levels.
[4] A Page Impression is in this sense a broader concept than Ad Impression, reflecting the opportunity to have seen the Ad itself. If an ad is put at the bottom, a page can be visited, but the ad not seen.

2.3 Sources of data

The measurement of online audiences can be categorized into three types:

- Site (or Ad) centric: measuring all traffic on sites (tagging), providing real (time) universe analytics on pages viewed, unique visitors, and so on
- User centric, passive registration: providing behaviour of panellist who allowed software to run on (some of) their devices to visit the web
- User centric, recall questionnaires: providing reported web behaviour (such as last time visited) using the recall capabilities of respondents

All of these measurements try to estimate the same from a different angle: the behaviour of a population on the web, and all have their pros and cons.

Site centric measurement delivers real traffic figures, without any sampling or recall error, but is not able to deliver much information on the individual site visitor, other than its ip-address and possible location. Socio demographic description or motivational / attitudinal information is mostly unknown from the anonymous site visitor, and cookie policy makes it even more difficult to link site traffic to unique visitors.

Panels with passive registration are able to connect real online behaviour with socio-demographic target group information but these solutions also have drawbacks: the demand to install software can lead to selective panels, which can be skewed on several characteristics: who are the respondents who are willing and able to put software on all or some of their devices. The rise in the use of smartphones and tablets has increased the fragmentation of media behaviour, which makes a panel centric registration method even more problematic. To get single source information we should know the digital behaviour on all platforms and devices to distinguish unique and overlap patterns of media usages.

User centric recall questionnaires, such as EMS, can provide the unique single source information on both socio demographic, motivational/attitudinal and media specific behavioural variables. Here however the limitations are also in the level of detail a respondent is able to recall, the length of the questionnaire and the possible selectiveness of the survey if too much information is asked in one survey[5]. In the past 15 years the strength of EMS has been to be able to deliver multi-media R&F on a group of international Print and TV brands within interesting target group, the top 20% household income, which no one else could deliver. Although there is a dominant trends towards passive measurement within TV and Web, for EMS the recall methodology has still been accepted. However when it comes to pages viewed, no one would believe that this is possible with recall. So page measurement has to be introduced. On the other hand there is no single provider who is able to deliver data on the affluent EMS population. The solution therefore has to be a hybrid one, using both user centric recall and passive information.

2.4 A Hybrid approach

In the case of EMS the best of both worlds is the combination of rich target group information of user centric survey with rich behavioural data from a site centric and / or passive measurement, to be merged into one data base using data fusion. To build such a hybrid solution describing the top income main income earners in 21 European countries, there are four elements to be solved:

- Decide on the metrics: when using multiple sources of data, which one will provide the leading metrics
- Find a provider of passive measured and/or site centric metrics for all EMS sites, within the EMS population
- Find a powerful data fusion protocol which respects the integrity of the data sets involved
- Find a proper algorithm to combine probabilities of EMS with page views

2.4.1 Leading metrics

If EMS was a general population survey the choice for the metrics would be easy: site centric metrics are indisputably more accurate than surveys with recall questionnaires and calibration of user centric data collection with targets from site centric measurement would be the way to go ahead. However, EMS describes the ‘affluent’ universe, a part of the population which can’t be constructed on site centric measures alone. On a site it is impossible to identify ‘affluent’ site traffic from the rest. Therefore the choice was twofold:

[5] The redesign of the EMS survey (described in “Less questions, more data”) dealt with this issue by dramatically decreasing the interview length to maintain respondent cooperation. Only half of all media information was asked, and later on the missing information was added by data fusion.

- EMS metrics as daily, weekly and monthly reach (web visits) will be leading. Although recall methodology has its drawbacks we think that questions like recency (last time site visited) and frequency (how often do you usually visit site) are good enough to produce reliable probabilities of site visit.
- Page views can't be asked in a recall questionnaire. It's clear that this level of detail is wanted and has to be provided by the external date provider. It is also clear that site centric information alone is not enough, because the connection should be made between site behaviour, such as pages viewed, with relevant socio-demographic information. Passive measurement from an external should be able to deliver this.

