Assessing the Validity of Commercial and Municipal Food Environment Datasets in Vancouver, Canada

Madeleine I.G. Daepp[1*] and Jennifer Black[2]

[1]Department of Urban Studies & Planning, Massachusetts Institute of Technology, Cambridge MA 02139

[2]Food, Nutrition and Health, Faculty of Land & Food Systems, University of British Columbia, Vancouver, BC, Canada

*corresponding author: mdaepp@mit.edu

This is the accepted manuscript version of an article published in revised form *Public Health Nutrition* published by Cambridge University Press, August 2017.

Recommended citation: Daepp MIG, Black J (2017). Assessing the validity of commercial and municipal food environment data sets in Vancouver, Canada. Public Health Nutrition 20(15): 2649-2659. https://doi.org/10.1017/S1368980017001744.

Abstract:

Objective: This study assessed systematic bias and the effects of dataset error on the validity of food environment measures in two municipal and two commercial secondary datasets.

Design: Sensitivity, positive predictive value (PPV), and concordance were calculated by comparing two municipal and two commercial secondary datasets with ground-truthed data collected within 800m buffers surrounding 26 schools. Logistic regression examined associations between sensitivity and PPV with commercial density and neighborhood socioeconomic deprivation. Kendall's Tau estimated correlations between density and proximity of food outlets near schools constructed with secondary datasets versus ground-truthed data.

Setting: Vancouver, Canada.

Subjects: Food retailers located within 800m of 26 schools

Results: All datasets scored relatively poorly across validity measures, though overall, municipal datasets had higher levels of validity than did commercial datasets. Food outlets were more likely to be missing from municipal health inspections lists and commercial datasets in neighborhoods with higher commercial density. Still, both proximity and density measures constructed from all secondary datasets were highly correlated (Kendall's Tau> 0.70) with measures constructed from ground-truthed data.

Conclusions: Despite relatively low levels of validity in all secondary datasets examined, food environment measures constructed from secondary datasets remained highly correlated with ground-truthed data. Findings suggest that secondary datasets can be used to measure the food environment, though estimates should be treated with caution in areas with high commercial density.

Keywords: built environment, public health, food environment, data validation

Introduction

Many countries including the U.S. and Canada have seen dramatic increases in rates of childhood obesity, type 2 diabetes, and other diet-related health conditions in recent decades[1, 2]. Researchers have argued that improvements to the wider food environment including the availability, accessibility, or affordability of healthy food[3] could contribute to public health strategies aimed at reducing barriers to healthy eating[4-6]. Recent studies and policy interventions have focused in particular on measuring and assessing the potential impact of the "community nutrition environment" surrounding schools[7], defined by Glanz et al. as "the number, type, and location and accessibility of food outlets" [8]. For example, Los Angeles recently banned fast-food outlets from opening in South Los Angeles[9], in part to reduce children's access to and intake of minimally nutritious foods. In Canada, the only G8 country without a federal school lunch program, students may be particularly likely to purchase minimally nutritious foods from food vendors near schools: Héroux et al.[10] report that Canadian children are more frequent school-day patrons of food retailers than are American children. However, large gaps remain in the evidence base regarding the ways Canadian children's dietary choices are shaped by community nutrition environments surrounding schools (or homes), in part due to difficulties associated with the collection of community nutrition environments data.

The majority of peer-reviewed studies on the community nutrition environment obtain food outlet data from either: (1) "ground truthing", the systematic surveying of a region to identify and classify food retailers, (2) commercial database providers, and (3) government sources[11]. Ground truthing is considered the gold standard[12-13], but the approach is resource intensive and infeasible for the assessment of past years. Commercial datasets often require less time and cost to obtain, and many are available for historical periods (e.g. DMTI Spatial, Inc. 2003[14], 2006[15], and 2009[16]) but such datasets are constructed for business purposes and may not achieve levels of quality necessary for research[11]. To date, many Canadian studies of the community nutrition environment surrounding schools have relied on Yellow Pages (commercial) food outlet directories[10, 17-19]. A recent review, however, found that Yellow Pages directories perform poorly in measures of validity compared with more expensive commercial sources[12]. Municipal datasets like health inspections listings or business registries are frequently free, and could have fewer missing data points because of the legal requirements associated with government data collection[20, 21], but government agencies vary in their efforts to maintain and update registries[12].

A 2013 systematic review identified 19 studies that tested the validity of commonly-used community nutrition environment data sources[12], generally comparing the data source of interest with ground-truthed data. Researchers then rely on validity measures including sensitivity, positive predictive value (PPV), and concordance (Table 1) to characterize levels of overcounting (including stores that have closed or do not exist) and undercounting (failing to include existing stores). Data validation studies also often test for systematic error in secondary datasets, evaluating associations between error rates and neighborhood characteristics[12]. Both random and systematic errors are of interest because random measurement error would add noise that obscures the associations of the community nutrition environment and outcomes of interest, while systematic error would contribute bias that could lead researchers to incorrect results. There is thus a need to understand both the magnitude and the nature of error in commonly used community nutrition environment datasets.

Systematic error is of particular concern because of its potential to produce misleading results. Most studies have not found evidence of systematic bias according to neighborhood socioeconomic status (SES)[22-28] or neighborhood racial demographics[24, 26, 29], but several studies show evidence of systematic bias in relation to urbanicity or commercial density. Four studies in the United States identified statistically significant differences in validity levels in association with urbanicity or density[24, 30-32] though no significant associations were identified in two UK studies[25, 27] and the direction of the association varies across studies. But the datasets examined in the aforementioned studies are often specific to the United States or Europe. In Canada, data validation research has focused on two targeted geographic areas (the city of Montreal[22, 28] and the province of Ontario[33]), limiting generalizability to other regions like Vancouver, where there has been recent interest in food environment research and policy[34]. Moreover, to our knowledge, no Canadian study has tested for systematic bias in validity scores according to commercial density. This is an important gap given the evidence from other countries of associations between validity and commercial density[24, 30-32] as well as the possibility that error, if systematic, may bias research results.

