United States Department of Agriculture

Economic Research Service

Technical Bulletin Number 1949

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

United States Department of Agriculture

Economic Research Service
www.ers.usda.gov

Recommended citation format for this publication:

Levin, David, Danton Noriega, Chris Dicken, Abigail M. Okrent, Matt Harding, and Michael Lovenheim. *Examining Food Store Scanner Data: A Comparison of the IRI InfoScan Data With Other Data Sets, 2008–2012,* TB-1949, U.S. Department of Agriculture, Economic Research Service, October 2018.

United States Department of Agriculture

Economic Research Service

Technical Bulletin Number 1949

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

Abstract

USDA's Economic Research Service (ERS) has purchased proprietary retail scanner data (InfoScan) since 2008 to examine food policy questions. To determine how representative the data are, this report compares the number of stores and sales revenue reported in the InfoScan data with the same information from other datasets. The InfoScan data purchased by ERS are limited to a subset of stores that agree to release their data and cover only food sales, while many of the other sources cover total sales. In addition, InfoScan includes only grocery stores having annual sales greater or equal to $2 million, while some of the other sources are not so limited. The subset of InfoScan stores in the ERS dataset results in a lower store count relative to other datasets, and coverage varies geographically. However, the sales reported in the ERS subset of InfoScan better align with sales reported in other datasets than do estimates of store counts, since InfoScan encompasses larger retail stores. The report discusses implications for using InfoScan for food economics research.

Keywords: IRI, InfoScan, scanner data, food expenditures, County Business Patterns, Economic Census, National Establishment and Time Series (NETS), TDLinx, food at home (FAH), food sales.

Acknowledgments

The authors thank the following peer reviewers: Christian Gregory (ERS), Chen Zhen, (University of Georgia), Kristen Giombi (RTI International), Gary D. Thompson (University of Arizona), and Katherine Ralston (ERS). We also thank Eliana Zeballos (ERS), Megan Sweitzer (ERS), Tim Park (ERS), and Constance Newman (ERS) for their comments, and Dale Simms and Curtia Taylor (ERS) for editorial and design services.

Contents

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.

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

Introduction

Over the past decade, ERS has acquired proprietary household scanner data—the retail purchases reported by a panel of over 120,000 households—to create data products and conduct research on food policy issues. For example, ERS researchers and collaborators have used household scanner data to investigate food policy questions such as demand for sugar-sweetened beverages (Smith et al., 2010; Zhen et al., 2014), fruits and vegetables (Dong and Lin, 2009), dairy products (Davis et al., 2010), and whole-grain products (Mancino et al., 2008); the healthfulness of food purchases (Volpe and Okrent, 2012); organic price premiums (Carlson and Jaenicke 2016); local foods (Low et al., 2015); and food-based tax policies (Harding and Lovenheim, 2017).

To conduct more in-depth analyses on the food environment, ERS purchased retail scanner data called InfoScan from IRI for food and alcohol purchases starting in 2008.[1] InfoScan contains sales revenue, quantity, and store location data for products sold by participating stores.

ERS, as a principal statistical agency of the Federal Government, provides objective and credible economic statistics to the public. ERS and collaborators have analyzed the sample design, collection methods, and statistical properties of the household scanner data (Muth et al., 2007; Zhen et al., 2009; Einav et al., 2008; Sweitzer et al., 2017). However, little is known about the statistical properties of retail scanner data like InfoScan, as none of the aforementioned publications examined retail scanner data. A recent ERS study describes the methodology and characteristics of the retail scanner data, and compares it at the national level with one other data set (Muth et al., 2016).

Further analysis comparing retail scanner data to other datasets would more definitively assess the representativeness of the InfoScan data both nationally and regionally. This is important because researchers may encounter geographic variability when examining issues of food access, market structure, and nutrition and diet quality. Understanding the representativeness of InfoScan data will help researchers assess its suitability for their particular research questions and help ensure that empirical results based on the data are interpreted appropriately.

Strengths of InfoScan for Food Policy Research

The InfoScan retail scanner data are strengthening ERS's efforts to conduct food policy research relating to the U.S. food environment. For example, these data are being used to (1) examine the effect of USDA's Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) on the retail price of infant formula, (2) evaluate the Food Safety Modernization Act, (3) estimate the relationship between food access and price-cost markups, (4) determine the effects of retailer concentration in the dairy industry, (5) examine geographic price variation for Thrifty Food

[1] Prior to 2008, ERS only acquired retail point-of-sale scanner data for certain food products, such as infant formula, nonalcoholic beverages, and specific vegetables.

Plans, and (6) report fruit and vegetable prices to coincide with each new version of the *Dietary Guidelines for Americans*. These particular areas of inquiry are central to ERS' mission, which is to anticipate trends and emerging issues in agriculture, food, the environment, and rural America and to conduct high-quality, objective economic research to inform and enhance public and private decision-making.

The availability of UPC-level transaction data in InfoScan allows analysis of specific food products or food categories; other available datasets report sales revenues only at the most aggregated levels. The ability to attribute sales to specific store locations and/or specific retailer chains in InfoScan enables researchers to conduct analysis on narrower geographic areas than would be possible using other datasets. Also, the geographic specificity of UPC-level transaction data enables researchers to explore other avenues of research, such as how the entry of a new food retailer affects the broader retail food market (Martinez and Levin, 2017).

The large number of observations and reported product characteristics in retail scanner data have been cited by the U.S. Bureau of Labor Statistics and other statistical agencies for increasing the accuracy of price index formulations (U.S. BLS, 1996; Ivancic et al., 2011). The granularity of the data at the product and spatial level also allows for the construction of spatial price indices, which have implications for the value of participant benefits in food assistance programs (Cakir et al., 2018). Finally, the availability of nutritional information and health/nutrition-related claims on packaging (for the UPCs in the InfoScan dataset) makes it possible to incorporate dietary, nutritional, and public health issues into research projects. Martinez and Levin (2017), for example, compared the nutrient content of newly introduced and discontinued products, and analyzed the results in the context of *Dietary Guidelines* recommendations.

Objectives and Approach

This report examines the geographical coverage of the InfoScan data in terms of store counts, as well as the sales revenue of the store locations. These results are compared to those of other datasets, including the U.S. Census Bureau's 2012 Economic Census (EC) and County Business Patterns (CBP), Walls & Associates National Establishment Time-Series Database (NETS), and Nielsen's TDLinx.

The Economic Census is the U.S. Government's official 5-year measure of U.S. business and the economy, and is the comprehensive source of economic information about American businesses at the national and local level. All U.S. stores with payroll are required by law to report sales, business activity (i.e., industry), employment, payroll, and location information to the U.S. Census Bureau. The Economic Census only collects information for a sample of non-payroll stores (e.g., stores that do not employ workers) and, in conjunction with IRS administrative data, imputes national and State-level estimates for non-payroll stores (Census Bureau, "Economic Census FAQ"). The EC is conducted every 5 years (those ending in 2 and 7), and data are released 3-4 years after the reference year. While the EC includes both store count and sales revenue information, the quinquennial frequency means that only 1 year of this survey (2012) falls within the time span of this analysis (2008-12). The Economic Census data report information at the industry (e.g., North American Industry Classification System, NAICS) and State levels; further disaggregation of the data is not available in the public-use data.

The limited availability of EC data is addressed in part by another U.S. Census Bureau dataset, the County Business Patterns (CBP). CBP is conducted annually and contains economic data by

industry at the county level, including the number of stores, employment during the week of March 12, first quarter payroll, and annual payroll (but no sales revenue data). These data can be used to examine the economic activity of regional, State, and local areas; observe economic changes over time; and as a benchmark for other statistical series, surveys, and databases between economic census years (Census Bureau, "County Business Patterns: About This Program").

In addition to these two Government datasets, two proprietary datasets are also included in this comparative analysis. TDLinx is a dataset maintained by Nielsen that consists of food retailers and additional information on retailer size, location, and store characteristics. The NETS is a database of store-level information by Walls and Associates that contains detailed characteristics about each store. Both TDLinx and NETS include store counts and sales revenue data, though TDLinx uses discrete variables encompassing ranges of sales revenue while NETS uses a single continuous variable. Both data sets contain annual total sales for individual stores; TDLinx is released 1 month after the reference month, while NETS is released 12 months from the reference month, but neither distinguish between food and nonfood sales. Unlike the Economic Census and CBP, TDLinx and NETS can be used to conduct analysis at the store level.

Data and Methods

To evaluate InfoScan data coverage, this analysis compares the number of stores and sales revenue in the InfoScan data with Nielsen TDLinx, the Economic Census (EC), the CBP, and the Walls and Associates NETS (table 1). Our analysis covers 2008-12, but years included in each comparison with InfoScan vary by dataset. This section describes differences across data sources in terms of establishment classification, time period of analysis, and construction of the store count/sales revenue estimates, as well as potential limitations of each dataset.

Table 1
Summary of data sources selected for comparison

	InfoScan (ERS subset)	Economic Census (EC)	County Business Patterns (CBP)	TDLinx	National Establishment Time Series (NETS)
Data	Store counts, sales of food and alcohol by UPC, UPC descriptors, industry/ retail format, other store characteristics	Store counts, total sales by aggregate product, industry/ retail format, employment, payroll	Store counts, industry/ retail format, employment, payroll	Store counts, total sales, industry/retail format,other store characteristics	Store counts, total sales, industry/ retail format, other store characteristics
Target population	Retail establishments	Business establishments	Business establishments	Business establishments	Business establishments
Geographic coverage	National, RMA, store-level[1]	National, State, county, metropolitan area[2]	National, State, county, metropolitan area, ZIP Code, Congressional district[2]	National, State, county, store-level	National, State, county, store-level
Industry/ retail classification system	Proprietary[3]	North American Industry Classification System (NAICS)	NAICS	Proprietary[3]	NAICS
Sample size (store counts, all channels, 2012)	59,538	402,159	400,952	229,797	269,698
Dates data cover	2008-12	2012	2008-12	2008-12	2008-09
Frequency	Weekly (sales revenue), annual (store counts); released to ERS 4 months after each reference year	Annual information reported every 5 years (years ending in 2 and 7), released 3 years after each reference year	Annual, reported 16 months after each reference year	Monthly, reported annually 1 month after the reference month	Annual, reported 12 months after each reference year

— continued

Table 1
Summary of data sources selected for comparison— continued

	InfoScan (ERS subset)	Economic Census (EC)	County Business Patterns (CBP)	TDLinx	National Establishment Time Series (NETS)
Source	IRI	U.S. Census Bureau	U.S. Census Bureau	Nielsen	Walls & Associates
Year survey started	2008[4]	1810[5]	1964[6]	1995	1990

[1] RMA = retail marketing areas; RMAs are defined differently by each retailer. Some retailers provide IRI with data for each store location, while others provide only aggregate data at the RMA level. UPC = Universal Product Code.

[2] To protect respondent confidentiality, some data is censored at smaller geographic areas.

[3] InfoScan and TDLinx each use their own unique proprietary classification systems. Grocery stores in InfoScan are limited to those with annual sales revenues of at least $2 million, while grocery stores in TDLinx include those with sales revenue of at least $1 million. Retail channels other than grocery stores are not limited by sales revenue.

[4] 2008 is the earliest year for which ERS has access to InfoScan data.

[5] The scope of the Economic Census has changed over time.

[6] Similar data have been reported for various periods since 1946.

Source: USDA, Economic Research Service.

IRI InfoScan

The time span of InfoScan evaluated here is 2008 through 2012, though 1 or more years of InfoScan are omitted when a comparison dataset lacks corresponding data for that year.[2] IRI maintains agreements with retail establishments, which provide weekly sales and quantity data for products with UPCs and random-weight (or perishable). ERS receives the data 4 months after the reference year. For example, the 2016 InfoScan data were released to ERS at the end of April 2017. IRI classifies the stores (or channels) in InfoScan as grocery, drug, convenience, mass merchandiser, club, dollar, and defense commissary stores. Though beyond the scope of this analysis, product characteristics and nutritional information are available for the UPCs in InfoScan.[3]

There are multiple limitations to the InfoScan data acquired by ERS. First, IRI's definition of the grocery store channel includes only stores that have $2 million or more in annual sales, so grocery store counts will be underreported in InfoScan. The 2012 EC reported that 98,385 grocery (NAICS 4451) and specialty food (NAICS 4452) stores operated the entire year; of those stores, 57,135 (57 percent) generated revenue less than or equal to $2.5 million. Also, the InfoScan data acquired by ERS include only food products, and many stores also sell nonfood products. For example, the 2012 EC reported that total sales for grocery (NAICS 4451) and specialty food (NAICS 4452) establishments (with payroll) were $578 billion; of these sales, $420 billion (73 percent) were for food. For other types of stores like warehouse clubs and mass merchandisers, food's percentage of sales is even smaller.

Second, InfoScan includes two components, which IRI calls the "census" and "sample" components. "Census" stores are members of retail chains that provide IRI with sales data for all of their store locations. The remaining stores are randomly selected from the remaining universe of retail

[2] ERS has purchased InfoScan data for 2013-17, which are excluded from this analysis.

[3] See Muth et al. (2016) for details regarding what product characteristic/nutrition information is available.

