Published in final edited form as:
J Nutr Educ Behav. 2013 ; 45(5): 435–442. doi:10.1016/j.jneb.2013.01.021.

Characterizing the Food Retail Environment: Impact of Count, Type and Geospatial Error in Two Secondary Data Sources

Angela D. Liese, PhD, MPH[1], **Timothy L. Barnes, MPH**[1], **Archana P. Lamichhane, PhD**[1], **James D. Hibbert, MS**[1], **Natalie Colabianchi, PhD**[2], and **Andrew B. Lawson, PhD**[3]

[1]Center for Research in Nutrition and Health Disparities and Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA

[2]Institute of Social Research, University of Michigan, Ann Arbor, MI, USA

[3]College of Medicine, Medical University of South Carolina, Charleston, SC, USA

Abstract

Objective—Commercial listings of food retail outlets are increasingly used by community members, food policy councils, and in multi-level intervention research to identify areas with limited access to healthier food. This study quantified the amount of count, type and geospatial error in two commercial data sources.

Methods—InfoUSA and Dun & Bradstreet (D&B) were compared to a validated field census and validity statistics calculated.

Results—Considering only completeness, D&B data undercounted 24% of existing supermarkets and grocery stores and InfoUSA 29%. Additionally, considering accuracy of outlet type assignment increased the undercount error to 42% and 39%, respectively. Marked overcount existed as well and only 43% of existing supermarkets were correctly identified with respect to presence, outlet type, and location.

Conclusions and Implications—Relying exclusively on secondary data to characterize the food environment will result in substantial error. While extensive data cleaning can offset some error, verification of outlets with a field census is still the method of choice.

Keywords

retail food environment; secondary data sources; validity; geography

INTRODUCTION

Access to healthier food retailers is a topic of public health and political interest. Over the past decade, an increasing number of studies have characterized the food environment and evaluated its influence on health behaviors and health outcomes.[1–4] Via the Food, Conservation, and Energy Act of 2008, the United States (US) Congress directed the US

Corresponding Author: Angela D. Liese, PhD, MPH, Center for Research in Nutrition and Health Disparities, Arnold School of Public Health, University of South Carolina, 921 Assembly Street, Columbia, SC 29208, USA, Phone: (803) 777-9414; Fax: (803) 777-2504; liese@sc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Department of Agriculture (USDA) "to assess the extent of areas with limited access to affordable and nutritious food, identify characteristics and causes of such areas, consider how limited access affects local populations, and outline recommendations to address the problem."[5] Since then, a variety of approaches to the identification of so called "food deserts" or, conversely, environments supporting healthy food choices have been proposed.[6–9] Interactive websites, such as the USDA Food Environment Atlas and the Food Desert Locator, provide geographic information on food access and the spatial distribution of food retailers.[10–12]

Local food policy councils are increasingly advocating for improvements in food access, including spatial access to healthier retail outlets. Furthermore, multi-level nutrition interventions frequently entail an assessment of and changes to the retail food environment. In response, a number of toolkits have been developed that assist community members in mapping and evaluating their local retail food environment.[13, 14]

Inherent in the aforementioned efforts is the need to identify specific types of retail outlets, such as supermarkets or grocery stores. Government reports and websites have been based on readily available commercial (e.g. Dun & Bradstreet, InfoUSA) or public secondary data. Most commercial databases include an outlet type designation such as the North American Industry Classification System (NAICS) code or Standard Industry Codes and consider their code assignments proprietary.[15]

This group of investigators has previously explored the completeness of several secondary databases' listings of food retail outlets, noting marked overcount and undercount of outlets.[16] At that time, the assignment to outlet type categories was based on a research-intense approach, not an automated algorithm that utilized the NAICS codes contained within the databases. Because national policies on spatial food access are largely directed at specific food outlet types and based on secondary data without further validation, this study extends research to a comprehensive evaluation of the validity of Dun & Bradstreet and InfoUSA data. The purpose of the present study was to quantify sequentially the impact of errors due to the number of food retails outlets (count), type of retail outlet, and errors in location (geospatial error) in these two secondary data sources by comparison to a field census of food outlets that was validated in person for both location and type. Additionally, this study explored whether the errors differed across a spectrum of Census tract demographic and socioeconomic characteristics, because this type of differential misclassification could potentially lead to biases in etiologic research and undermine the identification of neighborhoods which are particularly disadvantaged with respect to their food environment.[17–19]

METHODS

This study was part of a larger effort aimed at developing spatial accessibility measures of the built food environment for urban and rural areas in South Carolina.[16] The study region consisted of a geographically contiguous area of 5,575 square miles, including one urban county and seven rural counties.

