



Find the full report at www.ers.usda.gov/publications/eib-economic-information-bulletin/tb-1942

Understanding IRI Household-Based and Store-Based Scanner Data

Mary K. Muth, Megan Sweitzer, Derick Brown, Kristen Capogrossi, Shawn Karns, David Levin, Abigail Okrent, Peter Siegel, and Chen Zhen

What Is the Issue?

USDA's Economic Research Service (ERS) purchases proprietary household and retail scanner data that are an integral resource for many policy-relevant research projects. ERS obtained data for 2008-12 from IRI, a market research company, on household food purchases (called Consumer Network) and retail food sales (called InfoScan). While ERS has purchased and evaluated similar household data from other vendors, differences in how the data are processed by vendors could have implications for research programs at ERS. Additionally, ERS purchased comprehensive store-level scanner data and product dictionaries, including nutrition and health claims data, and little is known about the attributes of these data. To help users better understand the limitations of these data for food policy research, and in accordance with Office of Management and Budget specifications, this report documents the characteristics and examines the statistical properties of these datasets. This is the first in a series of ERS reports examining the statistical properties of the IRI datasets.

What Did the Study Find?

The IRI household and retail scanner data and associated files can be an extensive, impactful resource, but researchers should understand the complexity and different properties of these datasets. The Consumer Network household scanner data are derived from over 120,000 households who report what food products they purchased, when they shopped, and where they shopped. These households also report demographic information, and a subset of households report health and prescription drug information. The household purchase data can be linked to product characteristics (e.g., brand) and nutrition data, which gives a robust picture of the type of products households are purchasing. Researchers, however, should be aware of how well the household panel reflects the demographic makeup of the U.S. population and how the methods used to construct prices and demographic variables may affect analyses. In particular:

- The data include survey weights, which can be used to produce estimates for the total U.S. population. However, total U.S. expenditures reported by households in the weighted IRI data are less than those in other nationally representative datasets.
- Certain households are less likely to report purchases consistently, including households with heads under age 35, households in the lowest income bracket, and households with children. Hence, researchers should use caution in interpreting findings based on the data for certain population subgroups.

ERS is a primary source of economic research and analysis from the U.S. Department of Agriculture, providing timely information on economic and policy issues related to agriculture, food, the environment, and rural America.

- The household demographics file for 2008-12 is a snapshot of household demographic characteristics as of 2012, and changes in demographic characteristics over the time period cannot be determined.
- For the majority of products, IRI assigns prices using InfoScan data collected from stores, so many prices may not represent the exact value a household paid for an item. In addition, researchers should subtract the value of coupons from prices paid by households to calculate net amounts paid.
- Quantities purchased are not available for random-weight items (i.e., products purchased by the pound or unit, rather than by the package), which limits the usefulness of the data in food economics research involving fresh fruits and vegetables, meats and cheeses, and bakery items.

The InfoScan retail scanner data cover a large portion of retail food sales in the United States and contain billions of transactions by outlet type (i.e., grocery, convenience, dollar, drug, liquor, mass merchandiser, and club stores) and market area. Like the Consumer Network data, the InfoScan data can also be linked to nutrition and product characteristics data, enabling researchers to examine sales of products with particular characteristics geographically as well as by outlet type. However, these data also have limitations researchers should consider when using them to conduct food economics analysis. In particular:

- The IRI data obtained by ERS are a subset of the total data in InfoScan due to restrictions from IRI and retailers on what data may be released. As a result, the retail store set is a subset of IRI's fully projected market tracking service, and survey weights are unavailable to produce nationally representative estimates.
- Some retailers release data for each individual store, while others release data for retailer-defined retail marketing areas. Because these geography-based aggregations vary by retailer, it can be difficult to examine geographic variation or conduct analyses by geographic area for certain retailers.
- Some retailers limit the release of data on private-label products to broad categories rather than individual Universal Product Codes (UPC), limiting the scope of analysis for research on private labels.

Lastly, the product dictionaries and nutrition and health claims data provide information about the items households are purchasing and retailers are selling. The product dictionaries give detailed descriptions of the products, including flavor, brand, style, and type for items with UPCs. The nutrition and claims data contain information on the Nutrition Facts panel and front-of-package health claims. However, researchers should note the following when using these datasets:

- Limited product information is available for random-weight perishable products, such as bulk or loose produce; uniform-weight perishable products, such as bagged produce; and private-label products from certain retailers.
- Only 41 percent of the UPC products in the retail store set have any nutrition and/or claims data; however, these products make up about 81 percent of total sales in the InfoScan data.
- IRI provides substantially better nutrition information coverage for private-label products than other commercial nutrition datasets; however, IRI's nutrition coverage is less for private-label UPC products than for branded UPC products.

How Was the Study Conducted?

Researchers from ERS and RTI International examined the contents of the IRI datasets, initial documentation provided by IRI, and documentation prepared by ERS. Detailed discussions were conducted with IRI, including discussions on a set of questions developed under the study, and additional documentation was obtained from IRI on specific questions. Researchers documented their findings, prepared summaries of the data, and compared certain components of the datasets with Government or commercial data sources.