Spatial Aggregation Method for Anonymous Surveys

Article *in* Transportation Research Record Journal of the Transportation Research Board · January 2016

DOI: 10.3141/2598-04

4 authors, including:

Nima Amini
UNSW Sydney
9 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Lauren M. Gardner
UNSW Sydney
74 PUBLICATIONS **426** CITATIONS
SEE PROFILE

Spatial Aggregation Method for Anonymous Surveys

Case Study for Associations Between Urban Environment and Obesity

Nima Amini, Taha Rashidi, Lauren Gardner, and S. Travis Waller

Obesity and other chronic diseases are becoming more prevalent in affluent countries such as Australia. Researchers are trying to understand and combat this trend. One related growing stream of research explores the role of the built environment and transport system on an individual's weight. However, results from many studies conducted have been contradictory. A primary cause of these contradictions is due to how neighborhood areas are defined, which directly affects how the built environment variables are calculated in geographic information systems. The potential impacts on regression analysis resulting from different data aggregation methods are well documented in spatial studies, geography, and regional planning fields, and the problem is primarily referred to as the modifiable aerial unit problem. In this paper, the focus is on reducing the error caused by the modifiable aerial unit problem by introducing a new data aggregation method. Individual health and lifestyle data are obtained from the survey of households, income, and labor dynamics in Australia, and the relationship between the built environment and obesity is evaluated by using a discrete choice model. The proposed aggregation method is evaluated across three spatial scales and compared against a conventional data aggregation method (i.e., using predefined administrative boundaries such as census tracts). The results reveal a stronger relationship between land use variables and obesity when the proposed aggregation method is implemented. This paper is relevant primarily to researchers because it provides an improved aggregation method to deal with some privacy restrictions of surveys. It is also relevant to practitioners and policy makers by its quantification of the association between specific built environment variables and obesity.

Obesity and other chronic diseases are becoming more prevalent in affluent countries such as Australia and America (*1–4*), thus researchers are trying to understand and combat this trend (*5*). One related growing stream of research explores the role of the built environment on physical activity (*6–9*) and the individual's health (*10, 11*). However, results from many of the studies conducted have been contradictory (*12*). One of the primary causes of these contradictions can be attributed to the way the neighborhoods are defined, which directly affects how the built environment variables are calculated in geographical information systems (GIS) (*12–14*).

Research Center for Integrated Transport Innovation, School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales, 2052, Australia. Corresponding author: N. Amini, n.amini@unsw.edu.au.

Transportation Research Record: Journal of the Transportation Research Board, No. 2598, Transportation Research Board, Washington, D.C., 2016, pp. 27–36.
DOI: 10.3141/2598-04

Spatial analysis is the foundation of many studies within the transportation field. Spatial analysis implicitly defines some of the underlying assumptions of a study. One typical spatial analysis is the aggregation of data within a geographical boundary. For example, in the traditional four-step models the boundaries for traffic analysis zones (TAZs) are defined as varying shapes and sizes by assuming that people within a particular TAZ have similar transportation behavior. It is easy to imagine that inappropriate selection of the TAZ boundaries can have significant consequences in the final outputs of the transportation model.

Urban analysis studies typically aggregate survey data at predefined administrative boundaries (e.g., census tract, state boundary, country borders). Sometimes a boundary is selected intentionally in order to conduct the evaluation at a particular administrative or geographical level; however, other times survey data are preaggregated to relatively small predefined geographical boundaries (because of confidentiality restrictions) before the survey results are released. In the latter case, the predefined geographical boundaries are typically defined for a different purpose and are not necessarily representative of the phenomena being evaluated, which may result in errors (*15*). The geographical boundary used to aggregate the original data can influence the results of some statistical analyses. This influence was reported as early as 1934 (*16*) and continues to be an issue primarily because of the confidentiality restriction of surveys (*17, 18*). There are three major spatially related challenges in analyzing the association between the urban environment and the behavior of people. First, the zoning system definition can affect the influence of built form variables. This issue is targeted in the current paper. Second, the spatial correlation between the zones can influence the precision of the models. Third, an endogeneity problem can result from the unobserved behavior of people when deciding about their residence (*12, 18, 19*). This third challenge is also partially considered in this paper.

The modifiable aerial unit problem (MAUP) is defined as "a problem arising from the imposition of artificial units of spatial reporting on continuous geographical phenomenon resulting in the generation of artificial spatial patterns" (*20*). The MAUP is typically described in the literature as composed of scalar and zonal effects (*18*). The scalar effect relates to the chosen size of the zone used to aggregate the data; for example, selecting larger zones tends to increase internal variance and decrease external variance of the aggregated variable (*13*). The zonal effect relates to the shape of the boundary used to aggregate the data. Evaluation of the MAUP has shown that some derived statistics such as means and variances are resistant to the aggregation method in contrast to dramatic effects exhibited in regression coefficients and correlation statistics (*18*).

The MAUP is particularly evident in studies that attempt to capture the urban characteristics surrounding the residential location of an individual. The typical hypothesis of such studies is that the characteristic of the urban environment surrounding the residential location of an individual is associated with some phenomenon (for example, the level of active transport or the probability of being obese). The ideal method of capturing the boundary of the urban area that has an influence on the behavior of an individual is by considering the perceptual map of the area surrounding their place of residence (this area is henceforth referred to as the "neighborhood area"). For example, an individual may have preferences for certain routes or have limited awareness of the built environment surrounding their place of residence. Since perceptual maps of individuals are not typically captured in surveys, the majority of studies assume that the most appropriate way to define the neighborhood area is by drawing a buffer around each individual's residential address. Sometimes these boundaries are defined by using the street network (typically 500 m in every direction along the road network). The more recent studies indicate the necessity of evaluating each urban variable at different buffer distances.

