ORIGIN VID MODEL PROOF OF CONCEPT

Stage 2: Integration of TV data with Online

A report prepared for:

ISBA

December 2021

RSMB Limited, Savoy Hill House, 7-10 Savoy Hill, London WC2R 0BU
Tel: + 44 (0)20 7808 3600 Email: contact@rsmb.co.uk Web: www.rsmb.co.uk

Page

Table of Contents

1. Introduction

Origin is a cross industry initiative led by ISBA on behalf of UK advertisers to establish a cross media measurement service for campaign planning, reporting and evaluation, and aligned to the WFA Framework for cross media measurement.

As part of this initiative, Origin seeks to test and validate key components proposed as part of a technical blueprint that can provide for advertisers needs. One of these key components is the Virtual ID Model.

RSMB had previously completed a comprehensive methodology review and concept assessment concluding that the approach had potential merit and a recommended proof of concept (POC) be established to test the Virtual ID Model including how measured panel data for Linear TV advertising data could be integrated into the framework. This universe is defined by the broadcast TV commercial logs used as input to this POC. In particular, it excludes non-linear BVOD and addressable advertising. Return path data and BVOD census data are not a feature of this currency.

RSMB were commissioned by ISBA to undertake this Proof of Concept. The proof of concept consisted of two stages:

Stage 1 Integration of TV data

The first stage was to evaluate how data from a television panel could be adapted into a Virtual ID model and how well this would retain results seen in the original data. Given the hypothesis that channels could be taken as surrogates for websites, another important corollary outcome from this work was whether the VID model could work for online. Campaigns are restricted to spots on broadcast TV.

Stage 2 Integration of TV data with Online

The second stage is to evaluate if the Virtual ID model could be successfully extended across TV and Online data. The work would compare various model scenarios to understand if any could retain cross media campaign relationships.

Each stage is based upon a different snapshot of single source panel data provided by Ipsos MORI.

This report relates to work undertaken and completed for Stage 2, although in addition to this the model parameters (relative rates of advertising exposure) estimated from Stage 1 were applied to a new dataset in order to assess if rates sourced from one period are still applicable to another.

2. Executive Summary

The Proof of Concept (POC) was to test the Virtual ID component of the WFA technical blueprint for cross media measurement.

The POC used single source data from Ipsos MORI's Compass panel. Panel data was treated as census data in the design of the experiment in order to evaluate the model against 'true' census data.

The work was undertaken in two stages. This report relates to the second stage but it is worth recapping the findings from Stage 1:

The first stage was to evaluate how well data from a TV panel can be integrated into this VID framework whilst preserving TV campaign results. A secondary outcome from this TV data was to give an indication whether the model would potentially work for online (the focus for Stage 2) given the relationship between channels across campaigns may be analogous to relationships between websites. [The TV campaign data related to linear campaigns only; return path data, addressable advertising and BVOD were not part of this evaluation]. For the first stage, only data for television linear campaigns was used. The conclusion from this first stage was:

- The allocation model is able to address marked duplications seen across television campaigns producing acceptable results for key components (e.g. overall reach build, demographic profile, channel's unique contributions). The model closest to the WFA specification performed best but the other scenarios performed satisfactorily and therefore were continued to be assessed for Stage 2 as they might provide better cross media estimates when online campaigns are introduced into the assessment.

- The standalone results for television only campaigns have produced encouraging signs that the approach may preserve duplications between TV channels and websites for joint media campaigns given similar interdependent relationships between TV channels may also exist between TV channels and websites.

For the second stage, online campaign viewing data was assessed in conjunction with the TV campaigns. For each campaign the data only related to impressions from a single website.

To recap, the WFA proposed VID model allocates website impressions to a virtual respondent (VID) based on a probabilistic rate and potentially a website cookie (or equivalent personal identifier). For the second stage of the POC, 6 different scenarios were evaluated based on 2 types of rates of advertising exposure and using alternative 'cookie' (i.e. identifier) scenarios.

The two types of rates were as follows:

Abstract	this follows WFA framework and based on theoretical rates designed to give the best fit when training the model using aggregate campaign results from a single source panel.
Real	calculated from relative propensity to view advertising for each individual based on panel respondent level data from the single source panel.

For the cookie scenarios, the following were used:

'Total TV' cookie	For TV the panel ID information was preserved *across all impressions across all channels* – this was a new component to be evaluated in Stage 2. Once an impression was allocated to a VID all subsequent impressions for the panel member were allocated to the same VID. For digital, cookie information was used in the same way noting that panel members could have multiple cookies and therefore VIDs so preservation was only at the cookie level rather than the panel member level.
'Channel' cookie	For TV the panel ID information was only preserved *within channels*. So, once an impression for a panel member was allocated to a VID all subsequent impressions for the panel member *for that channel* were allocated to the same VID. For digital/online, cookie information was allocated as for the 'Total TV' scenario.

'No cookie Information from panel IDs or website cookies was not used so impressions from the same ID/cookie could be allocated to different VIDs.

In total 88 mixed campaigns were able to be evaluated. Although there were a relatively small number of campaigns, this was considered adequate for evaluating the performance of the scenarios as they:

- had a broad range of reach for the TV component, the Online component and permutations of both
- had schedules with similar GRPS that had varying levels of reach
- had a higher than expected number of schedules that exhibited a non-random relationship between tv and online components

For total campaign reach all models that used some form of cookie performed well. The two cookie models with abstract rates and the model with a total tv 'cookie' and real rates returned results closest to the benchmark 'census' data.

Allocation was undertaken within gender and age demographics and this resulted in good retention of demographic profiles for campaigns from the modelled data.

Week to week build of campaigns was largely preserved for the best scenarios and showed a similar degree of differences for each modelling scenario as was seen across the full period of the campaign.

As noted above the campaigns exhibited non-random overall reach given the television and online components (i.e. reach between media was not independent). Regression to the mean (RTM) measurements were used to evaluate how much of the non-random duplication between TV and online was lost by the VID model. By design, this is a stern test of any probabilistic data integration model and perfection is by no means achievable. In our opinion, the RTM results are acceptable and show that the principles and performance of the VID model are sound in statistical terms. The abstract models perform best. Within this there is a small but noticeable advantage for the TV channel by channel cookies compared to the Total TV cookie, however the trade-off is a small but noticeable distortion in the TV reach currency. Correction of this distortion is likely within the range of a post-analysis calibration tool, but this does increase the complexity of the analysis delivery system.

In summary, the Proof of Concept for Stage 2 has determined that the allocation model performs well under suitable conditions. These conditions are as follows:

- When integrating TAM panel campaign data into the VID model the use of the panel ID is essential. In the interests of currency preservation, the ideal is to carry across the single source campaign data for individuals ("Total cookie" scenario), effectively expanding the published data into the VID framework. An alternative is to only use this panel ID for impressions across the same channel. This is slightly weaker for currency preservation but the trade-off is slightly better estimates of duplications between television and online.

- The use of a "cookie" or other identifier for online is also essential as without this reach is overestimated and incremental reach is poorly estimated.

- Demographic labels are also required in order to be able to apply the model at a demographic level to produce suitable reach estimates.

- A high-quality single source panel is required for model training.

- The VID rates of advertising exposure are ideally calculated at an abstract level following the recommended method in the WFA Framework. This is essentially creating rates for groups following training on aggregated campaign reach results. It should be noted that this was a small scale test and it is unclear how this methodology might hold up when faced with a sterner test of increased media channels. A sophisticated optimisation routine is required for model training. Creating real rates based on individual panel members' propensity for advertising exposure on that channel/website is simpler but appears to be a weaker alternative and may need further modification/evaluation before it can be considered as a viable simpler method. In this respect it should be noted that there is potential, within the VID framework, to fine tune the model components.

Under the above conditions the allocation model is able to preserve campaign reach well overall, as schedules build, by demographic, and for linear television and online components of the schedule. In addition, regression to the mean (RTM) evaluations of TV/online duplications indicate that the best model overall has levels of RTM that are as low in our experience as other established data integration models, whilst exactly preserving the TV reach currency. Alternative scenarios can reduce this RTM but the penalty is a small distortion in the TV reach currency. The most appropriate trade-off has to be set against the expected use cases.

In RSMB's opinion the proof of concept demonstrates that, as statistical models go, the VID framework works well and retains acceptable accuracy whilst providing a practical and usable system.

3. Methodology

The framework set out by the WFA has a multitude of components in the cross media solution. However, at the heart of the solution is the Virtual ID model. This is the focus of the Proof of Concept.

The Virtual ID model is straightforward in nature. It can broadly be described as follows:

- A ‘respondent’ ID listing is created equal to the Census population.
- Gender and Age Group classifications are assigned to these based on universes sizes.
- A homogeneous group indicator is assigned to these indicating groups of similar behaviour for channels in the media campaign.
- Relative rates of advertising exposure to each channel/website are assigned to these homogeneous groups.
- Given Census campaign impressions, these are assigned in a probabilistic way using an allocation algorithm.
- Cookie information collected for the Census impressions may also be used in this allocation process.

The inputs for this process are acquired by training the model using Single Source Panel (SSP) data.

The framework allows and invites modifications to the process for example, in terms of how these rates and groups are created and whether cookie usage is beneficial.

4. Dataset

In order to conduct this evaluation a Single Source Panel dataset is required that is able to measure campaign data for both Television and Online. The Compass panel run by Ipsos MORI was used for this process. The panel consists of around 3,000 respondents aged 18+. Data is measured across TV, Radio and Online. For Stage 1, only data for the TV element was used. Online data was introduced for Stage 2.

The data used in Stage 1 was for 4 w/e 31st January 2021. For Stage 2 the TV data period was 4 w/e 12th April 2021, the online period was 8 w/e 10th May 2021. The online period was then proportionally reallocated into the 4 weeks of TV data, respondents that did not report in the 4 week TV period had their online data removed from the dataset.

For this Proof of Concept, for the sake of clarity in diagnostic evaluation, the TV channels were restricted to 6 channels. For each channel, respondent level data comprised impressions for every campaign and every spot transmitted during the period. For online, for each campaign, impressions were restricted to those on a single website in that campaign. It was accepted that some of these would constitute only partial campaigns but this data was still useful for training the model and calculating propensity to view for panel members. In the actual evaluation of the models only campaigns that consisted of an appreciable proportion of the original campaign were used.

The demographics used in the test were as follows:

Men 18-34	Women 18-34
Men 35-54	Women 35-54
Men 55+	Women 55+

The sample was restricted to a consistent cohort of Adults who had reported for at least 21 days out of the 28 days.

Online campaigns are sourced as detailed previously, with cookies based on cookie distribution data provided by Ipsos MORI. On average, each panel member (respondent) generates 1.13 cookies, although some panellists generate significantly more than this. Each website impression is assigned a campaign ID, with 88 of these campaigns linked to a TV campaign (either by a direct match or with a suitable surrogate). All of these 88 campaigns recorded at least 50 online impressions within the panel; a further 78 online campaigns recorded at least 50 impressions but no TV campaign could be linked to them.

For reference and to put any modelling differences into perspective, a campaign reach of 50% would have the following sampling errors:

- All Adults Sampling error = 0.8 95% Confidence Interval = (48.4,51.6)
- Demographic Sampling error = 2.2 95% Confidence Interval = (45.6,54.4)

5. Design of Experiment

Stage 1 of the Proof of Concept related to the integration of television data into the VID model.

For this stage, the evaluation was mainly concerned with how TAM panel data could be transformed into the Virtual ID framework, using elements of the methodology that was set out, while still retaining the general viewing metrics of the original source. In order to evaluate this, the panel data was used in two different ways to create the component input datasets required by the model. In the first instance, campaign reach data was generated at an aggregate level and respondent level; the former to generate abstract rates and the latter 'real'. Secondly, the panel data was treated as though it was the census population, generating both reach and impressions. The allocation algorithms could then be used to assign these impressions to this Census population using the VID model. The reach data could then be compared back to the 'real' Census benchmark.

Stage 2 built on this evaluation by introducing online campaign data; this was treated in largely the same way as the TV channels however in addition cookie deletion behaviour had to be allowed for in the VID model.

There were two key methodological considerations in the allocation process: the determination of VID relative rates of exposure and the use of a 'cookie' (to help control the reach currency), which for TV panel data was the respondent ID.

5.1 Individual rates

For each VID, the model parameters comprise a rate of exposure relative to the population average, for each channel and website.

For this evaluation two types of rates were used: Abstract and Real rates.

Abstract rates

These followed the protocols set out in the WFA framework (see Appendix A). Essentially these were designed to be based (or rather trained) on the panel campaign aggregate results rather than using any personal panel rates per se. So essentially an optimisation routine was undertaken to segment the VID population into (so called 'Dirac') groups with equal rates such that the expected results from the allocation process would be as close as possible on average to campaign data. The model was optimised with respect to the following reach components of the campaigns:

- Total campaign
- Demographics (Gender by Age Group)
- Weekly build
- Pairwise Channels

It is accepted that this level of training was achievable for this experiment but may be more challenging for a larger scale application.

Real rates

The real rates used were much simpler. Here each respondent was effectively in their own unique ('Dirac') group and generated a unique rate based on their propensity to view advertising on that media channel across the campaigns.

Note that in both cases, the VID model is guaranteed to replicate the impressions currency for each demographic group.

5.2 'Cookie'/Identifier

In the WFA framework, for Online the recommendation for the algorithm was to use the cookie link in the allocation process rather than treat the data as individual separate impressions. For TV data there is no such thing as a 'Cookie' but analogously the unique panel identifier can be used, principally to retain observed campaign frequencies by individuals. For the end-to-end project, 3 options related to cookies are considered:

Cookies across Channels ("Total TV")

Here is where each television impression uses the respondent ID as the "cookie" link across all campaign spots. This option by definition preserves the campaign data seen for the panel so was redundant for Stage 1 as it will match exactly. However, it is an important option in Stage 2 as online cookies will be allocated separately from TV cookies (in the same way as the "Cookies within Channels" option) and so the cross platform reach between TV and online will not necessarily be preserved. [It should also be noted that even for 'TV preservation', in the real world there may be some modelling adjustment to the TAM data that is not (or cannot be) accounted for so this may not *exactly* preserve but it will be very close].

Cookies within Channels

Here, each television impression uses the respondent ID as the "cookie" link for campaign spots within the same channel. So, for example if two spots were watched on a channel by a single respondent, the Virtual ID respondents that were allocated the first impression would also have the second. For the online component each cookie for a campaign is allocated one by one, with all campaign impressions associated with that cookie assigned to the same Virtual ID.

No Cookie

The final option is where each impression is treated independently in the process. Clearly, under this method there is a potential loss in preservation of the measured campaign reach and frequency data because spots viewed by the same individual are not linked directly. However, this option is still considered as the loss here may be offset if there are gains in accuracy of modelled cross media interactions.

So, the following 6 scenarios are considered and detailed below along with their pros and cons:

	Pros	Cons
Scenario 1: Abstract/Total TV Cookies	Follows WFA Framework Preserves TV reach perfectly	Dependent on availability of cookies. Rates based on training – limit to how much can be controlled.
Scenario 2: Abstract/Cookies By Channel	Follows WFA Framework	Dependent on availability of cookies. Rates based on training – limit to how much can be controlled.
Scenario 3: Abstract/No Cookies	Not dependent on availability of cookies.	Loses the benefit of cookie information.

Scenario 4: Real/Total TV Cookie	Uses observed data Preserves TV reach perfectly	Dependent on availability of cookies Range and mixture of rates limited by panel sample size.
Scenario 5: Real/Cookies by Channel	Uses observed data.	Dependent on availability of cookies Range and mixture of rates limited by panel sample size.
Scenario 6: Real/No Cookies	Not dependent on availability of cookies Uses observed data	Loses the benefit of cookie information

Work was undertaken to produce the two types of rates. Allocation algorithms as specified in the WFA framework were then applied using the appropriate inputs and protocols for the six scenarios.

