# Variation in low food access areas due to data source inaccuracies

**Xiaoguang Ma, MD, MPhil**[a], **Sarah E. Battersby, PhD**[b], **Bethany A. Bell, PhD**[c], **James D. Hibbert, MS**[d], **Timothy L. Barnes, MPH**[a], and **Angela D. Liese, PhD, MPH**[a,d,*]

[a]Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA

[b]Department of Geography, University of South Carolina, Columbia, SC, USA

[c]College of Education, University of South Carolina, Columbia, SC, USA

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

## Abstract

Several spatial measures of community food access identifying so called "food deserts" have been developed based on geospatial information and commercially-available, secondary data listings of food retail outlets. It is not known how data inaccuracies influence the designation of Census tracts as areas of low access. This study replicated the U.S. Department of Agriculture Economic Research Service (USDA ERS) food desert measure and the Centers for Disease Control and Prevention (CDC) non-healthier food retail tract measure in two secondary data sources (InfoUSA and Dun & Bradstreet) and reference data from an eight-county field census covering169 Census tracts in South Carolina. For the USDA ERS food deserts measure accuracy statistics for secondary data sources were 94% concordance, 50–65% sensitivity, and 60–64% positive predictive value (PPV). Based on the CDC non-healthier food retail tracts both secondary data demonstrated 88–91% concordance, 80–86% sensitivity and 78–82% PPV. While inaccuracies in secondary data sources used to identify low food access areas may be acceptable for large-scale surveillance, verification with field work is advisable for local community efforts aimed at identifying and improving food access.

## Keywords

food environment; food desert; policy; field census; infoUSA; Dun & Bradstreet

*Corresponding Author*: Angela D. Liese, Department of Epidemiology and Biostatistics, Center for Research in Nutrition and Health Disparities, Arnold School of Public Health, University of South Carolina, 921 Assembly Street, Columbia, SC, USA, 29208, liese@sc.edu, Phone: (803) 777-9414, Fax: (803) 777-2524.

**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.

**Conflict of Interests**
None.

## Introduction

Neighborhood characteristics have been shown to be associated with community food access, which in turn can influence healthy dietary behaviors (Edmonds, Baranowski, Baranowski, Cullen, & Myres, 2001; Moore, Diez Roux, Nettleton, & Jacobs, 2008; Morland, Wing, & Diez-Roux, 2002). A number of studies have shown that low access to healthier food outlets, specifically supermarkets, can contribute to poor diet quality, e.g. lower intake of fruits and vegetables and higher intake of calories from dietary fat (Franco, et al., 2009; Laraia, Siega-Riz, Kaufman, & Jones, 2004; Larson & Story, 2009). Additionally, access to unhealthy food outlets, such as convenience stores and fast food restaurants, also contributes to poor diet quality (Jago, Baranowski, Baranowski, Cullen, & Thompson, 2007; Pearce, Hiscock, Blakely, & Witten, 2008).

Improving access to healthy and affordable food is an explicit goal of several federal and state policy initiatives in the United States (US), including the Healthy Food Financing Initiative (HFFI), a partnership of the Department of Agriculture (USDA), Department of the Treasury (Treasury), and Department of Health and Human Services (DHHS) ("Health food financing initiative (HFFI)", 2010). In addition, the Centers for Disease Control and Prevention (CDC) ("CDC's State-Based Nutrition and Physical Activity Program to Prevent Obesity and Other Chronic Diseases", 2011; "Communities Putting Prevention to Work", 2011) and a variety of state efforts, such as the Pennsylvania Fresh Food Financing Initiative (FFFI) (Steenland, Henley, Calle, & Thun, 2004) have initiated several food environment initiatives. In order to identify areas eligible for the federal support initiatives, several agencies have developed spatial measures of community food access, including the USDA Economic Research Service's (ERS) food desert (FD) (Ver Ploeg, et al., 2009) and CDC's healthier food retail tract (HFRT) ("Children's Food Environment State Indicator Report", 2011; "State Indicator Report on Fruits and Vegetables, 2009", 2009). While the CDC's measure focuses on HFRT, the counterpart, non-healthier food retail tracts (NHFRT), provide a measures of low access similar to the USDA's FD.

