

Database Paper

The IRI Marketing Data Set

Bart J. Bronnenberg

University of Tilburg, Warandelaan 2, Tilburg, The Netherlands, bart.bronnenberg@uvt.nl

Michael W. Kruger

Analytics Research and Development, Information Resources, Incorporated, Chicago, Illinois 60661,
mike.kruger@infores.com

Carl F. Mela

The Fuqua School of Business, Duke University, Durham, North Carolina 27708,
mela@duke.edu

This paper describes a new data set available to academic researchers (at the following website: <http://mktsci.pubs.informs.org>). These data are comprised of store sales and consumer panel data for 30 product categories. The store sales data contain 5 years of product sales, pricing, and promotion data for all items sold in 47 U.S. markets. In two U.S. markets, the store level data are supplemented with panel-level purchase data and cover the entire population of stores. Further information is available regarding store characteristics in these markets. We address several potential applications of these data, as well as the access protocol.

Key words: price; promotion; product; distribution; sales

History: Received: June 6, 2007; accepted: June 9, 2008; processed by John Hauser.

Introduction

This document outlines a broad, new consumer packaged goods data set available for distribution to academics in 2008. These data are intended to enable academic researchers to study important research topics in marketing and economics that are of concern to practitioners, policy makers, and scholars. In the following paper we detail the broad class of studies the data set was intended to facilitate, the structure of the data, and the process for disseminating data and topics to the academic community. More specific information on the variables and data can be found in the Technical Appendix at <http://mktsci.pubs.informs.org>.

Research Issues

IRI convened an assembly of industry practitioners from leading retailers and manufacturers and a number of academics to overview issues of potential interest that would guide the construction of these data. The result of this process was a list of relevant domains of interest to IRI, its clients, and academics. Table 1 overviews the research domains raised by these participants during the construction of the marketing data set.

One dimension in which these issues are categorized in Table 1 is by retailer/manufacturer. It is worth noting that Ataman et al. (2007) find that there is considerable variation in market shares over regions that can be ascribed to retailer behavior, and they call for more work in this area. By conglomerating data that span a plenitude of product categories

across multiple retailers and regions, researchers will be able to explore the factors underlying variation in retailer shares over time and space. Prior data sets, because they are more limited in scope, make such effects harder to identify. Likewise, on the manufacturing side, Bronnenberg et al. (2007) find considerable variation in manufacturer market shares and market strategies over space, and such effects will be easier to explain with these data.

The second dimension along which these issues can be categorized is internal/external, wherein internal factors are aspects of marketing strategy within the control of the firm and external factors are aspects beyond the control of the firm. It is in the latter area that we expect these data to lead to especially novel findings. For example, competition in the grocery channel is defined over many categories and regions, whereas most data are limited in terms of categories and regions. Store branding strategies are often implemented across categories and these branding strategies have long-term implications for differentiating stores, and Corstjens and Lal (2000) indicate that more research is needed on the effect of store brands on store competition across categories. These data will facilitate the exploration of such problems. The long duration of the data over a large number of categories will enable researchers to assess the potential biases arising from left-censoring in purchase or adoption data (Bronnenberg and Mela 2004). By integrating advertising data and many natural experiments on product additions and deletions, this data set will also enable researchers to obtain a more complete view of

Table 1 Research Domains for the IRI Data Set

	Retailer	Manufacturer
Marketing mix		
Product	Category management	Product management
Price	Market baskets	Brand management
Promotion		Umbrella branding
Place		
Marketing environment		
Customers	Trips	Usage
Competitors	Stores/channels	Firms
Geographies	Trade areas	DMAs, regional markets
Time	Long-term store equity	Long-term brand equity

the relative effects of marketing mix on brand equity and market share. Sriram et al. (2007) and Pauwels et al. (2007) have both called for more data to address these issues.

The various cells of Table 1 suggest other issues that can uniquely be addressed by these data, such as the effects of category management when retailers interact with firms across multiple categories. Likewise, it is possible to assess the benefits of a corporate branding strategy, whereas past research has been limited to two categories (Erdem and Sun 2002). Related, one can assess the degree of interfirm competition across categories. Overall, these examples suggest that the breadth of these data should prove useful in advancing our knowledge of marketing strategy beyond a limited number of categories, markets, and years. Examples of other papers using similar data to address related marketing problems include McAlister (2007), Ailawadi et al. (2007), and Nijs et al. (2007).

