MASSIVE MODELING: A NEW MEDIA RESEARCH CHALLENGE

Gilles Santini, Vintco

1-Introduction

Although modeling was not part of the main stream of Media Research it played (since the first World Wide Readership symposium in 1981) a very special role paving, over the past 25 years, the way for innovation.

We got to learn how to fit specific functions to audience accumulation using various possible curve shapes and also how to simulate individual audience behavior through the use of personal probabilities ascribed on the basis of observed individual data. Those efforts implied to identify, estimate and control the proper parametric models.
We used variance decomposition when looking for segments of population with distinct behaviors.
We also came across non parametric geometric models when we wished to create respondents clusters with similar characteristics and today more frequently than ever we rely on topological distances calculations for data fusion and similar types of data merging.

These modeling tasks where performed on modest size data sets, collected on samples ranging from a few hundred to a few thousand respondents.
Today the proliferation of massive data sets brings with it a new challenge for media research.

The purpose of this paper is to make our community of researchers aware of the fact that the naive belief that "*we now have big enough computers to handle it*" is just wrong and that a benign neglective attitude such as "*the IT technical team will know how to solve the issue*" bears a major risk for the survival of our media research expertise.

2-Data avalanche

Our media industry is in the midst of a "data avalanche" which could very well sweep out our ability to keep control on how media data is used.

Such a situation occurs in a wide range of contexts such as out of home audience, web sites frequentation, or return path data capture, to name a few.

It has been triggered by the availability of audience passive means of data collection at a time when the way people consume media is also radically changing.
It has been swelled by the acceleration of marketing decisions and the advent of micromarketing.

Although the ability to process and model massive audience data sets start to be at the core of many new business service offers, very little has been published or uncovered on those issues which are considered as corporate assets by the industry players.

The inflation of the datasets size may have several factors which can be listed according to their increasing effect roughly measured as a magnitude of 10:

- Sample size :
  - from a traditional probability sample to an Internet or database sample magnitude increase is 2, to a census magnitude increase is 4
- Number of relevant entities :
  - from magazines to websites magnitude increase is 2, to in-store media magnitude increase is 5
- Time :
  - from daily data one month, to second/second continuous data, magnitude increase is 6
- Trips:
  - considering n sites the number of paths between them amounts quickly to a magnitude larger than n

Clearly if we consider situations where several of those factors come in play jointly, we rapidly get pharaonic data set sizes. This is obviously the case for RPD capture which produces terabytes of data for analysis but imagine what could be the size of a data set captured by your 3G phone service provider : petabytes (10^{15}) most likely.

So, with the extension of the role of mobile screens, one should recognize the fact that no media will stay unscathed in the long run from this data avalanche.

But the size of the data set is not the most complicated issue to handle, after all, some expensive parallel data warehouse systems do exist which can store and retrieve quickly any piece of information.
What is a much greater challenge is the processing of the data itself and in the case of a continuous capture to do it on the fly in streaming mode.

3-Processing paradigms

Historically two different paradigms have been brought to the problem of knowledge retrieval from a data set.

One can consider that the available data is a faithful image of what is "true" and the goal is then to tabulate and summarize it in such a way that relevant patterns will catch the researcher's attention.
The belief is that such patterns being "true", reflect a state of the world that can be used to support decisions.
We call this paradigm the **geometric** one since it provides a static insight by imposing geometry on the data.

Alternatively one may view the data as an instance of underlying probabilistic phenomenon. In such a case it is necessary to identify their structure and estimate their driving parameters in, other words, to model them.
The belief is that such models are latent to reality and can be used to project likely outcomes.
We call this paradigm the **probabilistic** one since it provides a dynamic insight by modeling the phenomenon.

Clearly in both cases noisy or uncertain data may lead to spurious results but careful data mining can protect against such risk.

Those two perspectives are not incompatible and a good statistical education typically covers both. The difficulty comes from the scale issue: when there is a large spectrum of situations, one cannot easily extract the few interesting tables that may exist nor discover what the underlying models are.

To illustrate this data modeling challenge we will present a case study on Internet audience data for which we conducted the work in 2008 and 2009 with Gaël Crochet from KMR Software/JFC Paris.

This case study has not been selected for its footprint, which is not gigantic compared to other ones we have come across with far more impressive sizes, but to illustrate the scale gap with a similar objective in the context of a print audience study.

4-Modeling Internet sites audience

4-1 The print case

The route to media planning for print media is (almost) a standard nowadays.

It can be sketched as follows when the RR collection method is used for audience measurement:

For each vehicle (a magazine or a newspaper) one collects:

- An overall audience population filter
- A declared "frequency of reading"
- A "has read during the last period" question

Each respondent within the sample should be ascribed a probability of reading.
Any respondent outside the audience overall bounds gets a zero probability.

The process used to ascribe the probabilities to the respondent is quite simple:

Every respondent which has declared to read with a certain frequency gets a probability equal to the ratio of
the number of respondents which have declared reading with that frequency and have read within the last period over the number of those which have declared reading with the same frequency.

To achieve a better precision the above calculation may be done separately on population sub groups either defined a priori or produced by statistical segmentation.

Audience accumulation is done using at respondent level a binomial distribution.

The number of pages read is taken into account when appropriate data is available but this is not frequent.

Overall audience accumulation is obtained by summation over all respondents in the target group.

All the above calculations are very straightforward and can be processed in a few seconds of elapsed time.

