THE REINVENTION OF JICREG

Keith Donaldson, JICREG
Kirsty Ford, RSMB

1. Background

JICREG has provided regional media audience data since 1990. It has a solid and credible reputation for print measurement as well as an innovative approach to the multi-media marketplace (Locally Connected).

JICREG's membership comprises 30 publishers (representing some 700 titles) and its data is purchased by all of the leading communication agencies.

The basic concept of JICREG is that it applies Readers Per Copy (RPC) estimates to copy distribution for individual newspapers at a postcode sector level (there are around 9,000 postcode sectors in the Great Britain). These can be summed to produce estimates of readership for any regional geography.

In the past the RPC figure came from one of two sources. Where individual titles had commissioned JICREG approved research then those titles took their RPC's from their research. For all other titles the researched titles data was pooled and used to produce RPC models that could be applied to titles that hadn't commissioned research.

The methodology for digital audiences used the same concept applying adults per browser figure to audited website figures and using either researched or modelled data to estimate reach as well as net reach across platforms.

For various reasons, not least the commercial environment for regional newspapers the volume of surveys being commissioned declined and those that were carrying out research weren't representative of all regionals, thus making it difficult to update the models.

In 2013 after consulting with publishers the board decided that an alternative solution was needed to maintain the credibility of JICREG. Discussions with research agencies provided a number of solutions but from a cost point of view these were unacceptable and at that stage it was decided to approach the NRS about including local newspaper readership on the NRS questionnaire.

How the NRS integrated local newspaper readership

The postcode sector database of local newspaper distribution allows the interviewer to prompt for readership of titles that are relevant for that particular geographic sector.

With many titles having very tight geographical circulation footprints and circulation figures ranging from 1,164 to 230,674 it is unrealistic to expect the data to allow stand alone un-modelled readership estimates by title.

What the NRS data has allowed is the construction of models based on all JICREG titles as a whole rather than the limited number who carried out their own research.

Data collection via NRS

The first stage was to create a database of JICREG titles, so that readership questions can be asked for the specific titles available in the area of each individual NRS interviewer assignment.

A 12 month readership filter is asked, followed by recency and frequency of readership for each title read in the last 12 months. Regional dailies are prompted on regional EML screens within the main NRS media list. Each screen is used to prompt for six titles. There are 10 different regional screens, titles are prompted in their core region and any neighbouring regions where they may have some penetration.

Regional weeklies appear after the main EML sequence has been completed, this allows customised prompting of the titles relevant for each specific NRS assignment.

The central file including colour mastheads is updated twice a year, when the circulation data is updated.

Once the NRS interviewers are issued with pre-selected addresses to contact, the selected postcodes are checked against the database in order to identify weekly titles with a penetration of over 5% in any of the postcodes selected for that particular assignment.

Using the data to build models

After a tendering process RSMB were appointed to use the data to build the new models for daily, paid and free weekly titles.

2. The modelling process

The following sections of the paper detail the different stages of the modelling process RSMB undertook to produce local newspaper readership models from the NRS survey. A number of models exist within the JICREG data, namely Readers per Copy (RPC), demographic, duplication and CUME models, however this paper focuses on the primary model to estimate Readers per Copy. At the time of writing this paper we are currently finalising the models, however this has been an iterative process and a number of models have been produced and adapted in order to reach the final models; the same process was adopted for each model.

2.1 Consideration of inputs into the models

The following datasets were available as inputs into the modelling:

- NRS data: readership and respondent data received every quarter
- News Media Association database: circulations and title information by postcode sector received bi-annually
- JICPOPS: postcode sector populations by sex, age and social grade received annually
- Census 2011 data: postcode sector data for population density, English proficiency and mode of transport to work
- Commitment to local area: whether a title is linked to a large sports team, updated yearly

The first stage was to explore the inclusion of explanatory variables into the model. The following conditions needed to be satisfied to incorporate a variable into the models:

a) available for all JICREG titles,
b) available at title or postcode sector level
c) relevant in regards to the time period the data was collected, i.e. the data was not too old
d) accurate
e) relatively easy to obtain and update

The above constraints meant that some variables could not be considered for inclusion. In particular, a number of title variables were excluded due to missing or out-of-date data, e.g. number of columns and cover price.