Field Census of Food Outlets

In preparation for the field census, i.e. direct observation and verification of all food outlets, data from Dun & Bradstreet, InfoUSA and the Licensed Food Services Facilities Database from the South Carolina Department of Health and Environmental Control had been obtained and were utilized to generate a comprehensive master listing (Figure 1, step 1).[16] Duplicate entries and food outlets that were ineligible had been removed prior to merging the three data sources into a single file by name and address. Certain types of food outlets were excluded, such as those only sporadically open, food outlets that serve special populations such as school cafeterias or cafeterias in nursing homes, assisted living facilities, or institutionalized settings, military settings, food preparation facilities for catering businesses that have no publicly accessible retail store, alcoholic beverage drinking places, and liquor stores.

The fieldwork was conducted by six persons who were trained under a standardized protocol; they took 114 trips which covered nearly 7,000 miles (Figure 1, step 2). Counties were treated individually in the field census and trips varied from two per county (Calhoun County) to 27 (Richland County). The fieldwork began in September 2008 and concluded in July 2009. The location (latitude and longitude) was recorded using a global positioning system (GPS, Trimble Juno ST GPS; 3–5 m spatial accuracy; Trimble Navigation Ltd., Sunnyvale, California and Arc-Pad 7.1 software, ESRI, Redlands, California).

Name-Based Outlet Type Assignment

To assign each food outlet to a retail type, a name-based approach was developed (Figure 1, step 3).[16] An algorithm was programmed in SAS 9.2 (released 2008, Cary, NC, USA) which assigned outlets based on their business name into one of 14 specific outlet types (supermarkets, supercenters, grocery stores, warehouse clubs, convenience stores, dollar/variety stores, drug/pharmacy, meat market, seafood market, green grocers, bakeries, confectionary stores, full service restaurants including cafeterias and limited service restaurants). For instance, “McDonald’s”, “Burger King” etc., were assigned the outlet type “limited service restaurants”;“Kroger”, “Publix”, “Bi-Lo”, “Piggly Wiggly” etc., were assigned to the outlet type “supermarkets”. This assignment was conducted entirely independent of existing NAICS codes. Any unassigned outlets were reviewed by team members to manually identify and assign the outlet type. For all outlets that still could not be assigned with certainty, additional internet research was conducted and the outlet facilities were called to self-identify. For outlets newly discovered, the outlet types were assigned during the verification effort in the field census.

Geospatial Analyses and Assignment of Census Tract Characteristics

All geospatial analyses were conducted within ArcGIS software (Version 9.3, ESRI, Redlands, CA) using TIGER 2008 street network data.[20] The GPS location taken during the field census served as the reference for the geospatial analyses, against which the commercial geocodes listed in Dun & Bradstreet or InfoUSA were compared.

US Census tract data on demographic and socioeconomic characteristics were obtained from Summary Files 1 and 3 from the U.S. Bureau of the Census for 2000.[16] Characteristics included racial composition (majority white, majority black, and mixed defined as greater than 60% of one race or else coded as mixed), median household income (categorized into tertiles <$28,829 (low), $28,829–$36,875 (mid), >$36,875 (high)) and poverty status (dichotomized as greater than 20% population below the 2000 poverty level into poor and non-poor).[21, 22] Outlets identified by the field census and outlets identified in either secondary data source were assigned to a Census tract whereby outlets inherited the attributes of their respective tracts based on their relative locations (point-in-polygon join).

Statistical Analyses

Both Dun & Bradstreet and InfoUSA contain a large amount of characteristics on each listed retail outlet. Geo-coordinates (latitude, longitude) and NAICS codes had been retained (Figure 1, step 1). For the purpose of this analysis, the NAICS codes were used to assign each listed outlet to an outlet type in preparation of the statistical analyses (Table 1). This process was automated in SAS using only the first NAICS code (primary code) as the basis of food outlet type assignment, similar to previous work.[19] About 2% of outlets in Dun & Bradstreet and InfoUSA had NAICS codes that did not match with the any of the codes outlined and were dropped from the analyses because they could not be assigned a specific outlet type.