The Dimensions of the Marketing Data Set

The foregoing issues suggest several dimensions define the structure of the data: the number of items, the number of markets, the number of stores, the number of categories, and if panel data are included, the number of panelists. We discuss each of these below. Detailed information on measures and category breadth can be found on the Marketing Science website at <http://mktsci.pubs.informs.org>.

Items and Categories

A large number of categories enables the analysis of store choice, purchase behavior across categories, and market basket effects. Accordingly, the data cover 30 categories, chosen to maximize variety on a number of dimensions (see Table 2).

Having the full set of items (defined as UPCs or SKUs) affords insights into category management, the effects of attribute innovation, and assortment and line management. Accordingly, the data for each category will be comprised of a complete set of UPCs. For all UPCs, the data will include key information from the UPC data dictionary to obtain information about the attributes of these items.

Stores

A large number of stores ensures greater market coverage and allows one to explore spatial competition across stores and channel choice across store formats. Given potential issues regarding the effect of aggregating store level data to chain level, the data are available at the store level for each chain. Further information regarding the store (e.g., size, format), the demographics of its customers, and aspects of the location (to infer characteristics of the neighborhood) are available to help explain store choice and heterogeneity in behaviors across stores. The data include only chains and not independents, as the latter are less important for competition in most markets. Observations are drawn from IRI's national sample of stores, except for IRI's Behavior Scan Markets, wherein a population of stores is available along with information regarding their locations. These markets are useful for modeling store competition. Chains can not be identified by chain name. Rather, to maintain information about market structure and store ownership in geographic retail markets, each store is endowed with an IRI-generated unique chain-alias that maps one-to-one onto the chain names. That is, whereas there is no data about chain names, membership to masked chain names is part of the data.

Markets

A broad array of markets is desirable to explore issues such as product roll-out and differences in firm, retailer, and consumer behavior across regions. The cross-market analog to time series analysis is spatial analysis. As such, complete geographic coverage of the United States is desirable for estimating spatial and market effects, just as a complete time series is desirable for understanding time series effects.

There are a total of 64 IRI markets. These markets are geographic units defined typically as an agglomeration of counties, usually covering a major metropolitan area (e.g., Chicago, IL) but sometimes covering a part of a region (e.g. New England). Markets with the highest retailer concentration are not included in the data, as these markets would inadvertently reveal information about retailer chain names and information regarding their operations. In practice, to protect the confidentiality of these chains, markets in which the top grocery chain has more than 50% of the grocery market are omitted. This reduces the coverage to 47 markets. For similar concentration reasons, mass merchandiser data is not available in most markets.

Panel Data

Panel data are useful for understanding issues such as how loyalty patterns shift over time, how brand penetration is influenced by marketing, commonalities in behavior across categories, and store switching. Moreover, for chains such as Wal-Mart, the only information available about volume and in-store environment is from the panel data (having these data