Each of these measures of community food access was operationalized using geographic information system (GIS)-based approaches (McEntee, & Agyeman, 2010; Hubley, 2011). These approaches relied on different sources of secondary data to locate and classify retail food outlets. For instance, USDA ERS used the database of stores authorized to receive Supplemental Nutrition Assistance Program (SNAP) benefits and data from Trade Dimensions TDLinx (New York, NY) in 2006 to define the FD ("Food Access Research Atlas data download and current and archived version of documentation", 2013; "Food Desert Locator documentation", 2011). CDC used the Dun & Bradstreet (D&B) data (Short Hills, NY) for locations of supermarkets in 2007 to define HFRT ("State Indicator Report on Fruits and Vegetables, 2009", 2009). In recent years, the validity of secondary retail food data sources has been evaluated in different geographic settings in studies using ground-truthed field census validation (Fleischhacker, et al., 2012; Powell, et al., 2011; Gustafson, et al., 2012; Liese, et al., 2013; Liese, et al., 2010). The ground-truthed field census has been considered as the gold standard for measuring food environment in such studies and the validity measures were estimated for the secondary data sources. Although the findings were inconsistent across studies, all those studies have consistently reported that secondary data sources such as Dun & Bradstreet and InfoUSA (Omaha, NE) contain substantial amounts of error, including undercounts and overcounts of outlets, geospatial inaccuracies, and incorrect assignments of store types (Fleischhacker, et al., 2012; Gustafson, et al., 2012; Liese, et al., 2013; Liese, et al., 2010; Powell, et al., 2011). Errors in these secondary data may introduce bias into studies focusing on individual behaviors and also may affect policy-level food environment indicators such as FD and NHFRT. To date, very little is known

about how inaccuracies in secondary data sources may influence the designation of a Census tract as an area of low food access.

The purpose of this study was to examine the variation in designation of low food access areas due to data source inaccuracies and to quantify the magnitude and direction of the inaccuracies. This study identified low access areas according to two agency-developed community food access measures (FD and NHFRT), using two secondary data sources (D&B and InfoUSA) and data from a validated field census.

## Materials and Methods

### Study area

The study area included eight contiguous counties in the Midlands of South Carolina (Figure 1). The area covers approximately 5,575 square miles and includes a population of more than 720,000, which accounts for about 16% of the total population of South Carolina. Geographically, according to the 2010 U.S. Census, this area includes 169 Census tracts.

### Data sources

**Field census on food outlets (reference data)—**A field census of retail food outlets that included direct observation and verification of all food outlets using geographic positioning systems (GPS) was conducted from September 2008 to July 2009 (Liese, et al., 2013; Liese, et al., 2010). The type of food outlet was assigned using a name-based approach described previously (Liese, et al., 2010). For all listed food outlets, multiple team members carefully reviewed NAICS codes to remove obvious type assignment errors and then assigned the outlets based on the business name and local knowledge of the food retail outlets to an outlet type. For outlets whose type could not be ascertained, Internet research was conducted, and the outlets contacted by telephone. Newly discovered outlets were assigned a type during ground-truthing. This verified dataset was considered the "gold-standard" (and is referred to as the reference data) in the description of the replication of the two measures of community food access.

**Secondary data on food outlets—**Two commercially-available secondary data sources, D&B and InfoUSA, were used to designate Census tracts in the study area according to the two measures of community food access, FD and NHFRT. Both datasets were from 2008–2009 which matches temporally with the field census data and allows us to look at their impact on the food access measures. The two datasets had been obtained in the context of the study described above, i.e. immediately prior to the start of the field census (Liese, et al., 2013; Liese, et al., 2010). Both data sources list businesses according to the North American Industry Classification System (NAICS) ("North American Industry Classification System (NAICS)", 2002) and include geo-coordinates of each outlet and other outlet information, such as number of employees and sales volume. The NAICS codes were used to assign each listed outlet to an outlet type as described previously (Liese, et al., 2010). For the purposes of this analysis, only the NAICS codes 445110, 452910, 452990, and 453998, corresponding to supermarket and grocery stores (including stores retailing a general line of food, supercenters, and warehouse clubs), and code 445230, corresponding to green grocers, were relevant.

**US Census data—**Population and demographic data were obtained from the U.S. Census 2010. Household income was obtained from Census 2010 American Community Survey. Additionally, 1km × 1km gridded population data were obtained. These data were downloaded from the Socioeconomic Data and Applications Center (SEDAC) hosted at Columbia University (Seirup & Yetman, 2006; "Socioeconomic data and applications center

at Columbia university", 2010). The details of how these data were used in the definitions of the two community food access measures are described below.

### GIS computation of community food access measures

We replicated two community food access measures (USDA ERS FD and CDC NHFRT) sequentially using the three data sources, i.e., the field census, D&B and InfoUSA. This allowed us to focus exclusively on the impact of the inaccuracies in the data sources on the measures of community food access. Figure 2 illustrates the process of the GIS analyses, the data sources, and a brief summary of the criteria used for identifying low access areas. The algorithms were implemented using ESRI's ArcGIS (version 10.0) software and related extensions. We refer to the areas designated by each of the two measures (i.e. FD, NHFRT) as low access areas.

In order to account for stores that lie just outside the boundaries of our study area, which will give rise to so called edge effects, before the replication of each food access measure, a 10-mile exterior buffer corridor (grey area in Figure 1) was created around the study area in ArcGIS, using two sources of readily available and existing but not ground-truthed data (InfoUSA and the Licensed Food Services Facilities Database from the South Carolina Department of Health and Environmental Control) (Liese, et al., 2010).

