Data Needs for Consumer and Retail Firm Studies

Jeffrey M. Perloff and Mark Denbaly

May 2007

Jeffrey Perloff is professor, Department of Agricultural and Resource Economics, and member of the Giannini Foundation, University of California at Berkeley. Mark Denbaly is Deputy Director for Data and Web Communication, Economic Research Service, U.S. Department of Agriculture. The opinions expressed in this paper are those of the authors and not necessarily those of the United States Department of Agriculture or any of its members.

Paper to be presented at the American Agricultural Economics Association Meetings, Portland, Oregon, 2007.

Data Needs for Consumer and Retail Firm Studies

Growing concentration in the retail grocery sector raises new economic questions that are difficult to answer with existing data sources. In part because of concentration in the retail data industry as well the fact that these data are not primarily collected for academic research purposes, currently available grocery-level datasets are extremely expensive, not properly randomized, and lack critical information.

To focus our discussion, we talk about data needs for industrial organization and marketing, nutrition and food safety, and government policy studies. The growing concentration at the grocery retail level raises a variety of industrial organization and marketing questions, such as: Did the greater concentration change market power? Did it change the vertical relationship between manufacturers and other suppliers with retailers? Did the entry of low-price superstores fundamentally change the services provided, the degree of product differentiation, the provision of private label products, and other actions by traditional supermarkets? What caused the mergers to occur?

Consequently, we want to know if these various changes in retailing affected the nation’s nutrition and food safety. As firms become more concentrated, are catastrophic food safety disasters more likely? Did the increased product differentiation and lower prices from changes in retailing contribute substantially to alarming increases in rates of overweight and obesity?

Finally, we want to know how government rules and regulations affected these markets and consumers**.** To protect consumers’ health, the government has introduced a number of measures concerning restrictions on selling certain goods when food safety issues arise (e.g., mad cow disease and e. coli in lettuce and spinach). The government also provides nutritional and other information (e.g., concerning health foods and organic foods) to help consumers make better food choice decisions. What were the effects of these laws and regulations on markets and on the health of various groups of consumers?

We discuss the increase in concentration at the retail level, concentration in data provision, data needs for a number of important research areas, and possible solutions.

Concentration in Retail Markets

Grocery retailing markets are much more concentrated today than they were a couple of decades ago. This increase has altered the relationship between manufacturers and retailers. Although most existing empirical studies based on grocery scanner data implicitly presume that manufacturers set prices and retailers passively add on a competitive markup, there is substantial evidence (e.g., Villas-Boas) that such a description of the market is no longer true—if it ever was.

Mergers and acquisitions by large grocery retailers, including Kroger Co., Albertson's, Ahold USA, and Safeway, have significantly increased concentration ratios. Between 1997 and 2000, more than 4,100 U.S. supermarkets were acquired, representing $69 billion in sales. The four-firm concentration ratio increased from 16.6 percent in 1992 to 35.5 percent in 2005: See Figure 1. This trend toward increased concentration has continued with the acquisition of third-ranked Albertson's in 2006 and the growth of Wal-Mart (Kaufman, 2007).

Companies that were not involved in the food business a couple of decades ago, such as Wal-Mart and Target, now account for a significant share of consumers' food-at-home expenditures. Since 1994, nontraditional food retailers (supercenters, warehouse clubs, mass merchandisers, drugstores, and dollar stores) have steadily increased their market share (by about 28 percentage points) to 31.6 percent in 2005. Led by Wal-Mart, most of this growth is attributed to supercenters that command 17.1 percent of the food-at-home retail markets in 2005 (Kaufman, 2007).

It took Wal-Mart just four years of aggressive supercenter growth to become the largest U.S. grocery chain by 2002. Wal-Mart's large share is due to its relatively low prices, which are driven by scale economies and efficient operations based on buying products directly from suppliers. Wal-Mart's approach has started a domino effect, significantly changing the retail food markets landscape. Warehouse club and mass-merchandisers have adopted this strategy, further intensifying the price competition as more consumers have switched from shopping at supermarkets to low-price, large scale operations.

Many supermarkets and other traditional grocery retailers have reacted by expanding their operations through merger and acquisition strategies, introducing a wider variety of new products (e.g., organic and natural foods, upgrade store brands, and convenience foods), promoting new store formats, introducing self-checkout stations, expanding frequent shopper card programs, and offering online home shopping services.