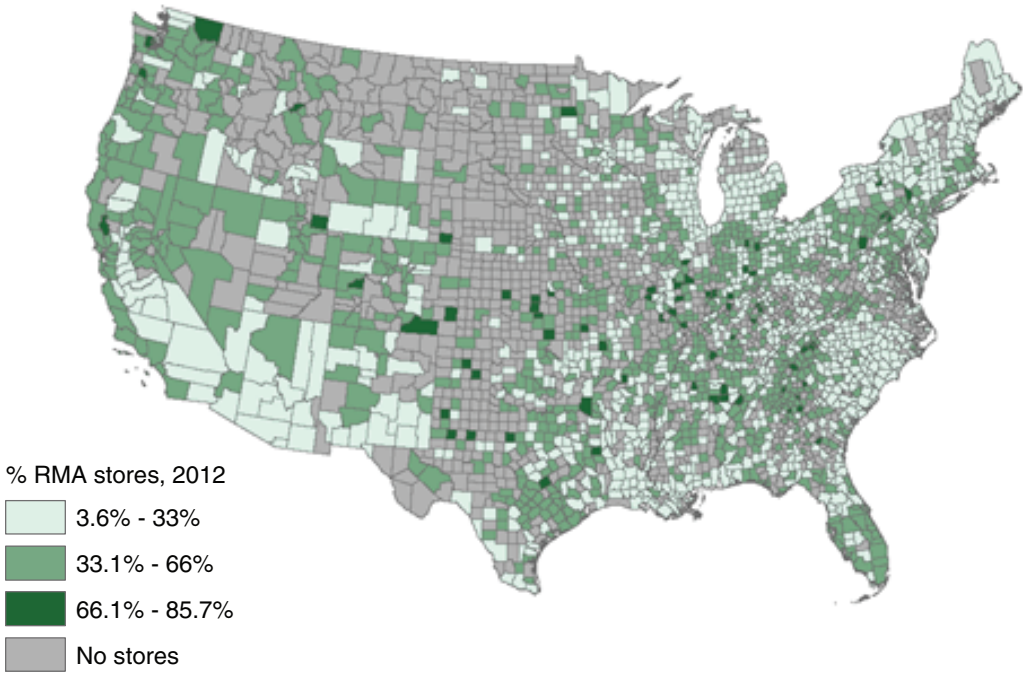
locations by IRI, which contracts with each store's parent chain to receive data from the selected stores. Because IRI does not sell its sample component, the InfoScan data ERS has acquired include only the "census" component.

Finally, some of the InfoScan data ERS receives are at the store level, while other data are at the retail marketing area (RMA) level; the latter is the case when a retailer has not approved the release of store-level data to ERS. RMA retailers still provide weekly UPC-level sales and quantity data, but the data are aggregated across all stores within the RMA, preventing the attribution of sales and quantities of a UPC to any specific store(s) within the RMA. Each retailer defines its own RMAs geographically, meaning that the differences in RMA definitions from one retailer to another make it extremely difficult to conduct regional analyses in which RMAs are compared across retailers.[4]

Figure 1 shows the percentage of stores by county in the InfoScan that are RMA retailers. The grey areas indicate counties where InfoScan has no coverage. Only a few counties in the InfoScan have more than two-thirds of stores as RMA retailers. Figure 2 shows the percentage of stores by county in the InfoScan that are non-RMA retailers.

Figure 1

Percentage of InfoScan stores that are restricted to retail marketing areas (RMA), 2012



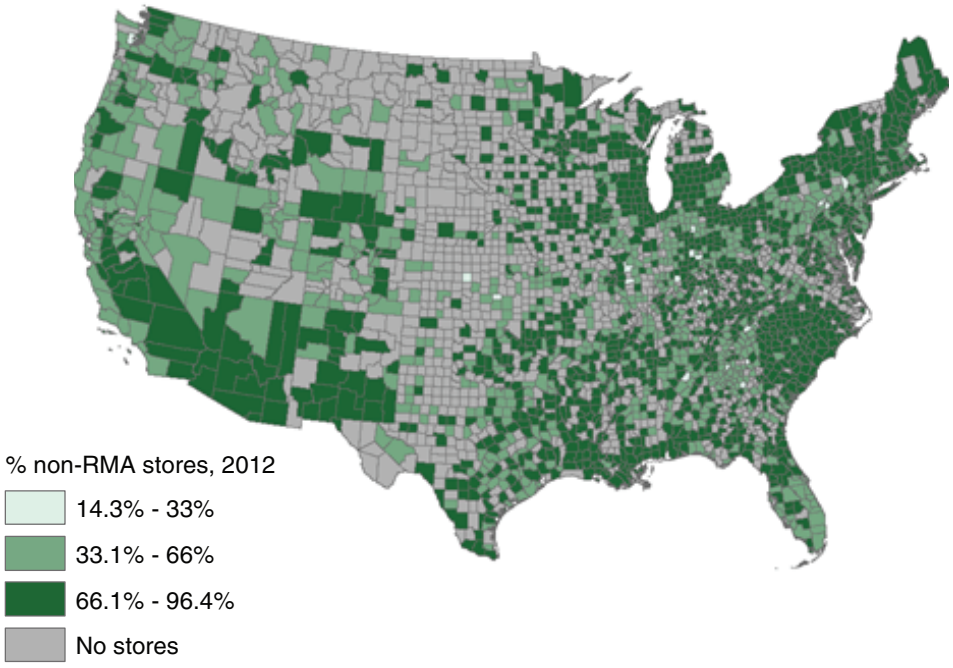
Note: Retail marketing areas (RMAs) are geographic areas defined by each retailer in which their UPC-level data are reported at the aggregate level rather than at the individual store level.

Source: USDA, Economic Research Service calculations, IRI InfoScan.

[4] See Muth et al (2016) for a specific example of differences in RMA definitions across retailers.

Figure 2

Percentage of InfoScan stores that are unrestricted (non-retail marketing area), 2012



Note: Retail marketing areas (RMAs) are geographic areas defined by each retailer in which their UPC-level data are reported at the aggregate level rather than at the individual store level.

Source: USDA, Economic Research Service calculations, IRI InfoScan.

U.S. Census Bureau Economic Census

The EC is released every 5 years, coinciding with our InfoScan data in 2012. The EC is the dataset that most closely approximates the universe of economic activity in the United States. The EC is organized along the hierarchical structure of the NAICS, allowing one to move up to broader NAICS codes or down to narrower ones. It was not possible to calculate sales revenue at the 5- or 6-digit NAICS codes because those data are censored "…to avoid disclosing data for individual companies" when there are very few businesses at a particular NAICS code in a given county. This censoring happens frequently enough that it is necessary to aggregate the sales revenue data to the 4-digit NAICS code. Although there is virtually no censoring in the EC data for the number of establishments at the 5- and 6-digit NAICS codes, for consistency with the sales revenue estimates we also calculate the number of unique stores for each 4-digit NAICS at the county level.

We also compare InfoScan data to national-level EC estimates of *total* and *food and beverage only* sales revenue in 2012 for payroll-only establishments. As noted above, ERS only purchases data on food products (including beverages) in InfoScan, representing 73 percent of total sales at grocery store establishments reported in the 2012 EC. Also, the grocery store channel in InfoScan includes only stores with more than $2 million in sales, while the grocery store channel in TDLinx includes only stores with more than $1 million in sales; these limitations mean that InfoScan and TDLinx grocery stores are more likely to be larger grocery stores with payroll. For example, in the 2012 EC, all non-payroll grocery stores had sales revenue less than $2.5 million. Hence, it may be better to measure the representativeness of the InfoScan sales revenue data (purchased by ERS) by comparing

it with food and beverage sales for payroll establishments in the EC at the national level. Because of differences in the discrete sales revenue variables between the EC and TDLinx—and also the differences in coverage limitations for the grocery store category between InfoScan, TDLinx (see TDLinx section below), and the EC—comparing sales revenue from InfoScan to that of the EC for food and beverage sales for payroll establishments having annual food and beverage sales revenue of $2.5 million or more is the most realistic comparison possible.

U.S. Census Bureau County Business Patterns (CBP)

The CBP is an annual series that provides data on the number of establishments, employment, and payroll by State and ZIP Code at the 2- through 6-digit NAICS levels. Hence, we use the estimates of the number of establishments (EST) between 2008 and 2012 from this dataset for the comparison with InfoScan. The CBP was mostly used as a data quality check when comparing InfoScan, NETS, and TDLinx to the EC. The EC limits the years for comparison because it only is collected every 5 years, and the CBP data are selected to line up with the dates that overlap with the EC.

Nielsen TDLinx

The period of study for the comparison between TDLinx and InfoScan is 2008-2012. TDLinx data provide a comprehensive listing of all grocery, club, convenience store, and small-format food selling locations in the United States. TDLinx collects total sales and geographic location data (e.g., FIPS (Federal Information Processing Standards), city, State, census tract), updating its listings monthly (Nielsen, 2010), though ERS receives annual TDLinx updates from Nielsen. On average, TDLinx includes 202,415 unique stores during the study period. The data are organized into two broad categories (“channels”)—grocery and convenience—and 10 narrowly defined subcategories (“sub-channels”)—limited assortment grocery, natural/gourmet foods, warehouse, superette, supermarket, supercenter, military commissary, gas station/kiosk, conventional convenience, and military convenience.[5]

TDLinx reports annual store sales revenue categorically, which allows annual sales revenue to be reported for every store without censorship. The interval values mask the precise performance of a specific store in any given year. The interval values provided in the TDLinx were converted to minimum and maximum values (table 2), which allowed the construction of county-level sales revenue bandwidths. We top-coded the largest sales revenue interval to be $1 billion. With no specific details, a high, but feasible, upper bound keeps the bandwidth from shifting too significantly upon the inclusion of this top-coded interval.

One limitation of the TDLinx data is that they only include grocery stores with more than $1 million in sales (Nielsen, 2010). According to the 2012 EC, about 44 percent of grocery (NAICS 4451) and specialty food (NAICS 4452) establishments have less than $1 million in sales, meaning TDLinx, like InfoScan, misses a sizeable number of smaller grocery stores. Gordon-Larsen et al. (2015) compared food outlet data from TDLinx and Dun and Bradstreet’s Duns Market Identifiers File to a field-based census of food stores and restaurants (often referred to as “ground-truthing”) in 31 census tracts in Durham, North Carolina. They found that 111 (64 percent) and 95 (55 percent) of the food stores identified in their census of stores through ground-truthing methods were listed in

[5] The TDLinx data were organized into 2 categories and 10 subcategories for this analysis; however, the original TDLinx data contain additional categories/subcategories not mentioned here.

TDLinx and Dun and Bradstreet, respectively. TDLinx identified 76 percent of convenience stores and 63 percent of grocery and supermarkets.

Walls and Associates National Establishment Times Series (NETS)

The shared period of study between the NETS and InfoScan data is 2008-12. Walls and Associates creates the NETS by appending establishment information like estimated annual sales, years active, and type of legal establishment (i.e., proprietorship, partnership, corporation) to the Dun and Bradstreet data that contain over 52.4 million establishments between 1990 and 2012 and their addresses (Walls and Associates, 2012). The NETS data are updated yearly based on a January "snapshot" of the Dun and Bradstreet database. Each establishment is uniquely identified by a DUNS number, and these unique IDs are site-specific. For example, companies or chains with multiple physical store locations would each have their own DUNS number. The DUNS numbers are never reassigned, and stores are never removed from NETS. If establishments close, they are not removed from the data but are coded as no longer being in business.[6]

Each DUNS number is assigned to a six-digit NAICS code. We included DUNS numbers with NAICS codes pertaining to grocery, convenience, and other food-related stores, and ultimately aggregated the data to the 4- rather than 6-digit NAICS code (the mapping section discusses in more detail the NAICS code included in this study). For example, 4-digit NAICS code 4451 (grocery stores) contains 6-digit NAICS codes 445110 (supermarkets and other grocery stores) and 445120 (convenience stores). Excluding DUNS for establishments that have closed, we calculated the number of unique stores (or DUNS numbers) for 4-digit NAICS codes per county, per industry, and by year. NETS provides a single point value estimate for annual sales revenue. The source of the point values varies by store, and flags are provided to distinguish the source, which can be either 0 = Actual, 1 = Bottom of Range, 2 = D&B Estimate, or 3 = Walls Estimate. A point value for sales revenue is always provided for any store currently in business, but the source may vary. In this report, each point estimate is taken at face value when aggregating to estimate sales revenue by county and by industry.

