November 2021

User Documentation: Store Weights for InfoScan Data, 2012–2018

Contract No. GS-00F-354CA, Task 3

Final Report

Prepared for

Matthew MacLachlan
Anne Byrne
U.S. Department of Agriculture
Economic Research Service
355 E Street SW
Washington, DC 20024-3221

Prepared by

Mary K. Muth, RTI
Saki Kinney, RTI
Marissa Trotta, RTI
Charlotte Looby, RTI
Peter Siegel, RTI
RTI International
3040 E. Cornwallis Road
Research Triangle Park, NC 27709

RTI Project Number 0216662.000.003

Contents

Tables

1. Introduction

Store-level weights for the IRI InfoScan data allow researchers to develop projections from the retail stores currently included in the data purchased by the Economic Research Service (ERS) to the population of stores in the United States.[1] IRI prepares the InfoScan datasets from data provided by retail establishments across the United States that have agreed to provide weekly retail sales data (revenue and quantity) for products with Universal Product Codes (UPCs) and random-weight (or perishable) products. The types of stores covered include grocery, drug, convenience, mass merchandiser, club, dollar, and defense commissary stores. IRI provides some of the InfoScan data to ERS at the store level, but in cases where the retailers did not approve release of their data at the store level, IRI provides the data at the retailer marketing area (RMA) level. Unlike the Consumer Network household scanner data purchased by ERS, IRI does not provide weights for stores in the InfoScan data.

The importance of weights was recently highlighted in the National Academies of Sciences, Engineering, and Medicine's (NASEM's) report *A Consumer Food Data System for 2030 and Beyond* (see Recommendation 4.9 in NASEM [2020]). Store-level weights can be used in a wide variety of research projects to calculate sales quantity, sales value, or other estimates that are representative of the population of stores. The store-level weights are useful because they allow analysts to create population estimates of sales quantity and sales value for foods and beverages at the national level or for major metropolitan areas. Without weights, estimates would underrepresent the total quantities or values.

The purpose of this report is to describe the approach and results of developing weights for stores in InfoScan for 2012 through 2018 and provide a user's guide. This report will be updated in the future to include weights for 2019 and 2020 and replicate weights that can be used in variance estimation.

1.1 Overview of Data Used in Developing Weights

As documented in Muth et al. (2016), InfoScan data comprise a nonprobability (convenience) sample of weekly retail sales (revenue and quantity) for UPC and random-weight (or perishables) products for retailers in the following industry channels:

- convenience stores (with scanning capability)
- dollar stores[2]
- drug stores

[1] The data provided to ERS, referred to as the IRI "census component," contain censuses of stores within firms that agree to provide all of their sales data for all of their locations to IRI.
[2] Data for dollar stores in InfoScan are extremely limited after 2016.

- grocery stores (with $2 million or more in annual grocery sales)
- mass merchandisers (including supercenters)
- club stores

Note that for 2017 and 2018 InfoScan data, the data or geographic coverage is insufficient to make inferences about dollar stores, so we did not construct weights for dollar stores in those years. Food and beverage sales[3] represented in InfoScan data comprise sales of branded UPC products, private-label UPC products, and random-weight items. Tables 1-1 and 1-2 summarize the total food and beverage sales volume and number of stores reporting sales in each category for individual stores and RMA stores across years. Comparing Table 1-1 with Table 1-2 indicates that total dollar sales volumes tend to be higher for stores reported at the RMA level even though there are fewer total stores; this difference suggests that RMA stores are larger, on average, than individually reported stores. Because some stores do not report private-label or random-weight sales in InfoScan, we imputed the missing volumes, as described in Section 2.

Table 1-1. InfoScan Store Annual Sales Summary: Individual Stores

Year	Quantity	Club	Mass Merchandisers	Dollar	Drug	Grocery	Convenience
2012	# Stores	—	3,140	8,237	12,381	7,098	9,613
	Sales ($M)	—	$15,862	$2,604	$8,072	$104,419	$5,551
2013	# Stores	—	3,319	8,704	12,443	7,037	9,564
	Sales ($M)	—	$12,365	$2,803	$8,134	$106,495	$5,724
2014	# Stores	—	3,249	9,189	12,523	6,972	10,951
	Sales ($M)	—	$12,219	$3,038	$8,032	$105,111	$6,134
2015	# Stores	—	3,144	9,080	12,478	6,135	13,163
	Sales ($M)	—	$12,637	$3,228	$8,272	$100,870	$7,233
2016	# Stores	214	3,111	9,131	12,375	6,583	12,595
	Sales ($M)	$6,528	$14,285	$3,239	$8,149	$104,900	$7,515
2017	# Stores	215	2,920	574	12,302	7,897	13,968
	Sales ($M)	$6,534	$14,145	$191	$7,884	$132,362	$8,415
2018	# Stores	216	2,687	539	12,304	5,711	13,409
	Sales ($M)	$6,516	$13,763	$184	$6,878	$89,286	$8,871

[3] Throughout this report, food and beverage sales exclude alcoholic beverage sales.

Table 1-2. InfoScan Store Annual Sales Summary: RMA Stores

Year	Quantity	Club	Mass Merchandisers	Dollar	Drug	Grocery	Convenience
2012	# Stores	605	3,897	—	7,386	5,774	—
	Sales ($M)	$21,002	$87,673	—	$4,143	$116,006	—
2013	# Stores	645	4,170	—	7,864	5,752	—
	Sales ($M)	$21,689	$89,558	—	$4,537	$112,808	—
2014	# Stores	659	4,367	—	7,912	5,915	—
	Sales ($M)	$22,515	$94,063	—	$4,649	$123,724	—
2015	# Stores	645	4,452	—	7,848	5,681	—
	Sales ($M)	$23,073	$98,674	—	$4,784	$131,783	—
2016	# Stores	650	4,621	—	7,930	5,935	—
	Sales ($M)	$19,608	$98,779	—	$4,558	$140,176	—
2017	# Stores	650	4,610	—	7,888	6,950	1,672
	Sales ($M)	$25,505	$123,256	—	$4,335	$156,617	$6,015
2018	# Stores	593	3,918	—	7,990	6,812	1,950
	Sales ($M)	$22,884	$127,966	—	$4,254	$157,616	$4,441

As previously mentioned, InfoScan is not a representative sample of stores in the United States. InfoScan stores are heavily skewed toward large chain stores, causing some industry channels to be predominantly represented by only one or two chains. IRI has provided proprietary maps of their store data markets that show that InfoScan data are obtained primarily from major metropolitan areas (Levin et al., 2018). Several states in the north-central area of the country have limited or no coverage, while other states, particularly in the Southwest, Northeast, and upper Midwest, have extensive coverage.

The source of population information used to compute the InfoScan store-level weights was TDLinx. TDLinx is a Nielsen data product that contains up-to-date information on over 400,000 retail stores in the United States (Muth et al., 2019). TDLinx is considered to approximate a census of retail food and beverage stores with $1 million or more in annual sales including grocery, club, convenience, and other types of stores (Levin et al., 2018).[4] The data in TDLinx include estimates of total annual sales, of which food and beverage sales is a component.

We selected TDLinx as the population dataset because TDLinx offers the following benefits: 1) data are published annually, 2) the population covered contains the same types of stores and similar industry channels as InfoScan, 3) over 97% of InfoScan stores can be linked directly to TDLinx, and 4) it includes additional predictor variables useful for imputing food and beverage sales. The primary limitation of TDLinx is that food and beverage sales are not broken out separately from total sales; hence, a high proportion of food and beverage

[4] InfoScan excludes grocery stores with less than $2 million in sales, so the exclusion of small stores with less than $1 million in sales in TDLinx is not a concern for the weighting procedures.

sales must be imputed to compute control totals. However, the available data provide us with high-quality information for the imputations.

Before working with the IRI InfoScan or TDLinx data, researchers may want to familiarize themselves with the contents of prior documentation of the data. In particular, we recommend the following publications:

- Muth, M. K., Sweitzer, M., Brown, D., Capogrossi, K. L., Karns, S. A., Levin, D., Okrent, A., Siegel, P., & Zhen, C. (2016, April). *Understanding IRI household-based and store-based scanner data* (ERS Technical Bulletin 1942). https://www.ers.usda.gov/publications/pub-details/?pubid=47636
- Levin, D., Noriega, D., Dicken, C., Okrent, A., Harding M., & Lovenheim, M. (2018, April). *Examining food store scanner data: A comparison of IRI InfoScan data with other datasets, 2008-2012* (ERS Technical Bulletin 1949). https://www.ers.usda.gov/publications/pub-details/?pubid=90354
- Muth, M. K., Okrent, A., Zhen, C., & Karns, S. A. (2019). *Using scanner data for food policy research* (1st ed.). Elsevier Academic Press. https://www.elsevier.com/books/using-scanner-data-for-food-policy-research/muth/978-0-12-814507-4

Additional details regarding the IRI data can be found on the ERS website "Using Scanner Data" at https://www.ers.usda.gov/topics/food-markets-prices/food-prices-expenditures-costs/using-scanner-data/.

1.2 Organization of this Report

In Section 2, we describe the data preparation and weighting procedures. Appendixes include supplementary tables related to the weighting procedures: variables available for imputation (Appendix A), imputation validation tables (Appendix B), control totals (Appendixes C and D), unequal weighting effects (UWEs) (Appendix E), and a user's guide for the resulting datasets (Appendix F).

2. InfoScan Weighting Procedures and Results

In this section, we describe the procedures used for calculating weights for stores in InfoScan and for validating the results. The process for developing the weights was as follows:

1. Merge InfoScan data with TDLinx using ERS-provided linking files and with public-use American Community Survey tract-level data.
2. Impute missing random-weight and private-label sales for InfoScan stores.
3. Multiply impute food and beverage sales for TDLinx stores not in InfoScan. For each implicate, compute control totals and average totals across implicates.
4. Compute preliminary weights as the ratio of population to sample totals.
5. Create final weights by raking preliminary weights so that both weighted store counts and weighted food and beverage sales sum to population totals with constraints to limit UWEs.[5]

ERS's goal is to use weights to make subnational estimates by IRI industry channel at the finest geographic level possible. The selected geography uses the top 10 metropolitan regions (New York, Los Angeles, Chicago, Houston, Dallas, Miami, Atlanta, Philadelphia, Detroit, and Boston) as determined by the number of food and beverage retail stores in TDLinx, across the years examined (2012 through 2018).[6] We grouped stores that are not in one of these metropolitan regions into their respective Census regions.

After developing the weights, we conducted quality control checks and comparisons across years. In the subsections below, we describe the preparation of the InfoScan data for weighting, calculation of control totals using TDLinx, the weighting procedures, and comparisons of the control totals to Census data. Finally, we discuss the limitations of the weighting procedures given the available data sources.

