REVIEWED PAPERS

ORGANIZING SCAN DATA FOR MARKET RESEARCH

by

David B. Eastwood, Professor*
Department of Agricultural Economics and Rural Sociology
The University of Tennessee
Knoxville, TN 37901-1071

Abstract

Two areas of practical problems in creating scan data sets for demand and marketing research are discussed. The first pertains to organizing scan data for variable weight items into consumer demand categories. The second is a set of problems associated with the creation of an advertising data set that can be merged with scan data to assess marketing strategies.

Introduction

The introduction of scanners into the supermarket checkout process has received a lot of attention in the popular press, food marketing publications, and research journals. Significant benefits were expected. Immediate gains included decreased checkout time, increased accuracy, decreased theft, better estimates of shrink, and checkout clerk efficiency. Longer term benefits were to be managerial in nature, such as better scheduling of labor, improved ordering and inventory control, shelf space allocation, direct product profit, and marketing.

New analytical approaches to demand research are possible with these data. Relationships among substitutes and complements can be examined to obtain better estimates of the trade offs consumers make when selecting food. The impacts of advertising and other promotional strategies can be examined (Capps, 1987). Capps (1988) has suggested a way of organizing scan data for managerial decision making. Capps (1989), Lesser and McLaughlin, and Jensen and Schroeter have examined some ways of using such data for demand analyses of variable weight items. Limited progress has been made in achieving the long-term benefits (Food Marketing Institute, Lesser and McLaughlin, and Wittink et al.). Reasons for this situation have been explained elsewhere (Capps, 1987, Eastwood). The end result is that many supermarket chains provide scan data to data processing firms that turn around and sell processed data back to these chains as well as to manufacturers, distributors, and other special interest groups.

While data processing firms provide a valuable service, supermarkets could perform this function internally. One advantage of doing so is the ability to understand the data and control how they are manipulated. Having data that were not generated by a "black box" should facilitate

*The author has benefited from the useful comments of Morgan Gray and John Brooker.

comparisons among stores within the same chain, identify trends at the store and more aggregated levels, and the monitoring of management practices. A chain's ability to analyze its own data also serves as an independent check on manufacturers' and food distributors' claims about sales.

The organizational structure of a supermarket chain reflects the evolution of this type of outlet. Management tends to focus on short-run decision making such as the ordering of foods, inventory control, labor requirements, and advertising (Capps, Carmen and Figueroa). Merchandising is divided into traditional areas of responsibility that have arisen due to managing the flow of foods into the supermarket. But related products from the consumer's perspective, such as meat, are found in several departments. The core of successful food marketing is attention to consumer demand. After all, these stores are providing goods and services to food shoppers. Thus, there is a need to have the capability of analyzing consumer demand via scan data in ways that are amenable to the corporate structure.

This paper discusses two kinds of practical problems encountered when individual chains generate scan data sets. The purpose is to acquire data suitable for long-run managerial decision making. The first set of problems occurs when implementing Capps's framework which focuses on areas of managerial responsibility. This approach organizes scan data and analyzes them for policy decisions at the corporate level through operational decisions at the store level. Problems here include: 1) lack of a methodology to identify product groups from the consumer's perspective, 2) adjustments for missing data, 3) creation of an analytical framework for item movement, and 4) alternatives for grouping data. The second set of problems is associated with the creation of a marketing/promotion database that can be merged with the scan data to assess marketing strategies.

Attention has centered on meat products for several reasons. Fresh meats make up a large proportion of the product group. They are variable weight items which have been very difficult to incorporate into scan databases. Furthermore, variable weight items are important components of supermarket revenues and profits (*Supermarket Business*). Another factor is that meats are found in many places throughout the supermarket, thereby necessitating methodologies for combining products across departments.

Scan Data Problems

Weekly scan data have been received from five stores in the Knoxville area. They are outlets of a single chain. These are raw data in the sense of representing information sent by all stores with scanners to corporate headquarters. Item movements and prices for all bar codes are included. The weekly format is suitable from both consumer and managerial decision making perspectives. No adjustments or manipulations have been made to the data prior to receipt. Thus, there are no unanswered questions as to how missing data were handled, new products included, products discontinued, or product group sums generated. The data are comparable to those typically received by corporate management.

Initial work centered on developing programs for managing the data and identifying meat products. The codes were written for the Statistical Analysis System (SAS) because both parties had access to it. A side benefit is that the programming steps can be easily replicated by any SAS user.

