MODELLING AUDIENCE DATA FOR NEWSPAPERS LOCAL EDITIONS MEDIA-PLANNING

Gilles Santini, IMS/IPSOS

Abstract

Since January 1999, magazine readership surveys in France have been using the CAPI double screen system. The researchers use a micro computer that asks the questions, handles the random order and filters, and logs the answers. It is connected to an electronic laptop that presents the respondents with the magazine mastheads and answer cards. This article explains why the AEPM adopted this system and provides the initial readership survey interview data gathered between January and June 1999.

I- Context

In France as in most countries regional dailies are circulated through local editions. From an area to another significant differences do exist in terms of audience level and structure of the readership. So, what is usually called the audience of a regional daily is in fact the consolidation of many local situations.

Although the generally available readership surveys are by design appropriate to provide the global average audience of regional dailies such surveys cannot produce reliable audience levels for each edition, due to their sample size practical and economic limitations.

In order to overcome this limitation an attempt has been made in France in cooperation with the industry bodies in charge of the regional dailies readership survey to experiment with a new approach that would produce reliable estimates for each local edition of a given regional daily with known total audience level and audited local circulation.
This method which has been developed by IPSOS/IMS over the past two years has received careful testing and has been approved for implementation by the SPQR study committee in charge of regional dailies audience measurement.
It is known and implemented under the Celtic acronym GAEL which stands for "Generation des Audiences des Editions Locales".

GAEL is a top down modelling process based on an "a priori" model that factors out four exploratory components. We will explain next how this model is built and how its parameters are estimated. But it is worth to point out immediately that such an approach is quite different in nature with statistical model estimation : in this later case, we seek relationships inside a data set and we settle for the most likely one, if any. Here, we postulate a scheme of causality and we attempt to apply it under known constraints to draw quantified inferences (specifically the edition audience levels).
In a way one could say that GAEL is more of a macro-economic modelling nature than of a statistical nature, although statistics are used in the fitting process.

II - Model

Let A_T be the total audience of a regional daily and N be the size of the total zone covered by the local editions. For a given local edition covering a zone k with population size N_k a simplistic evaluation of the edition audience is :

$$A_k = A_T \frac{N_k}{N_T}$$

This calculation would reflect the belief that audience is homogeneous across all geo-demographic zones and is dependent of the considered edition.

Equivalently we could work with the primary audience[1] itself P_T and introduce the ratio $\lambda = 1 + \frac{S_T}{P_T}$ that exhibits the swelling of the audience when the secondary readers S_T are added to the primary one to amount to the total audience

$$\boxed{A_k = \lambda \; P_T \frac{N_k}{N_T}}$$

GAEL makes the assumption that four factors should be taken in account in the calculation of A_k from A_T keeping as a constraint that the sum of all parts equals the total :

For K non overlapping zones and local editions, we have :

(1) $N_T = \sum_{k=1}^{K} N_k$ (size constraint)

(2) $P_T = \sum_{k=1}^{K} P_k$ (primary readers constraint)

(3) $A_T = \sum_{k=1}^{K} A_k$ (total audience constraint)

In order to understand the nature and the role of those four coefficients, the model is best described as a flow chart.

We start with the known total primary audience P_T and work out the local primary audience P_k :

$$\boxed{P_T \rightarrow P_k = \alpha_k \omega_k \mu_k \frac{N_k}{N_T} P_T}$$

This calculation involves three factors which are designed to be equal to 1 if the behaviour the k^{th} edition is similar to the overall daily behaviour.

- α_k measures the level of appeal of the daily to the population to whom edition k is circulated.
- ω_k measures the relative strength of the paid circulation of edition k compared to the total paid circulation.
- μ_k measures the relative level of the free circulation of edition k compared to the total free circulation.

Once P_k is known it is possible to determine the audience of edition k thanks to a local version of the λ ratio :

$$\boxed{P_k \rightarrow A_k = \lambda_k P_k}$$

So much for how these four factors operate, which is in essence quite simple since they are merely normalized multiplicative factors meant to express the local specificity compared to the average total situation of the daily.
We must now be more specific about how to calculate them.

For that purpose, we have to introduce the notion of "localized Primary Audience Level" P_k^* which is an estimation of what would be the level of the primary audience if the local population structure of the zone where edition k is circulated was the same as that of the global sample.This estimate can be obtained from the general audience survey by a sample reweighing technique as long as the structure of the population for major demographics are known from a reliable statistical source such as the census bureau.

P_k^* being known, the level of appeal of the daily for the population to whom edition k is circulated can be expressed as :

$$\alpha_k = \left[\frac{P_k^* / N_k}{P_T / N_T} \right]^m$$

where m is a parameter that fixes the intensity of the α factor effect.

[1] For the purpose of this work we call primary reader any reader who is either a buyer or belongs to a household where there is a paid subscription.

Also one should note that as required $\alpha_k=1$ if the structure of the population covered by the k^{th} edition is similar to the one of the total population concerned by the difference of the daily, since in such a case $P_k^*/N_k=P_T/N_T$ by construction.