This study sought to fill gaps in the literature through an evaluation of food outlet data sources for the city of Vancouver, Canada. The study's objectives were threefold: (1) to assess the validity of two commercial and two municipal secondary data sources in comparison with ground-truthed data; (2) to test each dataset for evidence of systematic bias in association with neighborhood socioeconomic deprivation or commercial density; and (3) to compare community nutrition environment measures constructed from secondary commercial and municipal datasets with gold standard ground-truthed data. Objective (1) provides results that can be compared with findings from previous data validation research in other countries and cities, while objectives (2) and (3) offer novel methods to help researchers understand how over- or undercounting of outlet listings may be affecting community nutrition environment research.

Methods

Data

This study examined the community nutrition environments surrounding schools in Vancouver, Canada. Vancouver is a coastal city with one of the most densely populated metropolitan areas in North America[35]. Food outlet data were obtained from five sources: (1) ground-truthed primary data, (2) (municipal) Business Licenses[36], (3) (municipal) Vancouver Coastal Health inspections lists[37], (4) (commercial) Pitney Bowes Software's Canada Business Points[38], and (5) (commercial) DMTI Spatial, Inc.'s Enhanced Points of Interest[39]. An overview of these datasets is provided in Table 2.

The ground-truthed data were obtained through systematic surveying between June 29th and September 30th, 2015. A purposive sampling approach was used to select 26 schools across the Vancouver School Board's six geographic sectors (detailed in previous papers[40, 41]) located in neighborhoods with diverse levels of commercial density and socioeconomic status.

Following a surveying protocol adapted from similar research[42] (Supplementary File 1), two researchers visited all commercial streets located within an 800m line-based buffer surrounding schools, a buffer size chosen because it is the distance most frequently examined in research on the community nutrition environment surrounding schools[43]. The researchers identified, photographed, and classified all food outlets; a single researcher also identified, photographed, and classified any outlets along each residential street included in the sample. The surveyors collected outlet GPS coordinates with a Garmin eTrex 20x Worldwide Handheld GPS Navigator. One school buffer zone was visited twice by two separate surveying teams, and the results were compared with Cohen's Kappa to assess inter-rater reliability in surveyors' store classifications.

The two municipal datasets—Business Licenses and Vancouver Coastal Health inspections lists—were obtained from the Vancouver Open Data Catalogue and from the inspections website for Vancouver Coastal Health, respectively, in October 2015. For the Business Licenses, historical records allowed this study to examine both 2015 and 2012 data to consider the potential impacts of temporality of data on validity measures. The inspections lists included records from health inspections of all restaurants and food facilities conducted by Vancouver Coastal Health, the health authority for the region within which this study was conducted. The organization's inspection lists comprised food service establishments, food stores, and food processors in the city of Vancouver, classified by "service type." The Business Licenses data were similar, though they offered a more fine-grained "business sub-type" classification system for identifying convenience stores, grocery stores, and produce outlets.

The most recent commercial data sources to which we had access were Canada Business Points data from 2012 and Enhanced Points of Interest data for 2013. Both datasets included geographic locations, Standard Industrial Classification (SIC) codes, and North American Industry Classification System (NAICS) codes—two federal coding systems that classify businesses according to industry. The NAICS codes are a more recent classification system that has replaced SIC codes for many government agencies in Canada, the United States, and Mexico[(44)].

The 2015 Business License Data[(36)] were also used to measure commercial density—defined as the total number of businesses of any type located within the 800m buffer surrounding schools—based on their performance in the validation study (see Results). Relative socioeconomic deprivation was assessed with the Vancouver Area Neighborhood Deprivation Index (VANDIX), an area-based index of deprivation constructed from seven variables—proportion of the population with less than a high school education, proportion with a university degree, unemployment rate, proportion lone-parent families, average income, proportion of home owners, and labor force participation rate—obtained from the 2006 Census of Canada[(45, 46)]. For this study, the VANDIX was calculated for dissemination areas, 400 to 700-person regions comprising the smallest available census geography[(47)]. The 26 schools examined in this study, which were mapped with data from the Vancouver Open Data Catalogue[(48)], were assigned a "high", "medium" or "low" VANDIX tertile based on the VANDIX scores of the dissemination area directly surrounding the school. "High" scores indicate the most socioeconomically deprived and "low" scores indicate the least deprived areas.

Cleaning and Classification of Food Outlets

The secondary datasets were carefully examined and listings that were outdated, duplicated, or lacking geographic information were deleted following standard procedures used in similar research[22, 28, 31, 49]. For the Vancouver Coastal Health inspections lists, which did not include geographic coordinates, an address locator[50] geolocated outlets with 98% accuracy; manual address matches were identified for the remaining 2% of outlets. For each of the four secondary community nutrition environment datasets, outlets located within 800 meter line-based buffers[51] surrounding each of the 26 schools of interest were extracted for comparison with ground-truthed outlets located within the same buffers. All geographic data were projected to the NAD83 / UTM zone 10N coordinate system with ArcGIS[52].

This study compared three classes of outlets: (1) limited-service food outlets, restaurants or coffee shops where customers order at a counter and pay before consuming food or beverages; (2) convenience stores, which included retail stores primarily offering snack foods or beverages, possibly attached to a pharmacy or gas station; and (3) grocery stores or supermarkets, comprising retail food stores with the departments of a traditional grocer (dairy, bakery, butcher, deli and produce). These three store types were selected because they are the most commonly used store types in the literature on community nutrition environments surrounding schools[43], and definitions were adapted from previous research[42, 49, 53]. Outlets were classified following a modification of the flowchart used by Clary and Kestens[28] (included in Supplementary File 1). For the 2012 and 2015 Business Licenses, "Business Type" and "Business Subtype" were used to classify listings. The "Facility Type" classification included in the Vancouver Coastal Health inspections lists was too coarse-grained to identify each of the three outlet classes and the SIC/NAICS codes provided in the commercial Canada Business Points and Enhanced Points of Interest were inadequate for classification (e.g. McDonald's and other well-known fast food outlets were listed as full-service restaurants, and the codes often failed to discriminate between convenience stores and small grocery outlets). This study thus supplemented the "Facility Type" and SIC/NAICS codes with the application of a name-based classification scheme (Supplementary File 2) following previous studies[27, 28].

