Published in final edited form as:
Health Place. 2012 November ; 18(6): 1341–1347. doi:10.1016/j.healthplace.2012.06.016.

Improving retrospective characterization of the food environment for a large region in the United States during a historic time period

Amy H. Auchincloss[(a)], **Kari A. B. Moore**[(b)], **Latetia V. Moore**[(c)], and **Ana V. Diez Roux**[(b)]

Amy H. Auchincloss: aha27@drexel.edu; Kari A. B. Moore: kbrunn@umich.edu; Latetia V. Moore: lvmoore@cdc.gov; Ana V. Diez Roux: adiezrou@umich.edu

[(a)]Department of Epidemiology and Biostatistics, School of Public Health, Drexel University, Philadelphia, Pennsylvania, USA

[(b)]Center for Social Epidemiology and Population Health, University of Michigan Dept. of Epidemiology, Ann Arbor, Michigan, USA

[(c)]Division of Nutrition, Physical Activity, & Obesity, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control & Prevention, Atlanta, Georgia, USA

Abstract

Access to healthy foods has received increasing attention due to growing prevalence of obesity and diet-related health conditions yet there are major obstacles in characterizing the local food environment. This study developed a method to retrospectively characterize supermarkets for a single historic year, 2005, in 19 counties in 6 states in the USA using a supermarket chain-name list and two business databases. Data preparation, merging, overlaps, added-value amongst various approaches and differences by census tract area-level socio-demographic characteristics are described. Agreement between two food store databases was modest: 63%. Only 55% of the final list of supermarkets were identified by a single business database and selection criteria that included industry classification codes and sales revenue >=$2 million. The added-value of using a supermarket chain-name list and second business database was identification of an additional 14% and 30% of supermarkets, respectively. These methods are particularly useful to retrospectively characterize access to supermarkets during a historic period and when field observations are not feasible and business databases are used.

MeSH Keywords

Residence characteristics; validity; reliability; food; geography; environment

Corresponding author: Amy Auchincloss, PhD, MPH, Department of Epidemiology and Biostatistics, School of Public Health, Drexel University, Philadelphia, PA 19102, USA, aha27@drexel.edu, Tel: (215)762-2056, Fax: (215)762-1174.

Author contributions: AH Auchincloss designed the study, supervised data compilation and analyses, and drafted the paper. EAB Moore contributed to study conceptualization, analyzed data, and contributed to writing the paper. LV Moore contributed to study conceptualization, interpreted findings, and edited drafts. AV Diez Roux contributed to study conceptualization, interpreted findings, and critically reviewed drafts.

INTRODUCTION

Factors related to access to healthy foods have received increasing attention due to growing prevalence of obesity and other diet-related health conditions. A recent U.S. review on this topic concluded that neighborhood residents tend to have healthier diets, including higher intakes of fruits and vegetables, and lower rates of overweight and obesity, when they have better access to supermarkets (Larson et al., 2009). Many studies have characterized healthy food access based on the types of stores available. Supermarkets offer a wide variety of unhealthy foods but appear to be relatively good proxies for availability of healthy food because compared to other grocery stores they tend to offer a larger volume of healthy choices and prices are often lower (Andreyeva et al., 2008; Ellickson and Misra, 2008; Franco et al., 2008). Nevertheless, important questions remain regarding the impact of the local food environment (or supermarkets in particular) on diet and related conditions in part due to major obstacles in obtaining valid measures of the local food environment.

There are multiple ways to identify locations of food retailers and supermarkets in particular including field observations, lists of food licenses from local departments of health or agriculture, or national lists compiled by commercial entities. Outside the U.S., government food license databases(Pearce et al., 2008) or field observations(Macdonald et al., 2011) have primarily been used although lists compiled by commercial entities may be improving and thus could become useful to health researchers (Bisnode, 2011; D&B, 2011a).

Food environment field audits are not possible for characterizing food establishments during historic periods and are typically impractical for characterizing establishments in very large regions. In the U.S., the most feasible approach for identifying supermarkets for very large regions and for historic time periods, is to use commercially compiled lists (Ver Ploeg et al., 2009). Most large-scale U.S. health studies examining food environment and health have used one of two business databases: Dun & Bradstreet (D&B, 2011b) and InfoUSA (Glanz, 2009; InfoUSA, 2011; Powell et al., 2007b). Businesses register with these companies in order to obtain a tax identification number and in this process they are assigned an industry classification code based on their primary activity (generally the activity that generates the most revenue for the establishment) (D&B, 2011c). The commercial database company compiles additional information on each business location using information that the business establishment provides on surveys, census forms, or administrative records.(US Census Bureau, 1993, 2011a) In the absence of direct observation, researchers purchase commercial databases and rely on industry classification codes to identify food stores;. A couple of studies have used a diversity of methods and study designs to examine the agreement between supermarkets on business lists vs. those observed in the field and found moderate (around 40% or higher (Powell et al., 2011)) to high agreement (about 80% (Bader et al., 2010; Liese et al.)). Commercial databases are prone to error due to companies' reporting bad or obsolete data, or misunderstanding questions (especially industry definitions), or clerical errors(Evans, 1987; Hoehner and Schootman, 2010). Nevertheless, data constraints have driven public health researchers to continue to rely on these databases to characterize the retail food environment (Liese et al., 2010; Michimi and Wimberly, 2010; Powell et al., 2007a; Powell et al., 2007b).

