American Journal of Epidemiology
Vol. 172, No. 11
DOI: 10.1093/aje/kwq292
Advance Access publication:
October 20, 2010

Practice of Epidemiology

Validation of 3 Food Outlet Databases: Completeness and Geospatial Accuracy in Rural and Urban Food Environments

Angela D. Liese*, Natalie Colabianchi, Archana P. Lamichhane, Timothy L. Barnes, James D. Hibbert, Dwayne E. Porter, Michele D. Nichols, and Andrew B. Lawson

* Correspondence to Dr. Angela D. Liese, Center for Research in Nutrition and Health Disparities, Arnold School of Public Health, University of South Carolina, 921 Assembly Street, Columbia, SC 29208 (e-mail: liese@sc.edu).

Initially submitted April 30, 2010; accepted for publication August 4, 2010.

Despite interest in the built food environment, little is known about the validity of commonly used secondary data. The authors conducted a comprehensive field census identifying the locations of all food outlets using a handheld global positioning system in 8 counties in South Carolina (2008–2009). Secondary data were obtained from 2 commercial companies, Dun & Bradstreet, Inc. (D&B) (Short Hills, New Jersey) and InfoUSA, Inc. (Omaha, Nebraska), and the South Carolina Department of Health and Environmental Control (DHEC). Sensitivity, positive predictive value, and geospatial accuracy were compared. The field census identified 2,208 food outlets, significantly more than the DHEC ($n = 1,694$), InfoUSA ($n = 1,657$), or D&B ($n = 1,573$). Sensitivities were moderate for DHEC (68%) and InfoUSA (65%) and fair for D&B (55%). Combining InfoUSA and D&B data would have increased sensitivity to 78%. Positive predictive values were very good for DHEC (89%) and InfoUSA (86%) and good for D&B (78%). Geospatial accuracy varied, depending on the scale: More than 80% of outlets were geocoded to the correct US Census tract, but only 29%–39% were correctly allocated within 100 m. This study suggests that the validity of common data sources used to characterize the food environment is limited. The marked undercount of food outlets and the geospatial inaccuracies observed have the potential to introduce bias into studies evaluating the impact of the built food environment.

environment; food; geography; reproducibility of results; residence characteristics

Abbreviations: D&B, Dun & Bradstreet, Inc.; DHEC, Department of Health and Environmental Control; GIS, geographic information systems; GPS, global positioning system; NAICS, North American Industry Classification System; PPV, positive predictive value; RUCA, Rural-Urban Commuting Area.

Researchers and health professionals increasingly recognize that the built environment influences health and health behaviors, including dietary intake. Studies have shown a relation of availability and proximity of certain types of food outlets in a neighborhood to dietary behaviors (1–3). The vast majority of epidemiologic studies have relied on readily available, secondary data sources, such as commercial data obtained from InfoUSA, Inc. (Omaha, Nebraska) or Dun & Bradstreet, Inc. (D&B) (Short Hills, New Jersey), telephone company directories (yellow pages), Internet listings, or state agency databases. The addresses are then geocoded without any on-the-ground verification (''ground-truthing'') (1, 3, 4). Through the use of geographic information systems (GIS), the geospatial locations of food outlets can be identified in order to quantify exposures such as distance to the nearest supermarket. To date, only single-site or smaller-scale studies have ground-truthed measures of the food environment (5–7).

Given epidemiologists' general concern about measurement accuracy, surprisingly little information on sources of measurement error exists in this field (8, 9). To our knowledge, only 1 study has examined the validity of a commercial database and an Internet listing of food outlets, by conducting an independent ground-truthing effort in 12 Montreal, Canada, census tracts (10). Indirect evidence comes from a study on the agreement of a local government listing of

