Worldwide Readership Research Symposium 2005

HOMELAND SECURITY: ARE READERSHIP SURVEYS SAFE?

Roland Soong, Lindsey Draves & Hugh White, KMR/MARS

Synopsis

Using a database with magazine readership data and ailment conditions, we attempt to see if commercially available consumer databases can be used for effective predictive modelling.

Background

In the United States, the Department of Homeland Security has gone through a series of data mining projects by the names of CAPPS ("Computer-Assisted Passenger Pre-Screening System"), TIA ("Total Information Awareness", later re-branded as "Terrorism Information Awareness"), the MATRIX ("the Multi-state Anti-Terrorism Information Exchange"), and so on. The older CAPPS-I system is based on watch lists and security triggers (which include buying a one-way ticket and paying with cash) and flags about 15% of all passengers as suspects for further checking. This is considered to be an excessive inconvenience.

A key assumption behind some of the more recent projects is that large-scale consumer databases (Mena (2004), Chapter 2) can be used for predictive purposes to identify and interdict potential criminals. For example, Delta Air Lines, Jet Blue Airways and American, Frontier, Continental and America West airlines provided data to the Transportation Security Administration to be augmented by data from Acxiom in order to test the CAPPS-II system, which hopefully can reduce the flagging rate down to 5%.

Do we know that data mining works for homeland security? The performance characteristics of the Homeland Security projects are obviously classified secrets (see Jehl (2005)), but we assume (and hope) that there has to be some empirical basis behind the decisions to continue to invest in these projects.

If these large-scale consumer databases are such powerful predictors of behaviour, then perhaps their power can be harnessed for readership surveys. For example, can we accurately predict readership behaviour and product usage in their entirety through these large-scale consumer databases? If so, will we ever need to do another readership survey?

Before we even describe our empirical work, we can say with high confidence that we will never ever reach a point in which we can say who the readers of a magazine are (in the sense of producing the list of the 10 million people who read the last issue of a magazine). Or at least we hope that the time will never come, for it would be a predictable world with no free will or choice.

Study Design

The object of our study is to see the extent to which readership survey data are predictable from commercially available data on people. For our purpose here, we will use the 2005 MARS OTC/DTC Pharmaceutical Study. This is a mail survey with 21,216 adult respondents conducted during the first quarter of 2005. The relevant portions of the data here are readership for 100 magazines, 54 ailments within the past 12 months and demographic questions.

Our goal was to see how well we can predict magazine readership and ailment conditions within the total population. Those will be our dependent (or outcome) variables. We will be using three sets of predictor variables.

The first set of predictor variables will be demographic variables collected in the MARS survey itself (age, sex, race, education, household income, occupation, geographical region). They are the most commonly used demographic variables for profiling magazine audiences and ailment incidents, and we can assume that they are powerful predictors. Of course, information of this quality is not available for the population at large. However, we are using these predictors here in order to set up a baseline against which the other predictor sets can be compared.

The second set of predictor variables were obtained from Acxiom. We sent the names and addresses of our survey respondents for matching, and we received the InfoBase Premier package of data elements in return. The details of the subset of the data elements used as predictors here are given in Appendix A.

Although the list of data elements on InfoBase seems very extensive, we must point out that it is incomplete. For example, nobody knows how many people play state lotteries. This is one aspect about consumer database that people often misunderstood. Perhaps it would be better if the variable in the example were "flagged by the database compiler as a player of state lottery (which may or may not be true)" instead of "a player of state lottery." There will be errors of the 'false positive' and 'false negative' types. For predictive modelling, truth and accuracy in the predictor variables are immaterial – if variable X works wonders, then it is an effective predictor regardless of what it really means.

The third set of predictor variables were obtained from Acxiom and Claritas. From Acxiom, we obtained the Geo-Plus data elements. For each MARS respondent, we know the address (including a nine-digit postal code). Acxiom 'geo-codes' the address and assigns it to the finest Census geographic unit possible (block first; if not, then block group; if not, then tract; if not then zipcode). Then Acxiom attaches the Census data for that geographic unit. The details of the set of data elements used as predictors are given in Appendix B.

From Claritas, we obtained the PRIZM NE clusters. For each MARS respondent, we know the nine-digit postal code which Claritas has assigned to one of different PRIZM NE clusters. This clustering system has been used by many people for sales and marketing applications (see Turow (1997), p.44).

The Geo-Plus and Claritas variables represent aggregate-level data that can be determined solely from the address alone. They are therefore very convenient. When applied to individual persons, there may be considerable for error as neighborhoods are often mixed. But we repeat what we said for predictive modelling – if variable X works wonders, then it is an effective predictor regardless of what it really means.