Some have called for using multiple data sources in order to supplement incomplete business directories (Forsyth et al., 2010; Ver Ploeg et al., 2009). However, many researchers only use a single business directory (Black et al., 2010; Franco et al., 2009; Li et al., 2008; Moore and Diez Roux, 2006; Neckerman et al., 2010). Due to the dominance of large corporations in the supermarket sector, a number of previous studies have tried to improve supermarket identification by using a list of chain-name supermarkets. However details are often lacking on the origin of these lists (for example: (Hoehner and Schootman,

2010; Moore and Diez Roux, 2006; Moore et al., 2008a; Moore et al., 2008b; Morland et al., 2002; Rundle et al., 2009; Smiley et al., 2010)). With the exception of a few studies,(Liese et al., 2010; Moore and Diez Roux, 2006; Rundle et al., 2009) researchers have published little information on the origin of supermarket chain-name lists, algorithms for data identification, or the contribution of one supermarket identification approach versus another.

This study responds to the call from expert panels for transparency in data compilation and improved data classification in order to accelerate knowledge acquisition and improve food environment study comparability (Story et al., 2009). This study details issues that arise when deriving a supermarket directory for a large region in the U.S. for a historic time period. The study describes methods for preparing and combining two commercial datasets and examines overlaps and added-value amongst various ways to identify supermarkets. Differences in datasets by area-level socio-demographic characteristics are also assessed.

METHODS

This work is part of a larger study that is characterizing food environments during a 10 year period (2000–2011) for participants in the Multi-Ethnic Study of Atherosclerosis (MESA) (Bild et al., 2002). This study developed a method of identifying supermarkets for a single historic year, 2005, which will subsequently be applied to other years to characterize food environments for the cohort. This study area includes 19 counties in 6 states: California (Los Angeles and Orange); Illinois (Cook and DuPage); Maryland (Baltimore, Howard, Carroll, and Baltimore City); Minnesota (Ramsey and Washington), North Carolina (Forsyth, Davidson, Davie, Yadkin, and Stokes); New York (Manhattan, Bronx, Queens, and Kings). In total, this area included 7008 census tracts and 11,408 square miles (29,546.584 kilometers).

Data

Year 2005 food store databases were obtained from two business databases.

Nielsen/TDLinx Retail Site Database (TD)—The Nielsen/TDLinx's Retail Site Database (TD) uses the Food Marketing Institute definition of a supermarket: stores selling a full line of food products and generating at least $2 million in yearly revenues (Nielsen Company, 2008; Progressive Grocer, 2011). The registry is updated twice per year. These data are widely used by marketing firms, food manufacturers and media including *Marketing Guidebook* and *Market Scope* publications and *Progressive Grocer*. Nielsen does its own research to catalogue stores as supermarkets and does not rely on US Census industry classification codes. This study purchased 3002 supermarkets which included the following categories: conventional supermarkets (e.g., Albertsons, Food Lion); limited assortment supermarkets which carry most grocery products but have less variety than conventional supermarkets (e.g., Whole Foods, Trader Joes) and may display their items in cardboard shipping boxes (e.g., Save-A-Lot, Aldi Food Store); warehouse supermarkets (e.g., Cash & Carry, Smart & Final); and supercenters (e.g., SuperTarget Center, Wal-Mart Supercenter).

A *supermarket "chain-name list"* was derived from a frequency of names in the TD data for use later in analyses. In order to maximize the sensitivity of the list, we pooled data from two years (2000 and 2005) in generating this list. Chains have been defined as a grouping of at least 10 (Nielsen Company, 2008) or 11 stores (Merrefield, 1998). For this study, we lowered the definition slightly reflecting that there will be fewer stores in our study area than the nation. If the name was repeated in at least 8 distinct locations in TD it was added to the chain name list which resulted in a supermarket chain-name list of 161 names (AKA "supermarket chain names").

Dun and Bradstreet - National Establishment Time Series (NETS)—The second database was the National Establishment Time Series (NETS) database from Walls & Associates(Walls & Associates, 2010) who, each year convert Dun and Bradstreet (D&B) archival establishment data into a time-series database. Data in each business included standard industry classification code (SIC)(US Census Bureau, 2011b), company name, trade name, sales revenue, and employee size. SIC codes are very similar to North American Industry Classification System (NAICS) and were used by this study in order to compare to previous work that used SIC codes (Moore and Diez Roux, 2006). The parent study purchased 110,917 food and drink related businesses including SIC codes 5411 [Grocery Stores], 5421 [Meat and Fish Markets], 5431 [Fruit and Vegetable Markets], 5441 [Candy, Nut, and Confectionary Stores], 5451 [Dairy Product Stores], 5461 [Retail Bakeries], 5499 [Miscellaneous Food Stores], 5812 [Eating Places], 5813 [Drinking Places], and 5921 [Liquor Stores]. This study further classified records as SIC "grocery or supermarket" using SIC 54110000 (Grocery Stores), 541101 (Supermarkets), 54119900 (Grocery Stores, nec), 54119901 (Cooperative food stores), 54119904 (Grocery stores, chain), and 54119905 (Grocery stores, independent). Due to concerns with relying on SIC codes alone to classify stores as supermarkets, this study identified supermarkets in NETS by combining various criteria. Supermarkets were defined as businesses that had a "grocery or supermarket" SIC code and had sales revenue ≥$2 million or employed ≥25 persons. Employee size has been used in prior studies because sales revenue is sometimes not available or is unreliable (Evans, 1987; Moore and Diez Roux, 2006; Rundle et al., 2009). A suitable cut-point (≥25 employees) was determined from the observed distribution of chain-name supermarkets. In addition food/drink stores and eating places whose company name or trade name was on the "supermarket chain-name list" were also classified as supermarkets regardless of their SIC code.