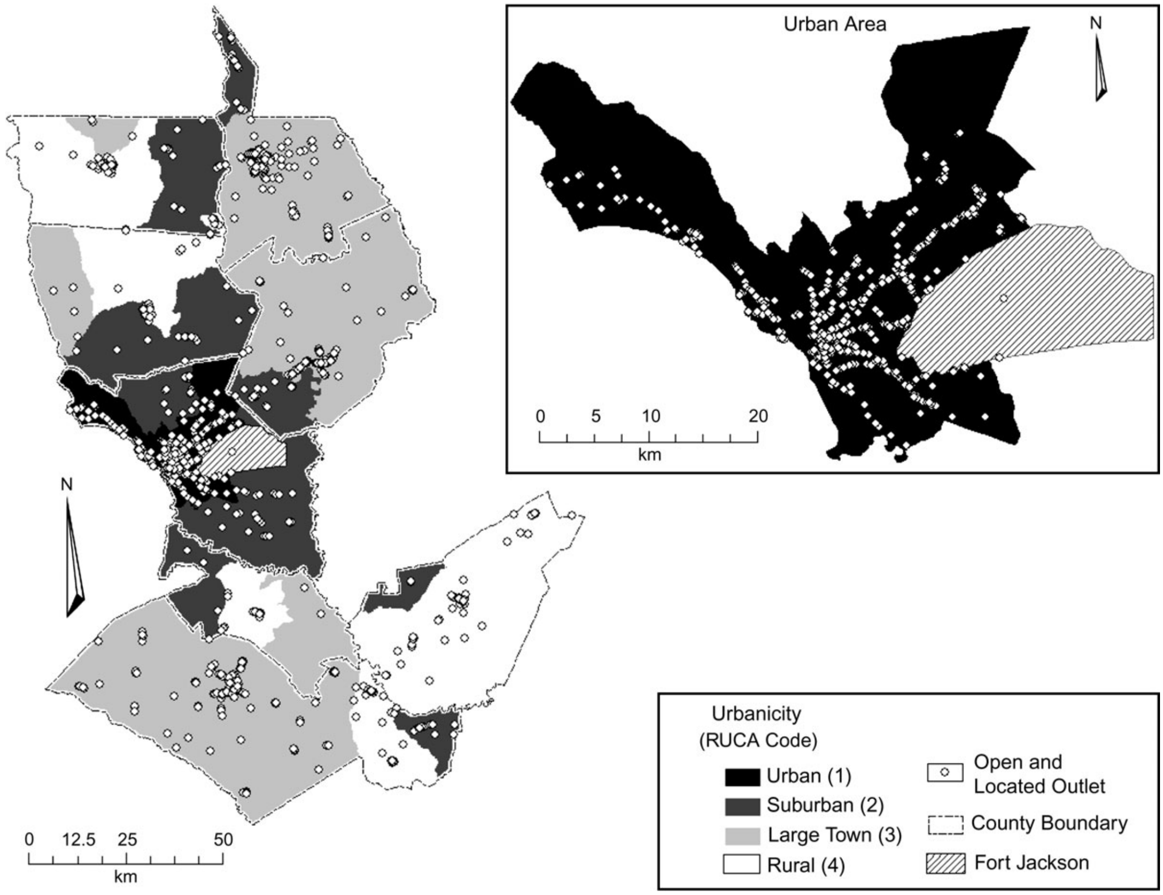

Figure 1. Eight-county study area with open and located food outlets, South Carolina, 2008–2009. The shading represents aggregated Rural-Urban Commuting Area (RUCA) codes. The inset shows the urban area around Columbia.

food outlets with field observation in Glasgow, United Kingdom (11) and the agreement between InfoUSA and field observation in Chicago, Illinois (12). These studies were conducted in large urban areas and suggest that the secondary data sources were quite complete. For instance, Paquet et al. (10) reported sensitivities and positive predictive values (PPVs) of 84% and 90%, respectively, for food stores in Montreal. The level of agreement between listing and actuality was reported at 88% in Glasgow (11) and above 80% in Chicago (12). However, no published validity studies have been conducted in rural environments or larger urban areas in the United States. Moreover, previous experience by our group in 1 rural county suggests that some secondary data sources may not be complete (13).

Another concern with secondary data is its geospatial or positional accuracy (14), a function of the completeness of the address information and the road network database which feed into the geocoding process. The accuracy of several commercial geocoding vendors has been described in detail (15). Many secondary data sources used to characterize the food environment provide geocoded addresses, but, to our knowledge, no study has specifically focused on geospatial error related to geocoded food outlets.

The purpose of this study was to assess the validity of 3 readily available, secondary data sources on food outlets within an 8-county region of South Carolina by conducting a comprehensive field census, including ground-truthing and ascertainment of geographic coordinates. We assessed validity by considering both count accuracy and geospatial accuracy and the influence of level of urbanization.

MATERIALS AND METHODS

Study region

This study was part of a larger effort to develop spatial accessibility measures of the built food environment for urban and rural areas. The geographically contiguous area included 1 urban county (Richland) and 7 rural counties (Calhoun, Chester, Clarendon, Fairfield, Kershaw, Lancaster, and Orangeburg) in the Midlands region of South Carolina (Figure 1). The study area covered 5,575 square miles (8,920 km^2) and a population of more than 620,000. The study was reviewed by the institutional review board of the University of South Carolina and considered exempt.

Data sources

Data on food outlets were obtained from 3 secondary data sources. We obtained the Licensed Food Services Facilities Database, which we have used previously (13), from the South Carolina Department of Health and Environmental Control (DHEC) in the summer of 2008. This database lists all facilities that sell prepared foods in South Carolina. Simultaneously, we obtained commercial data from D&B. We obtained commercial data from InfoUSA in February 2009 after recognizing that the addition of another data source would facilitate our fieldwork.

D&B and InfoUSA listings were queried for specific North American Industry Classification System (NAICS) codes corresponding to facilities that sell food. These included supermarkets and other grocery stores retailing a general line of food (445110), convenience stores (445120), pharmacies and drug stores (446110), gas stations with convenience stores attached (447110), other gas stations (447190), discount department stores or dollar stores (452112), warehouse clubs (452910), supercenters (452910), all other general merchandise stores (452990), specialty food stores (e.g., meat (445210), fish (445220), or fruit/vegetable (445230) markets, bakeries (445291), confectionery stores (445292), or other specialty stores), all other miscellaneous retailers except tobacco stores (453998), full-service restaurants (722110), commercial cafeterias (722212), limited-service restaurants (722211), and snack and nonalcoholic beverage bars (722213). The D&B listing contained up to 5 NAICS codes per food outlet, while the InfoUSA data contained 2 codes. The DHEC database was queried for code 206 (food service facilities) and code 211 (grocery stores).

Because of our focus on the retail food environment, 2 types of outlets were ineligible: 1) sporadic or temporary food vendors operating at sports stadiums or theme parks and 2) outlets that served special populations (e.g., cafeterias in schools or nursing homes, assisted living facilities or institutionalized settings, military settings, and catering businesses without a retail store). We further excluded bars/nightclubs (722410) and liquor stores (445310).

Each database was reviewed separately, and duplicate entries (based on name and address) and outlets that were ineligible because of geography or outlet type were removed. The databases were then merged by name and address into a single comprehensive database that listed each food outlet only once. We started with the DHEC data, into which we merged D&B data and subsequently InfoUSA data. However, we retained all attribute information from each source for each outlet, including a variable indicating the data source from which a given attribute had originated. This comprehensive master listing of unique food outlets served as the basis of our subsequent validation effort. Because data cleaning and merging was very time-consuming, the data cleaning and managing process was conducted on a county-by-county basis so that fieldwork could start as soon as the first counties' data had been managed.