**USDA ERS food deserts**—According to the USDA ERS, a FD is defined as a low-income Census tract where a substantial number or share of residents has low access to a supermarket or large grocery store. This definition is informed by the USDA ERS report *Access to Affordable and Nutritious Food - Measuring and Understanding Food Deserts and Their Consequences* (Ver Ploeg, et al., 2009). A tract is considered as low-income if 20 percent of residents live below the poverty line, or the tract's median family income is less than or equal to 80 percent of the State-wide median family income, or the tract is in a metropolitan area and has a median family income less than or equal to 80 percent of the metropolitan area's median family income. A tract is considered to be low-access if at least 500 people and/or at least 33 percent of the Census tract's population reside more than 1 mile (for urban tracts) or 10 miles (for rural tracts) from a supermarket or large grocery store ("Food Desert Locator documentation", 2011).

First, we identified the low income Census tracts. Then, polygonal 1km × 1km SEDAC population grids were used to evaluate distance to supermarkets or grocery stores. To examine the distance, we converted the SEDAC grids to point data using a centroid approach, retaining the SEDAC population estimates of all people living within each grid cell (Seirup & Yetman, 2006). Distance from each SEDAC grid cell centroid to the nearest food outlet was calculated in miles using Euclidean (straight-line distance) and network (shortest street distance) approaches. For network distance, street centerlines from Streetmap Premium (ESRI, 2011), based on commercial street centerline data from NAVTEQ and Tom Tom, were used. Distances were calculated using the Network Analyst (ESRI, 2011) extension for ArcGIS. Low access was evaluated differently according to USDA guidelines for urban and rural areas. Urbanicity was determined by the intersection of tract centroids with Census-designated urban areas. A tract was considered "urban" if its centroid fell within an urban area; otherwise, the tract was considered to be "rural." SEDAC population data points located in low income tracts that exceeded a threshold distance of 1 mile (urban) or 10 miles (rural) were summed within their corresponding tract boundary to obtain a total population of low-access individuals. If the number of summed population in the low income tracts was more than 500 people, or accounted for more than 33 percent of the Census tract's population, the tracts were defined as FDs.

**CDC non-healthier food retail tracts**—In CDC's *2009 State Indicator Report on Fruit and Vegetables* ("State Indicator Report on Fruits and Vegetables, 2009", 2009), the percentage of a state's Census tracts supporting healthier food choices was used as an indicator to quantify access to fruits and vegetables in the neighborhood. This measure defines a Census tract as being healthier based on availability of healthier food retailers (e.g. supermarkets, large grocery stores, warehouse clubs and fruit and vegetable markets) located within the tract or within a half-mile buffer surrounding the tract boundaries. In order to make it comparable to the USDA ERS food desert measure, we used the logical counterpart to the healthier tract, the NHFRT. The NHFRT was defined as a tract without any healthier food outlets within the tract or within a 0.5 mile buffer surrounding the tract boundary. Counts of food outlets were determined using a spatial join between the tract buffers and food outlets.

## Statistical analysis

The Census tract was the unit of analysis. First, we described the number and percentage of low access tracts identified using the methodology outlined above applied to the reference data, D&B, and InfoUSA. Subsequently, we estimated the influence of inaccuracies in the secondary data on the ability to identify Census tracts with low and non-low food access by using common accuracy statistics. These included the count of agreement on low access areas (+agree), count of agreement on non-low access areas (– agree), count of disagreement (disagree), percentage of concordance, sensitivity, specificity, positive predicted value (PPV), and negative predicted value (PPV). In this study, statistics below 30% were considered poor, 31–50% fair, 51–70% moderate, 71–90% good, and over 90% excellent. This classification method has been used in several studies (Paquet, Daniel, Kestens, Leger, & Gauvin, 2008). We calculated 95% confidence intervals (CI) for each of these proportions by approximating the binomial distribution with a normal distribution. Statistical analyses were conducted using Stata (version 11.0, College Station, TX).

## Results

The number of food outlets by type is shown according to data source in Table 1. The reference data (field census) identified fewer supermarkets and fruit and vegetable markets but more supercenters than listed in either secondary data source. It also included outlets in the categories warehouse club and large grocery store, which were not distinguishable in the secondary data sources because of lack of specific NAICS codes.

Compared to the reference data, D&B data identified fewer tracts as low access for USDA ERS FD (D&B: 11 tracts, InfoUSA: 15 tracts vs. reference data: 14 tracts out of the 169 tracts) (Table 1). For the CDC NHFRT, using either secondary data source to identify areas with low access yielded very similar results to the reference data.

Table 2 shows the ability of the secondary data sources to correctly identify areas designated as low access and as non-low access compared to the reference data. USDA ERS FD showed excellent overall geographic concordance between each of the two secondary data sources and the reference data (93.5%). The concordance of the CDC NHFRT was good but somewhat lower for both D&B and InfoUSA data (87.6%–90.5%). For both community access measures, the secondary data sources identified fewer areas with low food access than the reference data, i.e., sensitivities were below 90%. For example, using either secondary data source, the CDC NHFRT measure would have missed about 15% to 20% of existing low access tracts in the study area, according to the reference data. The USDA ERS FD measure would have missed 35% of low access areas using InfoUSA data, but 50% using D&B data. Specificities were very high and ranged from 90.8% to 97.4%. The PPV values were close to 80% for NHFRT for both data sources, which implies that about 20%

of areas designated as low access by a secondary data source were not low access according to the reference data. The PPV for FD were moderate (60.0%–63.6%), which means that about 40% of areas designated as FD by a secondary data source were not FD according to the reference data.