For example, Zhang and Kukadia used eight different aggregation methods to examine the MAUP effect of the scale and geographical boundary definition (*12*). They found major differences in the significance of the association between the built variables and mode choice across the eight aggregation methods. Similarly, studies examining the effects of the MAUP when the association between built environment variables and health outcomes is evaluated have shown that adopting different aggregation methods for defining the neighborhood area results in major differences in the significance of the regression. More important, the method can even affect the direction of the association between urban variables and active transport or health (*8, 13, 14*). These studies conclude that the aggregation zone should be built around each individual's residential location. Leal and Chaix indicate that the majority of studies continue to use predefined administrative boundaries as the neighborhood area possibly because the residential address of an individual is typically preaggregated to predefined geographical boundaries because of the confidentiality restrictions of the surveys (*10*). It is postulated that without a careful definition of the neighborhood area, the MAUP may be introduced in the analysis; this could be the reason for some of the differences in the association of the built environment and health outcomes across studies. In fact, few studies have considered the MAUP in reporting their findings of such definitions of the neighborhood area (*21*).

One solution has been to aggregate urban variables by using uniform grids. This method has shown improvement in the significance of the correlates between urban environment and mode choice (*12*). Other solutions include statistical corrections for specific urban variables such as dividing a land use mix variable by the area within the predefined geographical boundaries (*22*). However, these solutions do not apply to studies in which the residential addresses are preaggregated or only apply to a specific built environment variable. Solutions for the MAUP are not widely prevalent because (*a*) researchers have only begun to unpack the effects of the MAUP on analysis and (*b*) few generic and practical solutions exist (*23*). The MAUP is an active area of research and new methodologies are required to address the problem under difference circumstances (*13*).

The contributions of this paper are to (*a*) define a new aggregation method for the neighborhood area to reduce the MAUP when the residential location of individuals is preaggregated at relatively small zones, (*b*) collate and process a large set of GIS data to describe the urban characteristics of Sydney, Australia (to the authors' knowledge, this work has not been done for Sydney previously), (*c*) examine the association of the urban characteristics and obesity by using a logit model, and (*d*) compare the results obtained for the proposed aggregation method with those for a conventional aggregation method (i.e., using the predefined zone as the neighborhood area).

DATA

This section describes the study area, data source, and processing procedures used in this study. The study area (Figure 1) is defined by the Sydney Statistical Division, which is a large administrative zone that encompasses the city of Sydney, Australia. Only the urbanized areas were included in this study. The urban areas are based on the definition provided by the Australian Bureau of Statistics (ABS) and are composed of the urban centers with population of greater than 100,000 (*24*). The study area includes 2,653 participants who responded to the survey (referred to here as "respondents") within 639 census collection districts (CCDs).

A significant data acquisition effort was conducted to obtain the relevant GIS data from multiple state and federal agencies for the study area. The GIS data were subsequently processed in ArcGIS to calculate a number of built environment variables (introduced in the next section) that represented the urban characteristics of Sydney, Australia. The data processing was primarily built on similar work conducted in Adelaide, Australia, by Coffee (*11*); however, a number of new GIS procedures were developed in this study for further refinement when there were differences between the raw data received from different Australian agencies. However, for brevity only the properties of the final variables are discussed here.

Built Environment Variables

The "six Ds" principle, which accounts for the main ways in which the built environment is expected to influence travel behavior, was used as a guideline for the selection and development of the built environment variables included in this study. The six Ds principle is composed of density, diversity, design, destination, distance to transit, and demand management (*25, 26*). As in previous studies, it is hypothesized that an urban area that is more conducive to active transport is more likely to result in health benefits such as a reduction in obesity (*11*).

Density and diversity are considered by the following three built environment variables obtained from the 2011 census: population density, dwelling density, and employment density. Population density and dwelling density are obtained from the ABS and were preaggregated to Statistical Area 1. In order to convert these two variables to the CCD zones, the proportion of the residential land use overlapping the two boundaries was calculated. Employment density is obtained from the Bureau of Transport Statistics and preaggregated at Travel Zone 2011 (TZ11). This variable was converted to CCDs by averaging the density of all TZ11s that were intersecting or adjacent to the CCDs. The employment density was aggregated at a larger scale since other studies have shown that densities of services and retail jobs are significant at larger areas (*12, 27*).

The design principle is represented by four built environment variables: proportion of intersections with four or more legs, average block length, count of signalized intersections, and major-road density.