Evidence suggests that Dun and Bradstreet data, which is the starting point for the NETS, has—like TDLinx—limited coverage. Unlike TDLinx and InfoScan, the Dun and Bradstreet data do not specify a monetary cutoff for food establishments to be included. Gordon-Larson et al. (2015) found that Dun and Bradstreet identified 55 percent of all food outlets, 53 percent of convenience stores, and 65 percent of grocery and supermarket stores in their study of North Carolina. However, Liese et al. (2010) conducted a field census in one urban county (Richland) and seven rural counties (Calhoun, Chester, Clarendon, Fairfield, Kershaw, Lancaster, and Orangeburg) in the Midlands region of South Carolina to verify the presence and location of each food outlet listed in the Dun and Bradstreet database and to identify new, unlisted outlets. Sixty-three percent of the ground-truthed retail food stores were identified by Dun and Bradstreet, with 56 percent of convenience and 76 percent of supermarket and grocery stores included in the Dun and Bradstreet data. Results were

[6] The NETS data used in this study included an unusually small number of store locations for the drug store category – approximately 250 drug stores – while Cho et al. (2018) found between 30,000 and 40,000 drug stores in the NETS data used in their report. Attempts to reconcile the different number of drug stores across the two studies were unsuccessful; it is likely that the version of NETS used in this study differs from that used by Cho et al. As a result, this report excludes store count and sales revenue information for the drug store category in NETS from all comparisons. The number of store locations and sales revenue for the other categories in NETS were unaffected.

similar from Powell et al. (2011) in the Chicago area[7] and Fleischhacker et al. (2012), who used ground-truthed census data from seven American Indian communities in North Carolina.[8]

Concordance Across Data Sets

Complications in harmonizing data arise because establishments are classified into industry groups differently across datasets. Recall that InfoScan and TDLinx classify establishments according to an "in-house" coding system (known as "channels" in both datasets) while the NETS, CBP, and EC use the NAICS. To compare store counts and sales across datasets that do not share a unique identifier, we constructed a relation matrix.

A unique identifier for each establishment shared between datasets would be ideal for the mapping (e.g., TDLinx store code). Failing that (as is the case for the EC and CBP), a relation matrix between classification schemes must be constructed.

A relation matrix linking the InfoScan and TDLinx classification schemes to NAICS was constructed by first manually assigning channels (in both the InfoScan and TDLinx) to NAICS codes. For example, the TDLinx channel 07 (convenience store) was mapped to NAICS code 445120 (convenience store). Likewise, TDLinx channel 05 (grocery) was mapped to NAICS code 445110 (super markets and other grocery stores, except convenience). These assignments are simply due to the similar labels of the two classifiers. If all mapping proceeded in this way, according to a subjective best-fit criteria of the labels where the researcher uses her/his own best judgment, mapping the datasets based on NAICS classification (i.e., NETS, CBP, and EC) to TDLinx or InfoScan would be straightforward. However, this intuitive mapping violates the assumptions that establishments are accurately and consistently classified across datasets. In other words, an establishment considered a "grocery store" in one dataset will also be considered a "grocery store" in another, which often is not the case.

Such discrepancies between how stores are classified between datasets are largely driven by how grocery and convenience stores are classified. To address this misclassification problem, we combine a few related categories and then compare the combined categories, which produces much more consistent results.

To understand the likelihood of misclassification, we construct a matching algorithm that identifies establishments existing in both TDLinx and NETS. Searching was constrained by ZIP code, and establishments were matched using business name, street address, and latitude/longitude coordinates.

Across these three matching criteria, business name proved the most reliable and consistent criterion. Matching by street address achieved the highest hit rate but was less reliable. For example, shopping centers with different stores can have the same address, leading to a mismatch. Matching on latitude/longitude coordinates has similar issues to matching by street address but with the

[7] Powell et al. (2011) found that the Dun and Bradstreet data provided fair to poor coverage at lower levels of disaggregation in the classification.

[8] Fleishhacker et al. (2012) found the coverage of food retail outlets to be quite low at 41 percent, ranging from 32 percent for convenience stores to 81 percent for grocery stores.

advantage of being fast, being less prone to text errors, and allowing pairwise searches based on the spherical law of cosines.[9]

To illustrate, table 2 shows how the same store can be classified differently in NETS and TDLinx. Each panel shows the results of ZIP code-constrained matches by business name, taking the establishment names in one dataset and then searching the other for a match (e.g., taking establishment names in NETS of a particular ZIP code then searching for a match in TDLinx within the same ZIP code).

Table 2
Comparing classifications of grocery stores and convenience stores across NETS and TDLinx

NAICS (NETS)		Channel (TDLinx)	Number of matched stores
Establishment name-ZIP Code in NETS matched with TDLinx			
445120	Convenience stores	07 - convenience store	13,120
445110	Supermarkets and other grocery stores (excluding convenience stores)	05 - grocery	2,889
445110	*Supermarkets and other grocery stores (excluding convenience stores)*	*07 - convenience store*	*2,725*
447190	Gas stations without convenience stores	07 - convenience store	1,652
447110	Gas stations with convenience stores	07 - convenience store	336
452910	Superstores and warehouse clubs	01 - wholesale club	206
445120	*Convenience stores*	*05 - grocery*	122
452990	All other general merchandise	07 - convenience store	102
Establishment name-ZIP Code in TDLinx matched with NETS			
445120	Convenience stores	07 - convenience store	19,950
445110	Supermarkets and other grocery stores (excluding convenience stores)	05 - grocery	8,728
445110	*Supermarkets and other grocery stores (excluding convenience stores)*	*07 - convenience store*	4,374
447190	Gas stations without convenience stores	07 - convenience store	1,850
447110	Gas stations with convenience stores	07 - convenience store	566
445120	*Convenience stores*	*05 - grocery*	*219*
722110	*Full-service restaurants*	*07 - convenience store*	*213*
452910	Superstores and warehouse clubs	01 - wholesale club	208

Notes: Rows in italics indicate instances of misclassification. Establishments were matched where possible by a combination of name and ZIP Code. Totals shown include only establishments that were successfully matched in both datasets.

Source: USDA, Economic Research Service calculations, IRI InfoScan, Nielsen TDLinx, U.S. Census Bureau EC and CBP, Walls & Associates' National Establishment Time-Series (NETS).

What is considered a grocery store in NETS (NAICS code 445110) appears just as likely to be classified as a convenience store in TDLinx as it is to be classified a grocery store. In particular, 2,889 establishments that are classified as grocery stores (channel 05 and NAICS code 445110)

[9] The spherical law of cosines (or the great circle formula) is a method for calculating distance between two points on a spherical surface. This is the formula needed to calculate the straight-line distance between any two points on the earth's surface.

and 2,725 establishments that are classified as NAICS 445110 (supermarkets and other grocery stores (excluding convenience stores)) in NETS are classified as channel 07 (convenience store) in TDLinx. Additionally, establishments coded as *convenience stores* (channel 07) in TDLinx are often labeled as supermarkets and other grocery stores (excluding convenience stores) (NAICS code 445120) or as *gasoline stations with a convenience store* (NAICS code 447110) in NETS.

This misclassification implies that a direct comparison between classification groups across datasets would lead to inconsistent store counts, which we find to be the case. Table 4 provides an example of such misclassification for several FIPS codes in Alabama. Higher grocery store counts are followed by lower convenience store counts across datasets—and vice versa. For example, for FIPS 1003, NETS identified 75 establishments as channel 05 (grocery stores), whereas the number of establishments identified as NAICS 445110 (supermarkets and other grocery stores, excluding convenience stores) in the EC and CBP and as channel 05 (grocery stores) in TDLinx was substantially lower (22-42). However, the number of establishments classified as channel 07 (convenience stores) in NETS (81) is much lower than the number of establishments in EC and CBP classified as NAICS 445120 (convenience stores, 114-122) and in TDLinx classified as channel 07 (139).

Table 3
Number of establishments classified as "grocery store" or "convenience store" across datasets in 2007 for selected FIPS codes in Alabama

FIPS code	Classification[1]	NETS	Economic Census (EC)	County Business Patterns (CBP)	TDLinx
1001	Grocery	18		5	7
1001	Convenience store	16	25	33	35
1003	Grocery	75	25	42	22
1003	Convenience store	81	114	122	139
1005	Grocery	32		7	9
1005	Convenience store	19	19	22	31
1007	Grocery	10		7	5
1007	Convenience store	11		19	19

[1] Grocery is 6-digit NAICS code 445110—supermarkets and other grocery stores (excluding convenience stores)—for NETS, EC, and CBP and is channel 05—convenience store—for TDLinx. "Convenience stores" is 6-digit NAICS code 445120—convenience stores—for NETS, EC, and CBP and is channel 07—convenience stores—for TDLinx.

FIPS = Federal Information Processing Standard; NAICS = North American Industry Classification System.

Source: USDA, Economic Research Service calculations, Nielsen TDLinx, U.S. Census Bureau Economic Census, and County Business Patterns, Walls & Associates National Establishment Time Series (NETS).

The misclassification issue was mitigated by taking the union of groups based on higher level NAICS codes. For example, the 4-digit NAICS code 4451 (grocery) is the root classifier for 445110 (supercenters and other grocery stores, except convenience stores) and 445120 (convenience stores). Therefore, any TDLinx or InfoScan subchannel that maps to the 6-digit NAICS codes but is rooted in the *4-digit* codes must be combined to map to higher level codes. Doing so reduces specificity but leads to estimates with less misclassification.

Combining TDLinx and InfoScan groups to match 4-digit NAICS codes leads to more comparable counts (table 4). Furthermore, using 4-digit NAICS codes as the comparison level provides more

consistent counts for the EC values (the EC will not publish data at the 6-digit level if there are too few firms). Appendix tables A1-A3 show the full relation matrices between all the datasets.

Table 4
Number of establishments classified as "grocery and convenience stores" across datasets in 2007 for selected FIPS codes in Alabama

FIPS code	Classification[1]	NETS	Economic Census (EC)	County Business Patterns (CBP)	TDLinx
1001	Grocery and convenience	34	29	38	42
1003	Grocery and convenience	156	166	168	161
1005	Grocery and convenience	51	20	29	40
1007	Grocery and convenience	21	19	26	24

[1] Grocery and convenience is 4-digit NAICS code 4451—grocery stores—for NETS, EC, and County Business Patterns and channels 05 and 07—grocery stores and convenience stores—for TDLinx.
FIPS = Federal Information Processing Standard; NAICS = North American Industry Classification System.
Source: USDA, Economic Research Service calculations, Nielsen TDLinx, U.S. Census Bureau Economic Census and County Business Patterns, Walls & Associates National Establishment Time Series (NETS).

InfoScan Coverage Assessment at the National Level

We measure representativeness of the InfoScan data by comparing the number of stores and sales revenue across datasets for three store categories: (1) drug; (2) grocery, convenience, dollar, club, mass merchandiser, and commissary; and (3) liquor. While the second category represents the vast majority of sales revenue for food products, drug stores and liquor stores are also of interest with respect to food sales. Drug stores—particularly the larger retail drug store chains—participate in the Supplemental Nutrition Assistance Program (SNAP), making the inclusion of drug stores necessary when conducting analyses of the types of food products sold by SNAP retailers. Liquor stores are relevant because the consumption of alcohol is a public health issue, and excluding liquor stores from analyses of alcohol sales may result in an incomplete picture of the alcohol market.

The number of stores in InfoScan is compared to other datasets at the county and national levels; sales revenue is compared to the other data sets at the national level only because of the aggregated (RMA) data. To compare the sales revenue in InfoScan with other datasets at a more local level, we also construct estimates of sales revenue for two case studies that are inclusive of all RMA and non-RMA data, and compare these with the other data sets.

Number of Stores

Table 6 shows the number of stores (or "retail channels") by year in the IRI InfoScan data from 2008 to 2012.[10] The table presents store counts among seven retail channels: convenience, defense commissary, dollar, drug, grocery, liquor, and mass merchandiser/club. Counts are calculated separately by store-level and RMA-level reported data. Stores that do not provide data on UPC or random-weight purchases are excluded.

Table 5
Total number of stores in InfoScan by retail channel, 2008-12

		Number of stores by retail channel							
Year	Dataset	Convenience	Defense commissary	Dollar	Drug	Grocery	Liquor	Mass merchandiser/ club	Total
2008	Store-level	6,372	259	7,364	11,998	7,478	251	3,001	36,723
	RMA-level	0	10	0	7,341	5,743	487	0	13,581
2009	Store-level	8,529	255	7,392	12,276	7,463	269	3,058	39,242
	RMA-level	0	10	0	7,341	5,743	464	4,520	18,078
2010	Store-level	9,416	254	7,538	12,375	7,382	290	3,075	40,330
	RMA-level	0	10	0	7,358	5,743	464	4,520	18,095
2011	Store-level	9,579	514	7,808	12,414	7,165	318	3,109	40,907
	RMA-level	0	10	0	7,358	5,743	464	4,520	18,095
2012	Store-level	9,613	515	8,237	12,497	7,100	341	3,140	41,443
	RMA-level	0	10	0	7,358	5,743	464	4,520	18,095

RMA = Retail marketing area.
Source: Muth et al. (2016)

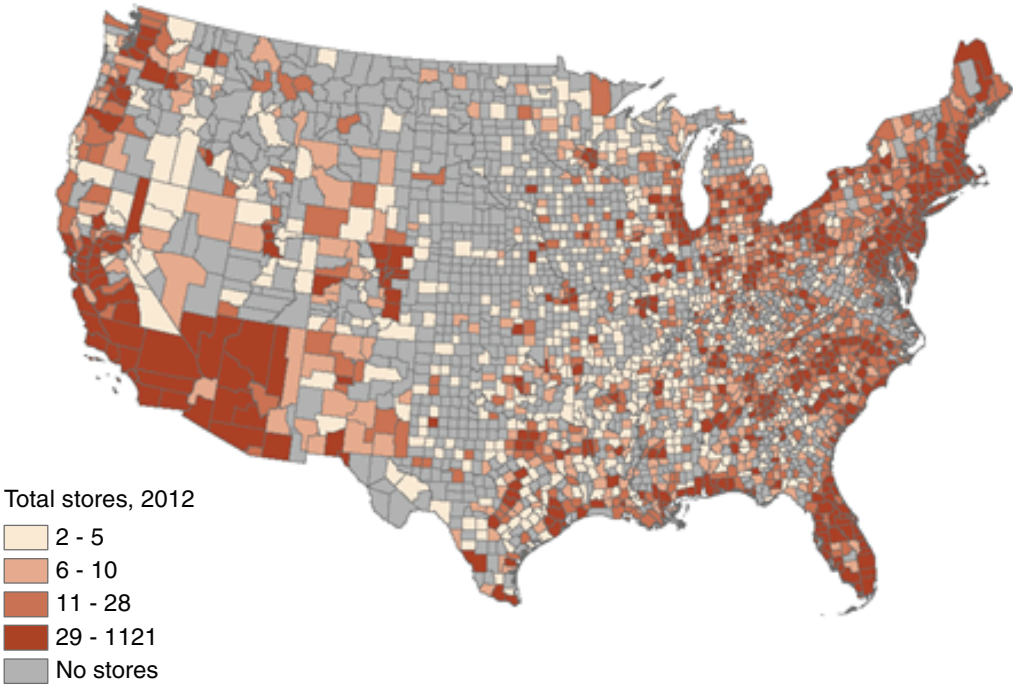
[10] This table is reproduced from table 5a in Muth et al. (2016).