Validation effort

We conducted the field census to verify the presence and location of each food outlet listed in our comprehensive database and to identify new, unlisted outlets. A navigational system was used to locate listed outlets during the ground-truthing trips. Six persons were trained under a standardized protocol, and they took 114 trips entailing 7,000 miles (11,200 km), averaging 2–3 trips per week. The fieldwork began in September 2008 and concluded in July 2009.

Once the address and food outlet were located, the global positioning system (GPS) coordinates were recorded using a Trimble Juno ST GPS receiver (3–5 m spatial accuracy; Trimble Navigation Ltd., Sunnyvale, California) and ArcPad 7.1 software (ESRI, Redlands, California). Facilities were classified as ''located and open'' (outlet was in database and found open for business), ''closed'' (outlet located but closed permanently), ''not found'' (outlet not located at the reported address), or ''ineligible'' (outlet located but gave no indication of being a retail food outlet (e.g., private residence, gas station without a convenience store)). GPS coordinates and outlet name, type, and address were also recorded for new food outlets discovered during the fieldwork (i.e., outlets not listed in the databases). In Table 1, these new outlets are shown as ''found but not listed'' relative to each of the 3 databases.

For outlets not found in a first ground-truthing attempt, we conducted additional Internet queries and, if needed, contacted the outlet. A second systematic ground-truthing attempt was made on all of these outlets toward the end of data collection. This effort would also have captured outlets that had relocated in the interim. If this second attempt was successful, an outlet was classified as ''located and open'' and, if not, as ''not found.'' For the small number of food outlets (D&B, $n = 37$; InfoUSA, $n = 16$) with only a post office box listed, an attempt was made to locate them.

Classification of food outlet type

To differentiate the types of food outlets, we used NAICS definitions as the basis of outlet type groups (''supermarkets and grocery stores,'' ''convenience stores,'' ''pharmacies and drug stores,'' ''dollar and variety stores,'' ''warehouse clubs,'' ''specialty stores,'' ''full-service restaurants,'' ''franchised limited-service restaurants,'' and ''nonfranchised limited-service restaurants'') but with a number of refinements. Warehouse clubs and supercenters were grouped with supermarkets and grocery stores. Snack and nonalcoholic beverage bars were assigned to the limited-service restaurant category, which was further divided to differentiate franchised outlets (Arby's, Burger King, etc.) from nonfranchised outlets.

For all listed food outlets, the NAICS codes were reviewed carefully by multiple team members and corrected manually as needed to remove obvious assignment errors. For all outlets that could not be assigned with certainty, we conducted Internet research and ultimately telephoned the outlet. For newly discovered outlets, the type was assigned during ground-truthing.

Table 1. Disposition of Food Outlets Listed in Secondary Data Sources After an 8-County Field Census, South Carolina, 2008–2009

Data Source and Type of Food Outlet	No. of Outlets Listed	Disposition, %				No. of Outlets Found but Not Listed
		Located and Open	Closed	Not Found	Post Office Box	
South Carolina Department of Health and Environmental Control						
All food outlets	1,694	89.2	4.2	6.6	0.0	696
Stores	417	92.6	1.9	5.5	0.0	513
Supermarket and grocery	122	95.9	1.6	2.5	0.0	44
Convenience	271	91.1	1.5	7.4	0.0	257
Dollar and variety	6	100.0	0.0	0.0	0.0	114
Drug and pharmacy	0	0.0	0.0	0.0	0.0	78
Specialty	18	88.9	11.1	0.0	0.0	20
Restaurants	1,277	88.1	5.0	6.9	0.0	183
Full-service	653	85.3	5.5	9.2	0.0	93
Franchised limited-service	291	95.2	2.1	2.7	0.0	36
Nonfranchised limited-service	333	87.4	6.6	6.0	0.0	54
Dun & Bradstreet, Inc.						
All food outlets	1,573	77.7	7.6	12.3	2.3	985
Stores	751	75.8	5.3	14.8	4.1	330
Supermarket and grocery	157	77.7	7.0	12.7	2.5	39
Convenience	383	74.9	5.7	16.4	2.9	217
Dollar and variety	99	82.8	1.0	10.1	6.1	38
Drug and pharmacy	96	66.7	4.2	18.7	10.4	14
Specialty	16	87.5	12.5	0.0	0.0	22
Restaurants	822	79.6	9.6	10.1	0.7	655
Full-service	389	75.8	11.0	11.8	1.3	356
Franchised limited-service	212	93.4	1.9	4.7	0.0	115
Nonfranchised limited-service	221	72.8	14.5	12.2	0.5	184
InfoUSA, Inc.						
All food outlets	1,657	86.5	3.5	9.0	1.0	774
Stores	672	81.8	3.1	12.8	2.2	349
Supermarket and grocery	136	84.5	3.7	11.0	0.7	46
Convenience	426	82.4	3.0	13.6	0.9	153
Dollar and variety	7	100.0	0.0	0.0	0.0	113
Drug and pharmacy	88	71.6	2.3	14.8	11.4	15
Specialty	15	93.3	6.7	0.0	0.0	22
Restaurants	985	89.7	3.7	6.4	0.1	425
Full-service	481	87.3	5.2	7.3	0.2	231
Franchised limited-service	267	94.4	1.1	4.5	0.0	61
Nonfranchised limited-service	237	89.4	3.8	6.7	0.0	133

Census tract characteristics

Outlets were assigned to their US Census tract and a corresponding level of urbanization based on the 2000 Rural-Urban Commuting Area (RUCA) codes, obtained from the US Department of Agriculture (16). The 10 tiers were consolidated into 4 as described previously (17): urban core (RUCA 1), suburban areas (RUCA 2), large towns (RUCA 3), and small towns/isolated rural areas (RUCA 4).