In the past, variables such as pagination and Advertising/Editorial ratio have proved difficult to collect accurately across all titles. These data has since become available from Ebiquity and we are currently cleaning the data with the intention of adding these variables into the final models.

Information on mode of transport to work, population density and English proficiency within the Census 2011 data was also explored. The main issue with census data is the survey was carried out four years ago in 2011 and new updated data may not be available for a number of years. Investigating an area’s English proficiency was of particular interest as it was believed that newspaper readerships would be lower in areas with large non-English speaking communities.

When exploring the inclusion of demographic variables, these needed to be available at postcode sector level, e.g. the percentage of a postcode sector aged between 15-24 years. Even though there were many demographic variables in the NRS data at respondent level, the models would be applied to postcode sector information to create the readership estimates. This limited the amount of possible demographic variables for modelling as only age, sex, social class and adults per household were available at postcode sector level.

The full list of variables considered for modelling are detailed in Appendix A.

2.2 Data Cleaning

A large part of the modelling process, and indeed a very important one, was 'cleaning' the separate datasets to ensure the data was correct before proceeding with the modelling. This was a time consuming process which included the following steps:

- Any explanatory variables with incomplete or inaccurate data were removed from the modelling, although efforts were made to clean the data for variables with only a small amount of missing data.

The following data cleaning was carried out on the NRS data:

- Whilst the regional weeklies were surveyed according to their distribution postal areas, the regional dailies were surveyed within relevant Government Office Regions and the Sundays across all regions. For each Daily and Sunday title, records were removed where a respondent's postcode did not fall into the title's circulation area.

- Some titles were not separated into editions (e.g. Mon-Thurs and Fri-Sat). The 'Day of Interview' field was used to separate the readerships into the appropriate editions.

- Some titles had conflicting paper IDs between the NRS data and the JICREG boundary files (a bi-annual file containing title circulations by postcode sector). The paper IDs were cleaned to ensure the correct paper IDs were matched between the two datasets.

- NRS records were removed from the modelling dataset where the title in question was not found on the JICREG boundary files. These titles were either closed or had no audited circulation figure.

- For the purpose of modelling, any mixed distributions, either at overall newspaper type or at postcode sector level, were excluded from the modelling datasets. The distribution of a title in a particular postcode sector was classed as 'Paid', for example, if its overall newspaper type was paid and at least 90% of the circulation in that particular postcode sector was also paid. The same applied for free and pickup distributions. Any distributions not falling into 'Paid', 'Free' or 'Pickup' were classed as a 'mixed' distribution and removed from the modelling datasets.

2.3 Linking datasets

Once the separate datasets had been cleaned, the next stage was to link the separate datasets listed in section 2.1.

Each respondent in the NRS survey had an individual record for each regional title they were surveyed (up to 8 title records), and recorded whether they had read the newspaper yesterday. Each record contained demographics of the respondent, the paper ID of the title being surveyed and the respondent's postcode sector. The paper IDs were used to link title information to the record and the postcode sector was used to link census and population data.

Since the new readership estimates are based on NRS data collated over a two year period, accounting for changes in distribution, circulation and title variables over the time frame was imperative to producing reliable and unbiased estimates. To account for these changes, each individual readership response was linked back to title and population data as close as possible to the time of survey.

2.4. Limitations of the data with regards to the modelling

Another important part of the modelling process was to identify the limitations of the data and what impact this could have on the success of the modelling.

2.4.1 Explanatory variables and distinguishing between titles

As mentioned in section 2.1, a number of variables did not satisfy the conditions required to be included in the modelling. The main limitation was the availability and reliability of data and some of the title data was either out of date or contained too many missing values.

