EMERGING DATA ISSUES IN APPLIED FOOD DEMAND ANALYSIS

Proceedings of a Workshop Held by the S216, Food Demand and Consumption Behavior Regional Commitee

October, 1993

David Eastwood and Benjamin Senauer, Editors,
Department of Agricultural Economics, The University of Tennessee
and
Department of Agricultural and Applied Economics, The University of Minnesota

TABLE OF CONTENTS

AUTHOR AFFILIATIONS

Margaret Andrews is an Economist in the Food and Nutrition Service, U. S. Department of Agriculture.

Peter Basiotis is an Economist/Chief Diet Appraisal Research Branch in the Human Nutrition Information Service, U. S. Department of Agriculture.

Oral Capps, Jr. is a Professor in the Department of Agricultural Economics, Texas A and M University.

Wen Chern is a Professor in the Department of Agricultural Economics and Rural Sociology, Ohio State University.

David Eastwood is a Professor in the Department of Agricultural Economics and Rural Sociology, University of Tennessee.

Mary Hama is an Economist in the Food Consumption Research Branch, Human Nutrition Information Service, U. S. Department of Agriculture.

Helen Jensen ia an Associate Professor in the Department of Economics, Iowa State University.

Vickie McCracken is an Associate Professor in the Department of Agricultural Economics, Washington State University.

David Price is a Professor in the Department of Agricultural Economics, Washington State University.

Dorothy Price is a Professor in the Department of Human Development, Washington State University.

David Smallwood is an Economist in the Economic Research Service, U. S. Department of Agriculture.

EDITORS' NOTE

This Tennessee Experiment Station Bulletin is the edited collection of seven papers presented by members of the Changing Patterns of Food Consumption (S216 Regional Committee) at a 1993 Workshop held by the Regional Committee. They focus on a variety of emerging issues associated with data sets used in applied demand analysis. These pertain to topics that are not discussed in the extant literature but are quite germane to the extension of empirical models of food consumption.

CHARACTERISTICS OF SUPERMARKET SCAN DATA AND THEIR IMPLICATIONS FOR APPLIED DEMAND ANALYSIS

David B. Eastwood[1]

Before one can consider the feasibility of using scan data for empirical demand analysis, it is necessary to have a clear understanding of what they are. The purpose of this paper is to provide an overview of the different types of scan data that are available. Once this has been completed, readers should be able to make a decision as to whether familiarity breeds contempt or beauty is in the eye of the beholder with respect to scan data and their relevance for applied demand analysis.

Bar Code Structure

Much of the confusion about scan data stems from confusion about bar codes. The uniqueness of bar codes and the information they contain have to be recognized in order to understand the characteristics of scan data. The present discussion is limited to bar codes that are used at the retail level. Particular attention is given to bar codes for food items.

Bar codes have two visual components. One is a series of alphanumeric characters that a person can read. The other is a corresponding set of rectangular printed bars and spacing, both of varying widths, that can be interpreted by optical scanners. A Universal Product Code (UPC) is a twelve digit bar code that conforms to standards established by the Uniform Code Council.

An illustration of the UPC structure is shown in Figure 1. Letters A-L represent the 12 digits of a UPC. The left-hand most digit (A in the figure) is called the number system character. If it is a 0, 6, or 7, then the respective UPC is a typical UPC. A 2 indicates it is a variable weight product. These are products, especially common in fresh meats, fresh produce, and deli departments. (The coding protocol for variable weight products is outlined subsequently.) UPCs beginning with a 3 denote drugs. When the initial digit is a 4, the corresponding product has in-store markings. Coupon

---

[1]David B. Eastwood is a Professor in the Department of Agricultural Economics and Rural Sociology, The University of Tennessee. Morgan D. Gray provided helpful comments during the preparation of the paper.

bar codes begin with a 5. Undesignated system characters are 1, 8, and 9.

