"DIFFERENT FROM YOU AND ME"

Ian Garland, Milton Data
Babette Lazarus, Condé Nast
Mark McLaren and Maria Cole, Ipsos MediaCT
Josh Chasin, comScore

"Let me tell you about the very rich. They are different from you and me."
– from "Rich Boy" by F. Scott Fitzgerald, 1925

1) Introduction

In early 2013, IPSOS MediaCT and comScore jointly agreed to create a fused data set which integrated the 2012 Ipsos Affluent Survey: USA (formerly Mendelsohn Affluent Survey and referred to for the remainder of this paper as Ipsos Affluent Survey or IAS) with comScore data. Amongst other things, the integrated information will be used to estimate the online and cross-media behaviours of affluent US. While the data set is released as a syndicated service, the Conde Nast company has been closely involved in the early testing of the dataset. The fusion was conducted by Ian Garland of Milton Data, a media and marketing analytics company and consultant to Ipsos.

In setting up the parameters for the data integration, the stakeholders sought to:

a) Achieve full use of the Ipsos data set
b) Preserve the comScore currency data
c) Reflect the differences in behaviours, such that they exist, between different groups with the affluent strata of US

In addition to these objectives, the parties also wanted to try and match the cross-platform duplications suggested by the primary data in the Ipsos survey as far as possible. While not a control parameter, this last objective was seen as being part of the validation of the fused data set. The initial fusion was conducted using the 2012 Ipsos Affluent Survey and the comScore Media Metrix data from March 2013. This paper documents the various stages completed in developing the fusion of these two data sets and concludes with some thoughts on next steps and areas of further research.

2) Data Sets – 2012 Ipsos Affluent Survey

The Ipsos Affluent Survey (IAS) is designed to measure and analyze the U.S. population aged 18 years and over who live in homes with household income of more than $100,000 per annum. The field work for the 2012 Ipsos Affluent Survey was conducted from April to June 2012 and obtained completed interviews from nearly 14,000 respondents. The resulting data set, weighted to Census projections for the measured universe, provides an extensive collection of consumer data including:

a) General demographics such as age, gender, education, marital status
b) Measures of affluence such as household & personal income, net worth & liquid assets
c) Socio-economic details such as household/family composition, occupation, industry
d) Consumer insights regarding planned life events, sports, exercise & leisure activities, brand usage and spending patterns
e) Media use, including 40 advertising touchpoints, readership of 155+ publications, and use of specific TV networks and 300 websites

While the IAS includes information on a large number websites, the website usage data is based on the respondent's recalled use of the site over the last thirty days. The fusion exercise was designed to integrate the currency web activity data from comScore with the rich consumer profile data available in the Ipsos data set.

3) Data Sets – comScore Media Metrix Data (March 2013)

The comScore Media Metrix service provides reports for online usage and visitor demographics for home, work and university audiences across the U.S. national market. For the initial fusion, comScore provided the U.S. Media Metrix data for March 2013, which included monthly site visitation from a sample of approximately 200,000 panelists aged 2+. These respondents are weighted to U.S. Census demographic targets for general population and site visitation levels are adjusted to incorporate both panel- and server-based data using comScore's Unified Digital Measurement process. This approach, which calibrates the panel-based audience estimates to the census-level detail from website servers, ensures that the Media Metrix reports account for 100% of a site's audience. While the comScore data set has an extensive list of websites, the range of demographics available for the fusion process is relatively modest, comprising the following details:

Demographic	Notes
Panellist ID	Unique ID
Age	Eight age groups
Gender	Male, Female
HH Income	Six breaks, max $100k+
Presence of Children	Yes/No
HH Size	Five breaks to Size=5+
Region	Nine US Census Division breaks
Race	Black/Other
Ethnicity	Hispanic x 3/Other

The limited number of demographics also restricts the selection of linkage variables to be used in the fusion, particularly with regard to the parameters which can be used to the wealth and income characteristics in the comScore data set. Specifically the comScore data set has only one economic indicator variable for the panel (household income) and this variable has a maximum classification for panel members value of $100k+/year. There is no other income or economic variables in the comScore data which can be used to match the wide range of economic profile data available in the Ipsos survey. In addition, the lowest level of geographic classification available in the standard comScore data set is Census division, which meant that it was not possible initially to match the samples on designated market areas or to infer income breaks in the comScore data using zip codes.

In response, comScore provided Ipsos with a custom data extract which allowed the team to map every comScore respondent to a zip code. This meant that respondents could be classified using median household income data from the U.S. Census data or other sources zip code based income data. To understand the limits of zip code data it is worth reviewing what a listing of US zip codes can reveal. For example, the 2010 U.S. Census data shows under 3% US zip codes across the 50 states have a median income of $100K or more, with those 1,181 zip codes accounting for 5% of the US population.

US MEDIAN HOUSEHOLD INCOME DISTRIBUTION BY ZIPCODE (US CENSUS, 2010)

Median Income	#ZipCodes	%	Pop. (000s)	%
<$100k	39,879	97.1%	292,353	94.7%
$100K+	***1,181***	***2.9%***	***16,385***	***5.3%***
$100-$150K	1,025	2.5%	15,133	4.9%
$150-$250k	143	0.3%	1,248	0.4%
$250k+	13	0.0%	5	0.0%
Grand Total	***41,060***	***100.0%***	***308,738***	***100.0%***

To overcome the limits of resolution of individual zip code data, we sought to create a stratified wealth model which used a rank order of zip codes based on median household income data from U.S. Census, supplemented where possible from with zip code level household income data from the last three years of Ipsos Affluent Surveys. This ranking was then divided into three groups, based on estimated population, to create income "tritiles".

The process for establishing the tritiles was as follows:

a) Rank all U.S. zip codes in descending order by median household income using 2010 U.S. Census data and zip code based household income data collected in the last three years of Ipsos Affluent Surveys
b) Eliminate all zip codes that are not present in the respective Ipsos and filtered comScore data sets used in the fusion
c) Identify the zip codes which account for the top 20% of the rank order on an accumulated population basis based on 2010 U.S. Census data (this selection of zip codes collectively known as Tritile 1)
d) Identify the zip codes which account for the next 30% of the rank order on an accumulated population basis (Tritile 2)
e) Classify the remaining zip codes as belonging to Tritile 3

We initially investigated if it was appropriate to use a more uniform (quintile) distribution, but since a key objective of the fusion is to establish the extent to which the fused dataset can successfully distinguish between the Ultra-Affluents (Household Income $250k+/year) and the Mass-Affluents (Household Income $100-$249k/year), the disproportionate segmentation approach provided a greater level of discrimination during the fusion process. While precision of the comScore data set cannot be verified (since the actual distribution of income is not known for that sample) we verified that the tritile classification was a useful discriminating tool by looking at the cross-tab of the tritile classification (based on zip code) with the claimed household income data from the 2012 Ipsos Affluent Survey.

As the table below shows, the majority of the Ultra-Affluents (68% of that segment) are members of the top tritile, while conversely 67% of the lowest segment of the Mass-Affluents ($100-$150) have 67% of their members in the lowest tritile. We were thus satisfied that the tritile segments would assist in identifying which respondents in the comScore data set which were likely to be from zip codes with a high proportion of upper-income homes.

	Household Income Segments			
IAS 2012 Cross Tab	**$100-$150k**	**$150-250k**	**$250k+**	**$100k+**
Tritile 1 (Top 20%)	12%	18%	68%	20%
Tritile 2 (Next 30%)	21%	47%	28%	30%
Tritile 3 (Next 50%)	67%	36%	4%	50%
Total	***100%***	***100%***	***100%***	***100%***

4) Fusing Two Currencies

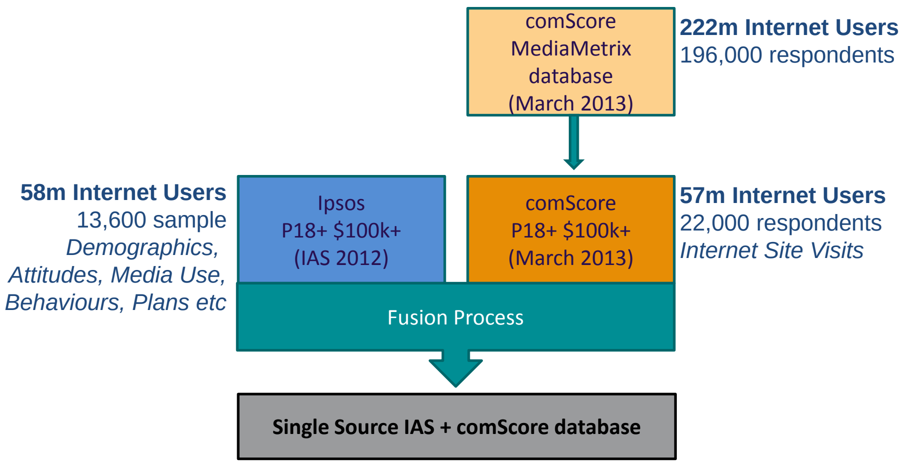
The primary objective of the any data fusion is to combine the two data sets in such a way that they appear to come from a single sample. During the fusion, individuals from one survey are matched to individuals in the other and the two sets of behaviours are jointly ascribed to the matched individuals. For sake of understanding the process, it is convenient to nominate one survey the **donor** and the other the **recipient**. In this fusion, we consider the comScore online data the donor survey and the Ipsos Affluent Survey the recipient. This interpretation reflects the result that, after the matching process, it can be imagined that the comScore panel members donate their online behaviour to the matched recipient individuals in the Ipsos survey. This then creates a set of pseudo-respondents who have all of the Ipsos characteristics (demographics, media usage, purchase behaviours, attitudes etc.) and the ascribed online behaviours from the comScore donor. A necessary pre-condition for the fusion is that the surveys come from (at least approximately) congruent universes and that the demographic and behaviour profiles of both surveys are sufficiently similar that they can be properly aligned.

As part of the data preparation, the following steps were taken to ensure that the two surveys could be integrated:

i) Defining the Universe
ii) Matching the Samples
iii) Linkage Variables
iv) Matching Process
v) Preservation of Currencies
vi) Data Validation

5) Defining The Universe

In an ideal world, the two surveys would be constructed from the same sample frame, would have exactly the same demographic profile and would be identical in size so that one individual in the Ipsos survey could be matched with one and only one respondent in the online panel. The reality in this case is that the universes and sample are slightly different.



As an initial step, since we are integrating internet users, we eliminated non-internet users from the Ipsos data set (less than 2% of the sample on a weighted basis). The comScore Media Metrix data for March 2013 was based on just under 200,000 respondents representing the general population (people age 2+). To align the two data sets, we created a custom extract of the comScore data which matched the Ipsos survey design (People age 18+, living in homes with household income over

$100k). These two data sets, one representing an estimated population of about 58 million and the other representing 57 million people, formed the primary data sets used in the fusion.

Since the Ipsos data is the recipient dataset in this fusion, by definition all of the characteristics of the Ipsos survey will be preserved following the fusion and cross tabs and demographic compositions from the original Ipsos survey will be recreated completely in the fused data set. To achieve a similarly precise outcome for the donor after fusion, an ideal situation would be that the data set from comScore would match the recipient Ipsos survey in the following ways:

a) Identical universe estimates
b) Identical demographic compositions
c) Identical sets of behaviours
d) Identical sample sizes with a comprehensive and mutually set of linkages between the respondents in both data sets.