Store counts remain largely stable across years but do grow slightly, from 50,304 in 2008 to 59,538 in 2012. Most of this growth occurs between 2008 and 2009 with the addition of over 4,500 RMA-level mass merchandise/club stores, which reflects the addition of Walmart to InfoScan (Muth et al., 2016). Year-to-year variation within each retail channel and level is relatively small. From 2009 to 2012, about 45 percent of the stores are RMA level, with the remainder being store level. Mass merchandise/club, liquor, grocery, and drug stores have a high share of RMA-level stores, while convenience, defense commissary, and dollar stores are almost wholly store level (table 5). The most prevalent store types in the data are convenience, dollar, drug, and grocery, and the composition of store types varies little across years.

Figure 3 shows the geographic distribution of IRI stores by county in 2012. The figure includes all stores, regardless of whether they report sales by store or at the RMA level. Stores are more prevalent on the coasts than in the middle of the country, although there are also many stores in parts of the Midwest and Southeast.

Figure 3

Total stores in InfoScan by county, 2012



Note: Store counts include those reporting at both the store and RMA (retail marketing areas) level and thus are total store counts.
Source: USDA, Economic Research Service calculations, IRI InfoScan.

The IRI InfoScan data consistently undercount stores relative to the other datasets (table 6).[11] IRI InfoScan stores are a subset of all stores in an area; the InfoScan data ERS purchases include only retailers who have agreements with IRI, and IRI does not provide projection factors that would

[11]As discussed previously, we aggregate most of the retail channels into one group to avoid problems associated with classification differences across datasets.

allow store counts in the data to be representative nationally.[12,13] In addition, InfoScan includes only grocery store retailers with over $2 million in annual sales, excluding many grocery stores without payroll. Thus, it is most informative to compare the differences in the extent of undercounting in InfoScan relative to the other datasets. For the largest category, grocery/convenience/dollar/club/mass merchandise/defense commissary stores, InfoScan includes between 14 and 19 percent of the stores in the other datasets, depending on the source and year. InfoScan captures about 22 percent of the number of drug stores in the CBP and just 2.5 percent of the number of liquor stores, but between 82 and 91 percent of the NETS totals (table 6).

Table 6
Total number of stores in InfoScan, Economic Census, County Business Patterns, TDLinx, and NETS by category, 2008-12

		Number of stores			
Year	Dataset	Drug	Grocery/convenience/dollar/club/mass merchandiser/defense commissary	Liquor	Total stores
2008	InfoScan	19,339	30,227	738	50,304
	EC	-	-	-	-
	CBP	88,445 (21.87%)	272,788 (11.08%)	30,714 (2.40%)	391,947 (12.83%)
	TDLinx	-	195,902 (15.43%)	-	195,902 (25.68%)
	NETS	-	263,415 (11.48%)	900 (82%)	264,315 (19.03%)
2009	InfoScan	19,617	36,970	733	57,320
	EC	-	-	-	-
	CBP	89,184 (22%)	271,986 (13.59%)	31,022 (2.36%)	392,192 (14.62%)
	TDLinx		194,421 (19.02%)		194,421 (29.48%)
	NETS	-	268,816 (13.75%)	882 (83.11%)	269,698 (21.25%)
2010	InfoScan	19,733	37,938	754	58,425
	EC	-	-	-	-
	CBP	90,104 (21.9%)	274,609 (13.82%)	31,491 (2.39%)	396,204 (14.75%)
	TDLinx	-	195,072 (19.45%)	-	195,072 (29.95%)
	NETS	-	268,816 (14.11%)	882 (85.49%)	269,698 (21.66%)
2011	InfoScan	19,772	38,448	782	59,002
	EC	-	-	-	-
	CBP	92,206 (21.44%)	275,543 (13.95%)	31,876 (2.45%)	399,625 (14.76%)
	TDLinx	-	203,424 (18.90%)	-	203,424 (29%)
	NETS	-	268,816 (14.30%)	882 (88.66%)	269,698(21.88%)

—continued

[12] IRI does not calculate projection factors for the "census" component of IRI, since data from all of a retailer's store locations are collected. IRI does calculate projection factors for the "sample" component, which are then applied to the data collected from the sampled locations of a retailer to make them representative of the retailer's full set of store locations. IRI does not sell any part of the "sample" component of InfoScan. ERS is exploring methods of creating projection factors that could be applied to the "census" component of InfoScan in order to make them representative of the universe of store locations.

[13] While it is possible to calculate the number of retailer chains in InfoScan, it is not possible to do so with either the EC or the CBP, as such respondent-identifying information is confidential.

Table 6
Total number of stores in InfoScan, Economic Census, County Business Patterns, TDLinx, and NETS by category, 2008-12—continued

Year	Dataset	Number of stores			Total stores
		Drug	Grocery/convenience/ dollar/club/mass merchandiser/defense commissary	Liquor	
2012	InfoScan	19,770	38,776	828	59,374
	EC	90,959 (21.83%)	278,575 (13.96%)	32,625 (2.47%)	402,159 (14.8%)
	CBP	92,423 (21.48%)	276,202 (14.08%)	32,327 (2.49%)	400,952 (14.85%)
	TDLinx	-	229,797 (16.92%)	-	229,797 (25.91%)
	NETS	-	268,816 (14.46%)	882 (91.27%)	269,698 (22.08%)

Note: Percentages in parentheses are calculated as the number of stores in the InfoScan category divided by the number of stores in the category in the other dataset.

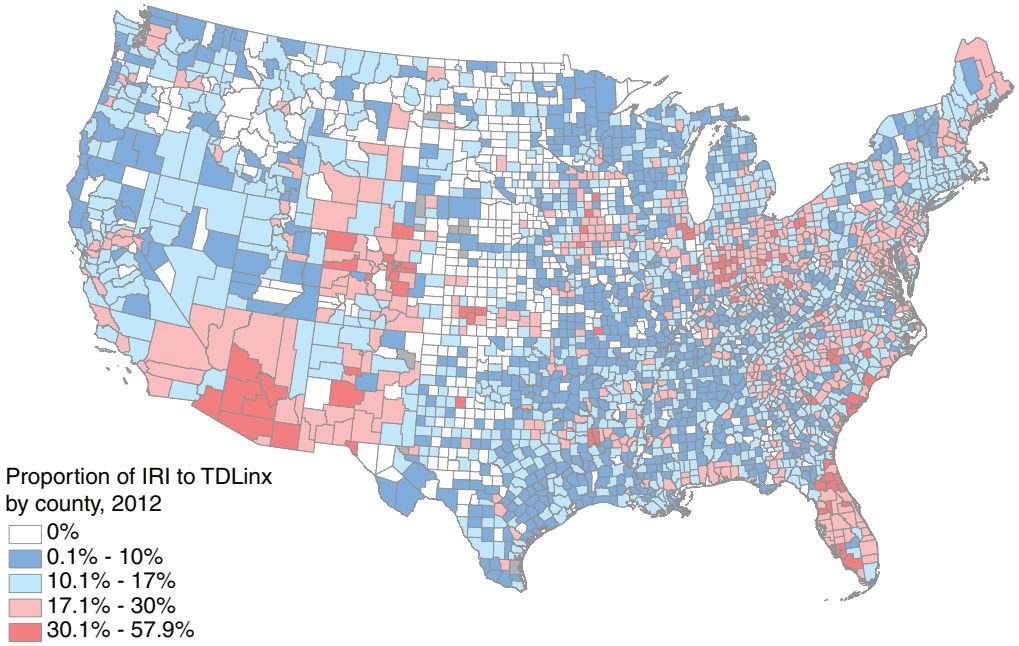
Source: USDA Economic Research Service calculations, IRI InfoScan, Nielsen TDLinx, U.S. Census Bureau Economic Census and County Business Patterns, Walls & Associates National Establishment Time Series (NETS).

The EC reports data for 2012 only, with two different counts of stores: total stores, and number of payroll stores. For the total number of stores, InfoScan includes nearly 22 percent of the drug stores, almost 14 percent of the grocery/convenience/dollar/club/mass merchandise/defense commissary stores, and 2.5 percent of the liquor stores found in the EC. Across all store types, InfoScan includes 14.8 percent of the total number of stores relative to the EC (table 6). When compared to the number of payroll stores in the EC, InfoScan's coverage is unchanged for drug and liquor stores, while the coverage of grocery/convenience/dollar/club/mass merchandise/defense commissary stores is marginally better, rising to 15.6 percent.

Cross-county variation in the coverage of stores in InfoScan relative to TDLinx is shown in figure 4. As with figure 3, the figure includes all stores regardless of reporting level. InfoScan store count coverage is highest in certain areas of the Southwest, Northeast, and Midwest, while it tends to be very low in the middle of the United States and in the Northwest. Comparing figure 4 to figure 3, the areas with most store-count coverage tend to be areas with more InfoScan stores in 2012, although the correlation is not perfect.

Figure 4

Proportion of number of stores in Infoscan to number of stores in TDLinx by county, 2012

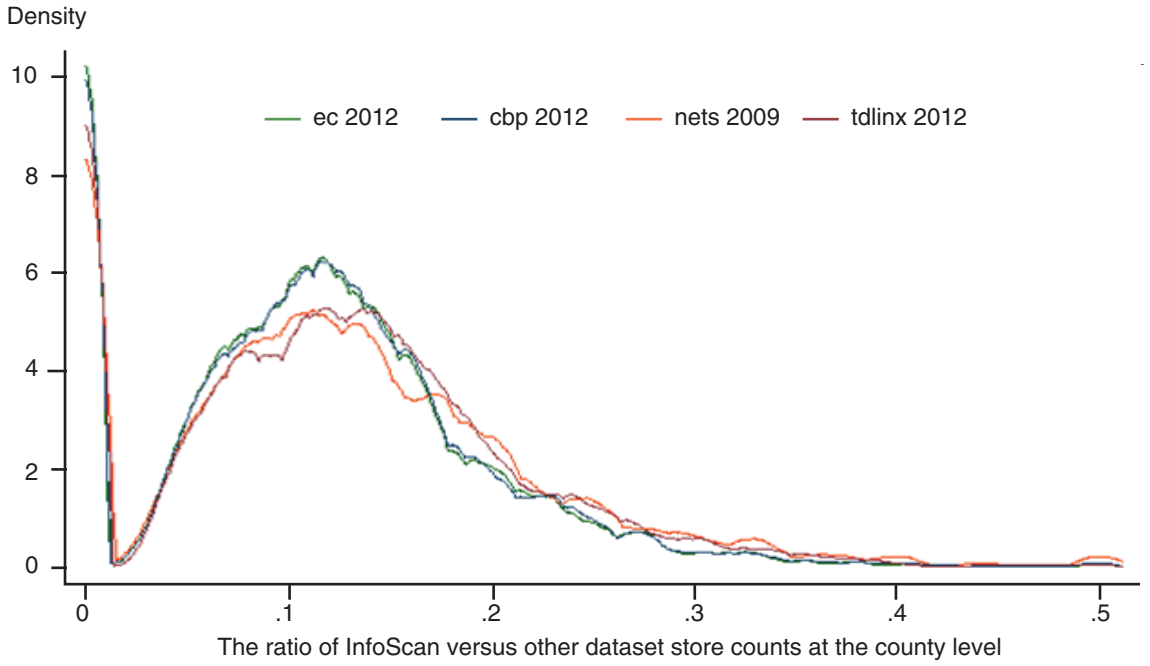


Note: Store counts include those reporting at both the store and RMA (retail marketing areas) level and thus are total store counts.
Source: USDA, Economic Research Service calculations, IRI InfoScan, Nielsen TDLinx.