Fixed Weight UPCs. For regular bar codes the remaining sequence of eleven digits is broken into two series of five digits and a final digit. The five digits (B-F) immediately to the right of the number system character designate the manufacturer of the respective product. The set is specified by the Uniform Code Council and is unique. Each manufacturer uses the next five digits (G-K) to designate specific products that it manufacturers. Assignment of these numbers is at the discretion of the manufacturer. The last digit's value (L) is based on the preceding eleven. It serves as a way of checking to make sure a scanner has interpreted the car code correctly.

Variable (Random) Weight UPCs. Variable weight UPCs have a different structure, which is displayed in below. Aside from beginning with a 2 and from having the twelfth digit be a check character, the remaining ten digits have the following configuration. The second digit from the left (B) identifies the packer. Values of 0-3 are for the retailer to assign, and values of 4-9 indicate the product was packed by a vendor. The next four digits (C-F) designate the product. However, the Uniform Code Council has only assigned ranges of values for products. Commodity groups, trade associations, etc. have the responsibility for assigning values within the ranges. To date these have not been standardized. Thus, retailers can designate specific items within the ranges at their own discretion. The seventh digit (G) is another internal check whose value is determined by the price of the respective product. The value of the package is contained in the next four places, with the decimal point assumed at two places.

Two key points follow directly from the present configuration of bar codes. One is that UPCs are not assigned according to a numerical scheme that permits sorting them in a way that conforms to food groups. Therefore, one usually has to resort to other ways of locating foods of interest for a particular study when using a scan data base. The other point is that the type of information a variable weight bar code contains is quite different from that of a fixed weight code. Although scanner software uses the value of

the package information, the computer programs that process the data beyond preparing customer bills vary considerably with respect to the amount of information they continue to process.

There is one further complicating factor. Cash registers have special keys to process frequently purchased items. The reason for these keys is to speed up the checkouts. They are called price look-up units, PLU, (or codes, PLC). Often these simplified codes become part of a scan data base rather than the corresponding UPC. The values of the PLUs are controlled by the respective retailer and are not standardized across chains.

Types of Scan Data

From the outset it is useful to bear in mind that scan data are not generated by retailers for the purpose of allowing economists to estimate demand relationships. Scanners were introduced to speed up and increase the accuracy of the checkout process. Capturing the data for subsequent analyses is secondary. Although these data are generated automatically, their subsequent use in operating food retail outlets varies considerably, and additional effort is required to transform the information into a data base that is suitable for estimating demand relationships. Furthermore, management considerations, including confidentiality and conflicts within the organizational structure, may inhibit a retailer’s ability to provide data to researchers.

Four points comprise the foundation for an understanding of scan data. First, there is no single type of scan data. Rather, the information that scanners capture and transfer to computer storage devices can take various forms. Second, the popular press has overstated the advances that have been made with respect to managerial decision making, especially in the area of food retailing. Third, there are some characteristics of bar codes that make the data base management of some products very tricky. Many of these characteristics are found in foods. Finally, the amount of information contained in retail bar codes varies by product, and there are differences in the capabilities of computer software that interprets the codes. These last

two observations, taken together, mean the amount of information that is transferred from the scanner through subsequent storage devices can changes considearbly. Information can be added or deleted during the data processing steps at the various stages.

Scanners read the bar codes of items that consumers want to purchase. Fixed weight UPCs are matched with those in a price file to generate customer bills. Variable weight UPCs have the values of the packages inbedded in the bar codes (H-K), and they are used directly in generating bills. Since computers are used, the information is automatically in a form for further processing and storage. The fundamental division for scan data occurs at this point -- either the data are stored by customer, or they are stored by bar code.

An easy way to conceptualize the different types of scan data is portrayed in Figure 2. It identifies the major places or ways in which scan data are held. The bottom path pertains to customer specific records, or the bundles of items that food shoppers purchase. Once a customer's bill is generated, the computer software marks the record and transfers the information to in-store computers. At regular intervals the customer specific records are uploaded from the outlets to the central management information system. These data may then be sold to market research companies (vendors) that further manipulate the data and sell the information and corresponding analyses. Behaviorscan and Infoscan are two examples. The most recent extension of customer records is the implementation of electronic benefit transfer programs (EBT) for WIC and Food Stamps. Because these are demonstration programs, a dashed line is used in the figure. In addition the EBT data at present do not cover all foods and only capture total program related expenditures.

