Public Health Nutrition: 20(15), 2649–2659 doi:10.1017/S1368980017001744

Assessing the validity of commercial and municipal food environment data sets in Vancouver, Canada

Madeleine IG Daepp[1,*] and Jennifer Black[2]

[1]Department of Urban Studies and Planning, Massachusetts Institute of Technology, 9-555, 77 Massachusetts Avenue, Cambridge, MA 02139, USA: [2]Food, Nutrition and Health, Faculty of Land & Food Systems, University of British Columbia, Vancouver, BC, Canada

Submitted 10 February 2017: Final revision received 29 May 2017: Accepted 16 June 2017: First published online 17 August 2017

Abstract

Objective: The present study assessed systematic bias and the effects of data set error on the validity of food environment measures in two municipal and two commercial secondary data sets.

Design: Sensitivity, positive predictive value (PPV) and concordance were calculated by comparing two municipal and two commercial secondary data sets with ground-truthed data collected within 800 m buffers surrounding twenty-six schools. Logistic regression examined associations of sensitivity and PPV with commercial density and neighbourhood socio-economic deprivation. Kendall's τ estimated correlations between density and proximity of food outlets near schools constructed with secondary data sets *v.* ground-truthed data.

Setting: Vancouver, Canada.

Subjects: Food retailers located within 800 m of twenty-six schools

Results: All data sets scored relatively poorly across validity measures, although, overall, municipal data sets had higher levels of validity than did commercial data sets. Food outlets were more likely to be missing from municipal health inspections lists and commercial data sets in neighbourhoods with higher commercial density. Still, both proximity and density measures constructed from all secondary data sets were highly correlated (Kendall's $\tau > 0{\cdot}70$) with measures constructed from ground-truthed data.

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

Keywords
Built environment
Public health
Food environment
Data validation

Many countries including the USA 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, in part to reduce children's access to and intake of minimally nutritious foods[9].

In Canada, the only G8 country without a federal school lunch programme, 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 data on community nutrition environments.

The majority of peer-reviewed studies on the community nutrition environment obtain food outlet data from: (i) 'ground-truthing', the systematic surveying of a region to identify and classify food retailers; (ii) commercial database providers; or (iii) government sources[11].

*Corresponding author: Email mdaepp@mit.edu

Ground-truthing is considered the gold standard[12,13], but the approach is resource-intensive and infeasible for the assessment of past years. Commercial data sets 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 data sets 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 data sets 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 nineteen 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 data sets, evaluating associations between error rates and neighbourhood 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 with 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 data sets.

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 neighbourhood socio-economic status[22–28] or neighbourhood racial demographics[24,26,29], but several studies show evidence of systematic bias in relation to urbanicity or commercial density. Four studies in the USA identified statistically significant differences in validity levels in association with urbanicity or density[24,30–32] although no significant associations were identified in two UK studies[25,27] and the direction of the association varies across studies. But the data sets examined in the aforementioned studies are often specific to the USA 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.

Table 1 Classifications and definitions of data set validity

Classification	Definition	Measurement
Sensitivity	Proportion of outlets observed during ground-truthing that were listed in the secondary data set	$\frac{TP}{TP+FN}$
Positive predictive value (PPV)	Proportion of outlets listed in the secondary data set that were observed during ground-truthing	$\frac{TP}{TP+FP}$
Concordance	Proportion of the total number of observed or listed outlets that were both listed in the secondary data set and observed during ground-truthing	$\frac{TP}{TP+FP+FN}$

TP, true positive; FN, false negative; FP, false positive.

The present 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: (i) to assess the validity of two commercial and two municipal secondary data sources in comparison with ground-truthed data; (ii) to test each data set for evidence of systematic bias in association with neighbourhood socio-economic deprivation or commercial density; and (iii) to compare community nutrition environment measures constructed from secondary commercial and municipal data sets with gold-standard ground-truthed data. The first objective provides results that can be compared with findings from previous data validation research in other countries and cities, while the second and third objectives offer novel methods to help researchers understand how over- or undercounting of outlet listings may be affecting community nutrition environment research.