Census data—American Community Survey 5-year census tract data for 2005–2009 (US Census Bureau, 2010) were used to describe the areas where supermarkets were located. Census tracts were classified as high poverty if ≥30% of households had incomes under the poverty threshold (for reference, in 2009 14% of US households were poor), mid-high population density defined as ≥1000 persons per square mile (1 square mile is 2.6 square kilometers, U.S. average in 2010 was 84.4 persons per square mile), and high proportion of minority residents defined as <40% of residents in the tract were non-Hispanic White (U.S. proportion in 2010 was 64%). These cut points approximated the median value in our sample and/or approximate definitions that have been used by others (Jargowsky, 1997; Moore and Diez Roux, 2006; US Census Bureau and Bishaw, 2011).

Analyses

Each dataset was prepared for merging by cleaning the data, standardizing supermarket names, applying the supermarket chain-name list, and exploring the added value of using other criteria (sales revenue, number of employees, and chain name list) for supermarket identification. The process for merging/comparing the two supermarket data sources is described below. Finally, analyses were performed to compare the datasets by demographic characteristics of areas.

Preparing the data for merging—Extensive work was done to clean the data; this work was primarily performed on the NETS database because the TD database had minimal irregularities. Duplicates were removed from NETS (n=1691, 1.5% different unique identifiers were found for the same name and location) and company headquarters were eliminated since they do not provide direct services to the public (n=1796, 1.6%). Word lists for synonyms, spelling irregularities, and exclusions were written and iteratively refined. Examples of words from the synonym list are: 'ALDI INC' and 'ALBRECHT

DISCOUNT', 'RALPHS MARKET' and 'RALPHS FRESH'; examples of words from the spelling variants list are: 'BILO' and 'BI-LO', 'SAFEWAY' and 'SAFE WAY'; examples of exclusion words are: AUTOMOBILE, DESIGN, MEDICAL. All records were geocoded using TeleAtlas EZ-Locate web-based geocoding software (Tele Atlas, 2011; Zhan et al., 2006); 99% of establishments in both databases were geocoded to the street level.

Matching name and location—The agreement between the two databases was investigated by comparing the 3002 supermarkets identified through TD to 2962 supermarkets identified through NETS. We first investigated the extent to which businesses in both databases matched in geographic location and then matched by name. Geographic location matching used the following location match categories: "excellent" (or perfect) was defined as building number and street address match, "good" was defined as <0.001 mile (1.6 meter) distance between latitude/longitude, and "fair" (or poor) defined as >0.001 mile and <0.30 mile (<483 meters) distance between latitude/longitude. In the fair/poor category, only 24 records were 0.001–0.01 miles, 19 were 0.01–0.1 miles and only 12 were between 0.1–0.3 miles. The cut point of <0.30 miles has been used by others (Ver Ploeg et al., 2009) and results were almost identical to a cut point of <0.50 miles [80 meters]. Name matching used an algorithm that calculated the asymmetric spelling distance score between records (the SAS SPEDIS function, details are provided elsewhere (Gershteyn, 2000; Roesch, 2011)). The possible range of spelling distance scores was 0–100 with lower scores indicating smaller spelling distance thereby a better match. A score of 10 was used by others as the cut point for good spelling distance(Gershteyn, 2000) but there is no agreed upon standard. For this study, cut-points were defined after examining the empirical distribution of the score and manual inspection. Three categories were defined as perfect or "excellent" 0-≤15, "good" >15 –≤45, and "fair" (or poor) >45–100. Records that had "good" or "fair" name match scores were manually inspected. (See Table 1 and the table footnotes for the combinations of acceptable matches.) Percent agreement was used to describe the agreement between supermarkets identified in NETS vs. in TD. Finally, census tract information was used to describe the areas that the stores were located in and to examine whether census characteristics differed for the two databases.

RESULTS

Supermarkets identified via TD or NETS

Among the 3002 TD records purchased for this study, TD classified 85% as conventional supermarkets, 11% as limited assortment supermarkets, 3% as warehouse clubs, and 1% as supercenters. Chain supermarket companies dominated the supermarket retailers: 75% of TD supermarket locations were operated by one of the 161 companies that had at least eight outlets.

Among the 107,430 NETS records (unduplicated and non-headquarters), 18,886 (17.6%) had a SIC code indicating that they were a grocery store or supermarket. Among those, 2454 (13%) met at least one of the supermarket criteria of annual sales volume of $2 million or more, 25 or more employees on the annual payroll, or the store name was on the chain-name list (Figure 1). Most of these met the sales criteria (83% out of 2454). The employee size criteria added little over and above sales; <1% of the 2454 supermarkets were added-in via employee size. The chain-name list criteria made a fairly substantial contribution to supermarket identification. A total of 413 supermarkets out of the 2454 (16.8%) were identified via the chain-name list. Utilizing the chain-name list criteria, 84 supermarkets were also added-in that did not have a grocery store or supermarket industry classification code. For example, "Publix Supermarket", "Whole Foods Market" and "Trader Joes" were large stores that were identified via the supermarket chain-name list criteria (not via industry

classification code) because they often had industry classification descriptors of “Retail Bakery” or “Miscellaneous Food Store” (see details in Supplement Table 2).