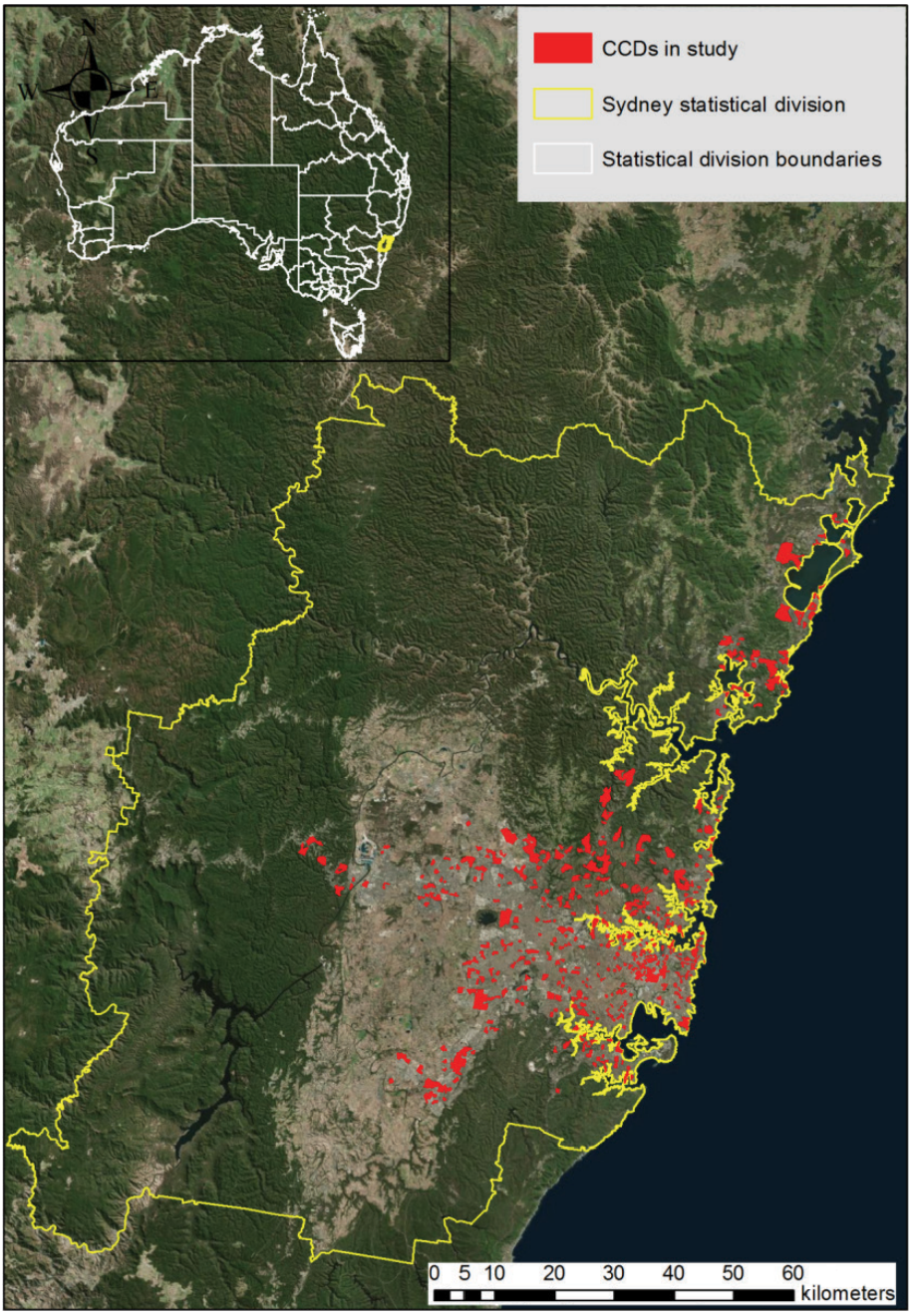

FIGURE 1 Study area: CCDs included in this study (shown in red).

The GIS data for the road network are obtained from the New South Wales (NSW) Land and Property Information (LPI), and the location of the signalized intersections is obtained from the Roads and Maritime Services. The average block length is defined as the length of all roads in a neighborhood area divided by the total number of intersections in the neighborhood area. The major-road density was calculated by dividing the total length of all major roads within a neighborhood area by the total area of the neighborhood area. The definition of major roads is based on the road hierarchy classification provided by LPI, such that roadways that are classified by LPI as expressways, arterials, and collector roads are considered major roadways (*28*).

The destination principle was considered by using the variable "distance to supermarket." A code was developed in Python to obtain the walking distance to the nearest Coles, Woolworths, or ALDI supermarkets from Google Maps API. These are the three major supermarkets in Australia that supply fresh food (*29*). It is assumed that smaller fresh food outlets are visited relatively less often compared with the three major supermarkets, since they contain less variety of stock and are typically more expensive. The actual measure used is the inverse of the distance to the closest of the three major supermarkets. Distances greater than 5 km are set to zero, since the likelihood of walking distances longer than 5 km is very low. This definition is based on the methodology adopted by Walkscore.com (*30*).

The "distance to transit" principle is measured separately for buses and trains and considers both distance to stops and the regularity of the service. A total of four built environment variables are developed for the distance to transit: number of bus stops, number of train stops, number of bus services, and number of train services. The transit data are referred to as the General Transit Feed Specification and are maintained and provided by the Transport for NSW. The number of stops is counted within the neighborhood area. The number of services is counted by adding the number of daily services (operating on weekdays) at each stop within the neighborhood area. The train services include the ferry services as well since train and ferry services are considered to be relatively similar as compared with bus services.

Health and Lifestyle Survey

The health and demographic variables used in this study came from the Household, Income and Labour Dynamics in Australia (HILDA) survey, which is a panel study that started in 2001 with 7,682 households across Australia. The household response rate of 66% is comparable with that of other international panel studies (*31, 32*). The data are collected annually with face-to-face interviews and a self-completion questionnaire. The focus of the HILDA survey is on family, household formation, income, and work. Every year the survey focuses on a specific area of life. The 2009 survey (ninth wave) focused on health, which is also the focus of this study (*33*).

The variables used from the HILDA survey included body mass index (BMI), gender, age, number of exercise sessions per week, vegetable servings consumed per week, index of relative socioeconomic disadvantage 2011 (IRSD), and place of residence preaggregated to CCD level. Obesity was defined by using a binary variable that was set to 1 if the self-reported BMI is larger than or equal to 30 kg/m^2 and zero otherwise.