The first step was to identify all meat products. Since the bar codes, which are read by scanners, are not ordered by product, there was no way to sort the data on this basis. Rather, the chain's list of departments associated with inventory/accounting records was used to list all products in meat related departments on the first tape received. Then, these products were regrouped according to the way agricultural economics researchers have traditionally grouped meats to conduct demand analyses. This initial list contained approximately 400 items. Some were deleted because they were not meat products even though they were listed under meat inventory categories. Consumer demand category codes were given to each product. In this way a master list of meat products was begun. Bar codes on subsequent tapes for the same inventory groups

were compared to the master list, and new bar codes were added and given a demand category. An important point is the fresh meat codes are only UPC-like codes. That is they are bar codes the chain has assigned to these products for stores in the Knoxville area. (They are sometimes called price lookup codes.)

Product descriptions and the sizes of products are included on each tape. Anticipating a need for products grouped on the basis of size and price per unit, the sizes were standardized into ounces. It is very difficult for computers to read and interpret combinations of letters, numbers, and other symbols. Sizes found on the tapes contained a variety of forms. The easiest way to generate a standard measure was to convert to the number of ounces as an additional piece of information with the master list. Size information was manually inserted when a demand category was assigned to each meat product.

Missing data cause difficulties when conducting analyses. This problem arose when a store's weekly sales were transmitted to corporate headquarters. If the data were missing due to store level failures, then the data were lost. If the problem was with the communication network or corporate level malfunctions, then the stores held on to the data, kept accumulating them, and then transmitted the combined sales the following week. How should this situation be coded? There are several sophisticated statistical approaches. The one used here is quite simple and is motivated by an effort to minimize the distortion. Missing weeks are left missing. Those weeks that have accumulated totals are divided by the number of weeks involved, and the result is just used to replace the one week. This imputation method is felt to be better than replacing all missing values with incorrect data.

The other major difficulty centered on variable weight items. Fresh meats do not currently have UPCs. Bar codes have been given to these products by the chain. As one reads a code from left to right, the five right most digits are the price of the package. In order to combine like products into weekly totals, the cost information is stripped, and the data only record the number of times the scanners read the bar code for a particular cut. Programming and hardware changes to permit the expenditure information to be stored have not occurred. However, we have worked on developing a framework that permits the use of item movement for demand and marketing analyses. Discussion of the approach is available in Eastwood, Gray and Brooker.

With the master list of products and the weekly files of scan data, an historic record is being generated with uniform sizes, item movement, and demand category definitions. The capability of looking at any group of meat products for any subset of weeks is now possible. Furthermore, the methodology can be extended to other product groups or stores.

Advertising Data Problems

The second area of practical problems has to do with the creation of a data set for the chain's advertising. Several media are involved: newspaper, television, ratio, and some point-of-purchase. They are all considered because food shoppers are exposed to all of them. The goal is to relate these weekly advertising efforts to item movement. This process is time-consuming because advertising and promotions are not tied directly to the bar codes. Products are not identified by their codes, and price change instructions are programmed via inventory codes which are unrelated to bar codes. Thus, there is no direct link between advertised products and scan data. The preliminary focus is on the effects of these types of advertising and promotions on the demand categories.

Most of the chain's newspaper advertising is through supplements. Various measures of advertising--the page, color versus black and white, and square inches--are recorded for meats. These are combined across products within each demand category to form weekly totals. Gross rating points are the radio and television measures. The chain provides information on the type of ad and products involved. These are also recorded by demand category. Later on, as needs dictate, bar codes could be attached to selected product advertising in these media.

Some point-of-purchase data are also being collected. For selected fresh meat products, the sizes of in-store signs are recorded. However, with respect to a presence/absence measure, there is always at least one bar code within each demand category, so no variation in this measure occurs for the groups.

Results

Seven inventory categories contained meat products (canned meat, canned fish, frozen prepared foods, diet and health foods, prepared foods, meats, and deli). Table 1 displays the consumer meat demand categories and the number of bar codes for each group as of September 29, 1989. A brief description of the categories is provided as well. The table shows a large number of products within each group. Altogether 1,744 meat bar codes had been identified between May 14, 1988 and September 29, 1989.

Illustrative summary information for a demand category for four weeks is displayed in Table 2 for each of the five stores. The table illustrates the potential of scan data by showing that even elementary information can be very enlightening. The "Total" columns reflect total item movement for a demand category for each store. By displaying the minimum and maximum item movements within a group, one can note that some products are not moving very well. For example, in week one store A's scanners read 1,641 bar codes for this group. Among individual bar codes within this demand category the lowest item movement is one and the highest is 68. Clearly, some products are not moving well. It would be relatively easy to identify these products, track them for a few weeks and use the data for shelf space allocation decisions. Summary data also provide a quick view of one store's sales vis-a-vis other selected stores.

Figure 1 displays the item movement for one of the demand categories. It is a five-store average. Such diagrams can be generated easily via SAS or spreadsheet programs. Trends and seasonal factors can often be observed.