Figure 5 shows the cross-county distribution of the ratio of store counts in InfoScan to those in the other four datasets. InfoScan store counts are universally lower than those from any other dataset. The distributions are also extremely similar across datasets, with the InfoScan count generally 20 percent or lower than the count from the other four datasets. A sizable number of counties contain no InfoScan stores. Furthermore, the distributions are rather tight: the means shown in table 6 do not hide much heterogeneity in the extent of undercounting across U.S. counties.

Figure 5

Cross-county distribution of InfoScan store counts as a proportion of store counts in other datasets, 2009-12



Note: The figure shows kernel densities of the ratio of InfoScan to other dataset store counts for every county in the United States. Kernel densities were done using an Epanechnikov kernel.

EC 2012 = U.S. Census Bureau Economic Census; CBP 2012 = U.S. Census Bureau County Business Patterns; TDLinx 2012 = Nielsen TDLinx; NETS 2009 = Walls & Associates National Establishment Time Series.

Source: IRI InfoScan, Nielsen TDLinx, U.S. Census Bureau Economic Census and County Business Patterns, Walls & Associates National Establishment Time Series.

Sales Revenue

National aggregate sales revenues from the InfoScan, TDLinx, EC, and NETS are shown in table 7. Since TDLinx reports annual store sales revenue categorically, TDLinx interval values were converted to minimum and maximum values, which leads to the county-level sales revenue bandwidths used in table 7. As in table 6, store types are split into three categories to avoid complications related to classification differences across surveys.

Table 7

Total sales in InfoScan, Economic Census, TDLinx, and NETS by category, 2008-12

Year	Dataset	Sales ($ million)			
		Drug	Grocery/convenience/ dollar/club/mass merchandiser/defense commissary	Liquor	Total sales
2008	InfoScan	8,318	179,216	1,719	189,253
	EC	-	-	-	-
	TDLinx (max)	-	1,575,901 (11.37%)	-	1,575,901 (12.01%)
	TDLinx (min)	-	871,381 (20.57%)	-	871,381 (21.72%)
	NETS	-	658,349 (27.22%)	895 (192.07%)	659,244(28.71%)
2009	InfoScan	8,020	277,061	1,930	287,011
	EC	-	-	-	-
	TDLinx (max)	-	1,708,820 (16.21%)	-	1,708,820 (16.8%)
	TDLinx (min)	-	901,441 (30.74%)	-	901,441 (31.84%)
	NETS	-	666,791 (41.55%)	833 (231.69%)	667,624(42.99%)
2010	InfoScan	8,346	282,061	2,108	292,515
	EC			-	-
	TDLinx (max)	-	1,678,512 (16.8%)	-	1,678,512 (17.43%)
	TDLinx (min)	-	909,329 (31.02%)	-	909,329 (32.17%)
	NETS	-	666,791 (42.30%)	833 (253.06%)	667,624 (43.81%)
2011	InfoScan	8,953	291,695	2,278	302,926
	EC	-	-	-	-
	TDLinx (max)	-	1,724,153 (16.92%)	-	1,724,153 (17.57%)
	TDLinx (min)	-	934,231 (31.22%)	-	934,231 (32.43%)
	NETS	-	666,791 (43.75%)	833 (273.47%)	667,624(45.37%)
2012	InfoScan	14,046	311,443	2,761	328,250
	EC (all sales)	255,203 (5.5%)	1,398,641 (22.27%)	33,772 (8.18%)	1,687,616 (19.45%)
	EC (food sales for payroll establishments)	9,970 (140.88%)	627,345 (49.64%)	1,911 (144.48)	639,226 (51.35%)
	TDLinx (max)	-	1,895,576 (16.43%)	-	1,895,576 (17.32%)
	TDLinx (min)	-	1,047,428 (29.73%)	-	1,047,428 (31.34%)
	NETS	-	666,791 (46.71%)	833 (331.45%)	667,624 (49.17%)

Note: Percentages in parentheses are calculated as the total sales of stores in the category in InfoScan divided by the total sales (or food sales) in the category in the other dataset. InfoScan includes only food sales, while TDLinx and NETS include both food and nonfood sales.

Source: USDA, Economic Research Service calculations, IRI InfoScan, Nielsen TDLinx, U.S. Census Bureau EC, Walls & Associates National Establishment Time Series (NETS).

InfoScan sales revenues for the grocery/convenience/dollar/club/mass merchandise/defense commissary category are considerably less than in other datasets because store counts in InfoScan are much lower (table 6). In addition, ERS only purchases InfoScan data for food and alcohol products, while sales for the other datasets reflect food and nonfood purchases. However, the share of sales revenue reported in InfoScan relative to the sales revenue reported in the comparison datasets is larger than

the share of stores, implying that the stores missing from InfoScan are smaller retailers. For most years, the InfoScan sales are between 16 and 32 percent of TDLinx sales, depending on whether the minimum or maximum is used in the latter dataset. There are 26-29 percent as many stores in InfoScan as in TDLinx, which is within the range of sales categories in table 7. Depending on the year, aggregate sales in the InfoScan data are between 29 and 49 percent of reported sales in NETS. This percentage is larger than the store count comparisons in table 6, suggesting that the InfoScan stores are higher volume than the NETS establishments.

Prior to 2012, the only other dataset that contains sales revenues for liquor stores is NETS. Sales revenues from the InfoScan data are two to three times larger than in the NETS data, though there are more stores in NETS than in InfoScan (table 6).

The EC reports sales revenue in 2012 only, and two different sales revenue figures were available from this dataset: all sales, and food sales for payroll establishments. Across store types, the sales revenue in the EC for all sales is far larger than in InfoScan, particularly among liquor stores and drug stores. For liquor stores and the aggregate grocery store group, these differences approximate those in table 6, while the sales revenue disparity is much larger than the store number disparity among drug stores. When comparing food sales for payroll establishments only (EC), InfoScan's coverage improves dramatically. InfoScan covers 41 and 44 percent more sales than the EC for drug and liquor stores, respectively, while covering almost half of the sales from the grocery/convenience/dollar/club/mass merchandise/defense commissary category relative to the EC. Overall, InfoScan includes about 51 percent of the food sales for payroll establishments in the EC when using this narrower definition of sales (table 7).

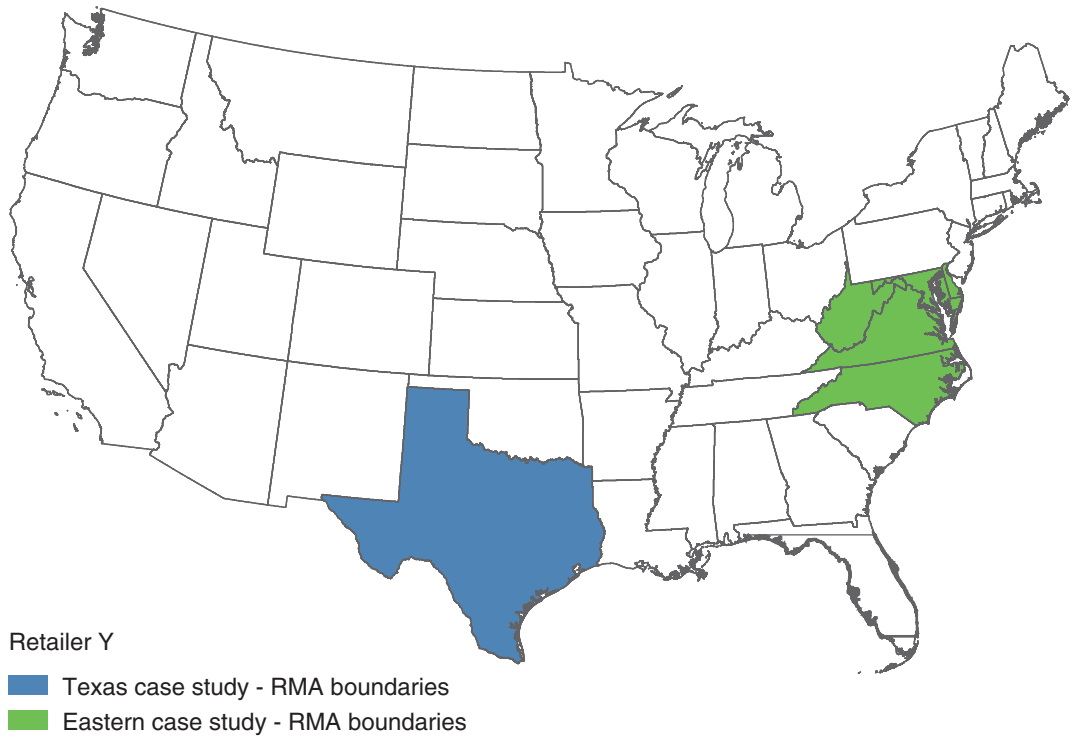
Case Studies of InfoScan Coverage

The county-level estimates in figures 3-5 include data both from companies that provide store-level data and those that provide RMA level data. However, State or county-level sales information can only be calculated from retailers reporting at the store level, since many companies that provide RMA-level data define their RMAs using boundaries that cross multiple counties and/or States.

In order to compare store counts and sales revenue at the subnational level, we perform two case studies of scenarios in which retailer RMAs line up with a State border and/or a set of contiguous county borders. This allows us to add together the aggregated store-level data from each of these regions and the RMA-level data such that the store count and sales information refer to the same geographic areas. Each retailer sets its own RMA, and typically this varies across retailers. We chose areas that align with State and county borders for three of the largest U.S. retailers that report data at the RMA level only. Figure 6 shows the two RMAs—one for all of Texas and one for Virginia, North Carolina, West Virginia, Maryland, Delaware and Washington, DC—for one of the retailers (Retailer Y). These two regions are examined as the Texas and Eastern case studies, respectively. Retailer X (figure 7) has several RMAs within Texas, and an Eastern RMA that is geographically smaller than Retailer Y's.[14]

Figure 6

Retailer Y's Retail Marketing Areas (RMAs) for case studies

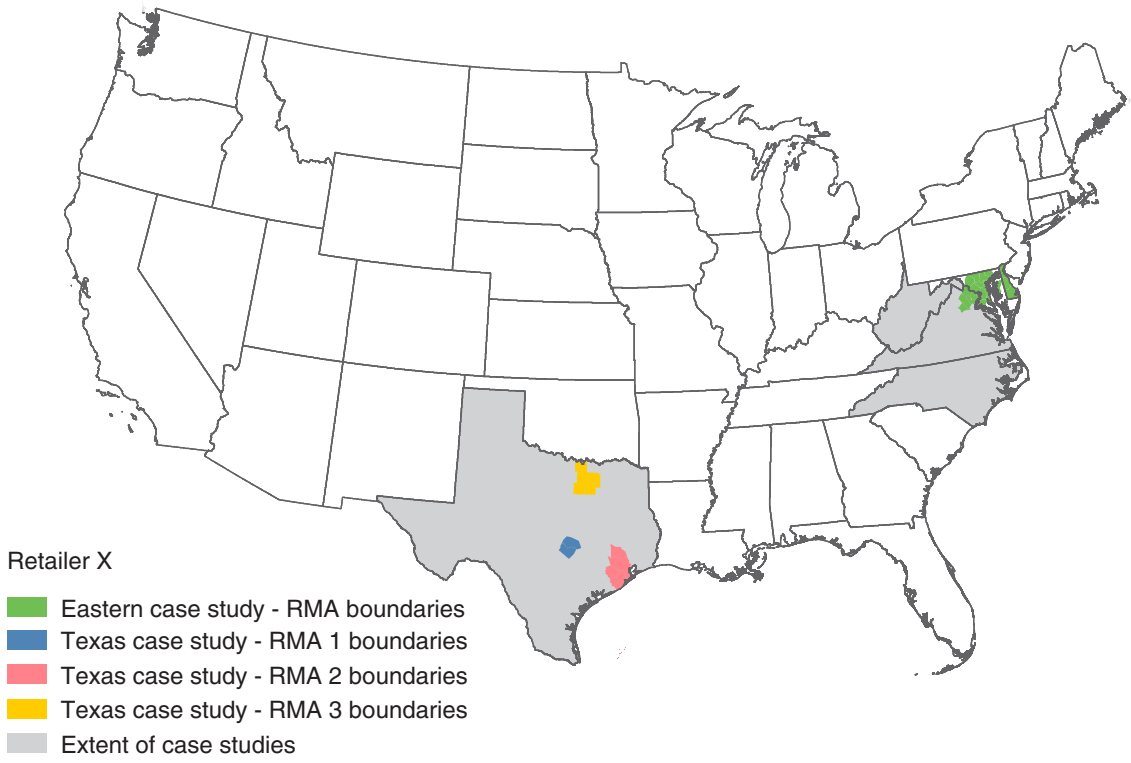


Source: IRI InfoScan.

[14] RMA boundaries are determined by the individual retailers. Retailer Y's RMA boundaries encompass individual States or groups of States, whereas Retailer X's boundaries are clusters of counties. Thus, even though the respective RMAs are geographically concurrent, the geographic extent of RMAs is different.

Figure 7

Retailer X's Retail Marketing Areas (RMAs) for case studies



Source: IRI InfoScan.

Store counts for the Texas and Eastern case studies—including all RMA-level stores as well as store-level stores in each region—are shown in tables 9 and 10, respectively.