5.3 Abstract Rates vs Real Rates

The abstract Dirac groups were selected to best model the reach behaviour of the panel; however in theory this can be done without creating a Dirac group for each individual panellist. Below details the number of abstract Dirac groups in each of the abstract methods in each demographic. Finally, the number of groups for 'real' rates are listed; these are used in all the real scenarios and here each panellist forms its own Dirac group so it is also gives the sample size of respondents that received at least one impression:

	Abstract Total TV Groups	Abstract By Channel Groups	Abstract No Cookie Groups	Real Groups
Men 18-34	39	57	13	178
Men 35-54	19	21	12	469
Men 55+	50	26	14	604
Women 18-34	29	29	10	350
Women 35-54	50	18	13	516
Women 55+	49	22	17	493

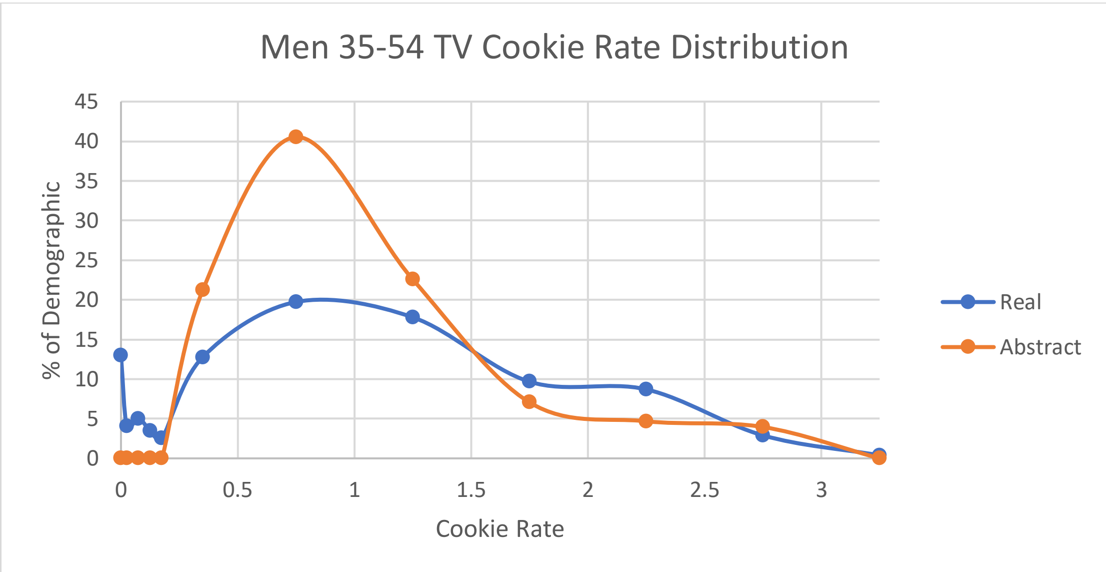
As there is no limit on the number of groups applied (aside from computational power) there is a risk of overfitting, however as the groups do not be of uniform size it is likely any overfitting would be represented by a small Dirac group that should have little impression on the overall output.

It should be noted that the abstract rates are not designed to mirror the true rate distribution of the census population (hence why the rates are abstract), they are designed solely to match the reach targets as accurately as possible. This is best demonstrated in the Men 35-54 demographic where the distribution of the real group's cookie rates and the abstract cookie rates are as follows:

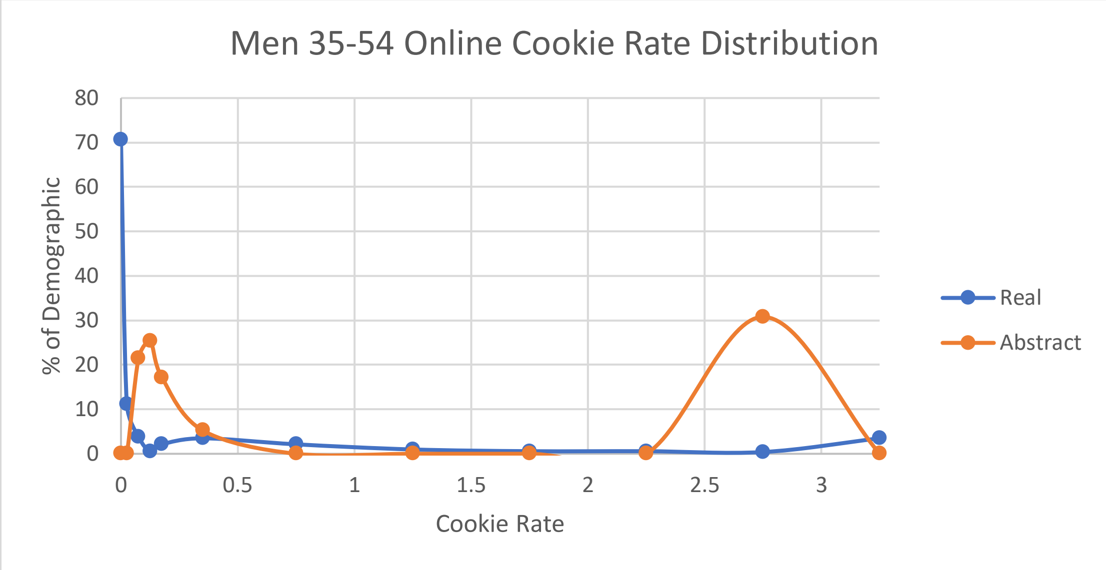
% frequency distribution

	Real		Abstract	
Rate	**TV**	**Online**	**TV**	**Online**
0	13	71	0	0
0-0.05	4	11	0	0
0.05-0.1	5	4	0	21
0.1-0.15	3	1	0	25
0.15-0.2	3	2	0	17
0.2-0.5	13	3	21	5
0.5-1	20	2	41	0
1-1.5	18	1	23	0
1.5-2	10	1	7	0
2-2.5	9	1	5	0
2.5-3	3	0	4	31
3+	0	3	0	0

This is more easily visualised as follows for TV:



And for Online:



Clearly the abstract rates have a distribution (particularly for online) which is unlikely to be observed in real life; however, as the focus of the model is to predict reach, the fact the abstract rates are not restricted to a distribution that you might obtain from a panel potentially could allow it to more closely model reach to the actual reach of a campaign.

6. Results – Evaluation of Mixed Campaigns

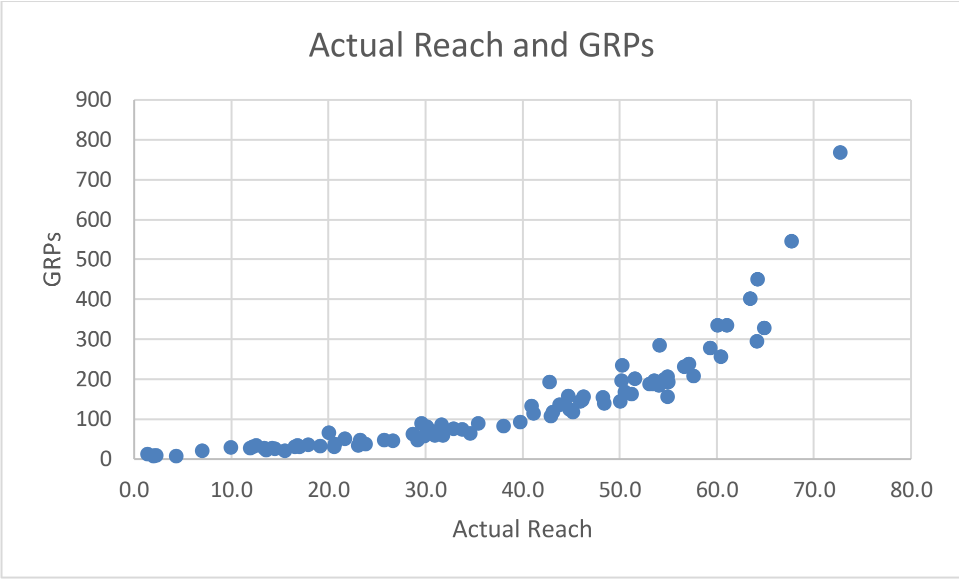
Stage 2 introduced an online element to the campaign dataset with impressions from 6,100 campaigns used to train the abstract rates; only 156 of these campaigns had at least 50 impressions, so these are the online campaigns that the analysis initially focuses on. Of these online campaigns 88 were linked to a TV campaign that was present in the dataset (either by a direct match or as a suitable surrogate). There were in addition a further 845 TV only campaigns which helped train the rates for the TV elements; of these 400 had at least 50 spots, only these 400 TV campaigns are included in any analysis. In addition based on the findings from Stage 1 the methodology for the abstract scenarios was adjusted slightly in an attempt to improve the accuracy of higher GRP campaigns.

For Stage 2 there are two questions of interest. The first is whether mixed campaign (comprised of both TV and online) elements can be modelled well using the VID model and the second is whether incorporating the online element has compromised the model performance (linked to this is an assessment of if the improvements to the model following the Stage 1 report have been successful). The focus is on the 88 campaigns that have both an online and a TV element to them. It is necessary to concentrate on these campaigns as modelling the duplication between different types of media is a key aim of Project Origin. It is important to note that in most of these campaigns the actual reach for TV is much higher than online. Reach interactions between websites cannot be assessed because each campaign has only impressions for a single website.

6.1 Total Campaign Reach

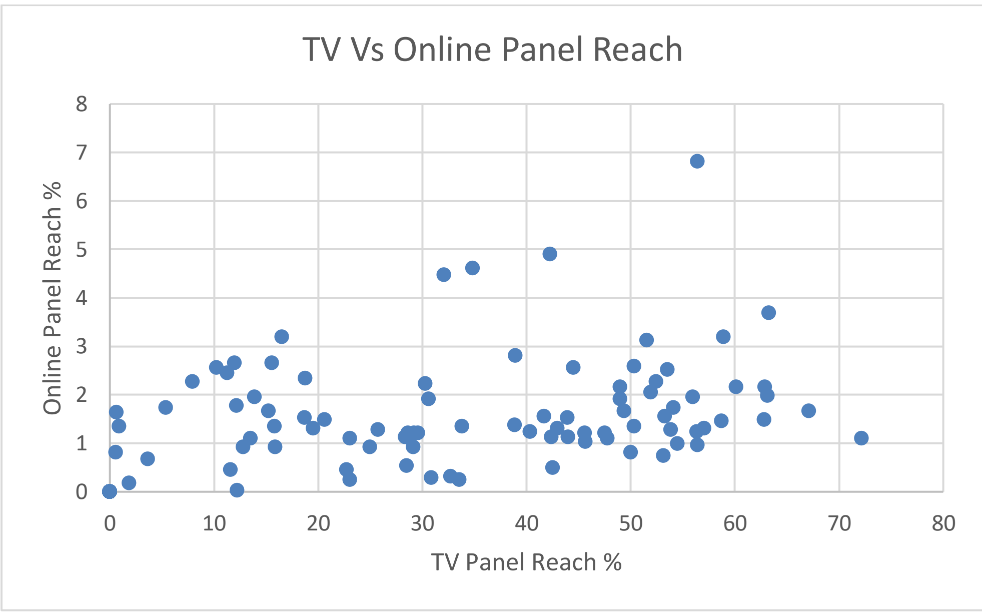
The first thing to examine is the performance of the modelling for overall reach for mixed campaigns.

Whilst there are only 88 mixed campaigns there is a good spread of reach and GRP values, also a variation of reach levels for campaigns with similar GRPs. The table below illustrates this:



Thus, the dataset is providing a reasonable variety of campaigns and differing reach and GRP relationships in order to test and evaluate the models.

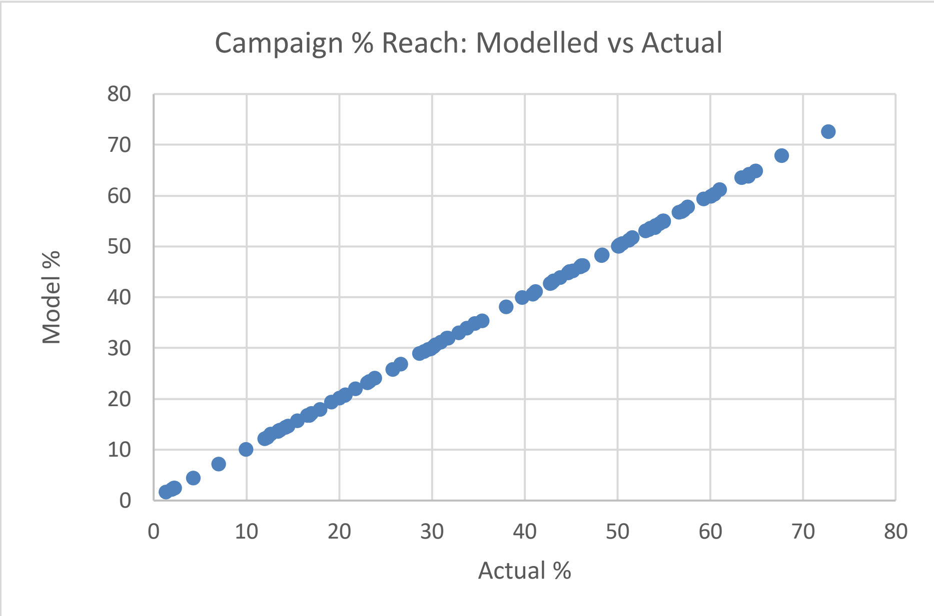
Comparing the panel TV element reach against the panel online element reach there is again a good spread of campaigns with differing levels of reach for each, although campaign reach is dominated by TV:



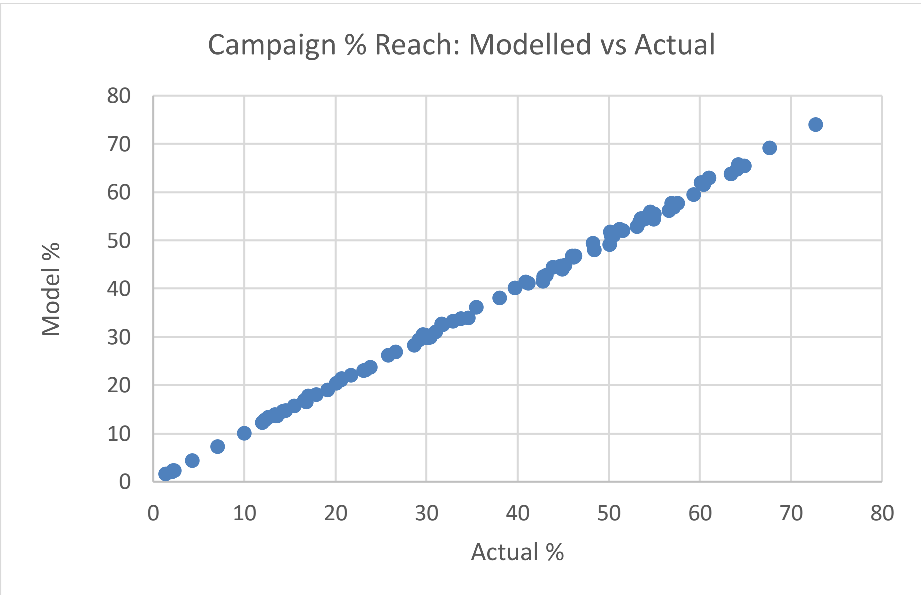
The overall total reach of these mixed campaigns will be dominated by TV reach. For the evaluation, the evaluations will assess how well the models can preserve TV reach, Online reach and Total campaign reach.