Matching

Based on the criteria described above, 3002 supermarkets were identified in the TD database and 2454 supermarkets were identified in NETS. Table 1 shows name/location matches between NETS and TD. Of the 3652 supermarkets identified by either database, overall, 62% had excellent or good location agreement (1571+686) and 57% had excellent or good name agreement (1870+203). A total of 690 TD supermarket locations did not match any NETS supermarket name/location. A total of 650 NETS supermarkets did not match any TD supermarket name/location. (See Figure 1 and also Supplement Table 3 for examples of the unmatched stores that were included in the final file.) Finally, another 424 supermarkets in NETS were identified by matching location and name NETS records without a grocery or supermarket SIC code to a TD supermarket location and name.

Final supermarket list

Table 2 shows the final tally of 3652 supermarkets by database and criteria of inclusion. The percent agreement between the two databases was 63% (2312/3652) leaving 1340 (37%) of all records in either database but not both. If only industry classification codes and sales volume criteria had been utilized, only a bit over one-half of total supermarkets would have been identified (2024/3652=55%). Supplementing industry classification codes and sales volume with a supermarket chain-name list, resulted in identification of an additional 497 supermarkets (14%) or combined 69% of all supermarkets identified ([2024+413+84]/3652). Supplementing this information with a second business database, TD, identified another 12% of supermarkets (n=424, these records were in NETS but not identified through SIC codes or names on the chain name list but were included in the TD database as supermarkets). Finally, using the supermarket list from TD, another 19% of stores were identified (n=690, stores that were not in NETS at all). In sum, the added-value of a second business database (TD) was identification of another 30% of supermarkets ([424+690]/3652).

Census data

Table 3 shows that 13% of census tracts the study area were very poor (census tract poverty density ≥30% of households); 39% had fairly high population density (defined as ≥1000 persons per square mile); and about half had a high proportion of minorities (defined as <40% of residents were non-Hispanic white). TD supermarkets and NETS supermarkets were located in census tracts that were similar to each other on these dimensions.

DISCUSSION

This study developed a method to characterize supermarkets for a single historic year, 2005, in 19 counties in 6 states using a number of variables from two business databases as well as a supermarket “chain-name list”. This study provides evidence of the added-value of using multiple data sources to identify supermarkets. Agreement between the two food store databases was modest (63%) indicating that a single business database is unlikely to provide a complete inventory of supermarkets. Using criteria most often employed in public health research to identify supermarkets – a combination of industry classification codes and sales revenue ≥$2 million – only 55% of the final list of supermarkets would have been identified. The added value of a supermarket chain-name list was that it identified an additional 14% of supermarkets on the final list. The added-value of a second business database (TD) was identification of another 30% of supermarkets.

In sum, this study suggests that using multiple criteria and datasets will improve retrospective characterization of supermarkets during a historic period; these methods are particularly useful when field observations are not feasible and when using business databases. While it is improbable that measurement error in characterizing food stores will ever be completely eliminated, use of certain methods -- such as using multiple attribute fields, multiple databases, and a chain-name list -- may reduce measurement error. It is rare for researchers to use multiple classification methods or datasets to classify supermarkets. To distinguish U.S. supermarkets from other food stores, previous studies primarily used commercial databases and identified supermarkets using industry classification codes or either sales or employee size (Block and Kouba, 2006; Kowaleski-Jones et al., 2009; Liese et al., 2010; Michimi and Wimberly, 2010; Powell et al., 2007a; Powell et al., 2007b; Wang et al., 2006; Wang et al., 2007). A few studies used a combination of multiple criteria: classification code, sales and/or employees, and chain name (Moore et al., 2008a; Rundle et al., 2009) and a few studies used multiple databases (Hoehner and Schootman, 2010; Kowaleski-Jones et al., 2009; Liese et al., 2010; Rundle et al., 2009; Ver Ploeg et al., 2009; Wang et al., 2006). In our study, there was significant overlap in supermarket identification when using sales revenue ≥$2 million and employee size ≥25; 90% of records with high sales also had high employees. Thus, in future work if sales is not available, employee size may be an adequate proxy for sales.

A number of previous studies mentioned that they used some sort of supermarket chain-name list but details on the list were largely absent (Hoehner and Schootman; Moore and Diez Roux, 2006; Rundle et al.). Chains increasingly dominate retail food shopping and chains and non-chains may be quite different in the breadth of their products and in their prices so it could be beneficial to differentiate supermarket subtypes. A systematic method for developing a name list -- that can be used to improve supermarket identification and also differentiate chain from non-chain -- can be done following what was done in this study: by examining high-frequency store names within a food store database (provided that the names are standardized and that the dataset is large enough such that frequencies are meaningful).

Chain stores may be more common in high income and non-minority communities (Chung and Myers, 1999; Powell et al., 2007b). Our study used chain name as another tool for supermarket identification, thus sensitivity of stores may have been higher in more advantaged commuhnites. We examined whether supermarket identification by database differed according to whether the tract was poor, densely populated, or had a high concentration of non-White residents; on average no differences were found although aggregation of the data could potentially have masked local differences.

Because our objective was to identify supermarkets for a historical time period and over a large region, we had no way of validating the commercial database. Results assumed that validity increased when multiple databases were merged however, without a “gold standard” we cannot prove that this was the case. The current study’s use of two data bases likely resulted in higher sensitivity but this may have occurred at the expense of specificity. A small handful of validation studies have been done that utilized D&B data (Hoehner and Schootman, 2010; Liese et al., 2010; Powell et al., 2011). Criteria used to define a supermarket in D&B varied across studies and validity of data may change depending on study location among other factors. One study in the St. Louis region study estimated that after merging/appending one commercial food store database to another, sensitivity of identifying a food store may rise from <40% up to 90% (Hoehner and Schootman, 2010). A validation study of food stores in North Carolina found food stores about twice as likely to be undercounted as opposed to being over-counted (closed or otherwise not in existence) (Liese et al., 2010). However, a recent validation study by Powell et al was less conclusive

regarding the directionality of the errors. The study was conducted in a large metropolitan area in the U.S., examined agreement between supermarkets identified in a field audit vs. those identified by industry classification codes in the commercial database D&B. Depending on the stringency of matching criteria, both sensitivity and specificity of supermarkets both ranged from about 40%-70% (Powell et al., 2011).

