The Age of the Algorithm

BY | Andrew Green and Mario Paic

Introduction

An algorithm is a process or set of rules followed in calculations or other problem-solving operations, especially by a computer. They are all-pervasive in the digital world, determining, for example, whether banks will lend or employers will pick people for interview.

In the online world, Google 担 PageRank algorithm determines how websites in the company's search engine results are ranked. Meanwhile, Facebook's News Feed algorithm controls who sees what within the social network.

In the Audience Measurement area, algorithms are commonly used in schedule analysis software, employed by media planners, buyers and sellers to estimate the number of people likely to see or hear an advertising message.

They are also central to more advanced statistical techniques being used increasingly in audience measurement including data ascription, data fusion and audience modelling.

Survey data alone is no longer enough to capture the complexities of a fast-changing media environment, with more content choices and more distribution platforms from which to access them. We are demanding more and more from our respondents at a time when many are becoming less willing to participate in long or boring surveys. Employing statistical techniques allows us to collect less data from individuals and to tolerate less precision in their answers.

The design and execution of these statistical methods demands a high level of expertise and skill. For users of research data, they have become increasingly critical in the race to keep abreast of changing audience characteristics and habits.

“ Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”

H.G. Wells

The Context

We don’ t need to extract all the blood from somebody in order to determine their blood type. In the same way, we can learn a lot from taking a sample of a population and questioning them or tracking their behaviour. Good market research practice demands that a sample should represent as much of the variability in the population being measured as possible to be effective.

In the audience measurement domain, we know that gender, age, education level, occupation, household size, income levels and the region where people live will all affect the probability that they read certain newspapers or magazines, watch given programmes or listen to the radio. So these characteristics must be faithfully reflected in the composition of any sample purporting to measure media usage.

Non-demographic factors can be important too. For example, when asking about digital behaviour, it is important to properly represent the number and range of media-receiving devices (PCs, tablets, Smartphones etc.) owned by households, which are likely to be associated with usage levels.

But while surveys have proven very valuable to companies and governments over the years, they have their limitations. One limitation is that they can never be perfectly representative of a population, unless we talk to everybody and unless they all answer with total honesty and perfect recall.

Statistical weighting is often used in market research to correct these sorts of imbalances.

The more important limitation for the purposes of this paper is how we extract as much information as possible from our respondents. There are natural limits to how much time people are prepared to spend answering questions. Given that few are blessed with perfect recall, we also have to be careful to word questions clearly and not to ask for information it is unreasonable to expect people to remember.

The challenge today is that clients want us to collect more information, not less.

Where Statistical Adjustment Can Help

Marketers want to know everything they can about their target consumer in order to maximise the return on their research investment:

- Who they are (demographics, geo-demographics, psychographics etc.)
- What they think about brands in the category they are asking about
- How they behave (purchasing levels, brand choice etc.)
- What they intend to purchase in the future

Those planning and buying advertising campaigns need to uncover the best ways of reaching and influencing their target audiences:

Which media do they come into contact with at different times of day (TV programmes, newspapers, magazines, radio stations, web sites, apps, poster panels…)?	Which media are they more or less attentive to or engaged with at different times of day?	When is the best time to reach people with an advertising message (message receptiveness, when are they in the market to buy…)?

But no individual respondent will agree to answer such a large number of questions. And many of the questions will be impossible to answer accurately. There are two closely related statistical techniques used to help address this: data ascription and data fusion.

“ An algorithm must be seen to be believed.”

Donald Ervin Knuth

Data Ascription

Where answers to a survey are missing or incomplete, it is possible to infer what those answers would be by looking at answers given by similar survey respondents. These responses may be missing by accident (people forgot or omitted to answer them) or by design (where we had too many questions to ask, so split the questionnaire between different sub-samples).

In this case, we design two (or more) questionnaires, each sharing certain core questions in common, but with separate sets of questions on other topics. These questionnaires can either be served simultaneously to separate but similar samples of people or they could be asked at different times (e.g. Questionnaire A can be asked for a month, then Questionnaire B for a month and so on).

The assumption is that we can then ‘match’ people answering the different questionnaires using the known demographic and other characteristics of each sample member, as well as the answers they give to other common questions.

We then take answers to the first set of questions and ascribe them to matching respondents who answered the second set of questions and vice versa. This gives us a larger database of answers than we could have had with a single sample of people.

Examples of Data Ascription

In **Brazil**, the Ipsos EGM survey has a lengthy questionnaire administered face-to-face in all the major regions of the country. There are two versions of the questionnaire, with media and demographic questions identical on both, but different questions asked about usage of various brands and products, as well as about attitudes. Each version of the questionnaire is served in alternate half years, allowing us to merge answers to all the questions into a single database for 12 month periods.