2.1 Preparation of InfoScan Data

The data preparation tasks involved two steps:

- For some retail chains, IRI provides aggregate sales totals by RMA rather than by store, so we disaggregated the sales estimates into individual stores (see Section 2.1.1).
- For stores that do not report sales for private-label UPCs or random-weight products, we imputed the missing sales values (see Section 2.1.2).

These data preparation tasks were necessary because the weighting procedure requires that each InfoScan store has its own individual value for total food and beverage sales. The final

[5] As discussed in Section 2.3, generalized raking ensures that weighted counts, as well as weighted sales totals, sum to the population (Folsom & Singh, 2000).
[6] Because InfoScan data do not contain sufficient numbers of stores in some geographic locations to implement the weighting procedure, we limited the geographic designations to the top 10 metropolitan areas and the four Census regions.

delivery file of InfoScan store weights includes the imputed sales values and disaggregated RMA sales along with the weights (see Section 3).

Before imputing missing sales values, we created a dataset of characteristics for each InfoScan store using TDLinx and characteristics of the census tract in which the store is located. InfoScan data were linked to TDLinx using ERS-provided linking files. We used census tract IDs available in TDLinx and InfoScan to merge tract characteristics onto the file.

2.1.1 Allocation of Sales to Individual Stores in RMAs

To allocate total RMA food and beverage sales values in InfoScan to individual stores in a retail chain, we used TDLinx data to calculate the proportions of sales for an RMA attributable to each store. Although the sales value in TDLinx represents total sales of all products,[7] we assumed that the within-RMA proportions of sales across all products are an appropriate proxy for the within-RMA proportions of sales for foods and beverages. In comparing chains for which we have both TDLinx sales and InfoScan sales, this assumption appears to be reasonable. After calculating the proportions using TDLinx data, we allocated the total sales estimate for the RMA in InfoScan to each individual store based on the proportions. In rare cases where a value was unavailable in TDLinx, we assumed the store's total sales equaled the average of the other stores in the RMA.

Note that in 2012 the InfoScan data do not indicate which stores within an RMA are active; therefore, we used 2013 InfoScan data to determine which stores to assume were active in 2012. Out of 18,257 RMA stores in 2012, 76 were determined to be inactive (i.e., out of business). InfoScan data for 2013 onward indicate which stores in the RMA are active.

2.1.2 Imputation of Missing Sales Values

Total food and beverage sales are calculated from InfoScan data as the sum of branded, private-label, and random-weight product sales. A handful of stores are missing private-label or random-weight product sales, so these were imputed before computing total food and beverage sales. We first identified which of these stores were likely missing values because of the absence of any private-label or random-weight sales and imputed these values to be zero. In the case of random-weight products, we assumed that convenience stores and drug stores typically do not sell random-weight products but that all other store types do.[8] In addition, we identified a few mass merchandiser chains that typically do not sell random-weight products. The remaining stores missing private-label or random-weight sales had these values stochastically imputed.

[7] For supercenters, TDLinx reports total sales for the "grocery equivalent" rather than the whole store.
[8] We believe it is a reasonable assumption that convenience stores and drug stores do not sell random-weight products (with a few exceptions) because doing so would require that each store have a label-printing scale system.

The InfoScan database together with variables from TDLinx and census tract characteristics provided a rich source of data to impute the missing sales data. Appendix Table A-1 lists the variables used to impute missing sales data, the source of data for each variable, and a description. Imputations were created using PROC MI (SAS Institute, 2015). The imputation procedure used is referred to as full conditional specification, chained equations, or originally as sequential regression (Raghunathan et al., 2001). This procedure cycles through all variables in the dataset with missing values, imputing each variable conditional on all others. This process repeats a few times until the change between iterations is minimal. We conducted imputations by year separately for the store file and RMA file and by channel for the store file. For 2013 on, private-label sales were missing for all RMA drug stores, so we imputed them using individual store data.

We compared the distributions for private-label and random-weight sales before and after imputation to ensure that the imputation did not substantially alter the distribution of values. Appendix Table B-1 provides the imputation rates and compares observed and imputed means by industry channel and year. Given the low imputation rate in most cases, it is unsurprising that these means match closely. Only drug stores for 2013 and on have a high imputation rate; however, the pre- and post-imputation means are similar.

2.2 Calculation of Control Totals Using TDLinx Data

As discussed above, we used TDLinx to define the population of stores for which the InfoScan data were weighted to represent. We kept all TDLinx stores in the ERS data extract, with the exception of liquor stores, defense commissaries, and stores in Puerto Rico, and merged these with census tract–level data using tract IDs and with the InfoScan data using ERS-supplied ID links.

For non-IRI stores, we mapped TDLinx industry channels to IRI channels using the assignments in Table 2-1. This channel mapping is based on the observed correspondence of IRI channels and TDLinx channels among stores linked across both datasets. About 99% of stores followed these assignments in each year. Nearly all the classification discrepancies were between grocery stores and mass merchandisers.

Table 2-1. TDLinx and IRI InfoScan Industry Channel Mapping

TDLinx Channel	TDLinx Subchannel	IRI Channel
01—Wholesale club	1—Conventional club	Club store
03—Drug	1—Rx only and small independent	Drug
03—Drug	3—Conventional drug	Drug
05—Grocery	1—Supermarket—Limited assortment	Grocery
05—Grocery	2—Supermarket—Natural/gourmet	Grocery
05—Grocery	3—Warehouse grocery	Grocery

(continued)

Table 2-2. TDLinx and IRI InfoScan Industry Channel Mapping (continued)

TDLinx Channel	TDLinx Subchannel	IRI Channel
05—Grocery	4—Superette	Grocery
05—Grocery	5—Supermarket—Conventional	Grocery
05—Grocery	6—Supercenter	Mass merchandiser
07—Convenience store	7—Conventional convenience	Convenience
08—Mass merchandiser	3—Dollar store	Dollar
08—Mass merchandiser	4—General merchandise	Dollar
08—Mass merchandiser	8—Conventional mass merchandise	Mass merchandiser

We computed annual total sales by multiplying the weekly total sales volume reported in TDLinx by 52. Between 900 and 2,100 stores in each year had values of total sales in TDLinx that were lower than the reported InfoScan food and beverage sales, which should always be less than or equal to total sales. For the purposes of imputing food and beverage sales that are less than or equal to total sales, we deleted the total sales values for these stores from the merged file and imputed them along with InfoScan sales.

After linking to InfoScan, food and beverage sales were available for about 20% of stores in TDLinx each year (corresponding to roughly 58% of estimated population food and beverage sales); hence, about 80% of store food and beverage sales had to be imputed. This number does not reflect that we calculated some food and beverage sales in InfoScan from imputed private-label or random-weight sales as described in Section 2.1. The distribution of industry channels in InfoScan differs substantially from TDLinx in that InfoScan predominantly contains larger stores. Consequently, the missing data rate is lower for club stores, drug stores, and mass merchandisers, while the missing data rate is much higher for convenience, grocery, and dollar stores. Appendix Table B-2 provides the proportions of TDLinx stores in InfoScan by channel and year.

We imputed food and beverage sales using multiple imputation, conducted using similar procedures to the single imputation procedures described in Section 2.1.2. We then used these imputed sales to generate population totals. Multiple imputation provides an advantage over single imputation in that averaging over many imputations allows us to minimize the random error due to imputation. We obtained control totals by calculating total food and beverage sales by channel and geographic area within each of 100 imputations and then averaging the totals across imputations.

Appendix Table A-2 lists the variables used as predictors. Food and beverage sales were missing for all non-InfoScan stores, while the remaining variables had few, if any, missing values. We specified predictive mean matching as the modeling approach in PROC MI for food and beverage sales, and we used PROC MI defaults for other variables with missingness. We conducted imputation separately for each channel and year.

Because RMA stores do not have sales reported at the store level in InfoScan, it is possible that the disaggregated store-level sales (described in Section 2.1.1) might not accurately reflect store-level relationships. Thus, when possible, we did not use RMA stores for imputation; however, in most cases there was no sensible alternative. For example, before 2016, all club stores were reported only at the RMA level, and beginning in 2016, the individually reported stores comprised only a fourth of InfoScan club stores. In other cases, the RMA stores were substantially different from the individually reported stores; thus, leaving them out would omit important information. After imputation, we returned the RMA stores not used in imputation to the population dataset. The RMA-level sales values are as accurate at the RMA level as the individual-store sales are at the store level; therefore, the disaggregation of RMA-level sales affects the control totals only when RMAs span across geographic areas (i.e., the disaggregation).[9]

Table 2-2 summarizes the imputed totals by channel and metro region across years. Appendix C provides the complete set of control totals resulting from the imputation process for each year.

Table 2-3. Imputed Food and Beverage Sales Totals by Channel, Metro Region, and Year ($K)

Channel/Metro Region	2012	2013	2014	2015	2016	2017	2018
Total	613,339	630,605	646,030	681,411	693,122	791,234	754,314
Channel							
Club store/mass merchandiser	159,166	158,232	167,988	180,423	177,357	213,999	230,359
Dollar store	8,496	9,927	10,513	11,447	12,047	11,795	13,122
Drug store	14,987	15,394	15,032	15,655	15,151	14,711	12,823
Grocery store	353,394	364,428	372,431	401,075	403,115	451,134	411,005
Convenience store	77,297	82,624	80,066	72,811	85,452	99,595	87,005
Metro Region							
New York	35,438	37,712	37,418	34,986	39,624	40,632	39,451
Los Angeles	22,612	23,317	23,770	25,992	26,195	30,021	27,470
Chicago	17,350	17,888	18,419	20,065	19,293	24,970	19,483
Houston	11,800	12,508	12,711	13,898	14,272	15,381	14,283
Dallas	12,084	12,450	12,708	14,353	14,764	18,170	16,381
Miami	12,571	12,442	11,044	12,736	14,233	15,046	14,347
Atlanta	11,091	10,353	10,801	12,444	12,094	13,291	13,679
Philadelphia	10,665	11,913	13,150	12,861	13,959	15,945	16,310
Detroit	8,433	8,776	9,244	9,383	9,207	9,550	9,983
Boston	10,666	11,055	11,461	11,261	12,060	13,544	12,089
Region 1 Northeast	54,983	56,803	58,016	58,896	60,776	67,703	66,245

(continued)

[9] Approximately half of the RMAs are almost entirely contained in one geographic region. The other half of RMAs have stores in more than one region but generally fewer than a hundred stores.