Methods

Data

The present 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: (i) ground-truthed primary data; (ii) (municipal) Business Licences[36]; (iii) (municipal) Vancouver Coastal Health inspections lists[37]; (iv) (commercial) Pitney Bowes Software's Canada Business Points[38]; and (v) (commercial) DMTI Spatial, Inc.'s Enhanced Points of Interest[39]. An overview of these data sets is provided in Table 2.

The ground-truthed data were obtained through systematic surveying between 29 June and 30 September 2015. A purposive sampling approach was used to select

https://doi.org/10.1017/S1368980017001744 Published online by Cambridge University Press

Table 2 Sources of data for food outlet locations in the city of Vancouver, Canada

Data source	Description	Classifiers	Year
Gold standard			
1. Ground-truthed primary data	Original data collected for the present study; identified retailers within 800 m buffers surrounding twenty-six Vancouver schools	Classification scheme (see online supplementary material, Supplementary File 1)	2015
Municipal			
2. City of Vancouver Business Licences	Records of businesses operating in the City of Vancouver; required under License By-Law No. 4450	Business Type Business Subtype	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 codes SIC codes	2012
5. DMTI Spatial, Inc. Enhanced Points of Interest	Vector GIS database of recreational places and businesses across Canada	NAICS codes SIC codes	2013

GIS, geographic information system; NAICS, North American Industry Classification System; SIC, Standard Industrial Classification.

twenty-six schools across the Vancouver School Board's six geographic sectors (detailed in previous papers[40,41]) located in neighbourhoods with diverse levels of commercial density and socio-economic status.

Following a surveying protocol adapted from similar research[42] (see online supplementary material, Supplementary File 1), two researchers visited all commercial streets located within an 800 m 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 using Cohen's κ to assess inter-rater reliability in surveyors' store classifications.

The two municipal data sets – Business Licences 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 Licences, historical records allowed the present 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 inspections lists comprised food-service establishments, food stores and food processors in the city of Vancouver, classified by 'service type'. The Business Licences data were similar, although 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 data sets 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 USA and Mexico[44].

The 2015 Business Licence data[36] were also used to measure commercial density – defined as the total number of businesses of any type located within the 800 m buffer surrounding schools – based on their performance in the validation study (see 'Results'). Relative socio-economic deprivation was assessed with the Vancouver Area Neighbourhood 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 of lone-parent families, average income, proportion of home owners and labour force participation rate) obtained from the 2006 Census of Canada[45,46]. For the current study, the VANDIX was calculated for dissemination areas, 400- to 700-person regions comprising the smallest available census geography[47]. The twenty-six schools examined in the 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

https://doi.org/10.1017/S1368980017001744 Published online by Cambridge University Press

dissemination area directly surrounding the school. 'High' scores indicate the most socio-economically deprived and 'low' scores indicate the least deprived areas.

Cleaning and classification of food outlets

The secondary data sets 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 data sets, outlets located within 800 m line-based buffers[51] surrounding each of the twenty-six 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].

The present study compared three classes of outlets: (i) limited-service food outlets, restaurants or coffee shops where customers order at a counter and pay before consuming food or beverages; (ii) convenience stores, which included retail stores primarily offering snack foods or beverages, possibly attached to a pharmacy or gas station; and (iii) 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 the online supplementary material, Supplementary File 1). For the 2012 and 2015 Business Licences, '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). The present study thus supplemented the 'Facility Type' and SIC/NAICS codes with the application of a name-based classification scheme (see online supplementary material, Supplementary File 2) following previous studies[27,28].