For our analyses on the accuracy of food outlet listings, their type and their location in the two secondary data sources (Dun & Bradstreet, InfoUSA) (Figure 1, step 4), the data from the field census served as the comparison reference. The analyses were conducted in distinct steps: First, validity statistics were calculated considering exclusively the completeness of each data source (i.e. count accuracy).The results shown are virtually identical to those published previously,[16] the small differences being due to minor corrections after publication. Next, both count and type assignment accuracy were considered. Sensitivity (the fraction of truly existing food outlets that was captured in a given secondary data source) and positive predicted values (PPV, the fraction of the food outlets listed in a commercial data source that was found to be open) were calculated and 95% confidence intervals (CI) estimated for each of these proportions by approximating the binomial distribution with a normal distribution. Both the undercount (100 minus sensitivity) and overcount (100 minus PPV) were calculated. To explore the impact of neighborhood characteristics on validity, the analyses for the supermarket and grocery store category were repeated, stratifying by levels of three Census-based characteristics: race (White, Black, Mixed), income (low, mid, high), and poverty level (poor, not poor). Fisher's exact tests were used to test differences in validity between the levels. Finally, the evaluation was extended to include consideration of geospatial accuracy in addition to type and count accuracy. The geospatial position listed in the Dun & Bradstreet and InfoUSA data was considered accurate if it was situated within 0.5 miles of their Census tract location determined in our field census, to be consistent with several policy-level food access.[8, 23] Statistical analyses were conducted using SAS software (version 9.2; released 2008, SAS Institute, Inc., Cary, North Carolina).

RESULTS

The field census identified a total of 1,697 food outlets in the categories shown in Table 2, Dun & Bradstreet identified 1,448 outlets, and InfoUSA identified 1,583 outlets in the listed categories.

Compared to consideration of only the count error, additional consideration of type assignment errors resulted in poorer validity statistics for all food outlet types (Table 3). For instance, for supermarkets and grocery stores, the sensitivity was reduced from 0.76 to 0.58 in Dun & Bradstreet and from 0.71 to 0.61 in InfoUSA. This corresponds to a notable undercount of existing supermarkets and grocery stores (42% in Dun & Bradstreet and 39% in InfoUSA) if relying exclusively on NAICS codes. Specifically, incorrect designations of supermarkets and grocery stores as dollar stores, convenience stores, and pharmacies by Dun & Bradstreet and as department stores, convenience stores and a few other store types by InfoUSA were observed. The PPV statistics were reduced from 0.73 to 0.39 for Dun & Bradstreet and from 0.83 to 0.57 for InfoUSA for supermarkets and grocery stores when relying exclusively on NAICS codes. This corresponds to a marked over-assignment (overcount) of outlets into the supermarket and grocery store category (61% in Dun & Bradstreet and 43% in InfoUSA) when relying on NAICS codes. Specifically, incorrect designations of dollar stores and convenience stores (in Dun & Bradstreet) and convenience stores (in InfoUSA) into the supermarkets and grocery store category were observed.

With respect to full service restaurants, InfoUSA data exhibited a moderate undercount of restaurants (35–38%), performing better than Dun & Bradstreet (65% undercount). Limited service restaurants, however, were virtually not identifiable using NAICS codes in InfoUSA, with an undercount of 92%, because many existing limited service restaurants were designated incorrectly as full service restaurants.

As shown in Table 4, for Dun & Bradstreet, no evidence for systematic differences in the validity statistics of supermarkets and grocery stores were found between levels of neighborhood characteristics. For InfoUSA, a significantly higher undercount of supermarkets and grocery stores in predominantly white neighborhoods compared to predominantly black neighborhoods (46% vs. 25%), high and mid-income compared to low income neighborhoods (41% vs. 18%) and non-poor compared to poor neighborhoods (45% vs. 22%) was observed. Overcount error also varied across neighborhood socioeconomic characteristics in InfoUSA, with higher error in the mid and high-income neighborhoods compared to the low income neighborhoods (43% vs. 23%).

Evaluating count and type accuracy and geospatial accuracy (Table 5) revealed that of the supermarkets and grocery stores existing in the study area, only 43% were listed and located accurately in both Dun & Bradstreet and InfoUSA, i.e. were listed with the correct outlet type and at a location within 0.5 miles of their actual Census tract. All other outlet types faired worse. As shown in Table 4, both Dun & Bradstreet and InfoUSA data exhibited about 57% undercount of supermarkets and grocery stores, and even higher levels for other outlet types.

DISCUSSION

Of the many types of food retail outlets, this paper focuses on the subset that are directly relevant to current policy-level indicators of community food access.[6–8] Consistent with the report by Powell et al.,[19] this study found that validity statistics for supermarkets and grocery stores outranked convenience stores, specialty stores, and full and limited service restaurants. Sensitivity estimates reported for supermarkets and grocery stores by Powell et al. (0.62 in Dun & Bradstreet and 0.74 for InfoUSA) and by Lisabeth et al. (0.85 for ReferenceUSA, which is part of the InfoUSA product lines) were very similar to ours when using a research intense, name-based typing approach and focusing only on count accuracy, i.e. ignoring any potential inaccuracies of the NAICS codes.[16, 19, 23] This suggests that of the various outlet types, supermarkets and large grocery stores can be reasonably accurately identified in both Dun & Bradstreet and InfoUSA data, especially if using a name-based typing approach similar to this and other studies.[16, 23] Other outlet types, including convenience stores, small grocery stores, and fast food restaurants which may be used in the characterization of non-health-promoting environments[8] will likely harbor substantially more error.