In reality there is never a perfect match and some adjustments are required as the donor (comScore) universes are adjusted to accommodate the different universes found in the recipient (Ipsos) data set. The following table compares the universes between comScore and Ipsos data

	Universe Estimates P18+ in $100k+ Homes (000s)		
	IAS	COM	%Dif
Female	***28,548***	***26,218***	***109%***
18-34	7,215	7,955	91%
34-54	13,982	10,882	128%
55+	7,351	7,381	100%
Male	***29,661***	***30,831***	***96%***
18-34	8,830	9,639	92%
34-54	12,379	11,950	104%
55+	8,453	9,242	91%
Total	**58,209**	**57,049**	**102%**

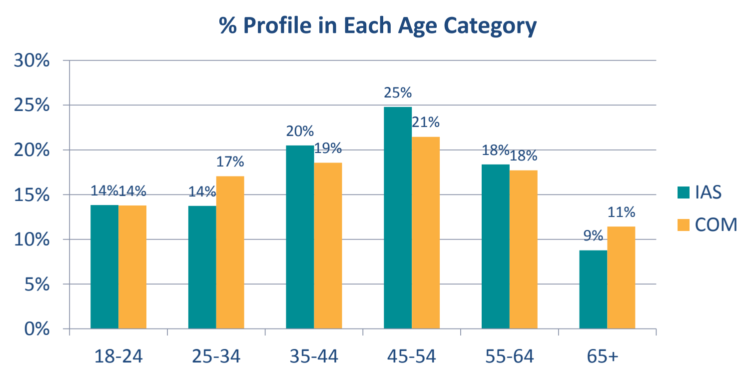
Sources: IAS=2012 Ipsos Affluent Study UE; COM=March 2013 UE

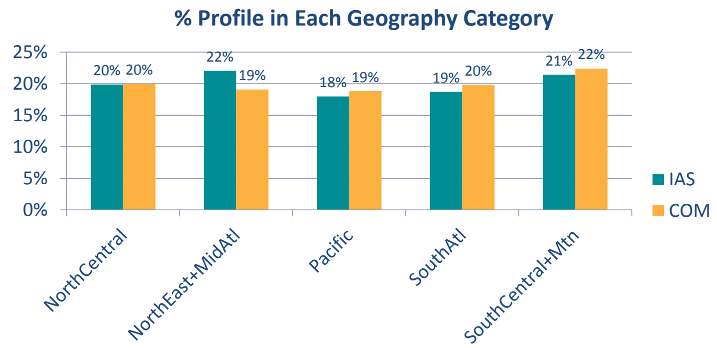
It shows that, while the total number of people in the two surveys is very similar, there is a significant difference in the surveys in certain age and gender cells. For example, the table above shows that while there is a modest 2% difference between the comScore and Ipsos universes (58 million active internet users in the Ipsos survey compared to 57 million in the comScore data), the Ipsos data shows 9% more active female internet users than comScore and the discrepancy is greatest in the middle age female group where the Ipsos data has a potential audience 28% higher than the comScore data.

These difference mean that behaviours transferred from comScore may change accordingly in absolute level (thousands) due to the relative differences in universe estimates for these key demographic cells. We quantify the effect of these differences on an overall basis later in the paper (see section 9).

6) Profile of Samples

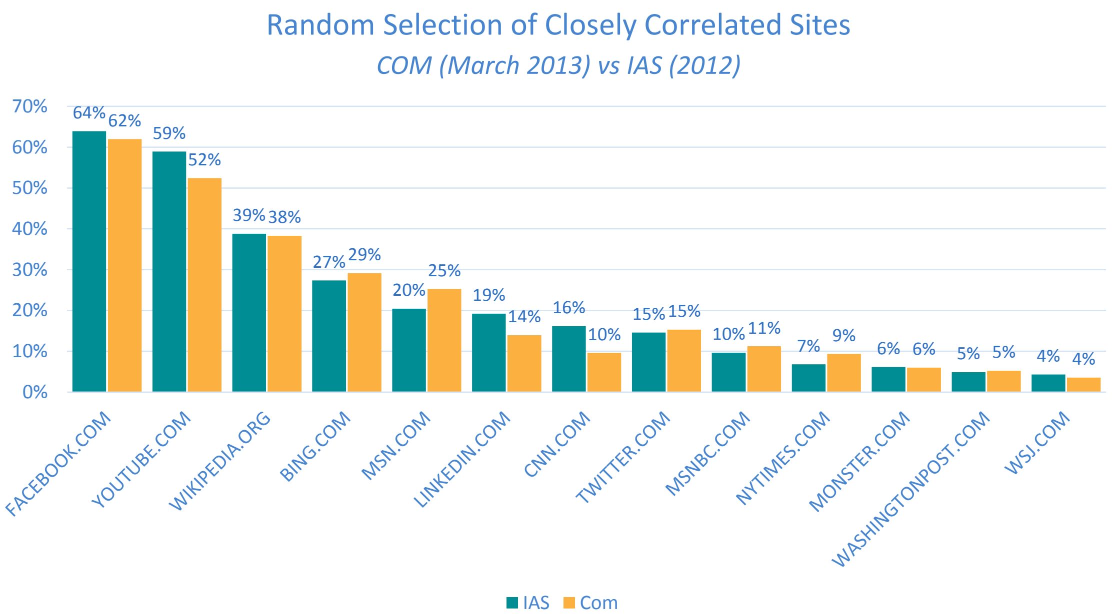
To establish the extent to which the two data sets showed concordant profiles and behaviours, the fusion team reviewed a wide range of characteristics for both Ipsos and comScore samples, including profile data for demographics, geography and online behaviours.



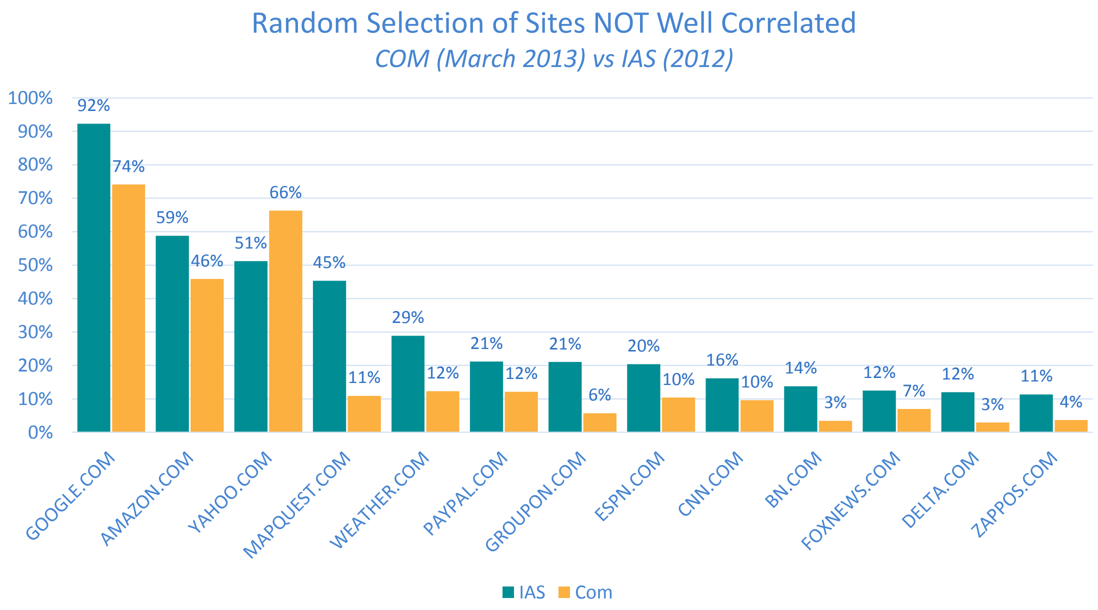


Apart from the previously noted differences within age bands, the Ipsos and comScore sample compositions were very similar for most other broad demographics. While the age and geography comparison data is shown in the charts above, we also compared other demographic groups including ethnicity, race and presence of children, and found no other significant demographic profile differences.

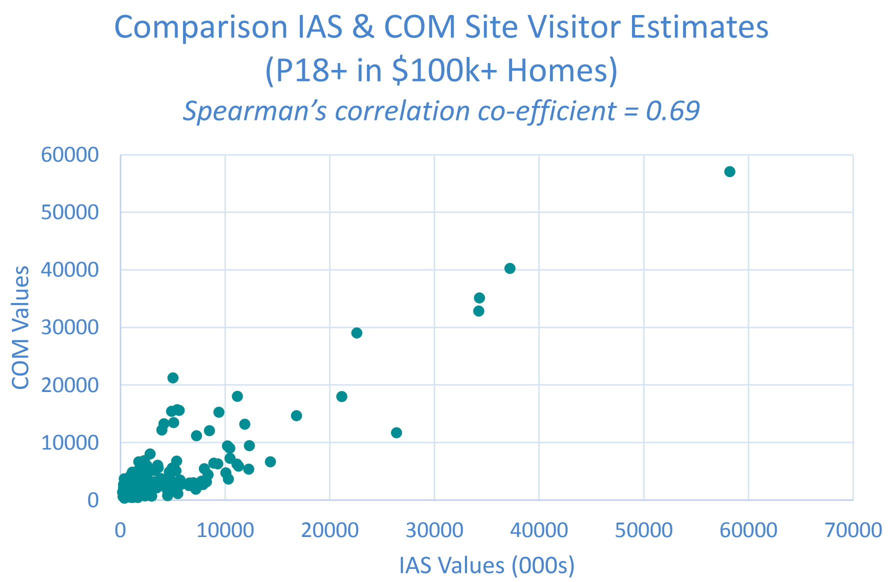
In terms of behavior, the main metric common to the Ipsos and comScore data sets was estimates of the number of website visitors, with both surveys providing estimates of 30 day usage for a range of sites. In some cases, the number of visitors (expressed in the charts as penetration against the active internet universe in each sample) were quite similar, despite the differences in data collection for the recall-based Ipsos survey and the observation-based data of the comScore panel (see chart below).



There were also a number of sites with significant differences. While some of these differences were attributed to definitional issues (for example the prospect of confusion between a major portal site and the search function of that site), the sites with the most significant differences tended to occur in sites that may have benefitted from a market-leader or big marketing presence affecting the ability of an Ipsos respondent to accurately remember when they last used the site.



While there were about 300 sites in Ipsos survey on which usage data was collected, the test was limited to 166 sites that were present in both Ipsos and comScore surveys and for which the sample count of active users in the comScore data was greater than 100 so that sample variability would not be significant. A scatter plot of the Ipsos and comScore data shows some dispersion, reflected in the correlation of 0.69 calculated using a Spearman rank correlation.

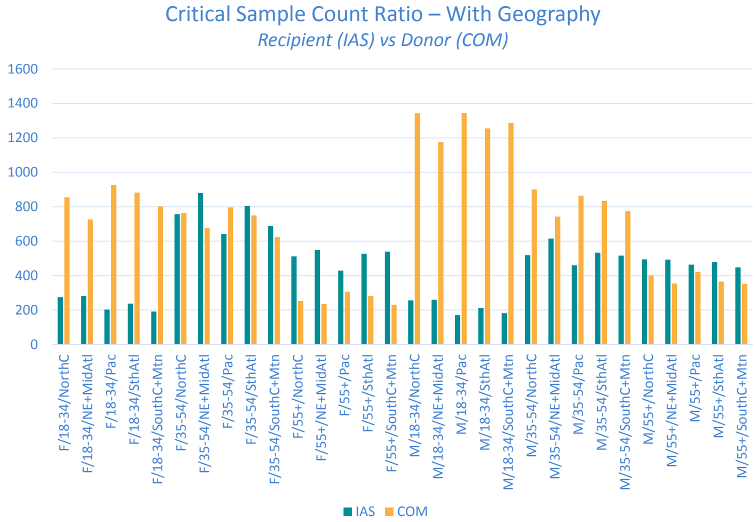


7) Linkage Variables

Ideally, the linkage variables selected should accurately discriminate the incidence and frequency of the various behaviours being donated to the recipient data set. While it is tempting to include as many variables as possible, various studies (e.g. Baker in particular) have shown that an abundance of matching variables tend to result in poorer fusions. In the case of the Ipsos/comScore fusion the list of demographic variables available for matching is relatively modest due to the small number of demographic classification that are common to the comScore and Ipsos surveys. Following established protocols, we selected ***critical*** and ***matching*** variables.