Outlet matching approach

Two approaches were applied to match outlets in the commercial and municipal data sets with outlets in the ground-truthed data set. 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 version 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 data set, geographic coordinates were used to identify all stores in the secondary data set located within 100 m 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 100 m 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 data sets was assessed with the ground-truthed data set serving as the gold standard. For each of the commercial and municipal secondary data sets, a matched store was considered a true positive (TP) if it was listed in both the secondary data set 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 data set. Sensitivity, PPV and concordance (defined in Table 1) were then calculated as measures of the validity of each secondary data source. A listing was considered a true positive even if it had a different name in the secondary data set from that in the ground-truthed data, if the two listings included identical addresses and classifications. As a sensitivity analysis, 'strict' true positives were calculated omitting stores with highly dissimilar names.

Second, logistic regressions examined whether the odds of false positives or false negatives increased in association with neighbourhood socio-economic deprivation or commercial density to assess systematic biases. Regressions were fitted for all stores in the ground-truthed data set 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 data set 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 800 m 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 data sets were compared with measures from the ground-truthed data set using Kendall's τ, a non-parametric measure of correlation[59]. ArcGIS was used to calculate density (the total number of outlets located within each 800 m 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 $P < 0{\cdot}05$ was used for determining statistical significance for all analyses.

Results

Assessment of data set validity

Table 3 reports the counts of food outlets for each of the municipal and commercial secondary data sets and results from comparisons between ground-truthed and secondary data sources. Ground-truthing identified 267 limited-service food outlets, 124 convenience stores and sixty-four grocery stores or supermarkets located within 800 m of the sample of twenty-six schools. For outlets classified by two surveyors, percentage agreement was 91 % and Cohen's κ was 0·88, indicating strong inter-rater reliability[61].

The 2015 Business Licences had the highest overall scores for sensitivity, identifying 69 % of the ground-truthed stores. This data set'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 Licences 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 data sets in 2015 were higher than those obtained for either of the two commercial data sets in all cases except for the sensitivity estimates for grocery stores.

With strict name matching, the 2015 Business Licence data lost twenty-eight outlet matches, leading its sensitivity to drop to 0·62 while PPV decreased to 0·50. The 2012 Business Licence data lost thirty-four matches (sensitivity = 0·51, PPV = 0·42), the Vancouver Coastal Health data lost fifteen matches (sensitivity = 0·50, PPV = 0·57) and the Enhanced Points of Interest lost twenty-seven matches (sensitivity = 0·33, PPV = 0·32). Canada Business Points had the fewest matched outlets with different names, with just seven outlets failing the stricter name-based standard (sensitivity = 0·40, PPV = 0·42). Regardless of the approach to matching store names, the municipal data sets performed better in terms of overall sensitivity, PPV and concordance than did the commercial data sets.

Assessment of systematic bias

Tables 4 and 5 report findings from bivariate logistic regression analyses examining associations of commercial density and socio-economic status with false positive and false negative listings in each secondary data set.

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 the city of Vancouver, Canada

	Municipal			Commercial	
	Business Licences		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

†Number of total unique food outlets listed in each data set located within 800 m of twenty-six schools.

https://doi.org/10.1017/S1368980017001744 Published online by Cambridge University Press

Table 4 Results from bivariate logistic regression analyses examining the associations of commercial density or socio-economic status with false positive (FP) listings in each secondary data source, city of Vancouver, Canada

	Municipal						Commercial			
	Business Licences				Vancouver Coastal Health		Canada Business Points		Enhanced Points of Interest	
	2012		2015		2015		2012		2013	
	OR	95 % CI	OR	95 % CI	OR	95 % CI	OR	95 % CI	OR	95 % CI
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

VANDIX, Vancouver Area Neighbourhood Deprivation Index.
*$P < 0·05$.
†Calculated in the 800 m region surrounding each school.
‡Calculated in the dissemination area surrounding each school; 'high' indicates most deprived.
§Number of outlets in each secondary data set; outlets in two buffer zones are counted twice.