This study revealed a number of NAICS code assignment issues specific to each of the secondary data sources. In Dun & Bradstreet and InfoUSA, convenience stores were often incorrectly designated as supermarkets and grocery stores. This practice resulted in a very poor overall PPV (0.39 for Dun & Bradstreet, 0.57 for InfoUSA) for the supermarket and grocery store category when relying on NAICS codes. Furthermore, both our study and that of Powell et al.[19] suggest that InfoUSA systematically assigns the full service restaurant NAICS code to traditional fast food restaurants, thereby resulting in a low PPV for full service restaurants (0.45) and a low sensitivity for limited service restaurants (0.08, undercount 92%). Taken together, these results suggest relying exclusively on NAICS codes for assignment of outlet types will result in a large number of systematic errors.

Previous research had already pointed toward one specific NAICS coding issue related to the convenience store category, which could have resulted in markedly higher underascertainment of that cateogry.[19, 24] However, by including not only one but three NAICS codes for the convenience store category, i.e. additionally including gas stations with convenience stores and so called “other gas stations” in our data request, we were able to prevent additional errors.

This study also sheds light on the combined impact of count, type and geospatial error. The Census tract plus a 0.5 mile buffer was chosen as a relevant geography to be consistent with the geographies of the CDC’s healthier retail tract indicator and the modified retail environment index and other policy indicators.[6, 8] The findings have the following implications: For instance, if one were to rely exclusively on information contained in either Dun & Bradstreet or InfoUSA, only 43% of existing supermarket or grocery stores would actually have been found in the database with accurate outlet type and geo-coordinates that placed them in the correct Census tract including a 0.5 mile buffer. This implies an undercount in the databases of 57%. What impact these inaccuracies would have on Census tract-based policy indicators remains to be evaluated.

Assuming one had to rely on a single data source, the results suggest that – due to lower count, type and geospatial errors – there may be a small advantage of using InfoUSA data for the identification of supermarkets and grocery stores, convenience stores, specialty stores, and full service restaurants. For the identification of limited service restaurants, Dun & Bradstreet clearly outranks InfoUSA. This study furthermore suggests that review and re-assignment of NAICS code assignments or use of a name-based assignment method can reduce the amount of type error. Some differences in accuracy associated with neighborhood characteristics for the supermarket and grocery store category in the InfoUSA but not the Dun & Bradstreet data were observed. To date, the literature has been very inconsistent with respect to the presence and direction of differences in the accuracy of data sources across levels of neighborhood characteristics.[17–19, 23] Thus, this database characteristic may not be a distinguishing feature. In practice, other considerations may also be important for the choice of database, such as availability of archived data for years past, availability or completeness of other store attributes such as employee number. or cost considerations.

There are several limitations and strengths to this study. Unlike other studies,[19] in this study the field census teams did not enter the stores to conduct an independent, objective assessment of the outlet type. Instead, information embedded in the outlet name, common knowledge of large franchised food outlets, and internet and phone research was used, similar to the process outlined by Lisabeth et al..[23] Secondly, the possibility that some food outlets discovered may have been listed in a secondary data source, but under a NAICS code that we did not request (e.g. 446191 Food/health supplement stores) cannot be ruled out. Furthermore, the analysis was based exclusively on each outlet’s primary NAICS code, because that is thought to be the overarching type designation. Lastly, while the field census attempted to be comprehensive, the possibility that some outlets were overlooked cannot be excluded.

The strengths of the study include that validity statistics were presented based on two distinct approaches that may be conceptualized as different levels of data cleaning. The NAICS code-based approach to type assignment probably represents a process that will be needed for national or large-scale studies. The name-based approach represents a level of data cleaning possible for smaller scale, in-depth studies.[16, 23] Some partial assignments based on names have also been employed by previous efforts.[19, 24] Furthermore, this study included geospatial information which was obtained during the field census using GPS and thus was able to establish the location using this objective method, which additionally allowed an evaluation of the joint impact of count, type and geospatial inaccuracies.