Table 2-4. Imputed Food and Beverage Sales Totals by Channel, Metro Region, and Year ($K) (continued)

Channel/Metro Region	2012	2013	2014	2015	2016	2017	2018
Region 2 Midwest	104,098	106,803	108,187	116,606	115,391	125,462	124,844
Region 3 South	186,614	189,534	193,956	202,295	205,402	232,786	230,900
Region 4 West	114,935	119,051	125,147	135,637	135,852	168,734	148,848

We evaluated the quality of the imputations using comparisons of pre- and post-imputation means for the population as well as a "validation sample," shown in Appendix Tables B-2 and B-3. Because InfoScan is dominated by larger stores while smaller stores are underrepresented, we expect the mean for the population to be lower than that for InfoScan, although we do not know for certain how much lower it should be. Hence, we supplemented this comparison of food and beverage sales means by comparisons of the model predictors, in particular, the total sales reported in TDLinx, as well as a validation sample comparison.

Total sales from TDLinx are a key predictor of food and beverage sales, so comparing TDLinx sales for InfoScan and non-InfoScan stores, combined with the percentage of InfoScan stores in TDLinx, provides an indication of the extent to which the imputation models are extrapolating and whether we should expect the pre- and post-imputation distributions to differ. The rightmost columns of Table B-2 compare the means of TDLinx sales for InfoScan and non-InfoScan stores. Sales are seen to be substantially higher in most cases for InfoScan stores, suggesting that food and beverage sales should also be higher for InfoScan stores.

We based the validation sample in each on duplicated records for which the InfoScan sales are known, but with the InfoScan sales deleted so that they could be imputed. We added the validation sample to the dataset before imputation and removed it from the imputed population after imputation. The validation sample pre- vs. post-imputation comparisons, summarized in Table B-3, suggests the imputation models are generally doing a good job of imputing food and beverage sales for non-InfoScan stores that are similar to InfoScan stores. Because the pre-imputation means for the validation sample are the known values of the means, we expect them to match the post-imputation means more closely than in Table B-2.

2.3 Weight Factor Construction

The goal for weight factor construction is to produce weights such that weighted food and beverage sales estimates match the control totals derived in Section 2.2.2 and shown in Appendix C and such that the weighted store counts match the control totals in Appendix D.

The control totals for store counts come directly from TDLinx.[10] We constructed preliminary weights by dividing the imputed population food and beverage sales control totals by the InfoScan sales totals for each industry channel and geography subgroup, which are easily computed from the InfoScan data. By construction, these factors produce weighted sales totals that match the population totals for food and beverage sales; however, sums of the weights do not match the population counts of eligible stores. To ensure that weighted counts, as well as weighted sales totals, sum to the population, we created weight adjustments using generalized raking, as proposed in Folsom and Singh (2000), implemented with SUDAAN procedure PROC WTADJUST (RTI International, 2012).

Raking, or iterative proportional fitting, is commonly used to adjust sampling weights when sample characteristics are known to be different from the population and to correct for nonresponse bias and adjust nonprobability samples. The procedure involves repeatedly estimating the adjustments until they converge to values satisfying the constraints imposed by the control totals. Bounds are placed on the weight adjustments to limit UWEs that can yield high standard errors.

We applied a raking model to the food and beverage sales totals and store counts for each channel and geographic area. Because of the small number of club stores, we combined those stores with mass merchandisers. The tables in Appendix C provide weighted sample estimates alongside the control totals for food and beverage sales and control totals for store counts. These illustrate that the raking process was successful in achieving its goals. In some cases, the totals do not match because some regions needed to be combined within a channel to allow the raking model to converge with moderate UWEs. For example, the Los Angeles metro has few InfoScan dollar stores, so for each year Los Angeles was merged with its Census region, Region 4 West, before raking. Similar issues occurred for grocery stores requiring metros to be merged with Census regions during raking, including Detroit with Region 2 Midwest (2012–2013), Miami with Region 3 South (2015), and Los Angeles with Region 4 West (2016).

2.4 Comparing Control Totals with Bureau of the Census Data

This section compares the imputed sales control totals and TDLinx store count totals with publicly available Census of Retail Trade (CRT) information for 2012 and 2017. We obtained food and beverage sales totals at the national level by North American Industry Classification System (NAICS) code from the Bureau of the Census website. The product line totals are less accurate than the store sales totals because the underlying data for product lines are sparse; however, the published totals are still considered to be high-quality estimates because the Bureau of the Census has good-quality information and expertise to

[10] For the stores not matched to TDLinx, we assumed that these exist in the population but are missing TDLinx IDs. Therefore, we did not add these stores to the control totals computed from TDLinx. The sample totals were constructed using the entire set of eligible InfoScan stores in the weighting file, which includes stores not matched to TDLinx.

produce them. Totals are not available by metro region; however, at the Census division level, food and beverage sales totals (product line sales for categories 20100 and 21100) are available for grocery stores (NAICS 44511) only. Variation in industry category definitions across products makes it challenging to align our totals with the Census totals; however, overall, our totals appear to be in the right ballpark.

We can assess the extent to which we should expect the control totals to align with CRT totals by comparing the distributions of store counts and total sales by industry in each data source. Table 2-3 shows fewer stores and higher total sales in the 2012 CRT compared with TDLinx. One factor in the discrepancy is the NAICS classification of mass merchandisers. Many are classified as department stores; however, department stores are outside the universe of interest. In addition, there appears to be inconsistency in what constitutes a gas station convenience store. There is a separate category for gas station convenience stores in TDLinx that is excluded from the data we have, yet the store counts of convenience stores in TDLinx in contrast with Census suggest that many stores with NAICS code 44711 (gas stations with convenience stores) are in fact included in TDLinx.

Although the 2012 total store count is higher for TDLinx than CRT (Table 2-3), the imputed total food and beverage sales are lower compared with CRT (Table 2-4). Because of industry coding differences, Table 2-4 groups channels to make the best possible comparison. Despite these differences, we see that in aggregate the totals are within a few percentage points of each other, and overall, the total food and beverage sales imputed for TDLinx seem reasonable. The imputed 2017 control totals compared with the 2017 CRT (Table 2-5) show similar results for the same industry groups, while the grand totals match quite closely. A fuller examination of TDLinx versus CRT industry coding and product line coding is needed to better assess the accuracy.

Table 2-5. TDLinx and Census of Retail Trade Total Sales and Store Count Comparison, 2012

		Sales ($K)		Store Counts	
IRI Channel	**NAICS**	**TDLinx**	**CRT Total**	**TDLinx**	**CRT**
Grocery, convenience	44511,[a] 44521, 44522, 44523, 44512, 44711, 44719	793,647,869	1,066,490,550	194,348	184,178
Drug, dollar, mass merchandiser, club store	44611, 45291, 45299, 45211	579,081,672	804,530,052	74,500	73,758
Total		1,372,729,539	1,871,020,602	268,848	257,936

[a] We restricted grocery store counts and total sales to stores with at least $1M in sales.
Note: CRT store counts and total sales are only for stores that reported product line sales for product line 20100. Food and beverage sales are totals for product lines 20100 and 21100 combined.

Sources: U.S. Bureau of the Census and TDLinx

Table 2-6. Food and Beverage Sales by Channel, Census of Retail Trade vs. Control Totals, 2012

		Census of Retail Trade		Imputed	
IRI Channel	**NAICS**	**Food and Beverage Sales ($K)**	**% of Total Sales**	**Control Total ($K)**	**% of Total Sales**
Grocery, convenience	44511,[a] 44521, 44522, 44523, 44512, 44711, 44719	453,143,793	71.1	430,690,929	70.2
Drug, dollar, mass merchandiser, club store	44611, 45291, 45299, 45211,	184,217,209	28.9	182,648,353	29.8
Total		637,361,002	100.0	613,339,282	100.0

[a] We restricted grocery store counts and total sales to stores with at least $1M in sales.

Note: Census store counts and total sales are only for stores that reported product line sales for product line 20100. Food and beverage sales are totals for product lines 20100 and 21100 combined.

Source: U.S. Bureau of the Census and imputed TDLinx and InfoScan data

Table 2-7. Food and Beverage Sales by Channel, Census of Retail Trade vs. Control Totals, 2017

		Census of Retail Trade		Imputed	
IRI Channel	**NAICS**	**Food and Beverage Sales ($K)**	**% of Total Sales**	**Control Total ($K)**	**% of Total Sales**
Grocery, convenience	445120, 445220, 445230, 447110, 445210, 557190, 445110	570,839,626	72.1	550,728,686	69.6
Drug, dollar, mass merchandiser, club store	452311, 446110, 452319, 452210	221,121,514	27.9	240,505,550	30.4
Total		791,961,140	100.0	791,234,236	100.0

Note: CRT food and beverage sales include North American Product Classification (NACPS) collection codes 2001575000, 2001450000, 2001425000, 2001400000, 7000025000, 5000125000, 5000020000, 5000100000, 5000225000, 5000050000, 5000075000, 5000025000, 5000175000, 5000150000, 5000250000.

Source: U.S. Bureau of the Census and imputed TDLinx and InfoScan data

Although public-use CRT data are not available by Census division for all NAICS codes, they are available for grocery stores, so Table 2-6 compares 2012 grocery store totals by Census division for the CRT with the imputed totals by Census division. Although the dollar amounts differ overall and by division, the percentage distributions are quite close.

Table 2-8. Grocery Store Food and Beverage Sales by Census Division: Census of Retail Trade vs. Control Totals, 2012

Census Division	Census of Retail Trade		Imputed TDLinx	
	Food and Beverage Sales ($K)	% of Total Sales	Control Total ($K)	% of Total Sales
New England	26,662,378	6.8	24,311,539	6.9
Middle Atlantic	60,211,000	15.4	53,148,644	15.0
East North Central	50,332,387	12.9	47,260,482	13.4
West North Central	23,943,596	6.1	20,242,901	5.7
South Atlantic	77,455,630	19.8	73,269,818	20.7
East South Central	17,307,360	4.4	16,302,474	4.6
West South Central	38,493,620	9.9	31,560,634	8.9
Mountain	24,259,383	6.2	23,904,249	6.7
Pacific	71,963,336	18.4	63,392,989	17.9
Total	390,628,690	100.0	353,393,731	100.0

Sources: U.S. Bureau of the Census and imputed TDLinx and InfoScan data.

2.5 Discussion of Limitations of Weighting Procedures

Using the weights constructed provides an advantage over unweighted analyses in that the results will be more representative of the population. Weighted variance estimates should also be used. These can be computed using any statistical software that handles weights (see Section 3 for example programming code). These weights are appropriate for analyses of the subgroups defined by the variables used in the weighting procedure, namely, the metro/region geographic areas and industry channel.[11] Use of the weights in analyses of smaller areas may be misleading. This section describes additional sources of uncertainty that we have not accounted for.