The IRSD was included because other studies have shown an association between neighborhoods with low socioeconomical position and health outcomes (*34*). The IRSD is obtained from attributes such as low income, low educational attainment, and high unemployment by using principal component analysis. A low score on this scale indicates more disadvantage, and a high score indicates less disadvantage (*35, 36*). The IRSD is preaggregated at Statistical Area 1 and was converted to a CCD by using the proportion of residential area overlapping across the two zones.

Data Filters

In order to reduce the interhousehold effects on weight gain (*37*), the data set is filtered so that no more than one man or woman is included from each household. This filtering is achieved by selecting the man or woman with the longest time at the current address. Participants living at the same address for less than 1 year were excluded from the analysis since the effect of the built environment on a weight gain or loss is estimated to be for at least 1 year (*38, 39*). A total of 37 individuals who had been pregnant within the last 12 months were also excluded from the analysis because the BMI may be affected before, during, and after pregnancy. Last, an age restriction of 18 to 65 (inclusive) is applied to the data set to exclude children and elderly from the study since the role of the built environment at different life stages maybe different (*14, 38, 40*).

The self-selection problem can be described as an artificial increase in association between the built environment and travel behavior because some individuals with strong transport preferences choose their neighborhood to be able to realize those preferences. For example, "residents who prefer walking may consciously choose to live in neighborhoods conducive to walking, and thus walk more" (*19*). A number of studies have shown the effect of self-selection in the association between the built environment and travel behavior (*26*) as well as the association between the built environment and obesity (*41*).

In order to reduce the effect of the self-selection problem, other questions in the HILDA survey were used to understand the reasons for moving. This question is used to exclude the individuals who may have an attitudinal preference pertaining to active transport (and are more likely candidates to have a self-selection bias). The question asks, "What were the main reasons for leaving the last address?" The responses are classified into a list of 32 reasons for moving, which are used to exclude the individuals who are likely to have a self-selection bias using the following criteria. The individuals who selected being close to amenities, services, public transport, work, or school were excluded regardless of their other reasons for moving. The individuals who chose lifestyle, health, and neighborhood quality were only excluded if they had selected another reason for moving that was related to transportation. Approximately 40% of the entries were removed to ensure that the self-selection bias was not included in the analysis.

The final data set used in this study contains a total of 477 individuals across 193 CCDs and is composed of 58% men and 18% classified as obese.

METHODOLOGY

The purpose of this study is to evaluate the association of urban environment characteristics and obesity in Sydney, Australia, and to examine the importance of spatial aggregation methods in the significance of this association. Within this context the proposed aggregation method for the definition of the neighborhood area is compared with a conventional aggregation method used when the residential location of individuals is preaggregated at an arbitrary geographical boundary.

Proposed Aggregation Methodology

The residential location of individuals is preaggregated to CCDs in the HILDA data used in this study. Although CCDs are relatively small zones (similar to U.S. census tracts), it is not possible to define the neighborhood area for each individual by using a road buffer from the address of each individual. Thus, the only link between the HILDA data and the built environment variables is the CCD. The CCDs were defined for the purposes of data collection during the census (i.e., CCDs are designed to contain approximately 200 to 250 households) (*42*).

The ideal aggregation methodology would define the neighborhood area on the basis of how each individual interacts with the physical environment surrounding his or her place of residence; this method would remove the associated MAUP errors (*12*). This problem is referred to as the uncertain geographic context problem, which requires that the travel trajectory of each individual represent the urban area of influence within which an individual would have interactions (*43, 44*).

A number of studies have shown that aggregating the built environment variables by using road buffers from an individual's address (as compared with a straight-line buffer or administrative boundaries) has a stronger association with transport behavior and health outcomes (*8, 45, 46*). This relation is because the defined neighborhood area is more likely to reflect the area in which an individual had direct interactions as compared with an administrative boundary or a straight-line buffer (*22*). However, the exact residential location of an individual is required for the road buffer aggregation method, and this location is unavailable; thus an alternative method is proposed.

The proposed aggregation method relies on the assumption that the neighborhood areas are better represented by small geographical areas based on residential land use boundaries (i.e., residential clusters). The aggregation method therefore initially disaggregates the CCD into clusters of residential areas within the CCD (Figure 2*b*). This process allows each neighborhood area to originate at the center of its residential cluster, and road buffers can then be used to define the boundary of the neighborhood area. Finally, the built environment variables for each residential cluster are aggregated back to the CCD level by using a weighted-average approach. The proposed method utilizes the ability to overlap layers in the GIS; in this case the CCD and the land use layer are overlapped. As shown in Figure 2*a*, one of the land use classifications is Residential. The following steps are used for calculating the built environment variables with the proposed aggregation method:

1. Overlap the CCD and the land use layer containing the residential areas (i.e., the red areas in Figure 2*a*);
2. If there are residential clusters that are composed of many smaller residential areas, combine the residential areas that are within 50 m of each other into one polygon (Figure 2*b*);
3. Calculate the internal centroid of the residential clusters (Figure 2*b*);
4. Create road buffers from the centroids of the residential clusters (Figure 2*c*);
5. Calculate the built environment variable by using the road buffers for each residential cluster; and
6. Aggregate the built environment variables from each residential cluster back to the CCD level; this step can be achieved by using a weighted average based on the size of the residential cluster.

$$U_c = \frac{\sum_{r=1}^{N_c} (U_r \times A_r)}{\sum_{r=1}^{N_c} A_r}$$

where

U_c = weighted average of built environment variable for CCD = c,
U_r = urban variable calculated with road buffer for residential cluster = r,
A_r = area of residential cluster = r (m^2), and
N_c = number of residential clusters within CCD = c.