Statistical and geospatial analysis

To characterize the validity of each database against the field census, we calculated sensitivity as the fraction of open food outlets that were listed and found to be open (i.e., ''located and open''/(''located and open'' + ''found, not listed'')). The PPV was calculated as the fraction of all listed food outlets that were ''located and open'' during the field census (i.e., ''located and open''/(''located and open'' + ''closed'' + ''not found'')). Because of structural zeroes, chance-adjusted kappa statistics could not be computed. We calculated confidence intervals for each of these proportions by approximating the binomial distribution with a normal distribution. Fisher's exact tests were used to evaluate accuracy. Analyses were conducted using SAS software (version 9.2; SAS Institute, Inc., Cary, North Carolina).

All distance analyses were conducted within ArcGIS software (version 9.3; ESRI) using 2008 street network data from the Topologically Integrated Geographic Encoding and Referencing System (18). We computed the geospatial accuracy of outlets by calculating the Euclidean distance between the geocoded outlet location and the GPS location recorded in the field. This analysis was limited to located and open outlets because of the need to have both geocodes from the database and the GPS coordinates from the field census. Similarly, correct allocation of outlets to US Census tracts was determined through comparison of tracts of geocoded outlets with tracts of GPS-verified outlets. Finally, to combine our evaluation of count accuracy with the geospatial accuracy, we calculated the proportion of open outlets that had been both listed in the respective database and geocoded to a position less than 100 m from the actual GPS-recorded location. While this calculation is identical to sensitivity, it could be conducted only on the outlets that were located and open (i.e., a subset of the validity analysis data).

RESULTS

The validation effort identified 2,208 open food outlets, including 160 supermarkets/grocery stores, 504 convenience stores, 120 dollar/variety stores, 79 drug stores, 36 specialty stores, 650 full-service restaurants, 312 franchised limited-service restaurants, and 347 nonfranchised limited-service restaurants. Fifty-two percent of all food outlets were located in Richland County, an urban area.

Table 1 shows the results of the validation effort relative to the secondary data sources. The DHEC database had the largest number of listed outlets ($n = 1,694$), including 417 stores and 1,277 restaurants. InfoUSA listed 1,657 outlets (672 stores and 985 restaurants), and D&B listed 1,573 (751 and 822, respectively). Of the outlets listed by the DHEC, only about 11% could not be confirmed by the field census, because they were either not found (6.6%) or closed (4.2%). The InfoUSA database was similar: Approximately 14% of listed outlets were not confirmed, largely because they were not found (9%) or closed (3.5 %). In contrast, D&B had the highest proportion of outlets that could not be confirmed (22%), with 12.3% not being found, 2.3% not being found because of a post office box address, and 7.6% being closed.

We found 183 food outlets during the validation effort that were not listed in *any* of the 3 data sources. The number of outlets newly found ranged from 696 for the DHEC to 774 for InfoUSA to 985 for D&B. The majority of outlets discovered relative to the DHEC were stores, which is not

Table 2. Validity of Food Outlet Locations Listed in Secondary Data Sources in an 8-County Region as Compared With a Field Census, South Carolina, 2008–2009

Type of Food Outlet	South Carolina Department of Health and Environmental Control				Dun & Bradstreet, Inc.				InfoUSA, Inc.				Dun & Bradstreet and InfoUSA[a]			
	Sensitivity	95% CI	PPV	95% CI	Sensitivity	95% CI	PPV	95% CI	Sensitivity	95% CI	PPV	95% CI	Sensitivity	95% CI	PPV	95% CI
All food outlets	0.68	0.66, 0.70	0.89	0.88, 0.91	0.55	0.53, 0.57	0.78	0.76, 0.80	0.65	0.63, 0.67	0.86	0.85, 0.88	0.78	0.76, 0.80	0.79	0.78, 0.81
Stores	0.43	0.40, 0.46	0.92	0.90, 0.95	0.63	0.60, 0.66	0.76	0.73, 0.79	0.61	0.58, 0.64	0.82	0.79, 0.85	0.81	0.78, 0.83	0.77	0.74, 0.79
Supermarket and grocery	0.73	0.66, 0.80	0.96	0.92, 1.00	0.76	0.70, 0.82	0.78	0.71, 0.84	0.71	0.64, 0.80	0.84	0.78, 0.91	0.90	0.85, 0.95	0.77	0.71, 0.84
Convenience	0.50	0.45, 0.53	0.91	0.88, 0.94	0.56	0.53, 0.61	0.75	0.70, 0.79	0.70	0.66, 0.74	0.82	0.79, 0.86	0.80	0.77, 0.84	0.76	0.73, 0.80
Dollar/variety					0.68	0.60, 0.77	0.83	0.75, 0.90	0.06	0.02, 0.10	1.00	1.00, 1.00	0.70	0.62, 0.78	0.83	0.76, 0.91
Drug and pharmacy					0.82	0.73, 0.90	0.67	0.57, 0.76	0.80	0.70, 0.90	0.71	0.62, 0.81	0.88	0.81, 0.95	0.68	0.58, 0.77
Specialty	0.44	0.28, 0.61	0.89	0.74, 1.04	0.39	0.23, 0.55	0.87	0.71, 1.04	0.39	0.23, 0.55	0.93	0.80, 1.06	0.64	0.48, 0.79	0.92	0.81, 1.03
Restaurants	0.86	0.84, 0.88	0.88	0.86, 0.90	0.50	0.47, 0.53	0.79	0.77, 0.82	0.67	0.65, 0.70	0.90	0.88, 0.92	0.76	0.74, 0.79	0.81	0.79, 0.84
Full-service	0.85	0.83, 0.90	0.85	0.83, 0.88	0.45	0.41, 0.50	0.76	0.71, 0.80	0.64	0.61, 0.68	0.87	0.84, 0.90	0.74	0.70, 0.77	0.79	0.76, 0.82
Franchised limited-service	0.88	0.85, 0.92	0.95	0.92, 0.97	0.63	0.58, 0.69	0.93	0.90, 0.97	0.80	0.76, 0.85	0.94	0.91, 0.97	0.88	0.85, 0.92	0.92	0.89, 0.95
Nonfranchised limited-service	0.84	0.80, 0.90	0.87	0.84, 0.91	0.47	0.41, 0.52	0.73	0.67, 0.79	0.61	0.56, 0.66	0.89	0.85, 0.93	0.71	0.66, 0.76	0.77	0.72, 0.81

Abbreviations: CI, confidence interval; PPV, positive predictive value.