In **Australia**, our emma (Enhanced Media Metrics Australia) study measures both media consumption and product usage. Because there are a large number of product categories covered, all participants in the study are asked about top line product category usage (e.g. do they drive a car? what kinds of food and beverages do they consume?).

Two matched samples are each then asked detailed brand questions about half of the product categories. The responses from each half of the sample are then ascribed back to the other half of the sample, producing a final integrated dataset where we have detailed brand information for the entire sample.

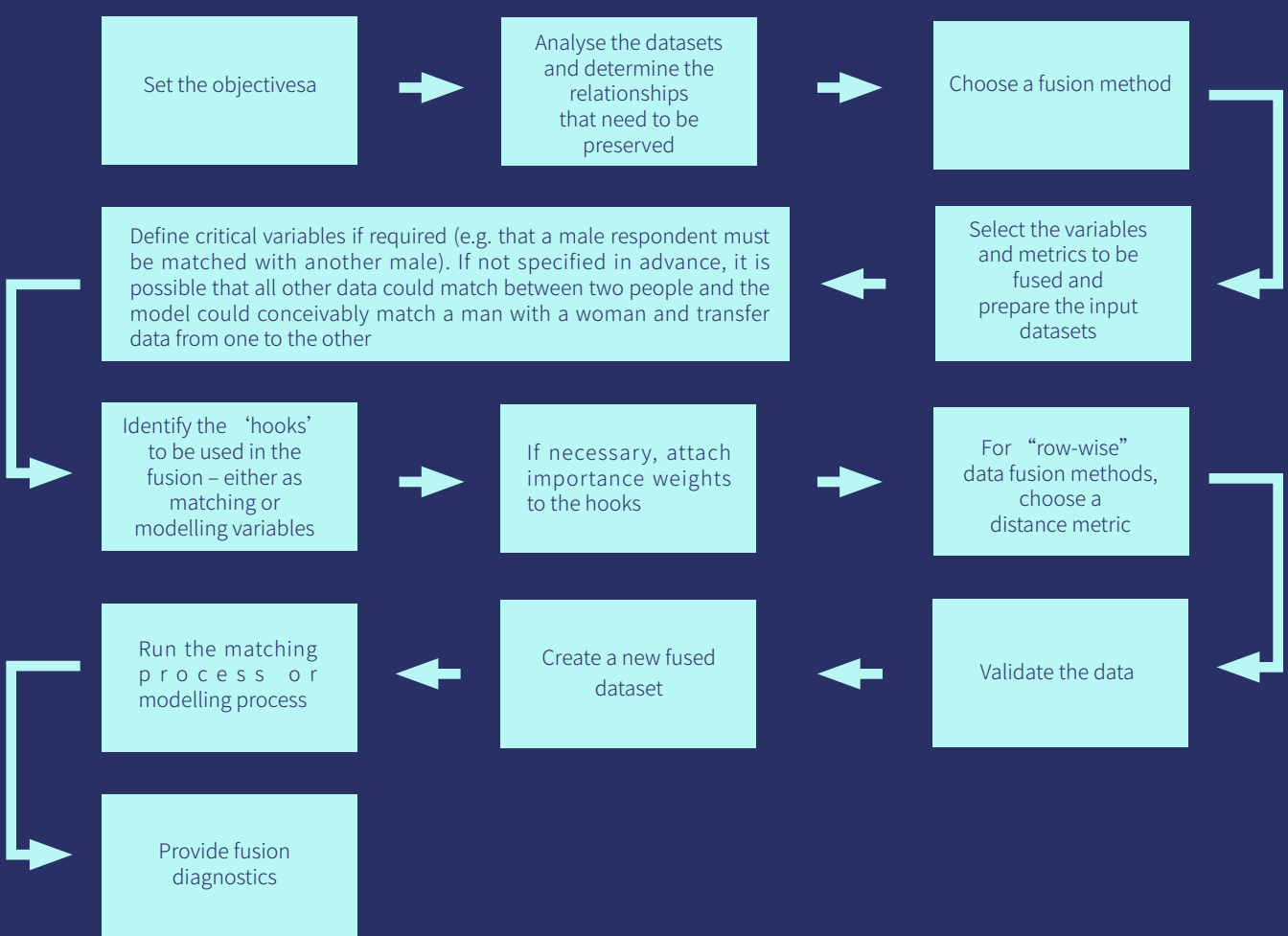
Data Fusion

A related technique is data fusion of two or more separate surveys. In this case, a survey designed for one purpose is joined with a survey carried out for a different purpose – as long as there is sufficient information common between both studies for survey participants to be matched. This generates what looks like a single-source database containing all the previously separate information.