Texas

In the Texas case study, InfoScan undercounts the number of drug stores relative to the CBP by twice as much as it does at the national level (table 6); InfoScan includes 10 percent of the drug stores in the CBP in the case study, compared to 21 percent at the national level. While InfoScan also undercounts liquor stores in this case study relative to the CBP and the EC (2012), it does so to a lesser extent than at the national level, including between 6 and 8 percent of the stores in the two other surveys versus approximately 2.5 percent at the national level. InfoScan's coverage of liquor stores exceeds that of NETS for the case study, including two to four times as many stores as NETS, while at the national level InfoScan includes 82-91 percent of the stores in NETS. For the grocery/convenience/dollar/club/mass merchandise/defense commissary category, InfoScan's coverage of stores in this case study is similar to that at the national level, relative to the CBP; compared to TDLinx and NETS, however, InfoScan includes slightly fewer stores in this case study than it does at the national level. Across all store categories, InfoScan undercounts the number of stores in this case study relative to all of the other datasets by a larger margin than at the national level.

Table 8

Texas case study: number of stores in InfoScan, Economic Census, County Business Patterns, TDLinx, and NETS, 2008-12

Year	Datasets	Number of stores			Total stores
		Drug	Grocery/convenience/ dollar/club/mass merchandiser/defense commissary	Liquor	
2008	InfoScan	648	2,586	111	3,345
	EC	-	-	-	-
	County Business Patterns	6,412 (10.11%)	21,884 (11.82%)	1,752 (6.34%)	30,048 (11.13%)
	TDLinx	-	17,696 (14.61%)	-	17,696 (18.9%)
	NETS	-	23,014 (11.24%)	47 (236.17%)	23,061 (14.51%)
2009	InfoScan	681	2,871	118	3,670
	EC	-	-	-	-
	County Business Patterns	6,448 (10.56%)	21,964 (13.07%)	1,744 (6.77%)	30,156 (12.17%)
	TDLinx	-	17,620 (16.29%)	-	17,620 (20.83%)
	NETS	-	23,948 (11.99%)	44 (268.18%)	23,992(15.30%)
2010	InfoScan	690	2,864	130	3,684
	EC	-	-	-	-
	County Business Patterns	6,612 (10.44%)	22,274 (13.86%)	1,768 (7.35%)	30,654 (12.02%)
	TDLinx	-	17,796 (16.09%)	-	17,796 (20.70%)
	NETS	-	23,948 (11.96%)	44 (295.45%)	23,992 (15.36%)
2011	InfoScan	694	2,915	141	3,750
	EC	-	-	-	-
	County Business Patterns	6,839 (10.15%)	22,568 (12.92%)	1,824 (7.73%)	31,231 (12.01%)
	TDLinx	-	18,632 (15.65%)	-	18,632 (20.13%)
	NETS	-	23,948 (12.17%)	44 (320.45%)	23,992(15.63%)
2012	InfoScan	710	2,984	165	3,859
	EC	6,891 (10.3%)	23,138 (12.9%)	1,971 (8.37%)	32,000 (12.06%)
	County Business Patterns	6,960 (10.2%)	22,955 (13%)	1,957 (8.43%)	31,872 (12.11%)
	TDLinx	-	21,433 (13.92%)	-	21,433 (18%)
	NETS	-	23,948 (12.46%)	44 (375%)	23,992(16.08%)

Percentages in parentheses are calculated as the number of stores in the category in InfoScan divided by the number of stores in the category in the other dataset.

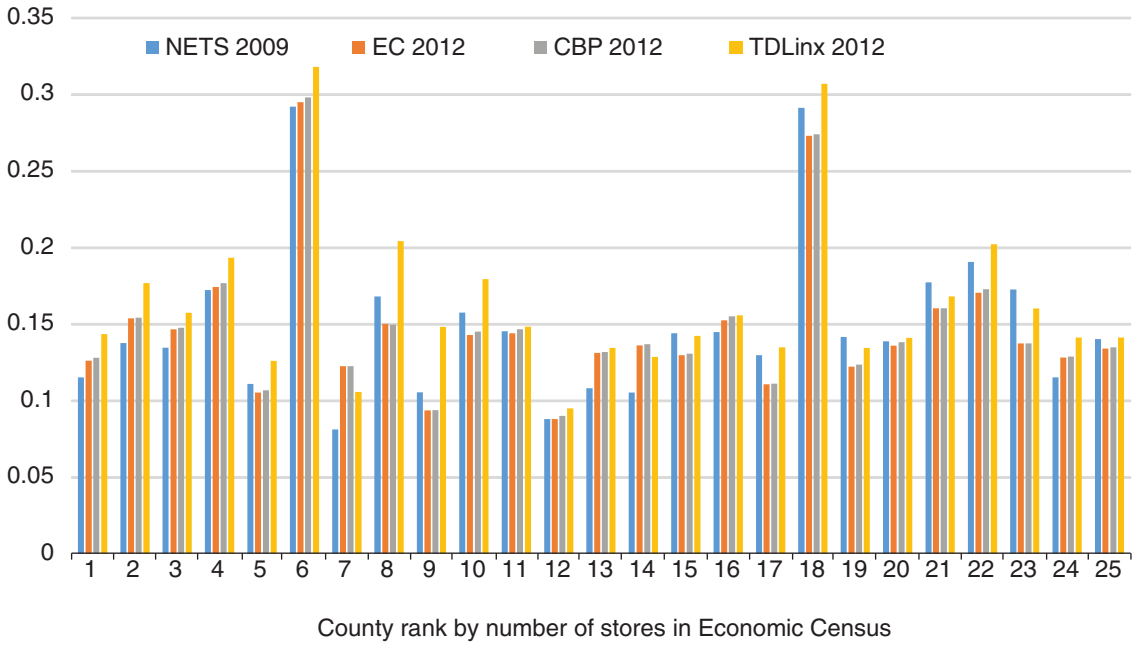
Source: USDA, Economic Research Service calculations, IRI InfoScan, Nielsen TDLinx, U.S. Census Bureau Economic Census and County Business Patterns, Walls & Associates National Establishment Time Series (NETS).

Figure 8 presents store counts in InfoScan as a proportion of store counts in the other four datasets among the 25 largest counties in the Texas case study. The counties are ranked by the number of stores in the 2012 EC. Store counts by county for these 25 counties are shown in appendix table A4. Consistent with figure 5, the InfoScan data systematically undercount the number of stores. Importantly, the InfoScan undercount is consistent across datasets and varies little across counties. In particular, there is no relationship between the extent of undercounting and the number of stores

in the county (according to the EC data). Figure 8 also highlights the remarkable consistency of the undercount ratio across datasets: it is not the case that the InfoScan data are more accurate in a meaningful way using some datasets than in others.

Figure 8

InfoScan store counts as a proportion of store counts in other datasets in the 25 largest counties in the Texas case study, ranked by Economic Census store counts, 2009-12



Note: See appendix table A.4 for store counts and county names. Store counts include all stores regardless or reporting level. Counties are ranked by the number of stores in the 2012 Economic Census.

EC 2012 = U.S. Census Bureau Economic Census, CBP 2012 = U.S. Census Bureau County Business Patterns, TDLinx 2012 = Nielsen TDLinx, NETS 2009 = Walls & Associates National Establishment Time Series.

Source: IRI InfoScan, Nielsen TDLinx, U.S. Census Bureau Economic Census and County Business Patterns, Walls & Associates National Establishment Time Series.

Eastern

In the Eastern case study, InfoScan includes 15 percent of the drug stores in the CBP, while at the national level it includes 22 percent. Similarly, comparing InfoScan's coverage of Eastern drug stores relative to the EC (2012) reveals that the former undercounts the latter by a larger margin than at the national level. InfoScan's coverage of liquor stores in this case study is considerably worse than at the national level, including just 0.1 percent of the stores in the CBP and 2 to 64 percent of the stores in NETS, while at the national level those figures are around 2 percent and 80 to 91 percent for the respective datasets. InfoScan's coverage of liquor stores improves considerably in 2012, but remains markedly lower than at the national level for all the other datasets.

The opposite pattern emerges for the grocery/convenience/dollar/club/mass merchandise/defense commissary category, where InfoScan still undercounts the number of stores relative to the other datasets, but to a lesser degree than at the national level. Across all store categories, InfoScan's coverage in the Eastern case study relative to that at the national level is mixed; compared to TDLinx, InfoScan generally includes 0.5 to 2 percent fewer stores in the case study, while including

slightly more stores than the CBP (NETS comparisons vary depending on the year compared to InfoScan). That the coverage of store counts in InfoScan for the Texas case study differed from the coverage in the Eastern case study indicates that there is geographic heterogeneity in InfoScan's coverage of store counts.

Table 9
Eastern case study: number of stores in InfoScan, Economic Census, County Business Patterns, TDLinx, and NETS, 2008-2012

Year	Datasets	Number of stores			
		Drug	Grocery/convenience/ dollar/club/mass merchandiser/defense commissary	Liquor	Total stores
2008	InfoScan	1,152	3,656	1	4,809
	EC	-	-	-	-
	CBP	7,753 (14.86%)	25,630 (14.26%)	2,597 (0.04%)	35,980 (13.37%)
	TDLinx	-	19,094 (19.15%)	-	19,094 (25.19%)
	NETS	-	24,479 (14.94%)	49 (2.04%)	24,528 (19.61%)
2009	InfoScan	1,167	4,090	3	5,260
	EC	-	-	-	-
	CBP	7,761 (15.04%)	25,490 (16.05%)	2,607 (0.12%)	35,858 (14.67%)
	TDLinx	-	18,959 (21.75%)	-	18,959 (27.74%)
	NETS	-	24,754 (16.52%)	47 (6.38%)	24,801 (21.21%)
2010	InfoScan	1,190	4,145	3	5,338
	EC	-	-	-	-
	CBP	7,849 (15.16%)	25,608 (16.19%)	2,638 (0.11%)	36,095 (14.79%)
	TDLinx	-	19,012 (21.8%)	-	19,012 (28.08%)
	NETS	-	24,754 (16.74%)	47 (6.38%)	24,801 (21.21%)
2011	InfoScan	1,203	4,198	2	5,403
	EC	-	-	-	-
	CBP	8,036 (14.97%)	25,587 (16.41%)	2,655 (0.08%)	36,278 (14.89%)
	TDLinx	-	19,762 (21.24%)	-	19,762 (27.34%)
	NETS	-	24,754 (16.96%)	47 (4.26%)	24,801 (21.79
2012	InfoScan	1,217	4,206	30	5,453
	EC	7,905 (15.4%)	25,911 (16.23%)	2,693 (1.11%)	36,509 (14.94%)
	CBP	8,088 (15.05%)	25,682 (16.38%)	2,683 (1.12%)	36,453 (14.96%)
	TDLinx	-	22,613 (18.6%)	-	22,613 (24.11%)
	NETS	-	24,754 (16.99%)	47 (63.8 3%)	24,801 (21.99%)

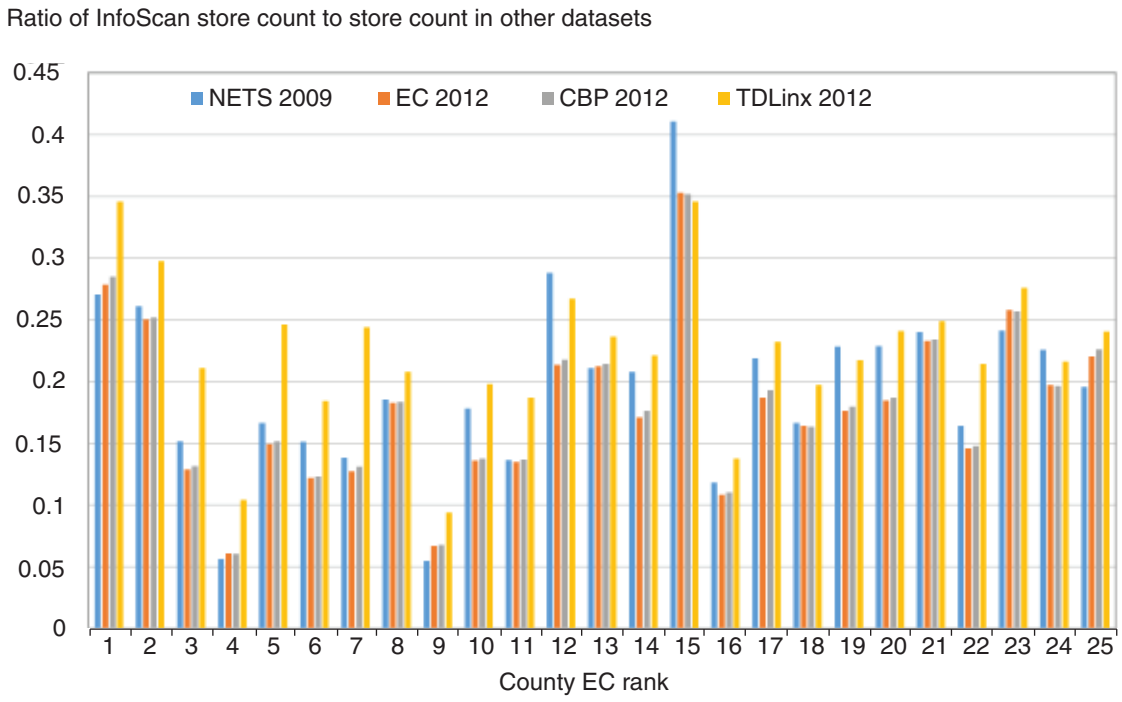
Note: Percentages in parentheses are calculated as the number of stores in the category in InfoScan divided by the number of stores in the category in the other dataset.