Table 5 Results from bivariate logistic regression analyses examining the associations of commercial density or socio-economic status with false negative (FN) listings in each secondary data source, city of Vancouver, Canada

	Municipal						Commercial			
	Business Licences				Vancouver Coastal Health		Canada Business Points		Enhanced Points of Interest	
	2012		2015		2015		2012		2013	
	OR	95 % CI	OR	95 % CI	OR	95 % CI	OR	95 % CI	OR	95 % CI
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
$n_{outlets}$§		788		788		788		788		788

VANDIX, Vancouver Area Neighbourhood Deprivation Index.
*$P < 0·05$, **$P < 0·01$.
†Calculated in the 800 m region surrounding each school.
‡Calculated in the dissemination area surrounding each school; 'high' indicates most deprived.
§Number of outlets in each secondary data set; outlets in two buffer zones are counted twice.

Table 6 Kendall's τ correlations between measures of the community nutrition environment surrounding schools (n 26) evaluated with ground-truthed data and measures constructed from secondary data, city of Vancouver, Canada

	Municipal						Commercial			
	Business Licences				Vancouver Coastal Health		Canada Business Points		Enhanced Points of Interest	
	2012		2015		2015		2012		2013	
	Kendall's τ	95 % CI	Kendall's τ	95 % CI	Kendall's τ	95 % CI	Kendall's τ	95 % CI	Kendall's τ	95 % CI
Density within 800 m 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

*$P < 0·05$, **$P < 0·01$, ***$P < 0·001$.
†Evaluated with τ_b due to ties.
‡Evaluated with τ_a.

Neighbourhood socio-economic 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 (*v.* true positive) listings in all secondary data sets except the municipal Business Licences data. An increase of 100 stores within an 800 m 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 data sets

Across all secondary data sources, density measures were highly correlated with measures from the ground-truthed data (Kendall's $\tau_b \geq 0·87$ for all outlets). The strength of the correlations between proximity measures from secondary and ground-truthed data was slightly lower, with Kendall's τ_a falling between 0·61 for the 2012 Business Licences (95 % CI 0·37, 0·84) and 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. Although both commercial data sets performed comparably to the municipal data sets 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

The present study assessed the validity of two municipal and two commercial community nutrition environment data sources compared with a gold standard, ground-truthed data set 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 data sets, and offering insight from a large Canadian city.

The study found that all data sets were subject to high levels of error: data sets both (i) failed to include at least 20 % of outlets observed in the field and (ii) consisted at minimum of 25 % listings not found in the field. The 2015 Business Licence data and the Vancouver Coastal Health data had sensitivity and PPV values in the range of

https://doi.org/10.1017/S1368980017001744 Published online by Cambridge University Press

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, USA[42], and to a sensitivity estimate (0·66) for city council data in Newcastle, UK[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 data sets 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 the present 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 the present study, suggesting that the difference in results may be explained by the depreciation of data quality over time.

Nevertheless, the current study found that overall both municipal data sets outperformed commercial data sets in measures of validity, even when the 2012, rather than 2015 Business Licence data were 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 data sets examined here[43]. Our study suggests that municipal data sets can provide adequate alternatives that may offer higher-quality data than many of the data sets on which the community nutrition environment literature currently relies.

The present study also evaluated associations between neighbourhood socio-economic deprivation and commercial density with the odds of incorrect listings. This examination was valuable because systematic error in data sets could bias research findings: if data sets consistently fail to identify existing food retailers in low-income neighbourhoods, 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 neighbourhood socio-economic deprivation and the odds of false negative or false positive associations – are reassuring for researchers because they suggest that results regarding socio-economic 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 socio-economic deprivation and levels of commercial data set validity[22,23,26–28].