[a] A listing in either database would have been included.

surprising since many stores do not fall under the licensing regulations enforced by the DHEC. However, 183 restaurants were discovered that were not listed by the DHEC. The InfoUSA database was missing 349 stores and 425 restaurants. The D&B database was missing 330 stores and a very large number of restaurants ($n = 655$). Combining D&B and InfoUSA data sources would have yielded 2,170 unique, listed outlets (944 stores and 1,226 restaurants), of which 1,727 were found and open during the field survey (data not shown). Relative to this combined listing, 481 new outlets were discovered during fieldwork.

Validity statistics are shown in Table 2. For all outlets combined, the sensitivity (i.e., the ability to capture food outlets that truly existed in the area) was moderate to fair (68% for the DHEC, 65% for InfoUSA, and 55% for D&B). Both D&B and InfoUSA exhibited moderate sensitivities for food stores (63% for D&B and 61% for InfoUSA) and ranked significantly better than the DHEC (43%). This implies notable undercounting of existing stores for both commercial databases—approximately 37% for D&B and 39% for InfoUSA. Combining the 2 commercial databases would have resulted in a significant improvement in sensitivity for food store identification (81%). Likewise, for supermarkets and grocery stores, all 3 databases had similar levels of sensitivity, ranging from 71% to 76%. The combination of InfoUSA and D&B data would have resulted in a marked improvement in supermarket sensitivity (90%). Combination of DHEC and InfoUSA data or DHEC and D&B data or data from all 3 sources would have resulted in very good to excellent sensitivity for the identification of supermarkets: 86%, 91%, and 97%, respectively (data not shown). No data are shown for DHEC with respect to dollar/variety stores and drug stores/pharmacies because of the DHEC's focus on licensing prepared food outlets. Had we excluded dollar/variety stores and drug stores/pharmacies entirely from the DHEC analysis, the sensitivity of the overall food store category would have improved to 54% (95% confidence interval: 51, 58). Of the other types of food stores, D&B and InfoUSA demonstrated very good sensitivities at 80% or above for drug stores and pharmacies.

With respect to restaurants (lower portion of Table 2), the DHEC database had very good sensitivity of 86% (i.e., an undercount (discovery rate) of only 14%). InfoUSA (67%) and D&B (50%) performed significantly worse. With respect to ranking of the 3 databases across type of restaurants, the findings were consistent. Combining the 2 commercial databases would have resulted in a significant increase in sensitivity compared with using either one alone, but sensitivity would not have reached the level of the DHEC database. However, the combination of DHEC data with InfoUSA or D&B data or the combination of all 3 databases would have resulted in excellent sensitivity values for restaurants (91%, 92%, and 94%, respectively; data not shown).

Table 2 also shows PPVs, which can also be interpreted as verification rates—that is, the likelihood that a listed food outlet actually existed and was open. PPVs ranged from good to very good: 78% for D&B, 86% for InfoUSA, and 89% for DHEC, for all food outlets combined. For stores, the PPV was highest (92%) for the DHEC database, significantly better than for any other database (D&B, 76%;

Table 3. Validity of Secondary Data Sources for Locations of Food Outlets in an 8-County Region as Compared With a Field Census, by Level of Urbanicity, South Carolina, 2008–2009

Type of Food Outlet and Data Source	Level of Urbanicity															
	Urban (RUCA 1)				Suburban (RUCA 2)				Large Town (RUCA 3)				Small Town and Rural (RUCA 4)			
	Sensitivity	95% CI	PPV	95% CI	Sensitivity	95% CI	PPV	95% CI	Sensitivity	95% CI	PPV	95% CI	Sensitivity	95% CI	PPV	95% CI
Stores																
South Carolina DHEC	0.37	0.32, 0.42	0.94	0.90, 0.98	0.59	0.50, 0.69	0.95	0.89, 1.00	0.46	0.41, 0.52	0.93	0.89, 0.97	0.37	0.30, 0.45	0.86	0.77, 0.95
D&B	0.68	0.63, 0.73	0.84	0.79, 0.88	0.54	0.45, 0.64	0.78	0.69, 0.88	0.65	0.60, 0.70	0.71	0.65, 0.76	0.56	0.48, 0.64	0.67	0.58, 0.75
InfoUSA, Inc.	0.62	0.56, 0.67	0.90	0.87, 0.94	0.54	0.45, 0.64	0.76	0.66, 0.86	0.65	0.60, 0.70	0.82	0.78, 0.87	0.56	0.48, 0.64	0.67	0.60, 0.75
D&B and InfoUSA	0.83	0.79, 0.87	0.85	0.80, 0.88	0.70	0.61, 0.79	0.75	0.67, 0.84	0.84	0.80, 0.88	0.75	0.70, 0.79	0.76	0.70, 0.83	0.66	0.60, 0.73
Restaurants																
South Carolina DHEC	0.87	0.85, 0.89	0.89	0.87, 0.92	0.88	0.82, 0.94	0.90	0.83, 0.95	0.82	0.78, 0.86	0.86	0.82, 0.90	0.90	0.84, 0.95	0.84	0.78, 0.90
D&B	0.57	0.54, 0.61	0.79	0.76, 0.83	0.33	0.25, 0.42	0.83	0.71, 0.94	0.45	0.39, 0.50	0.79	0.73, 0.85	0.37	0.29, 0.45	0.78	0.68, 0.89
InfoUSA, Inc.	0.68	0.65, 0.71	0.90	0.88, 0.93	0.50	0.41, 0.59	0.90	0.83, 0.97	0.69	0.64, 0.74	0.90	0.87, 0.94	0.75	0.68, 0.83	0.84	0.77, 0.91
D&B and InfoUSA	0.78	0.75, 0.81	0.81	0.78, 0.84	0.60	0.51, 0.69	0.84	0.76, 0.92	0.75	0.71, 0.80	0.83	0.78, 0.87	0.83	0.76, 0.89	0.80	0.73, 0.87