## Discussion

In this study, secondary data sources such as D&B and InfoUSA, identified a similar number of low food access areas compared with field census data; however the low food access areas were not the same across different data sources. A much lower percentage of Census tracts were designated as low food access areas by USDA ERS FD definition than by CDC NHFRT definition. Compared to reference data, secondary data sources had good to excellent concordance for both FD and NHFRT, and had moderate sensitivity and PPV for FD, and good sensitivity and PPV for NHFRT.

Epidemiological studies have relied on commercial secondary databases such as D&B and InfoUSA to measure food access (McKinnon, Reedy, Morrissette, Lytle, & Yaroch, 2009). Likewise, government agencies have utilized secondary databases to develop community food access measures ("Children's Food Environment State Indicator Report", 2011; "Food Desert Locator documentation", 2011; "State Indicator Report on Fruits and Vegetables, 2009", 2009; Ver Ploeg, et al., 2009). While we have previously reported that secondary databases (i.e., D&B, InfoUSA, and the Licensed Food Services Facilities Database from the South Carolina Department of Health and Environmental Control) were more likely to undercount food outlets and have geospatial inaccuracies compared to our reference data (Liese, et al., 2010) (Liese, et al., 2013), the effect of these inaccuracies on the identification of low access areas as defined by policy-relevant measures of community food access has not been evaluated. In the present study, we found that using D&B tended to identify fewer low food access areas than our reference database for USDA ERS FD. Moreover, using either secondary data source did not consistently lead to the identification of the same tracts as low access areas. We used the same GIS algorithm and Census data when identifying each community food access measure for secondary databases and reference data. Therefore, any differences observed are the result of differences in the count and geographic accuracy of secondary data listings of supermarkets and large grocery stores (and fruit and vegetable markets for NHFRT). The under-ascertainment of low access areas is likely attributable to the geographic inaccuracies of secondary databases and the undercount of food outlets (D&B 24% and InfoUSA 29%) (Liese, et al., 2010).

Even though D&B and InfoUSA located almost the same number of NHFRTs as the reference data, these NHFRTs were not geographically identical. Less than 90% concordance was obtained between these two secondary datasets and the reference data. Both sensitivity and PPV were approximately 80% for the secondary databases. According to the results of this study, using secondary data sources without validation would introduce errors in both directions when identifying NHFRTs. Compared to USDA FD, NHFRT was based solely on food outlet data and not Census data. Thus, the accuracy of NHFRT designations was only dependent on the accuracy and validity of the food outlet database. All the errors in the database would be transferred into the designation of NHFRT.

Based on this study, using InfoUSA resulted in higher sensitivity when identifying FD compared to D&B. In our previous study, compared to the reference data, InfoUSA and D&B had similar sensitivity (71% and 76%, respectively) when identifying supermarket and grocery stores (Liese, et al., 2010). Therefore, the relatively low sensitivity for identifying FD with D&B (50.0%) versus with InfoUSA (64.3%) could not be fully explained by inaccurate counts of food outlets in the databases. In our recent study, considering errors in

both count and type of food outlet, we stratified our validity analysis by income and poverty and found that D&B was more likely to undercount supermarkets and grocery stores in low-income and poor areas (Liese, et al., 2013). According to USDA ERS's definition, all FDs were low-income areas. Therefore, it is likely that using D&B supermarket and grocery store listings for low income areas led to a differential (disproportional) under-ascertainment of FD. In general, under-ascertainment of areas with low food access was a somewhat more pronounced problem, especially for the USDA ERS FD measure, than incorrect assignment of high food access areas to be low food access areas.