The dearth of small stores (i.e., those with less than $2 million in annual sales) in InfoScan means that for these segments the imputed control totals could be biased because the imputation models may be missing important information. For example, there are small store subchannels in TDLinx with few if any InfoScan stores, including limited assortment, warehouse, superette, and natural/gourmet grocery stores. If food and beverage sales for these stores differ meaningfully from other grocery stores in their relationship with total sales and other variables, this will not be reflected in the control totals or weighted estimates. However, if these stores represent only a small portion of the sales in the domain being analyzed (which we believe to be likely), their absence may be unimportant.

The use of multiple imputation for the control totals minimizes random errors due to imputing values but does not eliminate the possibility of bias due to the coverage errors known to exist in InfoScan and TDLinx. If food and beverage sales are *missing at random,*

[11] The primary purpose of the InfoScan weights is to calculate population estimates. As argued in Solon et al. (2015), the weights may not be necessary when estimating regression models.

that is, noninclusion in InfoScan can be explained by the available predictor variables, then the imputed control totals are considered to be unbiased estimates of the true population values. Stores with lower food and beverage sales are less likely to be included in InfoScan, which means that food and beverage sales could be considered to be not missing at random; however, this is ameliorated by the high correlation of food and beverage sales with total sales and other available predictors. Nonetheless, given the severe overrepresentation of large stores and urban areas in InfoScan, it is possible that some bias remains.

The overrepresentation of large stores in InfoScan also increases the UWE (i.e., the variance inflation due to unequal weights). The calibration process yielded variable weights with an overall UWE in each year between 3.8 and 4.8. Ideally, the UWE would be 1, meaning all the weights are the same as in a simple random sample, but for a well-designed complex sample survey, we would expect it to be closer to 2 or 3. Given the heavy overrepresentation of large stores in the sample, the larger UWE may be appropriate. Achieving a lower UWE requires a trade-off with bias; that is, we would need to reduce the number of control totals.

As seen in Appendix E, for many industry–region subgroups, the UWE is, in fact, close to 1, while in others it is quite large, even above 40. This result is due to a combination of factors, including possible inaccuracy in the control totals, but primarily discrepancy between the proportion of population stores represented by the InfoScan sample and the proportion of total population sales represented by the sample. For example, in 2012, the IRI data contain 23% of the population of Los Angeles grocery stores, which have a UWE of 39.29, but because of the overrepresentation of large stores, these represent about 49% of Los Angeles grocery store food and beverage sales. In contrast, the 2012 IRI data contain about 4% of Boston convenience stores, which have a UWE of 1.02, and these represent about 4% of Boston convenience store food and beverage sales.

Weights for probability samples are usually interpreted as the number of units in the population represented by a given sample unit. This interpretation does not hold for these weight factors because, across years, close to 30% of them are below 1. That is, while the sum of weights in a given industry and geographic region represents the total number of stores in that industry and geographic region, an individual sample store's weight does not represent the number of stores in the population represented by that sample store.

References

Briesch, R. A., Chintagunta, P. K., & Fox, E. J. (2009). How does assortment affect grocery store choice? *Journal of Marketing Research 46*, 176–189.

Carlson, A. E., Page, E. T., Zimmerman, T. P., Tornow, C. E., & Hermansen, S. (2019, March). *Linking USDA nutrition databases to IRI household-based and store-based scanner data, TB-1952*. U.S. Department of Agriculture, Economic Research Service. https://www.ers.usda.gov/webdocs/publications/92571/tb-1952.pdf?v=3582.9

Folsom, R., & Singh, A. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the Section on Survey Research Methods*, 598–603.

Levin, D., Noriega, D., Dicken, C., Okrent, A., M., Harding, M., & Lovenheim, M. (2018, October). *Examining food store scanner data: A comparison of IRI InfoScan data with other data sets, 2008-2012*. (ERS Technical Bulletin 1949). U.S. Department of Agriculture, Economic Research Service.

Muth, M. K., Okrent, A., Zhen, C., & Karns, S. A. (2019). *Using scanner data for food policy research*. (1st ed.). Elsevier Academic Press.

Muth, M. K., Sweitzer, M., Brown, D., Capogrossi, K. L., Karns, S. A., Levin, D., Okrent, A., Siegel, P., & Zhen, C. (2016, April). *Understanding IRI household-based and store-based scanner data* (ERS Technical Bulletin 1942). https://www.ers.usda.gov/publications/pub-details/?pubid=47636

National Academies of Sciences, Engineering, and Medicine. (2020). *A consumer food data system for 2030 and beyond*. The National Academies Press. https://doi.org/10.17226/25657

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85–95.

RTI International. (2012). *SUDAAN user's manual*. Release 11.0. RTI International.

SAS Institute Inc. (2015). *SAS/STAT® 14.1 user's guide*. SAS Institute Inc.

Taylor, R., & Villas-Boas, S. B. (2016). Food store choices of poor households: A discrete choice analysis of the National Household Food Acquisition and Purchase Survey (FoodAPS). *American Journal of Agricultural Economics, 98*(2), 513–32.

Zhen, C., Muth, M., Okrent, A., Karns, S., Brown, D., & Siegel, P. (2019). Do differences in reported expenditures between household scanner data and expenditure surveys matter in health policy research? Health Economics, 28(6), 782–800.

Appendix A:
Variables Available for Imputation

This appendix provides the list of variables used for imputation as described in Section 2. Table A-1 lists the variables used for imputing missing sales volumes for stores in InfoScan, and Table A-2 lists the variables used for imputing control totals.

Table A-1. Variables Used for Imputing Random-Weight and Private-Label Sales for Stores in InfoScan

Variable Name	Source	Description
Brand_Sales	Derived	Total sales of national branded products
chain	TDLinx	Indicates store is part of a chain or an independent
ChannelID	IRI	Industry channel
Med_Inc_yy	Census	Census tract median household income (in inflation-adjusted $) for current year
metroregion	IRI	Metropolitan region
Pop_10_Inside_Urb	Census	Percentage of census tract population inside urbanized areas
Pop_10_Urb	Census	Percentage of census tract population in urban areas
Pop_10_Urb_Clst	Census	Percentage of census tract population inside urban clusters
Pop_yy	Census	Census tract total population for current year
sftemploy	TDLinx	Total of full-time employee and part-time employee equivalents
snmchkout	TDLinx	Number of checkout registers in the store
ssqft	TDLinx	Selling square footage of the store in thousands of square feet
swklyvol	TDLinx	Estimated average weekly all commodity volume of the store ($K)

Table A-2. Variables Used for Imputing Control Totals from TDLinx

Variable Name	Source	Description
chain	TDLinx	Indicates store is part of a chain or an independent
ChannelID	InfoScan	Industry channel; derived for non-InfoScan stores
Med_Inc_yy	Census	Census tract median household income (in inflation-adjusted $) for current year
metroregion	TDLinx	Metropolitan region
Pop_10_Inside_Urb	Census	Percentage of census tract population inside urbanized areas
Pop_10_Rural	Census	Percentage of census tract population in rural areas
Pop_10_Urb	Census	Percentage of census tract population in urban areas
Pop_10_Urb_Clst	Census	Percentage of census tract population inside urban clusters
Pop_yy	Census	Census tract total population, current year

(continued)

Table A-2. Variables Used for Imputing Control Totals from TDLinx (continued)

Variable Name	Source	Description
sbeer	TDLinx	Indicates the store sells beer
sftemploy	TDLinx	Total of full-time employee and part-time employee equivalents
slat	TDLinx	Store latitude
sliquor	TDLinx	Indicates the store sells liquor
slong	TDLinx	Store longitude
snmchkout	TDLinx	Number of checkout registers in the store
ssqft	TDLinx	Selling square footage of the store in thousands of square feet
subchannel	TDLinx	Industry subcategory
swine	TDLinx	Indicates the store sells wine
swklyvol	TDLinx	Estimated average weekly all commodity volume of the store ($K)

Appendix B:
Imputation Tables

This appendix contains tables evaluating the quality of the imputation models used to impute random-weight sales and private-label sales for stores in InfoScan (Section 2.1) and to impute food and beverage sales for stores not in InfoScan (Section 2.2).

Table B-1. Pre- vs. Post-imputation Comparison for Random-Weight and Private-Label Sales ($K), by Channel and Year

		Random-Weight Sales					Private-Label Sales				
			Observed		Imputed			Observed		Imputed	
Year	Channel	Imputation Rate	N	Mean Sales ($K)	N	Mean Sales ($K)	Imputation Rate	N	Mean Sales ($K)	N	Mean Sales ($K)
2012	Club store	0.0%	—	—	—	—	15.2%	513	2.0	605	3.8
	Mass merchandiser	0.6%	6,993	2,033.5	7,037	2,029.0	0.3%	7,019	544.1	7,037	542.7
	Dollar store	0.0%	8,236	0.2	8,237	0.2	0.0%	—	—	—	—
	Drug store	0.0%	—	—	—	—	0.0%	19,766	40.2	19,767	40.2
	Grocery store	3.0%	12,492	4,587.5	12,872	4,581.3	0.0%	—	—	—	—
	Convenience store	0.0%	—	—	—	—	0.0%	—	—	—	—
2013	Club store	0.0%	—	—	—	—	0.0%	—	—	—	—
	Mass merchandiser	0.3%	7,450	2,047.5	7,475	2,040.7	5.2%	7,083	35.0	7,475	39.5
	Dollar store	0.1%	7,489	0.1	7,494	0.1	0.0%	—	—	—	—
	Drug store	0.0%	—	—	—	—	39.1%	12,331	73.6	20,233	72.3
	Grocery store	4.0%	12,233	4,790.8	12,739	4,761.7	0.4%	12,683	2,432.7	12,739	2,433.9
	Convenience store	0.0%	—	—	—	—	10.0%	8,543	20.1	9,495	19.3
2014	Club store	0.0%	—	—	—	—	0.0%	—	—	—	—
	Mass merchandiser	1.8%	7,476	2,181.0	7,614	2,143.5	0.1%	7,608	33.6	7,614	33.6
	Dollar store	0.0%	9,039	0.1	9,041	0.1	0.0%	—	—	—	—
	Drug store	0.0%	—	—	—	—	38.7%	12,522	68.4	20,434	67.9
	Grocery store	5.0%	12,249	4,992.0	12,887	4,906.9	2.7%	12,540	2,533.6	12,887	2,531.2
	Convenience store	0.0%	—	—	—	—	18.8%	8,889	23.1	10,951	20.3
2015	Club store	0.0%	—	—	—	—	0.0%	645	42.1	645	42.1
	Mass merchandiser	0.3%	7,570	2,324.6	7,595	2,317.0	0.0%	7,592	33.4	7,595	33.4
	Dollar store	0.1%	8,545	0.0	8,554	0.0	0.0%	—	—	—	—
	Drug store	0.0%	—	—	—	—	38.6%	12,478	63.1	20,326	63.2
	Grocery store	5.3%	11,195	5,496.7	11,816	5,434.7	0.0%	—	—	—	—
	Convenience store	0.0%	—	—	—	—	12.3%	11,549	25.2	13,163	22.7