The foregoing formula indicates that if a CCD does not contain more than one residential cluster, it is not known within which residential cluster the survey respondent resides. The probability of the survey respondent's residing in the larger residential cluster is higher, and thus the built environment variables are aggregated to the CCD level by applying a weighted-average approach based on the area (in square meters) of the residential clusters.

Three road buffer distances were developed for each centroid in order to evaluate the effects of the built environment variables on the basis of the road buffer distance. Figure 2*c* shows the difference between the road buffers created at 500 m (red), 1,000 m (blue), and 1,600 m (green) from the centroids. The distances represent 5 to 7, 10 to 12, and 15 to 18 min of walking at a normal pace of 80, 90, and 100 m per minute, respectively (*47*).

Statistical Analysis

To measure the association between the built environment and obesity, a binary logit model is developed. Within this context, four definitions of the neighborhood areas are used to calculate the built environment variables. The four definitions of the neighborhood area are the CCD boundary and the proposed aggregation method using three road buffer distances: 500, 1,000, and 1,600 m. Thus, for each individual address there are four sets of built environment variables, each set pertaining to a definition of the neighborhood area. Four logit models were developed to evaluate the four definitions of the neighborhood area as they relate to obesity.

The binary logit model is developed by first including the individual-level variables such as age and gender. A large variety of individual-level variables typically found in similar studies were tested and kept if they were highly significant and if they were not highly correlated with the urban variables (*5–9, 25, 41, 48*). In addition to the interaction between the variables, the underlying relationships of the individual-level variables with obesity were considered. In particular, the number of vegetable servings consumed per week and the number of exercise sessions per week were included in the final model to account for the input–output model of obesity (*2*).

The Spearman's correlation matrix was used to ensure that the built environment variables are not extremely multicollinear; multicollinearity is a typical issue encountered in built environment studies (*49*). The variables included in the final models had a correlation of 0.6 or less across the matrix. This test was repeated for all four aggregation schemes. When two or more urban variables had a correlation of 0.6 or higher, the urban variable with the higher significance was selected.

The variables included in the final models are shown in Table 1. In order to compare the four definitions of the neighborhood area, the same variables were specified across the four models.

RESULTS

Table 2 includes the descriptive statistics for the built environment variables calculated by using the four definitions of the neighborhood area. The first definition is the CCD boundary (based on the typical aggregation method of using the existing administrative boundary), and the other three definitions are the 500-m buffer, 1,000-m buffer, and 1,600-m buffer, which are based on the proposed aggregation methodology.

Table 2 indicates that the standard deviation decreases as the buffer distance of the neighborhood area increases for the following three variables: four-way intersection ratio, average block length, and density of major roadways. As the distance of the road buffer increases, adjacent areas are included in the neighborhood area and the differences between these design variables are reduced possibly because of the inclusion of the necessary services such as major roads (and hence reduction of the proportion of four-way intersections and