Customer purchases can also be added to running totals of each product carried by the retail outlet, and they become store level sales records. These totals can be either the number of times each of the bar codes (called item movement) is read by a store's scanners or in the case of variable items, they

can be accumulated item movements or package values. The flow is reflected in the upper path of Figure 1. Further processing of these data may occur as they enter the management information system. In addition to some software only keeping item movement, bar code activity may be placed in more aggregated categories called default codes. These data also can be sold to vendors which analyze them and sell the information. Examples of data that can be added are merchandising codes. They include variables for newspaper, in-store coupons, and point-of-purchase displays. Other information about the store can also be incorporated.

Scan Data and a Demand Equation

The functional form of a demand equation may be expressed as

$$q_{ijst} = f(Y_{it}, p_{i1st}, \ldots, p_{i1St}, \ldots, p_{iJSt}, V_{it}) \quad (1)$$

i = ith consumer unit.

t = time period.

j = jth good (of which there are J).

s = sth store (of which there are S).

Y = income.

p = unit price.

V = vector of other variables affecting demand.

Observations on q and p are market data. That is, they are derived from marketplace transactions representing interactions between buyers and sellers. Traditional data sets for estimating demand parameters typically involve several types of aggregation across the various subscripts. Problems and caveats associated with these procedures are discussed elsewhere (e.g., Buse, Eastwood, and Wahl).

The various types of scan data in Figure 1 can be related to equation (1). This is done in Figure 3. Customer specific records keep the i subscript active. Incorporating V is accomplished via frequent shopper programs. Food shoppers fill out questionnaires to be eligible for cards that give discounts or other incentives. The cards' magnetic strips are passed through card readers at each checkout position at the beginning or end of

customers' transactions. Thus, the i, j, and t subscripts are incorporated into the data. Shoppers can be tracked across the outlets of a chain, or the s subscript can vary somewhat. These is some evidence that suggests this may not be too limiting as only 27 percent of food shoppers compare prices from store to store (Cox and Foster).

Some arrangements have been made by market research companies to track shoppers across chains within a restricted market area. EBT programs operate in a similar fashion but require further coordination across retailers. The ability to track shoppers across chains involves a process similar to the one used by the credit card and banking industries to clear transactions. In addition, all bar codes, including variable weight items, in-store produced foods, and PLUs, would have to be the same or a master file of conversion codes used.

Data associated with the upper path in Figure 2 are characterized by aggregation across consumers. The aggregation also results in the loss of socioeconomic information, including Y. Some merchandising variables can be associated with these data if one has the ability to record the information and match it to the bar codes. These span the entire range of in-store characteristics and newspaper and broadcast advertising. To the extent that several outlets are included in a data base, the socioeconomic characteristics of the various locations can be incorporated.

The use of coupons may be part of the scan data base, which is particularly important for processed foods. All coupons have bar codes that begin with "5". But not all scanner software is capable of identifying and tracking the information. Those systems that do can create either customer specific or store record data. Double and triple couponing programs, especially those that are location specific, may have to be manually added to the data base.

Most promotional information is not automatically part of a scan data base. The situation is a direct result of scanners generating the data and of the organizational structure of most food retailers which evolved

independently of coordinated management information systems. Special efforts must be made to obtain the data and relate them to bar codes. Advertising/marketing departments typically do not record their activity on a bar code basis, although the situation should improve over time. In some instances, vendors have made arrangements in test market areas to obtain bar code specific promotions and/or have created experimental designs to estimate the impacts of alternative marketing strategies.