Some researchers contend mergers and acquisitions are driven by a search for efficiencies associated with consolidation as supermarkets are increasingly pressured to meet price competition from non-traditional food retailers, like Wal-Mart. Others contend mergers increase the market power of supermarkets and increase prices to consumers.

Growing retail concentration has not only changed the nature of competition at the retail level, it has greatly affected the vertical relations along the marketing chain. As a result of the competitive pressure pressures from Wal-Mart and other nontraditional formats, many firms in the grocery industry have resorted to, what the industry refers to as, efficient consumer response. The methods are designed to enhance timely, accurate, continuous, consistent flow of products that are matched to consumer demands. The initiative focuses on reengineering activities in four processes throughout the food supply chain: selection of product assortments, product replenishment, product promotions, and new product introductions. Data on the type and the extent of these business practices are not readily available to examine their impact on prices and consumer welfare.

Many believe that the now larger retail vendors are exercising their increased oligopsony power to lower prices paid to suppliers. It is also hypothesized that, in addition to lower prices, powerful retailers have increasingly been charging manufacturers slotting fees, which are lump-sum fees for carrying a new product or continuing to carry an existing one.

Concentration among Grocery Scanner Data Providers

For many years, agricultural economists have studied a variety of demand, health, marketing, and industrial organization questions using proprietary retail grocery scanner data. Today, the only two major firms providing such scanner data are Information Resources Inc. (IRI) and Nielsen (formerly known as ACNielsen). Their datasets are constructed primarily for marketing purposes and are used by retailers, manufacturers, and farm commodity groups. Usually, these firms charge researchers prices comparable to those they charge their commercial customers, so that a dataset covering only a few commodities for the most recent year may cost hundreds of thousands of dollars.

The current major point-of-sale or store scanner data sources are IRI's InfoScan, and Nielsen's ScanTrack. Store scanner data are collected at cash registers, while household scanner data are obtained from a sample of households that scan their purchases after each shopping trip.

Over the past ten years, IRI and Nielsen also have begun to track grocery purchases by specific households. Nielsen's household scanner dataset is Homescan, and IRI's is Consumer Network.[1] These datasets provide rich and detailed information on household demographic characteristics that are not available in store scanner data (Muth, et al., 2007). Demographic characteristics of Homescan and Consumer Network households who remain in the sample for multiple years are updated annually.

Because the household scanner data panelists are instructed (by IRI and Nielsen) to scan all purchases from all outlets, the datasets from household-based scanner data are more complete than grocery datasets of purchases of individual households collected through loyalty card users. These datasets do not include detailed information on household demographics and are potentially subject to more measurement errors due to infrequent use of loyalty cards or use of someone else's card for convenience. Grocery chains rarely make their databases available to researchers.

In addition to being expensive, commercial datasets come with significant restrictions on how they may be used (e.g., market shares of competing brands for specific markets may not be reported) and do not provide all critical information needed for all important research and topics. For example, although feasible, they do not have

[1] Knowledge Networks is also in the process of developing a household-based scanner data panel.

information on whether a specific low-income household is a WIC program participant, they do not have any detail on retailers' cost of operation (e.g., wholesale prices), and the household scanner databases lack prices of non-purchased items for demand studies.

Because scanner data are proprietary and are not primarily designed for academic research, detailed documentation on sampling and data collection procedures, and statistical properties of the data are not readily available. Although few academic papers that use IRI and Nielsen data discuss the quality of these data sets, there is good reason to question whether these firms use proper random sampling techniques. In the store-based scanner data, large, traditional supermarket chains are overrepresented (because they supply data and hence are included with certainty, as opposed to smaller stores that are sampled). In addition, store-based scanner data may not adequately include new sources of food sales (Wal-Mart supercenters and other big box stores, and WIC-only stores).

Muth, et al. (2007) document the data collection process for Nielsen's Homescan data and identify four potential sources of bias: sample design, self-selection, self-reporting, nonresponse, and attrition. However, no formal statistical studies have been conducted to measure the magnitude of the actual presence or magnitude of any potential bias. The households included in the sample are not probability based and randomly drawn from the community, and hence Homescan is a convenience sample.