2.4.2 Provider for external web metrics

The choice for a supplier for the passive measurement is based on:

- Availability in (almost all) the EMS markets
- Acceptance in the market place
- Enough demographic data available to serve as 'hooks' for the modelling process
- Adequate sample size / quality
- Experience with hybrid solutions, combining site and user centric data[6]

ComScore is meeting all these criteria, but there are still coverage issues remaining.

First of all not all 21 European EMS markets could be covered: no comScore data was available for Hungary and Czech Republic. This issue was solved by taking the four Central European markets, also including Poland and Russia, as one market in EMS. In fact this is done more often. The four countries form as C-EMS already a separate entity within EMS.

Another issue is that the specific EMS population is based on a threshold only including the top20% on household income[7]. The household income threshold is however not available for all markets in a comparable format and the sample size of the higher income segments within the comScore data was insufficient. A test was conducted to investigate the impact on page views of income information, compared to other socio-demographic variables. The conclusion of this test was that other socio-demographic variables were much more discriminating than the income threshold[8]. In other words, it is better to match the (other) socio-demographic variables than to focus on the income distribution alone.

The last coverage issue concerns the specific hybrid comScore methodology[9], in which the passive measurement in the comScore panel is used to enrich the site-centric figures. If it was the other way round (site centric figures are used as target for calibrating or re-weighting panel measurement results), the weighted pages of the passive measurement could be a reliable estimate of the total population page views. Since EMS is only reporting on the affluent population, it isn't necessary to reconstruct the total population, and it isn't possible either to reconstruct the EMS population. The purpose is not calibration of site centric figures, but to get variation in page views for different type of consumers for different sites.

We can conclude that comScore data may not be the ideal data provider, missing countries, household income information and weighted population projections; for EMS it is perfectly providing the required personal information on page views for the selected international brands.

2.4.3 Data fusion

Within audience research it is widely accepted to use Data fusion to avoid missing values in case of (partial) nonresponse. The need for large 'single source' multi-media data bases has even more stimulated the use of fused data bases. A good example is the IPA/TouchPoint practice to use a media diary as linking pin (hub survey) to attach all media currencies (radio, print, TV, web) into one data base.

There are several methods for data fusion[10], most commonly used are row-wise ascription methods based on distance functions; a complete record of information is captured from the donor and transferred to the nearest recipient in terms of distance on a fixed set of socio demographic variables. This technique has the advantage that within the record of information copied all relations and values are kept intact, they remain consistent and real (existing values), but has the disadvantage to rely on the strength of the relation between 'hook' (mostly socio-demographic) variables and the relevant media behaviour. Some socio-demographic variables can be a good 'predictor' for media variable XYZ, but not for ABC, and other socio-demographic variables might be good predictors for ABC and not for XYZ. It is impossible to find a set of hooks which are

[6] comScore Media Metrix Description of Methodology: Unified Digital Measurement, August 2013
[7] EMS is labeled as top 13% personal income, but the first selection is based on household income. In the weighting it is later on projected to the 13% highest personal income of all main income earners.
[8] The other conclusion on the test ran by Carthage was that it is better to have a larger sample size on a broader population to be able to match the right socio-demographic profile. Limiting the sample to the higher income group will introduce serious sample size issues, especially for the smaller titles.
[9] The IPA TouchPoint initiative in practice, Belinda Beeftink, PrintandDigitalResearchForum, San Francisco 2011
[10] A comprehensive (but conservative) overview can be found in the ARF guidelines for Data Integration

good 'predictors' for all media (target-) variables and above all, reducing these variables to one distance causes weak relations between the socio demographic variables and the fused media behaviour. One of the well-known problems of these techniques is what is called "Regression-to-the-mean": existing relations are flattened and existing variations are decreasing towards mean values[11].

Outside the strict domain of audience research there are also alternative fusion techniques[12], mostly called imputation. Imputation is often used. The advantage is that for each target variable a model is used to analyse the relationship of the target variable with the other variables in detail to reproduce this relationship. There are also some disadvantages (and a whole literature to fix these[i]), like the problem of non-existing values, high frequency of one imputed value, assumptions on (linear) relations, and so on. This column (model) based approach was the inspiration to build the EMS Data Fusion using Chaid trees as the model to find for each media title the most discriminating split per media variable for the donors. From the end-nodes, when the groups are too small, or no significant split is possible anymore, donors are ascribed to recipients with the same characteristics, based on the same tree. This last (hot deck) ascription within the end-nodes of the Chaid tree is done using the weights of respondents to 'marry' donors with a large weight to recipients with a large weight.