The ***critical variables*** are those chosen for which a match is mandatory. By common practice, and for obvious reasons of research integrity, gender is always a critical variable. While the selection of other critical variables is a matter of some discussion in the literature, there is no other variable apart from gender which is consistently used as a critical variable. In general, broad age bands, geography and socio-economic status are often candidates since they force a minimum level of consistency between the fused data set and the currency for key demographics. Since there was no reliable socio-economic classification for both Ipsos and comScore data sets, we reviewed the inclusion of broad age bands and geographic as critical variables.

While there are sufficient sample elements to provide a robust pool of donors and recipients, a number of the cells will result in a significant ascription ratio (e.g. each female 18-34 in the IAS data will obtain data from an average of 3.5 donors). Indeed, we experimented with a number of other critical cells including using geography (e.g. Census division), presence of children and ethnicity/race interlaced with the age/gender cells. In each case, the resulting cell count ratios in some critical cells became extreme and, post-fusion, resulted in a data sets that generally did not meet the wealth distribution sought by the stakeholders (see section 10 for further discussion on this element of validation).



In the end it was decided that the critical cells would comprise the following list:

a) Web Active (Yes)
b) Gender (Male/Female)
c) Broad Age (18-34, 35-54, 55+)

The following table shows the sample sizes available for fusion within each of the critical cells.

	Critical Cell Sample Counts (Web Actives)	
	IAS	**COM**
Female	***7,515***	***9,108***
18-34	1,189	4,192
34-54	3,770	3,611
55+	2,556	1,305
Male	***6,103***	***12,415***
18-34	1,082	6,403
34-54	2,643	4,115
55+	2,378	1,897
Total	**13,618**	**21,523**

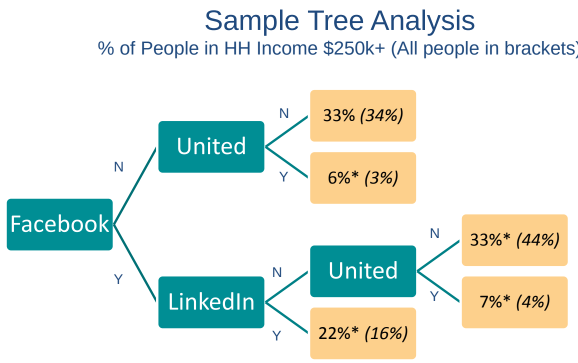
Sources: IAS=2012 Ipsos Affluent Study UE; COM=March 2013 UE

The ***matching variables*** are used to further predict or explain the variables being linked within each critical cell. Various approaches for the selection of matching variables have been proposed including a range of multivariate techniques (e.g. Wilcox using ANOVA, Doe and Collins using Principal Components analysis). In each of these approaches the generalized outcome is to identify a subset of independent variables that can most reliably determine dependent behaviours.

The initial attempt to identify the matching variables included preliminary analysis using both logistic regression and discriminant analysis. These were seen as being more appropriate given their applicability to the categorical variable data sets like the ones used in the fusion, and seen as preferable to using principal components (PC) given the mathematical principles underpinning principal components is largely dependent on modelling continuous variables. Tree classification techniques also provide another approach, with techniques such as AID and CHAID (Chi Square Automatic Interaction Detector) effectively conducting multiple significance tests of various combinations of the independent variables to establish the best predictor of the dependent variable(s).

Like the regression based approaches, the tree analyses are used to identify the independent variables that best account for variability in a dependent variable. Because of our desire to ensure the fusion retained any behavioural differences that existed within the different wealth strata the initial analyses were conducted to identify candidate variables that reliably segmented various facets of a respondent's wealth status without including in the independent variables specific reference to those same wealth measures.

For the analysis, a multinomial logistic discriminant analysis (a form of tree analysis) was conducted on a variety of wealth indicators (household income, net worth, liquid assets and personal income). As an example of the output, we have included one of the generated trees. In the example below, Facebook, LinkedIn and United Airlines' websites provided a segmentation that improved the precision of the (imputed) wealth measures for the donor data. For example in the tree below, the segmentation shows that 22% of the people living in homes with income of $250k (the Ultra Affluents) used both Facebook and LinkedIn over the last 30 days, a level which is significantly higher than the 16% of general population.



*Proportion of HH Income segment in each group (all people% in bracket); segments statistically different to all people market with **

The independent variables used in the initial modeling included a wide range of candidate websites that are common to the Ipsos and comScore surveys (e.g. Facebook, Twitter, Wikipedia as well as specific media titles such as FT.com and Bloomberg – Appendix 1 has the entire list of candidates considered). This process was repeated numerous times to identify the list of sites and demographics which were useful in developing predictor variables for key sample characteristics and behaviours we were seeking to preserve in the fusion (e.g. usage of specific sites classified by various wealth). Following these analyses, we selected the following critical and matching variables:

Critical: Web Active, Gender, Broad Age

Matching: Age Band, Age Group (finer details), census division, state, city, tritile, top 25 city status, top 10 city status, presence of children, household size, race (binary match, Ethnicity, Web H/M/L, various social media links including Facebook.com, LinkedIn.com, Usage of specific web categories (Business, News, Travel etc.), usage of specific online properties including selected media and commercial websites.

8) Matching Process

The final design decision associated with the fusion was to determine the distance metric to be used, to consider the role of importance weights and to decide on the data integration algorithm to be used.

Distance Metric

During the development of the fusion process we contemplated and tested a number of distance metrics. Initial work focused on Euclidean and Mahalanobis and a variation of the Manhattan block distance. We excluded the Mahalanobis metric for a couple of reasons, including the added complexity and processing time required to calculate the inverse of the co-variance matrix, and because there seems to be a significant difference of opinion in the literature on the subject to date. In particular, while there are some advocates (Doe, Wilcox) who argue for its use because it accounts for covariance while others (Jephcott & Bock) have a published opinion that they aren't convinced of its merits. Of more concern is that the mathematics behind Mahalanobis distance favours continuous variables while the linkage variables in this fusion are primarily categorical in nature. As a result, we used a variation of the Manhattan block distance which, given the binary nature of many of the linkage data variables gives a distance metric which is very similar to the Euclidean distance in this instance.

We have not discounted the use of the more complex distance metrics (like Mahalanobis) but need additional time to establish the extent to which the metric substantially improves the quality of the fusion given the cost of implementing the calculation.

Importance Weights

We also spent considerable time reviewing the alternative approaches for establishing importance weights. Again the literature offers mixed solutions and recommendations on the choice of importance weights. Both historical work (RSMB) and recent work (MRI, Nielsen) suggest that the average of the loadings found in a range of factor analyses, with key behavioural variables being the dependent variable, provide summary of the relative importance of each of the matching variables in the process. Our concern with this approach is that the weight for each variable can have a significant variance because of the relative and changing importance of each variable in its correlation with the dependent variables.

For example, during the early testing in the fusion we conducted multiple regressions of the candidate matching variables against a range of numeric dependent variables captured in the Ipsos survey (# sports played, # cars owned, # cultural events attended etc.) to establish the importance of each independent variable. The "importance" is an estimate of the sensitivity of the dependent variable to changes in the independent variables. For example, if a regression model shows that the effect of a small change in one variable is twice the effect of a similarly small change on another, then the importance of the first variable will be computed as twice that of the second.

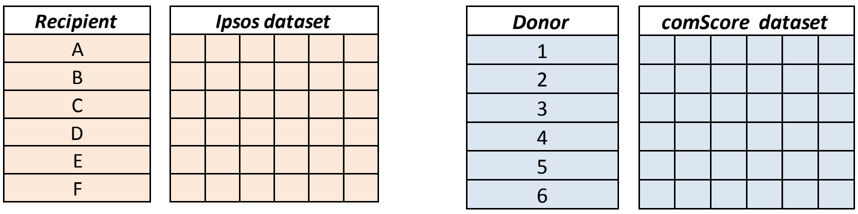
We experimented with determining weights for the independent variables by using multiple regressions on the normalised independent variables and calculated the average importance metric for the range of variables (an extract shown below). As the table shows in the Relative Standard Error column ("RSE"), an independent variable can show considerable range in terms of the importance of that variable to different dependent variables.

	Average of Importance Various Dependent Variables										Statistical Summary		
Various Demographics	# Sports	#Cars	#Cultural	#Gadgets	# Print titles	HH Income	HH Net Worth	Hours Radio	Hours TV	Pers Inc	*Average*	StdDev	RSE
Q18j. Age	9%	14%	5%	16%	8%	6%	30%	11%	16%	23%	***13%***	8%	59%
Q20m. DMA (Designated Market Area) Rankings	14%	15%	9%	9%	8%	18%	17%	16%	17%	6%	***12%***	4%	36%
Q18e. Number of People in Household	2%	32%	5%	9%	2%	7%	4%	6%	5%	13%	***8%***	9%	109%
Q18e. Number of Children Under Eighteen in Household	2%	15%	8%	8%	7%	2%	6%	13%	14%	5%	***8%***	5%	59%
Q20m. Census Division	6%	6%	4%	3%	3%	6%	9%	5%	7%	2%	***5%***	2%	41%
Q18a. Respondent Gender	17%	0%	2%	0%	4%	3%	1%	3%	2%	10%	***4%***	5%	126%
Q5a. Websites Visited in Past Thirty Days: Twitter.com	10%	1%	4%	6%	3%	1%	1%	3%	6%	1%	***3%***	3%	85%
Q5a. Websites Visited in Past Thirty Days: LinkedIn.com	2%	2%	5%	2%	2%	7%	1%	3%	2%	5%	***3%***	2%	64%
Q5a. Websites Visited in Past Thirty Days: Facebook.com	3%	2%	3%	8%	0%	2%	1%	3%	3%	2%	***2%***	2%	86%
Q18c. Hispanic Descent	0%	2%	1%	2%	2%	2%	3%	0%	1%	1%	***1%***	1%	72%

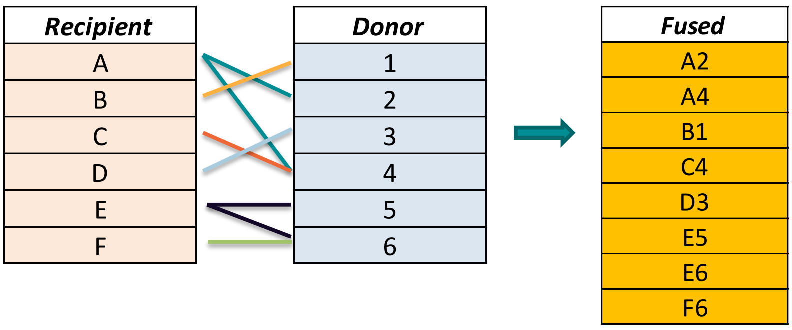
In addition to experimenting with importance weights derived in the fashion documented above, we also developed another set of importance weights which are proportional to the relative penetration of the various linkage variables. In both case we found that the quality of the fusion, in particular as measured by the median income ranking for various websites, was not improved and in some cases had a deleterious effect on the matching metric. Consequently, the critical and matching variables are allocated uniform importance weights for the beta fusion.