Empirical Results for Magazines

First, we deal with the subject of magazine readership. Within the MARS study, the respondents are asked about readership of 100 magazine titles through a frequency-of-reading question (number of issues read out of last four issues). Thus, we have 100 different outcome (or predicted) variables.

For the reasons that were explained in Soong and de Montigny (2003), our predictive models were developed with a split-sample design. The MARS survey respondents were randomly divided into two halves, one of which was the training set from which a predictive model was derived. The other half is the validation set upon which the performance of the predictive model could be assessed. Generally, the predictive model will be over-fitted in the training set, and its true performance is more realistically reflected in the validation set.

For each magazine, we run stepwise linear regressions with the three sets of predictor variables. We then have sets of predicted values for the respondents. We have no interest in the predicted values themselves, but we only want to use them to rank our respondents from high to low. As explained in Soong and de Montigny, this means that the specific method (which can any of the many described in Hastie, Tibshirani and Friedman (2001)) has very little impact and that is why we chose the most computationally efficient method of linear regression.

For each set of predicted values, we sort the respondents into deciles (top 10%, next 10%, … , bottom 10%). Within each decile, we compute the magazine audience rating. Then we express the decile magazine audience ratings as an index with respect to the total magazine audience rating.

In Table 1 below, we show the indices in the deciles averaged across the 100 magazines.

Table 1. Mean Magazine Audience Indices by Predicted Deciles for Training Sample

	MARS demographics	Acxiom/InfoBase	Geo-Plus/PRIZM NE
Decile 1 (top)	260	276	215
Decile 2	167	161	148
Decile 3	130	130	120
Decile 4	101	196	109
Decile 5	87	88	96
Decile 6	73	69	84
Decile 7	60	57	75
Decile 8	50	46	67
Decile 9	38	35	57
Decile 10 (bottom)	32	22	38

The patterns in the Table 1 match our expectations. The top deciles where the most likely readers are have higher indices, and vice versa.

But at the outset, we have stated that we use a split-sample design because the performance characteristics are likely to be overstated in a training sample. We have more Acxiom/InfoBase predictors, and by sheer numbers alone, its apparent leadership position here may be due to over-fitting.

In Table 2 below, we show the results for the validation sample. We applied the predictor model obtained from the training sample to the validation sample to obtain decile indices.

Table 2. Mean Magazine Audience Indices by Predicted Deciles for Validation Sample

	MARS demographics	Acxiom/InfoBase	Geo-Plus/PRIZM NE
Decile 1 (top)	247	216	165
Decile 2	163	151	122
Decile 3	129	128	110
Decile 4	103	108	101
Decile 5	87	95	95
Decile 6	76	82	89
Decile 7	62	72	86
Decile 8	53	60	80
Decile 9	45	50	79
Decile 10 (bottom)	36	38	73

Comparing Tables 1 and 2, we see that there is a 'pull-back' or 'regression to the mean' for the decile indices. Whereas in Table 1, the Acxiom InfoBase had the highest average top decile index, which we suspected could be due to over-fitting, the MARS demographics now do better under split-sample validation.

On a magazine-by-magazine basis, MARS demographics has the largest top decile index 63% of the time, Acxiom InfoBase in 34% of the time and Geo-Plus/PRIZM NE in 3% of the time.

When does MARS demographics win? Obviously, they should be best for anything that profiles directly and strongly for age, sex, race, education, household income, occupation and geographical region. Thus, they are best for genres such as men, women, parent/child, ethnic and business/finance.

When does Acxiom/InfoBase win? Obviously, they should be best when there are some specifics that are not covered by broad demographics alone. Here, it would do well to scan through the list of data elements in Appendix A, especially the section on lifestyle, interests and hobbies. It should come as no surprise that Acxiom/InfoBase does well for the home service, epicurean, and golf titles. It also does well for *Arthritis Today* and *Diabetes Forecast*, since those ailments involve something beyond standard demographics.

Against the individual-level predictors, Geo-Plus/PRIZM NE really does well only in special situations, such as the regional magazines.

Empirical Results: Ailments

Within the MARS survey, the respondents are shown a list of 54 ailment conditions and they are asked if they have experienced them in the past 12 months. We repeat the same exercise with the magazine readership variables being substituted by the ailment states. In Table 3, we show the decile indices for the training sample.

Table 3. Mean Ailment Incidence Indices by Predicted Deciles for Training Sample

	MARS demographics	Acxiom/InfoBase	Geo-Plus/PRIZM NE
Decile 1 (top)	208	300	207
Decile 2	162	164	147
Decile 3	134	125	120
Decile 4	119	101	110
Decile 5	100	86	96
Decile 6	80	66	84
Decile 7	70	55	75
Decile 8	57	46	67
Decile 9	47	35	57
Decile 10 (bottom)	34	22	37

By now, we know well enough that the results obtained from a training sample is likely to be overly optimistic, especially in favour of those larger sets of predictors. In Table 4, we show the decile indices for the validation sample.