To get a sense of how these household data sets compare to Census demographics, we compare U.S. Census demographic information to sample averages for IRI InfoScan in 1999 for various zip codes (Table 1). IRI values could differ from Census data either because only a subset of grocery stores is sampled within any given zip code or because the households sampled who shop at those grocery stores are not representative. In our sample, the IRI sample values have relatively large standard errors, so that we cannot conclude that the means of demographic variables in the Census and IRI datasets differ statistically significantly. However in most zip codes areas, the IRI households are larger, more likely to be white, and more likely to have children than are Census households. Moreover, the IRI households are much less likely to have really low or really high incomes.

Data Problems for Research

Purveyors of proprietary scanner data focus on the most recent marketing information for the industry and not on creating datasets that are ideal for research. In the proprietary datasets, short time series and lack of information from other levels of the production chain and other missing variables limit the type of academic studies that are possible.

Industrial Organization and Marketing Studies

These datasets do not include information that would facilitate studies of market power and vertical relations between manufacturers, and retailers (much less suppliers and manufacturers). Critical missing variables include the wholesale price, slotting allowances, and other transfers and restrictions between manufacturers and retailers.

Both to study markups over the food chain and to examine food safety questions, we would like to be able to trace goods from the farm to the consumer. Most industrial organization studies and many nutritional and other studies require one to estimate a system of demand equations. Doing so properly is often difficult with existing data bases for three reasons.

First, the relevant prices are not always available. Often, household datasets only include prices for purchased goods and not other available goods. In a few cases, researchers have matched store-level data with household data (or purchases by other households) to obtain the missing information.[2]

Second, actual transaction price is not always obvious from the reported information. It is not always possible to determine if the price reflect all discounts, coupons, taxes, and so forth. The commercial databases do not record whether the purchases were made using food stamps or WIC vouchers, which preclude studies of such programs and may bias standard demand equation estimates.

Third, the data bases do not report shelf space allocations, local restrictions or store warnings, all relevant advertising (e.g., fliers from the stores), information provided on the products (e.g., fat, health, safety, price per unit, and organic), and other factors that may influence demand.

Because the data bases cover only a nonrandom subset of stores, conducting industrial organization studies of horizontal competition between stores is difficult. In particular, we do not have a complete enough set of stores to conduct spatial studies of pricing. Such spatial information may have other uses as well, if it were available. Research findings in the economics of consumer behavior provide insights into the effects of neighborhood characteristics on consumers' choices in differentiated product markets (Waldfogel, 2003, and Stewart and Davis, 2005).

Similarly, studies of vertical relations are very difficult to undertake and require substantial ingenuity because of a lack of upstream data. Although we have a large amount of retail price and quantity information, we lack information about wholesale

[2] Disturbingly, the price data from the grocery dataset do not always match that from the household dataset, and no means of reconciling these differences are available.

prices, slotting allowances, and other evidence of the interactions between wholesalers or manufacturers and retailers.

Nutritional and Food Safety Studies

The high societal costs associated with obesity and overweight have intensified the need to identify and understand the factors that influence food choices and the effects of these choices on an individual's health. Extensive studies on consumer food demand show that food choices may depend on food prices, as well as on consumers' income levels, time available to shop and prepare meals, and human capital resources, such as education and type of employment. Economic studies of these issues are hampered by a lack of data. Matching datasets with nutritional information for processed foods are not readily available.

No single reliable data source currently provides or could provide all of the information required in such an endeavor. A number of data sources provide some of the information, but each is weak in critical areas. A 2005 report by the National Research Council of the National Academies (NRC) made recommendations that enhance usability of various critical data systems to support research on critical U.S. food and nutrition policies. Following through with NRC's recommendation to create integrated and consistent data will help researchers to better understand how consumers' food choices, diets, and health are affected by such factors as changes in food prices, neighborhood characteristics and access to food stores and restaurants, and participation in government food assistance programs.

The National Center for Health Statistics of Center for Disease Control measures food intakes and an array of health outcomes for a representative population, but no information on prices of foods eaten by survey respondents is collected. Adding price information from other exiting sources would enable research on drivers of consumer food choice and their connections to health outcomes for various population subgroups and regions over time. Measuring consumer price responsiveness is a critical component of a sound policy strategy. Beyond characterizing consumer preferences, information on price responsiveness enables researchers to evaluate the effects of taxes and subsidies on consumption of various foods and nutrients they contain.