IMPLICATIONS FOR RESEARCH AND PRACTICE

Local communities, state, and federal agencies are increasingly interested in assessing and improving the food retail environment and are frequently using commercially available, secondary data sources in these efforts. This study suggests that relying exclusively on available NAICS codes and geocodes in secondary data sources for the assignment of food outlets to type categories and Census tracts will result in substantial error. Extensive data cleaning efforts are essential for the reliable use of the two secondary data sources evaluated here, for both research and practice efforts. Ideally, however, a field census should be conducted to verify the presence of food outlets. Undoubtedly, the errors contained within secondary data sources will also affect policy-level food environment indicators, such as the measures of community food access proposed by the USDA and other agencies,[6, 8, 9] as these are typically entirely dependent on secondary data.

Acknowledgments

This project was supported by grant R21CA132133 from the National Cancer Institute. The authors thank Denise M. Hodo, Kristopher Corwin, and Dr. Andrey Bortsov for conducting the fieldwork, Michele Nichols for data analyses, and Xiaoguang Ma for assistance with manuscript preparation. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

1. McKinnon RA, Reedy J, Morrissette MA, Lytle LA, Yaroch AL. Measures of the food environment A compilation of the literature, 1990–2007. Am J Prev Med. 2009; 36(4S):124–33.

2. Larson N, Story M. A review of environmental influences on food choices. Ann Behav Med. 2009; 38:56–73.

3. Beaulac J, Kristjansson E, Cummins S. A systematic review of food deserts, 1966–2007. Prev Chronic Dis. 2009; 6(3):A105. [PubMed: 19527577]

4. Walker RE, Keane CR, Burke JG. Disparities and access to healthy food in the United States: A review of food deserts literature. Health Place. 2010; 16(5):876–84. [PubMed: 20462784]

5. Ver Ploeg, M.; Breneman, V.; Farrigan, T.; Hamrick, K.; Hopkins, D.; Kaufman, P. Access to Affordable and Nutritious Food—Measuring and Understanding Food Deserts and Their Consequences: Report to Congress. 2011. Report No.: AP-036

6. Centers of Disease Control and Prevention. [Accessed October 25, 2010] State Indicator Report on Fruits and Vegetables. 2009. http://www.fruitsandveggiesmatter.gov/downloads/StateIndicatorReport2009.pdf

7. Health Food Financing Initiative. US Department of Health and Human Services; 2010. http://www.hhs.gov/news/press/2010pres/02/20100219a.html [Accessed April 27, 2011]

8. Centers for Disease and Control and Prevention. Children's Food Environment State Indicator Report. Department of Health and Human Services; 2011. http://www.cdc.gov/obesity/downloads/ChildrensFoodEnvironment.pdf [Accessed March 28, 2012]

9. Califano, C. [Accessed March 28, 2012] Estimating Supermarket Access: Summary of TRF's Research and Analysis. 2009. http://www.trfund.com/financing/Healthy_food/EstimatingSupermarketAccess-1pg.pdf

10. US Department of Agriculture ERS. USDA Food Desert Locator. US Department of Agriculture, Economic Research Service. [Accessed July 22, 2011] 2011. http://www.ers.usda.gov/data/fooddesert

11. US Department of Agriculture ERS. USDA Food Environment Atlas. US Department of Agriculture, Economic Research Service. [Accessed July 22, 2011] 2011. http://www.ers.usda.gov/FoodAtlas/

12. The Reinvestment Fund. TRF Limited Supermarket Access Area Widget. The Reinvestment Fund; 2012. http://www.trfund.com/TRF-LSA-widget.html [Accessed December 18, 2012]

13. Centers for Disease Control and Prevention. Healthier Food Retail: Beginning the Assessment Process in Your State or Community. Centers for Disease Control and Prevention (CDC), Division of Nutrition, Physical Activity, and Obesity (DNPAO); 2011. http://www.cdc.gov/obesity/downloads/HFRassessment.pdf [Accessed November 5, 2012]

14. Nutrition Environments Measures Study. Nutrition Environment Measures Study for Stores and Restaurants. University of Pennsylvania; 2006. http://www.med.upenn.edu/nems/index.shtml [Accessed November 5, 2012]

15. North American Industry Classification System (NAICS). US Census Bureau. [Accessed March 28, 2012] http://www.census.gov/epcd/www/naics.html

16. Liese AD, Colabianchi N, Lamichhane A, Barnes TL, Hibbert JD, Porter DE, Nichols M, Lawson A. Validation of three food outlet databases: completeness and geospatial accuracy in rural and urban food environments. Am J Epidemiol. 2010; 172(11):1324–33. [PubMed: 20961970]