In the present study, specificity was aided by having to fulfill multiple industry classification code criteria to qualify as a supermarket and extensive work was performed to guard against duplication of records: careful data merging/matching, automated de-duplication procedures, and manual inspection. If over-counting occurred, we speculate it may have occurred among D&B records that did not match the TD data: a portion of the 650 [18%] of total supermarkets identified. D&B is a huge database with wide breadth in the diversity of establishments while the TD list is a relatively small list that is primarily compiled to support in-depth large company profiles (Nielsen Company, 2008) so validity may be somewhat higher. In summary, we must consider the possibility of both over- and under-identification errors and absent validation, there is no way of confirming where the largest problems occurred.

Groceries are increasingly being sold in non-traditional venues such as drug stores, gas stations, discount department stores or mass merchandise/dollar stores. Approximately 15% of supermarkets are limited assortment, warehouse, or supercenters (according to TD classifiers). Unfortunately, these non-traditional food venues are hard to identify via industry classification codes since they do not tend to be in the “grocery/supermarket” classification category. A recent study that compiled food store/restaurant data for 8 counties added-in industry codes for a number of non-traditional venues including discount stores, warehouse clubs, supercenters, and other general merchandise stores (Liese et al., 2010). If we had purchased all businesses for the regions and years of our study, D&B data may have been more complete, a greater proportion of TD supermarkets may have matched the NETS data and then the added value of TD supermarkets may lessened; however, the costs of adding those business categories was not feasible for this study.

Commercial interests have long been interested in characterizing the retail environment and this is increasingly being recognized as important for public health planning and health policy. The steps described in this study can be used to characterize access to supermarkets and ultimately link the food environment data to healthy behaviors and conditions to assess influences on dietary intake and health conditions. This study attempted to be as transparent as possible in order to facilitate use of these methods by other researchers. Our study identified supermarkets using commercial databases but the methods we developed could be broadly applied to other types of location data thus has potential relevance to multiple research areas.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: This research was supported by grant 2R01 HL071759 from National Heart, Lung, and Blood Institute at the National Institutes of Health.

Acknowledgements for contributors: Special thanks to Lu Mao for assisting with data compilation. The findings and conclusions are those of the authors and do not necessarily represent the official position of the CDC

References

Andreyeva T, Blumenthal DM, Schwartz MB, Long MW, Brownell KD. Availability and prices of foods across stores and neighborhoods: the case of New Haven, Connecticut. Health Aff (Millwood). 2008; 27:1381–1388. [PubMed: 18780928]

Bader MD, Ailshire JA, Morenoff JD, House JS. Measurement of the local food environment: a comparison of existing data sources. Am J Epidemiol. 2010; 171:609–617. [PubMed: 20123688]

Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, Jacob DR Jr, Kronmal R, Liu K, Nelson JC, O'Leary D, Saad MF, Shea S, Szklo M, Tracy RP. Multi-Ethnic Study of Atherosclerosis: objectives and design. Am J Epidemiol. 2002; 156:871–881. [PubMed: 12397006]

Bisnode. Bisnode Business Information Group database. Ratos and The Bonnier Group; 2011. http://www.bisnode.com/en/About-Bisnode/

Black JL, Macinko J, Dixon LB, Fryer GE Jr. Neighborhoods and obesity in New York City. Health & Place. 2010; 16:489–499. [PubMed: 20106710]

Block D, Kouba J. A comparison of the availability and affordability of a market basket in two communities in the Chicago area. Public Health Nutr. 2006; 9:837–845. [PubMed: 17010248]

Chung C, Myers SL. Do the Poor Pay More for Food? An Analysis of Grocery Store Availability and Food Price Disparities. Journal of Consumer Affairs. 1999; 33:276–296.

D&B. D&B WorldBase File. Dun & Bradstreet; 2011a. http://mddi.dnb.com/mddi/worldbase_file.aspx

D&B. Dun & Bradstreet business database. Dun & Bradstreet, Inc; 2011b. http://www.dnb.com/about-dnb/14881789-1.html

D&B. What's in a business credit profile?. Dun & Bradstreet; 2011c. https://smallbusiness.dnb.com/business-finance/business-loans-business-credit/12160-1.html

Ellickson PB, Misra S. Supermarket Pricing Strategies. Marketing Science. 2008; 27:811–828.

Evans DS. Tests of Alternative Theories of Firm Growth. Journal of Political Economy. 1987; 95:657–674.

Forsyth, A.; Lytle, L.; Riper, DV. Finding food: Issues and challenges in using Geographic Information Systems (GIS) to measure food access. 2010.

