The case for multi-sourcing prospect pool data

Received (in revised form): 6th January, 2005

Julian Berry

is a founding director of the customer partnership and has been employed in customer marketing since he started his direct marketing career in 1980 working for Christian Brann. His vertical market client-side experience is in financial services, and he has been involved in over 400 projects as a consultant. His current research interests are in bringing together insight tools, processes and decision-making capabilities to enable organisations to make the most effective possible marketing decisions.

Paul Rowson

is a senior consultant specialising in advising clients on how to derive the full business benefit from their customer data through understanding and predicting customer behaviour. He has successfully applied customer insight techniques across a wide range of industries. Previously, he was Head of Database Marketing at GE Capital where his team delivered added value database marketing services to UK store-card clients.

Abstract Among business-to-consumer (B-to-C) companies engaged in large scale recruitment activity, there is currently a significant shift towards developing prospect pools rather than being reliant on *ad hoc* list purchase. One key decision to be made by prospect pool operators is whether to appoint a single data provider (which often also provides analytical and IT services) or whether to multi-source. This paper looks at the pros and cons of each approach and offers case study and survey results to show the significant overall benefit of multi-sourcing.

THE BUSINESS CONTEXT

There are a number of reasons why an increasing number of large scale acquisition direct marketers are switching from using one-off list purchases to setting up long term prospect pools; the authors expect that the UK is the most advanced country in adopting this approach. The key reasons are:

- the reduced cost of data if it is bought on a longer term contract;
- the ease of generating large volume campaigns;
- the availability and comparatively low cost of the electoral roll (even with opt-outs);
- the availability of many different types of data that can be used to support targeting activities (including lifestyle,

credit, transactional data pools, suppression files and geodemographic data);

- the effectiveness of the predictive models that can be run on the prospect pool to predict response and sometimes value;
- reduced IT costs if names do not have to be continually de-duplicated;
- the ability to integrate customer with prospect data; and
- the ability to manage longer term prospect contact strategy.

Managing a large prospect pool requires considerable investment, however, and hence they are normally operated by larger companies prepared to spend upwards of £500,000 *per annum* on them. This means, in effect, that only

Julian Berry
The customer partnership,
212 Piccadilly,
London, W1J 9HG, UK.
Tel: +44 (0)208 878 9350;
e-mail: Julian.Berry@
thecustomerpartnership.com

Table 1: The size of prospect pool by number of separate data sources

Source	Number of prospects
Combined de-duplicated file from three suppliers	100
Electoral roll 1 (ER1)	75
Electoral roll 2 (ER2)	74
Non-electoral roll	31
Gain from multi-sourcing	25

those companies with a direct marketing spend of, say, £5m+ are going to wish to manage a pool covering a large proportion of the UK adult population, although more focused prospect pools, eg for the 'grey' market, would be smaller in cost and scale.

The critical issue in terms of deciding whether or not to use a prospect pool rather than purchase lists is the ratio of the volume of direct activity driven by the data to the size of the pool. As that ratio drops so the data element of the communications costs rise. Although there is not a clear cut-off, a ratio of less than 0.5 would indicate that the data costs from a prospect pool may start to become too high.

The B-to-C sectors where prospect pools are normally found include financial services (particularly credit cards and investments), mail order and utilities. To the authors' knowledge they are only occasionally used in the business-to-business (B-to-B) market particularly by utilities. Within this business context this paper examines the pros and cons of buying the data for a prospect pool from one source versus using multiple sources (ie situations where a vendor buys in data from a number of suppliers on behalf of a client and consolidates the data into a single supply for the client).

The scope of paper is restricted to the B-to-C market, which is where the greatest volume of direct marketing is taking place.

THE IMPACT OF MULTI-SOURCING ON THE PROSPECT POOL'S FOOTPRINT

Before the introduction of opt-outs on the electoral roll, there was nothing very significant in the choice of electoral roll supplier other than price and service; the content was very similar. Since the introduction of opt-outs, however, as many as 20 per cent to 30 per cent of voters opt-out of allowing their details to be made available to marketers. This varies according to the geographical area and the way in which the opt-out clause is worded and presented to the recipient.

All the main suppliers of electoral roll data have made strenuous efforts to fill the gaps left by those opting out. This has been done by a number of different techniques including filling the gaps with more historic electoral roll data and overlaying other consumer list data sources.

The results, however, have not been even.

Using electoral roll data

For the first case study, the authors purchased and matched electoral roll data from two of the UK's largest suppliers and found that both parties provided a very significant number of names that were not available on the other's list. Names were also purchased from a third supplier who was not providing data based on the electoral roll but, nevertheless, was able to provide some unique names. The

Table 2: The performance of a response model when provided with data from different combinations of data sources

Source(s)	Top 10%	Top 20%
Non-electoral roll	4.72%	3.73%
Electoral roll 2 (ER2)	5.27%	3.93%
Electoral roll 1 (ER1)	5.50%	4.01%
ER1 + non-electoral roll	5.73%	4.12%
ER2 + non-electoral roll	5.89%	4.05%
ER1 + ER2	6.10%	4.38%
All three sources	6.63%	4.58%

results are shown in Table 1 (indices have been used).

This result is significant because the more complete the electoral roll, and hence the larger the footprint, the more higher-performing names will be available. If the prospect pool operator is using propensity models for selections, then he/she will have to drill less far down the modelled deciles for a given prospect volume.