Currently, no dataset can trace the foods back to their sources. Plans of Wal-Mart and others to use radio signals (RFID tags) to track goods from the manufacturer to the retailer or final consumer raise privacy issues, but they also may provide a means to examine important questions concerning food safety, food quality, and various vertical issues. However, we know of no plans to make such information available to researchers. Indeed, for proprietary reasons, manufacturing and retailing firms may not want information about the extent, cost, and efficiency of these devices be disseminated.

Nutritional studies are hampered because of a lack of datasets that cover both food at home and food in restaurants. As Americans have increasingly switched from home-cooked meals to processed or restaurant meals, the substitution patterns between these types of meals has substantial public policy importance.

Government Programs

Apparently because their scanner datasets are voluminous, sample observations older than three years are usually discarded by IRI and Nielsen. Thus, many time series or historical studies of government laws and regulations are difficult or impossible to conduct. For example, data from these sources before and after the key changes in U.S.

rules on organic foods are generally not available either because datasets are short or because older data are discarded (cf. Kiesel and Villas-Boas, 2007).

Food assistance programs are designed to provide a nutritional safety net, guaranteeing a minimum level of access to essential nutrients for participants. Empirical evidence on the extent to which the programs affect consumption, nutrient intake, and overweight and obesity provides critical information about the current effectiveness of the programs. Combining the existing measures of consumption patterns and health status of program participants with this information on program records on benefit levels and duration of participation will help to make the critical link between food assistance programs and diet, nutrition, and health outcomes of program participants.

This link will be particularly valuable in strengthening examinations of how, and if, diet and health are influenced by benefit levels, duration of program participation, and cumulative level of program benefits. For example, recording how long participants have been in the sample can help researchers determine if the sizes of the program’s effects differ depending on duration of participation.

Just as important, this link will improve data accuracy. The National Health and Nutrition Examination Survey (NHANES) queries respondents about their program participation and benefits. However, studies show that self-reported information is systematically underreported in many surveys, including NHANES. For example, in 2004, the Current Population Survey captured 60 percent of average monthly caseload and 58 percent of annual benefits (Bollinger and David, 2005). Administrative records can be used to correct this underreporting and avoid analytical results that would otherwise be biased.

Supplementing the NHANES dataset with this information would allow one to study the connection between food choices and neighborhood characteristics, particularly for low-income households in urban and rural areas. To the extent that NHANES includes such households, researchers could correlate health and nutrition outcomes with household and location characteristics. A link between NHANES data and information on the location of food stores and eating places would also enhance efforts to understand better the effects of access on food choices and health outcomes. Information on locations and characteristics of food stores and foodservice establishments can be collected using proprietary sources, such as Spectra® and NPD. While locations of and access to food stores and restaurants could influence consumption, other community and social factors may also affect food choices and health outcomes. The “neighborhood effect” refers to the interdependence between individual decisions and the decisions and characteristics of others within a common neighborhood. Linking NHANES to household and local community descriptors in the Census’s American Community Survey will help researchers understand how neighborhood characteristics influence food choices and health outcomes.

Improving Scanner Data

We have a simple and obvious message. With more data, economists could analyze additional, important issues of economic theory and government policies.

Because data lack rivalry, society underprovides data. Relying on commercial vendors is unattractive because these firms charge very high prices, do not fully disclose the nature of their data, provide data for only very short periods, and report only variables that are important for commercial customers and not all variables that are important for researchers.

One approach to ameliorating data shortages for research would be to have government agencies or nonprofit organizations collect the ideal datasets or provide incentives to commercial providers. Fundamentally, researchers need access to unrestricted data based on proper random samples and that include all the relevant variables.

First, to enable unfettered assess, to improve content, and to obtain better prices, it may make sense for university and government researchers and organizations (the AAEA, government agencies, business school organizations, the American Economic Association, and others) to try to negotiate with private purveyors collectively. They might also negotiate to house at no or little cost historical data that are discarded so that longer time series and additional variables can be created. However, such collective action might raise antitrust issues.