Table 4. Mean Ailment Incidence Indices by Predicted Deciles for Validation Sample

	MARS demographics	Acxiom/InfoBase	Geo-Plus/PRIZM NE
Decile 1 (top)	191	192	139
Decile 2	156	149	122
Decile 3	132	127	109
Decile 4	109	105	102
Decile 5	104	101	99
Decile 6	88	82	83
Decile 7	80	75	89
Decile 8	61	68	88
Decile 9	49	56	84
Decile 10 (bottom)	38	45	76

On an ailment-by-ailment basis, MARS demographics wins 51% of the time, Acxiom/InfoBase wins 45% of the time and Geo-Plus/PRIZM NE wins 4% of the time.

When does MARS demographics win? It wins when the ailment is clearly associated with the demographics, especially age and sex. Examples are age-related memory loss, arthritis, erectile difficulty, hangover, heart disease, menopause/hormonal replacement, menstrual cramps, osteoporosis, overactive bladder, post-traumatic stress disorder and yeast infection.

When does Acxiom/InfoBase win? It wins when the ailment involves some lifestyle elements beyond demographics. For example, anxiety (panic disorder), anxiety (social anxiety disorder), bipolar disorder, sleeping difficult/insomnia and depression involve psychological and social factors. Acid reflux, diabetes (insulin and non-insulin), food allergy, heartburn, high cholesterol, hypertension, nutritional deficiency, obesity and tobacco dependency are related to personal consumption habits and they are covered in the lifestyle section.

When does Geo-Plus/PRIZM NE win? It wins for only two ailments, and both can be sexually transmitted: herpes and HIV/AIDS. For those two ailments, knowing about the social environment of the neighbourhood of the individual appears to be more important than knowing about the personal characteristics of that person.

Conclusions

The objective of our empirical analysis was to see how commercially available consumer databases can be used effectively to predict magazine readership and ailment conditions. We used two such databases: the Acxiom/InfoBase demographic and lifestyle data, and the Geo-Plus/PRIZM NE data. Our conclusion was that the Acxiom/InfoBase can

be quite effective when compared to a benchmark of traditional individual-level demographic variables, especially in cases where the outcome variable involves social and lifestyle factors. By comparison, the aggregate-level Geo-Plus/PRIZM NE data were much less effective here.

The title of our paper referred to Homeland Security. Could consumer databases be powerful enough to predict magazine readership and ailment conditions? While we may have found a significant amount of predictive power here, the practical implications are quite limited.

Consider the best case for the Acxiom/InfoBase predictors. Erectile difficulty is a condition that afflicts 7.1 million adults in the United States. The top decile index is 426, which means that 42.6% (or 3 million persons) are in this top decile. However, there are 216 million adults in the United States, and the top decile is therefore 21.6 million persons. Thus, somewhere within those 21.6 million adults are 3 million who have erectile difficulty. From the opposite view, 18.6 million (or 86%) are 'false positives.' Thus, while consumer databases are useful, we should not expect them to be able to provide us with a list of completely accurate target individuals.

We note that under the legislation H.R. 4567, Department of Homeland Security Appropriations Act, 2005 (Public Law 108-334), it is required that "the underlying error rate of the government and private data bases that will be used both to establish identity and assign a risk level to a passenger will not produce a large number of false positives that will result in a significant number of passengers being treated mistakenly or security resources being diverted."

References

Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome (2001) The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag New York.

Jehl, Douglas (2005) Four in 9/11 Plot Are Called Tied to Qaeda in '00. New York Times, August 9, 2005.

Mena, Jesús (2004) Homeland Security: Techniques and Technologies. Charles River Media.

Singel, Ryan (2004) CAPPS II Stands Alone, Feds Say. Wired News, January 13, 2004.

Singel, Ryan (2005) TSA Work Sloppy, but Not Illegal. Wired News, March 26, 2005.

Soong, Roland and de Montigny, Michelle (2003) Foundations of split-sample foldover tests. Worldwide Readership Research Symposium. Cambridge, Massachusetts.

Turow, Joseph (1997) Breaking Up America: Advertisers and the New Media World. The University of Chicago Press: Chicago.