Outlet Matching Approach

Two approaches were applied to match outlets in the commercial and municipal datasets with outlets in the ground-truthed dataset. First, outlets were compared by address and two outlets were matched if the listings included identical street names and numbers. This approach left some stores unmatched due to small inconsistencies, so an algorithm was encoded in R 3.2.4[54] to match each store according to name and geographic location, following previous studies[55, 56]. For each store in the ground-truthed dataset, geographic coordinates were used to identify all stores in the secondary dataset located within 100 meters of the ground-truthed store. The Levenshtein similarity, a similarity function based on the Levenshtein distance—the minimum number of edits necessary for one store name to become identical to the other[57]—was calculated for all potential matches within 100m with the RecordLinkage package for R[58]; the ground-truthed store was then matched with the outlet with the highest Levenshtein similarity score. Results from the address- and the name-based matching approaches were compared and, for ground-truthed outlets with different results across the two approaches, the best match was determined manually. For the Canada Business Points, which did not include addresses, the algorithm was applied twice and each entry was reviewed and, if necessary, matched manually.

Analysis

First, the validity of all secondary datasets was assessed with the ground-truthed dataset serving as the gold standard. For each of the commercial and municipal secondary datasets, a matched store was considered a true positive (TP) if it was listed in both the secondary dataset and the ground-truthed data with the same classification, a false positive (FP) if listed in the secondary data but not in the ground-truthed data, and a false negative (FN) if listed in the ground-truthed data but not in the secondary dataset. Sensitivity, positive predictive value and concordance (defined in Table 1) were then calculated as measures of the validity of each secondary data source. A listing was considered a TP even if it had a different name in the secondary dataset from that in the ground-truthed data, if the two listings included identical addresses and classifications. As a sensitivity analysis, “strict” TP's were calculated omitting stores with highly dissimilar names.

Second, logistic regressions examined whether the odds of FP or FN's increased in association with neighborhood socioeconomic deprivation or commercial density to assess systematic biases.

Regressions were fitted for all stores in the ground-truthed dataset with the outcome equal to 1 if the store was a false negative and 0 if the outlet was a true positive; the PPV analyses were run for all stores in each secondary dataset with the outcome equal to 1 if the store was a false positive and 0 if the store was a true positive. Each model was fitted with either VANDIX score tertile or commercial density (in units of 100 outlets) as independent variables. As a sensitivity analysis, models were also fitted with population density—measured as the average number of people per hectare located within the 800m line-based buffers surrounding each school—calculated from dissemination area-level data from the 2006 Census.

Third, community nutrition environment measures (density and proximity of outlets near schools) constructed from the commercial and municipal datasets were compared with measures from the ground-truthed dataset using Kendall's Tau, a non-parametric measure of correlation[(59)]. ArcGIS was used to calculate density—the total number of outlets located within each 800m line-based school buffer—and proximity, the shortest street-based distance from each school to a food outlet. Confidence intervals were calculated with the DescTools package in R[(60)] and $p<0.05$ was used for determining statistical significance for all analyses.

Results

Assessment of Dataset Validity

Table 3 reports the counts of food outlets for each of the municipal and commercial secondary datasets and results from comparisons between ground-truthed and secondary data sources. Ground truthing identified 267 limited-service food outlets, 124 convenience stores, and 64 grocery stores or supermarkets located within 800m of the sample of 26 schools. For outlets classified by two surveyors, percent agreement was 91% and Cohen's Kappa was 0.88, indicating strong inter-rater reliability[(61)].

The 2015 Business Licenses had the highest overall scores for sensitivity, identifying 69% of the ground-truthed stores. This dataset's sensitivity was highest for convenience stores (0.75) and limited-service outlets (0.72), and lower for grocery stores (0.42). Nevertheless, the Business Licenses generated the highest sensitivity for grocery stores among the secondary data sources examined. The Vancouver Coastal Health inspections list, in contrast, had the highest PPV (0.60) for all outlets combined. The validity estimates for each of the municipal datasets in 2015 were higher than those obtained for either of the two commercial datasets in all cases except for the sensitivity estimates for grocery stores.

With strict name matching, the 2015 Business License data lost 28 outlet matches, leading its sensitivity to drop to 0.62 while PPV decreased to 0.50. The 2012 Business License data lost 34 matches (sensitivity=0.51, PPV=0.42), the Vancouver Coastal Health data lost 15 matches (sensitivity=0.50, PPV=0.57), and the Enhanced Points of Interest lost 27 matches (sensitivity=0.33, PPV=0.32). Canada Business Points had the fewest matched outlets with different names, with just 7 outlets failing the stricter name-based standard (sensitivity=0.40, PPV=0.42). Regardless of the approach to matching store names, the municipal datasets performed better in terms of overall sensitivity, PPV, and concordance than did the commercial datasets.

Assessment of Systematic Bias

Tables 4 and 5 report findings from bivariate logistic regressions examining associations of commercial density and socioeconomic status with false positive (FP) and false negative (FN) listings in each secondary dataset. Neighborhood socioeconomic deprivation surrounding schools was not consistently associated with the odds of listings being false positives or false negatives. However, commercial density surrounding schools was significantly associated with the proportion of false negative (versus true positive) listings in all secondary datasets except the municipal Business Licenses data. An increase in 100 stores within an 800m buffer zone surrounding schools was associated with a 7% increase in the odds that a store in the ground-truthed data would be missing from the Vancouver Coastal Health Inspections lists (OR=1.07, 95% CI 1.01 - 1.14), 11% higher odds in the Canada Business Points (OR=1.11, 95% CI: 1.04 - 1.18), and 8% higher odds in the Enhanced Points of Interest (OR=1.08, 95% CI 1.01 - 1.15). Commercial density was not significantly associated with the odds of false positive listings, and no significant associations were observed in models fitted with population density rather than commercial density.