THE IMPACT OF MULTI-SOURCING ON THE EFFECTIVENESS OF PREDICTIVE MODELS

This next case study, shows the impact of different data sources on the overall predictiveness of response models. One of the key arguments made by data vendors in favour of their products is that they better enable targeting of prospects. When using prospect pools, most targeting is driven by propensity models; In reality, only 17 per cent of prospect pool operators work without using any predictive models at all).¹

Prospect pools can be developed using data of many types from many sources although their foundation is normally electoral roll data. Typically this is then overlaid by data from one or more of the following categories:

- Geodemographic★
- Demographic/lifestyle★

- Co-operative databases (eg of pooled orderers or transactions)★
- Credit
- Suppression
- Market specific (eg house values)

Asterisks denote data types used in this case study.

In this exercise, the data from the three data sources featured in the footprint exercise in the previous section were matched. The response models were then developed using data from the combined sources neutrally; in other words the predictiveness of the data was allowed to determine which fields were going into which model rather than the authors trying to influence the model in any way.

The modelling technique used was logistic regression.

Table 2 illustrates the impact of different combinations of data sources on the power of one of the response models. In the table, data in the 'Top 10%' and 'Top 20%' columns are the levels of response predicted at 10 per cent and 20 per cent penetration, respectively, of the combined file created by the matching of the three sources.

The results of the exercise are enormously significant. They show that by purchasing data from all three sources, and combining them, the result obtained from the top decile is 21 per cent better than that obtained from the best of the

Table 3: The number of variables from each source used in response models for five different products on the same prospect pool

	Electoral roll 1	Electoral roll 2	Non-electoral roll	Internal
Model 1	4	4	1	3
Model 2	3	7	3	1
Model 3	3	6	2	4
Model 4	4	3	3	2
Model 5	5	6	2	2

Table 4: The impact on the cost per customer acquired of increasing the amount of data in a prospect pool purchased from the primary supplier

% of data purchased from the primary supplier	20	40	60	80
Indexed cost of acquisition per customer	75	100	100	125

single sources and 41 per cent better than that obtained from the worst of the single sources. At the second decile level, the equivalent results are 9 per cent and 16 per cent. Moreover, in this exercise there was some previous customer history data included in the model, which uplifted the result; if that had been removed, the differentiation would have been even stronger.

The development of the model was further complicated by the problem of missing data (where individuals appeared on only one of the two electoral roll sources). It was found that the most efficient overall solution came from building separate models for individuals from each of the three electoral roll source combinations: ER1 only, ER2 only and ER1 + ER2.

It is also worth considering how the different data sources contributed to the models themselves. Here this is illustrated by looking at — for each of five separate response models produced from the same prospect pool and for five different products — the number of variables from each source that came into the final model. This demonstrates how all three sources were required to give the best results and illustrates that there is really no such a thing as a dominant data

source; the loss of any one of them would have had a very significant negative impact. The results shown in Table 3 are for the five models developed on individuals appearing in both electoral roll sources, as well as the results for non-electoral roll and internal data. Evidence from the 2004 Tank Survey into prospect pools² also supports the hypothesis that multi-sourcing leads to better targeting, and because the information is presented at the level of the cost of customer acquisition, it also takes account of any negative cost implications of multi-sourcing.

The results from the Tank Survey show that as the percentage of data coming from the primary supplier increases, so does the average acquisition cost per new customer (Table 4).

It must be added that in this exercise, the sample size was not large enough for these figures to be treated as anything more than indicative of a trend in favour of multi-sourcing.

THE COST IMPACT OF MULTI-SOURCING DATA

Clearly the benefits of using multi-sourced data have to be set against the costs involved. At very least there are

additional administration, data purchase and data matching costs involved. In fact, the average UK spend per annum on data for a prospect pool is £700,000, with the median just over £250,000. The average prospect pool contains 12 million records and 370 fields. The average volume of mailings produced from a prospect pool is 14 million.³

The pricing of data in the UK market is not consistent — suppliers quote wildly varying rates for what appears to be similar data and all prices are subject to negotiation. The effect of this is that purchasers of multi-sourced data might actually pay less for their data than those who opt for single sourced data, because of their ability to off-set one supplier against another. On the other hand, they might pay more.

The additional data processing costs for multi-sourced data also need to be considered. Benchmark data processing costs are a third to a quarter of data costs.⁴ If the data costs £700,000 per annum, then one can expect the data processing costs to be around £200,000. The additional cost of multi-sourcing is very unlikely to be more than £50,000.

If, however, one takes an average mailing volume of 14 million, then assuming an uplift due to better targeting of 15 per cent and an average cost per mail piece of 40p, then the average benefit, from this cause alone, of multi-sourcing to a prospect pool operator will be £840,000. This benefit is expressed as the potential cost saving from fewer mailings, rather than the benefit of increased customer acquisition, which will be worth several times that

figure. Neither does it take into account the benefit of the enlarged footprint of the multi-sourced prospect pool; waving 25 per cent more profitable names to target is a very strong advantage to an acquisition programme. In the authors' opinion, the only scenario in which the benefits of multi-sourcing are not likely to outweigh the data and data processing costs are when the use of the prospect pool diminishes to an uneconomic level.

CONCLUSIONS

In this paper, the authors have set out the results of their examination of the impact of multi-sourcing data for a prospect pool. The principal benefits come from the increased footprint and better targeting. The main additional costs are for data and data processing. Overall, the economics from this particular case come out strongly in favour of multi-sourcing.

In reality, every case will be different and situations in which the benefits were greater or smaller than those discovered by the authors can be envisaged. Given the relatively low cost of trialling multi-sourcing and of developing response models from the data obtained, it would be unwise for any business concerned with managing a prospect pool not to test it.

References

- 1 Patron, M. (2004) 'Prospect Database Report' Tank.
- 2 *Ibid.*
- 3 *Ibid.*
- 4 *Ibid.*