The form of the ω factor is more complex : it reflects the fact that if the circulation grows, the audience grows too but with an asymptote.

We use a logistic for such a purpose, which writes :

$$\omega_k = \frac{\alpha}{1+(\alpha-1)\exp(\beta(1-\delta_k))}$$

where $\delta_k = \frac{C_k / N_k}{C_T / N_T}$ with C_k = the paid circulation of the k^{th} edition.

Clearly if the local circulation is equal to the average one $\delta_k=1$ and $\omega_k=1$ as required. Also, if the k^{th} edition is massively being circulated δ_k will grows but ω_k will reach a ceiling.

Last the μ_k factor, which carries a total circulation effect, has a simple form.
It writes:

$$\mu_k = \frac{1+\frac{F_k}{C_k}}{1+\frac{F_T}{C_T}}$$

where as above C_k is the paid circulation and F_k the free circulation.

Globally the model depends on three parameters m, x and b. The value of which need to be fixed on the basis of the overall constraints.

Keeping for the time being m fixed equal to a small integer value (i.e. 1 to 4) the model reduces to a two parameters problem α and β. The two additive constraints (2) and (3) on A_T and P_T should in principle be sufficient to solve the problem but it turns out that the equations to solve are numerically very unstable.

In order to overcome such a difficulty we have followed what mathematicians call a regularization strategy : we include an additional non normalized multiplicative factor R in the model and solve for this augmented problem. Since it makes sense to keep the circulation ratio variability from one edition to another somewhat small we will use such a condition to reach an optimal solution.

Namely we have to find R, α and β that minimize :

$$E = (1-w)\left|P_T - \sum P_k\right| + w\sum\left(\frac{A_T}{D_T} - \frac{A_k}{D_k}\right)^2 \quad \left(D_T = C_T + F_T\right)$$

where w is a small contamination factor (e.g.=0.10), under the constraint that $\sum A_k = A_T$

This strategy works well as anticipated and the best solution can readily be achieved for m fixed.

Finally the value of m that leads to the smallest minimized E value is chosen.

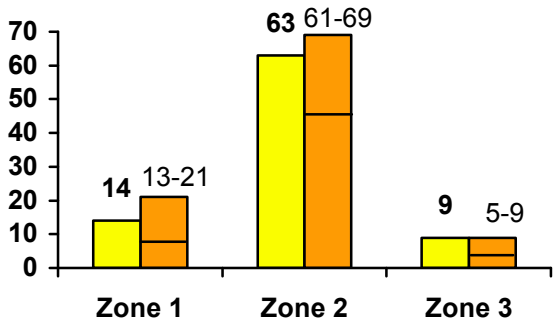
III- Control

In order to check the behavior of the above approach, we had the possibility to compare with the figures from private local oversamples recently commissioned to IPSOS MEDIAS by four regional dailies groups. Although the time reference was different and these surveys limited in extent, the comparison turned out to show excellent adequacy for almost all local situation. This was a very encouraging result since the four cases were very different in nature ranging from dominant newspapers to weaker ones with local editions covering rural and urban areas in different parts of the country.

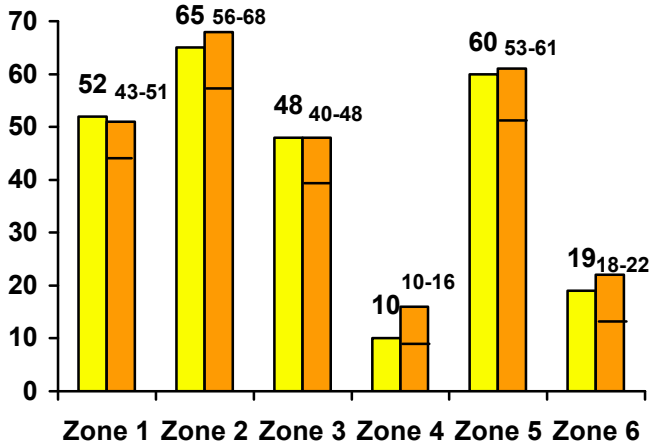
The following graphs show some of the findings which should be examined keeping in mind the fact that the data from these sub samples has not been used in the GAEL process that only relies on the general audience survey, census type data and local circulation figures.

GAEL calulated audience levels are shown on the left compared to the ad hoc surveys estimation intervals on the right

Daily A



Daily B



IV- Validation

Seeking for ways of additional validation of the GAEL method several proposals were examined by the industry. Although direct methods based on specific ad-hoc national survey were appealing, it turned out that these were neither economically affordable nor practically easy to design.

The GAEL validation working committee turned then towards an indirect methodology proposed by experts and commissioned it.