(continued)

Table B-1. Pre- vs. Post-imputation Comparison for Random-Weight and Private-Label Sales ($K), by Channel and Year (continued)

Year	Channel	Random-Weight Sales					Private-Label Sales				
			Observed		Imputed			Observed		Imputed	
		Imputation Rate	N	Mean Sales ($K)	N	Mean Sales ($K)	Imputation Rate	N	Mean Sales ($K)	N	Mean Sales ($K)
2016	Club store	0.0%	—	—	—	—	0.0%	—	—	—	—
	Mass merchandiser	0.4%	7,666	2,332.3	7,700	2,322.2	0.0%	—	—	—	—
	Dollar store	0.1%	8,246	0.0	8,252	0.0	0.0%	—	—	—	—
	Drug store	0.0%	—	—	—	—	39.1%	12,375	57.6	20,305	56.2
	Grocery store	5.2%	11,864	5,447.4	12,518	5,390.8	0.0%	—	—	—	—
	Convenience store	0.0%	—	—	—	—	12.6%	11,007	30.5	12,593	27.6
2017	Club store	0.0%	—	—	—	—	0.0%	—	—	—	—
	Mass merchandiser	0.4%	7,496	2,473.5	7,529	2,474.2	0.0%	—	—	—	—
	Dollar store	—	—	—	—	—	—	—	—	—	—
	Drug store	0.0%	—	—	—	—	39.1%	12,302	50.1	20,190	48.6
	Grocery store	15.4%	12,567	5,508.4	14,846	5,374.8	0.0%	—	—	—	—
	Convenience store	0.0%	15,620	0.0	15,620	0.0	11.8%	13,775	83.1	15,620	75.1
2018	Club store	0.4%	806	9,151.5	809	9,128.6	0.0%	—	—	—	—
	Mass merchandiser	0.5%	6,568	2,980.0	6,603	2,964.4	0.0%	—	—	—	—
	Dollar store	—	—	—	—	—	—	—	—	—	—
	Drug store	0.0%	—	—	—	—	39.4%	12,303	40.5	20,294	40.2
	Grocery store	6.7%	11,681	5,577.2	12,523	5,510.6	0.0%	—	—	—	—
	Convenience store	0.0%	—	—	—	—	0.9%	15,190	66.2	15,333	68.0

Table B-2. Pre- vs. Post-imputation Comparison of Food and Beverage Sales ($K), by Channel and Year

Year	Channel	% TDLinx Stores in InfoScan	InfoScan Food & Beverage Sales ($K)		Imputed Population Food & Beverage Sales ($K) (all imputations)		TDLinx Total Sales InfoScan Stores ($K)		TDLinx Total Sales Non-InfoScan Stores ($K)	
			N	Mean Sales ($K)	N per Imputation	Mean Sales ($K)	N	Mean	N	Mean
2012	Club store	48.3%	597	34,733.0	1,236	39,000.0	597	80,036.0	639	110,489.8
	Mass merchandiser	87.3%	6,944	14,785.0	7,952	13,953.0	6,940	32,539.1	1,012	16,516.9
	Dollar store	29.3%	7,436	325.0	25,361	335.0	7,436	1,330.5	17,925	1,650.9
	Drug store	48.8%	19,492	621.0	39,941	375.0	12,230	7,361.4	20,455	1,592.1
	Grocery store	27.8%	12,641	17,485.0	45,493	7,768.0	6,310	20,590.5	33,438	5,647.3
	Convenience store	6.3%	9,348	589.0	148,260	521.0	9,339	4,105.6	138,921	2,172.5
	Total	21.0%	56,458	6,455.4	268,243	2,286.2	42,852	12,643.4	212,390	3,013.9
2013	Club store	48.1%	608	33,901.8	1,265	40,969.3	609	80,211.5	656	113,284.3
	Mass merchandiser	89.1%	7,214	13,712.1	8,100	13,133.0	7,245	32,440.0	854	18,192.4
	Dollar store	25.1%	6,735	325.4	26,811	370.4	7,311	1,397.0	19499	1,647.3
	Drug store	47.6%	19,643	629.4	41,298	372.8	19,698	7,680.1	21623	1,566.1
	Grocery store	27.2%	12,495	17,239.6	45,954	7,930.8	11,700	22,911.1	33419	5,704.3
	Convenience store	6.2%	9,198	609.1	149,139	554.0	9,221	4,277.3	139905	2,249.0
	Total	20.5%	55,893	6,353.2	272,567	2,313.6	55,784	13,496.2	215,956	3,061.3
2014	Club store	46.6%	604	34,371.2	1,297	43,757.3	621	80,306.9	676	122,090.4
	Mass merchandiser	86.9%	7,145	13,981.8	8,218	13,535.5	7,340	32,317.7	878	19,420.9
	Dollar store	28.5%	8,089	335.1	28,345	370.9	8,756	1,390.8	19588	1,714.7
	Drug store	46.9%	19,450	622.8	41,482	362.4	19,956	7,668.2	21738	1,526.5
	Grocery store	26.7%	12,315	17,776.9	46,172	8,066.2	11,816	23,269.4	33676	5,862.3
	Convenience store	6.2%	9,377	615.3	150,974	530.3	9,601	4,636.9	141363	2,481.8
	Total	20.6%	56,980	6,321.1	276,488	2,336.6	58,090	13,285.6	217,919	3,279.2

(continued)

Table B-2. Pre- vs. Post-imputation Comparison of Food and Beverage Sales ($K), by Channel and Year (continued)

Year	Channel	% TDLinx Stores in InfoScan	InfoScan Food & Beverage Sales ($K)		Imputed Population Food & Beverage Sales ($K) (all imputations)		TDLinx Total Sales InfoScan Stores ($K)		TDLinx Total Sales Non-InfoScan Stores ($K)	
			N	Mean Sales ($K)	N per Imputation	Mean Sales ($K)	N	Mean	N	Mean
2015	Club store	47.6%	633	35,859.5	1,330	46,877.5	635	80,241.7	695	125,600.6
	Mass merchandiser	91.3%	7,387	14,780.8	8,095	14,586.3	7,414	32,390.0	680	24,839.9
	Dollar store	26.9%	7,893	362.3	29,388	389.5	8,345	1,399.9	21042	1,721.0
	Drug store	47.8%	20,050	644.8	41,970	373.0	20,111	7,670.2	21888	1,475.4
	Grocery store	25.2%	11,591	19,853.5	46,046	8,710.3	9,611	24,670.3	34419	6,233.5
	Convenience store	7.7%	11,807	591.7	153,255	475.1	11,853	4,994.7	141390	2,610.4
	Total	21.2%	59,361	6,482.1	280,084	2,432.9	57,969	12,995.5	220,114	3,436.1
2016	Club store	62.3%	844	30,380.5	1,355	41,159.5	849	82,367.0	506	154,540.7
	Mass merchandiser	90.8%	7,463	14,864.1	8,222	14,787.8	7,500	32,669.3	721	25,271.3
	Dollar store	24.4%	7,544	359.5	30,911	389.7	7,989	1,430.2	22922	1,715.7
	Drug store	45.8%	20,075	627.7	43,801	345.9	20,126	7,810.0	23700	1,444.5
	Grocery store	26.6%	12,265	19,668.7	46,195	8,726.4	11,012	24,324.1	33892	6,084.2
	Convenience store	7.3%	11,235	647.4	154,092	554.6	11,267	4,664.8	142795	2,709.6
	Total	20.9%	59,426	6,737.7	284,576	2,435.6	58,743	13,686.3	224,536	3,398.6
2017	Club store	62.3%	861	37,049.0	1,381	48,272.2	862	87,535.8	519	153,901.0
	Mass merchandiser	90.6%	7,344	18,585.0	8,105	18,178.3	7,368	34,257.1	734	24,558.7
	Dollar store	—	—	—	—	—	—	—	—	—
	Drug store	46.0%	19,978	607.9	43,388	339.1	20,012	7,886.8	23391	1,467.3
	Grocery store	29.9%	14,644	19,578.1	48,898	9,226.0	13,559	24,940.4	34224	6,335.8
	Convenience store	9.1%	14,173	998.5	155,035	642.4	13,825	4,711.2	140790	2,697.5
	Total	19.7%	57,032	8,440.8	289,123	2,736.7	56,031	15,886.9	231,569	3,370.4

(continued)

Table B-2. Pre- vs. Post-imputation Comparison of Food and Beverage Sales ($K), by Channel and Year (continued)

Year	Channel	% TDLinx Stores in InfoScan	InfoScan Food & Beverage Sales ($K)		Imputed Population Food & Beverage Sales ($K) (all imputations)		TDLinx Total Sales InfoScan Stores ($K)		TDLinx Total Sales Non-InfoScan Stores ($K)	
			N	Mean Sales ($K)	N per Imputation	Mean Sales ($K)	N	Mean	N	Mean
	Club store	60.5%	801	36,488.5	1,324	47,794.6	801	88,927.5	523	155,816.1
	Mass merchandiser	82.3%	6,424	21,954.9	7,801	21,417.6	5,805	34,688.0	1353	31,953.5
	Dollar store	—	—	—	—	—	—	—	—	—
2018	Drug store	47.0%	19,996	551.5	42,560	301.3	20,026	7,941.1	22546	1,401.4
	Grocery store	27.0%	12,281	19,947.0	45,560	9,021.2	11,827	25,140.2	33252	6,605.1
	Convenience store	9.8%	14,982	880.4	153,645	566.3	14,672	4,836.3	138585	2,756.5
	Total	19.2%	54,515	8,061.3	284,641	2,650.1	53,504	14,970.2	229,637	3,536.1

Table B-3. Food and Beverage Sales Imputation Validation Summary

Year	Channel	N	Mean Sales ($K)	
			Observed	All Imputations
2012	Club store	597	34,733.4	34,722.0
	Mass merchandiser	6,940	14,770.9	14,733.4
	Dollar store	7,436	325.0	325.7
	Drug store	12,230	654.5	656.0
	Grocery store	6,310	15,009.4	15,014.7
	Convenience store	9,339	587.7	590.0
2013	Club store	608	33,883.1	33,878.5
	Mass merchandiser	7,214	13,711.0	13,871.1
	Dollar store	6,735	329.1	316.1
	Drug store	12,225	654.9	659.6
	Grocery store	6,871	15,610.3	15,610.3
	Convenience store	9,198	607.8	607.5
2014	Club store	604	34,449.9	34,346.3
	Mass merchandiser	7,145	14,042.7	14,133.3
	Dollar store	8,089	337.7	327.0
	Drug store	12,083	642.5	642.5
	Grocery store	6,731	15,347.8	15,347.8
	Convenience store	9,377	614.7	616.2
2015	Club store	633	35,836.4	35,845.1
	Mass merchandiser	7,387	14,799.6	14,789.2
	Dollar store	7,893	362.5	351.3
	Drug store	12,354	665.9	672.8
	Grocery store	5,979	16,936.2	16,936.2
	Convenience store	11,807	591.3	603.1
2016	Club store	844	30,346.8	30,373.6
	Mass merchandiser	7,463	14,863.7	14,923.2
	Dollar store	7,544	358.3	347.2
	Drug store	12,287	659.7	665.5
	Grocery store	6,469	16,497.9	16,497.9
	Convenience store	11,235	645.5	652.9
2017	Club store	861	37,043.4	37,019.7
	Mass merchandiser	7,344	18,582.1	18,605.6
	Dollar store	—	—	—
	Drug store	12,170	644.0	651.1
	Grocery store	7,756	17,185.3	17,185.3
	Convenience store	14,173	729.7	730.4

(continued)

Table B-3. Food and Beverage Sales Imputation Validation Summary (continued)

Year	Channel	N	Mean Sales ($K)	
			Observed	All Imputations
2018	Club store	801	36,488.4	36,435.7
	Mass merchandiser	6,424	21,107.2	21,107.9
	Dollar store	—	—	—
	Drug store	12,116	563.9	573.0
	Grocery store	5,524	16,558.0	16,558.0
	Convenience store	14,982	728.6	725.0

Appendix C. Sales Control Totals

This appendix contains tables with the control totals for food and beverage sales by channel and metro region for each year.