The success of the modelling relied on the accuracy of title data and being able to include variables in the models which define a newspaper's quality. In particular we were interested in being able to discriminate between two similar titles in a given area; i.e. differentiating between a 'good' quality (higher Readers Per Copy) and 'poor' quality publication. The aim of the modelling was to find variables that explain as much of the variation as possible in readership between different newspapers. The models would not be expected to explain every variation between titles as there are other factors affecting readership that we cannot quantify. However the aim was to achieve fair and unbiased estimates of title readerships.

2.4.2 Sample sizes

The next stage was to assess the amount of available NRS data for each title. It was anticipated that the main constraint on the new readership models would be that the NRS survey would collate only a limited number of NRS records for each individual title. The intention was to model the data rather than create stand-alone un-modelled readership estimates directly from the NRS. There were two possible ways to model the NRS data: if sample sizes allowed the title data could be aggregated and modelled, if not, the model would be based on a dataset comprising individual respondent records (see section 2.5).

Table 1 below provides a summary of the final modelling datasets for each newspaper type after data cleaning.

The average AIRs per title confirms there was not enough data to calculate models on individual title data. Even though the Dailies had a reasonable number of interviews per title, there were only five titles with over 175 unweighted average issue readerships (AIR). 175 is the NRS's requirement for the publication of an estimate. While we are not publishing estimates directly from the NRS data, this gives an indicator of the sample sizes required to obtain reliable title estimates. The same was true for Sunday titles with only two titles having over 175 unweighted average issue readerships. However even with sufficient data for the Sunday titles, we would have been unable to model just eight titles.

Table 1: Modelling dataset for each newspaper type, after data cleaning (Q2 2013 to Q4 2014)

	Number of NRS records	Number of titles with NRS data	Average interviews per title	Average AIRs per title
Dailies	85,390	72	1,186	85
Paid weeklies	36,547	315	116	32
Free weeklies/Sundays	21,428	190	113	40
Paid Sundays	40,777	8	5,097	130
Free Pick up	878	12	73	13

The Paid Sunday titles accumulated a large enough sample to be able to produce a separate model for paid Sundays which had previously been modelled with the paid weeklies. There was not enough data however to produce a free pickup model therefore any free pickup distributions will continue to be assigned a Readers per Copy (RPC) of 1.

2.5 Developing the Readership Models

As mentioned in the previous section, there was not enough data to model aggregated title data. This was not necessarily the best modelling approach since title data would have been averaged over one to two years of survey data resulting in a loss of granularity. Modelling individual respondent records had the following advantages:

- More data was available to feed into the models, e.g. for Dailies 85,390 records as opposed to 72 titles
- Each readership value is related back to title and population information at the point of survey (or as close as possible to using NMA / JICPOPS data)
- Small vs large sample titles are naturally given an appropriate level of importance in the overall analysis
- Paid Sundays were able to be modelled separately from paid weeklies.

Each individual NRS readership record had a binary response, i.e. a respondent either read the newspaper (1= Success) or not (0= Failure). These individual records were modelled, based on the title and area variables linked to each readership record, to produce models to estimate the probability of reading a newspaper, i.e. AIR %.

The probability of reading a newspaper (AIR %) was modelled rather than readers per copy, however the two are linked by the simple equation:

$$\textit{Readers per copy (RPC)} = \textit{AIR \%} \times \frac{\textit{Population in circ area}}{\textit{Circulation}} \qquad \textit{where population and circulation are known quantities}$$

Since probability is restricted to values between 0 and 1 and a linear model can give rise to any value between $-\infty$ and $+\infty$, we used **multiple logistic regression**.
The aim was to produce four separate models for the different newspaper types: Dailies, Paid weeklies, Free weeklies and Paid Sundays. The free Sundays were modelled with the free weeklies since there was not enough data for a separate model.