Franco M, Diez-Roux AV, Nettleton JA, Lazo M, Brancati F, Caballero B, Glass T, Moore LV. Availability of healthy foods and dietary patterns: the Multi-Ethnic Study of Atherosclerosis. Am J Clin Nutr. 2009; 89:897–904. [PubMed: 19144728]

Franco M, Diez Roux AV, Glass TA, Caballero B, Brancati FL. Neighborhood characteristics and availability of healthy foods in Baltimore. Am J Prev Med. 2008; 35:561–567. [PubMed: 18842389]

Gershteyn, Y. Use of SPEDIS Function in Finding Specific Values; SAS Global Users Group Conference; Indianapolis, IN: SAS Institute Inc; 2000. [Paper 86-25] http://www2.sas.com/proceedings/sugi25/25/cc/25p086.pdf

Glanz K. Measuring food environments: a historical perspective. Am J Prev Med. 2009; 36:S93–98. [PubMed: 19285215]

Hoehner CM, Schootman M. Concordance of commercial data sources for neighborhood-effects studies. J Urban Health. 2010; 87:713–725. [PubMed: 20480397]

InfoUSA. InfoUSA Database. InfoUSA.com; Omaha, NE: 2011.

Jargowsky, PA. Poverty and Place: Ghettos, Barrios, and the American City. Russell Sage Foundation; New York: 1997.

Kowaleski-Jones, L.; Fan, JX.; Yamada, I.; Zick, C.; Smith, K.; Brown, B. Alternative Measures of Food Deserts: Fruitful Options or Empty Cupboards?. National Poverty Center Working Paper. 2009. http://www.npc.umich.edu/news/events/food-access/kowaleski-jones_et_al.pdf

Larson NI, Story MT, Nelson MC. Neighborhood environments: disparities in access to healthy foods in the U.S. Am J Prev Med. 2009; 36:74–81. [PubMed: 18977112]

Li F, Harmer PA, Cardinal BJ, Bosworth M, Acock A, Johnson-Shelton D, Moore JM. Built environment, adiposity, and physical activity in adults aged 50–75. Am J Prev Med. 2008; 35:38–46. [PubMed: 18541175]

Liese AD, Colabianchi N, Lamichhane AP, Barnes TL, Hibbert JD, Porter DE, Nichols MD, Lawson AB. Validation of 3 Food Outlet Databases: Completeness and Geospatial Accuracy in Rural and Urban Food Environments. Am J Epidemiol. 2010; 172:1324–1333. [PubMed: 20961970]

Macdonald L, Ellaway A, Ball K, Macintyre S. Is proximity to a food retail store associated with diet and BMI in Glasgow, Scotland? Bmc Public Health. 2011:11. [PubMed: 21208451]

Merrefield, D. Supermarket News. Penton Media, Inc; 1998. Defining independents and chains.

Michimi A, Wimberly MC. Associations of supermarket accessibility with obesity and fruit and vegetable consumption in the conterminous United States. International Journal of Health Geographics. 2010:9. [PubMed: 20156361]

Moore LV, Diez Roux AV. Associations of neighborhood characteristics with the location and type of food stores. Am J Public Health. 2006; 96:325–331. [PubMed: 16380567]

Moore LV, Diez Roux AV, Brines S. Comparing Perception-Based and Geographic Information System (GIS)-based characterizations of the local food environment. J Urban Health. 2008a; 85:206–216. [PubMed: 18247121]

Moore LV, Diez Roux AV, Nettleton JA, Jacobs DR Jr. Associations of the local food environment with diet quality--a comparison of assessments based on surveys and geographic information systems: the multi-ethnic study of atherosclerosis. Am J Epidemiol. 2008b; 167:917–924. [PubMed: 18304960]

Morland K, Wing S, Diez Roux A, Poole C. Neighborhood characteristics associated with the location of food stores and food service places. Am J Prev Med. 2002; 22:23–29. [PubMed: 11777675]

Neckerman KM, Bader MD, Richards CA, Purciel M, Quinn JW, Thomas JS, Warbelow C, Weiss CC, Lovasi GS, Rundle A. Disparities in the Food Environments of New York City Public Schools. Am J Prev Med. 2010; 39:195–202. [PubMed: 20709250]

Nielsen Company. Retail Site Database, The Ultimate Source Retail Site Database - The Ultimate Source Trade Dimensions, a subsidiary of Nielsen Company. 2008.

Pearce J, Day P, Witten K. Neighbourhood provision of food and alcohol retailing and social deprivation in urban New Zealand. Urban Policy and Research. 2008; 26:213–227.

Powell LM, Auld MC, Chaloupka FJ, O'Malley PM, Johnston LD. Associations between access to food stores and adolescent body mass index. Am J Prev Med. 2007a; 33:S301–S307. [PubMed: 17884578]

Powell LM, Han E, Zenk SN, Khan T, Quinn CM, Gibbs KP, Pugach O, Barker DC, Resnick EA, Myllyluoma J, Chaloupka FJ. Field validation of secondary commercial data sources on the retail food outlet environment in the U.S. Health Place. 2011; 17:1122–1131. [PubMed: 21741875]

Powell LM, Slater S, Mirtcheva D, Bao Y, Chaloupka FJ. Food store availability and neighborhood characteristics in the United States. Prev Med. 2007b; 44:189–195. [PubMed: 16997358]

Progressive Grocer. Progressive Grocer. Stagnito Media; 2011. Crunching the Numbers.