Fusion Algorithm

The Ipsos comScore data integration is classified as a constrained fusion since all of the recipient and donor individuals are used completely. Drawing on optimization techniques from Operations Research, the fusion approach uses a variation of the transportation algorithm to preserve the recipient and donor data sets and relativities.



As noted earlier, the matching process identifies the best match within the critical cells and – if necessary creates a partial record which ensures all donors and all recipients are used in the process.



Following this initial fusion, the Ipsos comScore fusion requires an additional step to correctly allocate the online audiences.

9) Preserving Currencies

To understand the challenge of preserving the online audience estimates, it is worth summarizing the issues associated with integrating the comScore unified data and the Ipsos survey. In simple terms, the comScore unified data approach means that each respondent in the comScore data file could account for a different number of visitors for different websites. This is in contrast to most other surveys where a respondent typically represents the same number of people for any observed behaviour. For example, and more typically, if a respondent in a survey was weighted to represent 2,000 real people, any of their activities (such as a visit to a site) would also be imputed to have occurred for the same number of real people.

In the comScore data, a specific respondent – who might account for 2,000 internet users for demographic purposes - might represent 1,000 visitors to website A, 2,000 visitors to website B and 3,000 visitors to website C. This causes problems in the transportation algorithm since it typically matches individuals and attributes the donor and recipient behaviours to the fused data set in a fashion which respects the respective weights of the donor and recipient. Since the standard transportation algorithm makes no accommodation for different levels of activity for different sites within the same individual, the standard approach could result in a fused data set which generally understates the visitor levels. Further, other volumetric measures like Page Views and Duration would be similarly inaccurate relative to the source data.

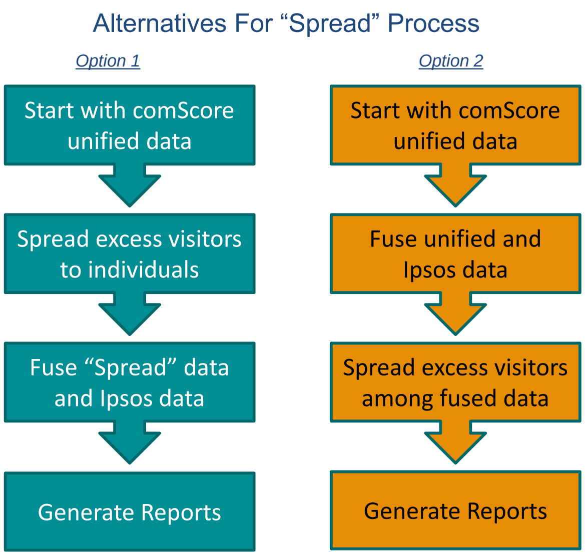
In general there are two solutions to this issue:

a) A scaling factor for each site, applied to a respondent's weight, resulting in a situation where each site visitation is treated as a volumetric measure

b) A modeled approach where the delta between the sum of weights for a collection of respondents and the unified audience figure is allocated to similar individuals in the comScore data set.

Early discussions with third-party processors indicated that the use of the first approach would be challenging since the systems generally could not use a volumetric measure in cross-tabs; further, the approach could lead to differences in audience calculations by different third-party processors (particularly in the processing of cumulative audience data) and may result in market inconsistencies. Consequently, we adopted the second, modeling-based approach which seeks to spread excess visitors in one individual to one or more other carefully selected respondents.

The model required consideration of two key design decisions. The first concerns when the modeling of the excess audiences is conducted – before or after the fusion – while the second concerns the process by which the excess visitor data is allocated.



Early experiments within the comScore fusion used the first approach (Option 1 in green boxes above) where excess individuals were allocated on a site by site basis, within critical cells and then the "spread" data was used in the fusion. However, post-fusion validation work showed that, while it was possible to achieve a high degree of concordance between the source unified data and the modelled "spread" data, the approach also tended to lose discrimination within the fusion, particularly with regard to income based measures. This occurred because the "spread" process sometimes resulted in excess visitors from a donor who was matched to a recipient in a higher income bracket being allocated to other donors who matched recipients in a lower income bracket. This effectively distributes the online behaviours that would have been within the upper income group to recipients in lower income groups, decreasing the discrimination that would have otherwise been preserved by the final matching process. While it is possible that the reverse could happen (individuals matched at a lower level "spread" to recipients in upper income groups), this trend is less likely on a probability basis since there are significantly more people in the lower income groups.

To overcome these issues we developed a process which saw the recipients stratified on various wealth indicators, fusing the Ipsos and the unified comScore data without specific regard to these strata then preferentially allocating the excess visitors in the donors to respondents linked to the same wealth strata.

Conceptual Process For "Spreading" Within Wealth Strata

Wealth Strata	Recipient
1	A
1	B
1	C
2	D
2	E
2	F

→

Wealth Strata	Fused
1	A2
1	A4
1	B1
1	C4
2	D3
2	E5
2	E6
2	F6

→

Donor	Wealth Strata
1	1
2	1
3	2
4	1
5	2
6	2

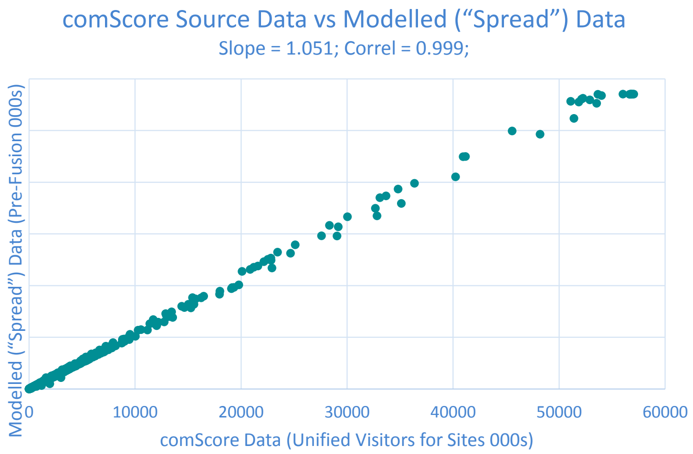
It is worth using a worked example to understand the process. In the diagram above we can imagine that recipients A, B & C (Ipsos respondents) belong to wealth strata #1 and the other recipients D, E & F belong to wealth strata #2. These strata are determined for recipients A to F using the richer data available within the Ipsos Affluent Survey (household income, net worth, liquid assets etc.). The recipients are matched with their comScore donors using the fusion process described earlier and in doing so, we identify that donors 1, 2 & 4 are matched with recipients in wealth strata #1 and that the other donors 3, 5 & 6 are matched to wealth strata #2. In this model, there is no occasion where a single donor (say 1) is matched to recipients who come from different strata (e.g. A & D). In such a case in the actual fusion, the donor is assigned to the higher of the two strata (i.e. to the wealthier strata).

For the sake of clarity, we will refer to the donor with the excess of visitors as the "ascriber" and the donor receiving the excess visitors as the "ascribee". Once all of the individuals in the donor file have been stratified, the "spreader" process attempts to allocate all excess visitors initially to ascribees within the ascriber's own wealth strata and then (if there are still excess visitors left) to ascribees in other wealth strata. The selection of an ascribee to be receive excess visitors from a specific ascriber is a function of:

a) The similarity of the ascribee to the ascriber (determined using a modified version of the fusion matching algorithm)
b) The extent to which as many sites can be transferred completely from the ascriber to the ascribee
c) The proximity of the ascribee's weight to the amount of visitors being transferred from the ascriber

The first criteria is essential to ensure that the attribution of websites remains, as far as possible, within groups identical to the demographic and behavioural profile of the ascriber. The second criteria tries to preserve the reach and cross-media behaviours of the ascribed sites. This initiative came out of an insight from the earlier "spreader" where we found it was possible for related sites (e.g. a media vehicle and its parent) to be ascribed to different individuals, potentially giving cumulative audience figures that were inconsistent. The final criteria is an attempt to minimize the number of ascribees an ascriber uses to allocate all of their excess visitors.

Having allocated all of the excess visitors, the other volumetric values (Page Views, Duration) are kept in alignment by applying the original ascriber's pages/visitor and duration/visitor ratios to the ascribee's visitor counts. Because of the granularity of the individual weights, it is not always possible for the modeled visitor data to match exactly the unified data, but the results in a very close match, with a correlation co-efficient of 0.999 on visitors (shown in graph below). The analysis of the modeled data also shows that the slope is about 5% greater than the diagonal because each site's modeled audience, calculated within critical cells, is scaled to the IAS universe estimate prior to fusion.



10) Data Validation

The main validation criteria used in assessing the quality of the fusions were the extent to which a fusion met the following criteria:

a) Full use of the Ipsos data set
b) The preservation of the comScore metric after fusion
c) The extent to which the relativities of the different websites with respect to a couple of key metrics, including the extent to which the websites after fusion reflected the expected rankings with respect to different measures of affluence

In addition, while not a specific validation criteria, we also sought to ensure that the fused data, as far as possible, reflected the cross media duplications evidence in the IAS data.

Each of these is explored below.

Full use of the Ipsos Data Set

The fused data set includes all of the demographics available in 2012 Ipsos Affluent Survey study combined with online data, giving a data set rich in its ability to profile the websites. The stakeholders reviewed a wide range of media titles, with a particular emphasis on comparing the reported behavior of the websites using various demographic and activity data with the expected profile of the title. For example, the following table shows the top 5 demographics by index for the New York Times.

NYTimes.Com	Index
Q20e. Household Income ($500k+)	168
Publications Read: Foreign Affairs	160
Publications Read: New York Magazine	156
Publications Read: Ivy League Network (net)	156
Publications Read: The New York Times (Sunday)	152

For ESPN, the indexes show results consistent with the media vehicle.

ESPN.COM	Index
Publications Read: ESPN The Magazine	155
Q18j. Age (30 to 34 years)	150
Cable Viewed: ESPNU	148
Q20e. Household Income ($500k+)	142
Q18a. Respondent Gender (Male)	141

For women's and lifestyle titles the stakeholders were comfortable that the indexes of behavior were generally in alignment with expectations Indeed, for the Martha Stewart web site, the readership of Martha Stewart Living magazine was the eighth most popular activity (while not shown in the table below it had an index of 146).

MARTHASTEWART.COM	Index
Cable Viewed: Ovation	187
Publications Read: Cooking with Paula Deen	161
Publications Read: Harper's Bazaar	156
Publications Read: Midwest Living	150
Publications Read: Family Circle	149

Even more appropriately, when the demographic data sets are restricted to a subset of the demographic (such as preferred leisure/entertainment options) titles also continue to reflect the sensibility of the media vehicle's objectives. The following demographic profile (top 5 opinion for a couple of Condé Nast titles) reflect this observation.