Abbreviations: CI, confidence interval; D&B, Dun & Bradstreet, Inc.; DHEC, Department of Health and Environmental Control; PPV, positive predictive value; RUCA, Rural-Urban Commuting Area.

Table 4. Geospatial Accuracy of Secondary Data Sources for Locations of Food Outlets (All Types Combined) in an 8-County Region, by Level of Urbanicity, South Carolina, 2008–2009

Data Source and Geospatial Accuracy (Euclidian Distance)	Total	Level of Urbanicity				Kruskal-Wallis *P* Value
		Urban (RUCA 1)	Suburban (RUCA 2)	Large Town (RUCA 3)	Small Town and Rural (RUCA 4)	
South Carolina Department of Health and Environmental Control						
No. of food outlets	1,507	762	167	438	140	
Median distance, m	76	58	110	109	181	<0.0001
Range, m	0.004–57,150	0.004–17,071	0.01–57,150	0.59–56,200	0.05–37,566	
% correctly allocated to within:						
<100 m	56.4	66.1	47.3	47.7	41.4	
100–499 m	24.3	24.3	31.1	23.7	17.9	
≥500 m	19.3	9.6	21.6	28.5	40.7	
% correctly allocated to census tract	82.9	87.1	83.8	76.5	78.6	
Dun & Bradstreet, Inc.						
No. of food outlets	1,213	643	88	353	129	
Median distance, m	92	81	111	105	160	<0.0001
Range, m	0.0009–47,240	1.12–25,968	0.0009–12,310	0.97–47,240	13–36,921	
% correctly allocated to within:						
<100 m	53.0	59.1	46.6	48.7	38.8	
100–499 m	31.2	32.2	35.2	29.7	27.1	
≥500 m	15.8	8.7	18.2	21.5	34.1	
% correctly allocated to census tract	85.2	85.8	90.9	81.9	86.8	
InfoUSA, Inc.						
No. of food outlets	1,434	717	112	430	175	
Median distance, m	92	69	192	117	408	<0.0001
Range, m	1.24–39,839	5.20–35,447	7.81–7,087	1.24–39,839	9.35–38,782	
% correctly allocated to within:						
<100 m	53.2	67.9	36.6	46.7	19.4	
100–499 m	29.0	25.7	26.8	33.3	33.1	
≥500 m	17.8	6.4	36.6	20.0	47.4	
% correctly allocated to census tract	84.0	80.2	94.6	88.6	81.1	

Abbreviation: RUCA, Rural-Urban Commuting Area.

InfoUSA, 82%). For restaurants, DHEC and InfoUSA data performed equally well with respect to PPV (88% and 90%, respectively), with D&B performing significantly worse (79%). This ranking between the databases remained consistent for all 3 restaurant types.

We subsequently evaluated potential differences in the validity of the 3 secondary data sources across levels of urbanization (Table 3). For stores, there were no marked differences between levels of urbanization in any of the 3 databases or the combined D&B and InfoUSA databases. A similar picture emerged for restaurants, the exception being significantly higher sensitivity in urban areas in the D&B data. We additionally evaluated the potential influence of tract racial composition or poverty on the validity estimates but found no evidence for any systematic differences (data not shown).

Geospatial accuracy statistics are shown in Table 4 and are limited to located and open outlets because of the need to have both geocodes from the database and GPS coordinates from the field census. The geospatial accuracy varied widely, with a median Euclidian difference of 76 m (n = 1,507) for DHEC, 92 m for D&B (n = 1,213), and 92 m for InfoUSA (n = 1,434). The percentage of outlets for which the geocoded position was less than 100 m from the GPS

Table 5. Percentage of Located and Open Food Outlets That Were Correctly Allocated to Within 100 m of Their Actual Position, by Secondary Data Source, South Carolina, 2008–2009

Outlet Type	South Carolina DHEC		Dun & Bradstreet, Inc.		InfoUSA, Inc.	
	%	95% CI	%	95% CI	%	95% CI
All food outlets	39	37, 41	29	27, 31	34	32, 36
Stores	24	21, 26	31	28, 34	29	26, 32
Restaurants	49	46, 52	28	25, 30	38	36, 41

Abbreviations: CI, confidence interval; DHEC, Department of Health and Environmental Control.

location ranged from 53% for both D&B and InfoUSA to 56% for DHEC. The correct allocation of outlets to census tracts was high (83%, 85%, and 84% for DHEC, D&B, and InfoUSA, respectively). No notable differences in the median distances were observed by type of food outlet; hence, all outlet types were combined. As expected, the Euclidian distance differences were lowest for the urban areas for all 3 data sources, intermediate for the suburban and large-town areas, and highest for small-town and rural areas ($P < 0.0001$ for all contrasts).