Economic models emphasize relative prices as a key factor in consumer choice. Scan data provide the requisite observations. This can be particularly important for some types of foods. Figure 4 displays price and item movement (the number of times scanners read a particular bar code) for a specific food -- an 18 ounce jar of a brand of chunky peanut butter. The diagram shows an expected price-quantity relationship and an expected price behavior. Lower prices are associated with higher sales, and price changes are infrequent. Figure 5, displays item movement and price for a cut of beef steak. Again there appears to be a negative own-price relationship. More importantly, during the two periods of price variation, the first is nearly a year and the second is considerably more than a year, significant price variation occurs. Many of these price changes are well over a dollar per pound and change from week to week. Other research (Eastwood, Gray, and Brooker) has found that there is little price correlation between fresh beef aggregates of ground, roasts, and steak. Such results suggest that relative prices change quite a bit from week to week (weekly pricing is used by the chain supplying the data). Perhaps even more important for demand analysis, is their preliminary conclusion that prices for aggregates such as hot dogs and roasts do not change in a proportional manner over time, thereby calling into question the assumption that the composite good theorem applies for aggregates which typically are found in traditional data bases.

Returning to Figure 3, which summarizes the relationships between the equation (1) and the types of scan data, notice the use of an *. It appears in the vendor data sets and denotes possible changes in the measures when the

data are transferred from retailers to market research companies. This does not mean that there are any errors in the data. Instead, it is not clear how various vendors treat some situations as the data are aggregated across retailers. Examples include the following. Chains have different seven day weeks necessitating some sort of designation of a common seven day period. With respect to variable weight foods and price look up codes, different chains capture different information regarding item movement, unit price, or value of the package. The nonreporting of stores creates missing data problems that require adjustment algorithms that may be hard to track down.

Four other features of scan data are particularly noteworthy. First, the level of detail allows for research on close substitutes and complements. Second, the time period also is more consistent with consumers' planning horizons. Most store records are aggregated on a weekly basis. Some are available daily. The records could be aggregated across time to larger periods such as months or quarters. Third, the data can be obtained much more quickly than traditional data sets. Fourth, it is possible to set up experimental designs to test various merchandising hypotheses under marketplace conditions.

Another important point is contained in Figure 6. There is an extended period of time for which no sales of eye of round steak occur. This was most likely due to the cut not being available for sale or a change in bar code. Consequently, food shoppers would have to either switch to another cut or make no purchase. This suggests that scan data may provide some opportunities to look at tradeoffs that are not possible with more aggregated data bases. It also suggests that to the extent food shoppers decide not to purchase, more aggregate data would simply show a reduction in demand.

Scan Data Caveats for Demand Analysis

In many respects the econometric problems of scan data are the same as those associated with traditional sources. Heteroscedasticity, autocorrelation, and multicollinearity can be present in the data. In addition, other problems more specific to scan data arise. A two-part way of

grouping them is by managerial induced problems and by those that are inherent in the type of data involved.

Managerial related problems pertain to data difficulties that arise from management decisions. They are included here to help researchers, who are interested in working with scan data, understand some implicit properties of the data they may obtain. Many of these problems are direct consequences of supermarket management not devoting the resources needed to take advantage of all the information that could be obtained from scan data (McLaughlin and Lesser). Recall that these data are automatically generated by the scanners, but there are opportunity costs associated with allocating the requisite resources to obtain a viable data base. The benefits, while substantial, are primarily long term, and progress in restructuring corporate cultures to take advantage of scan data has been slow (e.g., Shulman).

Scanners, computers, and software limitations may preclude capturing variable weight foods. For example, after a customer’s bill has been generated, the software could place some food items into default categories, or they could be deleted. Although these bar codes conform to UPC standards, the codes are not unique. The first digit of "2" denotes variable weight, and the next five digits denote specific items that must fall within fixed ranges, but the numbering is not necessarily common across outlets. UPCs that begin with a 4 denote foods that have in-store designations, which are unique to the outlet but not standardized across stores. Frequently purchased items may be given special keys (price look-ups) on cash registers to speed up the checkout process, and the software may not convert the PLUs to UPCs. Thus, comparability of some UPCs across retailers is tricky. Obtaining data directly from chains can circumvent these problems, and working with a specific management information department can be of assistance in learning how these foods are coded.

