United States Department of Agriculture

A report summary from the Economic Research Service October 2018

Examining Food Store Scanner Data: A Comparison of the IRI InfoScan Data with Other Data Sets, 2008–2012

David Levin, Danton Noriega, Chris Dicken, Abigail M. Okrent, Matt Harding, and Michael Lovenheim

What Is the Issue?

USDA's Economic Research Service (ERS) has purchased proprietary household scanner data for more than a decade, and started acquiring proprietary retail scanner data (InfoScan) from the market research firm IRI in 2008. Previous statistical evaluations of the household data have examined their usefulness in food policy analysis, but retail scanner data are less studied. This report explores the representativeness of the InfoScan data with regard to store counts and food sales, as well as its strengths and limitations in food policy analysis.

What Did the Study Find?

While the number of stores and sales revenue reported in InfoScan are generally lower than other datasets nationally, both measures of InfoScan coverage vary substantially by year and category (i.e., grocery, liquor, drug), and also across geographic areas. These differences are likely driven by the subset of store information released by InfoScan to ERS, which: (1) only includes stores that agree to release information to ERS for statistical purposes and does not include weights that can be used to project sales revenue to the national level, (2) only includes grocery stores with more than $2 million in annual sales revenue, and (3) excludes sales revenue for nonfood products.

The other datasets used in the comparison include the Economic Census, County Business Patterns (CBP), TDLinx, and the National Establishment Time Series (NETS). The following are some of the results of the comparison between InfoScan and the other datasets. For the combined category of grocery/convenience/dollar/club/mass merchandise/defense commissary, the Economic Census reported 402,159 stores, Nielsen's TDLinx 229,797 stores, the CBP 400,952 stores, and NETS 269,698 stores in this category in 2012, whereas InfoScan captured 59,374 stores in this category, corresponding to roughly 15 percent of the stores in the Economic Census, 26 percent of those in TDLinx, 15 percent of those in the CBP, and 22 percent of those in NETS.

National sales revenue data in InfoScan are better aligned with sales revenue reported in the other datasets than the store count data, which may reflect the fact that InfoScan picks up larger stores.

- Sales revenue reported in InfoScan, which covers food products only, represents nearly 50 percent of sales revenue in the most comparable subset of the Economic Census, food sales at payroll establishments.
- InfoScan has lower store counts compared to the other data sets for all counties in the United States, but the degree of undercounting of stores in InfoScan varies across counties.
- InfoScan coverage of sales revenue differs across geographic areas. In regional case studies of the Texas and Eastern areas, InfoScan coverage relative to other data sets in both areas was lower than the national average, but higher in the Eastern area than in Texas.

The limited coverage of the InfoScan data relative to the TDLinx, CBP, Economic Census, and NETS data (with respect to the number of establishments and sales revenue) means that for analysis of these metrics at the aggregate/national level, these other datasets may present a more representative picture. The geographic variability of InfoScan's coverage at the subnational level may also make such subnational analyses problematic, and the unavailability of weights for InfoScan may complicate attempts to conduct demand analysis.

InfoScan remains a valuable data source for analysis of topics requiring Universal Product Code (UPC)-level transaction data for food purchases, with the caveat that results are more relevant to larger stores. The combination of UPC-level transaction data with the ability to attribute sales to specific store locations and retailer chains opens additional avenues of research, though researchers should be mindful of the representativeness issues discussed in this report.

How Was the Study Conducted?

Researchers from ERS, Duke University, and the University of California-Irvine compared the store counts and sales revenue from a subset of IRI's InfoScan (clients who agree to release information) to those of several other national data sets, including U.S. Census Bureau's Economic Census and CBP, Walls & Associates National Establishment Time Series (NETS) database, and Nielsen's TDLinx. The Economic Census is a publicly available dataset that covers almost all industries and provides information at the county level; it is considered the "gold standard" for measuring overall economic performance of business in the United States, but is only conducted every 5 years. CBP is an annual series that provides subnational economic data by industry between each Economic Census. TDLinx and the NETS are proprietary datasets maintained by Nielsen and Walls and Associates, respectively, which contain more detailed information for each establishment.

The years 2008 through 2012 were examined for all datasets except the Economic Census, for which data only exist in 2012. Before the comparisons could be made, it was necessary to identify the same stores across all of the datasets, several of which used different schemes to classify store types. This was accomplished by constructing a relational matrix to bridge the various classification systems. Misclassification of store types across datasets, primarily affecting the convenience store and grocery store types, prevented comparisons of those individual store types. As a result, those two categories were combined with the dollar, club, mass merchandiser, and defense commissary store types into one larger category to allow meaningful comparisons across datasets.