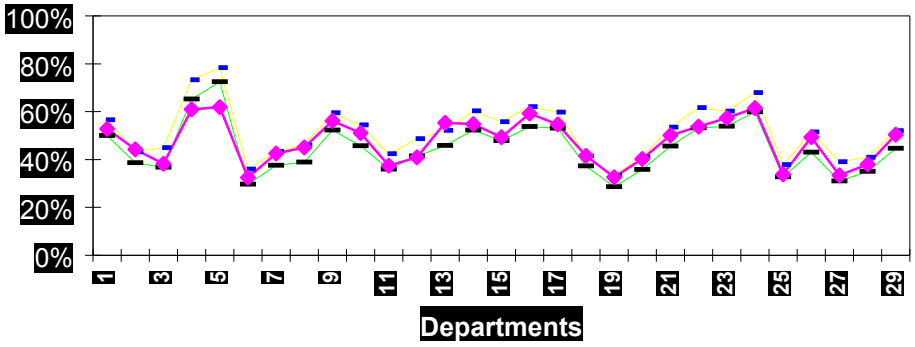
The idea behind this new approach was to magnify the sized the problem and instead of calculating the audience of all of the edition of a regional newspaper to calculate the audience in all the French counties (department) of the regional press considered as a whole. Although this was more difficult because of the higher heterogeneity of the situations, the nice thing was that the reference study could provide reliable estimate for comparison in 29 counties if we worked with a pooled database of the regional newspapers reference study over three years (1995 to 1999) and used the average circulation published over the same period.

The validation lead to a case by case comparison in each of these 29 areas as to check whether GAEL was indeed producing figures within statistical bounds of the survey itself and was not introducing any systematic bias.

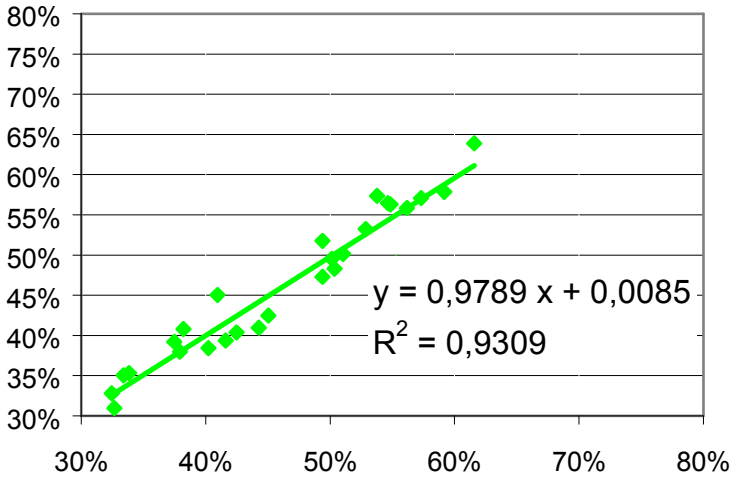
Departments	Sample size	GAEL calculated audience levels	observed audience	absolute value error	Departments	Sample size	GAEL calculated audience levels	absolute value error	Gap in absolute value
Department 1	906	53%	53%	0%	Department 16	558	59%	58%	1%
Department 2	1674	44%	41%	3%	Department 17	837	55%	56%	2%
Department 3	555	38%	41%	3%	Department 18	2172	42%	39%	2%
Department 4	501	61%	69%	8%	Department 19	1431	33%	31%	2%
Department 5	840	62%	75%	13%	Department 20	1329	40%	38%	2%
Department 6	852	32%	33%	0%	Department 21	630	50%	50%	1%
Department 7	1125	42%	40%	2%	Department 22	534	54%	57%	4%
Department 8	768	45%	42%	3%	Department 23	948	57%	57%	0%
Department 9	723	56%	56%	0%	Department 24	549	62%	64%	2%
Department 10	510	51%	50%	1%	Department 25	1377	34%	35%	1%
Department 11	897	37%	39%	2%	Department 26	540	49%	47%	2%
Department 12	729	41%	45%	4%	Department 27	546	33%	35%	2%
Department 13	957	55%	49%	6%	Department 28	1038	38%	38%	0%
Department 14	603	55%	56%	1%	Department 29	705	50%	48%	2%
Department 15	618	49%	52%	2%	Average discrepancy				3%

Except in two known special situation where it was not surprising to see a different behavior the results turned out to be strikingly good with a R^2 adequacy of 93% around the diagonal perfect concordance line.
The following couple of graphs speak by themselves.

GAEL calculated audience levels compared with reference surveys confidence intervals



OBSERVED VS CALCULATED AUDIENCE LEVELS



V- Conclusion

Implementing GAEL unusual top down approach, we were confronted with a rather difficult mathematical optimization case but once this difficulty overcome the outcome was very satisfactory and the method released for industry usage.

Clearly GAEL does not make up data which does not exist. It merely extend the benefit of the reference national audience survey in combination with external official or controlled data in particular audited circulation figures and public demographic data.
In our opinion this has only been made possible because of the quality of the various data inputs and of the meaningful identification of the underlying model.

Aside the practical use of the outcomes of GAEL this is probably one of the constructive findings of this effort and we may want to learn from it in other similar media research circumstances by moving from our usual sampled data best fit approach to a more integrated and causal treatment.