The below charts the percentage reach from the allocation against the actual total campaign reach for two of the scenarios:

Scenario 1 (Abstract/Cookie Total TV):



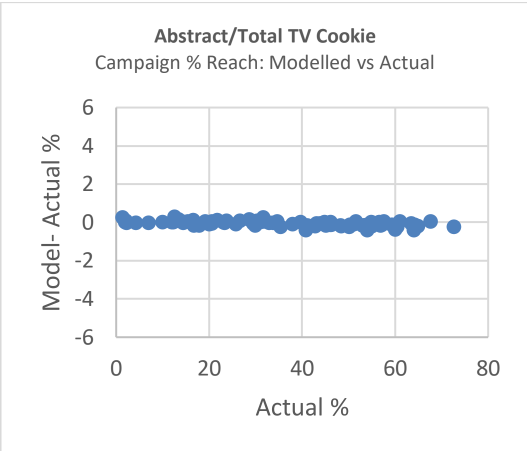
Scenario 2 (Abstract/Cookie by Channel):

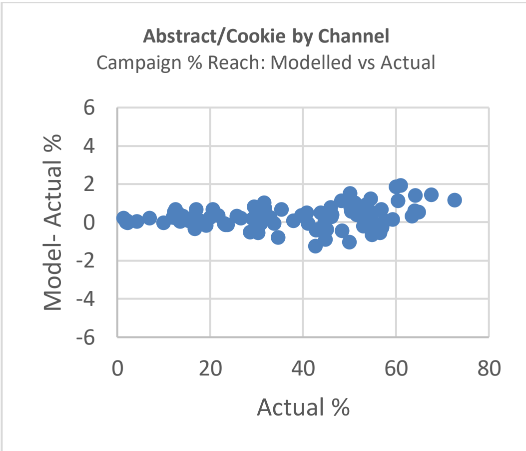


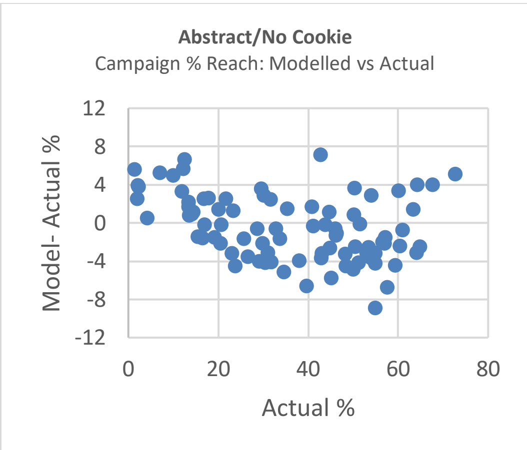
From the above, it can be seen that the modelled reach is close to actual reach; unsurprisingly the Total TV Cookie scenario shows a closer fit as TV reach is guaranteed to be matched in this scenario, so only the modelling of the online element can result in a non-perfect fit in the Total TV cookie scenario.

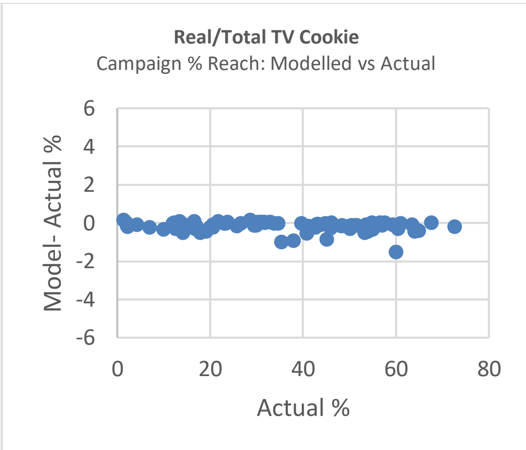
The key differences in goodness of fit are imperceptible from these types of charts and plotting the (Model – Actual) vs the Actual reach gives a clearer picture. This will better show any differences in the precision of modelled estimates between the scenarios and also any biases within the scenarios.

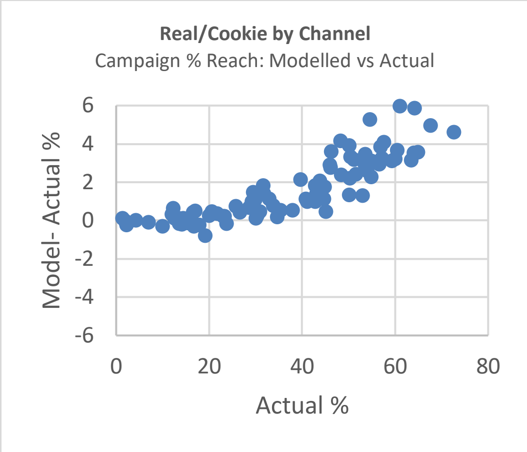
The below shows these for the 6 scenarios. Note that the y-axis is in terms of percentage points (e.g. if modelled reach was 11% and actual reach was 10% then Model – Actual % would be 1%):

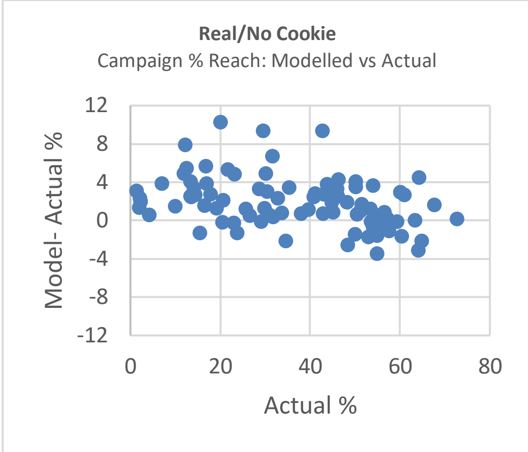










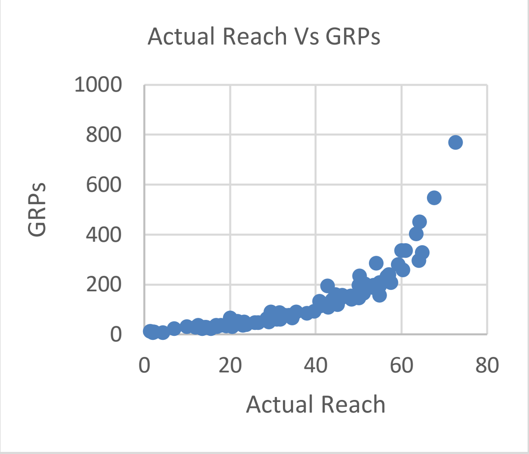


_A brief summary of the previous charts is given below:

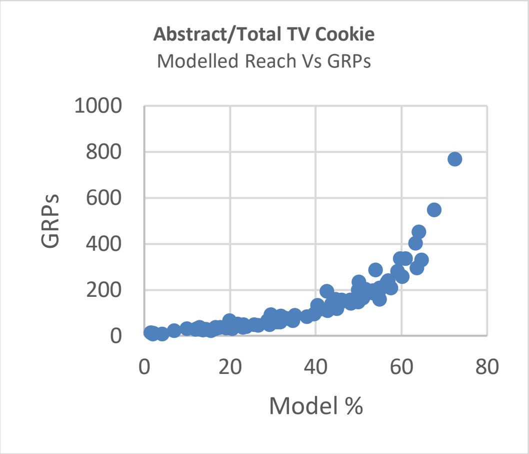
Scenario 1: Abstract/ Total TV Cookies	Closest fit, to be expected as only the online element can be inaccurate with this method, however no outliers are visible giving some indication that online is being modelled sensibly.
Scenario 2: Abstract/Cookies By Channel	Slightly more variability that scenario 1 but still a good fit with almost all reach estimates within 2 percentage points of actual. The modifications made following the findings from Stage 1 seemingly have helped remove the previous issue of overestimating high reach campaigns.
Scenario 3: Abstract/No Cookies	Noticeably more variation in results, with both overestimations and underestimations in reach present.
Scenario 4: Real/Total TV Cookies	A close fit in general (as to be expected as only the online component can't be perfectly estimated) however there seem to be some outliers, which could indicate online isn't always modelled well.
Scenario 5: Real/Cookies by Channel	Relatively close fit at low reach campaigns but a clear tendency to overestimate at higher reach campaigns.

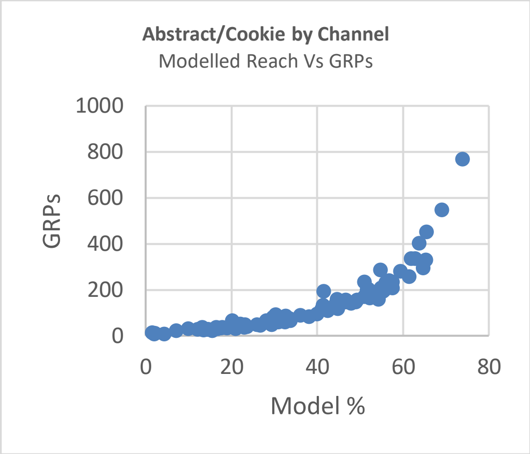
Scenario 6: Real/No Cookie	Notable variation in results like Scenario 3 although unlike scenario 3 there usually is an overestimation in reach.

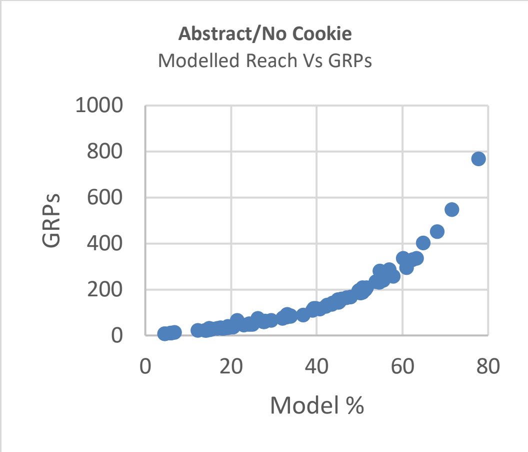
As a reminder GRPs plotted against Actual Reach was as follows:

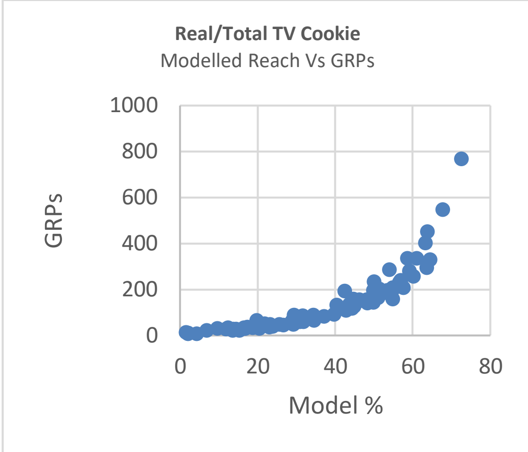


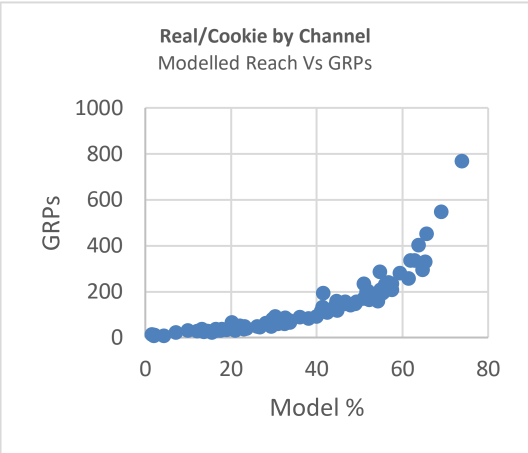
The modelled reaches when compared to GRPs look as follows:

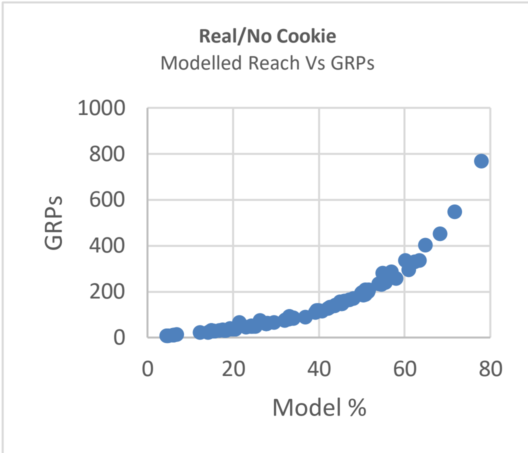












All methods show a similar shape to the actual reach, however the “no cookie” models show less dispersion than the actual reach curve has which could indicate that the “no cookie” models aren’t able to model the variation in reach as well as the cookie models.

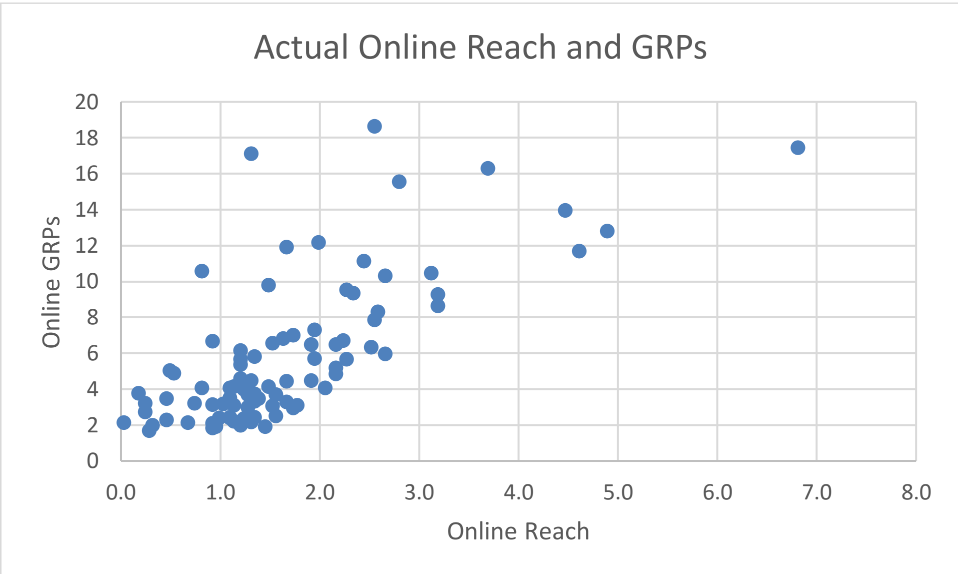
The table below summarises the performance across all scenarios for the 88 mixed campaigns.

	Rates	Cookie used	Diff	Abs Diff	Standard Deviation
Scenario 1:	Abstract	Total TV	0.0	0.1	0.1
Scenario 2:	Abstract	By Channel	0.3	0.5	0.6
Scenario 3:	Abstract	None	-0.6	2.9	3.4
Scenario 4:	Real	Total TV	-0.2	0.2	0.3
Scenario 5:	Real	By Channel	1.6	1.7	1.6
Scenario 6:	Real	None	1.9	2.5	2.7

Although averages can mask what happens at an individual level (e.g. Scenario 4's outliers), all scenarios perform well in terms of reach. The average difference shows that Abstract/Total TV Cookie has the least bias overall, due to the guarantee of TV reach matching but the Abstract Cookie by Channel model also performs very well. The other scenarios show a similar level of closeness of fit, although for Scenarios 5 and 6 the consistent figures for difference and absolute differences reinforce the findings seen in the charts whereby there is a consistent over estimation. The 'no cookie' models (especially Scenario 3) have higher levels of standard deviation, perhaps reflecting their inability to account for repeat viewers in the way cookie models can; this suggests that whilst on average the reach differences are small there will be many campaigns with a poorly modelled reach in these scenarios compared to the scenarios that utilise a cookie.