17. Paquet C, Daniel M, Kestens Y, Leger K, Gauvin L. Field validation of listings of food stores and commercial physical activity establishments from secondary data. Int J Behav Nutr Phys Act. 2008; 5:58. [PubMed: 19000319]

18. Bader MDM, Ailshire JA, Morenoff JD, House JS. Measurement of the local food environment: A comparison of existing data sources. Am J Epidemiol. 2010; 171(5):609–617. [PubMed: 20123688]

19. Powell LM, Han E, Zenk SN, et al. Field validation of secondary commercial data sources on the retail food outlet environment in the U.S. Health Place. 2011; 17(5):1122–31. [PubMed: 21741875]

20. 2008 TIGER/Line Shapefiles. US Census Bureau; http://www.census.gov/geo/www/tiger/ [Accessed October 7, 2011]

21. Dalaker, J. US Census Bureau. Current Population Reports, Series P60-214, Poverty in the United States: 2000. US Government Printing Office; Washington, DC: 2001.

22. Krieger N, Williams DR, Moss NE. Measuring social class in US public health research: concepts, methodologies, and guidelines. Annu Rev Public Health. 1997; 18:341–78. [PubMed: 9143723]

23. Lisabeth LD, Sanchez BN, Escobar J, Hughes R, Meurer WJ, Zuniga B, Garcia N, Brown DL, Morgenstern LB. The food environment in an urban Mexican American community. Health Place. 2010; 16(3):598–605. [PubMed: 20167528]

24. 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(1):23–9. [PubMed: 11777675]

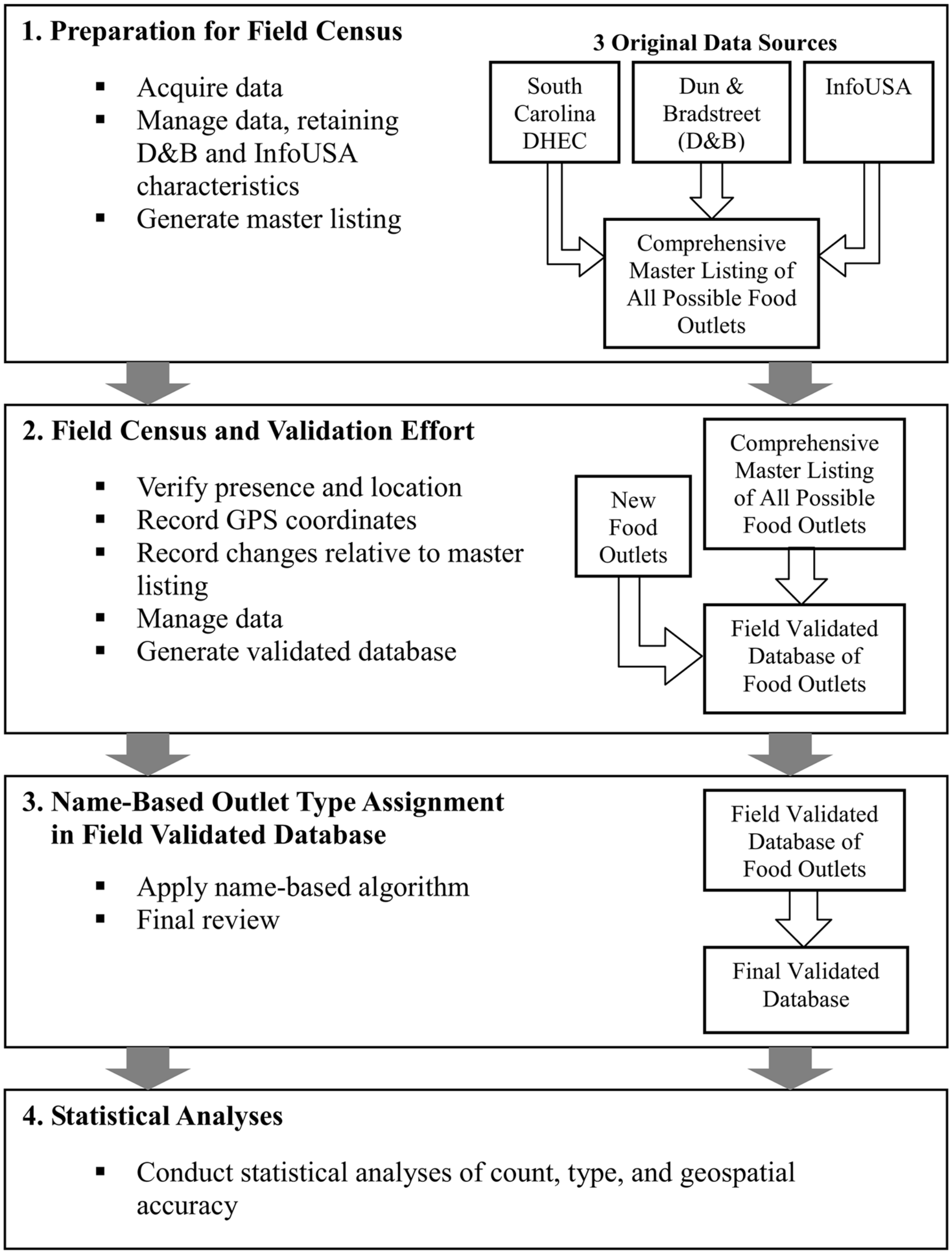
1. Preparation for Field Census
▪ Acquire data
▪ Manage data, retaining D&B and InfoUSA characteristics
▪ Generate master listing
3 Original Data Sources
South Carolina DHEC
Dun & Bradstreet (D&B)
InfoUSA
Comprehensive Master Listing of All Possible Food Outlets
2. Field Census and Validation Effort
▪ Verify presence and location
▪ Record GPS coordinates
▪ Record changes relative to master listing
▪ Manage data
▪ Generate validated database
New Food Outlets
Comprehensive Master Listing of All Possible Food Outlets
Field Validated Database of Food Outlets
3. Name-Based Outlet Type Assignment in Field Validated Database
▪ Apply name-based algorithm
▪ Final review
Field Validated Database of Food Outlets
Final Validated Database
4. Statistical Analyses
▪ Conduct statistical analyses of count, type, and geospatial accuracy