It is important to recognize that both the USDA ERS and the CDC have published information designating US Census tracts as low access areas at the Census tract level in the 2011 Food Desert Locator and at the state level in the State Indicator Report on Fruits and Vegetables ("State Indicator Report on Fruits and Vegetables, 2009", 2009) based on Census 2000 data. Applying the exact same time points and methods as the federal agencies, we re-analyzed our results using Census 2000 boundaries (results now shown). According to the agencies' reports, 23 and 31 tracts have been designated as FD and NHFRT in this study area, respectively ("Food Desert Locator documentation", 2011; "State Indicator Report on Fruits and Vegetables, 2009", 2009). For USDA ERS's FD, the tracts identified by the agency outnumber those identified by any of the three data sources in this study. However, the CDC's report lists far fewer tracts than were identified by either secondary data source in this study. USDA ERS used TDLinx and SNAP data in 2006 to identify the FDs in the 2011 Food Desert Locator ("Food Desert Locator documentation," 2011). The different food outlet databases might be the reason for the different number of FDs. However, we did not have access to the original secondary data utilized by USDA ERS. Our results suggest that it may be worthwhile to conduct a formal evaluation of the accuracy of the specific secondary data sources used by the USDA ERS food desert measure. In this study, the findings related to D&B are directly applicable and informative for the CDC's NHFRT, because D&B data were used in the original CDC publication ("State Indicator Report on Fruits and Vegetables, 2009", 2009). However, we identified many more NHFRTs, even using the same secondary database (D&B), possibly because we used data for a more recent year. In this study, a 10-mile buffer area was added around the study area; however, the number of NHFRT was not changed after removing the buffer area. We have tried to replicate the algorithms used by the two agencies to designate low food access areas; however, there are some proprietary aspects of the algorithms. Thus, even though the agencies published the general method, the datasets used as inputs (including the date of each dataset), and possibly some other GIS analysis settings cannot be known and replicated.

There were several strengths of this study. This study is the first to replicate the two agency standard food access measures using different food outlet data sources, and to examine the influence of the accuracy of the databases on the ability to identify Census tracts as having low food access. In addition, the reference database, field census data, was the most comprehensive effort of its kind to date. Moreover, we examined the two community food access measures which are most widely used to identify low food access areas by US government agencies, and used the two secondary food outlet databases, D&B and InfoUSA, most commonly used in epidemiological studies.

The results of our study should be interpreted in light of several limitations. First, in order to estimate the food access for residents living at the edge of the study area, we added a 10-mile buffer around the study area using the combination of InfoUSA and SC DHEC data. Because of this, not all food outlets in the reference data were spatially verified by field census. However, the buffer area was small compared to the whole study area and used the combination of two secondary databases to increase the sensitivity of identifying food outlets of interest (Liese, et al.). Second, some food outlets may be listed in the secondary

databases but under an NAICS code that we did not request (e.g., code 446191 – food/health supplement stores). However, the number of such food outlets should be very small. Third, we replicated the algorithms of identifying each community food access measure based on the descriptions from the documentations published by the agencies.

## Conclusions

Our results suggest that Census tracts identified as having low food access vary substantially depending on the secondary data source used and the particular community food access measure chosen. The amount and direction of error introduced due to using secondary data sources is not acceptable to designate the USDA ERS FD and is probably acceptable to designate the CDC NHFRT if these community food access measures are used for large-scale surveillance purposes, e.g., to estimate the size of populations or areas with low food access. However, if these measures are to be used to identify specific areas in need of intervention by local stakeholders or community efforts, information from secondary data sources probably should be verified by field work.

## Acknowledgments

## List of Abbreviations

| | |
|---|---|
| **HFFI** | Healthy Food Financing Initiative |
| **USDA** | U.S. Department of Agriculture |
| **Treasury** | Department of the Treasury |
| **DHHS** | Department of Health and Human Services |
| **FFFI** | Pennsylvania Fresh Food Financing Initiative |
| **CDC** | Centers for Disease Control and Prevention |
| **FD** | food desert |
| **ERS** | Economic Research Service |
| **HFRT** | healthier food retail tract |
| **NHFRT** | non-healthier food retail tract |
| **SNAP** | Supplemental Nutrition Assistance Program |
| **D&B** | Dun & Bradstreet |
| **GPS** | geographic positioning systems |
| **NAICS** | North American Industry Classification System |
| **SEDAC** | Socioeconomic Data and Applications Center |
| **PPV** | positive predicted value |
| **PPV** | negative predicted value |

# REFERENCES

Center for Disease Control and Prevention. [Accessed on 2012.04.27] CDC's State-Based Nutrition and Physical Activity Program to Prevent Obesity and Other Chronic Diseases. 2011. http://www.cdc.gov/obesity/stateprograms/index.html.

Centers for Disease Control and Prevention. [Accessed on 2012.11.05] Children's Food Environment State Indicator Report. 2011. http://www.cdc.gov/obesity/downloads/childrensfoodenvironment.pdf.

Centers for Disease Control and Prevention. [Accessed on 2012.05.24] Communities Putting Prevention to Work. 2011. http://www.cdc.gov/CommunitiesPuttingPreventiontoWork.

Centers of Disease Control and Prevention. [Accessed on 2010 10.25] State Indicator Report on Fruits and Vegetables, 2009. 2009. http://www.fruitsandveggiesmatter.gov/downloads/StateIndicatorReport2009.pdf.