Finally, to combine our evaluation of count accuracy with the geospatial accuracy, we calculated the proportion of open outlets that had been both listed in the respective database and geocoded to a position less than 100 m from the actual GPS-recorded location (Table 5). Overall, between 29% (D&B) and 39% (DHEC) of open food outlets were listed with geocodes that would place them less than 100 m from their actual location. DHEC performed best for restaurants (49%) and worst for stores (24%). D&B and InfoUSA did not differ significantly for stores (31% vs. 29%), but InfoUSA performed significantly better than D&B for restaurants (38% vs. 28%).

DISCUSSION

With increasing adoption of the socioecologic paradigm in public health, the number of studies employing GIS techniques has increased dramatically. As recently reviewed by McKinnon et al. (4), the majority of GIS studies of the food environment have relied on readily available, secondary data sources, with very few epidemiologic researchers conducting any verification (6, 19, 20). In contrast to previous findings (10–12), our study demonstrates that secondary data sources harbor substantial amounts of error, including undercounts, overcounts, and geospatial inaccuracies.

To our knowledge, the only other comprehensive validity study on the built food environment published to date was conducted in 12 census tracts in the Montreal metropolitan region (10). Paquet et al. (10) reported undercounts of 16% for a commercial listing of 171 food stores and 34% for an Internet-based listing of 123 food stores. In our study of 8 rural and urban counties and 2,208 food outlets, undercounts of food stores were much more common, with 37%–39% of food stores and 33%–50% of restaurants not being listed in the D&B and InfoUSA databases. Furthermore, we found the lowest undercount error in the DHEC database, from which only 14% of restaurants were missing (but 57% of stores, which largely are not subject to DHEC regulatory permitting). Similar to our study, Sharkey et al. (7) reported that 21% of the 213 food stores found during ground-truthing in a rural environment were not listed; however, they did not present any other validity statistics in the published article. In our study, combining both commercial databases would have decreased the undercount to 19% of stores and 24% of restaurants, and combining all 3 sources would have decreased the undercounts to 12% and 6%, respectively. Thus, combining multiple data sources and including data from state regulatory agencies is a strategy future investigators should consider in order to reduce undercount error.

Overcount error—that is, listed outlets' either not being present or being found to be closed upon ground-truthing—occurred less frequently (22% in D&B, 14% in InfoUSA, and 11% in DHEC). Had we not conducted substantial data cleaning efforts to remove ineligible outlets prior to the field census, the overcount error would have been much higher. In the Glasgow study, Cummins and Macintyre (11) aimed at characterizing food availability and prices in stores, reporting that approximately 88% of listed outlets were confirmed as open (i.e., an overcount error of only 12%). Unfortunately, combining multiple data sources would have a detrimental effect on this type of error, in that it increases the overcount.

Similar to a previous study of commercial listings of physical activity facilities by Boone et al. (21), our study suggests that, given the dominating nature of undercount error, it is unlikely that the effects of under- and overcounting would balance each other out. This implies that epidemiologic studies using commercial secondary data to characterize the food environments of participants would most likely have underestimated participants' true exposure status. Furthermore, our data suggest that the ratio of undercount error to overcount error may differ somewhat by type of food outlet. In future studies, investigators should consider carefully the amount and type of error specific to the food-outlet type under study in selecting secondary data sources.

We also evaluated the geospatial accuracy of the geocodes contained in each of the 3 databases. Geospatial accuracy factors into any analysis involving distances or precise geospatial location relative to predefined boundaries. For instance, in research on the built food environment, exposures typically include distance measures such as proximity to the nearest supermarket (6) or count or density measures such as number of supermarkets per square mile in a 1-mile (1.6-km) buffer around a residence (3) or within a census tract (1). We found varying levels of geospatial accuracy for the 3 data sources, using the conservative Euclidian (straight-line) distance estimate. Correct allocation to a census tract was high (above 80%), recognizing that in the rural parts of our area census tracts tended to be very expansive. However, on a smaller spatial scale, marked inaccuracies became apparent, with only about 50% of outlets being allocated correctly within 100 m. Furthermore, we attempted to estimate the overall amount of error by combining geospatial accuracy with an estimate of undercount.

The results were sobering: Only 29%–39% of outlets listed had geocodes that placed them within 100 m of their actual location. Therefore, studies relying on very small definitions of neighborhoods may incur larger amounts of error than those utilizing larger buffers or census tracts.

There were several limitations to our study. Given the large geographic area and the number of food outlets, data collection spanned 10 months. Therefore, our data represent more of a period prevalence, while other efforts represent a period as short as 1 month (10). While we attempted to be comprehensive in our field efforts, we cannot exclude the possibility that some outlets were overlooked. In addition, some food outlets discovered may have been listed in a secondary data source but under an NAICS code that we did not request (e.g., code 446191—food/health supplement stores). For the geospatial accuracy analyses, GPS data were used as the gold standard, although this method is subject to some error which can arise from issues including satellite-related errors, signal propagation errors, and receiver errors (22). Lastly, it is important to keep in mind that commercial databases are developed for purposes other than scientific research.

Our study is likely the largest of its kind to date, in terms of both geographic area and the number of food outlets. We included food stores and restaurants, since both have been of interest in research on the built food environment (3, 23), and rural areas, which have not been included in previous studies. We have presented our key results according to specific types of food outlets within the larger groups of stores or restaurants, to assist in future efforts that may be tailored to specific outlets.

In summary, relying exclusively on a single secondary data source for characterizing the built food environment may introduce substantial bias into epidemiologic studies evaluating the impact of the food environment on health behaviors or outcomes. In addition, in the area of policy research, secondary data sources increasingly serve as metrics for public health policy recommendations. For instance, the recent *State Indicator Report on Fruits and Vegetables* published by the Centers for Disease Control and Prevention (24) provides data on policy and environmental supports for fruit and vegetable consumption based on D&B data. In light of our findings, future investigators should consider combining at least 2 secondary data sources to improve levels of accuracy, if secondary data sources cannot be supplemented with extensive ground-truthing efforts.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology and Biostatistics and Center for Research in Nutrition and Health Disparities, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina (Angela D. Liese, Natalie Colabianchi, Archana P. Lamichhane, Timothy L. Barnes, James D. Hibbert, Michele D. Nichols); Department of Environmental Health Sciences, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina (Dwayne E. Porter); and Department of Biostatistics and Epidemiology, College of Medicine, Medical University of South Carolina, Charleston, South Carolina (Andrew B. Lawson).