Other managerial related caveats that can affect the type of data include the following. Default categories may contain particular foods of interest. If the software that stores the data places a food item into a

default (e.g., deli chicken salad into the deli department), a chain is very unlikely to alter the software to accommodate research interests. Bar code designations for variable weight and in-store prepared foods may change, necessitating a matching algorithm for a consistent time series. In addition, food processors can change the product (G-K, Figure 1) designations. Another concern is whether the time period for the scan data matches that of the advertising period. That is, the seven days that comprise the scan data week may not match the seven days used by the marketing department in its advertising. This problem is compounded if data are gathered from more than one chain.

The second group of caveats are specific to the data. Errors occur, especially with the store level records, due to the software that is used or to human error. Prices can be incorrect, especially with variable weight items. Often, these errors are corrected for computing customer bills, but the management information system software may not be corrected. These errors can be identified if the chain uses uniform pricing throughout a market area. Zero purchases may be due to the usual decisions not to buy or the product is not available. They may also reflect technical difficulties that preclude transmitting the data from the scanner through the subsequent storage locations. Usually, this situation is characterized by all the data being lost. Chain specific data are easily checked for this problem, but vendor level data may have made adjustments without providing any information.

Some management information system software does not capture the value of the respective package or the weight of the variable weight foods. Instead, the only measure is item movement. If the average size of a package does not change much from week to week, then changes in item movement can serve as a proxy for pounds sold. A related complexity is that some software captures different information for PLUs depending on whether they are fixed or variable weight foods.

The customer specific records are fairly similar to NFCS and CSFII data bases with respect to their tracking individual foods purchased by specific

food shoppers. Unfortunately, since the data are so detailed, price tradeoffs are difficult to capture because prices of foods not purchased are not part of the data base. Obtaining the missing price information may be possible in some situations, but would require considerable programming effort. Store level records, on the other hand, contain the prices of all the foods, but lack the customer specific socioeconomic information. However, to the extent that the respective stores cater to specific socioeconomic groups, the data can be reintroduced into the store level data through careful selection of the outlets. One possibility is to use locations that provide a variety of socioeconomic variation and to include them in (1). Another possibility is to select outlets that cater to specific types of consumers so there is little variation, and an intercept would pick up the common influence.

A final concern also points to the relevance of scan data for applied demand analysis at the store or market level of aggregation. If weekly or daily data are used, variations in customer counts are quite important. Figure 5 presents customer counts at one supermarket over a five and a half year period. Many factors determine patronage, including the competitive environment at each location. Furthermore, research shows that weekly customer counts are quite independent across outlets of the same chain within a metropolitan area (Eastwood, Gray, and Brooker). This suggests that much of the variation in market demand is due to variations in customer counts. Furthermore, if the data are for several stores, then aggregating quantities across stores may have a problem that the number of reporting stores varies due to technical difficulties when transmitting the data from the scanners through the various stages of the management information system. These considerations suggest that a per customer quantity measure for store level data is required.

References

Buse, R. C., D. B. Eastwood, and T. I. Wahl. "Sources of U. S. Food Consumption Data." Japanese and American Agriculture: Tradition and Progress in Conflict. Boulder, CO: Westview Press, 1993.

Cox. C. and R. Foster. "What's Ahead for the U. S. Food Processing Industry?" American Journal of Agricultural Economics. 67(1985):1155-7.

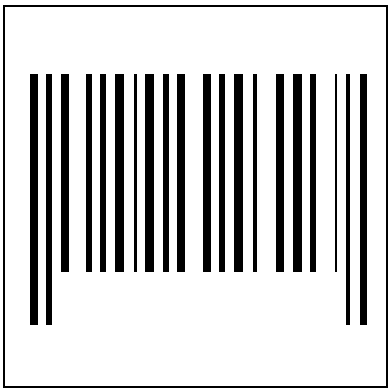
Eastwood, D. B., M. D. Gray, and J. R. Brooker. "An Empirical Test of the Composite Good Theorem Using Scan Data." Enhancing Consumer Choice. R. N. Mayer, Ed. Columbia, MO: American Council on Consumer Interests, 1991:1-8.