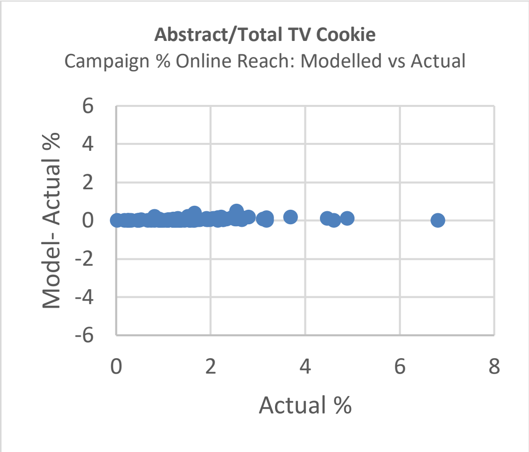
6.2 Online Campaign Reach

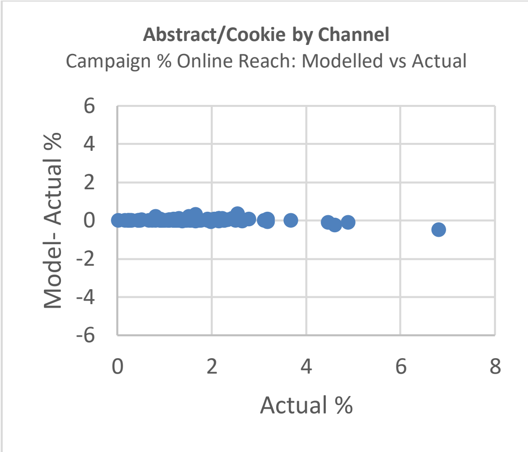
Given that the reach is dominated by TV it is also worthwhile to examine only the online elements of campaign reach (reach for other stations are examined across all campaigns in Section 7). The GRP distribution for the online campaigns is as follows:

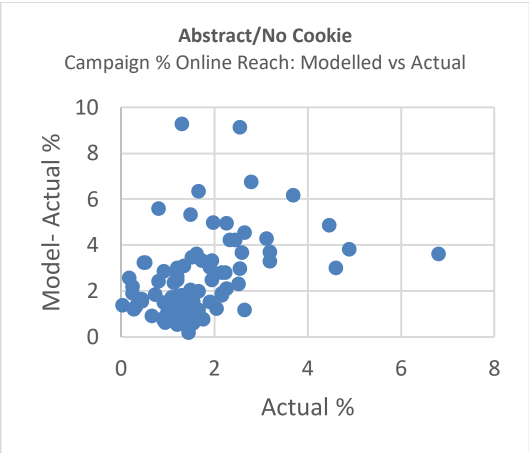


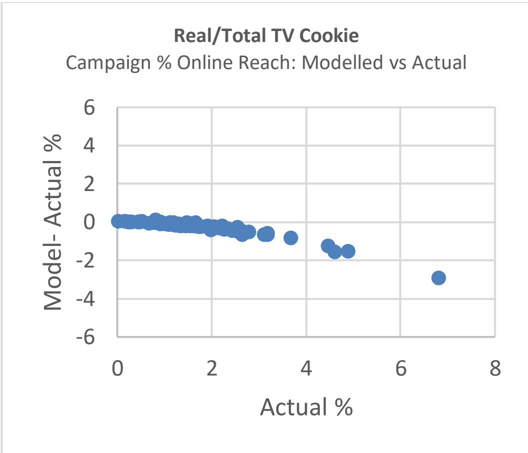
We can note that all the campaigns are relatively small, however if the models perform well on these campaigns it's a good indication that the model could also work as well for larger online campaigns. Whilst there is a clear relationship between GRPs and Online Reach we can also note that there are quite a few outliers (a non-cookie based model in particular may struggle to model these outliers well) – so a good test for model performance.

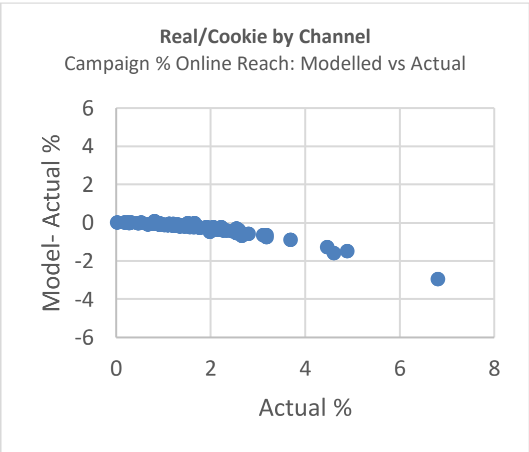
The (Model – Actual) Vs Actual plots are as follows:

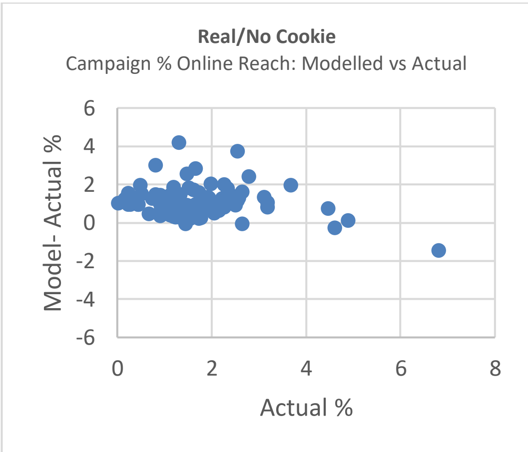












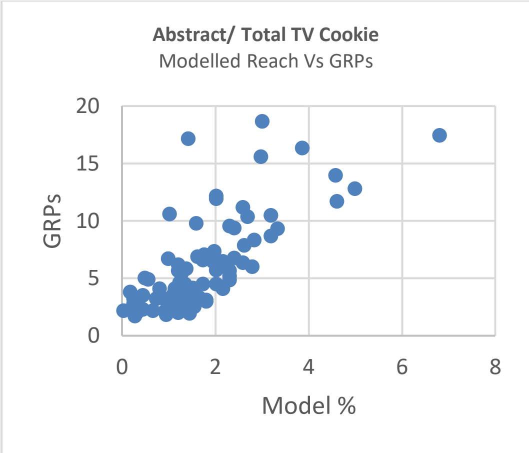
A brief summary of the above:

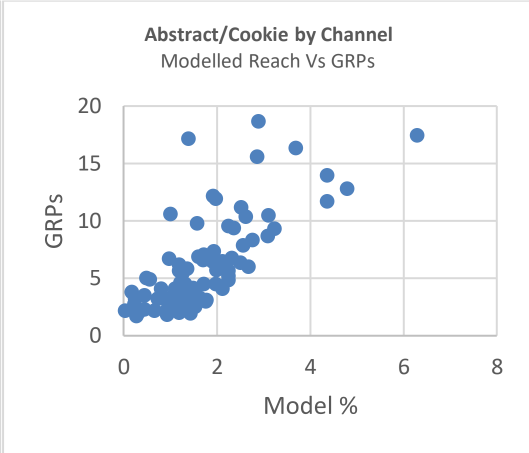
Scenario 1: Abstract/ Total TV Cookies	Very close to the true online reach in all 88 campaigns with no notable outliers.
Scenario 2: Abstract/Cookies By Channel	Very close to the true online reach in all 88 campaigns with no notable outliers.
Scenario 3: Abstract/No Cookies	Unacceptable levels of overestimation in online reach. In many places the modelled reach is more than double the actual.
Scenario 4: Real/Total TV Cookies	A close fit in general however there is a clear underestimation of reach in the higher online reach campaigns.
Scenario 5: Real/Cookies by Channel	This is identical to scenario 4 when looking at the online element only.
Scenario 6: Real/No Cookie	Seems to overestimate reach in most cases, some outliers with unacceptably high discrepancies.

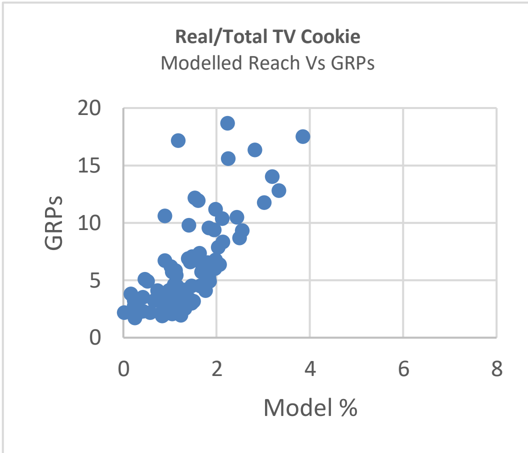
From the above it seems that cookies are a vital element of a VID model if we want to model online reach as the 'No Cookie' scenarios perform poorly when modelling online.

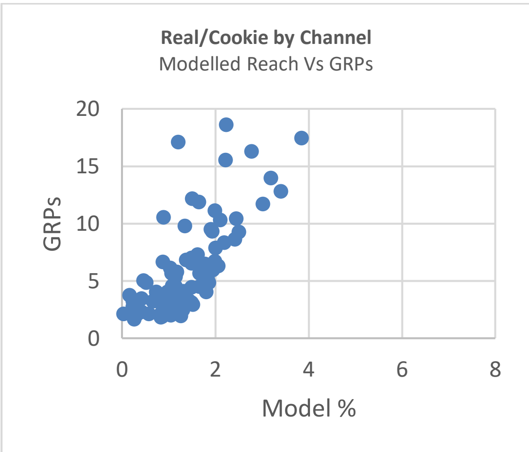
Scenarios 4 and 5 show that there is a clear, systematic trend to under-estimation of online reach when using real rates of exposure. It suggests that there is an implicit cap on reach build within the model. This could be caused by discreteness in the training dataset - for example, not enough respondents have small but non-zero rates of exposure. The abstract scenarios address this issue by creating rather than observing a spread of rates across the population.

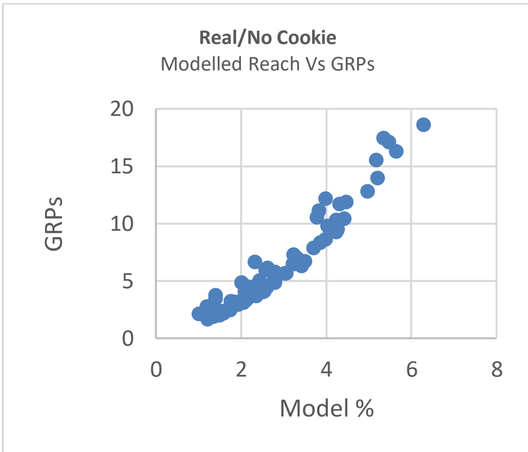
The online GRP plots for each scenario are as follows and give some indication as to why the 'no cookie' scenarios perform so badly:











We can see that the 'No Cookie' models have far too strong a relationship between Online Model Reach and Online GRPs, which explains why they perform so poorly. The 'No Cookie' scenarios are unable to account for the variation in GRP to reach relationships in different online campaigns and so it ends up trying to settle for somewhere in the middle, meaning that any campaign that doesn't fit this average behaviour is poorly modelled.

The average differences between online modelled and actual for each scenario are:

	Rates	Cookie used	Diff	Abs Diff	Standard Deviation
Scenario 1:	Abstract	Total TV	0.1	0.1	0.1
Scenario 2:	Abstract	By Channel	0.0	0.0	0.1
Scenario 3:	Abstract	None	2.5	2.5	1.8
Scenario 4:	Real	Total TV	-0.3	0.3	0.4
Scenario 5:	Real	By Channel	-0.3	0.3	1.0
Scenario 6:	Real	None	1.1	1.1	0.8

Whilst the scenario 3 and 6 differences don't appear too large in absolute terms, within the context of the size of the online campaign elements (the largest being less than 7%) these are relatively large differences that indicate that the 'no cookie' scenarios do not model online reach well. The performance of the abstract cookie scenarios (scenarios 1 and 2) in contrast are very close to actual and in the context of online reach these models perform very well.

6.3 Reach By Demographic

For this proof of concept the allocation process for TV campaign data was undertaken at a demographic level. The breakdown was as follow:

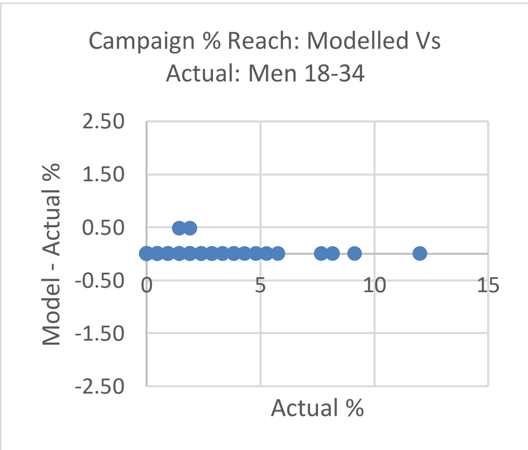
Men/Women * Age (18-34, 35-54, 55+)

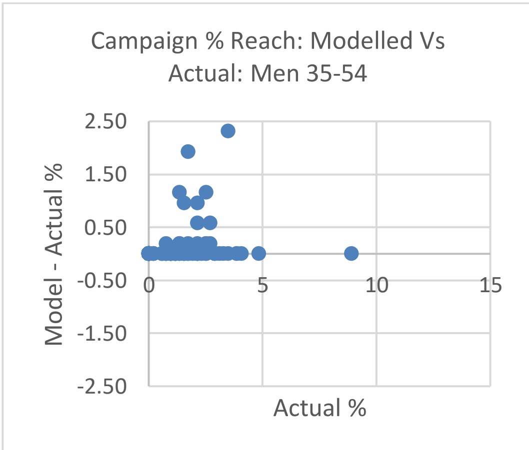
Similar to the channel data it is intuitive to examine if there are any performance differences in the allocation process.

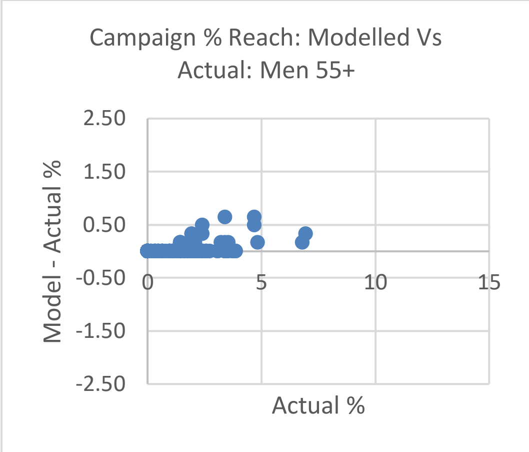
Data analyses were undertaken for all scenarios but only the Abstract/Total TV Cookie and Abstract/Cookie by Channel scenarios are detailed below.

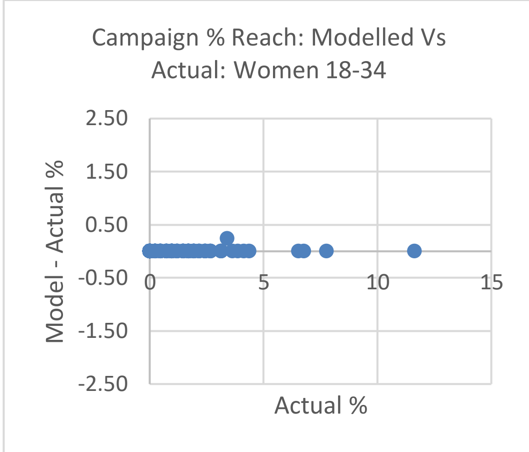
	Abstract/Total TV Cookie			Abstract/Cookie by Channel		
	Diff	Abs diff	St Dev	Diff	Abs diff	St Dev
All	0.0	0.1	0.1	0.3	0.5	0.6
Men 18-34	0.0	0.3	0.4	0.0	1.0	1.4
Men 35-54	0.0	0.2	0.4	0.2	0.9	1.2
Men 55+	0.0	0.2	0.3	0.0	0.9	1.1
Women 18-34	0.0	0.2	0.3	-0.3	0.7	1.1
Women 35-54	0.0	0.2	0.2	0.2	0.7	1.0
Women 55+	0.0	0.2	0.3	1.3	1.5	1.7

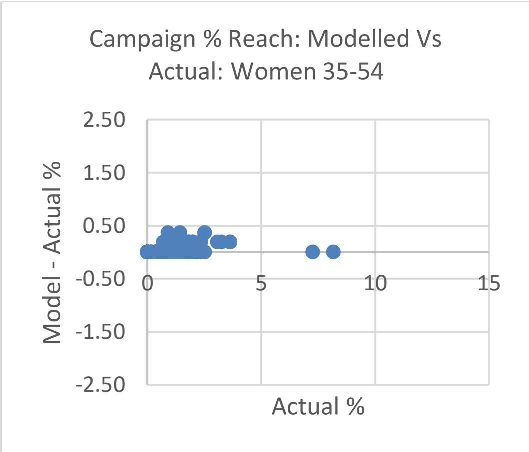
The demographic behaviour largely mirrors what we saw at a topline level (with the exception of women 55+ which strangely performs worse than the other demographics for Abstract Cookie by Channel) and this is observed for the other scenarios as well. What is of more interest within the 88 mixed campaigns though is how the demographics perform for the online element, unfortunately with such low levels of reach at a topline level these get even smaller at a demographic level, so plots of each demographic are more illustrative than definitive. Below we focus on Abstract/Total TV Cookie (however Abstract/Cookie by Channel looks very similar):

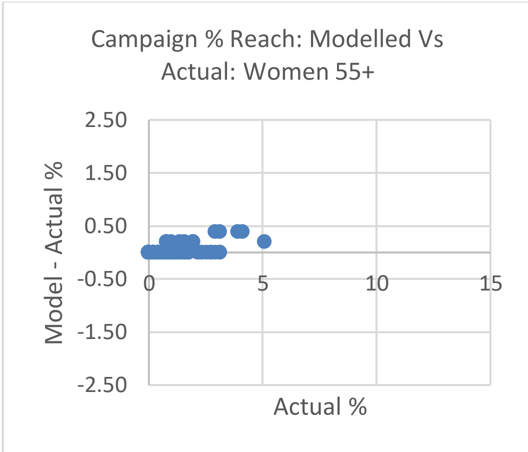










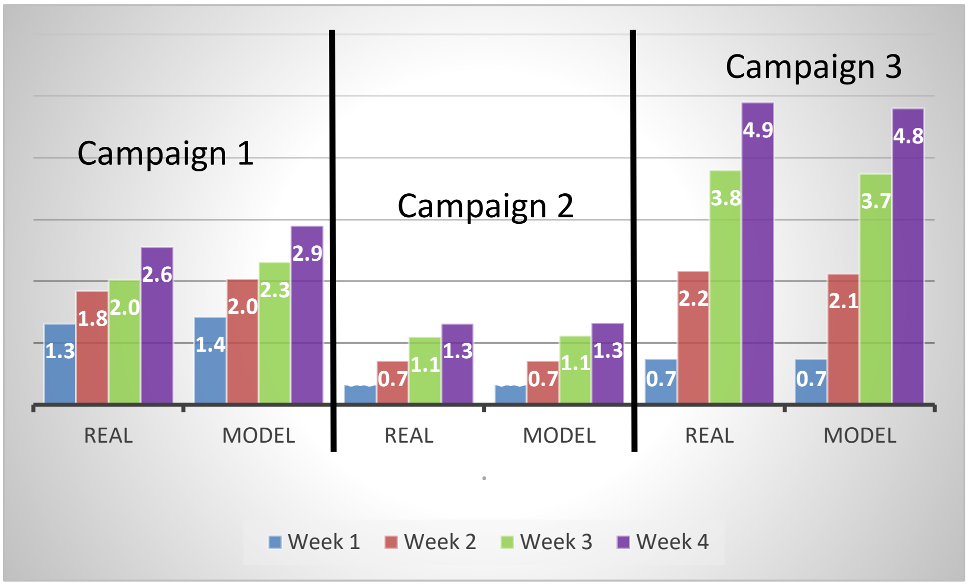


For 5 of the six demographics there doesn't seem to be any cause for concern, there is slightly more discrepancy than seen at a top line level but this is to be expected since it is dealing with smaller amounts of data.