The table below provides the overall unweighted AIR% (sum of AIR ÷ number of interviews) for each newspaper type. These AIR percentages for the NRS data are broadly in line with current JICREG data which is encouraging. The variability in AIR% between the separate newspaper types highlights the need for separate models, particularly being able to separate paid weeklies and Sundays.

Table 2: Final data set for modelling after data cleaning
(Q2 2013 to Q4 2014)

		Total
Total Sample		**62,759**
Dailies	Total records	85,390
	Read Past Year %	27%
	AIR %	7%
Paid Weeklies	Total records	36,547
	Read Past Year %	50%
	AIR %	27%
Free Weeklies	Total records	21,428
	Read Past Year %	48%
	AIR %	35%
Paid Sundays	Total records	40,777
	Read Past Year %	11%
	AIR %	5%

The total unweighted sample before data cleaning was 62,759; approximately 9,000 individuals per quarter. Each individual had up to eight ‘records’, one for each title they were surveyed.

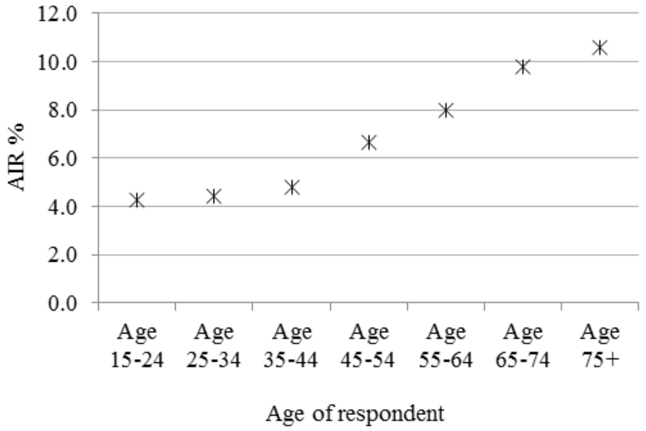
Descriptive statistics were carried out on the data to gain an understanding of the distribution of data for each variable and the relationship between readerships and the explanatory variables. Any categorical variables with small sample sizes were either re-grouped or removed. Quantitative variables which showed little variability were also removed from further analysis, for example male/female populations showed little variability across postcode sectors.

For each categorical variable, the AIR% was calculated for each group and for continuous variables, the values were grouped into intervals and the AIR% calculated for each interval. Charts were used to plot the average AIR% for each grouping, to highlight any possible relationships in the data.

As mentioned in section 2.1, only postcode sector level data could be included in the final models as the final estimates would be based on title and population data by postcode sector. Some demographic variables were available at respondent level from the NRS and at postcode sector level, such as gender, age, social class and adults per household. The relationships between readership and the demographics at respondent level were compared against the relationships at postcode sector level.

For example, for the Daily titles it was evident from the NRS data there is a strong relationship between AIR% and a respondent’s age as shown in Figure 1 below. The AIR% was lower for younger respondents compared to those over 45 years of age.

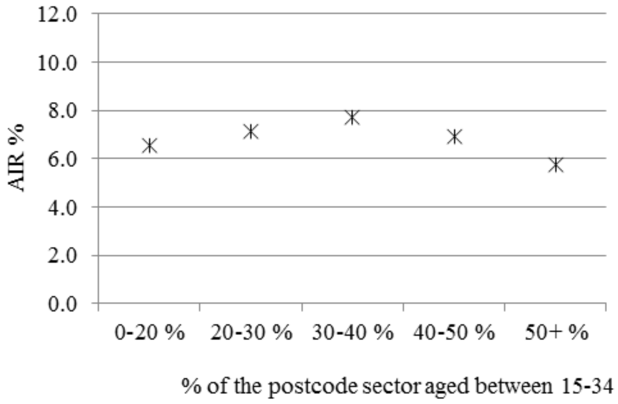
Figure 1: AIR% by respondents' age group



At postcode sector level, age information was available in the form of percentage profiles within a postcode sector, for example the percentage of the respondents' postcode sector aged between 15 and 34 years. Based on Figure 1 above, we would expect a postcode sector with a larger proportion of 15-34 year olds to have a lower readership.