Source: USDA, Economic Research Service calculations, IRI InfoScan, Nielsen TDLinx, U.S. Census Bureau Economic Census and County Business Patterns, Walls & Associates National Establishment Time Series (NETS).

Figure 9 presents similar information to figure 8 but for the Eastern case study. County-level store counts for these counties are shown in appendix table A5. The conclusions drawn from the two figures are extremely similar: InfoScan store counts are much lower than those in the other datasets, and the extent of undercounting varies little across counties and datasets and is unrelated to the number of stores in the county (according to the EC data). Figures 8 and 9 suggest that the national distribution of undercounting by InfoScan across counties shown in Figure 5 is not hiding any substantial heterogeneity in relative store counts across datasets.

Figure 9

InfoScan store counts as a proportion of store counts in other datasets in the 25 largest counties in the Eastern case study, ranked by Economic Census store counts, 2009-12



Note: See appendix table A.4 for store counts and county names. Store counts include all stores regardless or reporting level. Counties are ranked by the number of stores in the 2012 Economic Census.

EC 2012 = U.S. Census Bureau Economic Census, CBP 2012 = U.S. Census Bureau County Business Patterns, TDLinx 2012 = Nielsen TDLinx, NETS 2009 = Walls & Associates National Establishment Time Series.

Source: IRI InfoScan, Nielsen TDLinx, U.S. Census Bureau Economic Census and County Business Patterns, Walls & Associates National Establishment Time Series.

The sales revenue for the Texas and Eastern case studies is shown in tables 10 and 11, again revealing geographic heterogeneity in InfoScan's coverage. In the Texas case study, InfoScan undercounts sales revenues to a greater degree than at the national level (table 7) relative to TDLinx and NETS for 2008-12, and relative to the EC in 2012— InfoScan's coverage is notably poorer than at the national level. In the Eastern case study, InfoScan's coverage is similar to its national coverage for most years, and is noticeably better than in the Texas case study.

Table 10

Texas case study sales in InfoScan, Economic Census, TDLinx, and NETS, 2008-12

Year	Datasets	Total sales ($ million)
2008	InfoScan	8,832
	EC	-
	TDLinx (max)	135,601 (6.51%)
	TDLinx (min)	76,526 (11.54%)
	NETS	52,404 (16.85%)
2009	InfoScan	17,746
	EC	-
	TDLinx (max)	159,824 (11.1%)
	TDLinx (min)	77,757 (22.82%)
	NETS	52,087 (34.07%)
2010	InfoScan	17,860
	EC	-
	TDLinx (max)	143,339 (12.46%)
	TDLinx (min)	77,622 (23.01%)
	NETS	52,087 (34.29%)
2011	InfoScan	17,125
	EC	-
	TDLinx (max)	145,993 (11.73%)
	TDLinx (min)	80,011 (21.4%)
	NETS	52,087 (32.88%)
2012	InfoScan	20,091
	EC	117,807 (17.05%)
	TDLinx (max)	157,065 (12.79%)
	TDLinx (min)	86,619 (23.19%)
	NETS	52,087 (38.57%)

Note: Percentages in parentheses are calculated as the sales revenue in InfoScan divided by the sales revenue in the other dataset.

Source: USDA, Economic Research Service calculations, IRI InfoScan, Nielsen TDLinx, U.S. Census Bureau Economic Census, Walls & Associates National Establishment Time Series (NETS).

Table 11

Eastern case study: sales in InfoScan, Economic Census, TDLinx, and NETS, 2008-12

Year	Datasets	Total sales ($ million)
2008	InfoScan	19,334
	EC	-
	TDLinx (max)	145,249 (13.31%)
	TDLinx (min)	81,292 (23.78%)
	NETS	60,690 (31.86%)
2009	InfoScan	26,528
	EC	-
	TDLinx (max)	155,869 (17.02%)
	TDLinx (min)	84,790 (31.29%)
	NETS	62,013 (42.78%)
2010	InfoScan	26,956
	EC	-
	TDLinx (max)	155,138 (17.38%)
	TDLinx (min)	85,409 (31.56%)
	NETS	62,013 (43.47%)
2011	InfoScan	26,700
	EC	-
	TDLinx (max)	157,734 (16.94%)
	TDLinx (min)	87,507 (30.53%)
	NETS	62,013 (43.06%)
2012	InfoScan	29,565
	EC	145,316 (20.35%)
	TDLinx (max)	174,410 (16.95%)
	TDLinx (min)	98,614 (29.98%)
	NETS	62,013 (47.68%)

Note: Percentages in parentheses are calculated as the sales revenue in InfoScan divided by the sales revenue in the other dataset.

Source: USDA, Economic Research Service calculations, IRI InfoScan, Nielsen TDLinx, U.S. Census Bureau Economic Census, Walls & Associates National Establishment Time Series (NETS).

Discussion

At the national level, the coverage of store counts and sales revenue in the InfoScan data purchased by ERS is lower than other datasets by a large magnitude. However, coverage varies by geographic area and dataset. While InfoScan's coverage of store counts and sales revenues was generally worse than at the national level in the Texas case study, its overall coverage in the Eastern case study was better.

Many factors may explain InfoScan's undercounting, relative to other datasets, of both the number of establishments and sales revenues. The full InfoScan dataset includes both the "census" and "sample" components, whereas InfoScan data acquired by ERS includes only the "census" component. As a result, both the number of establishments and sales revenues for InfoScan used in this analysis are a subset of the full InfoScan dataset, and the degree to which InfoScan undercounts both metrics relative to the other datasets would be lower if it were possible to include the "sample" component in the comparisons.[15] Also contributing to the InfoScan undercounting is the specific definition used by IRI in their grocery store channel. Only stores having $2 million or more in annual sales are included in the grocery store channel of InfoScan. This exclusion of smaller grocery stores will inherently result in InfoScan undercounting both the number of establishments and sales revenue versus the other datasets. The exclusion of smaller grocery stores in InfoScan may have additional implications with respect to certain policy issues. For example, if low-income consumers shop more or less frequently at smaller grocery stores than is the case for the general population, then researchers may need to examine whether the InfoScan data are adequately representative of retail sales across consumer income groups.

One limitation applicable only to the sales revenue comparisons is that the InfoScan data acquired by ERS include only food products, while the other datasets include both food and nonfood purchases. However, the sales revenue figures in TDLinx, the CBP, and NETS do not distinguish between food and nonfood sales. As such, it is unsurprising that comparing the sales revenue of food products in InfoScan with the combined food and non-food sales revenue of the other datasets results in considerably lower values for InfoScan. This limitation is readily apparent in the comparisons of sales revenue in InfoScan with those of the EC, where two different sales figures were used—the combined food and nonfood sales—and the food sales at payroll establishments; InfoScan's coverage of sales more than doubled when the latter figure was used.

While two case studies (Texas, Eastern) showed that differences in the number of establishments and sales revenue between InfoScan and the other datasets varied geographically, there are limitations to this approach. The two case studies were selected solely because the retailer RMAs lined up well with State border(s) and/or contiguous sets of counties; due to each RMA-level retailer defining its own RMA boundaries, the number of additional case studies that could be examined using this method is extremely limited. As such, the results of these two case studies may not be representative of InfoScan's coverage in other areas of the Nation.

This report focused on geographic coverage; other considerations related to InfoScan's representativeness are also important. For example, researchers using InfoScan's nutritional and product claims information should be aware that existing research on the completeness of these variables

[15] A primary data delivery to IRI's clients is projected estimates combining census and sample components, but these are not included in the ERS dataset.

is limited. Muth et al. (2016) calculated the share of UPCs in InfoScan that had matching nutrition data but used the broadest definition of a match: a UPC that has at least one field of nutrition data. A comparison of the coverage of nutrition and claims data between InfoScan and Gladson, undertaken by Giombi et al. (2018) for the soup product category, found the coverage to be comparable, but it did not compare nutrition and claims data of other product categories between the two datasets or determine whether InfoScan's coverage of nutrition and claims data varies from one product category to another. Additionally, we are unaware of any data source that could serve as a reasonable proxy for the universe of UPCs, so there is no basis for assessment of the coverage of InfoScan in terms of UPCs.

Conclusion

InfoScan consistently reports fewer stores and lower sales revenues than the TDLinx, CBP, EC, and NETS datasets at the national level, though sales revenue in InfoScan is closer to those in the other datasets. The two case studies show that InfoScan's coverage of both store counts and sales revenue varies from one geographic area to another.

Finding comparable groups for gauging the representativeness of the sales revenue in the InfoScan data that ERS purchases is difficult. ERS only purchases information on food products in InfoScan, and the EC is the only other source available for this information. These data are only available nationally and are restricted to payroll establishments. Also, the RMA (regional) and non-RMA data available in InfoScan make comparisons of sales revenue at a more granular level of geography difficult as well because each retailer defines their own RMA boundaries, leading to inconsistent RMA definitions across retailers.

The limited coverage of the InfoScan data relative to the TDLinx, CBP, EC, and NETS data with respect to number of establishments and sales revenue means that at the aggregate/national level, these other datasets may be more representative. The geographic variability of InfoScan's coverage of store counts and sales revenues may also make subnational analyses—including research studies examining regional price variation (e.g., regional price indexes), other spatial analysis, and food access—problematic. Additionally, the unavailability of weights for InfoScan may complicate attempts to conduct demand analysis; the IRI household scanner data may be more appropriate for demand analysis (Sweitzer et al., 2017). This report demonstrates the need for users to carefully evaluate whether InfoScan's coverage of the number of stores and volume of sales is sufficiently representative of the universe for their particular research project.

InfoScan remains a valuable data source for analysis of topics requiring UPC-level transaction data for food purchases, with the caveat that results are more relevant to larger stores. Also, 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, such as how the entry of a new food retailer affects the broader retail food market, though researchers should be mindful of InfoScan's representativeness issues discussed in this report.

ERS is working with external colleagues to address some of the representativeness issues in the InfoScan data. With RTI International, ERS is exploring the construction of projection factors for the InfoScan data so that the sales and store counts in the "census" component of InfoScan that ERS purchases will be representative nationally and at the Census division level.

References

Bureau of Labor Statistics. 1996. "Briefing on the Consumer Price Index." Economic News Release, December 3.

Carlson, A., and E. Jaenicke. 2016. *Changes in Retail Organic Price Premiums from 2014 to 2010*. Economic Research Report 209. *U.S. Department of Agriculture, Economic Research Service.*

Çakır, M., T.K. Beatty, M.A. Boland, T.A. Park, S. Snyder, and Y. Wang. 2018. "Spatial and Temporal Variation in the Value of the Women, Infants, and Children Program's Fruit and Vegetable Voucher," *American Journal of Agricultural Economics* 100(3): 691-706.

Cho, Clare, Patrick McLaughlin, Eliana Zeballos, Jessica Kent, and Chris Dicken. "Capturing the Complete Food Environment with Commercial Data: A Comparison of TDLinx, ReCount, and NETS." U.S. Department of Agriculture, Economic Research Service. Technical Bulletin (under review).

Davis, Christopher G., Diansheng Dong, Don P. Blayney, and Ashley Owens. 2010. *An Analysis of U.S. Household Dairy Demand*, TB-1928, U.S. Department of Agriculture, Economic Research Service. Dec.

Dong, D., and B.-H. Lin. 2009. *Fruit and Vegetable Consumption by Low-Income Americans*. Economic Research Report 70. U.S. Department of Agriculture, Economic Research Service.

Einav, L., E. Leibtag, and A. Nevo. 2008. *On the Accuracy of Nielsen Homescan Data*. Economic Research Report 69. *U.S. Department of Agriculture, Economic Research Service.*

Fleischhacker, S.E., D.A. Rodriguez, K.R. Evenson, A. Henley, Z. Gizlice, D. Soto, and G. Ramachandran. 2012. "Evidence for validity of five secondary data sources for enumerating retail food outlets in seven American Indian Communities in North Carolina*," International Journal of Behavioral Nutrition and Physical Activity* 9 (137):1-14.

Giombi, Kristen C., Mary K. Muth, and David Levin. 2018. "A comparative analysis of hedonic models of nutrition information and health claims on food products: An application to soup products," *Journal of Food Products Marketing*. 24:7, 906-926.

Gordon-Larsen, P. P.E. Rummo, and S.S. Albrecht. 2015. "Field validation of food outlet databases: The Latino food environment in North Carolina," *Public Health Nutrition* 18(6): 977-982.

Harding, Matthew, and Michael Lovenheim. 2017. "The effect of prices on nutrition: Comparing the impact of product- and nutrient-specific taxes," *Journal of Health Economics* 53:53-71.

Ivancic, L., W.E. Diewert, and K.J. Fox. 2011. "Scanner Data, Time Aggregation, and the Construction of Price Indexes," *Journal of Econometrics* 161(1): 24-35.