_____. "Supermarket Patronage: An Analysis of Customer Counts Among Outlets within a Geographic Area." Dept. Ag. Ec. & Rur. Soc., Univ. of Tennessee, Working Paper (1993).

McLaughlin, E. and W. Lesser. "Experimental Price Variability and Consumer response: Tracking Potato Sales with Scanners." Dept. Ag. Ec., Cornell Univ., Working Paper (1986).

Shulman, R. "Will the ECR Report Evoke Effective Corporate Responses?" Supermarket Business. May, 1993.

Figure 1. UPC Bar Code Structure.



A = system character number, which has the following format:
0, 6, 7 are for regular UPCs,
2 is for variable weight items,
3 is for drugs,
4 is for in-store marking,
5 is for coupons, and
1, 8, 9 are unspecified

BCDEF = specific manufacturers/food processors.

GHIJK = manufacutrer’s products or value of package.

L = check number.

Figure 2. Types of Scan Data.

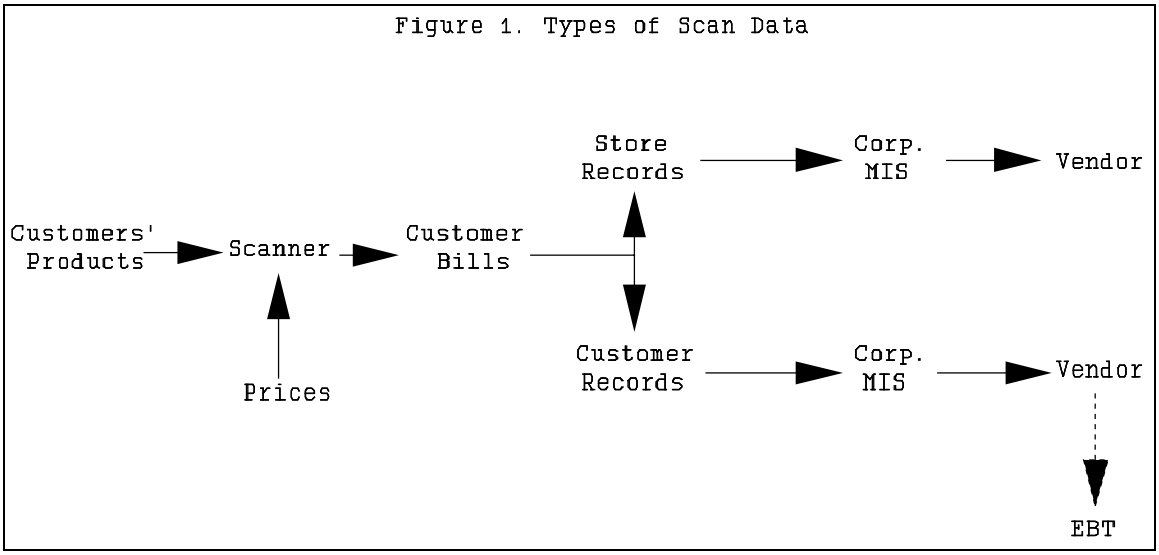


Figure 3. Types of Scan Data and Demand Variables.