Table 2 Product Categories

Category	Dollars per 1000 HH (\$)	Percent of HH's buying (%)	Purchase cycle, (days)	Perishability	Stockpilability	Percent of volume on any deal (%)	Average percent off price reduction (%)
Beer/ale/alcoholic cider	21,503	29.9	67	m	m	31.0	13.4
Carbonated beverages	76,567	91.9	40	l	m	58.2	23.6
Coffee	17,026	57.3	65	l	h	40.8	26.2
Cold cereal	46,555	87.2	48	m	m	43.4	30.7
Deodorant	6,020	53.4	94	l	h	35.5	28.0
Diapers	12,021	14.7	55	l	l	35.0	16.2
Facial tissue	8,611	59.9	70	l	h	38.9	25.1
Photography supplies	2,911	18.4	104	l	l	34.2	29.2
Frankfurters	9,896	65.8	82	h	l	46.9	31.1
Frozen dinners/entrees	51,552	80.3	51	l	l	40.7	25.7
Frozen pizza	19,087	63.4	64	l	l	50.2	26.3
Household cleaner	10,397	70.0	82	l	h	22.7	25.0
Mustard & ketchup	4,647	71.2	91	l	h	32.4	28.3
Mayonnaise	6,652	72.8	95	m	h	41.1	29.1
Laundry detergent	18,294	68.0	80	l	h	46.2	26.2
Margarine/spreads/butter blends	8,994	74.1	65	m	m	29.3	27.0
Milk	61,588	93.4	29	h	l	22.4	22.5
Paper towels	11,809	64.6	78	l	l	45.0	24.4
Peanut butter	6,311	61.0	82	m	h	32.9	25.3
Razors	1,258	9.2	87	l	h	34.0	20.6
Blades	4,448	28.5	106	l	h	20.3	21.5
Salty snacks	44,234	93.3	41	h	l	40.4	25.4
Shampoo	6,302	55.2	87	l	h	35.1	22.5
Soup	27,418	90.3	45	l	h	38.5	29.0
Spaghetti/Italian sauce	8,908	67.6	72	l	h	42.5	27.0
Sugar substitutes	2,731	21.7	82	l	h	14.4	23.2
Toilet tissue	24,189	75.3	67	l	l	45.4	23.9
Toothbrush	6,862	49.3	87	l	h	33.1	27.1
Toothpaste	7,997	62.8	89	l	h	40.1	25.8
Yogurt	23,556	71.8	50	h	l	34.7	24.3

Notes. Total U.S.—Grocery, drug, and mass excluding Wal-Mart. For 52 weeks, ending 6/25/2006. l = low, m = medium, h = high.
 Source: IRI Builders Suite.

overlap with census data would be especially helpful in this regard). Thus, panel data are included for the 30 categories in the two largest Behavior Scan markets using a yearly static sample over a 5-year duration. The yearly static nature of the sample means that for any given year, only households are included that have maintained in the panel for the duration of the entire 12 months. Panel recruitment and attrition are thus confined to the end-of-year time periods.

Duration and Sampling Frequency

The data cover is for 5 years, beginning January 1, 2001, to ensure a sufficient interval to analyze the impact of marketing strategy on brand performance. For example, the role of strategy on performance could pertain to either the effect of marketing strategy existing brands' equity or to the success of new brands (i.e., what factors ensure a successful launch). In addition, a long time horizon could also afford insights into how price elasticities vary over the life cycle of a product. It is intended that additional subsequent years will become available at later dates.

Advertising Data

Media data are available for two categories (salty snacks and beer). These data are available from TNS Media Intelligence (formerly Competitive Media

Reporting, CMR). The advertising data cover 16 measures at the market/brand/month level covering different media channels and different advertising weight metrics.¹ For salty snacks, advertising data are available from 2001 to 2005 and for most markets. The beer data cover the same markets and are available for 2001.

Process

Data Distribution and Conditions of Use

The distribution of the data is controlled to ensure its use is consistent with its mission of providing academics with the information to test their ideas and engage in new research. Upon visiting the *Marketing Science* website at <http://mktsci.pubs.informs.org>, users will be directed to a Web page maintained by IRI with additional instructions on obtaining these data. Users will be asked to download, sign, and return an agreement to IRI intended to exclude unapproved uses of the data, including for any commercial

¹ Total dollars, total units, spot TV dollars, spot TV units, newspapers dollars, newspaper units, Hispanic newspaper dollars, Hispanic newspaper units, national spot radio dollars, national spot radio units, local radio dollars, local radio units, U.S. Internet dollars, U.S. Internet units, outdoor dollars, and outdoor units.

purpose such as reselling or consulting, factual misrepresentation of IRI products, or mentioning retailers by name. Users must provide to IRI their academic affiliation. Users will further be asked not to distribute the data and to use it only for the project for which they register. This agreement is strictly between IRI and the authors using these data. *Marketing Science* and INFORMS play no role in this agreement or its enforcement. Failure to adhere to these policies will result in termination of privileges to use the data.

Upon receipt of this agreement, IRI will mail an external hard drive with the data to the academic researchers. For users of the database, there will be a nominal fee of \$500 payable directly to IRI to cover the administrative costs associated with archiving, distributing, and adding future years to the data. Without these funds, the likelihood of being able to obtain resources to update the data as more years become available is decreased.