IRI InfoScan Data, ERS-USDA version, 2012.

Liese, A.D., N. Colabianchi, A. P. Lamichhane, T.L. Barnes, J.D. Hibbert, D.E. Porter, M.D. Nichols, and A.B. Lawson. 2010. "Validation of 3 Food Outlet Databases: Completeness

and Geospatial Accuracy in Rural and Urban Food Environments," *American Journal of Epidemiology* 172(111): 1324-1333.

Low, S.A., Aaron Adalja, Elizabeth Beaulieu, Nigel Key, Steve Martinez, Alex Melton, Agnes Perez, Katherine Ralston, Hayden Stewart, Shellye Suttles, Stephen Vogel, and Becca B.R. Jablonski. 2015. *Trends in U.S. Local and Regional Food Systems: A Report to Congress*. Administrative Publication-068. U.S. Department of Agriculture, Economic Research Service.

Lusk, J., and K. Brooks. 2011. "Who Participates in Household Scanner Data?" *American Journal of Agricultural Economics* 93(1): 226-240.

Mancino, L., F. Kuchler, and E. Leibtag. 2008. "Getting consumers to eat more whole-grains: The role of policy, information, and food manufacturers," *Food Policy* 33: 489-496.

Martinez, Stephen W., and David Levin. 2017. *An Assessment of Product Turnover in the U.S. Food Industry and Effects on Nutrient Content,* EIB-183, U.S. Department of Agriculture, Economic Research Service, November.

Muth, M.K., P.H. Siegel, and C. Zhen. 2007. "ERS Data Quality Study Design." Report prepared for U.S. Department of Agriculture, Economic Research Service.

Muth, M., M. Sweitzer, D. Brown, K. Capogrossi, S. Karns, D. Levin, A. Okrent, P. Siegel, and C. Zhen. 2016. *Understanding IRI Household-Based and Store-Based Scanner Data*. TB-1942, U.S. Department of Agriculture, Economic Research Service, April.

Nielsen. 2010. "TDLinx Data Dictionary."

Powell, L.M., E. Han, S.N. Zenk, T. Khan, C.M. Quinn, K.P. Gibbs, O. Pugach, D.C. Barker, E.A. Resnick, J. Myllyluoma, and F.J. Chaloupka. 2011. "Field validation of secondary commercial data sources on the retail food outlet environment in the U.S.," *Health and Place* 17: 1122-1131.

Smith, Travis A., Biing-Hwan Lin, and Jonq-Ying Lee. *Taxing Caloric Sweetened Beverages: Potential Effects on Beverage Consumption, Calorie Intake, and Obesity*, ERR-100, U.S. Department of Agriculture, Economic Research Service, July 2010.

Sweitzer, Megan, Derick Brown, Shawn Karns, Mary K. Muth, Peter Siegel, and Chen Zhen. 2017. *Food-at-Home Expenditures: Comparing Commercial Household Scanner Data From IRI and Government Survey Data,* TB-1946, U.S. Department of Agriculture, Economic Research Service, Sept.

U.S. Census Bureau. 2015a. 2007 and 2012 Economic Census. FTP website.

U.S. Census Bureau. 2015b. County Business Patterns Complete County File dataset.

U.S. Census Bureau. 2016a. "County Business Patterns (CBP)." About This Program, 7 Sept.

U.S. Census Bureau. 2016b. "Economic Census." FAQ, 29 Aug.

Volpe, R., and A.M. Okrent. 2012. *Assessing the Healthfulness of Consumers' Purchases.* Economic Information Bulletin 102. U.S. Department of Agriculture, Economic Research Service.

Walls and Associates. 2012. "National Establishment Time-Series (NETS) Database: 2012 Database Description."

Zhen, C., J.L. Taylor, M. Muth, and E. Leibtag. 2009. "Understanding Differences in Self-Reported Expenditures between Household Scanner Data and Diary Survey Data: A Comparison of Homescan and Consumer Expenditure Survey," *Review of Agricultural Economics* 31(3): 470-492.

Zhen, C., E. Finkelstein, J. Nonnemaker, S. Karns, and J. Todd. 2014. "Predicting the Effects of Sugar-Sweetened Beverage Taxes on Food and Beverage Demand in a Large Demand System," *American Journal of Agricultural Economics* 96(1): 1-25.

Appendix

Table A1
Mapping between North American Industry Classification System (NAICS) and TDLinx channels

NAICS	NAICS description	Channel (TDLinx)	Map1 (TDLinx)
445110	Supermarkets and other grocery stores (excluding convenience stores)	05 - grocery	05 - grocery | 07 - convenience store
445120	Convenience store	07 - convenience store	05 - grocery | 07 - convenience store
445210	Meat markets	05 - grocery	05 - grocery | 07 - convenience store
445220	Fish and seafood markets	05 - grocery	05 - grocery | 07 - convenience store
445230	Fruit and vegetable markets	05 - grocery	05 - grocery | 07 - convenience store
445310	Beer, wine and liquor stores	02 - liquor trade	02 - liquor trade
446110	Pharmacies and drug stores	03 - drug trade	03 - drug trade
447110	Gasoline stations with convenience stores	07 - convenience store	05 - grocery | 07 - convenience store
447190	Gasoline stations without convenience stores	07 - convenience store	05 - grocery | 07 - convenience store
452112	Discount department stores	08 - mass merchandiser	01 - wholesale club | 08 - mass merchandiser
452910	Warehouse clubs and supercenters	01 - wholesale club	01 - wholesale club | 08 - mass merchandiser
452990	All other general merchandize stores	08 - mass merchandiser	01 - wholesale club | 08 - mass merchandiser
453991	Tobacco stores	04 - cigarette outlet	04 - cigarette outlet
4451	Grocery stores	05 - grocery | 07 - convenience store	05 - grocery | 07 - convenience store
4452	Grocery stores	05 - grocery | 07 - convenience store	05 - grocery | 07 - convenience store
4453	Beer, wine and liquor stores	02 - liquor trade	02 - liquor trade
4461	Pharmacies and drug stores	03 - drug trade	03 - drug trade
4471	Gasoline stations	05 - grocery | 07 - convenience store	05 - grocery | 07 - convenience store
4529	Other general merchandise stores	01 - wholesale club | 08 - mass merchandiser	01 - wholesale club | 08 - mass merchandiser
4539	Tobacco stores	04 - cigarette outlet	04 - cigarette outlet
4521	Department stores	08 - mass merchandiser	01 - wholesale club | 08 - mass merchandiser

Source: USDA, Economic Research Service calculations; Nielsen TDLinx.

Table A2

Mapping between North American Industry Classification System (NAICS) and InfoScan channels

NAICS	NAICS description	Channel (InfoScan)	Map2 (InfoScan)
445110	Supermarkets and other grocery stores (excluding convenience stores)	Grocery	Grocery ǀ Convenience ǀ Defense commissary
445120	Convenience store	Convenience	Grocery ǀ Convenience ǀ Defense commissary
445210	Meat markets	Grocery	Grocery ǀ Convenience ǀ Defense commissary
445220	Fish and seafood markets	Grocery	Grocery ǀ Convenience ǀ Defense commissary
445230	Fruit and vegetable markets	Grocery	Grocery ǀ Convenience ǀ Defense commissary
445310	Beer, wine and liquor stores	Liquor	Liquor
446110	Pharmacies and drug stores	Drug	Drug
447110	Gasoline stations with convenience stores	Convenience	Grocery ǀ Convenience ǀ Defense commissary
447190	Gasoline stations without convenience stores	Convenience	Grocery ǀ Convenience ǀ Defense commissary
452910	Warehouse clubs and supercenters	Club	Club ǀ Dollar
452990	All other general merchandize stores	Dollar	Club ǀ Dollar
452112	Tobacco stores	Mass merchandiser	Mass merchandiser
4451	Grocery stores	Grocery ǀ Convenience ǀ Defense commissary	Grocery ǀ Convenience ǀ Defense commissary
4452	Grocery stores	Grocery ǀ Convenience ǀ Defense commissary	Grocery ǀ Convenience ǀ Defense commissary
4453	Beer, wine and liquor stores	Liquor	Liquor
4461	Pharmacies and drug stores	Drug	Drug
4471	Gasoline stations	Grocery ǀ Convenience ǀ Defense commissary	Grocery ǀ Convenience ǀ Defense commissary
4529	Other general merchandise stores	Club ǀ Dollar	Club ǀ Dollar
4521	Department stores	Mass merchandiser	Mass merchandiser

Source: USDA, Economic Research Service calculations; IRI InfoScan.

Table A3

Mapping between TDLinx and InfoScan channels to allow comparison to North American Industry Classification System (NAICS) datasets

Map1 (TDLinx)	Map3 (InfoScan)
05 - Grocery | 07 - Convenience store	Grocery | Convenience | Defense commissary | Dollar | Club | Mass merchandiser
02 - Liquor trade	Liquor
03 - Drug trade	Drug
01 - Wholesale club | 08 - Mass merchandiser	Grocery | Convenience | Defense commissary | Dollar | Club | Mass merchandiser

Source: USDA, Economic Research Service calculations; IRI InfoScan and Nielsen TDLinx.

Table A4

Store counts in the 25 largest counties in Retail Marketing Area (RMA 1: TX, LA), ranked by the number of stores in the Economic Census, 2009-12

FIPS code	County	State	EC rank	NETS 2009	EC 2012	CBP 2012	TDLinx 2012	InfoScan 2012
48201	Harris County	TX	1	4033	3687	3631	3239	465
48113	Dallas County	TX	2	2223	1990	1984	1731	306
48439	Tarrant County	TX	3	1633	1500	1490	1397	220
48029	Bexar County	TX	4	1283	1267	1250	1143	221
48453	Travis County	TX	5	784	826	814	691	87
48141	El Paso County	TX	6	599	593	587	550	175
48215	Hidalgo County	TX	7	849	563	563	653	69
48085	Collin County	TX	8	464	519	521	382	78
48157	Fort Bend County	TX	9	408	459	458	290	43
48121	Denton County	TX	10	387	427	420	340	61
48339	Montgomery County	TX	11	344	347	341	337	50
48245	Jefferson County	TX	12	318	318	311	295	28
48355	Nueces County	TX	13	361	297	296	290	39
48061	Cameron County	TX	14	380	294	292	311	40
48027	Bell County	TX	15	257	285	283	260	37
48167	Galveston County	TX	16	297	282	277	276	43
48491	Williamson County	TX	17	239	280	279	230	31
22017	Caddo Parish	LA	18	254	271	270	241	74
48309	McLennan County	TX	19	226	262	259	238	32
48039	Brazoria County	TX	20	245	250	246	241	34
22019	Calcasieu Parish	LA	21	220	243	243	232	39
48423	Smith County	TX	22	194	217	214	183	37
48303	Lubbock County	TX	23	168	211	211	181	29
48479	Webb County	TX	24	217	195	194	177	25
22079	Rapides Parish	LA	25	171	179	178	170	24

Source: USDA, Economic Research Service calculations; IRI InfoScan, Nielsen TDLinx, U.S. Census Bureau Economic Census and County Business Patterns, Walls & Associates National Establishment Time Series.

Table A5

Store counts in the 25 largest counties in Retail Marketing Area 2 (RMA 2: DE, DC, KY, MD, NC, OH, TN, VA, WV), ranked by the number of stores in the Economic Census

FIPS code	County	State	EC Rank	NETS 2009	EC 2012	CBP 2012	TDLinx 2012	InfoScan 2012
37119	Mecklenburg County	NC	1	802	779	762	628	217
37183	Wake County	NC	2	709	739	734	622	185
24033	Prince George's County	MD	3	580	683	668	417	88
24510	Baltimore city	MD	4	725	671	676	394	41
51059	Fairfax County	VA	5	582	649	640	394	97
24005	Baltimore County	MD	6	502	623	617	412	76
24031	Montgomery County	MD	7	556	605	587	316	77
37081	Guilford County	NC	8	496	503	501	443	92
11001	District of Columbia	DC	9	547	447	443	319	30
24003	Anne Arundel County	MD	10	325	426	421	293	58
10003	New Castle County	DE	11	396	400	395	289	54
51810	Virginia Beach city	VA	12	271	365	358	292	78
37067	Forsyth County	NC	13	351	348	345	313	74
51087	Henrico County	VA	14	250	304	295	235	52
37051	Cumberland County	NC	15	251	292	293	298	103
51760	Richmond city	VA	16	245	268	263	211	29
51153	Prince William County	VA	17	224	262	254	211	49
37021	Buncombe County	NC	18	252	256	257	213	42
51041	Chesterfield County	VA	19	197	255	250	207	45
51710	Norfolk city	VA	20	201	249	246	191	46
37063	Durham County	NC	21	229	236	235	221	55
10005	Sussex County	DE	22	201	226	223	154	33
37071	Gaston County	NC	23	232	217	218	203	56
54039	Kanawha County	WV	24	186	213	214	194	42
37129	New Hanover County	NC	25	230	204	199	187	45

Source: USDA, Economic Research Service calculations; IRI InfoScan, Nielsen TDLinx, U.S. Census Bureau Economic Census (EC) and County Business Patterns (CBP, Walls & Associates National Establishment Time Series (NETS).