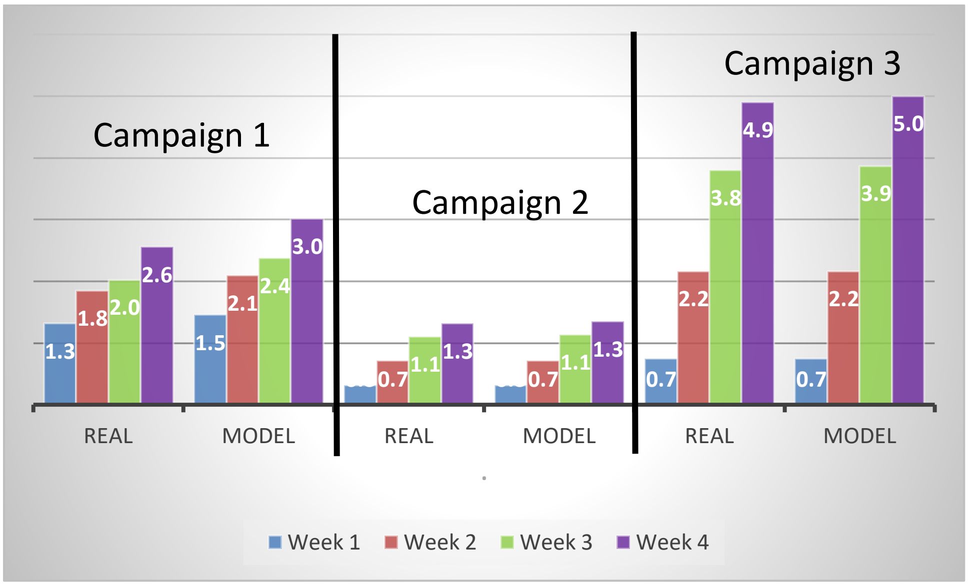
Men 35-54 has a few outliers when the reach estimation is poor; further investigation revealed this is down to high cookie volatility in the real data between campaigns for this demographic. For example one of the poorly performing campaigns had an actual reach sample of 180 despite having 380 cookies, this is in contrast to another campaign which had 360 cookies but also a reach sample of 360. As the allocation is based on cookies the model cannot suitably model the reach for both of these campaigns since they have similar levels of cookies. RSMB does not believe this a cause for concern; However, this is a symptom of the data size being so small which drastically increases the probability of an outlier campaign forming in the real data. With more data abnormal behaviour (such as a individual with a high number of impressions with frequent cookie deletion) it will have a much smaller impact on the total reach (which is why we don't see these reach outliers at a top line online level).

6.4 Week by Week Build

Below shows 3 example campaigns' week by week build for the online element of campaigns. First for Abstract/Cookie by Channel:



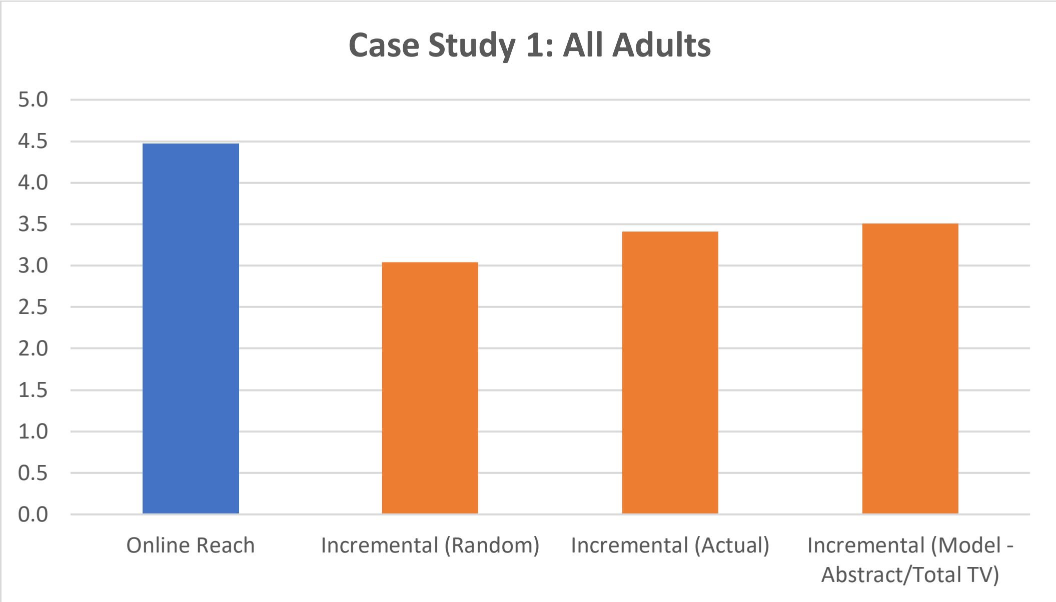
And for Abstract/Total TV Cookie:



The build in general looks consistent across the three examples, whilst there is a slight overestimation in example 1 this is present even in Week 4 (which is equivalent to total reach), so this doesn't indicate any issue with how reach built is developed.

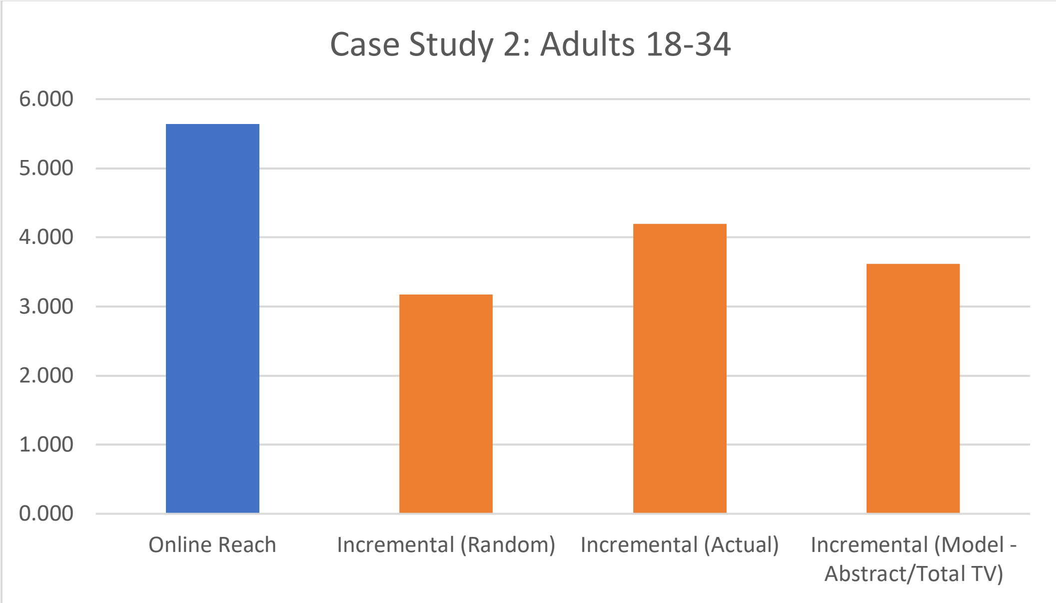
6.5 Regression to the Mean: Example Campaigns

In the next section the regression to the mean overall will be examined; however, it is worth first looking at a couple of real examples from the dataset to understand what this metric represents and to give an illustration of how the model performs.



The above shows the actual online reach of a campaign along with the actual incremental reach (i.e. how much of the online reach isn't duplicated with TV reach), this is then compared to the incremental reach that the model (here being Abstract/Total TV) produces. Also shown is the random incremental reach, this is the expected reach if we just assigned reach at random. If the model produces a figure similar to random then it isn't actually doing much to predict the marked duplication; however, in this case we can see the duplication has been predicted very well.

We now look at a second example, this time looking at Adults 18-34s:



Here we can see that whilst the model is greater than random the duplication isn't as close to actual as previously. We can use the regression to the mean statistic in order to approximate how "far" from actual the model has moved towards the random incremental reach.

Scenario 1: Abstract/Total TV Cookie

Total Reach	=	5.65%
Random duplication	=	5.65 – 3.17 = 2.48%
Actual total panel duplication	=	5.65 – 4.19 = 1.45%
Model duplication	=	5.65 – 3.61 = 2.03%
RTM	=	(1.45-2.03)/(1.45-2.48) = 56%

Using the random, actual and model duplication (the difference between incremental and total online) we calculate a regression to the mean figure of 56%, i.e. the duplication reach of the model has moved 56% of the way towards random from the actual duplication reach. Whilst the ideal is for this figure to be 0 anything below 100% represents an improvement from random allocation, indicating that the model provides some benefit even here.

6.6 Regression to the Mean: All Campaigns

Below shows the average regression to the mean figures for each method, in addition the campaigns are then split out by relative size of campaign and the average RTM is then recalculated based on that:

	Abstract			Real		
Size of Campaigns	**Total**	**Channel**	**None**	**Total**	**Channel**	**None**
Total	**42**	**38**	**463**	**57**	**40**	**-13**
High	30	20	421	50	29	-37
Medium	51	47	334	61	41	1
Low	64	78	748	72	69	35

We can see that the best performing scenario (by this metric) is Real/No Cookie; however this should not be taken to mean that it is the best scenario overall. We have already observed that this scenario fails to preserve online reach well, so even though it may have the lowest average RTM the poor prediction elsewhere means it should not be championed. Of the methods that did predict online reach well all show reasonably good RTM statistics, with the Abstract/Cookie by Channel performing particularly well. The largest campaigns are the most robust and give the best indication of the “true” regression to the mean statistic.

Looking now at the 18-34 demographic we can see a similar pattern, albeit with drastically worse RTM statistics at low sized campaigns; this isn’t a big concern as it is down to RTM being volatile when the reach of a campaign is very small:

	Abstract			Real		
Size of Campaigns	**Total**	**Channel**	**None**	**Total**	**Channel**	**None**
Total	**17**	**31**	**-389**	**31**	**39**	**5**
High	3	20	-488	19	28	-12
Medium	35	36	-468	41	52	12
Low	295	322	3280	320	278	464

We can create a more diverse test dataset comparing the incremental online reach against each possible combination of TV channels within the campaigns (so for example one data point would be online incremental compared to TV Station 1, another would be online incremental compared to TV Station 1 and TV Station 2 pairwise reach etc.):

	Abstract			**Real**		
Size of Campaigns	**Total**	**Channel**	**None**	**Total**	**Channel**	**None**
Total	**48**	**41**	**539**	**64**	**46**	**-11**
High	34	23	465	57	36	-34
Medium	65	57	543	71	54	4
Low	71	76	797	76	68	42

This produces a similar picture to the previous, albeit with higher levels of RTM compared to when we looked only at Total TV.

In general the findings from this section suggest that it is important that a cookie (or equivalent identifying parameter) is used in order to model online reach well. It is evident (based on section 6.1) that the abstract cookie methods are less prone to overestimating reach than the real equivalents (although this isn't as big a concern for total TV cookie at a top line level). Section 6.2 suggests that the abstract cookie methods are less prone to underestimating online reach (particularly at higher levels) than the real cookie methods and the abstract cookie methods perform just as well as the real cookie methods for regression to the mean tests. Therefore it seems that is a benefit to using abstract Dirac groups over real Dirac groups when incorporating an online element into the model.

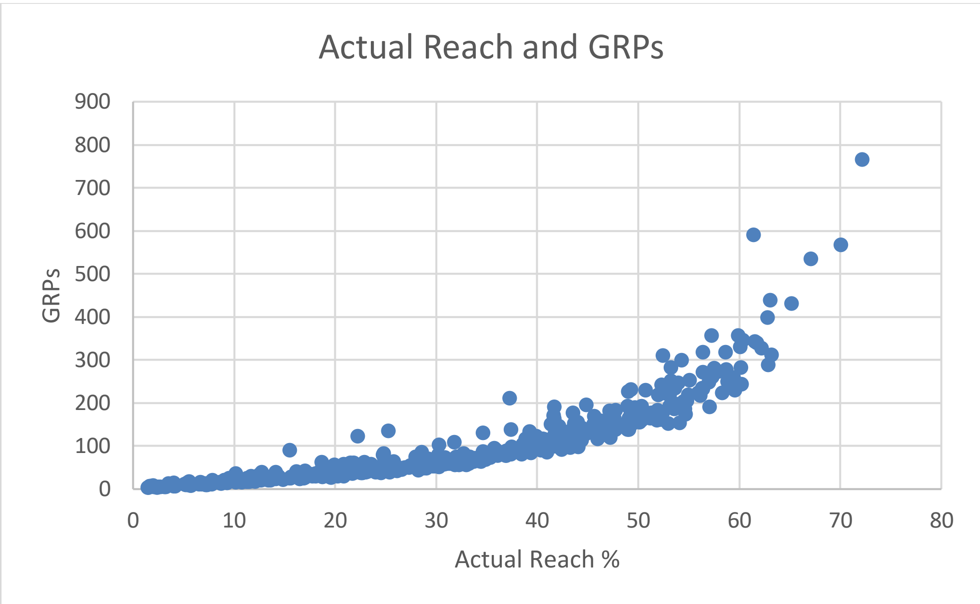
The analysis also shows that there is an improved estimation of online incremental reach when the TV "cookie" is broken down by channel. Of course this must be traded-off against the loss of control in Total TV reach for each campaign.

7. Domain of Study

Whilst in Stage 1 encouraging signs were shown towards the effectiveness of the methodology, one concern was that the period used for the training and test dataset were the same which is tautological. This meant there was a danger that the rates obtained could fail when applied to campaigns from a different period than those used to inform the rates. To evaluate this concern, the rates obtained from Stage 1 were applied to the campaigns used in Stage 2. This allows the performance of the rates to be evaluated against campaigns that did not contribute to the calculation of the rates. This reflects the intended real life application. All four scenarios were evaluated for 470 campaigns (N.B. the 'Cookies Across Channels' scenarios were not evaluated for Stage 1 as reach is guaranteed to be preserved).