Figure 2 below plots the AIR% for postcodes with varying percentages aged between 15-34 years. Those postcodes with more than 50% of the population aged between 15 and 34 indicated a lower AIR% on average.

Figure 2: AIR% by percentage of the respondents' postcode sector aged between 15 and 34 years



The distributions for postcode sector variables tended to be less variable than at respondent level data. There were some instances where demographic variables were significant at respondent level but not postcode sector level. Significance at postcode sector level also relied on variability in the demographic profiles between postcode sectors. For example, if all postcode sectors surveyed had a percentage of 15-34 year olds between 30-40% then the variable would unlikely to be statistically significant.

Univariable Analyses were performed to test the association of each explanatory variable with the outcome, i.e. AIR%. This was to help shortlist the variables for multivariable analysis. Any variables which did not show any significant association with AIR% on their own were unlikely to be associated with the outcome after adjusting for other variables.

Correlations were checked between explanatory variables. For any pair of highly correlated variables (correlation over 0.9), e.g. circulation density and household penetration, the variable which had the least reduction on the Residual Deviance was excluded from the model.
After shortlisting the variables from the univariate analysis and inspection of descriptive statistics, the remaining continuous variables were considered for transformation to ensure the relationship between the variables and AIR% was approximately linear.

Multiple logistic regression was used to model AIR% in the statistical analysis package 'R'. Backwards stepwise regression was used to obtain a final model, with the variables included being statistically significant at the 5% level. Models with a smaller residual deviance, compared to the null deviance, were favoured. The coefficients for the variables in the model were checked to make sure they were reasonable, e.g. we would expect an increase in circulation penetration in a postcode sector to increase an individual's chance of reading a newspaper, and therefore we would expect the coefficient for circulation density to be positive. The final models were checked against the descriptive statistics and univariate analyses to ensure similar relationships with readership were observed and any differences were investigated.

As in the current JICREG models, the main significant factor in all the models was circulation penetration or household penetration. As with previous modelling exercises it was difficult to find significant variables which improved the models and had a believable effect on AIR%.

The main variables used in the models were: household penetration / circulation density, age % populations, social grade % populations, whether a postcode sector contains over 5% of non-English speakers and population density. It is also anticipated to add Publication SCC and advertising ratio into the models.

2.6. Calculating Readership Estimates

Once the final models were produced, estimates were calculated for each title using the most up-to-date JICREG boundary file and area/census information.

For each individual title, the AIR% was calculated for each postcode sector in the title's distribution area by applying the appropriate JICREG model for the type of newspaper. Readerships and Readers per Copy (RPC) were then calculated using population and circulation figures. The estimates at postcode sector level were then aggregated to produce an overall title readership and Readers per Copy.

The models were found to overestimate title readerships in postcode sectors where the circulation was very small (less than 100) and circulation density was less than one percent. As a result a cap was applied to postcode sector readerships in these instances.

For mixed distribution titles, there was too much variability in the mix of distribution types to produce a separate model for combined distribution titles. Therefore the separate models for each distribution type were applied to the different circulation proportions (e.g. paid / free) at postcode sector level.

2.7. Validation of models

The models were validated and checked for accuracy in a number of ways, as detailed below.

a) Statistical model validation including checking model deviance against the null deviance and comparing estimated probabilities with actuals

b) Respondent demographic variables within the on NRS were used to validate relationships between postcode sector demographic variables and AIR%.

c) The final models were checked against the descriptive statistics and univariate analyses to ensure similar relationships with readership were observed.

d) Comparison of average RPC per newspaper type with current JICREG data

The average RPC from the new NRS estimates were compared against the current JICREG data. Since the NRS research was carried out face-to-face, comparisons could only be made with other face-to-face research, not telephone research. It was noted across all newspaper types, the RPC values for telephone research were on average higher and more variable than face-to-face or modelled.