Second, these research groups could try to make arrangements with individual firms to supply data. We know of at least two supermarket chains that have been willing to make such agreements in the past. The AAEA could lead the efforts to select representative samples of suppliers to collect details of proprietary transaction data and provide them to researchers so that privacy and confidentiality of the data are maintained.

Third, these research organizations could collaborate to collect data on their own. Even discussing this possibility may facilitate negotiations with commercial data purveyors.

On a less grand scale, we have a laundry list of new datasets that would be particularly useful. First, industrial organization and food safety studies require information at both the retail and upstream levels, including information about wholesale prices, food sources, and various slotting and tying relations.

Second, nutritional studies need datasets that combine information on food-at-home and away-from-home as well as the nutritional content of these various foods. Because consumer studies find substantial variation in nutritional consumption across demographic groups and neighborhoods, datasets are needed that cover a broad cross-section.

Third, health and nutrition studies would benefit substantially if we could link the intake and health data of with administrative food assistance records to add levels and duration of program assistance. Such a link would have to address two challenging issues: (1) privacy and confidentiality conditions under which states collect the administrative data must be met to access the data for linking purposes, and (2) data formats vary across states make linking these sets to survey data difficult. In addition, given the relatively small effects of price and income on food choices, addressing the obesity epidemic may require collection of new data on consumers' health and nutritional knowledge, attitude, and available time to shop and prepare meals to undertake economic studies to understand consumer dietary behavior.

References

Bollinger, C., and M. David. 2005. "I Didn't Tell, and I Won't Tell: Dynamic Response Error in the SIPP," *Journal of Applied Econometrics*, 20: 563-569.

Kaufman, P. R. 2000. "Consolidation in Food Retailing: Prospects for Consumers & Grocery Suppliers," *Agricultural Outlook*, Economic Research Service.

Kaufman, P. 2007. "Food Market Structures: Food Retailing." www.ers.usda.gov/Briefing/FoodMarketStructures/foodretailing.htm.

Kiesel, K., and S. B. Villas-Boas. 2007. "Got Organic Milk? Consumer Valuations of Milk Labels after the Implementation of the USDA Organic Seal." *Journal of Agricultural & Food Industrial Organization*, 5 www.bepress.com/jafio/vol5/iss1/art4.

Muth, M. K., P. H. Siegel, C. Zhen. 2007. "ERS Data Quality Study Design." Final Report, Research Triangle Institute, Project Number 210153.001.

Stewart, H., and D. Davis. 2005. "Price Dispersion and Accessibility: A Case Study of Fast Food." *Southern Economic Journal*, 71:784-799

National Research Council of National Academies. 2005. "Improving Data to Analyze Food and Nutrition Policies." Committee on National Statistics, Panel on Enhancing the Data Infrastructure in Support of Food and Nutrition Programs, Research, and Decision Making.

Villas-Boas, S. B. 2007. "Vertical Relationships Between Manufacturers and Retailers: Inference With Limited Data." *Review of Economic Studies*, 74:625-652.

Waldfogel, J. 2003. "Preference Externalities: An Empirical Study of Who Benefits Whom in Differentiated Product Markets." *Rand Journal of Economics*.

Figure 1. Top Four (C4), Eight (C8), and Twenty (C20) Firms' Share of U.S. Grocery Store Sales