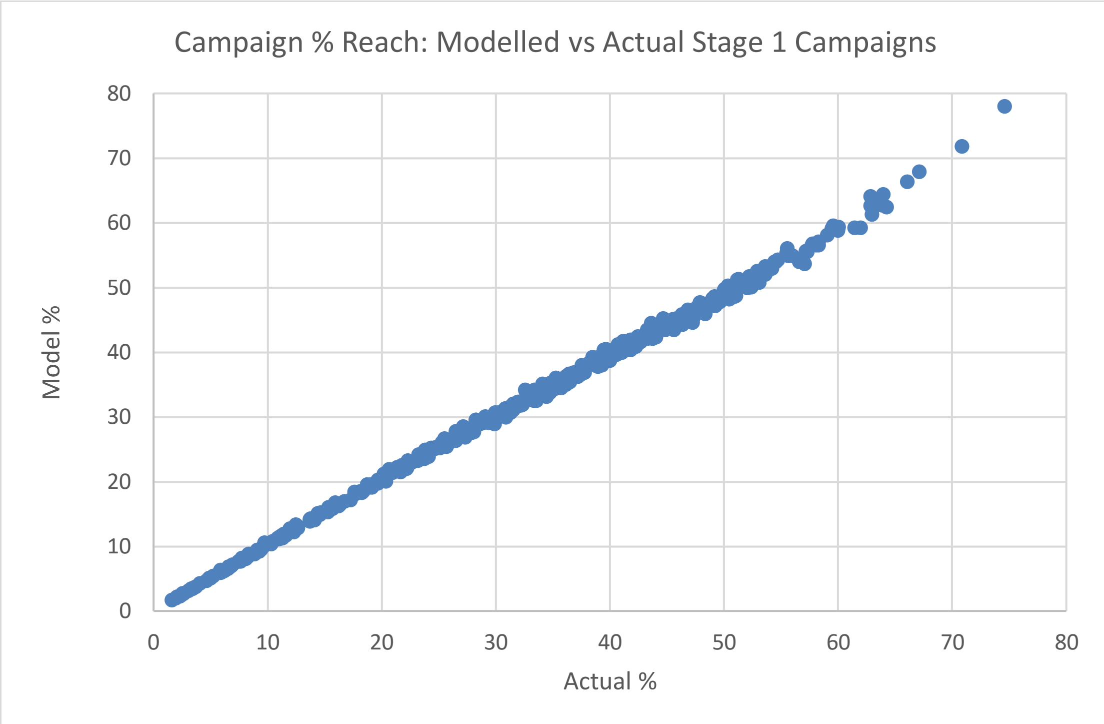
7.1 Total Campaign Reach

To repeat, a key requirement of the model is that it produces credible estimates for the overall campaign reach. The actual reach of each Stage 2 campaign (TV Only) compared to GRPs is shown below:

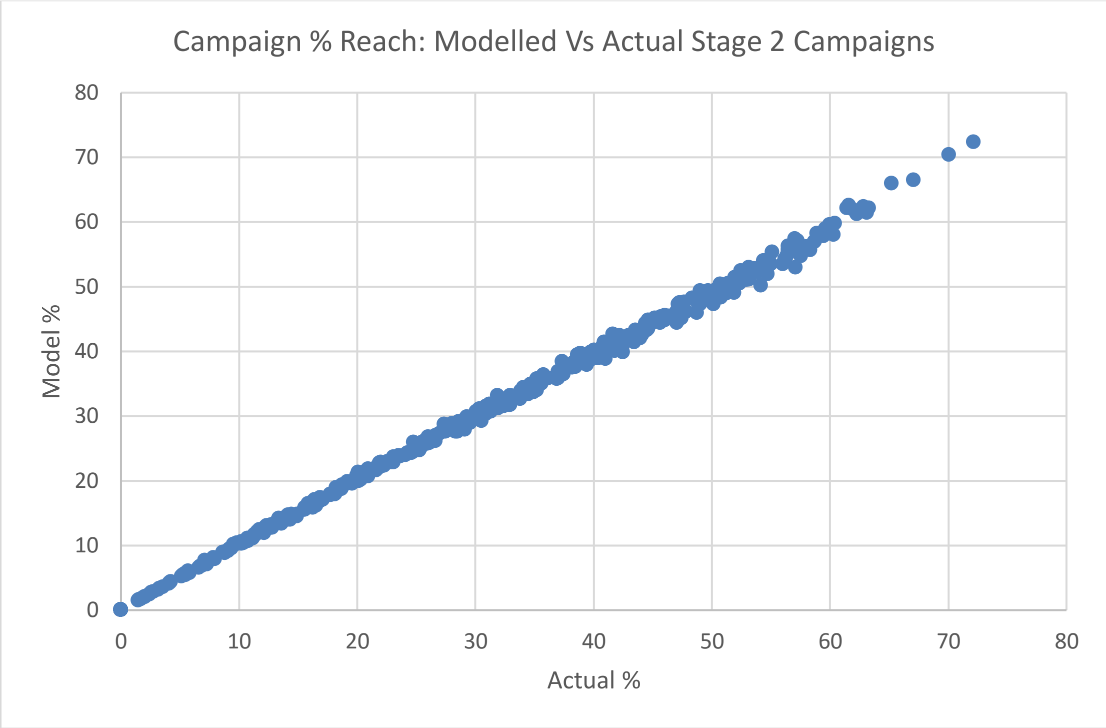


As with all statistical evaluations, the proof of concept can only assess model performance within the bounds of the input data. There is a clear relationship between Reach and GRPs but there is some noticeable variation in reach around the underlying curve. For example, at around 150 GRPs the reach ranges from 40% to 50%, with an average spread of around plus or minus 5 points. This significantly exceeds the sampling errors and confidence intervals cited in section 3 and reassures that the dispersion within the domain of study is systematic rather than random. Essentially the model has to work hard to reflect systematic variations in the reach to frequency relationships and the relatively low sampling error means that the test is quite powerful. Inevitably, systematic variations within demographics are more confounded with sampling error.

In Stage 1 the rates for Scenario 1 (Abstract/Channel Cookie) produced a close fit for the Stage 1 campaigns:

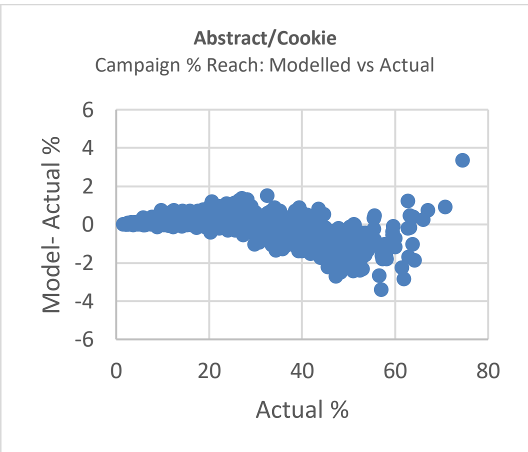


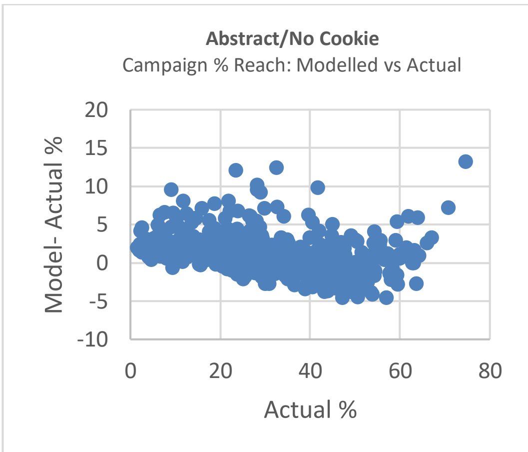
If we apply these same rates to the Stage 2 data a reassuringly similar story is seen:

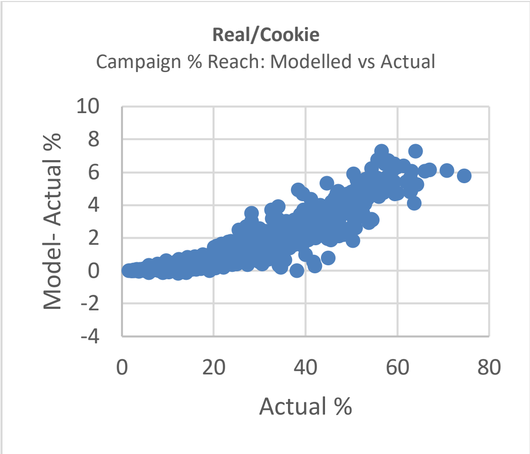


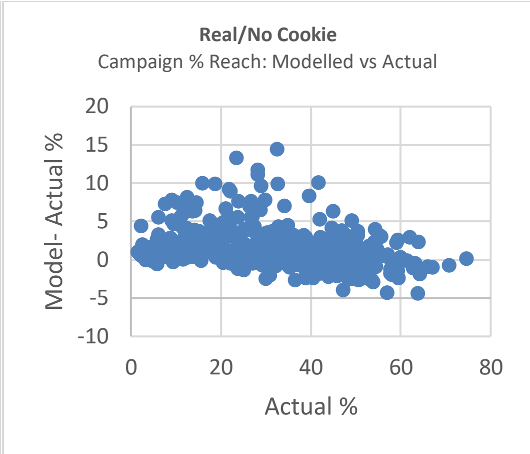
The other scenarios show a similar relationship between the Stage 1 campaign results and the Stage 2 campaign results.

The below shows (Model – Actual) vs Actual for the 4 scenarios using Stage 1 rates **(Jan)** modelling Stage 1 Campaigns **(Jan)**. Note that the y-axis is in terms of percentage points (e.g. if modelled reach was 11% and actual reach was 10% then Model – Actual % would be 1%):

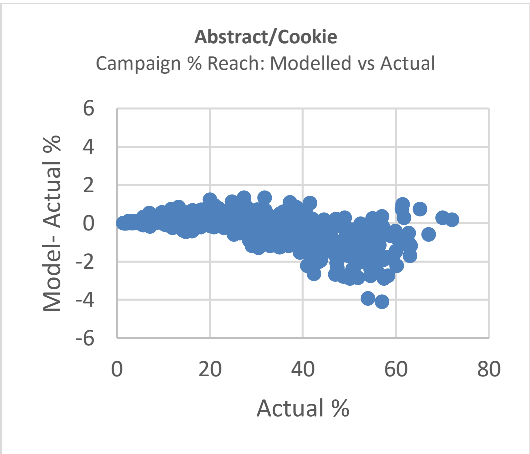


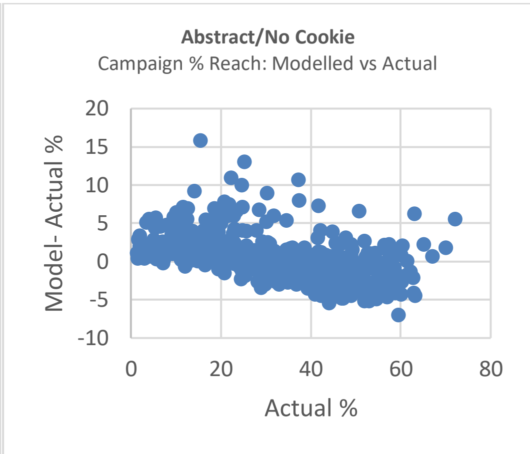


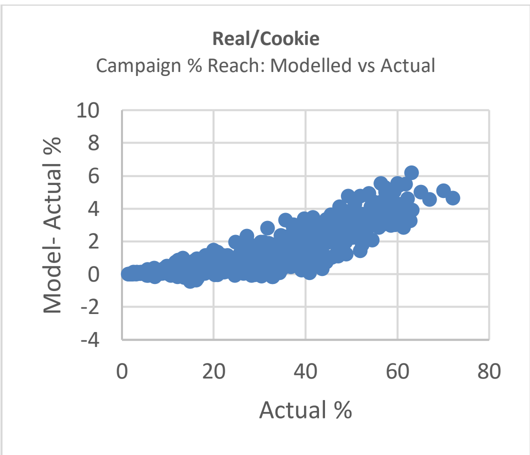


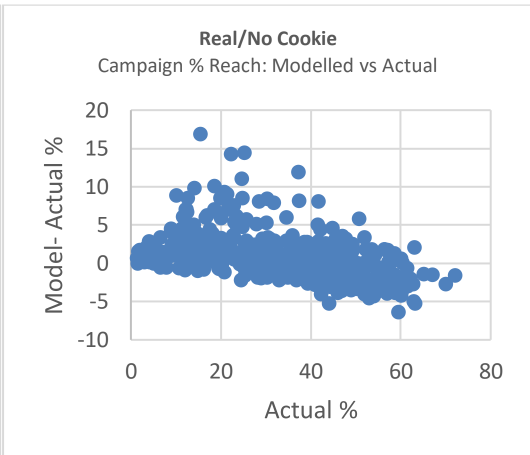


The equivalent charts when the Stage 1 rates **(Jan)** are used to model Stage 2 TV Campaigns **(Apr)** are as follows:









Whilst there are some small differences in distribution in general applying the Stage 1 rates to the Stage 2 data does not seem to produce notably different results, suggesting that they are still applicable at a top line level when applied to a different campaign period.

The table below summarises the performance across all scenarios for the campaigns in Stage 1:

Jan rates to **Jan** Campaigns

	Rates	Cookie used	Diff	Abs Diff
Scenario 1:	Abstract	Yes	-0.3	0.6
Scenario 2:	Abstract	No	1.0	2.1
Scenario 3:	Real	Yes	1.6	1.6
Scenario 4:	Real	No	1.4	2.0

The equivalent statistics when the rates are applied to Stage 2 TV Campaigns are as follows:

Jan rates to **Apr** Campaigns

	Rates	Cookie used	Diff	Abs Diff
Scenario 1:	Abstract	Yes	-0.4	0.7
Scenario 2:	Abstract	No	0.4	2.2
Scenario 3:	Real	Yes	1.5	1.5
Scenario 4:	Real	No	0.9	2.1

In general the results when applied to the Stage 1 Campaigns and Stage 2 television campaigns are similar. There are some slight improvements in the non-cookie methods in terms of Average Difference; however, as the Average Absolute Differences are similar this is likely just coincidental as the spread of reach outputs compared to actual is similar between the two campaign test sets.

7.2 Reach by Demographic

Data analyses were undertaken for all scenarios but only the Abstract/Cookie scenario is detailed below. For the Stage 1 campaigns the differences by demographic were as follows:

	Abstract/Cookie	
	Diff	Abs diff
All	-0.3	0.6
Men 18-34	0.5	0.9
Men 35-54	-0.4	0.9
Men 55+	-0.4	1.0
Women 18-34	0.1	0.8
Women 35-54	-0.7	1.1
Women 55+	-0.3	1.0

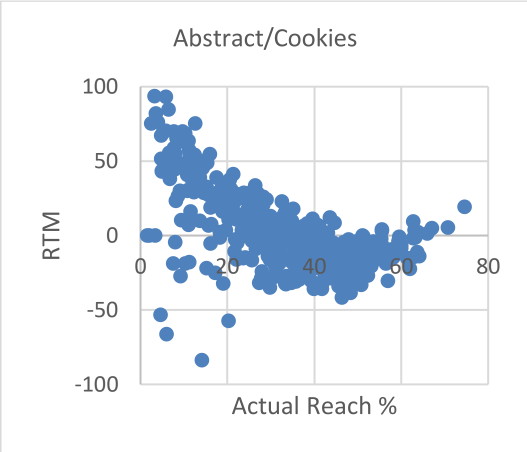
The differences for the Stage 2 TV campaigns are as follows:

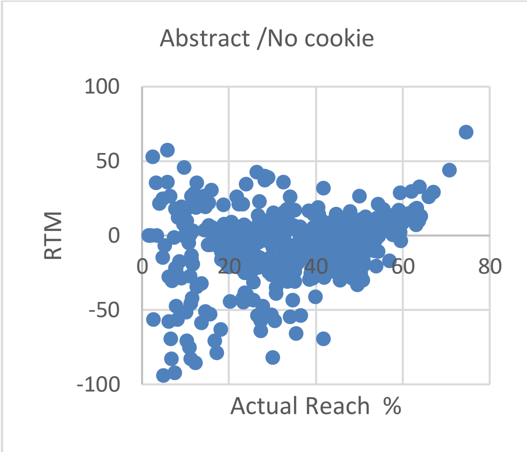
	Abstract/Cookie	
	Diff	Abs diff
All	-0.4	0.7
Men 18-34	0.3	0.9
Men 35-54	-0.5	1.1
Men 55+	-0.5	1.0
Women 18-34	0.2	0.7
Women 35-54	-0.7	1.2
Women 55+	-0.5	1.2

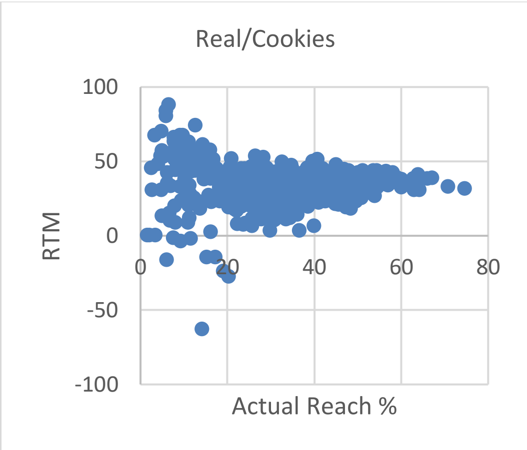
Differences are broadly similar for both the Stage 1 campaigns and the Stage 2 campaigns for this scenario. Other scenarios were similar in reflecting their corresponding patterns from Stage 1.