As an example, the Paid weekly comparisons are provided in Table 3 below.
It is important to note the averages of the face-to-face research are based on a smaller number of titles in comparison to the new model averages. We would expect other face-to-face research to possibly exceed the RPC ranges of the face-to-face research. Also the new models are based on 2013/2014 data and the majority of the current research was carried out between 2010 and 2012.

The new modelled RPCs have a larger range of values than the current models and an average RPC more in line with the current face-to-face research.

Table 3: Paid Weeklies Readers Per Copy by modelled, Face-to-Face and telephone research

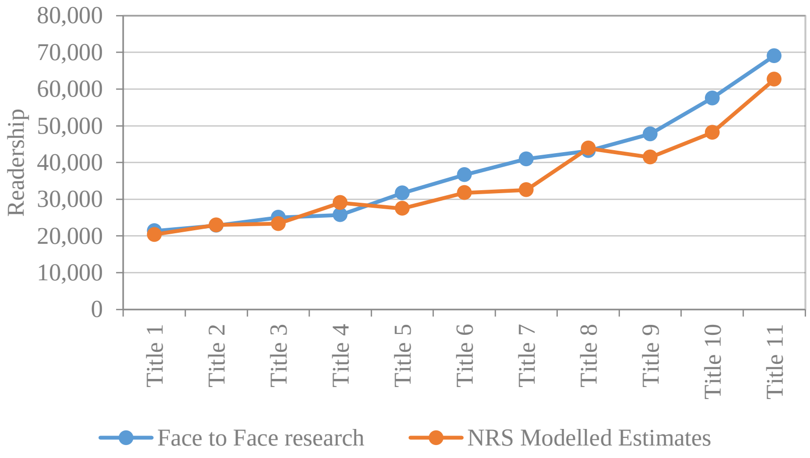
	Lowest RPC	Highest RPC	Average	Number of titles
All titles	1.9	7.1	2.9	321
Modelled	2.2	3.5	2.8	263
Face to Face Research	**2.5**	**4.2**	**3.3**	**11**
Telephone Research	1.9	7.1	3.5	47
New NRS models	**1.6**	**5.7**	**3.3**	**321**

e) Comparison of models with current JICREG face-to-face research including titles already measured on the NRS

The new model title estimates were compared to the current face-to-face research. The following charts show Paid Weeklies face-to-face research versus the new modelled figures. Each point represents a title, the blue dot representing the current Face-to-Face research and the orange dot representing the new modelled estimates. The chart has been sorted into ascending order for clarity and lines have been added to highlight the trend through the data points.

As mentioned in the previous section the current face-to-face research and the NRS estimates have been compiled from data surveyed at different time points so we would not expect the readerships to necessarily be the same, however we would expect the general ranking of the titles to be similar. Any changes in title distributions were taken into consideration when comparing research figures.

Figure 3: Paid weeklies: Comparison of **Title readerships** between Face-to-Face research and the NRS estimates



The general ranking of titles' readerships, AIR % and RPCs were consistent between the current research and the new modelled estimates.

Figure 4: Paid weeklies: Comparison of **Title AIR%** between Face-to-Face research and the NRS estimates