Appendix A: Description of Variables Used From Acxiom's Infobase Premier Product

- Age in two-year increments
- Sex
- Occupation (professional/technical, administrative/managerial; sales/service; clerical white collar; craftsman/blue collar; student; homemaker; retired; farmer; military; religious; educator.
- Education (completely high school; completed college; completed graduate school; attended vocational/technical)
- Number of adults present in household by sex and age range (Males 18-24; Males 25-34, etc)
- Children present in household by age range and sex (males 00-02; males 03-05, etc)
- Marital status (married; single)
- Working woman
- Credit card (bank card; gas/department/retail, travel/entertainment)
- Household income (Less than $15,000; $15,000 to $19,999; $20,000 to $29,999; $30,000 to $39,999; $40,000 to $49,999; $50,000 to $74,999; $75,000 to $99,999; $100,000 to $124,999; $125,000 or more)
- Home owner/renter
- Length of residence
- Dwelling type (single family/multiple family)
- Home equity available ($1-$4,999; $5,000-$9,999; $10,000-$19,999; $20,000-$29,999; $30,000-$49,999; $50,000-$74,999; $75,000-$99,999; $100,000-$149,999; $150,000-$199,999; $200,000-$249,999; $250,000-$499,999; $500,000-$749,999; $750,000-$999,999; $1,000,000-$1,999,999; $2,000,000 or more.
- Number of cars (1, 2, or 3+)
- Aggregate value of cars
- Type of vehicle
- Mail order buyer
- Mail order responder
- Mail order donor
- Lifestyle, interests and hobbies fashion; history/military; smoking/tobacco; celebrities; current affairs/politics; theatre/performing arts; community/charities; religious/inspirational; science/space; strange and unusual; career improvement; wines; arts; reading (general); reading (top sellers); reading (religious/inspirational); reading (science fiction); reading (magazines); reading (audio books); cooking (general); cooking (gourmet); cooking (low-fat); vegetarian; natural foods; travel (U.S.); travel (foreign); recreational vehicles; travel (family vacations); cruise vacations; running/jogging; walking; aerobic/cardiovascular; crafts; photography; aviation; auto work/mechanics; sewing/knitting; woodworking; board games/puzzles; home stereo owner; CD player owners; records/tapes/CDs owner; avid music listener; VCR/LD/DVD movie collector; cable TV; video game console player; VCR/LD/DVD player/owner; satellite dish owner; health/medical; dieting/weight loss; self improvement; pets (cat owner); pets (dog owner); pets (other); house plants; parenting; children's interests; grandchildren; racing spectator (auto/motorcycle); football spectator; baseball spectator; basketball spectator; hockey spectator; soccer spectator; tennis spectator; collectibles (general); collectibles (stamps); collectibles (coins); collectibles (arts); collectibles (antiques); investment (personal); investment (real estate); investment (stocks/bonds); PC owner; PC Internet/online service user; PC modem owner; PC game player; cellular phone owner; fishing; camping/hiking; hunting/shooting; boating/sailing; water sports; scuba diving; biking/mountain biking; environmental issues; tennis participant; golf participant; snow skiing participant; motorcycling participant; equestrian participant; home furnishings/decorating; home improvement; gardening; gambling (state lotteries); gambling (casinos); sweepstakes/contests; sports grouping; outdoors grouping; travel grouping; reading grouping; cooking/food grouping; exercise/health grouping; stereo/video grouping; electronics/computers grouping; home improvement grouping; investment/finance grouping; collectible antiques grouping.

Appendix B: Description of Acxiom's GeoPlus Data

For each record, there is a known street address with a nine-digit zipcode (=postal code). The street address would place the record into a particular Census block group, whose Census data are attached to the record. Where the street address is not known in the geo-coding system, an attempt is made to assign to a broader Census tract, whose Census data are attached to the record. If all of the above failed, the Census data for the larger zipcode is attached to the record. The general idea is to use the finest level of Census data possible.

The following variables were selected for this study.

- %male
- %white
- %black
- %persons of Hispanic origin
- Median age of population
- %population under 18
- %children under 7
- %children 7-13
- %children 14017
- %adults 65-74
- %adults 75+
- %single person households
- %average household size
- %owner occupied housing units
- %renter occupies housing units
- %single parent households
- %driving to work alone
- Mean travel time to work in minutes
- Median years of school
- %adults 25+ grade 0-8
- %adults 25+ some high school
- %adults 25+ completed high school
- %adults 25+ some college (no diploma)
- %adults 25+ associates degree
- %adults 25+ bachelors degree
- %adults 25+ graduate degree
- %adult males employed
- %adult female employed
- %management/business/financial operations
- %professional and related occupation
- %sales and related occupation
- %office/administrative support
- %households with income less than $15,000
- %persons below poverty level
- Median household income
- Median contract rent
- Median age of occupied structures
- %households with 1+ vehicles
- Median home values