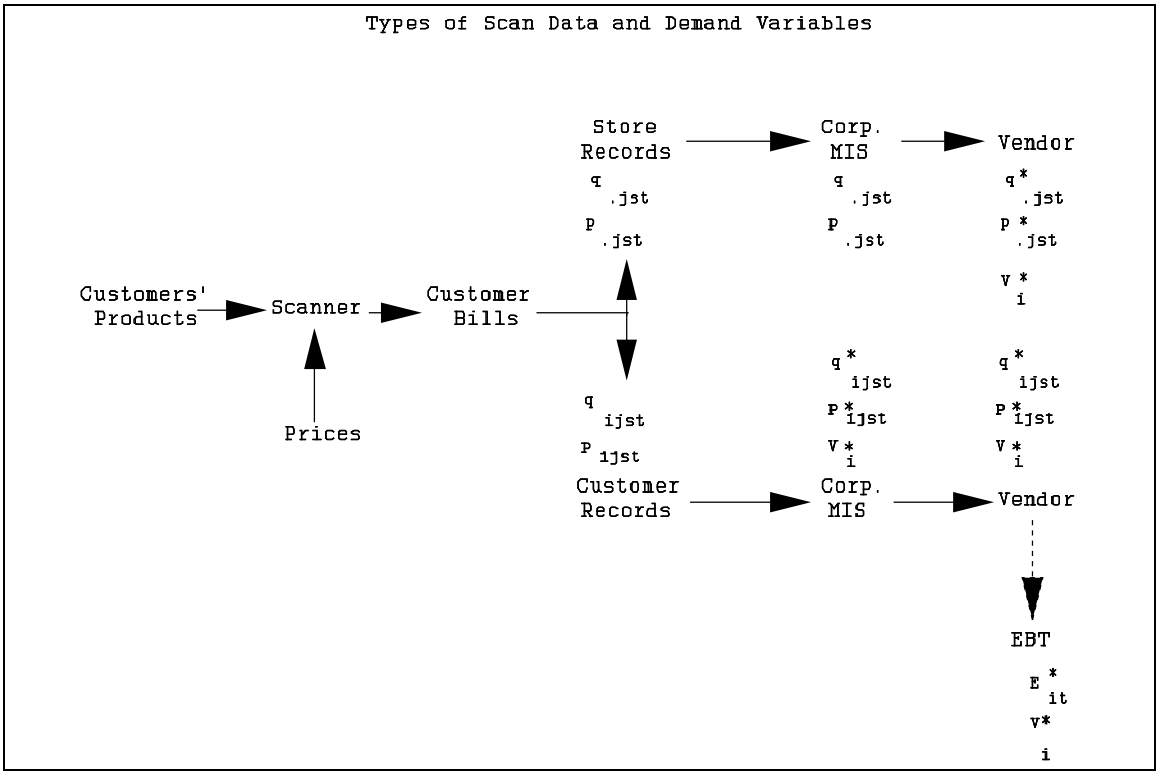


Figure 4. Peanut Butter Brand Item Movement and Price.