Edmonds J, Baranowski T, Baranowski J, Cullen KW, Myres D. Ecological and socioeconomic correlates of fruit, juice, and vegetable consumption among African-American boys. Prev Med. 2001; 32(6):476–481. [PubMed: 11394951]

Fleischhacker SE, Rodriguez DA, Evenson KR, Henley A, Gizlice Z, Soto D, Ramachandran G. Evidence for validity of five secondary data sources for enumerating retail food outlets in seven American Indian Communities in North Carolina. Int J Behav Nutr Phys Act. 2012; 9(1):137. [PubMed: 23173781]

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. American Journal of Clinical Nutrition. 2009; 89(3):897–904. [PubMed: 19144728]

Gustafson A, Lewis S, Wilson C, Jilcott-Pitts S. Validation of food store environment secondary data source and the role of neighborhood deprivation in Appalachia, Kentucky. BMC public health. 2012; 12(1):688. [PubMed: 22914100]

Hubley TA. Assessing the proximity of healthy food options and food deserts in a rural area in Maine. Applied Geography. 2011; 31(4):1224–1231.

Jago R, Baranowski T, Baranowski JC, Cullen KW, Thompson D. Distance to food stores & adolescent male fruit and vegetable consumption: mediation effects. Int J Behav. Nutr Phys Act. 2007; 4:35. [PubMed: 17850673]

Laraia BA, Siega-Riz AM, Kaufman JS, Jones SJ. Proximity of supermarkets is positively associated with diet quality index for pregnancy. Prev Med. 2004; 39(5):869–875. [PubMed: 15475018]

Larson N, Story M. A review of environmental influences on food choices. Annals of Behavioral Medicine. 2009; 38:56–73.

Liese AD, Barnes TL, Lamichhane AP, Hibbert JD, Colabianchi N, Lawson AB. Characterizing the food retail environment: impact of count, type and geospatial error in two secondary data sources. Journal Nutrition Education and Behavior. 2013 In Press.

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(11):1324–1333. [PubMed: 20961970]

McEntee J, Agyeman J. Towards the development of a GIS method for identifying rural food deserts: Geographic access in Vermont, USA. Applied Geography. 2010; 30(1):165–176.

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(4 Suppl):S124–S133. [PubMed: 19285203]

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 multiethnic study of atherosclerosis. Am J Epidemiol. 2008; 167(8):917–924. [PubMed: 18304960]

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. [PubMed: 12406805]

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]

Pearce J, Hiscock R, Blakely T, Witten K. The contextual effects of neighbourhood access to supermarkets and convenience stores on individual fruit and vegetable consumption. J Epidemiol Community Health. 2008; 62(3):198–201. [PubMed: 18272733]

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(5):1122–1131. [PubMed: 21741875]

SEDAC. Socioeconomic data and applications center at columbia university. 2010 http://sedac.ciesin.columbia.edu.

Seirup, L.; Yetman, G. U.S. Census Grids (Summary File 3), 2000. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC).; 2006. http://sedac.ciesin.columbia.edu/data/set/usgrid-summary-file3-2000. [Accessed on 2011.05.01]

Steenland K, Henley J, Calle E, Thun M. Individual- and area-level socioeconomic status variables as predictors of mortality in a cohort of 179,383 persons. Am J Epidemiol. 2004; 159(11):1047–1056. [PubMed: 15155289]

U.S. Department of Agriculture. [Accessed on 2013.04.13] Economic Research Service. Food Access Research Atlas data download and current and archived version of documentation. 2013. http://www.ers.usda.gov/data-products/food-access-researchatlas/download-the-data.aspx.

U.S. Department of Agriculture. [Accessed on 2012 03.13] Economic Research Service. Food Desert Locator documentation. 2011. http://www.ers.usda.gov/dataFiles/Food_Access_Research_Atlas/Download_the_Data/Archived_Version/archived_documentation.pdf.

U.S. Department of Commerce. [Accessed on 2011.05.24] United States Census Bureau. North American Industry Classification System (NAICS). 2002. http://www.census.gov/epcd/www/naics.html.

U.S. Department of Health and Human Services. [Accessed on 2011.04.27] Health food financing initiative (HFFI). 2010. http://www.hhs.gov/news/press/2010pres/02/20100219a.html.

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. 2009; Vol. AP-036:160.

**Highlights**

- We replicated two US food access measures comparing results from two commercially available data sources with field census data.

- Low food access areas vary substantially depending on secondary data sources used.