4-2 The internet case

The internet case is much more involved with several sources of complexity.

Firstly, we have to deal with a much larger number of sites than in the print case.
Secondly, we have to handle the number of visits as well as the number of pages looked through
Thirdly, we want to be able to assess when in time the exposures do occur
Fourthly, the surfing paths need to be properly accounted for

An extra complexity comes from the fact that sites may be very distinct and do play different roles in the internet user experience: visiting a search engine site, an air flight booking one, or tax revenue service facility implies different behaviors over time.

After several attempts we addressed the above as follows, the work being carried on the latest month of data:

Step one:
In order to have at hand a "summary" of the population habits, a set of explanatory variables is built using quartiles of the number of visits and the number of pages seen for each category of site.

Step two:
Using the previous summary variables in complement to the ordinary socio-demo attributes a specific segmentation of the population is performed for each site using a generalized AID algorithm in order to identify distinct and compact groups showing the same level of frequency of visits over the month.

Step three:
For each site and for each segment of population identified for that site at step 2, four distinct models incorporating a zero segment are tested:

- z-Binomial
- z-Poisson
- z-Markovian-Binomial
- z-Beta-Binomial

The best one is selected.

The rationale behind the choice of those models is that they cover a very wide range of phenomenon:

- z-Binomial accounts well for behavior characterized by a frequency level
- z-Poisson is good for time scattered events
- z-Markovian-Binomial takes in account time correlation of the events
- z-Beta-Binomial implies event exchangeability of the events over the period

The zero-segment (never) can be present or not depending of the segment.
Estimation is made with the objective to fit most exactly the monthly reach and the number of pages seen.
Selection between the models is based on the goodness of fit.

The calculation is rather intensive leading to the estimation of 150 000 models or so and can be performed in a matter of hours.

Once we have for each respondent his behavior for each site properly modeled we may consider doing some media planning.

The first difficulty is that the data is so big that it cannot be:

i. Delivered easily to your PC nor updated afterwards,
ii. Retrieved fast enough to be convenient in interactive mode

The solution is to organize the data in a way which takes full advantage of its specificities (in particular the way it has been created and its time structure) and to design an on-the-fly compression/decompression mechanism.

This is feasible but at a cost : since the generally available data handling tools, databases and retrieval algorithms cannot do it, one needs to design specific software for such a purpose.

The second difficulty is that standard media planning systems cannot handle, at the same time, time processes of different natures (and most of them stick to timeless Reach and Frequency evaluation): innovative R&D and its corresponding budget requirements is the answer there.

4-3 Complementary issues

The above methodology clearly follows the probabilistic paradigm.
Internet site audience data also raises challenges in the perspective of the geometric paradigm.

Here is an example.

Magazines publishers often also maintain a *sister-site* to the print product extending brand coverage.
The duplication between the readership of the magazine and the exposure to the corresponding site is often lower than expected. Such duplications are probably better captured by ad-hoc studies and it is interesting to use them for media planning.

Given the available studies, fusion is most often the unique solution to put data together. The difficulty is that one doesn't want to lose or distort such duplications in the fusion process.

We use for fusion a distance between donors and recipients which is quite complex to calculate, the Procrustean distance. It is not the purpose of this paper to describe such methods, that we have presented in Venice at the 2001 WWRS, but it is worth knowing that it requires computing the Eigen structure of a large matrix. Such a calculation is routine work for a matrix with 10 to 100 thousands entries but we have been faced in some circumstances to deal with matrices 100 times bigger. The work to calculate an Eigen structure for such an object is very large but can be addressed with a powerful processor. However, computations rapidly reach instability because of floating-point precision problems.
Computing techniques which are not widely available are necessary to break this computing barrier.

4-4 Lessons

Comparing the Print case to the Internet one we have learned that:

- Heterogeneity of the phenomenon calls for complex modeling methodologies
- Time dimension is the major source of data inflation
- Combination of behaviors require heavy geometrical analysis
- Data delivery of large quantities of data is not as easy as it seems
- Data storage and retrieval need to be carefully though of

Similar lessons can be learned from other cases that we have lately encountered in dealing with in-store data and product and brand usage studies.

All raise the same buzz word: "*scalability is an issue*" …

5-Conclusion

In the past few years online advertising has grown an order of magnitude faster than advertising on other media.
Passive electronic audience capture enables the media industry to envision the possibility of census collection instead of sampling.
Media usage needs be analyzed over time.
Cross media planning calls for better duplication estimation.
Multi-tasking and Out-of-home are becoming dominant behaviors.

As a consequence media research is faced with major complexity and scalability issues.

Two answers can be given to the challenge:

- **Discard the problems to another industry**:
  - streaming and parallel computing are in the scope of powerful IT companies which might be happy to control tools for data handling and analysis, influencing as a result market intelligence.

- **Handle the issue at industry level**:
  - the existing expertise needs to be recognized, strengthened and shared as a common industry asset since it is too heavy to grow and maintain for a single player even if he is a major one.

To conclude this paper we will quote Adam Jacobs, a senior software engineer at 1010Data Inc who worked previously at Cornell University and UCLA, from his latest paper "The Pathologies of Big Data" published in the last summer issue of The Communications of the ACM : *"[...],as analyses of ever-larger datasets become routine, the definition will continue to shift, but one thing will remain constant : success at the leading edge will be achieved by those developers who can look past the standard, off-the-shelf techniques and understand the true nature of the hardware resources and the full panoply of algorithms that are available to them.".*