The present study did, however, find positive associations between the odds of false positive listings and commercial density in three of four data sets. 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 neighbourhoods, 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 alternative measure of urbanicity, which did not find evidence of significant associations between that measure and odds of false positives or false negatives in any data set. We did not have access to data regarding business turnover, but hypothesize that more commercially dense Vancouver neighbourhoods (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 data sets were highly correlated with measures from ground-truthed data. This observation is consistent with findings of two other known studies examining the effect of data set validity on community nutrition environment measures: Ma *et al.*[63] found that measures of food deserts, which are low-income areas where residents lack access to grocery stores or supermarkets, created from two commercial data sets (InfoUSA and Dun & Bradstreet) had 93·5% concordance with comparable measures obtained from the US 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 data set (InfoUSA) had 86·9% correlation with estimates calculated from a gold standard data set (Boston Inspectional Services Department). The high levels of undercounting and overcounting estimated with low sensitivity and PPV, respectively, may offset one another, resulting in data that remain 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 data sets for research on the community nutrition environment[64].

Several notable limitations of the present 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 the study obtained for commercial data sets (collected in 2012 and 2013) in comparison with the municipal data sets, which were collected immediately after the completion of ground-truthing in 2015. However, the inclusion of both current (2015) and historical (2012) Business Licence data suggests that depreciation explains only part of the difference in validity. The two commercial data sets 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 Licences for 2012. Additionally, findings may not be generalizable to other cities because of variance in municipal data set quality, and the findings may overestimate validity for studies that do not follow the data cleaning and classification protocols used in the current 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 data sets. 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 the 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, the present 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 data sets and ground-truthed data could be replicated with data sets 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 data sets examined in the present study scored relatively poorly across validity measures. Three of the four data sets also had evidence of systematic bias in association with commercial density, although no data sets were systematically more likely to over- or undercount outlets in relation to neighbourhood socio-economic status. Nevertheless, community nutrition environment measures constructed from both municipal and commercial data sources were highly correlated with ground-truthed measures, suggesting that data sets with low validity scores may still offer reliable measures of community nutrition environment exposure.

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

Acknowledgements

Acknowledgements: Koharu Chayama and Cayley Velazquez assisted with the ground-truthing of food outlets in Vancouver. The authors would also like to thank Carol McAusland and Nadine Schuurman for guidance and comments. *Financial support:* This study received funding from the Canadian Institutes of Health Research (grant number FRN 119577). In addition, M.I.G.D. was funded by the University of British Columbia Li Tze Fong Fellowship (grant number #4895). The funding agencies had no role in the design, analysis or writing of this article. *Conflict of interest:* The authors declare that they have no competing interests. *Authorship:* Both J.B. and M.I.G.D. contributed to study design. J.B. was the principal investigator and supervised the research. M.I.G.D. developed and led the ground-truthing protocol, sourced secondary data sets, performed data analyses and drafted the manuscript. J.B. contributed to manuscript writing and editing, and both M.I.G.D. and J.B. reviewed the manuscript and approved the final version. *Ethics of human subject participation:* This study did not include human subjects research.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/S1368980017001744