Key to a fusion are the "hooks" common to the surveys being linked, designed to connect respondents in one survey with those from another (known as 'donor' and 'recipient' respondents, because data from one respondent will eventually be 'donated' to a matching respondent from another survey).

Obvious questions likely to be asked in both surveys include gender, age, education level and the region lived in. The fusion process seeks matches on these and other criteria – ideally we will then be able to match (for example) young, well-educated males living in the South-East on one survey to other young educated males in the same region on another survey.

Fusion is not a single technique – different approaches can be taken depending on the objectives. The principles for any approach are similar however and follow these general steps:



Examples of Data Fusion

In **Great Britain**, Ipsos Connect has designed a method for combining two sets of audience data – AMP (which collects readership information for newspapers and magazines) and UKOM (which tracks audiences to websites and apps).

Adding to the complexity of the fusion design was a requirement that each of the separate studies 'preserved' their individual audience estimates. It also needed to bring in cross-platform behavioural insights gleaned from a sub-set of individuals from the AMP study who had installed passive data capture software on their digital devices (in order to accurately measure reading across all the platforms where newspaper and magazine content can be found).

In **Australia**, Ipsos Connect has similarly built a data fusion approach which integrates outputs from the emma survey (which surveys media usage) with online audience information from a separate panel to create total audience estimates by platform for individual newspapers and magazines.

The data integration process developed involves a constrained[1] fusion methodology between the two data sources. Information included on both sources includes demographic, geographic and internet activity variables to produce the optimum solution.

[1] A detailed description of the various fusion techniques is beyond the scope of this paper, but options include 'constrained' or 'unconstrained' fusion (taking decisions on any survey data which must be preserved in the final results) and whether we use a 'row-wise' or column-wise' approach (different methods of handling the datasets).

Audience Modelling

Audience modelling is another weapon in our armoury. The idea is to take known information which can help predict the behaviour we want to measure and to organise it into a process which delivers credible results.

There are at least three reasons why we might want to use it to enhance our survey data:

To increase the scope of our survey coverage. For example, on a readership survey, we can increase the number of titles covered by including those far too small to register on a sample-based survey alone.

To increase the granularity of our reporting. Reporting frequency can be increased, as it will no longer be constrained by the need to build a large enough sample for robust reporting – instead it is limited by the reporting frequency of the model input data (such as circulation or sales information).

To improve the speed of reporting. A model does not depend on waiting for survey data to be collected and processed, it can be published far more quickly.

Examples of Audience Modelling

In **Australia**, we were faced with the challenge of providing robust readership estimates for more than 400 regional, local and community titles, including many selling fewer than 5,000 copies per issue. To obtain statistically valid readership estimates from a sample would have required unfeasibly large numbers of respondents to be polled.

Having noted previous work demonstrating the close correlation between newspaper sales (circulation) and readership levels, we built a model incorporating circulation and circulation splits across newspaper distribution areas, combined with the demographic profiles of the same distribution areas and information about the titles themselves.

In **Belgium**, new modelling techniques developed by Ipsos are helping measurement body CIM (Centre d' Information sur les Médias) to create the world' s first daily audience data for newspapers. The model is fed by a readership survey, alongside a daily SMS panel and daily sales data from publishers, which are then tied together to create dayby- day readership information for newspapers and specific issue readership for magazines.

In the **UK**, "multi-sensor" tracking devices are used on our Route Out-of-Home study to track the travel behaviour and movements of a large sample of people over two-week periods. These data are combined with information on traffic and pedestrian flows taken from local authorities, as well as detailed location information for every advertising panel in the country. From all this, we are able to generate estimates of the number and profile of people passing by any of these panels which is used to plan and buy space in the medium.

" An algorithm is like a recipe."

Waseem Latif

Conclusion

The digital revolution means advertisers and media companies need more information than ever before on media usage.

Yet people are less willing than they have been in the past to participate in surveys and, even when they can be persuaded to do so, want to be engaged rather than bored.

This conundrum of needing more information while finding it harder to collect from surveys alone is likely to get harder rather than easier as time passes.

In the audience measurement domain...

Techniques like data ascription, data fusion and audience modelling are allowing us to collect and report more and better data, enabling us to keep pace with increasingly complex client needs.

The practical application of data science demands a high level of skill and expertise, as well as experience – many of the decisions and choices made in building fusions and ascriptions, for example, are not black and white, demanding judgement and a deep knowledge of the context.

We believe it will play a growing role in the processing of audience measurement data.

“ Statistics is, or should be, about scientific investigation and how to do it better..”

George EP Box