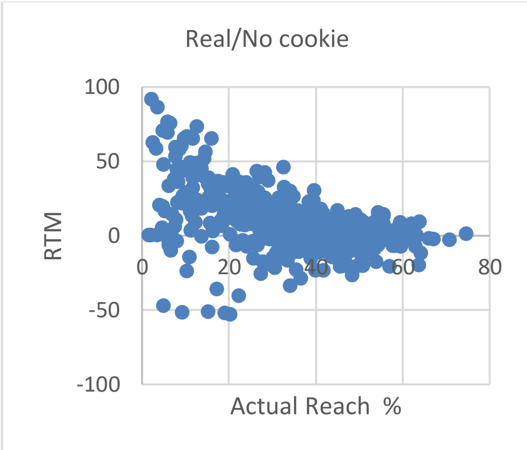
7.3 Overall Reach – Regression to the Mean

An important indicator of performance is regression to the mean. This calculation was previously undertaken for all 459 Stage 1 campaigns for each scenario. The following charts compare the regression to the mean against the reach of the campaign:



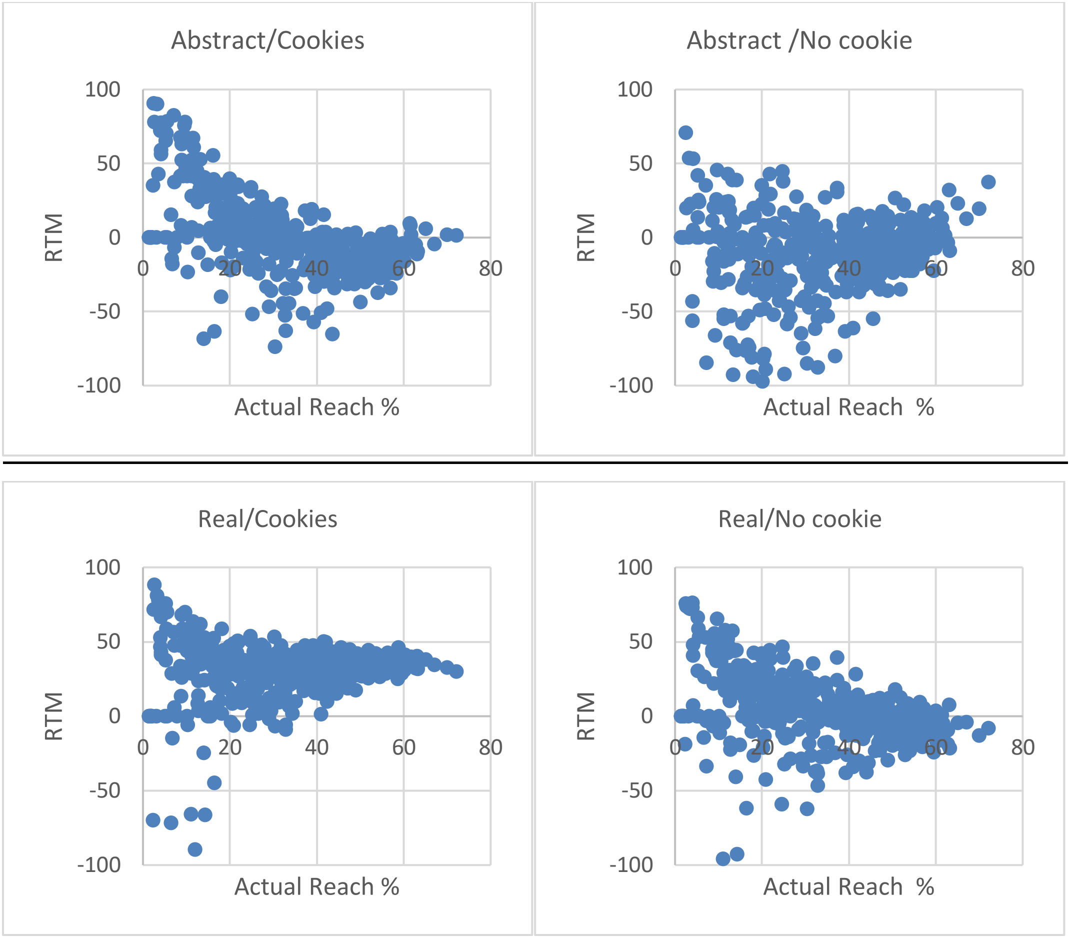






The charts above show varying patterns of regression to the mean for these scenarios. The important thing is that would the evaluation been different if we were not training and evaluating on the same period.

Applying the same rates to the Stage 2 TV Campaign Data (i.e. a different period) the RTM statistic plots are as follows:



As observed with the other comparisons there is little to distinguish between the output of applying the Stage 1 rates onto the Stage 1 TV campaigns and applying the Stage 1 rates onto the Stage 2 TV campaigns.

7.4 Summary

Overall there were no significant differences observed when using the Stage 1 rates to predict the Stage 2 Allocation. This is an important finding as, although in practice the model would be trained on one dataset of campaigns and then applied onto another set of campaigns, the results from the original Stage 1 evaluation whereby the data was trained and evaluated on the same period are representative and fears of tautology contaminating the evaluation are allayed. The Stage 2 results for mixed campaigns are based on training and evaluating on the same period but there is now confidence that this tautology is unlikely to be an issue here either and findings from this are valid too.

8. Summary

Using the Compass panel (run by Ipsos MORI) Stage 2 of Origin was evaluated by treating the panel as a census and applying the methodology to a set of campaigns with TV and online elements, with the TV element known to have been broadcast in the period of the dataset and the online element taken directly from the dataset. Six scenarios were assessed (Abstract rates to assign cookies (with Total TV viewing treated as a single cookie); abstract rates to assign cookies (with each channel's viewing treated as a separate cookie); abstract rates to assign impressions; panel based rates to assign cookies (with Total TV viewing treated as a single cookie); panel based rates to assign cookies (with each channel's viewing treated as a separate cookie); panel based rates to assign impressions) and their performance evaluated. Performance was assessed at a top line level for reach as well as by demographic, type and week by week. Regression to the mean was also used to assess model performance.

Results indicated that a cookie (or equivalent identifier that allows to at least partially link impressions from the same user together) is a vital component of modelling the online reach of campaigns. Without this identifier reach overestimates are common. It is recommended therefore that the model going forwards incorporates a cookie. It should be noted however that this identifying information does not need to necessarily encapsulate all of a person's viewing as the model is capable of assigning multiple cookies to a single panellist.

Of the models that do incorporate cookies/identifying information the abstract models perform better than the "real" models. This is because the real models are unable to account for the variation in campaigns in the same way that the abstract models are. This is especially apparent in the online element where reach is unacceptably inaccurate when a cookie isn't used, but there is also a clear benefit to using cookies for the TV element as well (a "cookie" for TV would be the viewing activity of a panellist). For the TV element ideally the viewing activity across all channels for a panellist ("Total TV Cookie") would be assigned to the same panellist. Linking only viewing on a single channel ("Cookie by Channel") also performs well, however some calibration would then be necessary to meet a TAM Panel Gold Standard. For these reasons, analysis determined that Abstract with Total TV Cookies is the best performing method, although Abstract with Cookies by Channel also performed well.

It should be noted that adding multiple website campaigns to the model may present additional challenges that could not be tested with the current data available. For this reason additional testing would be sensible in the future once a specialised panel has been created as this would provide a more robust dataset to test against. Despite this however the performance of the abstract cookie models on the data provided is good so this proviso should not be viewed as a reason not to approve the models and we would there recommend moving forward with this approach. In particular, the findings here point towards the Abstract Cookie models being suitable for modelling the reach of TV + Online campaigns and the duplication between them. However, it should be noted that this evaluation was on a small scale and it is unclear how this methodology might hold up when faced with a sterner test of increased media channels. For this reason, the other model scenarios cannot be discounted as viable alternatives although given the results observed it is accepted that further refinement may be needed for these.

Appendix A: Technical Description of VID Model Work

The abstract VID model training relies on mathematical optimisation routines; the algorithms are outlined in the WFA technical papers, but several parameters and also some operational requirements are not specified. Furthermore, some alterations were required to bring TV viewing into the model training process. This document provides a description of the process and describes some of the decisions that were made in training the VID model for the POC.

It is worth noting here that the term ‘cookie’ is used to refer to any persistent identifier, including registration IDs.

Website VID Model Training

For each campaign in the training period, we need aggregate information for each contributing website and for the campaign as a whole, taken from the panel. For each website, we need the count of registration IDs/cookies that have viewed at least one spot. For the campaign as a whole, we require the overall reach. In practice, each publisher will have access only to the overall number of cookies that have seen a campaign for their own content; they will not know to how many people this corresponds, nor will they have any information relating to other publishers.

The goal of the model training is to obtain the optimal set of model parameters. In this case, optimal means that the model minimises the difference between the modelled and target (i.e. panel) reach across all campaigns in the input dataset. A complete description of the optimisation routines is unnecessary for this document, but it is instructive to understand the calculation of the reach estimates. A key consideration for model training is that the underlying theory aligns with the intended application, i.e. the allocation process.

This is best illustrated by an example. With a single publisher initially, let’s assume that the VID model parameters are as follows:

Group	Proportion	Rate	Proportion*Rate
1	0.2	0.5	0.1
2	0.3	1.0	0.3
3	0.5	1.2	0.6

There are three usage groups in the model. The proportion column gives the proportion of the population that falls into each group, for example 20% of the population is in usage group 1. The total number of VIDs will equal the population. The rate columns tell us whether the group has heavy or light users of the website, for example group 1 are light users.

In the allocation process, a publisher considers each cookie/registration ID in turn. To select a VID, the first step is to select one of the groups. The final column indicates the probability of each group being selected: there's a 10% chance of being allocated to group 1, 30% to group 2, and 60% to group 3. Within a group, a VID is selected at random. All viewing from a single cookie goes to the same VID, but due to this randomness, some VIDs will get viewing from multiple cookies: this is intentional, to account for real phenomena, for example people using multiple devices/browsers.

The model training is based on aggregate panel data. We know the total number of cookies (weighted according to the panel weights) that have at least one campaign spot, and we also know the total reach. Let's assume that for a given campaign, the cookies per person figure (i.e. the cookie count divided by the total population) is 0.2, and that the reach is 15%, or 0.15 as a proportion. We need a formula that can estimate the number of VIDs that we expect to be allocated at least 1 cookie (i.e. the reach) if we actually went through the allocation process as described. This formula is given by:

$$Modelled\ Reach = \sum_{Usage\ groups} prop * (1 - e^{-rate*cookies\ per\ person})$$

While there will be some variation when actual allocations are performed, the differences are small and unbiased.

For our model parameters, the first group's reach would be:

$$Modelled\ Reach\ group\ 1 = (1 - e^{-0.5*0.2}) = 0.095$$

This figure represents the reach proportion within the group. The same calculation is done for groups 2 and 3, and the overall reach is obtained by multiplying these figures by the respective proportions of the population:

$$Overall\ Modelled\ Reach = 0.2 * (1 - e^{-0.5*0.2}) + 0.3 * (1 - e^{-1*0.2}) + 0.5 * (1 - e^{-1.2*0.2}) = 0.18$$

So the modelled reach is 18%, which is an overestimate compared to the panel reach target of 15%. Different sets of model parameters may produce better reach estimates. This covers the single publisher situation, which essentially aims to model the cookies to people relationships. The next step is to consider the multi-publisher situation, where the model must account for correlations between publishers.

The table below contains the VID model parameters, with an additional column for the rates of the second publisher:

Group	Proportion	Rate Publisher 1	Rate Publisher 2
1	0.2	0.5	0.5
2	0.3	1.0	1.5
3	0.5	1.2	0.9

Again, we have the overall reach and the cookie counts for both websites from the panel. Let's take the reach to be 0.3, the website 1 cookies per person to be 0.2, and the website 2 cookies per person to be 0.15.

The expected modelled reach is calculated as follows:

$$Modelled\ Reach = \sum_{Usage\ groups} prop * (1 - e^{-rate\ 1*CPP\ 1} * e^{-rate\ 2*CPP\ 2})$$

For group 1, the contribution is:

$$Modelled\ Reach\ group\ 1 = 1 - e^{-0.5*0.2} * e^{-0.5*0.15} = 0.16$$

And the overall modelled reach is given by:

$$Overall\ Modelled\ Reach = 0.2 * 0.16 + 0.3 * 0.35 + 0.5 * 0.31 = 0.29$$

Recalling that the actual panel reach is 0.3 for this campaign, we see that this set of model parameters will closely recreate the panel reach but is underestimating slightly. These figures must be calculated for all campaigns. The goal is to minimise the sum of differences between modelled and panel reach across all campaigns.

The optimisation routine operates by considering a set of rates, and then determining the best possible alpha values. This formulation is a standard optimisation problem, with many well-known solving algorithms.

Combinations of Channels

The previous section describes the general approach for model training, but as described, this only accounts for overall campaign reach and individual publisher reach. It's important that the model can also deliver accurate reach results for different combinations of publishers.

In practice, we will likely have more than 2 publishers contributing to each campaign. We could introduce targets for any combination of publishers. For example, if we have six publishers, we will have the overall reach as a target, and also the 6 individual publisher targets, but we could also consider all pairs of publishers too. This allows us to control the modelled reach to each pair of publishers and ensure that these results are close to the panel numbers. Taking this further, we can consider all sets of three publishers too, and all sets of four and five publishers. While it is straightforward to extend the theory as described above to these situations, it does introduce a large number of targets. With six publishers, there are 15 distinct pairs, and 20 distinct sets of three publishers. More generally, with n publishers, there are 2^n-1 combinations of publishers when considering all different set sizes. Clearly, this cannot be extended when the publisher count grows as it will become computationally infeasible.

This table shows some of the possibilities in terms of which combinations to include:

Approach	Targets	Target Count		
		6 publishers	10 publishers	n publishers
1	(1, n)	7	11	$n+1$
2	(1, n - 1, n)	13	21	$2n+1$
3	(1, 2, n - 1, n)	28	66	$\frac{1}{2}n^2+\frac{3}{2}n+1$
4	All	63	1023	2^n-1

Approach 2 in this table includes individual publisher targets and also all combinations of n-1 publishers. The latter targets are useful as they are equivalent to having incremental reach targets for each publisher on all others. As can clearly be seen in approach 4, the number of targets per campaign becomes very large even in a 10-publisher situation.

For the proof of concept, it was decided that approach 3 was an acceptable trade-off between practicality and controlling targets. It was found that the modelled reach for sets of 3 and 4 publishers, despite not being controlled directly, closely matched the panel reach figures.

Extension to TV

For a website publisher, the modelling input is the count of cookies that the publisher holds, and there are two goals of the modelling. The first goal is to create a cookie to individual model. The second is to ensure that any correlations between the publisher and others in the model training are preserved. For TV, the situation is slightly different. The input is the reach, rather than the cookie count: the modelling therefore only needs to account for correlations with online publishers, as the reach is already known.

Before considering the training algorithms, it's worth noting that we can either consider TV as a single entity and take the total TV reach to a campaign as the input, or we can fragment TV by channel, and take channel reach as the input (essentially treating each channel as if it were an individual publisher). The former preserves the true total TV reach but has the potential to be blunt as it cannot differentiate between two campaigns with the same overall reach but different channel composition (and therefore perhaps different correlations with websites). The latter allows for complex correlations to be accounted for but does not perfectly preserve total TV reach; it also increases the number of targets per campaign considerably. While both options were tested, it is simpler to consider the case of total TV for our purposes here.