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**, 161–175.
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**, 1172–1187.
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. Administrative Publication* no. AP-036. Washington, DC: US Department of Agriculture, Economic Research Service; available at 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. http://www.ca-ilg.org/sites/main/files/file-attachments/resources__hfhc_short_final.pdf (accessed May 2017).
6. Zenk SN, Thatcher E, Reina M *et al.* (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 *Local Food Environments: Food Access in America*, pp. 191–192 [KB Morland, editor]. Boca Raton, FL: CRC Press.
7. Mair JS, Pierce MW & Teret SP (2005) *The Use of Zoning to Restrict Fast Food Outlets: A Potential Strategy to Combat Obesity*, pp. 51–53. Baltimore, MD: The Center for Law and the Public's Health, Johns Hopkins & Georgetown Universities; available at 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 (Millwood)* **28**, w1088–w1097.
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**, 1240–1247.
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 *Local Food Environments: Food Access in America*, pp. 205–226 [KB Morland, editor]. Boca Raton, FL: CRC Press.
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**, 462–473.
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]. http://hdl.handle.net.ezproxy.library.ubc.ca/11272/NBRIL (accessed June 2016).
15. DMTI Spatial, Inc. (2006) Enhanced Point of Interest Layers [2006]. http://hdl.handle.net.ezproxy.library.ubc.ca/11272/KDY86 (accessed June 2016).
16. DMTI Spatial, Inc. (2009) Enhanced Point of Interest Layers [v.2009.3] http://hdl.handle.net.ezproxy.library.ubc.ca/11272/JGQ3B (accessed June 2016).
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**, 1384–1391.
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**, 41–44.
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**, 864–869.
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**, 527–528.
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**, 609–617.
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**, 956–962.
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**, 2.
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 (2015) Field validation of food outlet databases: the Latino food environment in North Carolina, USA. *Public Health Nutr* **18**, 977–982.
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**, 577–582.
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**, 1324–1333.
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 US. *Health Place* **17**, 1122–1131.

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**, 2601–2607.
34. City of Vancouver (2013) What feeds us: Vancouver Food Strategy. http://vancouver.ca/files/cov/vancouver-food-strategy-final.PDF (accessed February 2017).
35. Statistics Canada (2016) Population and Dwelling Count Highlight Tables, 2011 Census. http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/hlt-fst/pd-pl/Table-Tableau.cfm (accessed June 2016).
36. City of Vancouver (2015) Business Licences. http://data.vancouver.ca/datacatalogue/businessLicence.htm (accessed October 2015).
37. Vancouver Coastal Health (2015) Inspection Reports. http://www.vch.ca/your-environment/facility-licensing/residential-care/inspection-reports/ (accessed October 2015).
38. Pitney Bowes Software (2012) *Canada Business Data*. Troy, NY: Pitney Bowes Software Inc.
39. DMTI Spatial, Inc. (2013) EPOI v2013.3. http://hdl.handle.net.ezproxy.library.ubc.ca/ (accessed May 2015).
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**, 1308–1317.
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**, 359–374.
44. US Census Bureau (2016) North American Industry Classification System: Introduction to NAICS. http://www.census.gov/eos/www/naics/ (accessed June 2016).
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* **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**, 8 Suppl. 2, S28–S32.
47. Census Dictionary (2011) Dissemination Area (DA). https://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo021-eng.cfm (accessed December 2016).
48. British Columbia Ministry of Education (2016) BC Schools – School Locations. https://catalogue.data.gov.bc.ca/dataset/bc-schools-school-locations (accessed June 2016).
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**, 1332–1339.
50. DMTI Spatial, Inc. (2013) CanMap Streetfiles, v2013.3. http://hdl.handle.net.ezproxy.library.ubc.ca/ (accessed May 2015).
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. Environmental Systems Research Institute, Inc. (2015) *ArcGIS Desktop: Release 10.3.1*. Redlands. CA: ESRI.
53. Han E, Powell LM, Zenk SN *et al.* (2012) Classification bias in commercial business lists for retail food stores in the US. *Int J Behav Nutr Phys Act* **9**, 46.
54. R Core Team (2016) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing; available at 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**, 1341–1347.
56. Hoehner CM & Schootman M (2010) Concordance of commercial data sources for neighborhood-effects studies. *J Urban Health* **87**, 713–725.
57. Winkler WE (1990) String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, pp. S354–S369. Alexandria, VA: American Statistical Association.
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. *STATA J* **2**, 454–464.
60. Signorell A (2016) DescTools: Tools for Descriptive Statistics. https://cran.r-project.org/web/packages/DescTools/index.html (accessed September 2016).
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**, 666–673.
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 *et al.* (2017) A step-by-step approach to improve data quality when using commercial business lists to characterize retail food environments. *BMC Res Notes* **10**, 35.