Table C-1. Food and Beverage Sales Control Totals ($M) by Channel and Metro Region, 2012

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	5,412	206	1,342	25,842	2,636	35,438
Los Angeles[a]	4,505	—	778	15,788	1,448	22,519
Chicago	4,902	155	659	10,096	1,537	17,350
Houston	2,922	160	294	6,243	2,182	11,800
Dallas	4,445	191	249	5,331	1,869	12,084
Miami	2,545	91	477	8,179	1,278	12,571
Atlanta	2,671	165	168	6,800	1,289	11,091
Philadelphia	2,246	126	340	7,325	628	10,665
Detroit[a]	2,911	111	313	—	1,006	4,341
Boston	1,657	58	271	7,909	771	10,666
Region 1 Northwest	10,779	772	1,228	37,212	4,992	54,983
Region 2 Midwest	31,910	1,534	2,248	57,407	15,091	108,189
Region 3 South	54,162	4,256	3,894	93,754	30,549	186,614
Region 4 West	28,100	671	2,727	71,510	12,020	115,028
Total	159,166	8,496	14,987	353,394	77,297	613,339

[a] Los Angeles dollar stores were combined with Region 4 and Detroit grocery stores with Region 2.

Table C-2. Food and Beverage Sales Control Totals ($M) by Channel and Metro Region, 2013

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	6,204	250	1,464	27,060	2,734	37,712
Los Angeles[a]	4,491	—	756	15,738	1,555	22,540
Chicago	4,947	163	684	10,591	1,503	17,888
Houston	2,798	152	310	6,738	2,509	12,508
Dallas	4,181	200	260	5,782	2,027	12,450
Miami	2,534	87	547	8,109	1,165	12,442
Atlanta	2,701	173	172	6,108	1,198	10,353
Philadelphia	2,218	134	360	8,676	526	11,913
Detroit[a]	2,855	110	326	—	1,320	4,610
Boston	1,684	59	288	8,243	781	11,055

(continued)

Table C-2. Food and Beverage Sales Control Totals ($M) by Channel and Metro Region, 2013 (continued)

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
Region 1 Northwest	10,349	796	1,234	38,928	5,496	56,803
Region 2 Midwest	31,064	1,644	2,225	59,943	16,093	110,970
Region 3 South	53,377	4,492	4,048	95,648	31,969	189,534
Region 4 West	28,829	1,666	2,720	72,864	13,748	119,827
Total	158,232	9,927	15,394	364,428	82,624	630,605

[a] Los Angeles dollar stores were combined with Region 4 and Detroit grocery stores with Region 2.

Table C-3. Food and Beverage Sales Control Totals ($M) by Channel and Metro Region, 2014

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	6,718	304	1,471	26,430	2,494	37,418
Los Angeles[a]	4,931	—	734	16,261	1,614	23,540
Chicago	5,257	178	697	10,841	1,446	18,419
Houston	2,928	177	307	6,908	2,392	12,711
Dallas	4,339	223	254	5,982	1,910	12,708
Miami	2,662	99	464	6,621	1,197	11,044
Atlanta	2,876	180	169	6,292	1,284	10,801
Philadelphia	2,326	149	358	9,840	477	13,150
Detroit	2,992	123	316	4,345	1,469	9,244
Boston	1,785	65	290	8,624	696	11,461
Region 1 Northwest	10,996	942	1,195	39,796	5,087	58,016
Region 2 Midwest	32,290	1,808	2,178	56,144	15,768	108,187
Region 3 South	56,420	4,979	3,929	98,008	30,620	193,956
Region 4 West	31,516	1,284	2,673	76,291	13,613	125,376
Total	168,036	10,511	15,034	372,383	80,066	646,030

[a] Los Angeles dollar stores were combined with Region 4.

Table C-4. Food and Beverage Sales Control Totals ($M) by Channel and Metro Region, 2015

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	6,946	328	1,583	25,422	707	34,986
Los Angeles[a]	5,545	—	758	17,796	1,625	25,724
Chicago	5,616	192	715	12,129	1,413	20,065
Houston	3,234	199	317	8,080	2,069	13,898
Dallas	5,036	245	267	6,680	2,125	14,353
Miami[a]	3,027	118	560	—	324	4,030
Atlanta	3,016	197	175	7,351	1,706	12,444
Philadelphia	2,260	162	367	9,901	172	12,861
Detroit	3,273	131	310	4,194	1,476	9,383
Boston	1,609	69	296	8,882	404	11,261
Region 1 Northwest	12,099	1,015	1,211	41,329	3,241	58,896
Region 2 Midwest	34,881	1,974	2,184	60,455	17,112	116,606
Region 3 South	59,173	5,366	4,013	117,761	24,690	211,001
Region 4 West	34,708	1,453	2,898	81,097	15,748	135,904
Total	180,423	11,447	15,655	401,075	72,811	681,411

[a] Los Angeles dollar stores were combined with Region 4 and Miami grocery stores with Region 3.

Table C-5. Food and Beverage Sales Control Totals ($M) by Channel and Metro Region, 2016

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	8,429	340	1,593	27,774	1,489	39,624
Los Angeles[a]	5,198	—	722	—	1,758	7,679
Chicago	5,172	200	698	11,728	1,494	19,293
Houston	3,118	208	299	8,211	2,435	14,272
Dallas	4,871	267	261	6,777	2,587	14,764
Miami	3,285	118	505	9,167	1,158	14,233
Atlanta	2,894	204	169	7,000	1,827	12,094
Philadelphia	2,409	164	363	10,729	294	13,959
Detroit	3,031	132	297	4,249	1,498	9,207
Boston	1,796	69	295	9,152	748	12,060
Region 1 Northwest	12,074	1,047	1,175	40,646	5,834	60,776
Region 2 Midwest	33,977	2,104	2,084	59,295	17,932	115,391
Region 3 South	58,361	5,544	3,824	106,094	31,578	205,402
Region 4 West	32,741	1,650	2,866	102,293	14,820	154,369
Total	177,357	12,047	15,151	403,115	85,452	693,122

[a] Los Angeles dollar stores and grocery stores were combined with Region 4.

Table C-6. Food and Beverage Sales Control Totals ($M) by Channel and Metro Region, 2017

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	8,049	—	1,574	28,113	2,673	40,409
Los Angeles	6,131	—	714	20,983	2,005	29,832
Chicago	6,182	—	664	16,380	1,552	24,778
Houston	3,860	—	294	8,342	2,650	15,146
Dallas	6,061	—	260	8,541	3,048	17,910
Miami	3,806	—	510	9,277	1,307	14,901
Atlanta	3,483	—	165	7,307	2,108	13,061
Philadelphia	2,566	—	358	12,177	700	15,801
Detroit	3,631	—	279	4,180	1,330	9,420
Boston	1,713	—	289	10,344	1,131	13,477
Region 1 Northwest	13,916	—	1,148	43,501	8,175	66,739
Region 2 Midwest	41,904	—	1,988	61,067	18,413	123,372
Region 3 South	73,065	—	3,650	112,064	38,140	226,919
Region 4 West	39,633	—	2,819	108,857	16,363	167,672
Total	213,999	—	14,711	451,134	99,595	779,439

Table C-7. Food and Beverage Sales Control Totals ($M) by Channel and Metro Region, 2018

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	7,866	—	1,397	27,189	2,742	39,194
Los Angeles	6,275	—	643	18,339	2,014	27,271
Chicago	5,635	—	570	11,833	1,226	19,264
Houston	3,736	—	250	8,226	1,842	14,054
Dallas	6,734	—	231	7,163	1,989	16,116
Miami	3,726	—	400	8,935	1,153	14,214
Atlanta	3,723	—	133	7,802	1,780	13,438
Philadelphia	2,594	—	335	11,337	1,884	16,150
Detroit	4,170	—	247	4,367	1,039	9,824
Boston	1,961	—	265	8,879	911	12,016
Region 1 Northwest	15,378	—	1,053	41,080	7,582	65,093
Region 2 Midwest	45,877	—	1,721	59,767	14,911	122,276
Region 3 South	79,961	—	3,091	108,100	33,525	224,677
Region 4 West	42,723	—	2,486	87,988	14,407	147,605
Total	230,359	—	12,823	411,005	87,005	741,192

Appendix D. Store Count Control Totals

This appendix contains tables showing the control totals for store counts by channel and metro region by year.

Table D-1. Store Count Control Totals by Channel and Metro Region: 2012

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	240	435	4,002	4,504	6621	15,802
Los Angelesa	199	—	1,662	2,055	2,812	6,728
Chicago	236	446	991	1,318	2,838	5,829
Houston	128	442	863	636	3,387	5,456
Dallas	188	520	622	635	3,174	5,139
Miami	120	236	1,114	887	2,404	4,761
Atlanta	143	477	628	630	2,680	4,558
Philadelphia	144	319	972	835	1,799	4,069
Detroit[a]	146	327	841	—	1,840	3,154
Boston	95	154	590	582	2,063	3,484
Region 1 Northwest	740	2,205	3,631	4,087	12,719	23,382
Region 2 Midwest	2,188	4,894	6,358	8,387	25,951	47,778
Region 3 South	2,970	12,597	12,740	12,879	57,926	99,112
Region 4 West	1,655	2,309	4,933	8,644	22,055	39,596
Total	9,192	25,361	39,947	46,079	148,269	268,848

[a] Los Angeles dollar stores were combined with Region 4 and Detroit grocery stores with Region 2.