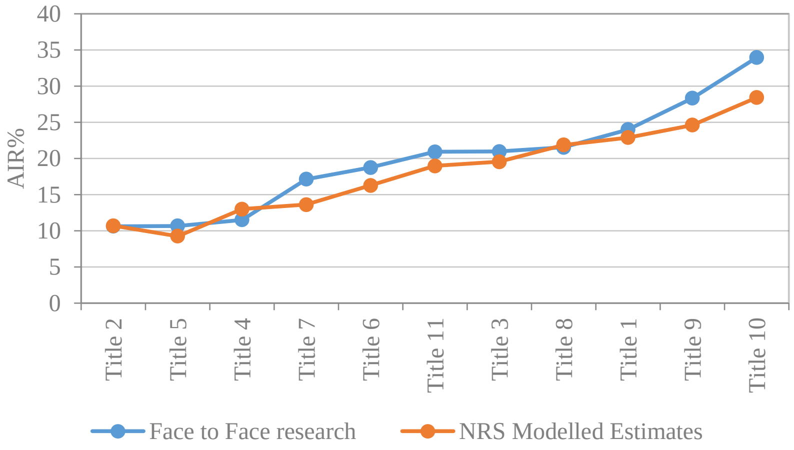
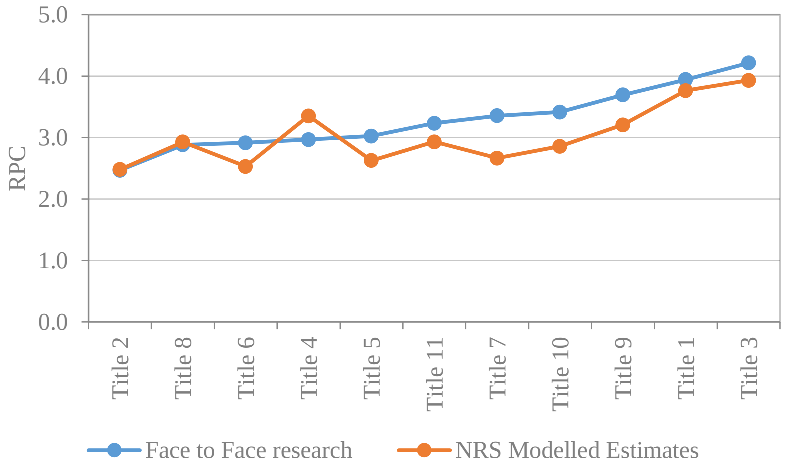


Figure 5: Paid weeklies: Comparison of **Title Readers Per Copy** between Face-to-Face research and the NRS estimates

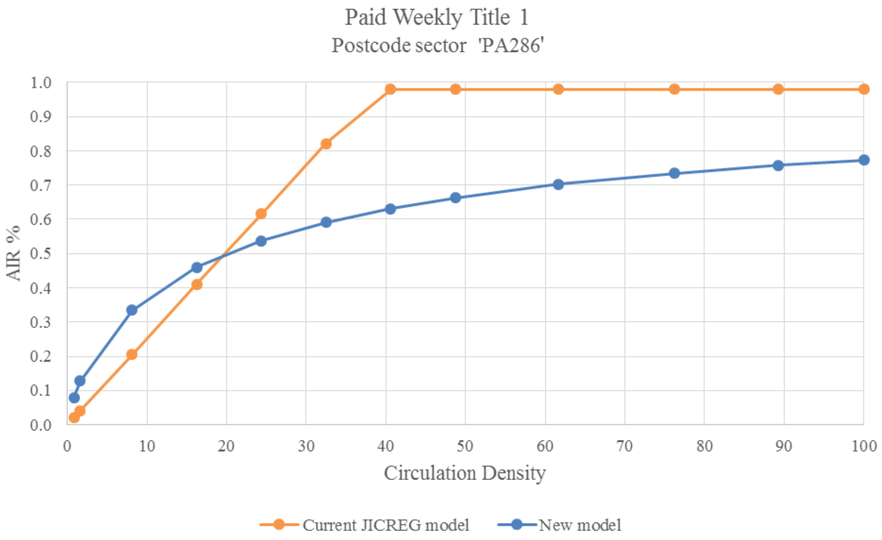


f) Larger differences in readership between the current JICREG models and the new NRS model estimates were investigated, particularly for the Paid weekly titles.

The main reason was found to be a difference in the relationship between circulation density and AIR% between the two models.

To investigate the effect of circulation density on readership estimates, one postcode sector was analysed within the circulation area for one title (Title 1). This ensured the values for all the other variables stayed constant in the model. The amount of circulation was then adjusted and the AIR% were calculated for each case. This was carried out for both the current JICREG model and the new model. The results can be seen in Figure 6 below.