VOGUE.COM (Opinions)	Index
People often ask my advice on fashion and what they should wear	129
People often ask my advice when they are considering where to vacation	128
I enjoy keeping up with the latest fashions and trends	123
I tend to take the lead in decision-making	118
People often ask my advice [for] entertainment choices and leisure activities	105

WIRED.COM	Index
People often ask my advice when [buying] technology or electronics products	135
I like to offer advice to others	119
I am usually one of the first [to try] new products or services	115
People often ask my advice [regarding] making a significant purchase	114
I consider myself an opinion leader	114

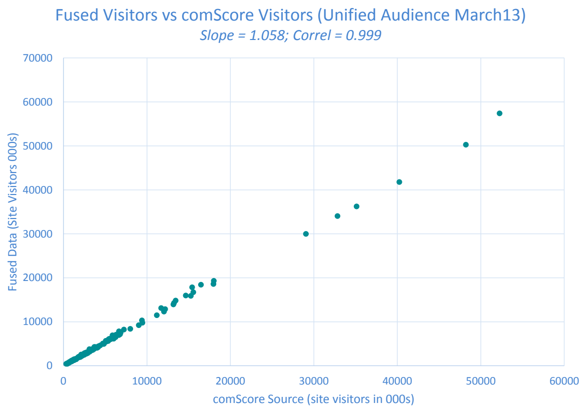
While a sampling of media vehicles does not constitute a comprehensive assessment of the fusion's ability to capture all of the IAS demographic detail and apply it to the fused online sites, it gave the stakeholder a satisfactory indication that the fused data will allow users full use of the Ipsos data set and more importantly provided top level data (indexes in this case) that were consistent with expectations.

Preservation of comScore Estimates Post-Fusion
In creating the fusion, there are a number of ways in which the fused data set could result in a website visitor count which is different to the source data:

a) The spreading of the unified data result in a value which is different
b) The fragmented data could then be scaled (up or down) due to differences in the universe estimate

The following chart compares the source data from the comScore unified audience estimates for March 2013 (the input data for the fusion) against the audience estimates for the same sites from the fused data set. While the correlation is very high (0.999), the slope of the line is 1.058. This indicates that on average, the fused data estimates are about 6% higher than the source data, a delta which is mostly attributable to the difference in universe estimates for the donor and recipient data sets (the 5% difference from the "Spread" modeling described in section 9). The remainder, which is less than approximately 0.7%, is due to a combination of factors, but mostly due to attribution of online behaviours from the donor sets to recipient individuals that have different weight ratios to the source data.

The chart on the following page demonstrates the concordance between the estimated site visitors after fusion and the source data from the Media Metrix service. The degree to which the fused data for website visitations matches the data from the source Media Metrix file (the high correlation and an understanding of why the differences exist between the absolute visitor levels) means that the stakeholders were happy to accept the beta fusion as having met the second criteria of preserving the comScore audience estimates.



Median Income Imputation Tests
As part of the fusion validation, we tracked specific metrics for 166 different websites for which actual and imputed wealth indicators could be established. Specific measures which were used in assessing the quality of the fusion included:

a) The extent to which the median household income of the fused website data matched the same metric for the same website as measured by recall data in the Ipsos survey
b) The extent to which the percentage of a website's audience that comes from homes earning over $250k/year in the fused data set is the same as the equivalent metric for the same website as measured by recall data in the Ipsos survey

In summary, for the 166 sites we obtained an average median income of $145,994 for the sites using fused data, which compared well to the $146,464 for the same metric using the Ipsos recall data for the same sites. In other measures, while fused median income was 99.7% of the median income from the source (recall) data, the Spearman rank correlations for the median income for the two data sets was 0.56 and the same measure for the percentage of a sites audience that came from homes earning $250k or more per year was 0.59.

While this is typically only a moderately strong concordance, it is worth re-iterating that the stakeholders acknowledged that the comparison between the two measures will never be exact. Apart from the fact that two surveys may differ by chance alone, the two surveys also attempting to measure the same behaviours (visits in the last thirty days to specific website) using different modes. In addition, we would generally expect to see median income, and other measures regress to the mean during any fusion and would expect this to become more pronounced with the expansion of reported audience (i.e. as more people are added to the list of visitors to the site, it becomes more difficult in general for the site's behavior to index higher than average). Consequently, while the correlation co-efficient of 0.59 might typically indicate a modestly strong correlation between the two data sets, the stakeholders recognized it as being acceptable for the reasons outlined above.

As part of the process for reviewing the extent to which the fused data matched the "wealth" profile of the source recipient data we also looked at specific websites.

This table below an extract of a more complete top 50 included in the Appendix and shows comparisons in the key metrics used to validate the fusion (eg comparisons of Median Household Income, $250k+ percentage, comparison of the fused and source audience data and information on the donor's sample size).

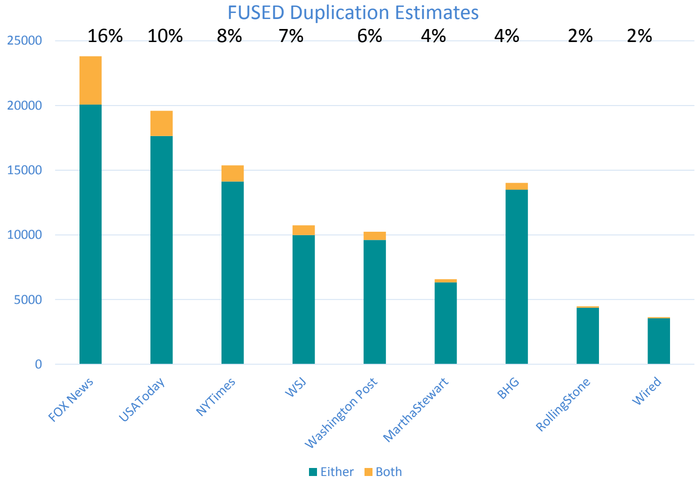
	Measures							
	Med. HHInc. ($)		% $250k+		Visitors (000s)		%Change Vis.	Donor Sample
Sites (Table 1)	IAS	FUS	IAS	FUS	COM	FUS	MAS->FUS	Size (n)
WSJ.com (The Wall Street Journal)	163,256	152,209	21%	18%	6,011	6,149	102%	565
FT.com (Financial Times)	159,542	148,797	26%	19%	665	677	102%	147
Bloomberg.com	158,377	158,077	20%	18%	4,020	4,051	101%	365
United.com (United Airlines)	157,827	156,217	20%	18%	2,909	2,990	103%	478
Businessweek.com	157,460	147,831	16%	14%	2,469	2,459	100%	235
CarandDriver.com	157,206	147,234	14%	13%	1,001	1,019	102%	132
Edmunds.com	156,296	148,919	15%	14%	2,040	2,048	100%	432
Sirius.com/XMRadio.com	155,938	144,786	15%	15%	705	738	105%	123
AutomobileMag.com	155,877	145,526	22%	16%	330	422	128%	162
MarketWatch.com	155,804	150,980	16%	16%	2,417	2,483	103%	239
Forbes.com	155,699	151,921	19%	16%	6,637	6,977	105%	890
LinkedIn.com	155,631	149,333	16%	15%	18,004	19,295	107%	2323
Finance.Yahoo.com	155,630	148,139	17%	15%	13,244	14,251	108%	1840
Delta.com (Delta Air Lines)	155,065	144,597	18%	13%	3,002	3,205	107%	485
LonelyPlanet.com	154,866	144,910	18%	13%	759	835	110%	103
Local.com	154,758	142,977	11%	12%	2,688	2,833	105%	718
CNBC.com	153,552	150,053	19%	15%	3,140	3,182	101%	320
Time.com	153,488	147,033	13%	14%	4,027	4,210	105%	534
FoxBusiness.com	153,237	147,082	16%	15%	1,528	1,589	104%	131
AA.com (American Airlines)	153,159	149,110	18%	16%	2,608	2,737	105%	436
NYTimes.com	152,207	151,896	18%	16%	12,174	12,870	106%	1758
SI.com (Sports Illustrated)	152,136	150,508	14%	16%	4,151	4,226	102%	489

In general, the proximity of the median household income calculations and upscale ($250k+/year) measures for the fused data to the "targets" established with the IAS recall data mean that the stakeholders believe the fused data has met the third criteria.

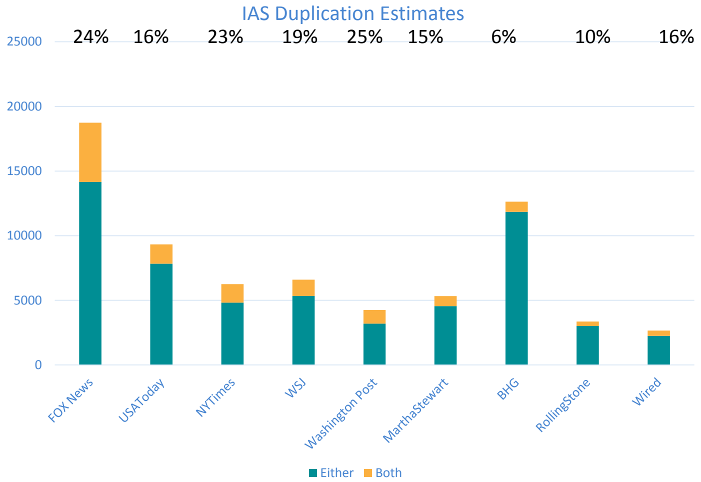
Media Vehicle Duplication

In addition to the test for comparisons between website profiles with respect to income and other measures of affluence, another validation available was to compare the extent to which the fused data set preserved the relative duplication between the main media vehicle and its online presence. This was possible, to some degree, since the Ipsos survey also includes an extensive array of websites for which recall behaviours are collected. While we did not expect the duplications to be preserved precisely (since, the size of the imputed audience after fusion was often substantially different to the Ipsos survey result), we started with the hypothesis that the fusion should preserve – as far as practical - the relativities between the duplications.

The following chart shows the duplication between a number of key media vehicles (both print and television) and the equivalent online properties. In the examples, we see that the broadcast and newspaper titles have the highest duplication (with duplication defined in this instance as the percentage of people who attend both the website and the primary media vehicle (print or television) as a percentage of the total audience to both platforms.



The same analysis using the recall based web data gives a similar but not quite identical set of duplications. Most of the differences have eventuate because the denominator has increased for most of the sites while the numerator has stayed relatively constant.



While the duplications are not identical, the trend is generally in the same direction and the stakeholders accept that the fusion has preserved, as far as possible, the duplications between the main media vehicle and the online presence.

11) Multi-Platform Duplication and Incremental Reach among Affluent Readers of Traditional Print

Of importance to all traditional publisher brands is the amount of duplication registered by readers of traditional print vehicles (printed magazines and newspapers) who also visit the print brand's associated website. Of equal importance is the amount of potential incremental reach afforded by adding a publisher brand's website-only visitors to the readership of its traditional print vehicle. Though there is availability in the marketplace for understanding duplication and potential incremental reach in the general US, The Ipsos Affluent Survey: USA/comScore Fusion offers for an understanding of duplication and reach among the US affluent audience, and among subsets thereof.

(Note: The following analysis is performed using The Ipsos Affluent Survey: USA/comScore Fusion dataset. See *Appendix #3* for a list of select print titles included in this analysis.)

Overview of Duplication and Incremental Reach

Among the US affluent, the average duplication of traditional print readers of magazine titles and visitation to their associated website (that is, the percent of 1+ print magazine title readers who also have visited the title's associated website in the past 30 days) is 7%. Thus one in fourteen readers of traditional print magazine titles has also visited the title's associated website. The potential incremental reach afforded by adding an associated website's visitation to the readership of a traditional print vehicle is on average 94%, almost doubling the potential reach for the magazine brand.

Thus for magazine brands, the duplication and incremental reach stories are compelling among the US affluent. A substantial percent of readers are consuming both traditional print and associated website content offering cross-marketing opportunities. And traditional magazine brands, are on average, nearly doubling their US affluent potential reach with their web-only consumers.