Comparison of Community nutrition environment Measures Across Datasets

Across all secondary data sources, density measures were highly correlated with measures from the ground-truthed data (Kendall's Tau-b≥0.87 for all outlets). The strength of the correlations between proximity measures from secondary and ground-truthed data were slightly lower, with Kendall's Tau-a falling between 0.61 for the 2012 Business Licenses (95% CI 0.37 - 0.84) to 0.74 for the Canada Business Points (95% CI 0.49 - 0.99). This suggests that in ranking schools by proximity, measures constructed from the Canada Business Points were 74% more likely to agree than to disagree with measures constructed from the ground-truthed data; rankings based on measures constructed from the 2012 Business Licences were only 61% more likely to agree than to disagree with measures constructed from the ground-truthed data.

Table 6 further illustrates differences in the correlations of community nutrition environment measures between data sources depending on the store type of interest. Though both commercial datasets performed comparably to the municipal datasets in estimating the density of limited-service outlets and convenience stores, rank-correlations were considerably lower for grocery store densities (0.56 and 0.51, respectively).

Discussion

This study assessed the validity of two municipal and two commercial community nutrition environment data sources compared with a gold standard, ground-truthed dataset in a large North American city. This research to our knowledge is the first to directly compare two commercial database providers—DMTI Spatial, Inc. and Pitney Bowes Software—which are among the most accessible proprietary sources of commercial food outlet data in Canada. The study adds to the literature by examining how error affects measures of community nutrition environment exposure surrounding schools, illuminating the nature and magnitude of error within secondary datasets, and offering insight from a large Canadian city.

The study found that all datasets were subject to high levels of error: datasets both (1) failed to include at least 20% of outlets observed in the field and (2) consisted at minimum of 25% listings not found in the field. The 2015 Business License data and the Vancouver Coastal Health data had sensitivity and PPV values in the range of 0.54 - 0.69 (for all food outlets), similar to results for local health department listings' sensitivity (0.66) and PPV (0.49) in North Carolina, U.S.[(42)], and to a sensitivity estimate (0.66) for city council data in Newcastle, U.K.[(62)]. The municipal data sources' PPV scores were lower, however, than those found in Newcastle city council data (PPV=0.92)[(62)] and for South Carolina Department of Health and Environmental Control data (PPV=0.89)[(31)]. These differences suggest that researchers should evaluate the validity of government data on a case-by-case basis, if possible, before choosing to use municipal datasets for research purposes[12].

Overall, the sensitivity, PPV, and concordance values for the commercial data sources were lower in Vancouver than reported in previous studies in other regions. For example, examining food outlets in the UK Points of Interest data for 2012, Burgoine and Harrison[27] obtained a sensitivity value of 0.60 and PPV of 0.75, significantly higher than the values observed for commercial data sources in this study; Clary and Kestens[28] similarly obtained higher PPV and sensitivity estimates (0.64 and 0.55, respectively) for their examination of the 2010 Enhanced Points of Interest data in Montreal. Both sets of researchers, however, had a smaller temporal difference between the last update of the secondary data source and their collection of ground-truthed data in comparison with this study, suggesting that the difference in results may be explained by the depreciation of data quality over time.

Nevertheless, this study found that overall both municipal datasets outperformed commercial datasets in measures of validity, even when the 2012, rather than 2015 Business License data was used for comparison. Much of the existing literature on the community nutrition environment surrounding schools has relied on commercial data sources such as the two datasets examined here[43]. This study suggests that municipal datasets can provide adequate alternatives that may offer higher quality data than many of the datasets on which the community nutrition environment literature currently relies.

This study also evaluated associations between neighborhood socioeconomic deprivation and commercial density with the odds of incorrect listings. This examination was valuable because systematic error in datasets could bias research findings: if datasets consistently fail to identify existing food retailers in low-income neighborhoods, for example, researchers might underestimate low-income communities' access to food retailers. In the absence of such bias, random error could create “noise” that weakens the magnitude of observed associations (i.e. type 2 error when true associations are not detected). Thus, the results obtained here—of no consistent associations between neighborhood socioeconomic deprivation and the odds of false negative or false positive associations—are reassuring for researchers because they suggest that results regarding socioeconomic disparities in food retail access are not subject to systematic bias. This finding is similar to the results of several previous studies that have reported no associations between measures of socioeconomic deprivation and levels of commercial dataset validity[22, 23, 26-28].

This study did, however, find positive associations between the odds of false positive listings and commercial density in three of four datasets. Similar results were reported in Chicago where more disagreement between secondary and ground-truthed data was found for stores closer to the city's central business district [(24)]. Areas close to the central business district are among the city's most commercially dense neighborhoods, so these results suggest that researchers would obtain lower validity scores in more commercially dense areas. It is worth noting that we conducted a sensitivity analysis using population density as an alternate measure of urbanicity, which did not find evidence of significant associations between that measure and odds of false positives or false negatives in any dataset. We did not have access to data regarding business turnover, but hypothesize that more commercially dense Vancouver neighborhoods (but not necessarily those with higher population densities alone) may have more outlets opening annually and thus more stores that can be missed. Researchers using commercial data to compare areas with higher and lower commercial density should therefore bear in mind potential impacts of such systematic error.

Despite the evidence of low levels of validity, community nutrition environment measures constructed from the commercial and municipal datasets were highly correlated with measures from ground-truthed data. This observation is consistent with findings of two other known studies examining the effect of dataset validity on community nutrition environment measures: Ma et al.[(63)] found that measures of food deserts—low-income areas where residents lack access to grocery stores or supermarkets—created from two commercial datasets (InfoUSA and Dun & Bradstreet) had 93.5% concordance with comparable measures obtained from the United States Department of Agriculture and the Centers for Disease Control and Prevention; and Lebel et al.[(64)] found that estimates of food stores per 1000 people constructed from a commercial dataset (InfoUSA) had 86.9% correlation with estimates calculated from a gold standard dataset (Boston Inspectional Services Department). The high levels of undercounting and overcounting estimated with low sensitivity and positive predictive values, respectively, may offset one another, resulting in data that remains representative of the true community nutrition environment. Thus low validity scores did not translate into low validity for measures of relative access to food outlets, leading researchers to underestimate the usefulness of secondary datasets for research on the community nutrition environment[(64)].