The various advertising measures are described in Table 3. They reflect the aggregations of the media measures across individual meats according to their respective demand categories. The electronic media values are just the sums of the gross rating points. Newspaper and point-of-purchase measures accommodate multiple ads because the chain often promotes more than one product within a demand group.

Table 4 illustrates the advertising data. Inspection of these data over the entire period suggests substantial variation. As the number of weeks increases, the possibility of relating these measures to demand category item movement may allow for an examination of the impacts of promotion alternatives or sales.

Figure 2 summarizes the potential impact of scan data. Through careful creation of a master list, grouping schemes for inventory management and food sales can be matched. A second advertising data set can also be established. Management could have the capability to do its own marketing analyses. The chain also can have an independent data source for responding to vendors' claims and data. There also is the capability of tracking the impacts of management policies on customer sales. Data for shelf space allocation and for direct-product-profit estimates can be generated using this approach. Consequently, the long-run benefits for introducing scanners could be achieved.

References

Capps, Oral Jr. "The Ultimate Source for Demand Analysis?" Working Paper, Texas A&M University, 1987.

_____. "Defining Management Responsibilities in the Retail Food Distribution Industry," *Journal of Food Distribution Research* 19(1988):29-34.

_____. "Using Scanner Data to Estimate Retail Demand Functions for Meat Product," *American Journal of Agricultural Economics*, 71(1989):750-60.

Table 1

Meat Demand Categories

Category	# Bar Codes	Category	# Bar Codes
Bacon	35	Lamb	19
Beef ground	14	Lunch[c]	182
Beef meal[a]	411	Meat (default)	4
Beef roast	27	Meat meal[d]	13
Beef steak	47	Other beef[e]	35
Bologna	50	Other chicken[f]	57
Chicken	46	Other pork[a]	73
Chicken meal[a]	176	Pork	38
Deli[b]	1	Poultry	25
Fish	190	Poultry meal[a]	36
Franks	60	Sausage	111
Ham	44		

[a]Prepared (nonfresh) and contain vegetables.

[b]All deli items are stored in this default code.

[c]Meats that are primarily used as cold cuts.

[d]Prepared (nonfresh) foods that do not specify type of meat.

[e]Veal and miscellaneous beef cuts.

[f]Prepared (nonfresh) chicken.

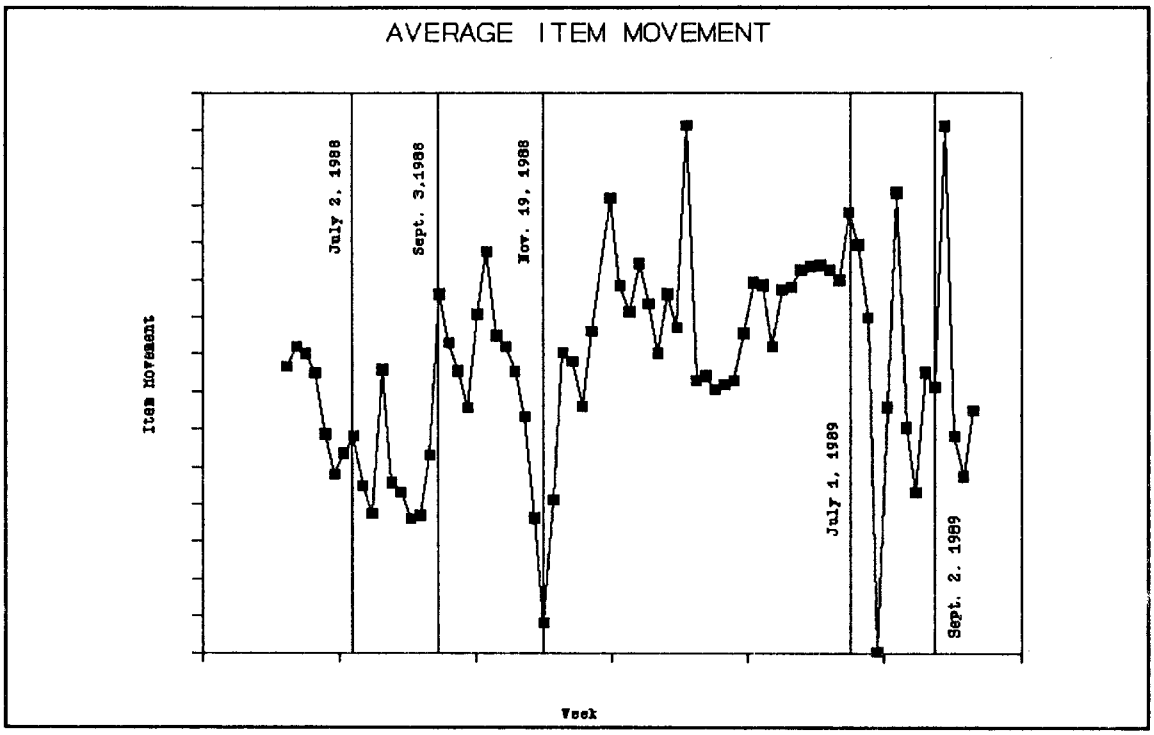
Table 2

Illustrative Weekly Item Movement Summary Information for a demand category[a]