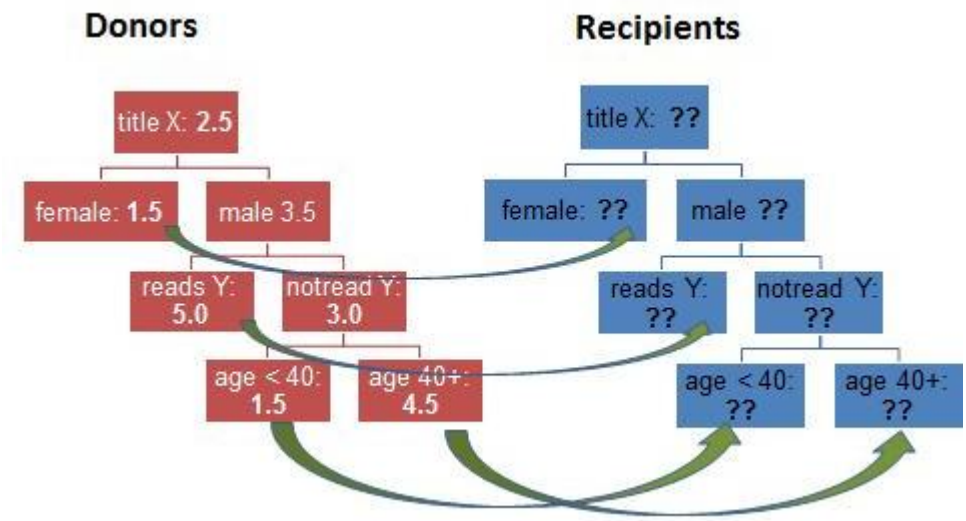

Figure 1. example of a Chaid based data fusion

In figure 1 this is worked out in a fictitious example: Most discriminating factor for title X is gender, for male we need to split between male respondents who read title Y and those who didn't. The last group can be split into age groups.
The same tree is built for the recipients and the information from the end node of the donors is copied to the recipients by systematically selecting records.

This Chaid based data fusion has several advantages[13]:

- All average levels are respected: recipients in the end-nodes of the tree get - on average - the same values as can be found among the donors.
- Ascription on end-node level copies the same variation in answer patterns: instead of assigning everyone within the end node with exactly the same value, the complete distribution and range of values of the donor are copied. So, no regression-to-the-mean.
- Chaid takes into account all possible relations (non-linear, linear), with socio-demos, but also with other titles
- It is an automated, unattended fusion process, finding only those splits in the trees that are relevant for each separate variable (title). Without prior ideas on the structure of the data, the Chaid analysis find those differences which can be significant later on, and with data fusion we construct the same structure back.

Chaid Based Data Fusion is ideal for matching the Average Page Views on the EMS set of probabilities:

- A traditional ascription method is impossible, the patterns of media (site) usage of individual respondents on the EMS and comScore can never be matched
- Only one simple variable (Average Page Views) has to be modeled (for each site)

[11] Data Fusion, A White Paper by Ipsos MediaCT, Trevor Sharot, UK 2011
[12] Statistical Analysis with Missing Data /Roderick J. Little & Donald B. Rubin, Wiley 2nd edition, Wiley Series in Probability and Statistics
[13] This is worked out in more detail in the San Francisco paper "Less questions, more data: ..".