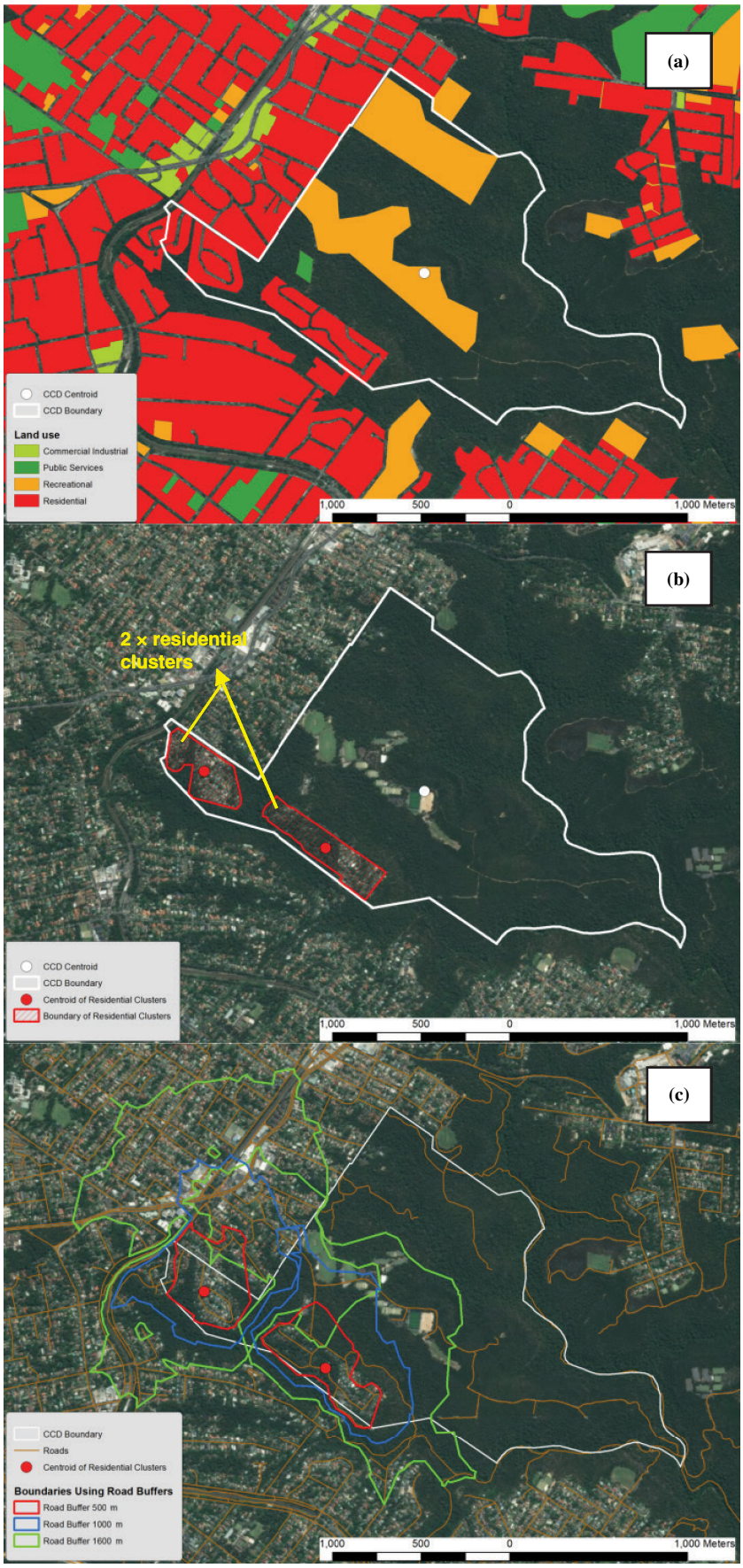

FIGURE 2 Aggregation clustering: (*a*) residential land use areas within CCD boundary (land use layer obtained from Office of Environment and Heritage, New South Wales, Australia); (*b*) two residential clusters within CCD boundary; and (*c*) proposed boundaries using three road buffer distances.

TABLE 1 Results of Four Logit Models Predicting Obesity

	Typical Aggregation Scheme	Proposed Aggregation Methodology		
Factor	CCD Boundary	500-m Buffer	1,000-m Buffer	1,600-m Buffer
Constant	−1.350 (0.371)	−2.254 (0.142)	−2.086 (0.181)	−1.686 (0.303)
Gender (female = 1)	0.296 (0.267)	0.357 (0.183)	0.296 (0.263)	0.312 (0.24)
Age	0.034 (0.001)	0.034 (0.002)	0.033 (0.002)	0.034 (0.002)
Vegetable servings per week	0.022 (0.159)	0.019 (0.205)	0.019 (0.215)	0.02 (0.196)
Exercise: none	0.614 (0.154)	0.628 (0.147)	0.658 (0.127)	0.628 (0.146)
Exercise: <1 per week	0.657 (0.06)	0.545 (0.122)	0.639 (0.066)	0.627 (0.072)
Exercise: 1–2 per week	0.436 (0.162)	0.387 (0.217)	0.450 (0.147)	0.439 (0.158)
IRSD	−0.002 (0.045)	−0.002 (0.048)	−0.002 (0.048)	−0.002 (0.038)
Supermarket (inverse distance)	0.111 (0.546)	0.112 (0.409)	0.125 (0.360)	0.121 (0.364)
Four-way intersection ratio	−1.087 (0.300)	−1.897 (0.088)	−1.337 (0.475)	−1.956 (0.399)
Average block length	0.0002 (0.865)	0.0020 (0.185)	0.0020 (0.470)	0.0020 (0.567)
Number of signalized intersections	−0.022 (0.088)	−0.177 (0.122)	−0.042 (0.188)	−0.017 (0.295)
Number of train stops	−0.283 (0.725)	−0.140 (0.878)	0.104 (0.679)	−0.042 (0.801)
Number of bus stops	−0.039 (0.536)	0.099 (0.114)	0.009 (0.720)	0.002 (0.860)
AIC	439.71	434.25	440.74	439.47
BIC	498.05	492.59	499.08	497.82
P-value of χ^2 test	.0022	.0003	.0031	.0020
Pseudo-R^2	.0728	.0851	.0705	.0733

NOTE: The values relevant to each variable are the estimated coefficient and the (*P*-value) in brackets. Number of included samples = 477. BIC = Bayesian information criterion.

TABLE 2 Descriptive Statistics

	Typical Aggregation Methodology		Proposed Aggregation Methodology[a]					
	CCD Boundary		500-m Buffer		1,000-m Buffer		1,600-m Buffer	
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Four-way intersection ratio	0.12	0.14	0.15	0.13	0.13	0.08	0.13	0.06
Average block length	262.86	116.37	211.64	75.91	212.65	56.86	219.50	54.89
Number of signalized intersections	12.41	17.37	1.11	2.47	4.85	8.10	12.45	17.40
Density of major roadways	2,638.0	2,755.9	2,546.2	2,101.2	2,309.6	1,351.9	2,200.2	1,074.9
Number of train stops	0.03	0.17	0.03	0.16	0.27	0.61	0.71	1.02
Number of train services	6.1	38.3	7.7	48.1	118.8	439.7	329.4	888.7
Number of bus stops	2.31	2.28	3.1	2.0	11.5	5.6	29.1	13.8
Number of bus services	259.7	487.8	445.1	903.5	1,846.1	2,834.2	4,707.1	6,046.6
Supermarket (inverse distance)	0.85	0.81	0.88	0.99	0.88	0.99	0.88	0.99
Population density	4,146.3	3,440.2						
Employment density	1,468.3	7,822.3						
Dwelling density	1,234.5	1,213.3						
IRSD	1,096.9	113.3						
Age	45.8	13.6						
Vegetable servings per week	16.1	8.5						

NOTE: Number of included samples = 477. SD = standard deviation.
[a]Empty cell = not based on aggregation method (individual-level data).

increase in the average block lengths). The CCD boundary aggregation method has a relatively high standard deviation across the three variables. This finding is particularly noticeable when compared with the 1,000-m buffer, which is closest to the CCD boundary in terms of spatial scale. The reason for the high standard deviation for the CCD boundary aggregation method may be that for some neighborhoods it includes irrelevant areas, such as nature strips as shown in Figure 2*a* (i.e., the undeveloped tree-lined areas visible in the aerial imagery).