Table D-2. Store Count Control Totals by Channel and Metro Region: 2013

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	239	492	4,145	4,506	6706	16,088
Los Angeles[a]	219	—	1,729	2,001	2,853	6,802
Chicago	237	471	1,031	1,306	2,856	5,901
Houston	131	459	937	653	3,438	5,618
Dallas	196	537	659	643	3,221	5,256
Miami	121	250	1,165	905	2,425	4,866
Atlanta	150	509	632	583	2,692	4,566
Philadelphia	144	339	1,035	829	1,797	4,144
Detroit[a]	146	356	881	—	1,828	3,211
Boston	96	162	603	595	2,084	3,540

(continued)

Table D-2. Store Count Control Totals by Channel and Metro Region: 2013 (continued)

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
Region 1 Northwest	749	2,315	3,708	4,134	12,833	23,739
Region 2 Midwest	2,216	5,123	6,519	8,400	26,038	48,296
Region 3 South	3001	13,230	13,118	12,734	58,139	100,222
Region 4 West	1,701	2,543	5,093	8,598	22,205	40,140
Total	9,346	26,786	41,255	45,887	149,115	272,389

[a] Los Angeles dollar stores were combined with Region 4 and Detroit grocery stores with Region 2.

Table D-3. Store Count Control Totals by Channel and Metro Region: 2014

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	246	538	4,264	4,586	6873	16,507
Los Angeles[a]	227	—	1,759	1,964	2,966	6,916
Chicago	240	491	1,026	1,268	2,887	5,912
Houston	134	481	968	666	3,533	5,782
Dallas	205	570	683	657	3,245	5,360
Miami	121	266	1,183	912	2,455	4,937
Atlanta	154	535	642	635	2,730	4,696
Philadelphia	141	361	1,057	851	1,830	4,240
Detroit	145	379	916	645	1,871	3,956
Boston	95	165	611	597	2,115	3,583
Region 1 Northwest	753	2,502	3,672	4,136	12,961	24,024
Region 2 Midwest	2,210	5,400	6,469	7,800	26,300	48,179
Region 3 South	3084	13,849	13,314	12,878	58,658	101,783
Region 4 West	1,741	2,779	5,094	8,556	22,530	40,700
Total	9,496	28,316	41,658	46,151	150,954	276,575

[a] Los Angeles dollar stores were combined with Region 4.

Table D-4. Store Count Control Totals by Channel and Metro Region: 2015

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	253	564	4,371	4,621	7072	16,881
Los Angeles[a]	228	—	1,782	1,934	3,030	6,974
Chicago	244	507	1,026	1,316	2,912	6,005
Houston	139	502	998	696	3,581	5,916
Dallas	212	602	713	640	3,376	5,543
Miami[a]	134	293	1,202	—	2,477	4,106
Atlanta	157	556	644	632	2,718	4,707
Philadelphia	138	377	1,070	835	1,883	4,303
Detroit	138	384	933	618	1,859	3,932
Boston	97	174	610	597	2,147	3,625
Region 1 Northwest	743	2,606	3,659	4,094	13,072	24,174
Region 2 Midwest	2,051	5,597	6,457	7,719	26,455	48,279
Region 3 South	3158	14,275	13,492	13,778	58,934	103,637
Region 4 West	1,728	2,940	5,036	8,469	22,675	40,848
Total	9,420	29,377	41,993	45,949	152,191	278,930

[a] Los Angeles dollar stores were combined with Region 4 and Miami grocery stores with Region 3.

Table D-5. Store Count Control Totals by Channel and Metro Region: 2016

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	254	595	4,521	4,715	7343	17,428
Los Angeles[a]	226	—	1,904	—	3,127	5,257
Chicago	245	529	1,085	1,328	2,998	6,185
Houston	144	531	1,053	700	3,664	6,092
Dallas	216	634	791	631	3,404	5,676
Miami	136	311	1,202	956	2,496	5,101
Atlanta	162	580	671	649	2,744	4,806
Philadelphia	138	391	1,136	835	1,940	4,440
Detroit	136	392	985	612	1,818	3,943
Boston	97	187	635	611	2,193	3,723
Region 1 Northwest	737	2,747	3,771	4,141	13,317	24,713
Region 2 Midwest	2,054	5,876	6,714	7,750	26,729	49,123
Region 3 South	3269	14,936	13,964	12,902	59,282	104,353
Region 4 West	1,743	3,176	5,386	10,342	22,911	43,558
Total	9,557	30,885	43,818	46,172	153,966	284,398

[a] Los Angeles dollar stores and grocery stores were combined with Region 4.

Table D-6. Store Count Control Totals by Channel and Metro Region: 2017

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	261	—	4,511	4,848	7515	17,135
Los Angeles	225	—	1,923	1,910	3,230	7,288
Chicago	238	—	1,057	1,333	3,008	5,636
Houston	149	—	1,020	710	3,725	5,604
Dallas	221	—	786	650	3,434	5,091
Miami	138	—	1,106	963	2,517	4,724
Atlanta	157	—	670	665	2,785	4,277
Philadelphia	138	—	1,155	843	2,021	4,157
Detroit	128	—	994	621	1,856	3,599
Boston	98	—	627	618	2,217	3,560
Region 1 Northwest	718	—	3,737	4,124	13,385	21,964
Region 2 Midwest	2,019	—	6,621	7,701	27,016	43,357
Region 3 South	3273	—	13,807	12,947	59,038	89,065
Region 4 West	1,722	—	5,389	8,522	23,176	38,809
Total	9,485	—	43,403	46,455	154,923	254,266

Table D-7. Store Count Control Totals by Channel and Metro Region: 2018

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	256	—	4,533	4,667	7512	16,968
Los Angeles	223	—	1,897	1,904	3,244	7,268
Chicago	230	—	1,003	1,266	3,004	5,503
Houston	149	—	951	699	3,766	5,565
Dallas	224	—	785	642	3,453	5,104
Miami	132	—	988	918	2,504	4,542
Atlanta	155	—	648	649	2,772	4,224
Philadelphia	130	—	1,118	815	1,989	4,052
Detroit	122	—	950	605	1,854	3,531
Boston	97	—	604	613	2,169	3,483
Region 1 Northwest	671	—	3,663	4,019	13,109	21,462
Region 2 Midwest	1,933	—	6,490	7,505	26,886	42,814
Region 3 South	3139	—	13,518	12,688	58,045	87,390
Region 4 West	1,663	—	5,260	8,516	23,144	38,583
Total	9,124	—	42,408	45,506	153,451	250,489

Appendix E. Unequal Weighting Effects

This appendix contains tables showing the UWEs by channel and metro region by year.

Table E-1. Unequal Weighting Effect by Metro Region and Channel: 2012

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	2.15	1.05	5.00	10.68	1.30	7.28
Los Angelesa	3.95	—	11.03	39.29	1.01	14.80
Chicago	1.09	1.02	1.77	34.76	1.02	11.11
Houston	1.00	1.26	14.50	11.55	1.09	4.63
Dallas	1.00	1.00	6.48	4.66	1.03	4.22
Miami	1.04	1.00	4.39	32.43	1.17	9.14
Atlanta	1.03	1.00	1.43	2.11	1.50	16.82
Philadelphia	1.14	1.03	1.34	4.16	8.76	47.42
Detroit[a]	1.56	1.59	8.85	—	1.05	3.97
Boston	8.73	2.47	1.10	5.81	1.02	4.52
Region 1 Northwest	1.04	1.49	1.36	7.08	1.02	4.68
Region 2 Midwest	1.00	1.06	3.65	10.73	1.03	4.00
Region 3 South	1.00	1.01	2.30	2.86	1.02	2.44
Region 4 West	1.06	1.00	2.11	8.95	1.06	3.17
Overall	1.23	1.09	3.85	12.34	1.56	4.84

[a] Los Angeles dollar stores were combined with Region 4 and Detroit grocery stores with Region 2.

Table E-2. Unequal Weighting Effect by Metro Region and Channel: 2013

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	5.06	1.22	9.57	12.74	1.31	8.45
Los Angeles[a]	2.47	—	13.16	19.00	1.00	8.79
Chicago	1.21	1.00	3.46	11.20	1.00	4.67
Houston	1.00	1.00	38.99	5.97	1.00	5.68
Dallas	1.00	1.00	13.62	3.14	1.00	4.39
Miami	1.02	1.05	7.59	40.84	1.01	9.83
Atlanta	1.02	1.01	1.74	11.47	1.00	13.87
Philadelphia	1.22	1.01	1.69	2.67	6.28	34.51
Detroit[a]	1.48	1.02	25.58	—	1.18	6.43
Boston	10.86	1.48	1.11	4.29	1.03	4.41
Region 1 Northwest	1.02	1.09	1.64	3.71	1.00	4.20
Region 2 Midwest	1.00	1.02	7.40	6.91	1.01	3.36
Region 3 South	1.00	1.00	3.49	3.47	1.03	2.55
Region 4 West	1.11	18.83	5.79	7.09	1.01	4.33
Overall	1.33	3.46	7.36	9.10	1.45	4.58

[a] Los Angeles dollar stores were combined with Region 4 and Detroit grocery stores with Region 2.

Table E-3. Unequal Weighting Effect by Metro Region and Channel: 2014

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	6.00	1.57	8.04	9.79	1.22	7.14
Los Angeles[a]	3.11	—	11.08	28.09	1.00	11.02
Chicago	1.26	1.03	2.23	39.22	1.00	13.90
Houston	1.01	1.01	42.86	7.01	1.02	5.96
Dallas	1.00	1.00	16.49	3.65	1.10	4.00
Miami	1.03	1.00	24.10	44.25	1.05	12.33
Atlanta	1.03	1.00	1.86	22.33	2.46	18.40
Philadelphia	1.25	1.01	1.70	2.04	3.25	18.05
Detroit	1.58	1.03	22.62	14.44	1.31	8.69
Boston	10.14	1.89	1.14	6.88	1.00	4.66
Region 1 Northwest	1.02	1.18	1.51	3.88	1.04	4.34
Region 2 Midwest	1.01	1.04	5.66	7.39	1.00	3.12
Region 3 South	1.00	1.02	2.98	4.28	1.00	2.28
Region 4 West	1.15	1.87	3.26	6.99	1.03	2.91
Total	1.39	1.16	7.04	10.77	1.43	4.40

[a] Los Angeles dollar stores were combined with Region 4.