2.4.4 R&F algorithm

After carefully evaluating the research proposals from the brief circulated to a small number of 'industry recognized' statisticians, Carthage was chosen for executing the modelling process. Not only does Carthage already perform the TV day part modelling for EMS, their proposal for website modelling was the most practical to implement. And above all, the Carthage method is not a black box:

- Presented at the 14th Worldwide Readership Research Symposium in October 2009 in Valencia, nominated for the best technical paper
- Has been exposed to critics and shared with all the players in media and was well received

The suggestions of the Carthage modelling are:

- Use probabilities
- Respect individual media performances
- Use page impressions with share of site (percentage of pages carrying the agreed advertisement) and capping (restricting the number of times (frequency) a specific visitor to a website is shown a particular advertisement)
- Add number of pages for Internet

These elements are all part of the solution chosen for the EMS / comScore webplanner. The exact implementation will be worked out in more detail in the next chapter.

3. Methodology / Implementation

3.1 Introduction

The basic idea of the EMS / comScore Webplanner is that page impressions can be split into two components:

- Probability visiting a site on a day
- Average pages visit per day for that site

The idea is as simple as it should be: if we know how many pages a person visits on each site, and we know how often he visits this site, we can easily calculate the number of page impressions generated from this person.

This chapter describes in more detail the steps from data through fusion and algorithm into planning.

3.2 Data: Probabilities and Page Views

3.2.1 Probabilities from EMS

In its fifteen years' existence EMS has always followed a probabilistic policy. For websites there is a filter on last 30 days visit, which means that people not visited a site in the last 30 days will get a probability of zero.

Probabilities are modelled in such a way that they represented adequate audience levels per day and week, based on mainly two questions:

- How often do you usually visit website A? (Frequency)
- When was the last time you visited website A? (Recency)

The recency levels for daily or weekly visits on each level of frequency are the input for the probabilities. After the modelling aggregated probabilities, representing the audience levels will be identical with the recency levels as measured in the survey. As an effect frequent (daily) visitors will have a higher probability than infrequent (monthly) visitors[14].

These probabilities only give an OTS to visit a site on an average day, but don't give information on pages viewed.

3.2.2 Page Views from comScore

Reports on site traffic from comScore follow a hybrid approach: a combination of site centric metrics with passive measurement on consumer panels[15], creating correct population figures enriched with a socio-demographic profile. For the EMS / comScore Webplanner only the passive data from their consumer panel in the European markets of EMS was needed to construct the average page views.

[14] More detail on the modeling of probabilities can be found in the Technical Description of the EMS Survey
[15] Media Metrix, Methodology Overview, Bas Bartholomeus, comScore 2013

ComScore provides 12 months of data on all domains which are part of the EMS website R&F, for all European markets[16]. The data itself contains all visit dates and number of pages viewed and a socio-demographic description of the panellist: Age, Gender, Region, Household income, Presence of Children, Education/Social Grade.

Some basic Data Processing was needed to construct Average Pages Viewed (APV):

- Monthly processing was needed to meet the criteria of the EMS probabilities. Not only because within EMS only for the people who have visited the site in the last 30 days a probability is assigned, but also because comScore itself uses months as processing entity: the pages of a person only enters the data if he or she is a full month period active member of the comScore panel.
- Aggregation of number of pages visited per site per month. If a person visits site ABC three times a month, once only one page and twice four pages, then the APV will be three (being equal to (1+4+4)/3).
- Outlier correction was performed to create stable data: APV larger than the 99th percentile are truncated.

3.3 Chaid based data fusion

Instead of providing all respondents (within a certain socio demographic target group) with one Average Page View, we would like to maintain the distribution of APVs within target groups. For instance, if French females tend to visit 2 pages on average of a certain site, we can provide all female French females with 2 APVs for that site, but because there are also within this group respondents with 1 APV, 2 APVs, 3 APVs we would like to respect this distribution by assigning the EMS respondents the same range of APVs as the comScore data. In fact what we see is that for most sites a lot of visitors usually only visit 1 page, and only few a lot of pages. Not taking this distribution into account in the modelling will result in totally different R&F although generating the same GRPs.