Several notable limitations of this study should be considered. Foremost, because ground-truthed data were collected in 2015, depreciation of data quality over time may contribute to the lower validity scores this study obtained for commercial datasets (collected in 2012 and 2013) in comparison with the municipal datasets, which were collected immediately after the completion of ground truthing in 2015. However, the inclusion of both current (2015) and historical (2012) Business License data suggests that depreciation explains only part of the difference in validity. The two commercial datasets still performed between 5 and 10 percentage points worse in PPV and nearly 20 percentage points worse in sensitivity scores compared with the municipal Business Licenses for 2012. Additionally, findings may not be generalizable to other cities because of variance in municipal dataset quality, and the findings may overestimate validity for studies that do not follow the data cleaning and classification protocols used in this research[(65)]. It should also be noted that the gold standard, ground-truthed data, is subject to error that could contribute to the low validity scores estimated for secondary datasets. Although inter-rater reliability in store classification was high, it remains possible that surveyors missed stores or that results were affected by turnover in Vancouver storefronts. Finally, our definition of the community nutrition environment was limited to publicly accessible food outlets; places with restricted access such as office cafeterias or school snack shops were not examined in this study because they are considered to comprise the "organizational" nutrition environment rather than the community nutrition environment[(8)].

Further research is still needed to understand why measures of proximity and density from secondary and ground-truthed data remained highly correlated despite low levels of sensitivity and PPV; researchers also need to continue working on classification schemes that could reduce the over- and undercounting attributable to reliance on industrial classification codes. And finally, studies are needed that examine how error may affect outcomes ultimately of interest—the associations between diet-related health outcomes and community nutrition environment exposures.

Nevertheless, this research remains relevant to researchers outside Vancouver in both its methods and its findings. The inclusion of multiple years of municipal data offers researchers insight into the effects of depreciation over time. The finding of an association between error and commercial density joins several studies suggesting that researchers should be concerned with the effects of commercial density on data quality. Furthermore, the method of calculating the correlation between community nutrition environment measures from secondary datasets and ground-truthed data could be replicated with datasets in other geographic and national contexts, an effort that would help bring researchers a step closer to understanding the impact of error on the results obtained in community nutrition environment studies.

Conclusions

All datasets examined in this study scored relatively poorly across validity measures. Three of the four datasets also had evidence of systematic bias in association with commercial density, though no datasets were systematically more likely to over- or under-count outlets in relation to neighborhood socioeconomic status. Nevertheless, community nutrition environment measures constructed from both municipal and commercial data sources were highly correlated with ground-truthed measures, suggesting that datasets with low validity scores may still offer reliable measures of community nutrition environment exposure.

The City of Vancouver Business Licenses outperformed other data sources in measures of sensitivity and in its lack of systematic error in association with neighborhood characteristics. Furthermore, community nutrition environment measures constructed from the Business Licenses and those constructed from ground-truthed data were highly correlated. This study thus suggests that the Business Licenses offer the best available dataset for community nutrition environment research in Vancouver. For studies using commercial data providers, this study suggests that researchers should be wary of systematic error in association with commercial density. While such datasets perform reasonably well for studies quantifying relative community nutrition environment exposures, they may be less useful for policymakers or planners seeking to identify specific food outlets.

References

1. Patterson C, Guariguata L, Dahlquist G, *et al.* (2014) Diabetes in the young - a global view and worldwide estimates of numbers of children with type 1 diabetes. *Diabetes Res Clin* Pract **103**, 161–75.
2. Ng M, Fleming T, Robinson M, *et al.* (2014) Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **384**, 766–781.
3. Caspi CE, Sorensen G, Subramanian, SV, *et al.* (2012). The local food environment and diet: A systematic review. *Health Place* **18**, 1172-87.
4. Ver Ploeg M, Breneman V, Farrigan K, *et al.* (2009). Access to affordable and nutritious food—Measuring and understanding food deserts and their consequences: Report to congress. Washington, DC: USDA Economic Research Service, Administrative Publication No. 036. Retrieved May 7, 2017 from https://www.ers.usda.gov/publications/pub-details/?pubid=42729.
5. Fluornoy R (2010). Health food, healthy communities: Promising strategies to improve access to fresh, healthy food and transform communities. PolicyLink. Retrieved May 7, 2017 from http://www.ca-ilg.org/sites/main/files/file-attachments/resources__hfhc_short_final.pdf.
6. Zenk SN, Thatcher E, Reina M, Odoms-Young A (2015) Local food environments and diet-related health outcomes: A systematic review of local food environments, body weight, and other diet-related health outcomes. In: Morland KB, editor. *Local Food Environments: Food Access in America.* Boca Raton, FL: CRC Press; 191–192.
7. Mair JS, Pierce MW, Teret SP (2005) The use of zoning to restrict fast food outlets: a potential strategy to combat obesity. The Center for Law and the Public's Health, Johns Hopkins & Georgetown Universities, 51-53. Retrieved May 7, 2017 from http://www.publichealthlaw.net/Zoning%20Fast%20Food%20Outlets.pdf
8. Glanz K, Sallis JF, Saelens BE, *et al.* (2005). Healthy nutrition environments: Concepts and measures. Am J Health Promot **19**, 330-333.
9. Sturm R, Cohen DA (2009) Zoning for health? The year-old ban on new fast-food restaurants in South LA. *Health Aff* **28**, w1088–97.

10. Héroux M, Iannotti RJ, Currie D, *et al.* (2012) The food retail environment in school neighborhoods and its relation to lunchtime eating behaviors in youth from three countries. *Health Place* **18**, 1240–7.

11. Moore LV, Diez-Roux AV (2015) Measurement and Analytical Issues Involved in the Estimation of the Effects of Local Food Environments on Health Behaviors and Health Outcomes. In: Morland KB, editor. *Local s: Food Access in America.* Boca Raton, FL: CRC Press; p. 205–226.

12. Fleischhacker SE, Evenson KR, Sharkey J, *et al*. (2013) Validity of secondary retail food outlet data: a systematic review. *Am J Prev Med* **45**, 462–73.