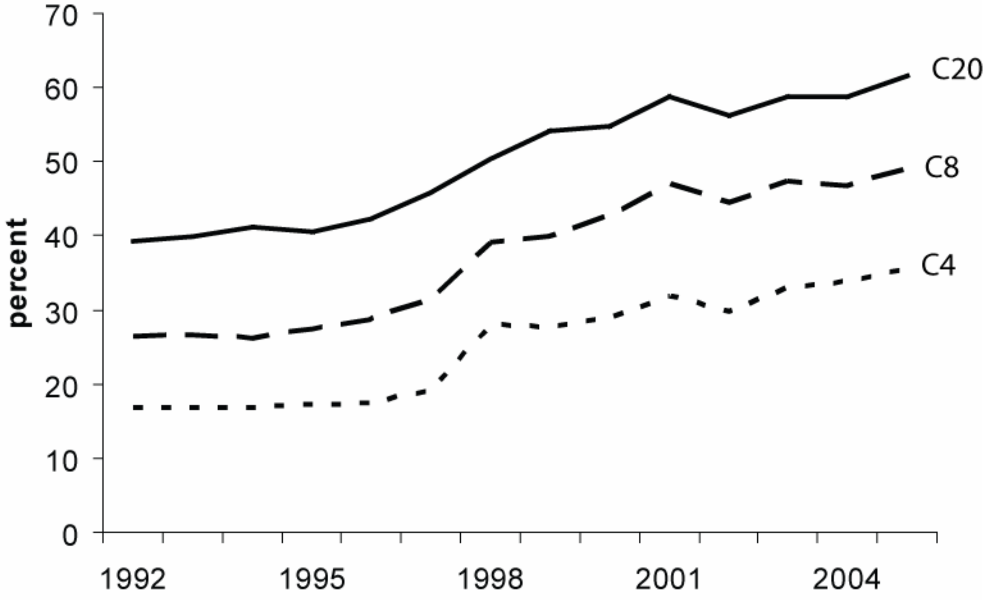

Table 1. Comparison of U.S. Census and IRI Demographic Data

HH with Individuals Younger than 18 (%)		Share of HH with Incomes (%)						Share White (%)		Average HH Size	
		< $10,000		$25,000-$34,999		≥ $100,000					
Census	*IRI*	*Census*	*IRI*	*Census*	*IRI*	*Census*	*IRI*	*Census*	*IRI*	*Census*	*IRI*
29.3	45.6	5.0	1.6	11.1	12.8	33.6	6.7	95.7	97.9	2.39	3.08
31.5	32.5	5.6	2.5	13.3	19.5	18.5	0.1	93.6	85.6	2.50	2.94
44.3	44.3	4.1	1.1	7.3	8.4	44.9	6.0	80.0	93.2	2.74	3.09
26.8	26.8	8.0	2.7	16.3	13.4	17.9	3.1	72.0	82.5	2.50	2.96
34.0	26.1	11.3	6.7	16.3	21.6	8.5	0.3	88.9	95.4	2.41	2.58
36.5	38.3	3.7	3.3	11.8	17.4	25.7	1.5	97.2	97.3	2.51	2.83
31.8	34.2	2.1	3.1	10.1	16.1	33.4	2.4	92.8	96.2	2.34	2.73
31.5	32.8	6.4	5.3	13.6	14.3	26.8	1.5	89.4	93.3	2.40	2.69
30.2	30.2	7.2	4.4	13.9	17.0	12.9	1.8	93.2	96.7	2.34	2.66
33.8	34.1	5.6	4.7	14.7	17.1	18.9	1.7	94.4	96.6	2.42	2.77
27.8	39.8	6.0	2.9	12.5	16.5	29.9	2.8	89.9	96.8	2.31	2.94
37.6	42.7	3.4	2.4	14.1	16.2	22.1	0.7	85.9	95.9	2.58	3.04
35.8	46.9	4.4	1.7	14.5	17.1	21.8	0.5	43.7	94.1	2.42	3.11
30.2	35.5	3.6	3.3	15.1	19.4	19.3	1.7	90.5	96.2	2.32	2.81
34.6	41.1	3.2	2.3	13.0	18.9	24.1	1.6	93.6	96.6	2.50	2.93
44.6	37.2	2.0	2.6	9.6	16.5	35.4	1.3	93.3	96.1	2.74	2.90
43.8	30.5	11.1	6.4	15.2	17.3	17.4	2.0	39.9	54.6	2.75	2.67
33.4	39.6	9.0	5.9	13.2	15.8	33.3	4.0	65.9	62.4	2.39	3.03
45.3	35.8	14.9	10.6	16.2	15.2	11.3	1.3	59.6	68.7	2.86	2.87
46.1	41.0	7.0	10.0	17.5	17.2	11.2	0.9	78.5	67.2	2.72	2.98
37.8	38.5	8.6	5.4	12.5	12.1	27.2	2.7	77.2	76.3	2.56	2.95
39.6	38.4	3.7	8.6	8.1	14.3	40.8	2.1	87.6	72.3	2.49	2.91

Notes: Each row represents a zip code region. IRI data are for 1999 and Census data are from 2000.