Let's assume that we have one online publisher and total TV only. As before, we consider a single campaign. From the panel, we need the overall reach, the total TV reach, the website cookie count, and the website reach. As proportions/per person figures, let's take these to be 0.4, 0.25, 0.2, and 0.16 respectively.

The format of the model is as before:

Group	Proportion	Rate Website	Rate Total TV
1	0.2	0.5	0.5
2	0.3	1.0	1.5
3	0.5	1.2	0.9

The allocation process for TV is somewhat different to the website version described previously. Instead of allocating a cookie and its viewing to a VID, we allocate a panellist's viewing to multiple VIDs, in order to match their panel weight, e.g. if a panellist has a weight of 5,000, we must select 5,000 VIDs. In this case, we do not allow a VID to receive viewing from multiple panellists, as this would deflate the reach. The selection process again uses a usage group's proportion multiplied by its rate, so the first group has a 10% chance of selection, the second a 45% chance, and the third a 45% chance also. As before, a VID is selected at random, but only from those that haven't already been assigned any viewing.

The next step is to establish formulae to estimate the number of VIDs allocated viewing in each group, given the input aggregate reach figure from the panel:

$$Modelled\ Reach = \sum\nolimits_{Usage\ groups} reach\,prop * rate$$

So for group 1, this is:

$$Modelled\ Reach\ group\ 1 = 0.25 * 0.5 = 0.125$$

The overall reach is given by the sum-product across the groups:

$$Overall\ Modelled\ Reach = 0.2 * 0.125 + 0.3 * 0.375 + 0.5 * 0.225 = 0.25$$

Unsurprisingly, this is the same as the input reach for total TV, so this is perfectly preserved as expected. More useful is expected reach across TV and the website:

$$Modelled\ Reach = \sum Prop * (1 - e^{-rate\ website * CPP} * (1 - rate\ TV * reach\,prop\ TV))$$

So for group 1, this is:

$$Modelled\ Reach\ group\ 1 = 1 - e^{-0.5*0.2} * (1 - 0.5 * 0.25) = 0.21$$

And the overall modelled reach is obtained by calculating the sum-product of these values and their respective population proportions:

$$Overall\ Modelled\ Reach = 0.2 * 0.21 + 0.3 * 0.49 + 0.5 * 0.39 = 0.38$$

The expected modelled reach is 38%, which is an underestimate compared to the panel reach of 40%. The model can also have an individual reach target for the website, but there is no need to have a reach target for total TV on its own as the approach guarantees that this is preserved.

It's worth noting that there are other approaches that can be taken when integrating TV data into the VID framework. Large-scale data sources such as RPD and STB data can be used, but our primary goal is to ensure that TAM Panel data can be brought into the system. Even when using panel data as input, there are other possible approaches, which are not ruled out for future use.

Demographics

The proof of concept (POC) limited the number of demographic groups to 6: a male/female split, with 3 age groups (18-34, 35-54, 55+). Note that for simplicity weight was not used in the POC. Several different approaches are possible when it comes to VID modelling with multiple demographic groups; these were investigated in the proof of concept:

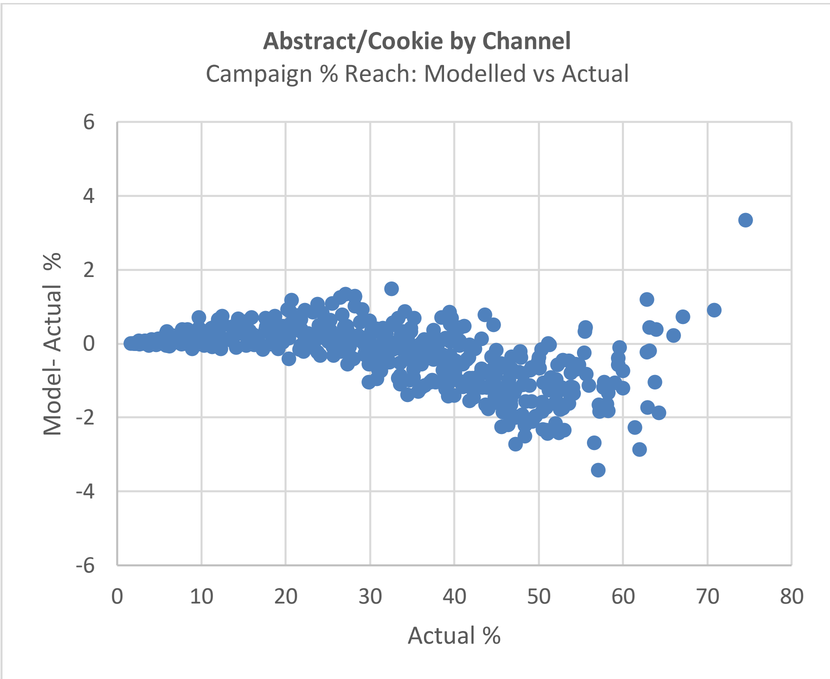
- One set of model parameters per demographic, each trained independently
- One set of model parameters per demographic, but trained simultaneously so that all adults reach targets may also be controlled
- The same set of model parameters used for all demographics

The third option is fairly straightforward to implement, but in practice, there are significantly different correlations and relationships observed for different demographic groups, so this approach does not produce good results. The second option is perhaps the most appealing, as it ensures the higher-level reach targets are controlled directly; however, it does increase the complexity of the model training considerably, and the number of targets, creating computational challenges. Option 1 is again more straightforward but allows total reach to float. In practice however, it was found that all adults reach results are well preserved using the option 1 approach, meaning that the complexity of option 2 is unnecessary. For the proof of concept, it was therefore option 1 that was implemented.

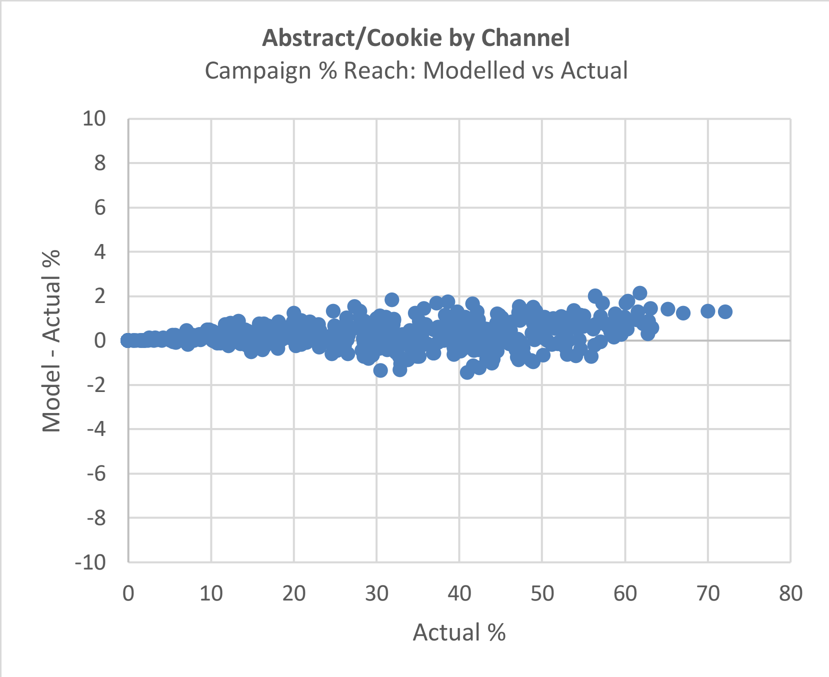
Appendix B: Abstract Cookies by Channel Improvement

In the Stage 1 the performance of TV only campaigns was assessed; whilst this research was repeated for Stage 2 it has not been reproduced in this document as the analysis was very similar to what was found previously, indicating that the addition of an online channel has had very little impact on the performance of the model when predicting TV campaigns. There is one exception to this - the abstract cookie by channel method; this is due to an improvement to the model that was made following the findings of Stage 1 rather than being a consequence of adding online in. This improvement is worth analysing though so the results are detailed below.

As a reminder this was the Stage 1 model – model (diff) vs actual reach differences for Total TV:

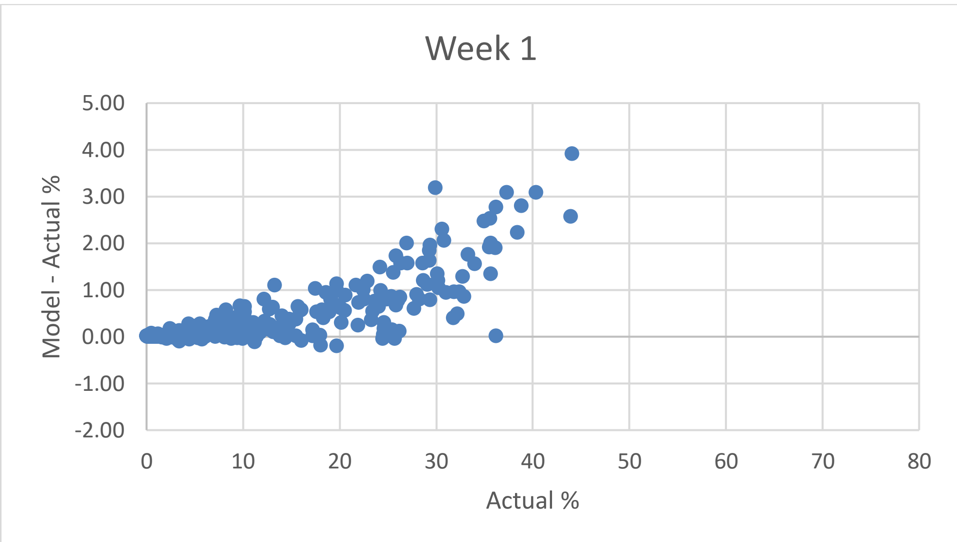


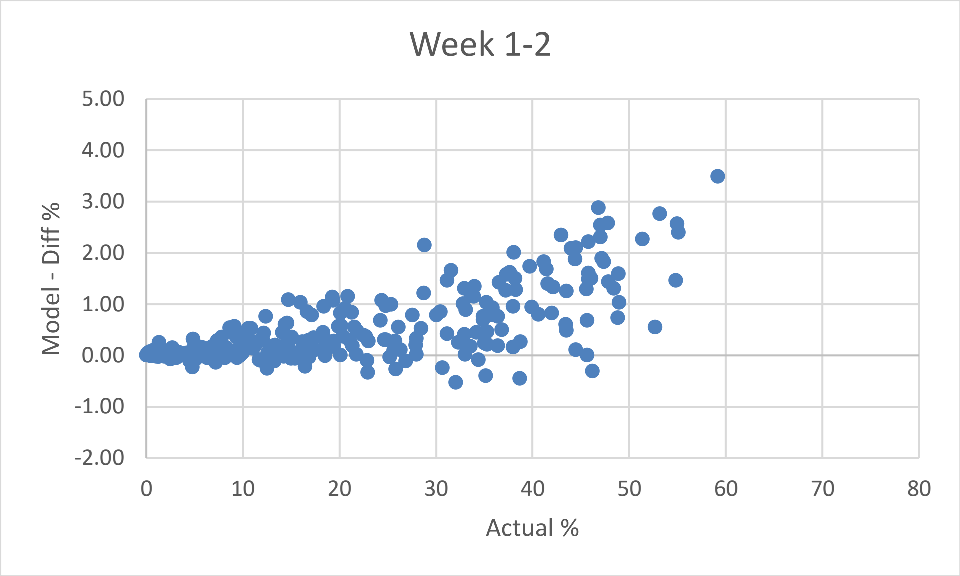
The equivalent Total TV reach plot for Stage 2 is as follows:

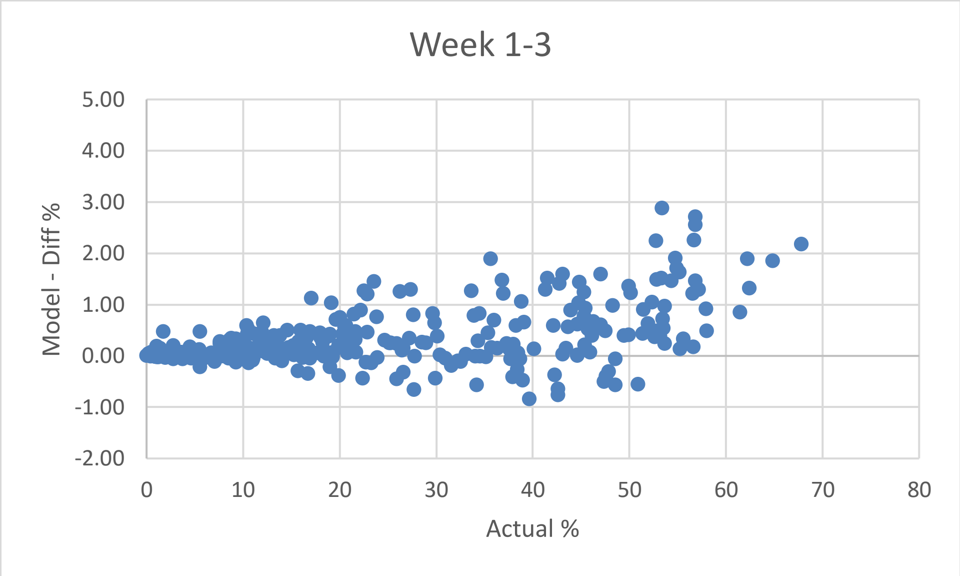


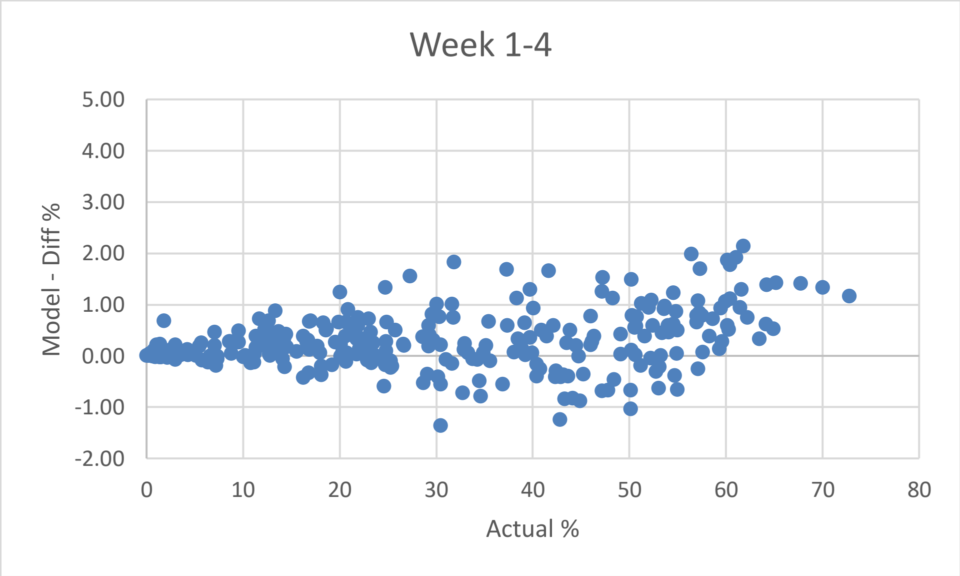
This indicates that the improvements made to the abstract rates have reduced the clear tendency to underestimate that was seen in Stage 1.

It should be noted however that this improvement applies only in campaigns where at least one abstract Dirac group is completely filled; as a result the larger campaigns in the first week or two have a tendency to overestimate before correcting in later weeks:




Week 1-2
Model - Diff %
Actual %
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
0
10
20
30
40
50
60
70
80


Week 1-3
Model - Diff %
Actual %
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
0
10
20
30
40
50
60
70
80


Week 1-4
Model - Diff %
Actual %
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
0
10
20
30
40
50
60
70
80

This identifies a slight caveat to the improvement, it seems that there is a tendency to overestimate the reach in the first week for the largest campaigns which is gradually corrected over time. By week 3 the plot begins to look sensible so it is possible that this is an issue that only affects the first week or two. This idea will not be further explored in the report but may be worth examining in the future.