Table E-4. Unequal Weighting Effect by Metro Region and Channel: 2015

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	4.17	1.27	5.99	7.93	1.59	7.24
Los Angeles[a]	2.93	—	10.22	34.23	1.03	12.09
Chicago	1.36	1.00	1.67	8.16	1.02	3.90
Houston	1.02	1.01	43.51	4.43	1.02	5.76
Dallas	1.01	1.00	15.42	2.41	1.13	3.25
Miami[a]	1.05	1.02	2.75	—	2.45	8.33
Atlanta	1.03	1.01	1.54	17.89	1.02	4.56
Philadelphia	1.19	1.01	1.59	1.47	1.00	6.11
Detroit	1.78	1.02	16.53	28.91	1.18	10.74
Boston	3.51	1.35	1.14	3.86	1.48	5.79
Region 1 Northwest	1.01	1.09	1.42	2.94	1.78	5.94
Region 2 Midwest	1.04	1.02	4.10	4.86	1.00	2.54
Region 3 South	1.00	1.01	2.36	2.91	1.22	2.24
Region 4 West	1.28	1.51	2.01	5.71	1.00	2.59
Overall	1.29	1.10	5.27	7.39	1.53	3.80

[a] Los Angeles dollar stores were combined with Region 4 and Miami grocery stores with Region 3.

Table E-5. Unequal Weighting Effect by Metro Region and Channel: 2016

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	3.01	5.91	6.33	9.61	1.11	7.01
Los Angeles[a]	4.72	—	12.13	—	1.03	4.24
Chicago	1.32	1.01	1.91	35.82	1.02	11.04
Houston	1.02	1.01	37.22	9.27	1.01	5.83
Dallas	1.01	1.00	13.90	2.35	1.02	2.93
Miami	1.05	1.04	3.28	20.76	1.15	6.24
Atlanta	1.01	1.00	1.65	41.35	1.01	5.91
Philadelphia	1.08	1.00	1.82	5.54	1.02	6.35
Detroit	3.43	1.02	27.91	32.83	4.18	17.63
Boston	1.44	1.83	1.17	5.58	1.02	3.94
Region 1 Northwest	1.01	1.09	1.50	4.52	1.04	3.55
Region 2 Midwest	1.03	1.05	4.28	5.28	1.02	2.61
Region 3 South	1.00	1.01	2.46	3.97	1.01	2.00
Region 4 West	1.18	11.24	2.51	9.32	1.03	4.18
Overall	1.26	2.48	5.69	9.99	1.24	3.77

[a] Los Angeles dollar stores and grocery stores were combined with Region 4.

Table E-6. Unequal Weighting Effect by Metro Region and Channel: 2017

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	1.37	—	14.73	8.82	1.14	7.38
Los Angeles	3.29	—	22.98	51.76	1.01	18.16
Chicago	1.28	—	5.61	6.73	1.00	3.70
Houston	1.02	—	17.51	11.19	1.02	4.10
Dallas	1.01	—	30.65	2.34	1.00	4.14
Miami	1.02	—	17.00	12.37	1.05	6.57
Atlanta	1.00	—	3.11	29.91	1.00	5.25
Philadelphia	1.04	—	6.13	7.36	1.50	6.61
Detroit	1.61	—	34.49	15.31	1.01	9.65
Boston	1.10	—	1.33	5.78	1.06	3.94
Region 1 Northwest	1.00	—	2.65	3.58	1.09	2.88
Region 2 Midwest	1.03	—	15.32	3.77	1.00	3.06
Region 3 South	1.00	—	7.43	2.73	1.03	2.25
Region 4 West	1.12	—	11.58	8.13	1.08	3.59
Overall	1.12	—	12.58	9.51	1.21	4.12

Table E-7. Unequal Weighting Effect by Metro Region and Channel: 2018

Metro Region	Club & Mass Merchandiser	Dollar	Drug	Grocery	Conv.	Overall
New York	1.38	—	19.65	9.33	1.49	8.14
Los Angeles	2.16	—	16.96	35.47	1.00	13.40
Chicago	1.37	—	3.38	20.50	1.02	7.31
Houston	1.01	—	15.29	4.96	4.31	10.01
Dallas	1.01	—	26.97	1.95	2.71	7.37
Miami	1.01	—	19.88	7.19	1.61	7.68
Atlanta	1.00	—	2.13	14.90	1.15	4.77
Philadelphia	1.08	—	2.53	4.03	1.84	2.88
Detroit	1.68	—	19.69	30.97	1.01	11.25
Boston	1.08	—	1.22	4.79	1.30	5.10
Region 1 Northwest	1.00	—	1.79	3.49	1.43	3.40
Region 2 Midwest	1.02	—	8.79	3.55	1.14	2.58
Region 3 South	1.00	—	3.65	2.29	1.47	2.54
Region 4 West	1.07	—	5.45	7.16	1.62	3.72
Overall	1.09	—	9.45	9.02	1.74	4.44

Appendix F. User's Guide

This document describes the datasets and provides the codebook and other details to assist users in understanding and using the weights when analyzing IRI InfoScan data.[12]

F.1 Weighting Files

The weighting data files are listed in Table F-1, and Table F-2 provides the codebook common to all files. The data files contain one record per establishment per year, with attributes such as firm ownership based on June of the given year. Files are available in both SAS and CSV format. The files in CSV format can be read by most statistical analysis software. Note that after 2016 there are few dollar stores in InfoScan and insufficient geographic coverage to construct weights; hence, these stores were excluded from the 2017 and 2018 weighting files.

Table F-9. Datasets for InfoScan Store Weights

File Name	Year	No. of Records
Final_weighted_2012	2012	58,110
Final_weighted_2013	2013	57,823
Final_weighted_2014	2014	61,210
Final_weighted_2015	2015	60,847
Final_weighted_2016	2016	61,967
Final_weighted_2017	2017	56,438
Final_weighted_2018	2018	55,120

Table F-10. InfoScan Weights Codebook

Variable Name	Variable Definition	Type
ChannelID	Industry channel[a]	Categorical
EstabID	Establishment identifier	Numeric
MetroRegion	Geographic area code	Categorical
MetroRegion2	Geographic area used in calibration	Categorical
MetroRegionName	Geographic area label	Text
MetroRegionName2	Geographic area code label	Text
StoreID	InfoScan store identifier	Numeric
geogkey	RMA identifier	Numeric

(continued)

[12] We created the weighting files under the assumption that users do not have access to the TDLinx dataset of store information. However, in some cases, users may want to combine fields from the TDLinx data with the InfoScan data for specific analyses.

Table F-11. InfoScan Weights Codebook (continued)

Variable Name	Variable Definition	Type
inRMA	Indicates store is in an RMA	Binary
sales	Total food and beverage sales ($K)	Numeric
weight	Weighting factor	Numeric

[a] Industry channel refers to store type (i.e., club store/mass merchandiser, dollar store, drug store, grocery store, or convenience store).

As described in the weighting file documentation, we combined stores in some metros and channels with Census regions for the purpose of calibration. The variables MetroRegion2 and MetroRegionName2 contain these geographic area definitions used in the calibration process. For 2017 and 2018, these are the same as MetroRegion and MetroRegionName.

It is possible for establishments in InfoScan, for 2013 on, to have multiple StoreIDs in a given year due to midyear changes such as changes in ownership. In addition, a StoreID may have multiple records due to changes such as switching from individual store to RMA reporting, or perhaps if some UPCs are reported at the RMA level while others are reported for the individual store. Like TDLinx, the weighting files differ from InfoScan data files in that they are establishment-level datasets; that is, they contain one record per establishment per year, with attributes such as firm ownership based on June of the given year. When multiple InfoScan StoreIDs were linked to the same TDLinx ID or when duplicate StoreIDs occurred in the InfoScan data, we aggregated the records into a single establishment by summing food and beverage sales across records and assigning the StoreID and characteristics of the StoreID with the largest sales total for the year. In case analysts need to separate these establishments into stores with the original InfoScan StoreIDs, supplementary establishment link files in SAS and CSV formats, using the naming convention establink*yy*, are provided that include the variables StoreID, EstabID, and geogkey. They can be linked to the weighting files using the EstabID variable. The EstabID variable is a randomly generated identifier. Files linking these IDs to TDLinx IDs are available to users authorized to access TDLinx.

As described in Section 2, we developed weights for each store in the InfoScan database that can be used to make population-based inferences. The exceptions are that we did not calculate weights for stores in Puerto Rico, defense commissaries, or liquor stores. The files include total food and beverage sales that were computed from InfoScan as the sum of random-weight, branded-product, and private-label sales. Sales for some stores were computed using imputed random-weight or private-label sales. For retailers that report sales at the RMA level, sales were disaggregated to individual stores.

F.2 Using InfoScan Weights

The weights for stores in InfoScan can be used in calculating point estimates representative of the nation or one of the metro-based geographic areas used to develop the weights. These include the 10 largest metropolitan areas, as defined by the number of retail food and beverage stores (New York, Los Angeles, Chicago, Houston, Dallas, Miami, Atlanta, Philadelphia, Detroit, and Boston), with the remaining stores grouped by Census region. We do not recommend using the weights to calculate point estimates for other definitions of geographic areas because the weights were constructed to align with control totals calculated using the top 10 metropolitan areas and the remainder in four Census regions.

The weights should be used the same way as sampling weights are typically used when analyzing survey data. For example, the total food and beverage sales in the United States can be computed by summing the product of weights and sales across stores (i.e., $\sum_i weight_i \cdot sales_i$ where i = stores). We recommend using software that can handle survey weights for computing standard error estimates and conducting other types of analyses. Table F-3 provides other examples for programming code using SAS-callable SUDAAN and Stata.[13]

Table F-12. Example Code Using Weights

Software	Sample Code
SUDAAN	proc descript data=final_weighted notsorted; nest _one_; var sales; weight weight; run;
Stata	use final_weighted.dta, clear svyset[pweight = weight] svy: mean sales

F.3 Alternative Weighting Approach

An alternative approach to weighting, which provides more flexibility in terms of defining geographic regions, is to use a sales-volume weighted approach. Using store sales volume as a weight takes the gravitational approach to food shopping: holding travel distance constant, consumers are more likely to shop at larger stores than at smaller stores because the former are likely to offer greater variety at lower prices (Briesch et al., 2009). For an application using this approach, see Appendix A in Zhen et al. (2019), in which the authors used sales volumes to calculate sales-weighted price indexes for the MSA where the

[13] Analyzing the data in this manner is equivalent to assuming the sample comes from a with-replacement sampling design in which stores are the primary sampling units with probability of selection equal to the reciprocal of the analysis weight. Uncertainty in the weights themselves is not accounted for in variance estimates.

household was located. Users can also develop more sophisticated weighting schemes that account for the distance between each store and a household under the assumption that a household is less likely to shop at a more distant store, ceteris paribus (Taylor & Villas-Boas, 2016).