	Store A			Store B			Store C			Store D			Store E		
Week	Total	Min	Max	Total	Min	Max	Total	Min	Max	Total	Min	Max	Total	Min	Max
1	1,641	1	68	1,895	1	106	1,301	1	67	684	1	83	missing		
2	1,738	1	70	1,255	1	62	missing			680	1	85	1,491	1	42
3	2,100	1	77	missing			1,440	1	61	748	1	61	1,712	1	79
4	1.170	1	46	missing			836	1	48	441	1	48	912	1	44

[a]These are not actual item movement data. The numbers have been changed to protect the confidentiality of the participating chain.

Figure 1



Carman, H. F. and E. E. Figueroa. "An Analysis of Factors Associated with Weekly Food Store Sales Variation," *Agribusiness*, 2(1986):375-90.

Eastwood, D. B. "Potential Consumer Benefits of Scanner Data: The Case of Supermarkets," *Proceedings of the 35th Annual Meeting of the American Council on Consumer Interests*, M. Carsky, editor, 1989:14-17.

Eastwood, D. B., M. D. Gray, and J. R. Brooker. "Using Scan Data to Evaluate Advertising: A Suggested Methodology and Preliminary Results," paper presented at the Annual meeting of the Southern Agricultural Economics Association (February, 1990.)

Food Marketing Institute, *Retailer Applications of Scanning Data*, 1985.

Jensen, H. and J. Schroeter. "Estimating Retail Beef Demand Using Household Panel Scanner Data," paper presented at the Annual Meeting, American Agricultural Economics Association, 1989.

Lesser, W. and J. Smith. "The Accuracy of Supermarket Scanning Data: An Initial Investigation," *Journal of Food Distribution Research* 17(1986):69-74.

Supermarket Business, *Supermarket Business* (September 1989).

Wittink, D. R., M. J. Addona, W. J. Hawkes, and J. C. Porter. "The Estimation, Validation, and Use of Promotional Effects Bases on Scanner Data," Working paper, Cornell University, Cornell University (1988).

Table 3

Advertising Measures for Meat Demand Categories

Media	Variable	Description
Television and Radio		Rating points for specific products for each type of commercial. If more than one product in a demand category was advertised, the sum was used. Categories are mutually exclusive.
	RU	Roll-up.
	TVP	TV personality for fresh meat products.
	BIGI	Buy-one, Get-one-free.
	FF	Frozen foods.
	H	Holiday season specials.
Newspaper		Weekly supplements are the primary vehicle.
	PAGE	Page on which a product appears. No ad=0, front=1, middle=2, other=3, both front and middle=4, other plus front and/or middle=5, and regular paper=6.
	SPACE	Square inches.
	COLOR	One black and white=0, one color=1, more than one black and white=2, more than one color=3, black and white, and color=4, no ad=5.
Point of Purchase	TAG	Whether a shelf label denoting a special price is present.
	SIGN	Square inches of special product signs.

Table 4

Advertising Data for Demand Categories for One Week

	Newspaper			Television						Point of Purchase	
Demand Category	Color	Space	Page	RU	TVP	F	BIGI	H	RADIO	TAG	SIGN
Bacon	1	20.00	M							1	
Beef Ground	0	2.44	M							1	42
Beef Meal	3	31.88	O			300				1	
Beef Roast	3	23.50	B							1	60
Beef Steak	4	23.50	M							1	60
Bologna	1	11.38	M							1	
Chicken	4	62.50	C	200						1	
Chicken Meal							300			1	
Deli	2	29.76	O							1	
Fish	4	55.94	C							1	
Franks	3	21.94	M							1	
Ham	0	2.44	M							1	
Lamb										1	
Lunch										1	
Meat Meal										1	
Other Beef	2	5.69	M							1	
Other Chicken	0	3.25	M							1	
Other Meat										1	
Other Pork	0	3.25	M							1	
Pork	1	11.38	M							1	
Poultry	0	3.25	M							1	
Poultry Meal										1	
Sausage	3	35.25	B							1	

Figure 2

Consumer-Supermarket Linkages: In-House Scan Data