Data Fusion is needed to attach Average Page Views to each EMS respondent with a –small or high- probability of visiting a site and to do this a Chaid Based Data Fusion is executed:

- EMS respondents are compared to comScore panel members on the characteristics which were consistent within comScore and between EMS and comScore[17]:
    - Quintiles of probability to have visited the website yesterday
    - Country
    - Gender
    - Age in classes
    - Children in HH
    - HH Size
- Chaid Trees are built for each title to find the discriminating groups for the (logged) APVs. The natural log transformation is used in this stage to have better (more normal) distributed variables to work with
- The modelling is done in several iterative runs, starting with only the set of six socio-demographic variables, then including also the relation with other media (international print titles, international TV channels, and other websites).
- Within the end nodes of the Chaid trees the data is imputed from the comScore to the EMS data set, with a systematic ascription: ordering the donors and recipients and assigning donors to recipients with fixed (fractional) steps, to secure the distribution of the Average Page Views.
- This data fusion matches typical EMS characteristics from the total population.

[16] Except Hungary & Czech Republic: for these markets the other Central-European markets were used
[17] The variables Region, Household Income and Education-level couldn't be matched with EMS variables.

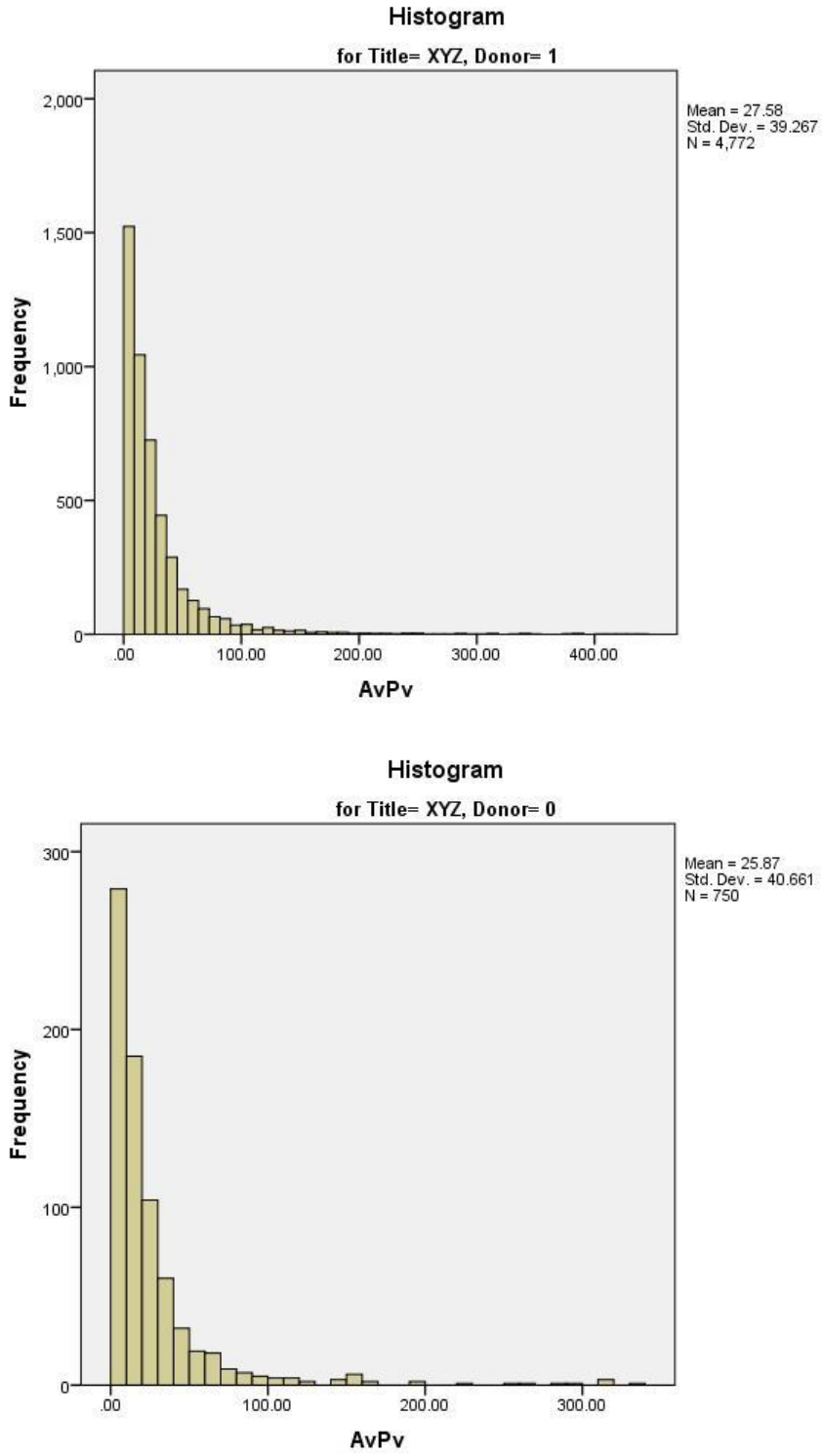

Figure 2a & 2b. example of the distribution of page view on the donor (2a) and recipients (2b) data

As can be seen in figures 2a and 2b, an example on a real but anonymous small site XYZ, the distribution of number of donor page views to a great extent resembles the distribution of the recipients. We also learn that the two distributions are very similar, but not exactly equal. The average pages viewed differs and this is logical, because the universe of the comScore (donor) respondents is different from the EMS (recipient) respondents. The data fusion process automatically detects which specific groups have more or less page views, but the relative size of this group might differ. If for example in an extreme case there is only a difference between male and female visitors for site XYZ, respectively 5 pages per day versus 1 pages per day, resulting in an average of 3 pages per day in the comScore data, because the male and female visitors are both exactly 50%. For the EMS site visitors this specific site might have twice as many male visitors as female visitors, therefore resulting in an average of 4 pages per day. The Chaid Based Data Fusion is in fact the re-weighting the comScore levels of Page Views