The transit-related variables calculated with the CCD boundary aggregation method have a relatively low mean and standard deviation compared with the proposed aggregation method. As the road buffer distance increases for the proposed aggregation method, the mean and standard deviation of the transit-related variables increase. It is possible that the CCD boundary aggregation method excludes adjacent sites containing infrastructure that is accessible to residents living close to the boundary of the CCD.

The "distance to major supermarkets" variable is almost the same when the distance from the centroid of the CCD is measured or the centroids of the residential clusters in the proposed aggregation method are used. In addition, during the modeling process this variable was found to be insignificant regardless of the aggregation method used. This result is an indication that the proposed aggregation method may not be suitable to define variables for which the distance from the origin (i.e., residential address) to a destination (i.e., major supermarket) is measured. This finding may be due to the averaging affect, that is, when the values of the variables calculated for each residential cluster are aggregated back to the CCD level. The average difference between the minimum and maximum distance to a major supermarket across residential clusters within a CCD was 238 m with a standard deviation of 461 m. This finding is significant considering that across the residential clusters the average distance to a major supermarket was 1,788 m, with a minimum of 24 m and maximum of 15.5 km.

Table 1 shows the results across the four models as discussed previously. From the individual-level variables the vegetable servings per week has a positive association with obesity, which may seem counterintuitive at the first glance (since vegetable intake is considered healthy). The amount of food intake is not controlled for, and thus the number of vegetable servings per week may be a better representation of the amount of food intake as opposed to the proportion of the food intake that is vegetables.

From the built environment variables, the number of bus stops has a positive association with obesity. The number of bus stops was only significant for the 500-m buffer model (*P*-value = .114). The number of train stops or number of train services was not significantly associated with obesity in any of the models. In general, people are more willing to travel a longer distance to train stations—in Sydney an average walking distance of 805 m (*50*)—and therefore the 1,000-m buffer maybe more appropriate for variables related to trains. Excluding train stops from the model did not result in any significant change and this variable was included in the model for discussion purposes.

To compare different definitions of the neighborhood, the levels of significance of the built environment variables within the models are typically compared (*8, 14*). From the four definitions of the neighborhood areas, the model with the lowest *P*-value for each built environment variable is highlighted in Table 1. The proposed aggregation method at the 500-m road buffer contains most of the urban variables with the lowest *P*-value, and thus it is superior to the other aggregation methods.

The overall comparative model improvement is also assessed by using the Akaike information criterion and the pseudo-R^2. The overall model performance indicators also highlight that the proposed aggregation method at the 500-m road buffer marginally outperforms the other aggregation methods.

DISCUSSION OF RESULTS

This study proposed, applied, and tested an aggregation methodology for anonymous surveys in which the residential location of individuals is preaggregated to an existing administrative boundary. The study assessed the association between built environment and obesity in Sydney, Australia. A significant data collection and processing effort was conducted to obtain GIS data necessary for the evaluation of the built environment variables. The variables were selected on the basis of the six Ds principle, which considers how the urban environment may affect travel behavior. The self-selection bias resulting from address selection was removed from the analysis by excluding individuals who chose their place of residence because of proximity to destinations.

Results from the study indicate that there is an association between built environment and obesity, particularly when the smaller 500-m road buffer is considered. In general, the proposed aggregation methodology increases the accuracy of the built environment variables by improving the definition of the neighborhood area. This improvement was achieved by defining residential clusters within each CCD, obtaining the built environment variables for each residential cluster, and aggregating back to the CCD level. The proposed aggregation methodology was applied by using three road buffer distances: 500 m, 1,000 m, and 1,600 m. The results illustrate that the proposed aggregation method at the 500-m road buffer is superior to the typical aggregation method (i.e., using predefined administrative boundaries).

The advantages of the proposed aggregation method over use of predefined administrative boundaries are that the neighborhood area is most likely to be the area in which an individual has regular interaction since it is centric to the residential land use and road buffers are used. The CCD boundary may include areas that individuals may not typically interact with or should be included in the calculation of some of the built environment variables (particularly for larger CCDs). In addition, the residential clusters that are on the edge of the CCD are likely to interact with the adjacent CCD. The boundary of the proposed aggregation method is not restricted to the boundary of the CCD. Finally, the proposed aggregation method allows the use of road buffers, and thus physical barriers such as rivers or railways that may block access from one side to the other are considered.

The main disadvantage of the proposed aggregation method is that the road buffer distances cannot be too large since a large road buffer distance may result in an overlap of the neighborhood area with relatively close residential clusters. In such a case, the overlapping area will be double-counted in the built environment variables. In addition, the final step of the proposed method requires aggregating the built environment variables back to the original CCD. If significant differences exist between residential clusters within a CCD, the averaging effect will dampen the final value of the built environment variable (this issue also exists in the traditional CCD aggregation method). However, given the spatial autocorrelation of urban characteristics (i.e., places close to each other have a tendency to be similar) and when it is applied across a large sample, the regression

results from the proposed aggregation method are expected to be marginally affected by this issue (however, this issue has not been specifically measured in this study). Alternatively, if a large sample is available, it is possible to remove the individuals living in CCDs with a large variation in built environment variables between the urban clusters present in the CCD.

CONCLUSION

Anonymous surveys tend to aggregate the residential location of survey respondents to predefined zones. These zones are typically not designed to capture the urban characteristics of a neighborhood. The issue associated with defining the spatial aggregation boundary is referred to as the MAUP and has been shown to have dire consequences, particularly in regression analysis. Given that the majority of built environment and health studies are based on anonymous surveys and thus are forced to use the predefined zones, it is not surprising that many contradictory findings are reported. One of the main aims of this study was to introduce and test a proposed aggregation method that focuses on the neighborhood areas within the predefined zones.