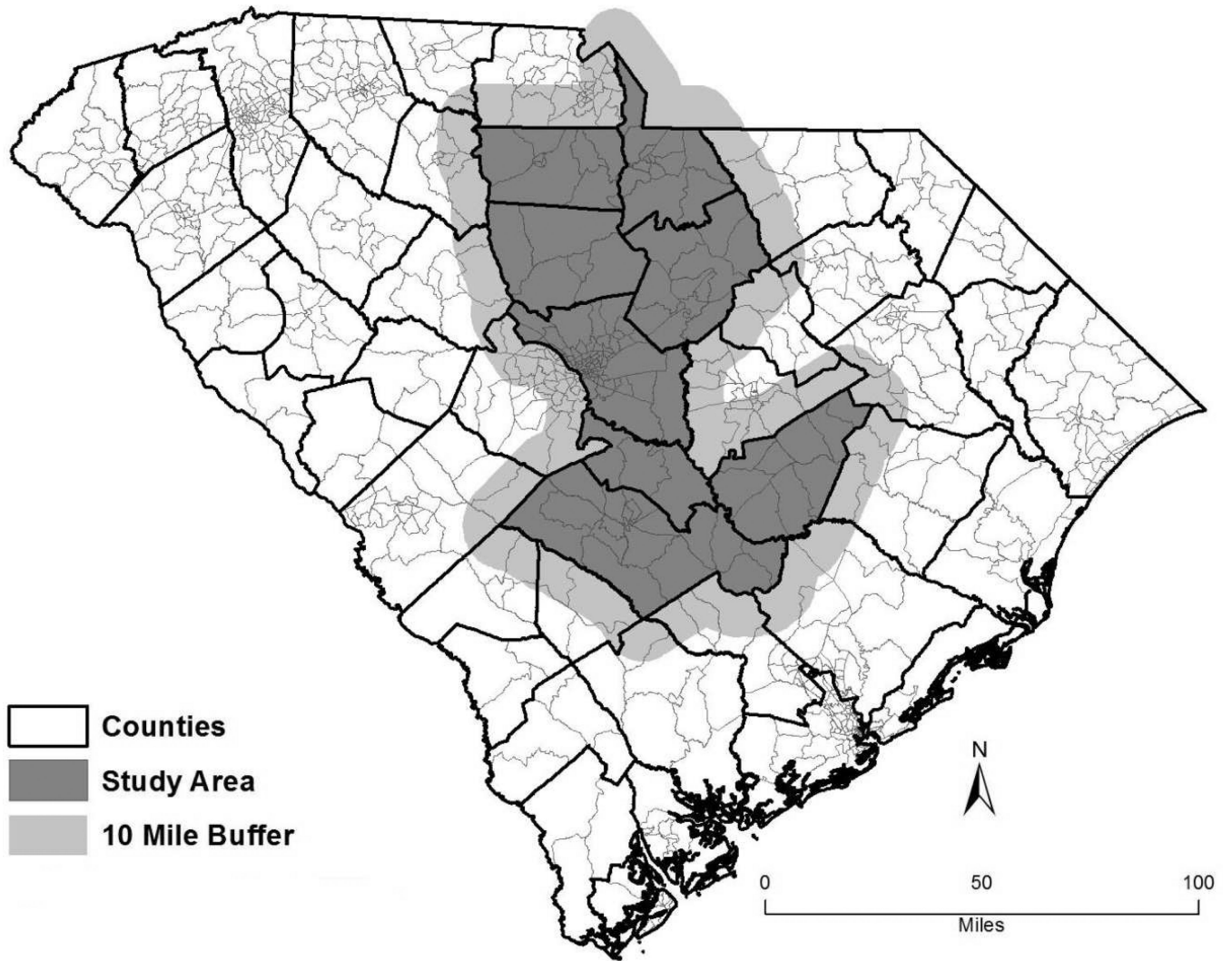- Verification of secondary data with field work is advisable for community efforts.
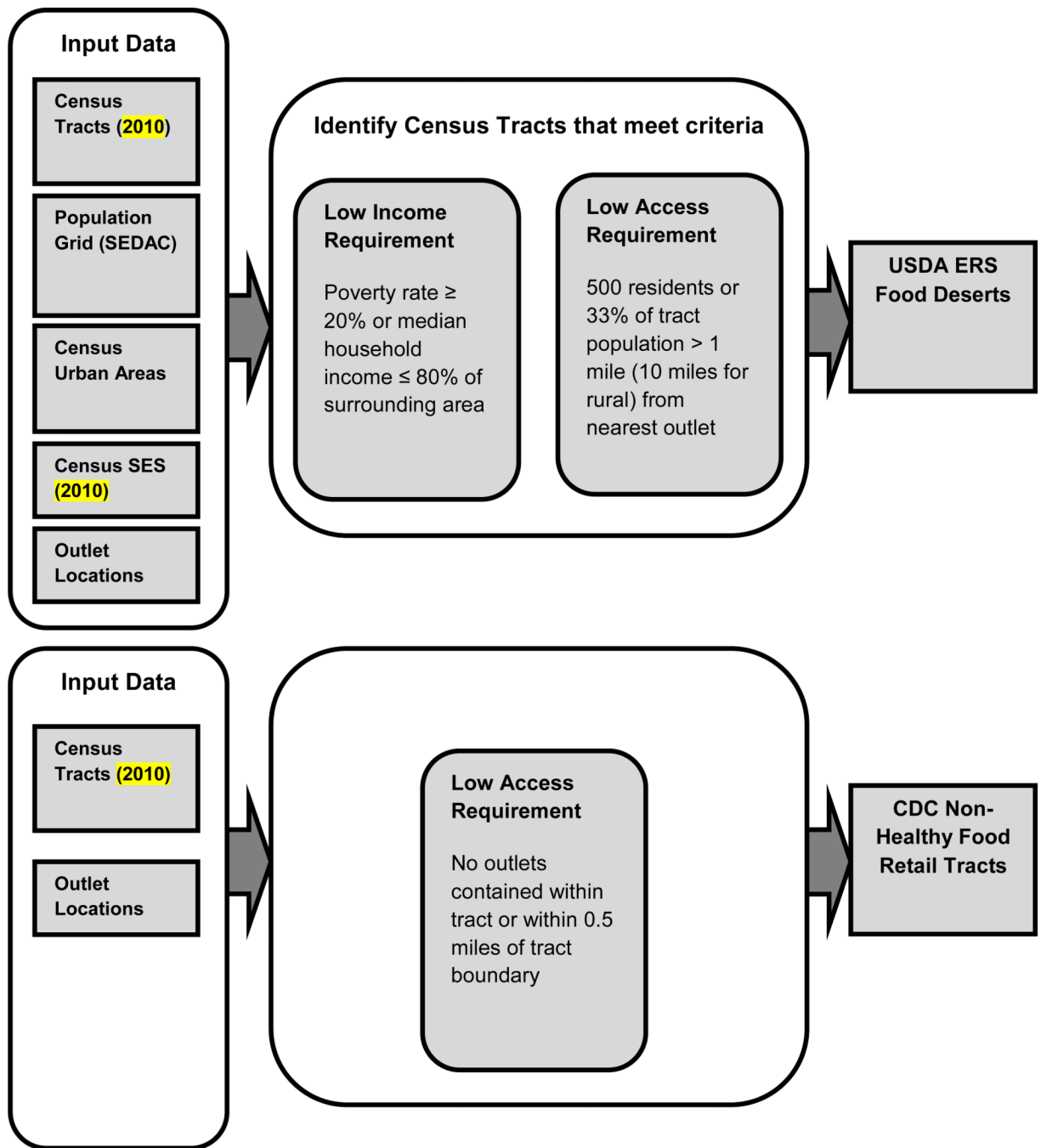
**Figure 1.**
South Carolina Study Area