Roesch, A. Matching Data Using Sounds-Like Operators and SAS® Compare Functions; Northeast SAS Users Group (NESUG) Conference; Portland, ME: SAS Institute Inc; 2011. www.nesug.org/Proceedings/nesug11/ap/ap07.pdf

Rundle A, Neckerman KM, Freeman L, Lovasi GS, Purciel M, Quinn J, Richards C, Sircar N, Weiss C. Neighborhood food environment and walkability predict obesity in New York City. Environ Health Perspect. 2009; 117:442–447. [PubMed: 19337520]

Smiley MJ, Diez Roux AV, Brines SJ, Brown DG, Evenson KR, Rodriguez DA. A spatial analysis of health-related resources in three diverse metropolitan areas. Health & Place. 2010; 16:885–892. [PubMed: 20478737]

Story M, Giles-Corti B, Yaroch AL, Cummins S, Frank LD, Huang TTK, Lewis LB. Work Group IV: Future Directions for Measures of the Food and Physical Activity Environments. American Journal of Preventive Medicine. 2009; 36:S182–S188. [PubMed: 19285212]

Tele Atlas. USA_Geo_002 [documentation for Tele Atlas products using Dynamap line files]. Tele Atlas North America, Inc; Lebanon, NH: 2011. http://www.geocode.com/documentation/USA_Geo_002.pdf

US Census Bureau. Collectibility of Data. U.S. Census Bureau, Economic Classification Policy Committee; 1993. [Issues Paper No. 3] http://www.census.gov/epcd/naics/issues3

US Census Bureau. American Community Survey 5-year small area estimates 2005–2009. U.S. Census Bureau; 2010. http://www.census.gov/acs/www/data_documentation/data_main/

US Census Bureau. North American Industry Classification System, Frequently Asked Questions (FAQs). U.S. Census Bureau; 2011a. http://www.census.gov/eos/www/naics/faqs/faqs.html

US Census Bureau. Standard Industrial Classification (SIC) System. U.S. Census Bureau; 2011b. http://www.census.gov/epcd/www/sic.html

Bishaw, A. US Census Bureau. American Community Survey Briefs. U.S. Census Bureau; 2011. Areas With Concentrated Poverty: 2006–2010. http://www.census.gov/prod/2011pubs/acsbr10-17.pdf

Ver Ploeg, M.; Breneman, V.; Farrigan, T.; Hamrick, K.; Hopkins, D.; Kaufman, P.; Lin, B.; Nord, M.; Smith, T.; Williams, R.; Kinnison, K.; Olander, C.; Singh, A.; Tuckermanty, E. Access to Affordable and Nutritious Food—Measuring and Understanding Food Deserts and Their Consequences: Report to Congress. United States Department of Agriculture; 2009.

Walls & Associates. National Establishment Time-Series (NETS) Database: Database Description. 2010. www.youreconomy.org/nets/NETSDatabaseDescription.pdf

Wang MC, Gonzalez AA, Ritchie LD, Winkleby MA. The neighborhood food environment: sources of historical data on retail food stores. Int J Behav Nutr Phys Act. 2006; 3:15. [PubMed: 16846518]

Wang MC, Kim S, Gonzalez AA, MacLeod KE, Winkleby MA. Socioeconomic and food-related physical characteristics of the neighbourhood environment are associated with body mass index. J Epidemiol Community Health. 2007; 61:491–498. [PubMed: 17496257]

Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH. Match rate and positional accuracy of two geocoding methods for epidemiologic research. Ann Epidemiol. 2006; 16:842–849. [PubMed: 17027286]

Highlights

- Methods identify supermarkets in a single historic year over a large area of the USA.
- Agreement between two food store databases was modest: 63%.
- 55% of the final list of supermarkets were identified by a single business database.
- The benefit of using multiple datasets may be even greater in poor areas