The story for traditional newspaper brands is even more compelling. Over one in four traditional print newspaper brand readers are also consuming the brand's content online, and the potential incremental reach of these brands is extended by over 200% by digital-only consumers of the brand's web content.

Total Print Duplication and Incremental Reach *(Readership/Web Visitation to 82 Reported Print Brands)*			
Publication Type	**Average Duplication***	**Average Incremental Reach****	**Titles Measured**
Total Print (Magazine and Newspaper)	10%	114%	82
Magazine Only	7%	94%	74
Newspaper Only	27%	223%	8

*Duplication = percent of 1+ print title readers who also have visited the title's associated website in the past 30 days. Average is weighted to traditional print readership.

**Incremental reach = Past 30-day Web-only visitors, as a percent of total traditional print (print-only and duplication) 1+ readership

Affluent Duplication and Incremental Reach by Genre

US affluent duplication and potential incremental reach perform differently for magazine titles across title genre. The Ipsos Affluent Survey: USA/comScore Fusion demonstrates that those brands operating in particularly time-sensitive genres tend to perform better on both duplication and potential incremental reach measures among the US affluent. For instance, magazine titles in the business and finance genre have the greatest duplication with nearly 2 in 5 readers of the traditional print content also consuming the brand's website content. It can be surmised that readers of traditional print business and finance titles are also looking to the print brand to provide time-sensitive information (e.g. stock quotes, loan rates, etc.) online.

Interestingly, this genre also has the largest increase in potential incremental reach, whereby the incremental reach of traditional print business and finance titles increases by 170% with the addition of web-only past 30-day visitors.

When newspaper titles are included in the business and finance genre with magazine titles, the duplication increases to 19% and potential incremental reach increases to 226%. Forbes leads the magazine brands in the business and finance genre with the highest duplication (20%) and extended reach (282%). Following Forbes closely in rankings are Bloomberg Businessweek and Bloomberg Markets which in particular has a very large extended reach (474%) afforded by its website. Two national newspapers rank higher than Forbes when newspaper brands are included as providers of business and finance content (New York Times and The Washington Post), while The Wall Street Journal ranks just below Forbes for duplication and lower on the rank for potential incremental reach.

Similarly, magazine titles in the time-sensitive sports genre have a very similar story for duplication (16%) and potential incremental reach (143%). Again, the time-sensitivity of sports content (scores/news) undoubtedly increases both the duplication and incremental reach for this genre, with readers going to the trusted brand's website for updates and reading the traditional print publication for in-depth sports analysis.

ESPN is the clear leader in the sports genre for duplication (36%) and potential incremental reach (341%). ESPN also has the highest affluent magazine total reach (traditional print and website) at just over 16 million. ESPN is followed by Sports Illustrated in rank on the three metrics above.

As one would expect, magazines titles in the news genre behave similarly to those in the business and finance, and sports genres. Those genres scoring lowest on rankings for duplication and potential incremental reach include automotive, fashion and beauty, political and commentary, shelter, and women's. Regional publications perform above average for potential incremental reach, as do fitness, health. In-Flight magazines perform admirably for both duplication and potential incremental reach among the affluent.

Print Brand Duplication and Incremental Reach *(Readership/Web Visitation to 82 Reported Print Brands)*			
Publication Genre	**Average Duplication**	**Average Incremental Reach**	**Titles Measured**
Automotive	2%	20%	4
Business, Finance	18%	170%	12
Cuisine	5%	38%	8
Entertainment	5%	44%	8
Fashion, Beauty	3%	17%	13
Fitness, Health	6%	112%	10
General Editorial	5%	57%	10
In-Flight	8%	82%	5
Lifestyle	5%	52%	7
Men's	11%	95%	13
News	12%	110%	12
Newspapers	27%	223%	8
Outdoor	6%	52%	3
Political and Commentary	4%	60%	6
Regional	6%	158%	4
Science, Technology	5%	82%	7
Shelter	4%	21%	6
Sports	16%	143%	6
Travel	5%	49%	14
Women's	3%	18%	18

Affluent Duplication and Incremental Reach for National Newspapers

Among national newspapers, The New York Times weekday and Sunday both rank highest for duplication (34%), while The Washington Post edges out The New York Times Sunday for potential incremental reach (310). USA today performs well on both duplication and incremental reach followed by The Wall Street Journal. Other financial newspaper brands perform lowest for both measures.

US Affluent Engagement with Online Media

While an understanding of reach and incremental reach among US affluent readers is of importance to US magazine and newspaper brands, of no less concern are the online behaviors of the US affluent consumer. To understand US affluent online engagement, we again turn to The Ipsos Affluent Survey: USA/comScore Fusion.

(Note: The following analysis is performed using The Ipsos Affluent Survey: USA/comScore Fusion dataset. See *Appendix #4* for a list of websites included in this analysis.)

The general comScore metrics of Time-Spent and Page Views are used to offer a measure of engagement with online media on the part of the US affluent population. And on average, the US affluent is engaged similarly with online media as compared to the general population. They spend slightly less time with online media overall than does the general population, and they visit slightly fewer pages. We understand that in general, the US affluent population is strapped for time, and their media choices are made for efficiency.

Online Engagement Metrics, Affluent Vs. General Population *(Visitation to 402 Select Websites)*	Affluent	General Population	Affluent Index
Average Monthly Time-Spent (Minutes)	51.4	53.1	96.7
Average Monthly Page Views	58.1	59.5	97.5

However, a measure of online engagement on the part of the US affluent is not average across all genres of online content, and there are topic areas with which this population spend more time and are, in general, more engaged than is the general population. One such genre of online content where US affluents spend the highest amount of time in comparison to the general population is in career service and development (comScore genre classifications).

The US affluent spends 23% more time with this category of website than does the general population. While visiting these websites, they view 22% more pages. Theladders.com and careerbuilder.com both wring 45% more time from affluent visitors than from general population visitors, and rack up 35% and 58% more page views respectively. Additionally, US affluents spend 24% more time and view 17% more pages on Linkedin.com, categorized as a social networking site by comScore, than does the general population.

Another website genre that performs particularly well for online engagement among the US affluent (and not as seemingly related directly to wealth creation as the previous genre) is entertainment. The US affluent population spends 21% more time with entertainment websites, and views 43% more pages in this category than does the general population. US affluents seek a wide variety of entertainment sources, and spend more time at billboard.com (14% more time), eonline.com (17%), nationalgeographic.com (26%), theonion.com (40%), vanityfair.com (16%), and VH1.com (16%) than does the general population. Many of these websites also see a higher number of page views by US affluents when compared to the general population, as do bbcamerica.com (21% more page views), comedycentral.com (12%), biography.com sites (21%), hbo.com (52%), and rollingstone.com (18%), and vimeo.com (30%).

And as it played an important part in an understanding of duplication and reach afforded by US affluents for traditional print brands, the sports category is another for which the affluent have a particular attraction demonstrated by high online engagement. Affluents spend 16% more time visiting sports websites and while there, view 12% more pages than does the general population. Leading affluent sports sites include si.com (Sports Illustrated – 63% more time, 37% more page views), foxsports.com (47% more time, 40% more page views), and espn.com (13% more time spent, 8% more page views) compared with the general population.

Average Monthly Time-Spent by Website Category: Affluent Vs. General Population (Visitation to 402 Select Websites)			
comScore Website Category	**Affluent**	**General Population**	**Affluent Index**
Auto	11.6	11.5	101
Business/Finance	14.9	14.2	105
Career Service and Development	13.5	11.0	123
Entertainment	80.2	66.5	121
Lifestyle	9.3	10.3	90
News & Information	23.5	22.3	106
Social Media	122.1	148.0	82
Sports	44.4	38.4	116
Technology	6.2	5.5	112
Travel	11.2	10.3	109

Average Monthly Page Views by Website Category: Affluent Vs. General Population (Visitation to 402 Select Websites)			
comScore Website Category	**Affluent**	**General Population**	**Affluent Index**
Auto	17.7	16.6	106
Business/Finance	23.1	22.3	104
Career Service and Development	20.9	17.1	122
Entertainment	74.8	52.2	143
Lifestyle	13.5	14.9	91
News & Information	19.1	17.9	107
Social Media	170.7	198.3	86
Sports	38.4	34.2	112
Technology	7.2	6.3	115
Travel	15.9	14.4	111

US Affluents and General Website Visitation

Despite the fact that the affluent spend less time on social media than does the general population, Facebook is the website with the highest time spent for this important online audience. Yahoo!, YouTube, Google, and REDDIT round out the top five websites for time spent. Websites falling within the top 20 include msn.com, eharmony.com, craigslist.org, and zynga.com.

More time on these sites translates to more page views as well, so many of these websites also sit within the top 20 for the highest number of largest number of average page views among the US affluent. Bucking this trend, several financial service and banking websites sit within the top 25 websites for US affluent page views, while not sitting among the top 25 for time spent. These include fidelity.com (#20 in rank for page views), bankofamerica.com (#21), pnc.com (#22).

Conclusions Drawn From the Fused Data

The US affluent consumer is a member of an important target for any number of advertisers and media content providers. A better understanding of this consumer's online experience will help content providers increase enjoyment of the online experience, and assist marketers in providing relevant, potent messaging.

With this understanding of the US affluents' multi-platform media consumption, the media and advertising communities can measure the total US affluent brand footprint across the web and other advertising channels (print as well as television), understand and quantify the level of engagement experienced by the US affluent audience across websites and website genres, and fundamentally promote the benefits to content providers and advertisers of reaching the US affluent across traditional media channels and associated online destinations.

12) Next Steps

The initial fusion of the 2012 Ipsos Affluent Survey and the March 2013 data set from comScore has provided a detailed and robust identification of the issues associated with developing a comprehensive data set which provides a rich insight into the online behaviours of affluent America.

This fusion was novel in the extent to which it sought to control for income and wealth factors in the fusion when the donor data set did not have extensive information on these characteristics. Since the stakeholders are satisfied that the fusion process can deliver the required data set, we are now looking to move on to the next stages of the service delivery, including the creation of a data base using the 2012 IAS study fused with the 2013Q3 Media Metrix data (April to June, 2013).

Once the Quarter 2 data set is with subscribers, there will be a new fusion conducted using the 2013 Ipsos Affluent Survey (due for release September 2013) with the 2013 quarter 3 data from comScore. We will use the next two fusions to revisit a number of areas where we believe the existing practices can be improved even further.

Fusion Improvement #1: Custom Matching Variables by Critical Cell
While we have demonstrated that the existing fusion, using a single set of matching variables, can deliver a data base which meets the criteria established at the outset (preserving the currency, duplicating the wealth parameters within the fused data set etc.), there is a belief that we might be able to improve the fusion by customizing the set of matching variables on a critical cell basis. Specifically, we are interested in testing whether we can improve the quality of the fusion by creating a different set of matching variables within each critical cell.

Fusion Improvement #2: The Best Wealth Metric
During the development of the fusion, the household median income became the primary metric against which the quality of the fusion was assessed with respect to wealth distribution. Unfortunately, the mechanics of calculating a median make it susceptible to noise and random events within the fusion process. When developing later fusions we believe it will be worth exploring if the quality of the fusion can be better assessed using simpler classification metrics like site composition based on homes earning $150k+/year and/or $250k+/year.

Fusion Improvement #3: The Best Distance Metric
We noted earlier that the initial fusion used simple distance metric, eschewing more complex measures such as the Mahalanobis distance metric for a number of reasons. These reasons included that there is no evidence (of which we are aware) demonstrating that the approach definitively improves the fusion quality; that the mathematics behind measures such as the Mahalanobis distance are more suited to applications involving continuous variables rather than the ordinal and nominal variables which dominate this fusion and that the method adds substantially to the time needed for each fusion.