Figure 1.

Flow diagram of study methods of a South Carolina food retail environment study

Table 1

Description and classification of food outlet types based on North American Industry Classification System (NAICS) codes

Outlet Types	**Corresponding Primary NAICS Codes**
Retail Stores	
Supermarket and grocery stores (includes stores retailing a general line of food, supercenters, and warehouse clubs)	445110, 452910, 452990, 453998
Convenience stores (includes gas stations)	445120, 447110, 447190
Specialty (includes meat markets, seafood markets, green grocers, bakeries, and confectionary stores)	445210, 445220, 445230, 445291, 445292
Restaurants	
Full service restaurant (includes sit down restaurants, cafeterias, and buffets)	722110, 722212
Limited service restaurant (includes franchised and non-franchised fast food)	722211, 722213

Table 2

Distribution of store types (number and percent) according to field census and secondary data sources

Outlet Types	Field Census[a]	Dun & Bradstreet[b]	InfoUSA[b]
Supermarkets & grocery stores	160 (9.4%)	239 (16.5%)	171 (11.0%)
Convenience stores	504 (29.7%)	321 (22.2%)	400 (25.0%)
Specialty stores	36 (2.1%)	7 (0.5%)	9 (0.1%)
Full service restaurants	650 (38.3%)	402 (27.8%)	898 (57.0%)
Limited service restaurants	347 (20.4%)	409 (28.2%)	61 (4.0%)
Other[c]	--	70 (4.8%)	44 (3.0%)

[a] Name-based outlet type assignment

[b] NAICS code-based outlet type assignment

[c] Other- Primary NAICS codes that did not fit into any of the listed categories

Table 3

Impact of count and type error on validity of secondary data sources on food environment compared to the validated field census data on food environment