to the EMS universe, with remaining the distribution of the page views. Most titles have many people only visiting one page, and only few visiting many pages; this distribution is respected. It is not necessary to calibrate the number of Page Views afterwards, because this is automatically done in the data fusion process; homogeneous groups of consumers with more or less the same number of page are the source of the data fusion, but the size of those groups can differ. Small differences between the distribution (and average number of page views) between donors and recipients are acceptable due to differences in population definitions.

3.4 Binominal R&F model

Adding APVs needed adjustments to the full-binominal model, normally used in the R&F planning tool. The Carthage model, based on the paper "Stop being Discrete", introduces an extra parameter. Besides the well-known insertion, which is the equivalent of the number of Campaign Days for a certain Web campaign, we need Share of Site (SoS) to be able to deal with APVs in the binominal model[18]. Share of Site, or SoS, is the part of the site the advertisement can be viewed, because it isn't shown on each page, or because it isn't shown all the time on the page, but only partly.

$PI = \sum_{i=1}^{n} CD \times SoS \times Wgt_i \times P_i \times APV_i$ (eq. 1.a)

$CD = {}^{PI}/_{\sum_{i=1}^{n} SoS \times Wgt_i \times P_i \times APV_i}$ (eq. 1.b)

$SoS = {}^{PI}/_{\sum_{i=1}^{n} CD \times Wgt_i \times P_i \times APV_i}$ (eq. 1.c)

Page Impressions, the GRP equivalent for web planning, corresponding with the total number of visitors to a page, can easily be derived by summing among all respondents the product of number of Campaign Days (CD), the Share of Site (Sos), the indidual Weights (Wgt), Probability (P) and Average Page Views (APV), as shown in eq. 1.a

In the world of web planning and capping Page Impressions are however used as an input. It will allow a planner to calculate reach and frequency based on the input of pages. Though it does not take into account the size of the ad or where it appears in the site, it should know the percentage of a site bought (Share of Site) or additionally the length of campaign (no. of days). This is because we should be aware that Page Impressions, Campaign Days and Share of Site are interconnected; when two of the three parameters are known the third one can be derived (as shown in eq. 1.abc). The several software packages providing this website R&F had to find a suitable solution to deal with this interconnection. A constraint is of course that Campaign Days can never be a decimal number. The easiest way to solve this is to round up the number of CD to an integer number (no decimal), and adjust the SoS accordingly[19].