However, we acknowledge that there is an intrinsic appeal in the Mahalanobis distance in that the metric accounts for co-variant behavior and that this may help improve the quality of the fusion. To that end we would eventually like to make a rigourous assessment of the merits of each approach to the distance metric given time and resources permit.

While none of these enhancements are essential to an operational fusion, establishing a definitive position on the merits or otherwise of each of these initiatives may improve the quality of the next set of fusions and provide valuable insight into areas of data integration that are rarely considered or reviewed.

THANK YOU
Thanks to the team at comScore, including Josh, Naresh & James for their patience and willing support; to Babette and the team at Condé Nast who have been both patient and enthusiastic supporters of the fusion initiative and to the Ipsos MediaCT team, including Mark, Donna and Maria for their assistance. Finally, thanks to Lorraine who started it all.

Appendix #1:
<u>List of 166 Websites Common Between Ipsos & comScore with comScore Sample Size > 100</u>

Sites Common to Comscore and Mendelsohn (Minimum Comscore Sample Size 100)							
	Sites		Sites		Sites		Sites
1	WSJ.com (The Wall Street Journal)	51	Southwest.com (Southwest Airlines)	101	EW.com (Entertainment Weekly)	151	UsMagazine.com
2	TheStreet.com	52	Economist.com	102	KBB.com (Kelley Blue Book)	152	AEtv.com
3	TheWeek.com	53	Spotify.com	103	ABC.com	153	Syfy.com
4	Fool.com (The Motley Fool)	54	Cnet.com	104	Etsy.com	154	DailyFinance.com
5	TheAtlantic.com	55	Cosmopolitan.com	105	Weather.com (The Weather Channel)	155	Hulu.com
6	FT.com (Financial Times)	56	WashingtonPost.com	106	MSN.com	156	Disney.com
7	Bloomberg.com	57	HowStuffWorks.com	107	Amazon.com	157	WomensHealthMag.com
8	United.com (United Airlines)	58	GolfChannel.com	108	EOnline.com	158	Nascar.com
9	Businessweek.com	59	MSNBC.com	109	FourSquare.com	159	BHG.com (Better Homes and Gardens)
10	CarandDriver.com	60	Zappos.com	110	TheKnot.com	160	Spike.com (TV channel)
11	Edmunds.com	61	MensHealth.com	111	Netflix.com	161	MarthaStewart.com
12	Sirius.com/XMRadio.com	62	Travelocity.com	112	HGTV.com	162	TLC.com (The Learning Channel)
13	AutomobileMag.com	63	USNews.com	113	MLB.com	163	AdultSwim.com
14	MarketWatch.com	64	IMDb.com	114	Nick.com (Nickelodeon)	164	PhotoBucket.com
15	Forbes.com	65	BusinessInsider.com	115	Dictionary.com	165	Univision.com
16	LinkedIn.com	66	ESPN.com	116	CBS.com	166	PlanetGreen.com (TV channel)
17	Finance.Yahoo.com	67	LivingSocial.com	117	Weatherbug.com		
18	Delta.com (Delta Air Lines)	68	Hotels.com	118	WebMD.com		
19	LonelyPlanet.com	69	Cars.com	119	PayPal.com		
20	Local.com	70	NationalGeographic.com	120	eBay.com		
21	Epicurious.com	71	Expedia.com	121	AutoTrader.com		
22	NYMag.com (New York Magazine)	72	NFL.com	122	YouTube.com		
23	CNBC.com	73	CitySearch.com	123	Google.com		
24	Time.com	74	Gawker.com	124	Salon.com		
25	FoxBusiness.com	75	Slate.com	125	WhitePages.com		
26	AA.com (American Airlines)	76	Ticketmaster.com	126	AOL.com		
27	Mint.com	77	Wired.com	127	MapQuest.com		
28	Smithsonian.com	78	FoxSports.com	128	All Internet		
29	USAToday.com	79	Fandango.com	129	AllRecipes.com		
30	NYTimes.com	80	Twitter.com	130	Answers.com		
31	SI.com (Sports Illustrated)	81	MotorTrend.com	131	CareerBuilder.com		
32	TripAdvisor.com	82	HotWire.com	132	Reddit.com		
33	NPR.org	83	CBSSports.com	133	Facebook.com		
34	HBO.com	84	Wikipedia.org	134	GameSpot.com		
35	VH1.com	85	AMCtv.com	135	People.com		
36	MyLifetime.com (Lifetime Television	86	Myspace.com	136	Yahoo.com		
37	NutritionData.com	87	Priceline.com	137	MyRecipes.com		
38	CNNMoney.com (Fortune/Money)	88	Flickr.com	138	FoodNetwork.com		
39	Buy.com	89	FOX.com	139	NewYorker.com		
40	BBC.com	90	iTunes.com	140	AARPmagazine.org		
41	RollingStone.com	91	FoxNews.com	141	Classmates.com		
42	Health.Discovery.com (TV channel)	92	TravelZoo.com	142	History.com		
43	TheOnion.com	93	Shape.com	143	DisneyChannel.com		
44	USAirways.com (US Airways)	94	ComedyCentral.com	144	Biography.com		
45	Match.com	95	Groupon.com	145	Ask.com		
46	NBA.com	96	ABCFamily.com	146	DiscoveryChannel.com (TV channel)		
47	CNN.com	97	CWtv.com	147	PetFinder.com		
48	Reuters.com	98	Monster.com	148	NBC.com		
49	Kayak.com	99	Pandora.com	149	MTV.com		
50	Orbitz.com	100	BarnesAndNoble.com	150	Oprah.com		

Appendix #2:
Top 50 Websites Ranked Descending MAS Median Income with Comparison to Fused Data

	Measures							
	Med. HHInc. ($)		% $250k+		Visitors (000s)		%Change Vis.	Donor Sample
Sites (Table 1)	IAS	FUS	IAS	FUS	COM	FUS	MAS->FUS	Size (n)
WSJ.com (The Wall Street Journal)	163,256	152,209	21%	18%	6,011	6,149	102%	565
TheStreet.com	163,188	150,407	22%	17%	2,895	2,876	99%	253
TheWeek.com	162,215	147,291	30%	15%	1,371	1,484	108%	192
Fool.com (The Motley Fool)	161,704	154,228	15%	20%	2,078	2,087	100%	331
TheAtlantic.com	161,503	150,576	16%	16%	2,706	2,700	100%	310
FT.com (Financial Times)	159,542	148,797	26%	19%	665	677	102%	147
Bloomberg.com	158,377	158,077	20%	18%	4,020	4,051	101%	365
United.com (United Airlines)	157,827	156,217	20%	18%	2,909	2,990	103%	478
Businessweek.com	157,460	147,831	16%	14%	2,469	2,459	100%	235
CarandDriver.com	157,206	147,234	14%	13%	1,001	1,019	102%	132
Edmunds.com	156,296	148,919	15%	14%	2,040	2,048	100%	432
Sirius.com/XMRadio.com	155,938	144,786	15%	15%	705	738	105%	123
AutomobileMag.com	155,877	145,526	22%	16%	330	422	128%	162
MarketWatch.com	155,804	150,980	16%	16%	2,417	2,483	103%	239
Forbes.com	155,699	151,921	19%	16%	6,637	6,977	105%	890
LinkedIn.com	155,631	149,333	16%	15%	18,004	19,295	107%	2323
Finance.Yahoo.com	155,630	148,139	17%	15%	13,244	14,251	108%	1840
Delta.com (Delta Air Lines)	155,065	144,597	18%	13%	3,002	3,205	107%	485
LonelyPlanet.com	154,866	144,910	18%	13%	759	835	110%	103
Local.com	154,758	142,977	11%	12%	2,688	2,833	105%	718
Epicurious.com	154,008	144,756	16%	12%	1,827	2,017	110%	186
NYMag.com (New York Magazine)	153,562	147,994	13%	17%	1,524	1,465	96%	151
CNBC.com	153,552	150,053	19%	15%	3,140	3,182	101%	320
Time.com	153,488	147,033	13%	14%	4,027	4,210	105%	534
FoxBusiness.com	153,237	147,082	16%	15%	1,528	1,589	104%	131
AA.com (American Airlines)	153,159	149,110	18%	16%	2,608	2,737	105%	436
Mint.com	152,927	148,787	12%	18%	450	426	95%	109
Smithsonian.com	152,643	146,208	14%	13%	1,038	1,186	114%	165
USAToday.com	152,622	156,477	16%	17%	3,134	3,294	105%	294
NYTimes.com	152,207	151,896	18%	16%	12,174	12,870	106%	1758
SI.com (Sports Illustrated)	152,136	150,508	14%	16%	4,151	4,226	102%	489
TripAdvisor.com	152,015	145,482	17%	14%	6,781	7,155	106%	1254
NPR.org	151,930	145,697	15%	15%	2,227	2,280	102%	443
HBO.com	151,905	148,199	16%	13%	1,361	1,374	101%	370
VH1.com	151,833	141,411	14%	9%	516	569	110%	109
MyLifetime.com (Lifetime Television)	151,779	138,775	15%	12%	516	581	113%	112
NutritionData.com	151,768	143,113	14%	10%	813	904	111%	102
CNNMoney.com (Fortune/Money)	151,761	156,319	14%	19%	4,809	4,953	103%	508
Buy.com	151,498	146,940	11%	14%	926	961	104%	201
BBC.com	151,478	147,628	16%	14%	5,785	6,133	106%	780
RollingStone.com	151,363	152,340	12%	14%	1,281	1,411	110%	225
Health.Discovery.com (TV channel)	151,220	149,155	14%	14%	718	823	115%	116
TheOnion.com	151,206	147,967	15%	15%	1,929	1,995	103%	306
USAirways.com (US Airways)	151,122	145,277	17%	13%	1,546	1,639	106%	231
Match.com	151,093	147,700	16%	16%	1,196	1,237	103%	239
NBA.com	150,669	146,293	12%	13%	2,498	2,804	112%	722
CNN.com	150,588	147,426	14%	14%	15,257	15,880	104%	1843
Reuters.com	150,235	149,993	16%	17%	4,848	5,007	103%	556
Kayak.com	149,948	146,323	15%	15%	3,296	3,445	104%	610
Orbitz.com	149,743	148,591	14%	16%	3,646	3,795	104%	665

Appendix #3

Traditional Print Brand with Associated Website Reported in The Ipsos Affluent Survey: USA/comScore Fusion for use in Duplication and Incremental Reach Analysis

Magazines with Corresponding Web Destination
AARP The Magazine
Allure
American Way (American Airlines)
The Atlantic
Automobile Magazine
Barron's
Better Homes and Gardens
Bloomberg Businessweek
Bloomberg Markets
Bon Appetit
Boston Magazine
Car and Driver
Cooking Light
Cosmopolitan
The Economist
Elle
Elle Decor
Entertainment Weekly
Entrepreneur
ESPN The Magazine
Esquire
Everyday Food
Fast Company
Financial Times
Food & Wine
Food Network Magazine
Forbes
Foreign Policy
Fortune
Glamour
Golf Digest
Golf Magazine
Good Housekeeping
GQ Gentlemen's Quarterly
Harvard Business Review
Health Magazine
Hemispheres (United and Continental Airlines)
Inc.
Investor's Business Daily
Kiplinger's Personal Finance
Marie Claire
Martha Stewart Living
Men's Health
Money
Motor Trend
National Geographic
National Geographic Traveler
New York Magazine
The New York Times (Sunday)
The New York Times (weekdays)
The New Yorker
Newsweek
O The Oprah Magazine
Parenting
Parents
People
Popular Mechanics
Popular Science
Real Simple
Redbook
Rolling Stone
Runner's World
Scientific American
Self
Shape
Sky (Delta Air Lines)
Smithsonian
Southern Living
Southwest Airlines Spirit (Southwest Airlines)
Sports Illustrated
This Old House
Time
US Airways Magazine (US Airways)
Us Weekly
USA Today
Vanity Fair
The Wall Street Journal
The Washington Post
WebMD The Magazine
The Week
Wired
Women's Health