**Figure 2.**
Dataflow diagram for identifying two community food access measures

Number of food outlets & Census tracts designated as low access according to community food access measure and data source

| | Field Census | D&B | InfoUSA |
|---|---|---|---|
| **Food Outlets in the Study Area, N** | | | |
| Supermarket | 81 | 89 | 82 |
| Supercenter | 13 | 1 | 0 |
| Warehouse Clubs | 1 | -- | -- |
| Large Grocery Store | 7 | -- | -- |
| Fruit and Vegetable Market (Green Grocers) | 6 | 11 | 17 |
| **Food Outlets in the Study Area with 10-Mile Buffer, N** | | | |
| Supermarket | 167 | 174 | 167 |
| Supercenter | 19 | 7 | 6 |
| Warehouse Clubs | 1 | -- | -- |
| Large Grocery Store | 7 | -- | -- |
| Fruit and Vegetable Market | 6 | 11 | 17 |
| **Identified Low Food Access Areas, N (%)** | | | |
| Food Desert Tracts[*] | 14 (8.3) | 11 (6.5) | 15 (8.9) |
| Non-Healthier Food Retail Tracts[*] | 49 (29.0) | 51 (30.2) | 50 (29.6) |

--: No NAICS codes for those categories in D&B and InfoUSA.

[*]: The food access measures are defined for Census tracts. There are 169 Census tracts in the study area.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

**Table 2**

Accuracy statistics of D&B and InfoUSA data compared to field census data for each community food access measure

| | +Agree[a] (count) | –Agree[b] (count) | Disagree (count) | Concordance[c] (%) | Sensitivity (95% CI %) | PPV (95% CI %) | Specificity (95% CI %) | NPV (95% CI %) |
|---|---|---|---|---|---|---|---|---|
| *Food Desert Tracts (Field Census: N=14/169)* | | | | | | | | |
| D&B | 7 | 151 | 11 | 93.5 | 50.0 (24.0, 76.0) | 63.6 (31.6, 87.6) | 97.4 (93.1, 99.2) | 95.6 (90.7, 98.0) |
| InfoUSA | 9 | 149 | 11 | 93.5 | 64.3 (35.6, 86.0) | 60.0 (32.9, 82.5) | 96.1 (91.4, 98.4) | 96.8 (92.2, 98.8) |
| *Non-Healthier Food Retail Tracts (Field Census: N=49/169)* | | | | | | | | |
| D&B | 42 | 111 | 16 | 90.5 | 85.7 (72.1, 93.6) | 82.4 (68.6, 91.1) | 92.5 (85.8, 96.3) | 94.1 (87.7, 97.4) |
| InfoUSA | 39 | 109 | 21 | 87.6 | 79.6 (65.2, 89.3) | 78.0 (63.7, 88.0) | 90.8 (83.8, 95.1) | 91.6 (84.7, 95.7) |

[a] "+Agree" means the agreement on low food access areas.

[b] "–Agree" means the agreement on non-low food access areas.

[c] Concordance = ("+Agree" + "–Agree")/ ("+Agree" + "–Agree" + "Disagree")