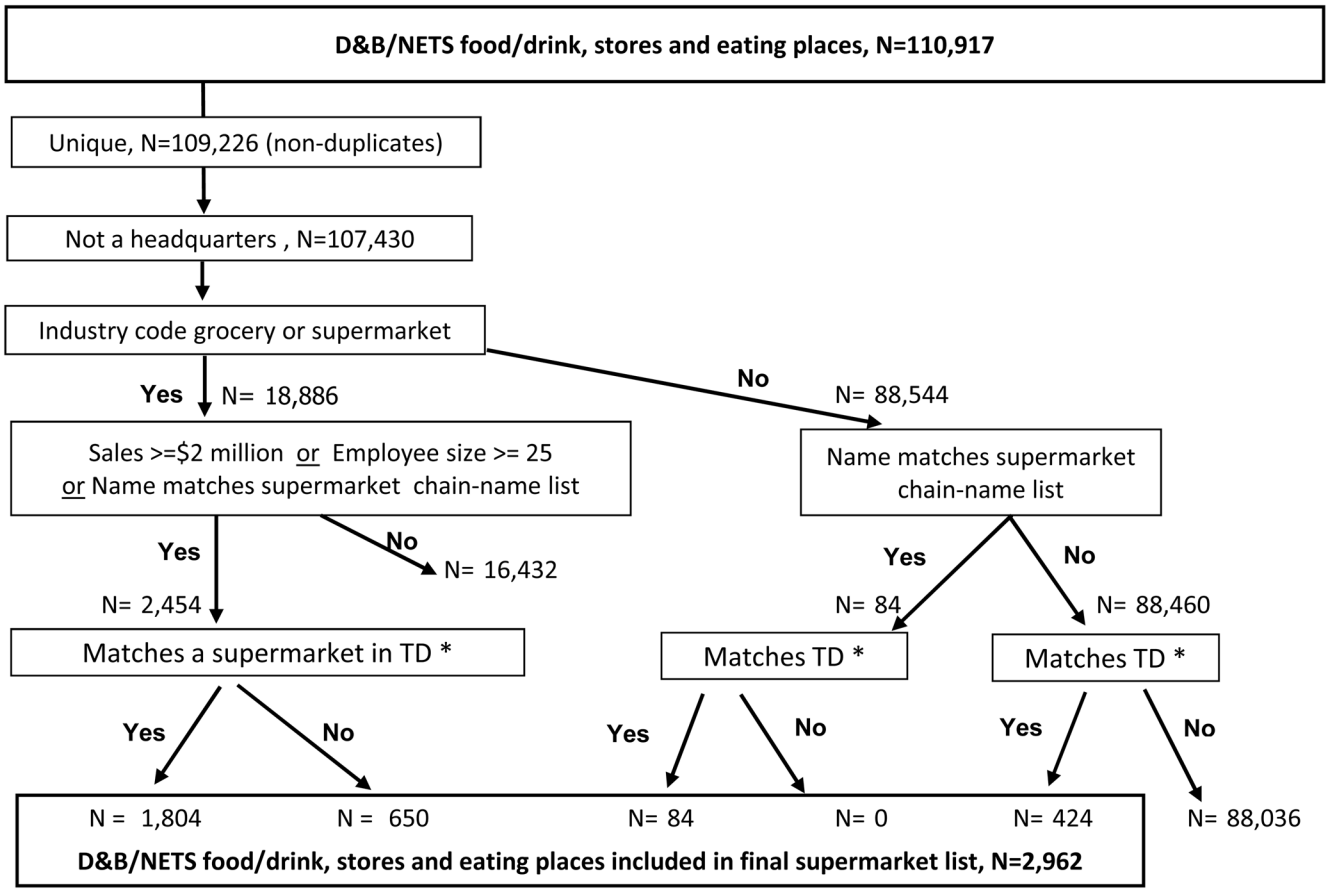

Figure 1.
Decision flow for Dun and Bradstreet food establishment archival data, National Establishment Time Series, year 2005
* Matches a supermarket in Trade Dimensions (TD)

Health Place. Author manuscript; available in PMC 2013 November 01.

Table 1

Name and location matching of 3002 records from Trade Dimensions (TD) supermarkets and 2962 records from Dun and Bradstreet food establishment archival data (National Establishment Time Series, NETS); 2005.

	Name Match *					
	Excellent	**Good**	**Fair**	**Does not match**		
Location match **						
Excellent	1250	143	178	0	1571	43%
Good	573	52	61	0	686	19%
Fair	47	8	0	0	55	2%
Does not match						
TD (db1) does not match any address in NETS (db2)	0	0	0	690	690	19%
NETS (db2) does not match any address in TD (db1)	0	0	0	650	650	18%
	1870	203	239	1340	3652	100%
	51%	6%	7%	37%	100%	

* The SAS SPEDIS function (standing for Spelling Distance) was applied to standardized addresses (Gershteyn, 2000 and Roesch, 2011).

"Excellent" matches had spelling match score 0<=15;

"Good" had spelling match score 15-<45 (were manually inspected and retained only if name was same despite spelling differences);

"Fair" had spelling match score => 45 (used only if the latitude/longitude matched and if the NETS record had SIC code of 5411).

** "Excellent" location matches had exact match on all of the following: ZIP code, building number and street address;

"Good" location matches had exact match on ZIP code and <0.001 miles (<1.6 meters) between latitude/longitude;

"Fair" location matches had perfect ZIP code match and <0.30 miles (<483 meters) between latitude/longitude. All these matches were manually inspected. If the suite number was specified and differed then they were not considered a match.

Table 2

Final list of supermarkets compiled using Trade Dimensions (TD), and Dun and Bradstreet food establishment archival data (National Establish Time Series, NETS).

	N	%
Final number of supermarkets	3652	100%
NETS record in grocery/supermarket SIC category and store has >1 other supermarket attribute (sales, employees, chain-name)	2454	67%
Sales >= $2million	2024	80%
Employees >= 25 (in addition to above criteria)	17	1%
Store name is on the *supermarket chain-name list* * *(in addition to above criteria)*	413	16%
NETS record in not in the grocery/supermarket SIC category (in convenience, deli, miscellaneous foods stores, eating places, liquor stores)	508	14%
Name was on the *supermarket chain-name list*	84	2%
Name and/or location mached ** TD supermarket	424	12%
Not in NETS but in TD supermarkets database	690	19%

* See 161 chain name supermarkets, Supplementary Table Chain Name List

** See matching details in Table 2.

Table 3

Census tract characteristics for the store datasets. Data from year 2005 Trade Dimensions (TD) and Dun and Bradstreet food establishment archival data (National Establishment Time Series, NETS), and 2005–2009 American Community Survey.

	Census tract-level		Store-level					
	Census tracts for all stores*		All stores *		TD		NETS	
	N	% of tracts	N	% of all stores	N	% of all stores	N	% of all stores
Total N	7008	100%	3652	100%	3002	82%	2538	69%
High poverty **								
Yes	916	13%	391	11%	305	10%	239	9%
No	6003	86%	3252	89%	2690	90%	2292	90%
Missing	89	1%	9	0%	7	0%	7	0%
Mid-high population density **								
Yes	2743	39%	1382	38%	1110	37%	879	35%
No	4136	59%	2256	62%	1882	63%	1649	65%
Missing	129	2%	14	0%	10	0%	10	0%
High proportion of minority residents **								
Yes	3679	52%	1771	48%	1441	48%	1164	46%
No	3254	46%	1873	51%	1554	52%	1368	54%
Missing	75	1%	8	0%	7	0%	6	0%

* "All stores" represents a combination of TD and NETS as shown in Table 3--4.

** High poverty area was defined as where >=30% of households in a tract have incomes under the poverty threshold, mid-high population density was defined as >=1000 persons per square mile, and high proportion of minority residents was defined as <40% of residents in the tract are non-Hispanic White.