13. Lucan SC (2015) Concerning limitations of food-environment research: a narrative review and commentary framed around obesity and diet-related diseases in youth. *J Acad Nutr Diet* **2**, 205–212.

14. DMTI Spatial, Inc (2003) Enhanced Point of Interest Layers [2003]. Retrieved June 30, 2016, from http://hdl.handle.net.ezproxy.library.ubc.ca/11272/NBRIL.

15. DMTI Spatial, Inc (2006) Enhanced Point of Interest Layers [2006]. Retrieved June 30, 2016, from http://hdl.handle.net.ezproxy.library.ubc.ca/11272/KDY86.

16. DMTI Spatial, Inc (2009) Enhanced Point of Interest Layers [v.2009.3]. Retrieved June 30, 2016, from http://hdl.handle.net.ezproxy.library.ubc.ca/11272/JGQ3B.

17. Seliske LM, Pickett W, Boyce WF, *et al*. (2009) Density and type of food retailers surrounding Canadian schools: variations across socioeconomic status. *Health Place* **15**, 903–907.

18. Seliske LM, Pickett W, Boyce WF, *et al*. (2009) Association between the food retail environment surrounding schools and overweight in Canadian youth. *Public Health Nutr* **12**, 1384–91.

19. Laxer RE, Janssen I (2013) The proportion of excessive fast-food consumption attributable to the neighbourhood food environment among youth living within 1 km of their school. *Appl Physiol Nutr Metab* **39**, 480–486.

20. Hosler AS, Dharssi A (2010) Identifying retail food stores to evaluate the food environment. *Am J Prev Med* **39**, 41–4.

21. Toft U, Erbs-Maibing P, Glümer C (2011) Identifying fast-food restaurants using a central register as a measure of the food environment. *Scand J Public Health* **39**, 864–9.

22. Paquet C, Daniel M, Kestens Y, *et al.* (2008) Field validation of listings of food stores and commercial physical activity establishments from secondary data. *Int J Behav Nutr Phys Act* **5**, 58.
23. Cummins S, Macintyre S (2009) Are secondary data sources on the neighbourhood food environment accurate? Case-study in Glasgow, UK. *Prev Med* **49**, 527–8.
24. Bader MDM, Ailshire JA, Morenoff JD, *et al*. (2010) Measurement of the local food environment: a comparison of existing data sources. *Am J Epidemiol* **171**, 609–17.
25. Lake AA, Burgoine T, Stamp E, *et al.* (2012) The foodscape: classification and field validation of secondary data sources across urban/rural and socio-economic classifications in England. *Int J Behav Nutr Phys Act* **9**, 37.
26. Rossen LM, Pollack KM, Curriero FC (2012) Verification of retail food outlet location data from a local health department using ground-truthing and remote-sensing technology: assessing differences by neighborhood characteristics. *Health Place* **18**, 956–62.
27. Burgoine T, Harrison F (2013) Comparing the accuracy of two secondary food environment data sources in the UK across socio-economic and urban/rural divides. *Int J Health Geogr* **12**, 1–8.
28. Clary CM, Kestens Y (2013) Field validation of secondary data sources: a novel measure of representativity applied to a Canadian food outlet database. *Int J Behav Nutr Phys Act* **10**, 77.
29. Rummo PE, Gordon-Larsen P, Albrecht SS (2014) Field validation of food outlet databases: the Latino food environment in North Carolina, USA. *Public Health Nutr* 2014 **6**, 1–6.
30. Longacre MR, Primack BA, Owens PM, *et al* (2011) Public directory data sources do not accurately characterize the food environment in two predominantly rural states. *J Am Diet Assoc* **111**, 577–82.
31. Liese AD, Colabianchi N, Lamichhane AP, *et al.* (2010) Validation of 3 food outlet databases: completeness and geospatial accuracy in rural and urban food environments. *Am J Epidemiol* **172**, 1324–33.
32. Powell LM, Han E, Zenk SN, *et al.* (2011) Field validation of secondary commercial data sources on the retail food outlet environment in the U.S. *Health Place* **17**, 1122–31.
33. Seliske L, Pickett W, Bates R, *et al.* (2012) Field validation of food service listings: a comparison of commercial and online geographic information system databases. *Int J Environ Res Public Health* **9**, 2601–7.

34. City of Vancouver (2013) What Feeds Us: Vancouver Food Strategy. Retrieved February 4, 2017 from http://vancouver.ca/files/cov/vancouver-food-strategy-final.PDF.
35. Statistics Canada (2016) Population and Dwelling Count Highlight Tables, 2011 Census. Statistics Canada. Retrieved June 19, 2016, from http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/hlt-fst/pd-pl/Table- Tableau.cfm.
36. City of Vancouver (2015). Business Licences. Vancouver, BC. Retrieved October 20, 2015, from http://data.vancouver.ca/datacatalogue/businessLicence.htm.
37. Vancouver Coastal Health (2015). Inspection Reports. Vancouver, BC; 2015. Retrieved October 20, 2015, from http://www.vch.ca/your- environment/facility- licensing/residential-care/inspection- reports/.
38. Pitney Bowes Software (2012) Canada Business Data. Pitney Bowes Software Inc., Troy, New York.
39. DMTI Spatial, Inc. (2013) EPOI v2013.3; Retrieved May 15, 2015, from http://hdl.handle.net.ezproxy.library.ubc.ca/.
40. Ahmadi N, Black JL, Velazquez CE, *et al* (2015) Associations between socio-economic status and school-day dietary intake in a sample of grade 5-8 students in Vancouver, Canada. *Public Health Nutr* **18**, 764–773.
41. Velazquez CE, Black JL, Billette JMM, *et al.* (2015) A Comparison of Dietary Practices at or En Route to School between Elementary and Secondary School Students in Vancouver, Canada. *J Acad Nutr Diet* **115**, 1308-17.
42. Fleischhacker SE, Rodriguez DA, Evenson KR, *et al.* (2012) 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* **9,** 137.
43. Williams J, Scarborough P, Matthews A, *et al.* (2014) A systematic review of the influence of the retail food environment around schools on obesity-related outcomes. *Obes Rev* **15**, 359–74.
44. United States Census Bureau (2016) Introduction to NAICS: North American Industry Classification System Retrieved June 30, 2016, from http://www.census.gov/eos/www/naics/.
45. Bell N, Schuurman N, Oliver L, *et al.* (2007) Towards the construction of place-specific measures of deprivation: a case study from the Vancouver metropolitan area. *Can Geogr-Geogr Can* **51**, 444–461.