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 their contributions to the fieldwork.

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.

Conflict of interest: none declared.

REFERENCES

1. Morland K, Wing S, Diez Roux A. The contextual effect of the local food environment on residents' diets: the Atherosclerosis Risk in Communities Study. *Am J Public Health*. 2002;92(11):1761–1767.
2. Laraia BA, Siega-Riz AM, Kaufman JS, et al. Proximity of supermarkets is positively associated with diet quality index for pregnancy. *Prev Med*. 2004;39(5):869–875.
3. Moore LV, Diez Roux AV, Nettleton JA, et al. 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*. 2008;167(8):917–924.
4. McKinnon RA, Reedy J, Morrissette MA, et al. Measures of the food environment: a compilation of the literature, 1990–2007. *Am J Prev Med*. 2009;36(4 suppl):S124–S133.
5. Zenk SN, Schulz AJ, Israel BA, et al. Neighborhood racial composition, neighborhood poverty, and the spatial accessibility of supermarkets in metropolitan Detroit. *Am J Public Health*. 2005;95(4):660–667.
6. Franco M, Diez-Roux AV, Nettleton JA, et al. Availability of healthy foods and dietary patterns: the Multi-Ethnic Study of Atherosclerosis. *Am J Clin Nutr*. 2009;89(3):897–904.
7. Sharkey JR, Horel S. Neighborhood socioeconomic deprivation and minority composition are associated with better potential spatial access to the ground-truthed food environment in a large rural area. *J Nutr*. 2008;138(3):620–627.
8. Oakes JM, Mâsse LC, Messer LC. Work group III: methodologic issues in research on the food and physical activity environments. Addressing data complexity. *Am J Prev Med*. 2009;36(4 suppl):S177–S181.
9. Lytle LA. Measuring the food environment: state of the science. *Am J Prev Med*. 2009;36(4 suppl):S134–S144.
10. Paquet C, Daniel M, Kestens Y, et al. Field validation of listings of food stores and commercial physical activity establishments from secondary data. *Int J Behav Nutr Phys Act*. 2008;5:58. (doi: 10.1186/1479-5868-5-58).
11. Cummins S, Macintyre S. Are secondary data sources on the neighbourhood food environment accurate? Case-study in Glasgow, UK. *Prev Med*. 2009;49(6):527–528.
12. Bader MD, Ailshire JA, Morenoff JD, et al. Measurement of the local food environment: a comparison of existing data sources. *Am J Epidemiol*. 2010;171(5):609–617.
13. Liese AD, Weis KE, Pluto D, et al. Food store types, availability, and cost of foods in a rural environment. *J Am Diet Assoc*. 2007;107(11):1916–1923.

14. Matthews SA, Moudon AV, Daniel M. Work group II: using geographic information systems for enhancing research relevant to policy on diet, physical activity, and weight. *Am J Prev Med*. 2009;36(4 suppl):S171–S176.
15. Whitsel EA, Quibrera PM, Smith RL, et al. Accuracy of commercial geocoding: assessment and implications. *Epidemiol Perspect Innov*. 2006;3:8. (doi: 10.1186/1742-5573-3-8).
16. Economic Research Service, US Department of Agriculture. *2000 Rural-Urban Commuting Area Codes*. Washington, DC: Economic Research Service; 2005. (http://www.ers.usda.gov/Data/RuralUrbanCommutingAreaCodes/2000/). (Accessed September 9, 2008).
17. Washington State Department of Health. *Guidelines for Using Rural-Urban Classification Systems for Public Health Assessment*. Olympia, WA: Washington State Department of Health; 2009. (http://www.doh.wa.gov/Data/guidelines/RuralUrban2.htm). (Accessed December 23, 2009).
18. Bureau of the Census, US Department of Commerce. *2008 TIGER/Line Shapefiles*. Washington, DC: US Census Bureau; 2008. (http://www.census.gov/geo/www/tiger/). (Accessed March 31, 2009).
19. Zenk SN, Schulz AJ, Hollis-Neely T, et al. Fruit and vegetable intake in African Americans: income and store characteristics. *Am J Prev Med*. 2005;29(1):1–9.
20. Ball K, Crawford D, Mishra G. Socio-economic inequalities in women's fruit and vegetable intakes: a multilevel study of individual, social and environmental mediators. *Public Health Nutr*. 2006;9(5):623–630.
21. Boone JE, Gordon-Larsen P, Stewart JD, et al. Validation of a GIS facilities database: quantification and implications of error. *Ann Epidemiol*. 2008;18(5):371–377.
22. Hofmann-Wellenhof B, Lichtenegger H, Collins J, eds. *Global Positioning System: Theory and Practice*. 5th ed. New York, NY: Springer-Verlag, Publishers; 1997.
23. Moore LV, Diez Roux AV, Nettleton JA, et al. Fast-food consumption, diet quality, and neighborhood exposure to fast food: the Multi-Ethnic Study of Atherosclerosis. *Am J Epidemiol*. 2009;170(1):29–36.
24. Centers for Disease Control and Prevention, US Department of Health and Human Services. *State Indicator Report on Fruits and Vegetables, 2009*. Atlanta, GA: Centers for Disease Control and Prevention; 2009. (http://www.fruitsandveggiesmatter.gov/downloads/StateIndicatorReport2009.pdf). (Accessed December 15, 2009).