The rules of calculation are discussed in detail in the Charton/Taconet paper. A short explanation would be that in general a Reach & Frequency model can be represented (using Probability Generating Functions), as a distribution vector with the following polynomial, using p as probability to visit a site on an average day and cd as the number of campaign days.

$$\prod\big((1-p)+pX\big)^{cd} \; or \sum_{k=0}^{c} \binom{cd}{k}(1-p)^{cd-k}p^{k} \quad \text{(eq. 2)}$$

For each (respondent i and site x) combination a Distribution Vector can be calculated:

$V_i^x[0]$ = probability being reached 0 times = $(1-p)^{cd}$

$V_i^x[k]$ = probability being reached k times (with n insertions) = $\binom{cd}{k}(1-p)^{cd-k}p^{k}$

$V_i^x[n]$ = probability being reached as many times as number of insertions = p^{cd}

[18] In the original paper of Taconet & Charton ("Stop being discrete") the suggestion is done to also incorporate the variation of page views. This is a clever observation because a person who is always visiting 2 pages on a day he or she visits a site is different from a person who most of time only visits 1 page, but once in the 10 days 10 pages. Adding variation in the R&F model would further complicate the calculation, and the basis of comScore data was too small to provide a reliable estimate of variation of page views on a day, so this suggestion is not (yet) implemented.

[19] So first the CDs are calculated as in 1.b using a fractional outcome, then the SoS is calculated using the rounded up number of CDs from the previous calculation.

Combinations can easily be calculated by combining distribution vectors of different sites (or instance x and y):

$V_i^{xy}[0] = V_i^x[0] * V_i^y[0]$

$V_i^{xy}[1] = V_i^x[1] * V_i^y[0] + V_i^x[0] * V_i^y[1]$

$V_i^{xy}[2] = V_i^x[2] * V_i^y[0] + V_i^x[0] * V_i^y[2] + V_i^x[1] * V_i^y[1]$

or

$V_i^{xy}[q] = \sum_{a,b} V_i^x[a] * V_i^y[b]$, where a+b=q

Extending the normal binominal R&F model to page views introduces the new parameters α and s.

$$\prod \left((1-p) + p\left((1-s) + sX \right)^{\propto} \right)^{cd} \quad \text{(eq. 3)}$$

For each respondent site we have:

The probability to visit the site on an average day (➔p)

The number of pages visited that day, or APV (➔α).

And for a given plan we have:

The number of insertions, or campaign days (➔cd)

The share of site, the part of the site bought by the campaign (➔s)

Naturally, when the complete site is bought (s=1), the above formula becomes

$$\prod \left((1-p) + pX^{\propto} \right)^{cd} \quad \text{(eq. 4)}$$

And also, when all respondents visit only one page the formula can be simplified to the well-known full binominal (eq. 2).

The exact way to work out all the details of this new R&F with Page Views will take ten pages of formulas and that goes beyond the scope of this article. A comprehensive document with rules of calculation is provided to all software houses working with the new model[20].

3.5 Alive and kicking

In 2011 the first version of the EMS / comScore Webplanner was produced unofficially for testing only. Planners working at media owners (and subscribing to EMS) were asked to evaluate the data, the model and the software. Deliverables in the software must allow commonly used online campaign metrics to be used as inputs; and common media campaign metrics to be derived as output. The input for website Reach and Frequency (R&F) normally is page impressions (or impacts), which allows the planner to calculate R&F (the output) based on the input of pages, but does not take into account the size of the ad or where it appears in the site.

The main issues from the first testing experience turned out to be the user-interface, and the way the different packages were dealing with the interrelated Campaign Days, Share of Site and Page Impressions. Some forced the user to work with a fixed selection of two (for instance Page Impressions in combination with Campaign Days), others left it open to the user to decide which two of the three were needed to specify a plan. Several workshops were organized with planners, software houses, and EMS representatives.

In 2012 and 2013 the official EMS releases are delivered with the so called EMS / comScore Webplanner. It is a new tool and planners need to become familiar with the model, but the general feedback was positive.

Three issues were mentioned which show the urgency to keep on developing on these kind of models: how to cope with the discrepancy between the EMS population and general population, the translation from Page Impressions to Ad Impressions and the on-going media fragmentation.