46. Bell N, Hayes MV (2012) The Vancouver Area Neighbourhood Deprivation Index (VANDIX): a census-based tool for assessing small-area variations in health status. *Can J Public Health* **103**, S28–S32.

47. Census Dictionary (2011) Dissemination Area (DA). Retrieved December 21, 2016, from https://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo021-eng.cfm.

48. BC Ministry of Education (2016) Schools. Vancouver, BC. Retrieved June 1, 2016, from https://catalogue.data.gov.bc.ca/dataset/bc-schools-school-locations.

49. Lucan SC, Maroko AR, Bumol J, *et al.* (2013) Business list vs ground observation for measuring a food environment: saving time or waste of time (or worse)? *J Acad Nutr Diet* **113**, 1332–9.

50. DMTI Spatial, Inc. (2013) CanMap Streetfiles, v2013.3. Markham, ON. Retrieved May 15, 2015, from http://hdl.handle.net.ezproxy.library.ubc.ca/.

51. Oliver LN, Schuurman N, Hall AW (2007). Comparing circular and network buffers to examine the influence of land use on walking for leisure and errands. *Int J Health Geogr* **6**, 41.

52. ESRI (2015) ArcGIS Desktop: Release 10.3.1. Redlands, CA.

53. Han E, Powell LM, Zenk SN, *et al.* (2012) Classification bias in commercial business lists for retail food stores in the U.S. *Int J Behav Nutr Phys Act* **9**, 46.

54. R Core Team (2016). *R: A Language and Environment for Statistical Computing.* Vienna, Austria. Available from: https://www.R-project.org.

55. Auchincloss AH, Moore KAB, Moore LV, *et al.* (2012) Improving retrospective characterization of the food environment for a large region in the United States during a historic time period. *Health Place* **18**, 1341–7.

56. Hoehner CM, Schootman M (2010) Concordance of commercial data sources for neighborhood-effects studies. *J Urban Health* **87**, 713–25.

57. Winkler WE (1990) String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Associatio,n* S354 – 369.

58. Sariyar M, Borg A (2010) The RecordLinkage package: Detecting errors in data. *R J* **2**, 61–67.

59. Newson R (2002) Parameters behind “nonparametric” statistics: Kendall’s tau, Somers’ D and median differences. *The STATA Journal* **2,** 454–64.

60. Signorell A (2016) DescTools: Tools for Descriptive Statistics. CRAN. Retrieved September 30, 2016, from https://cran.r-project.org/web/packages/DescTools/index.html.

61. McHugh ML (2012) Interrater reliability: the kappa statistic. *Biochem Med* **22** 276–282.

62. Lake AA, Burgoine T, Greenhalgh F, *et al.* (2010) The foodscape: classification and field validation of secondary data sources. *Health Place* **16**, 666–73.

63. Ma X, Battersby SE, Bell BA, *et al*. (2013) Variation in low food access areas due to data source inaccuracies. *Appl Geogr* **45**, 131–137.

64. Lebel A, Daepp MIG, Block JP, *et al.* (2017) Quantifying the foodscape: A systematic review and meta-analysis of the validity of commercially available business data. *PLoS ONE* **12**, e0174417.

65. Jones KK, Zenk SN, Tarlov E, Powell LM, Matthews SA, Horoi I (2017). A step-by-step approach to improve data quality when using commercial business lists to characterize retail food environments. *BMC Research Notes* **10**, 35.

Tables

Table 1. Classifications & definitions of dataset validity

Classification	Definition	Measurement[a]
Sensitivity	Proportion of outlets observed during ground-truthing that were listed in the secondary dataset	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
Positive Predictive Value (PPV)	Proportion of outlets listed in the secondary dataset that were observed during ground truthing	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
Concordance	Proportion of the total number of observed or listed outlets that were both listed in the secondary dataset and observed during ground-truthing	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$

[a]TP, True Positive; FP, False Positive; FN, False Negative

Table 2. Sources of data for food outlet locations in Vancouver, BC

Data Source	Description	Classifiers	Year
Gold Standard			
1) Ground-truthed primary data	Original data collected for this study; identified retailers within 800m buffers surrounding 26 Vancouver schools	Classification scheme (see Supplementary File 1)	2015
Municipal			
2) City of Vancouver Business Licenses	Records of businesses operating in the City of Vancouver; required under License By-Law No. 4450	Business Type Business Sub-type	2012 2015
3) Vancouver Coastal Health Inspections Lists	Health inspection records for restaurants, food stores, processors and other regulated facilities in the Vancouver Coastal Health service area.	Facility Type	2015
Commercial			
4) Pitney Bowes Software Canada Business Points	Geographic coordinates and attributes for businesses across Canada	NAICS[a] codes SIC[b] codes	2012
5) DMTI Spatial, Inc. Enhanced Points of Interest	Vector GIS database of recreational places and businesses across Canada	NAICS[a] codes SIC[b] codes	2013

[a]NAICS, North American Industry Classification System
[b]SIC, Standard Industrial Classification

Table 3. Sensitivity, positive predictive value (PPV), and concordance of two municipal and two commercial data sources compared with ground-truthed data (n = 455) for the locations of food outlets in Vancouver, BC

	Municipal		Commercial	
	Business Licenses	Vancouver Coastal Health	Canada Business Points	Enhanced Points of Interest