Figure 6: Paid weeklies - Comparison of the relationship between circulation density and AIR% between the current JICREG model and the new model



For the current JICREG model, the circulation density and AIR% have a linear relationship; as circulation density increases within the postcode sector, the AIR% rises towards 100% and is capped at 98% readership.

In the new model, circulation density and AIR% have a non-linear relationship; as circulation density increases the rate of increase in AIR% slows down. At 100% circulation density, the AIR% is estimated to be 77%.

The differences in the curves between the current JICREG model and new estimates, particularly for circulation densities over 30%, explain some of the differences between the two models.

The new model curve is believed to have a more realistic relationship between circulation density and readership than the current model. Even if more copies are allocated to a postcode sector, there is likely to be a limit on the number of people wanting to read the newspaper.

We are currently finalising the models but are confident that these new models are a good reflection of the current local print audiences.

3. Conclusion

The next steps before the introduction of the models is for them to be finalised with the addition of publication scc and advertising ratio data. Once this has been done a period of testing will take place with our software partners (Adweb) and publishers before signing off by the technical subcommittee and the board.

It is anticipated that readership estimates will be updated twice a year using updated title and population data including circulation figures and changes in distribution.

The new models provided a solution to a key issue for JICREG and the innovative collaboration between two JICs helped maintain the gold standard reputation of the currency.

Appendix A: Explanatory variables considered for modelling

Field	Description
	Title information
1.	Type of newspaper (Paid, Free door-to-door, Free pickup)
2.	Frequency (Morning, Evening, Weekly, Sunday)
3.	Publishing day
4.	Total newspaper circulation
5.	Circulation at postcode sector level
6.	Publisher
7.	Company
8.	Distribution % by type
9.	Issues per year
10.	Format (Broadsheet = 1, Tabloid = 0)
11.	SCC (single column centimetres) / unit cost
12.	Advertising Sales House
13.	Number of publishing days per week
14.	Mixed distribution title (1 or 0)
15.	Sunday title (1 or 0)
16.	Evening title (1 or 0)
17.	Circulation density
18.	Household penetration
19.	Pagination
20.	Advertising / editorial ratio
21.	Cover price
22.	ROP SCC advertising rate
23.	ROP full page advertising rate (pages)
24.	Tabloid equivalent pagination
25.	No of columns
26.	Column length
27.	Agency commission
28.	Paper Quality
	Competitive regional titles
29.	Number of Dailies
30.	Circulation - Dailies
31.	Number of Paid Weeklies
32.	Circulation - Paid Weeklies
33.	Number of Free Weeklies
34.	Circulation - Free Weeklies
35.	Number of Sundays
36.	Circulation - Sundays
	Area information (by postcode sector)
37.	Total Population (Adults)
38.	Number of households
39.	Adults per household
40.	Government Office Region
41.	Population Profile (%): Men
42.	Population Profile (%): Women
43.	Population Profile (%): 15-24
44.	Population Profile (%): 25-34
45.	Population Profile (%): 35-44
46.	Population Profile (%): 45-54
47.	Population Profile (%): 55-64
48.	Population Profile (%): 65-74
49.	Population Profile (%): 75+
50.	Population Profile (%): AB
51.	Population Profile (%): C1
52.	Population Profile (%): C2
53.	Population Profile (%): DE

Census data (by postcode sector)

54. % of non-English speakers
55. Population Density
56. Urban / Rural
57. % Work mainly at or from home
58. % Underground, metro, light rail, tram
59. % Train
60. % Bus, minibus or coach
61. % Taxi
62. % Motorcycle, scooter or moped
63. % Driving a car or van
64. % Passenger in a car or van
65. % Bicycle
66. % On foot
67. % Other method of travel to work
68. % Not in employment

Commitment to local area

69. Whether title is associated to a large local sports team