[20] Software houses that incorporated the new Website R&F are Telmar, IMS, Kantar, PeakTime, New Age, Sesame

4. Future development

4.1 EMS Population versus total population

The EMS / comScore Webplanner is aiming to provide the planner with a tool to evaluate plans and allow consistent multi-media evaluation and allow websites to be analysed on a stand-alone basis, or with other media in a multimedia campaign R&F.

The usual practice of working with Page Impressions is now part of the R&F model, but faces the planner with a dilemma: Page Impressions as input for a media-plan, as part of a contract with a site owner to stop an advertisement as soon as the targeted number of page views are reached, is ideal in the general population, but not for the higher end affluent population of EMS: on the site it is not possible to recognize an affluent visitor from a normal consumer. It is especially confusing when comparing sites with a totally different background, for instance a financial site which is mainly visited by senior management or C-Suite, versus a general sport site which is visited from low till high. Comparing the sites planning the same number of Page Impressions on both sites seems fair, but for the general site it takes much more Page Impressions in total to only deliver that amount of Page Impressions within the Affluent population.

The proposed solution at the moment is simply to learn to plan in Campaign Days and Share of Site, and letting Page Impressions (within the Affluent population) be part of the output rather than of the input. Campaign Days and Share of Site can of course be determined on general population media planning. Page Impressions are similar to GRPs, no more and no less.

A possible solution to be worked out in the future is to introduce factors in the model, representing the specific distribution of a site within the affluent population compared to the total population. This is not a difficult solution and could help making different sites better comparable, but the downsides of this solution are at the moment decisive. First of all, it will complicate the modelling, and the R&F model even more, possibly leading to lack of transparency in the communication of the R&F model and possible loss of credibility. A simpler, more robust model will gain easier acceptance. The second, and most important reason, is that it can only be done on total level, not allowing to do make specific target groups in the campaigns. This is from a user point of view an unacceptable limitation.

4.2 Ad versus Page Impressions

A similar problem is the wish to work with Ad Impressions instead of Page Impressions. And in the digital world of clicks and tags, this is a relevant issue, but for the EMS / comScore Webplanner this goes beyond the goal. It is clear that Ad and Page Impressions are not the same, but there is no reliable source of information available to create factors for each site representing the relation between the two. After discussion with the EMS subscribers the provisory solution was simply to put the relation to unity (one). This doesn't mean that in the future a different solution (with factors) can be worked out.

4.3 On-going Media fragmentation

During the time the EMS / comScore Webplanner was designed, implemented, tested and deployed for a web planning R&F tool, the digital world evolved further: a rapid rise in smartphone and tablet usage, the introduction of smart TVs, the increasing app usages, not only mobile devices, but also on PCs due to Windows8 and in the future on the TV set, E-reader and game console. It is clear that a solution which if fuelled with site traffic on PCs only is no longer sufficient and EMS is already active in incorporating the wide scope of digital platforms in its questionnaires. In the near future new solutions, sources and models must be worked out to catch the wide range of modern digital behaviour. Fortunately we see that partners like comScore are making the same judgements and are constantly investing in new sources of information.

5. Conclusion

Media behaviour is getting more and more fragmented, making recall questionnaires almost impossible. Passive measurement is able to deliver the highest level of granularity, but fails to deliver single source information in combination with the richness of target groups or additional motivational, attitudinal data. When we want to do multimedia planning with Page Impressions, the best has to come from both worlds in a hybrid solution, adding page views from passive measurement into the single source multimedia database.

In a hybrid solution data fusion is needed to connect different data sources. The EMS / comScore Webplanner was constructed to deliver page impressions for the affluent consumer in 21 European Markets, connecting site visit probabilities from the EMS survey, with the passively measured page views from comScore. Data fusion based on Chaid trees is an elegant way to do this.

EMS is reporting on the affluent space, and it is hard to find an external provider delivering exact the same population of the top 13% personal income. The solution was found in comScore, being able to provide the necessary information: how many pages on average do different consumers in Europe visit certain international sites.

To incorporate average page views in the R&F model, the full binomial needed some modification. Planning the web now involves not only Page Impressions, but also a mixture of Campaign Days and Share of Site.
Ipsos delivers with the EMS / comScore Webplanner an innovative solution to combine probabilities with page impressions, opening the world to true multimedia planning.