Outlet Types	Validity statistics	Count error considered		Count and type error considered	
		Dun & Bradstreet	InfoUSA	Dun & Bradstreet	InfoUSA
Supermarkets & grocery stores	Sensitivity (95% CI)	0.76 (0.69 – 0.82)	0.71(0.64 – 0.78)	0.58 (0.51 – 0.66)	0.61 (0.54 – 0.69)
	Undercount	24%	29%	42%	39%
	PPV (95% CI)	0.73 (0.66 – 0.80)	0.83(0.76 – 0.89)	0.39 (0.33 – 0.46)	0.57 (0.50 – 0.65)
	Overcount	27%	17%	61%	43%
Convenience stores	Sensitivity (95% CI)	0.57 (0.53 – 0.61)	0.70(0.66 – 0.74)	0.40 (0.36 – 0.45)	0.63 (0.59 – 0.68)
	Undercount	43%	30%	60%	37%
	PPV (95% CI)	0.71 (0.66 – 0.75)	0.82(0.78 – 0.86)	0.63 (0.58 – 0.69)	0.79 (0.74 – 0.83)
	Overcount	29%	18%	37%	21%
Specialty stores	Sensitivity (95% CI)	0.39 (0.23 – 0.55)	0.39 (0.23 – 0.55)	0.11 (0.01 – 0.21)	0.22 (0.09 – 0.36)
	Undercount	61%	61%	89%	78%
	PPV (95% CI)	0.88 (0.71 – 1.00)	0.93 (0.80 – 1.00)	0.57 (0.20 – 0.95)	0.80 (0.55 – 1.00)
	Overcount	12%	7%	43%	20%
Full service restaurants	Sensitivity (95% CI)	0.45(0.42 – 0.49)	0.65(0.61 – 0.68)	0.35 (0.31 – 0.39)	0.62 (0.59 – 0.66)
	Undercount	55%	35%	65%	38%
	PPV (95% CI)	0.72(0.68 – 0.76)	0.86(0.83 – 0.90)	0.56 (0.52 – 0.61)	0.45 (0.41 – 0.48)
	Overcount	28%	14%	44%	55%
Limited service restaurants	Sensitivity (95% CI)	0.55 (0.51 – 0.58)	0.71 (0.67 – 0.74)	0.41 (0.38 – 0.45)	0.08 (0.06 – 0.10)
	Undercount	45%	29%	59%	92%
	PPV (95% CI)	0.79 (0.76 – 0.83)	0.91 (0.89 – 0.94)	0.67 (0.62 – 0.71)	0.81 (0.71 – 0.91)
	Overcount	21%	9%	33%	19%

PPV: Positive predictive value

Non-overlapping confidence intervals are an indication that the accuracy statistics were significantly different

Table 4

Influence of neighborhood characteristics on validity of supermarkets and grocery stores listed in two secondary data sources, considering both count and type error

Validity statistics	Race			Income			Poverty	
	White	**Black**	**Mixed**	**Low**	**Mid**	**High**	**Poor**	**Not Poor**
				Dun & Bradstreet				
Sensitivity (95% CI)	0.57 (0.45–0.68)	0.67 (0.51–0.82)	0.55 (0.41–0.69)	0.62 (0.48–0.76)	0.53 (0.39–0.67)	0.53 (0.39–0.67)	0.65 (0.51–0.79)	0.56 (0.47–0.65)
Undercount	43%	33%	45%	38%	47%	47%	35%	44%
PPV (95% CI)	0.46 (0.35–0.56)	0.42 (0.29–0.55)	0.50 (0.37–0.63)	0.43 (0.31–0.55)	0.44 (0.31–0.56)	0.44 (0.31–0.56)	0.42 (0.30–0.53)	0.48 (0.39–0.57)
Overcount	54%	58%	50%	57%	56%	56%	58%	52%
				InfoUSA				
Sensitivity (95% CI)	**0.54 (0.43–0.65)**[a]	0.75 (0.61–0.89)	0.63 (0.49–0.76)	0.82 (0.71–0.93)	**0.59 (0.45–0.72)**[b]	**0.59 (0.45–0.72)**[c]	0.78 (0.66–0.90)	**0.55 (0.46–0.64)**[d]
Undercount	**46%**	25%	37%	18%	**41%**	**41%**	22%	**45%**
PPV (95% CI)	0.59 (0.47–0.71)	0.75 (0.61–0.89)	0.73 (0.59–0.86)	0.77 (0.65–0.89)	**0.57 (0.43–0.70)**[b]	**0.57 (0.43–0.70)**[c]	0.69 (0.56–0.82)	0.66 (0.56–0.75)
Overcount	41%	25%	27%	23%	**43%**	**43%**	31%	34%

Fisher's Exact test results

[a] White v. Black is significant;

[b] Mid v. Low is significant;

[c] High v. Low is significant;

[d] Not Poor v. Poor is significant

(Based on $p<0.05$)

Table 5

Undercount of supermarkets and grocery stores due to inaccuracies in count, type and geospatial attributes, by secondary data source

	Dun & Bradstreet		InfoUSA	
	Undercount		Undercount	
	%	95%CI	%	95%CI
Supermarkets & grocery stores	57.1	(49.5 – 64.8)	57.1	(49.5 – 64.8)
Convenience stores	67.1	(63 – 71.2)	63.7	(59.5 – 67.9)
Specialty stores	94.4	(87 – 100)	97.2	(91.9 – 100)
Full service restaurants	72.8	(69.3 – 76.2)	70.0	(66.5 – 73.5)
Limited service restaurants	67.3	(63.7 – 70.9)	97.3	(96 – 98.5)

Note: Geospatial attributes are defined as acceptable if the location fell within 0.5 miles of the actual Census tract