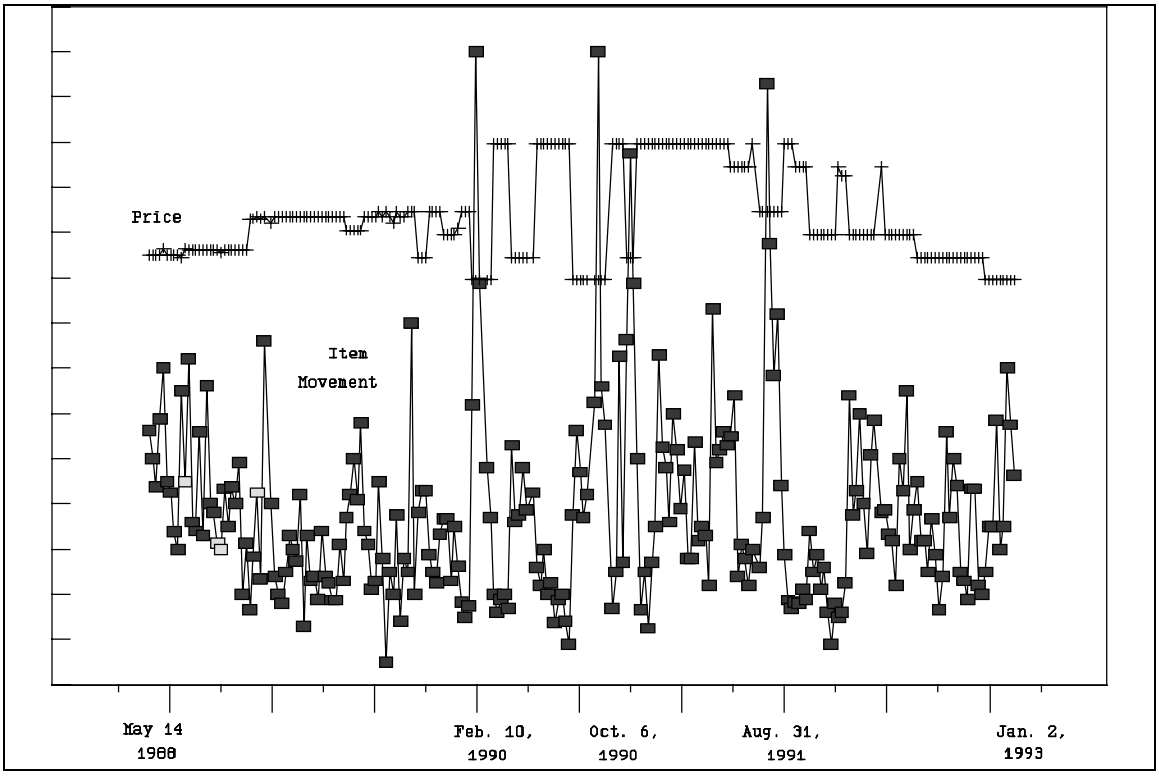


Figure 5. Eye Round Steak Item Movement and Price: Five Store Average.

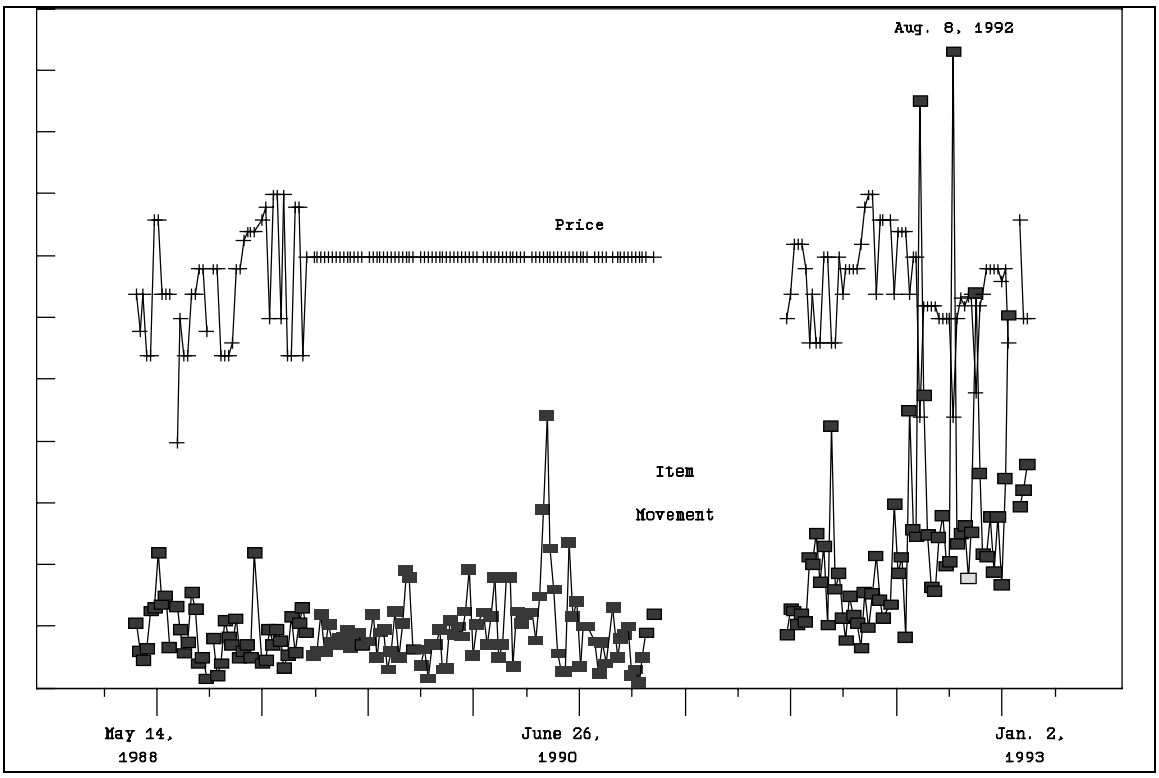


Figure 6. Weekly Customer Counts, One Store.