Appendix #4

Select Websites for Affluent Vs. General Population Analysis

Select Websites
4SHARED.COM
AA.COM
AARP.ORG
ABC.COM
ABOUT.COM
ACCUWEATHER.COM
ACTIVE.COM
ADOBE.COM
AETV.COM
ALIBABA.COM
ALLRECIPES.COM
ALLURE.COM
AMAZON.COM*
AMC Sites
AMCTV.COM
AMERICANEXPRESS.COM
Ancestry Sites
ANSWERS.COM
Apple.com Worldwide Sites
ASK.COM Sites
ATT.COM
ATT.NET
AUTOBLOG.COM
AUTOMOBILEMAG.COM
AUTOTRADER.COM
AVG.COM
AZLYRICS.COM
BABYCENTER.COM
BANKOFAMERICA.COM
BANKRATE.COM
Barnes & Noble.com Sites
BARRONS.COM
BBCAMERICA.COM
BEDBATHANDBEYOND.COM
BESTBUY.COM
BET.COM
BHG.COM
BILLBOARD.COM
BING.COM
Biography.com Sites
BIZJOURNALS.COM
BLOGGER.COM
Bloomberg.com Sites
BLUEFLY.COM
Bodybuilding.com Sites
BONAPPETIT.COM
Boston Globe Media Sites
BRASSRING.COM
BRAVOTV.COM
BREAK.COM
BUSINESSINSIDER.COM
Businessweek.com Sites
CAFEMOM.COM
CAPITALONE.COM
CARANDDRIVER.COM
CAREERBUILDER.COM
CARS.COM
CBS.COM
CBSSPORTS.COM
CHEAPOAIR.COM
CHEGG.COM
CHOW.COM
Citigroup Banking Sites
Citigroup Portal Sites
CITRIXONLINE.COM
CITYSEARCH.COM
CLASSESUSA.COM
Classmates.com Sites
CNBC.COM
CNET.COM
CNN.COM
CNNMONEY.COM
COLLEGEBOARD.COM
COLLEGEHUMOR.COM
Comcast.com Sites
COMEDYCENTRAL.COM
COOKINGCHANNELTV.COM
COOKINGLIGHT.COM
COOKS.COM
COOLMATH-GAMES.COM
COSMOPOLITAN.COM
COSTCO.COM
COUPONS.COM
CRAIGSLIST.ORG
CVS.COM
CWTV.COM
DAILYFINANCE.COM
DELL.COM
DELTA.COM
DICTIONARY.COM
DIRECTV.COM
DISCOVERCARD.COM
DIYNETWORK.COM
DOMINOS.COM
DRUDGEREPORT.COM
DRUGSTORE.COM
eBay Sites
ECONOMIST.COM
EDMUNDS.COM
EHARMONY.COM
ELLE.COM
ENTREPRENEUR.COM
EONLINE.COM
ESPN City Sites
ESPN.COM
ESQUIRE.COM
ETSY.COM
EVERYDAYHEALTH.COM
EVITE.COM
EW.COM
EXAMINER.COM
Expedia Sites

FACEBOOK.COM
FAMILY.COM
FANPOP.COM
FASTCOMPANY.COM
FEDEX.COM
FIDELITY.COM
FLICKR.COM
FLYERTALK.COM
FOOD.COM
FOODANDWINE.COM
FOODNETWORK.COM
FOOL.COM
FORBES.COM
FORCE.COM
FOREIGNPOLICY.COM
FOURSQUARE.COM
FOX.COM
FOXBUSINESS.COM
FOXNEWS.COM
FOXSPORTS.COM
FT.COM
G4TV.COM
GAMESTOP.COM
GAWKER.COM
GEICO.COM
GILT.COM
GIZMODO.COM
Glamour Sites
GODADDY.COM
Golf Digest Sites
GOODHOUSEKEEPING.COM
GOODREADS.COM
GOOGLE.COM
GQ.COM
GROUPON.COM
HALLMARK.COM
HBO.COM
HBR.ORG
HEALTH.COM
HEALTHGRADES.COM
HGTV.COM
HILTON.COM
History Channel Sites
HISTORY.COM
HOLLYWOODREPORTER.COM
HOMEDEPOT.COM
HOTELS.COM Sites
Hotwire.com Sites
HOWSTUFFWORKS.COM
HP.COM
HRBLOCK.COM
HUFFINGTONPOST.COM
IBTIMES.COM
IGN.COM
IHEART.COM
IKEA.COM
IMDB.COM
IMGUR.COM
IMVU.COM
INBOX.COM
INC.COM
INDEED.COM
INSTAGRAM.COM
INSTRUCTABLES.COM
INTELIUS.COM
INTUIT.COM
INVESTOPEDIA.COM
INVESTORS.COM
IO9.COM
IVILLAGE.COM
JAVA.COM
JCPENNEY.COM
JETBLUE.COM
JEZEBEL.COM
KAYAK.COM
KBB.COM
KIPLINGER.COM
KMART.COM
KOHLS.COM
LIFEHACKER.COM
LIFESCRIPT.COM
LINKEDIN.COM
LIVINGSOCIAL.COM
LOCAL.COM
LONELYPLANET.COM
MACRUMORS.COM
MACYS.COM
MAPQUEST.COM
MARIECLAIRE.COM
MARKETWATCH.COM
MARRIOTT.COM
MARTHASTEWART.COM
MATCH.COM
MCAFEE.COM
MEDICINENET.COM
MENSHEALTH.COM
MENUPAGES.COM
MERCHANTCIRCLE.COM
Meredith Food Sites
MICROSOFT.COM
MINT.COM
MLB.COM
MLIVE.COM
MNN.COM
MONSTER.COM
MORNINGSTAR.COM
MOTORTREND.COM
MOZILLA.ORG
MSN.COM
MTV.COM
MYLIFE.COM
MYLIFETIME.COM
MYSPACE.COM
MYWEBSEARCH.COM
NASCAR.COM
NATIONALGEOGRAPHIC.COM
NBC.COM
NBCSPORTS.COM
NETFLIX.COM
NEWEGG.COM
NEWSMAX.COM
NEWYORKER.COM
NFL.COM

NHL.COM
Nick.com Sites
NIH.GOV
NIKE.COM
NJ.COM Sites
NORDSTROM.COM
NORTON.COM
NPR.ORG
NYMAG.COM
NYTIMES.COM
OFFICEDEPOT.COM
OLDNAVY.COM
OPRAH.COM
ORBITZ.COM
ORIGIN.COM
OVERSTOCK.COM
Oxygen.com Sites
PANDORA.COM
PARENTS.COM
PAYPAL.COM
PBS.ORG
PCH.COM
PEOPLE.COM
PETFINDER.COM
PGATOUR.COM
PHILLY.COM
PIZZAHUT.COM
PNC.COM
POGO.COM
POPSCI.COM
POPULARMECHANICS.COM
PRICELINE.COM
Pronto Sites
QVC Sites
RANKINGSANDREVIEWS.COM
Real.com Media Sites
REALSIMPLE.COM
REALTOR.COM
REDBOOKMAG.COM
REDBOX.COM
REDDIT.COM

REFERENCE.COM
RETAILMENOT.COM
REUTERS.COM
ROADRUNNER.COM
ROLLINGSTONE.COM
RUELALA.COM
RUNNERSWORLD.COM
SALON.COM
SAMSCLUB.COM
SAMSUNG.COM
SBNATION.COM
SCIENTIFICAMERICAN.COM
SEARS.COM
SELF.COM
SENDORI.COM
SHAPE.COM
SHEKNOWS.COM
SHO.COM
Shopping.com Sites
SHOPZILLA.COM
SHUTTERFLY.COM
SI.COM
SIMPLYHIRED.COM
SIRIUSXM.COM
SKYPE.COM
SLATE.COM
SLIDESHARE.NET
SMARTMONEY.COM
SOFTONIC.COM
SOURCEFORGE.NET
SOUTHERNLIVING.COM
SOUTHWEST.COM
SPARKNOTES.COM
Sports Illustrated Sites
SPOTIFY.COM
SPRINT.COM
STAPLES.COM
STEAMCOMMUNITY.COM
STEAMPOWERED.COM
STUBHUB.COM
SUPERPAGES.COM

Syfy.com Sites
TARGET.COM
TBS.COM
TELEGRAPH.CO.UK
TELEMUNDO.COM
THEATLANTIC.COM
THEBLAZE.COM
THECHIVE.COM
THEDAILYBEAST.COM
THEFIND.COM
THEKNOT.COM
THELADDERS.COM
THEONION.COM
THESAURUS.COM
TheStreet Sites
THESTREET.COM
THEVERGE.COM
THEWEEK.COM
THISOLDHOUSE.COM
Ticketmaster Sites
TICKETMASTER.COM
TIGERDIRECT.COM
TIME.COM
T-MOBILE.COM
TOPIX.COM
TOSHIBA.COM
TOYOTA.COM
TOYSRUS.COM
TRAVELCHANNEL.COM
TRAVELOCITY.COM
TRAVELZOO.COM
Tripadvisor Sites
TRULIA.COM
TRUTV.COM
TURBOTAX.COM
TVGUIDE.COM
TWITTER.COM
TYPEPAD.COM
UNITED.COM
UNIVISION.COM
UPS.COM

URBANSPOON.COM	VICTORIASSECRET.COM	WIRED.COM
USAIRWAYS.COM	VIMEO.COM	WOMENSHEALTHMAG.COM
USANETWORK.COM Sites	VISTAPRINT.COM	WORLDSTARHIPHOP.COM
USATODAY Sites	VITALS.COM	XFINITY.COM
USATODAY.COM	WALGREENS.COM	Yahoo! Sites
USBANK.COM	WALMART.COM	YAHOO.COM
USMAGAZINE.COM	WASHINGTONPOST.COM	YOUTUBE.COM
USNEWS.COM	WASHINGTONTIMES.COM	YP.COM Sites
USPS.COM	WEATHER.COM	YPEEK.COM
USTREAM.TV	WEATHERBUG.COM	ZAP2IT.COM
VANITYFAIR.COM	WEBEX.COM	Zappos Sites
VERIZON.COM	WEBMD.COM	ZAZZLE.COM
VERIZONWIRELESS.COM	WEBS.COM	ZDNet Websites
VH1.COM	WHITEPAGES.COM	ZYNGA.COM
VICE.COM	WIKIPEDIA.ORG	

Bibliography

Baker, K., Harris, P. and O'Brien, J. (1989) **"Data fusion: an appraisal and experimental evaluation."** Journal of the Market Research Society, 31(2), 153-212.

Collins, J. and Doe, P. (2011) "**Making best use of brand target audiences",** Print & Digital Research Forum, San Francisco

Jephcott, J. and Bock, T. (1998) "**The application and validation of data fusion."** Journal of the Market Research Society, 40(3), 185-205.

Wilcox, S and Page, K. (2012) "**Developing a Cross Platform Audience Currency For Great Britain**", www.rsmb.co.uk (accessed July 2013)