At the request of IRI and its key clients, the most recent 2 years of the data will not be released. The initial data is from January 1, 2001, through the end of 2005. Each year that passes from the release of the data, an additional year of data will be added—one for download and one into the 2-year buffer. The latency in the data protects the propriety of the most recent data to IRI and its clients while ensuring that the data will eventually become available to the academic community.

Finally, the authors will open and maintain a separate online discussion group site in which we envision providing a set of basic scripts to read and process the data in SAS programming language across a simple menu of keys. It is requested that a new aggregation code, developed at individual users' needs, be sent back to the online discussion group to aid replication and extension of work published from these data. Details on accessing this site are to be provided on the database website (which can be accessed by using the *Marketing Science* link to the IRI Data Set, <http://mktsci.pubs.informs.org>).

Distribution of Papers to IRI

Keeping track of projects is desirable for the second aspect of process, namely, feeding back results to IRI and its clients. Users will be asked to make a working paper (prior to journal submission) and a final copy of the research available to IRI for its internal use, and IRI can use this list of projects to follow up on this opportunity. IRI intends to collect papers in a searchable data set on the website and categorized by the

key words on the website. This library will provide a cutting-edge resource to IRI and its clients. This does not limit the researcher in any way in making the paper available through other means of distribution or publication, except that IRI receives the opportunity to verify compliance with the terms of use. Authors should note that IRI will play no role in the editorial review process at *Marketing Science* or any other journal.

Summary

In sum, IRI makes available to the academic community a data set spanning 30 categories, 47 markets, and 5 years of weekly data. This is one of the largest consumer data sets ever to be released. It is our fervent hope that the broad use of these data will spark a major acceleration in research that will yield dividends to the practice of marketing for many years to come. We are especially grateful that IRI has had the vision for and made the effort to support these data and the attendant innovations they will bring.

Acknowledgments

The authors are grateful for the support and inspiration of IRI, Daniel Pagni, Arvid Johnson, Sunil Garga, members of Information Resources Incorporated Analytics Advisory Board, the Marketing Science Institute, and Gaurav Bhalla at TNS. Any errors in this manuscript are the sole responsibility of the authors, not of IRI.

References

- Ailawadi, K., B. A. Harlam, J. César, D. Trounce. 2007. Quantifying and improving promotion effectiveness at CVS. *Marketing Sci.* 26(4) 566–575.
- Ataman, B., C. F. Mela, H. J. van Heerde. 2007. Consumer packaged goods in France: National brands, regional chains, and local branding. *J. Marketing Res.* 44(1) 14–20.
- Bronnenberg, B. J., C. F. Mela. 2004. Market roll-out and retailer adoption for new brands. *Marketing Sci.* 23(4) 500–518.
- Bronnenberg, B. J., J.-P. Dubé, S. Dhar. 2007. Consumer packaged goods in the United States: National brands, local branding. *J. Marketing Res.* 44(1) 4–13.
- Corstjens, M., R. Lal. 2000. Building store loyalty through store brands. *J. Marketing Res.* 37(3) 281–291.
- Erdem, T., B. Sun. 2002. An empirical investigation of spillover effects of marketing mix strategy in umbrella branding. *J. Marketing Res.* 39(4) 408–420.
- McAlister, L. 2007. Cross-brand pass-through: Fact or artifact? *Marketing Sci.* 26(6) 876–898.
- Nijs, V. R., S. Srinivasan, K. Pauwels. 2007. Retail-price drivers and retailer profits. *Marketing Sci.* 26(4) 473–487.
- Pauwels K., S. Srinivasan, P. H. Franses. 2007. When do price thresholds matter in retail categories? *Marketing Sci.* 26(1) 83–100.
- Sriram, S., S. Balachander, M. Kalwani. 2007. Monitoring the dynamics of brand equity using store-level data. *J. Marketing* 71(2) 61–78.

The data set described in this paper is maintained by IRI and available through <http://mktsci.pubs.informs.org>. Any fees charged by IRI for the distribution of the data set will be used for the continual maintenance and updating of the data. Scholarships to cover IRI's fees (for those who need it) are available through the INFORMS Society for Marketing Science (ISMS). Please see the website above for further details.