	2012	**2015**	**2015**	**2012**	**2013**
Sensitivity					
All Outlets	**0.58**	**0.69**	**0.54**	**0.41**	**0.39**
Limited Service	0.62	0.72	0.55	0.40	0.37
Convenience	0.65	0.75	0.60	0.46	0.48
Grocery	0.31	0.42	0.34	0.36	0.25
PPV					
All Outlets	**0.48**	**0.55**	**0.60**	**0.44**	**0.37**
Limited Service	0.46	0.51	0.66	0.54	0.38
Convenience	0.53	0.60	0.54	0.39	0.34
Grocery	0.53	0.75	0.52	0.28	0.46
Concordance					
All Outlets	**0.36**	**0.44**	**0.40**	**0.27**	**0.23**
Limited Service	0.36	0.43	0.43	0.30	0.23
Convenience	0.41	0.50	0.39	0.27	0.25
Grocery	0.24	0.37	0.26	0.19	0.19
N†					
All Outlets	**552**	**567**	**405**	**426**	**473**
Limited Service	361	375	225	197	264
Convenience	153	156	138	148	174
Grocery	38	36	42	81	35

†Total unique food outlets listed in each dataset located within 800m of 26 schools

Table 4. Results from bivariate logistic regressions examining the associations of commercial density or socioeconomic status and false positive (FP)† listings in each secondary data source

	Municipal			Commercial	
	Business Licenses		Vancouver Coastal Health	Canada Business Points	Enhanced Points of Interest
	2012	2015	2015	2012	2013
Commercial Density‡ Per 100 Outlets	0.96 (0.91 – 1.01)	0.95 (0.90 – 1.01)	1.02 (0.95 – 1.10)	1.05 (0.98 – 1.12)	1.05 (0.99 – 1.12)
VANDIX§					
low	–	–	–	–	–
medium	0.97 (0.70 - 1.33)	1.05 (0.76 – 1.44)	0.86 (0.59 – 1.25)	0.70* (0.50 – 0.99)	0.74 (0.53 – 1.03)
high	1.07 (0.79 – 1.47)	0.98 (0.72 – 1.35)	1.20 (0.82 – 1.75)	0.85 (0.60 – 1.21)	0.86 (0.61 – 1.21)
N Outlets‖	929	923	677	778	851

Odds ratios with 95% confidence intervals in parentheses.
†FP, false positive
‡Calculated in the 800m region surrounding each school
§Calculated in the dissemination area surrounding each school; "high" indicates most deprived.
‖N outlets in each secondary dataset; outlets in two buffer zones are counted twice
**P*<0•05, ***P*<0•01, ****P*<0•001

Table 5. Results from bivariate logistic regressions examining the associations of commercial density or socioeconomic status and false negative (FN)† listings in each secondary data source

	Municipal			Commercial	
	Business Licenses		Vancouver Coastal Health	Canada Business Points	Enhanced Points of Interest
	2012	**2015**	**2015**	**2012**	**2013**
Commercial Density‡ Per 100 Outlets	0.97 (0.91 – 1.03)	0.95 (0.89 – 1.01)	1.07* (1.01 – 1.14)	1.11** (1.04 – 1.18)	1.08* (1.01 – 1.15)
VANDIX§					
low	–	–	–	–	–
medium	1.25 (0.89 – 1.77)	1.11 (0.78 – 1.58)	0.95 (0.68 – 1.34)	0.67* (0.47 – 0.94)	0.84 (0.59 – 1.19)
high	1.08 (0.76 – 1.53)	0.93 (0.65 – 1.33)	1.35 (0.96 – 1.92)	0.93 (0.66 – 1.33)	1.10 (0.78 – 1.56)
N Outlets[c]	788	788	788	788	788

Odds ratios with 95% confidence intervals in parentheses.
†FN, false negative
‡Calculated in the 800m region surrounding each school
§Calculated in the dissemination area surrounding each school; "high" indicates most deprived.
||N outlets in each secondary dataset; outlets in two buffer zones are counted twice
*P<0•05, **P<0•01, ***P<0•001

Table 6. Kendall's Tau correlations between measures of the community nutrition environment surrounding schools (n = 26) evaluated with ground-truthed data and measures constructed from secondary data

	Municipal			Commercial	
	Business Licenses		Vancouver Coastal Health	Canada Business Points	Enhanced Points of Interest
	2012	2015	2015	2012	2013
Density within 800m of schools†					
All Outlets	0.87*** (0.80 – 0.94)	0.90*** (0.83 – 0.96)	0.87*** (0.77 – 0.97)	0.94*** (0.88 – 0.99)	0.90*** (0.85 – 0.96)
Limited Service	0.85*** (0.77 – 0.92)	0.87*** (0.80 – 0.94)	0.83*** (0.72 – 0.95)	0.86*** (0.77 – 0.95)	0.91*** (0.84 – 0.97)
Convenience	0.70*** (0.55 – 0.86)	0.72*** (0.55 – 0.89)	0.57*** (0.36 – 0.79)	0.64*** (0.43, 0.84)	0.76*** (0.63 – 0.89)
Grocery	0.78*** (0.66 – 0.90)	0.80*** (0.69 – 0.91)	0.74*** (0.62 – 0.87)	0.56*** (0.34, 0.77)	0.51** (0.30 – 0.71)
Proximity to schools‡					
All Outlets	0.61*** (00.37 – 0.84)	0.72*** (0.51 – 0.94)	0.70*** (0.39 – 1.00)	0.74*** (0.49 – 0.99)	0.73*** (0.45 – 1.01)
Limited Service	0.57*** (0.39 – 0.74)	0.58*** (0.39 – 0.77)	0.71*** (0.47 – 0.95)	0.63*** (0.40 – 0.86)	0.72*** (0.50 – 0.93)
Convenience	0.61*** (0.36 – 0.86)	0.63*** (0.41 – 0.86)	0.68*** (0.46 – 0.91)	0.59*** (0.37 – 0.81)	0.67*** (0.46 – 0.87)
Grocery	0.38** (0.12 – 0.65)	0.54*** (0.31 – 0.77)	0.39* (0.05 – 0.72)	0.31* (0.03 – 0.60)	0.39* (0.04 – 0.75)

Kendall's Tau with 95% CIs in parentheses

†evaluated with Tau-b due to ties; ‡evaluated with Tau-a

**P*<0•05, ***P*<0•01, ****P*<0•001