In this study a variety of data sets describing the urban characteristics of Sydney, Australia, are collated, processed, and analyzed. The results from this analysis suggest that there is an association between the built environment and obesity among individuals living in Sydney. The strength of the association is dependent on the aggregation method used to calculate the built environment variables. The proposed aggregation methodology is illustrated to be better at capturing the association between certain built environment variables and obesity on the basis of the higher significance of the variables identified when the 500-m road buffer is used. As such, it can be applied to surveys in which privacy policy prevents individuals' residential locations from being released.

Future research will address the refinement of the road buffer to consider factors that may influence the directionality of travel. For example, the common routes to major public transport stations or supermarkets can be considered explicitly. The 360-degree street buffer from the user's address may not be a realistic representation of the user's perception of his urban environment. By considering behaviorally defined aggregation methods, the MAUP can be further reduced (*14*). Further, this study aims to expand research on jointly modeling other health-related variables while accounting for spatial correlation between the defined zones.

ACKNOWLEDGMENTS

The authors thank the Office of Environment and Heritage, Roads and Maritime Services, Bureau of Transport Statistics, Transport for New South Wales, New South Wales Land and Property Information, and the University of Melbourne for their support in providing the data and advice for this study.

REFERENCES

1. *Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks.* World Health Organization, Geneva, Switzerland, 2009.
2. Organisation for Economic Co-operation and Development. *Health at a Glance 2013: OECD Indicators.* OECD Publishing, Paris, 2013.
3. Finkelstein, E. A., J. G. Trogdon, J. W. Cohen, and W. Dietz. Annual Medical Spending Attributable to Obesity: Payer- and Service-Specific Estimates. *Health Affairs,* Vol. 28, No. 5, 2009, pp. w822–w831.
4. *Physical Activity and Health: A Report to the Surgeon General.* U.S. Department of Health and Human Services, 1996.
5. Grasser, G., D. Van Dyck, S. Titze, and W. Stronegger. Objectively Measured Walkability and Active Transport and Weight-Related Outcomes in Adults: A Systematic Review. *International Journal of Public Health,* Vol. 58, No. 4, 2013, pp. 615–625.
6. Samimi, A., A. Mohammadian, and S. Madanizadeh. Effects of Transportation and Built Environment on General Health and Obesity. *Transportation Research Part D: Transport and Environment,* Vol. 14, No. 1, 2009, pp. 67–71.
7. Frank, L. D., J. Kerr, J. F. Sallis, R. Miles, and J. Chapman. A Hierarchy of Sociodemographic and Environmental Correlates of Walking and Obesity. *Preventive Medicine,* Vol. 47, No. 2, 2008, pp. 172–178.
8. Coffee, N. T., N. Howard, C. Paquet, G. Hugo, and M. Daniel. Is Walkability Associated with a Lower Cardiometabolic Risk? *Health & Place,* Vol. 21, 2013, pp. 163–169.
9. Müller-Riemenschneider, F., G. Pereira, K. Villanueva, H. Christian, M. Knuiman, B. Giles-Corti, and F. Bull. Neighborhood Walkability and Cardiometabolic Risk Factors in Australian Adults: An Observational Study. *BMC Public Health,* Vol. 13, No. 1, 2013, pp. 1–9.
10. Leal, C., and B. Chaix. The Influence of Geographic Life Environments on Cardiometabolic Risk Factors: A Systematic Review, a Methodological Assessment, and a Research Agenda. *Obesity Reviews,* Vol. 12, No. 3, 2011, pp. 217–230.
11. Coffee T. N. *Constructing an Objective Index of Walkability.* Master's thesis. University of Adelaide, South Australia, 2005.
12. Zhang, M., and N. Kukadia. Metrics of Urban Form and the Modifiable Areal Unit Problem. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1902,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 71–79.
13. Riva, M., P. Apparicio, L. Gauvin, and J.-M. Brodeur. Establishing the Soundness of Administrative Spatial Units for Operationalising the Active Living Potential of Residential Environments: An Exemplar for Designing Optimal Zones. *International Journal of Health Geographics,* Vol. 7, 2008, p. 43.
14. Mitra, R., and R. N. Buliung. Built Environment Correlates of Active School Transportation: Neighborhood and the Modifiable Areal Unit Problem. *Journal of Transport Geography,* Vol. 20, No. 1, 2012, pp. 51–61.
15. Openshaw, S., and S. Alvanides. Applying Geocomputation to the Analysis of Spatial Distributions. In *Geographical Information Systems: Principles and Technical Issues,* Vol. 1, John Wiley and Sons, New York, 1999, pp. 267–282.
16. Gehlke, C., and K. Biehl. Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material. *Journal of the American Statistical Association,* Vol. 29, No. 185A, 1934, pp. 169–170.
17. Páez, A., and D. M. Scott. Spatial Statistics for Urban Analysis: A Review of Techniques with Examples. *GeoJournal,* Vol. 61, No. 1, 2004, pp. 53–67.
18. Amrhein, C. G. Searching for the Elusive Aggregation Effect: Evidence from Statistical Simulations. *Environment and Planning A,* Vol. 27, No. 1, 1995, pp. 105–119.
19. Mokhtarian, P. L., and X. Cao. Examining the Impacts of Residential Self-Selection on Travel Behavior: A Focus on Methodologies. *Transportation Research Part B: Methodological,* Vol. 42, No. 3, 2008, pp. 204–228.
20. Heywood, D. I., S. Cornelius, and S. Carver. *An Introduction to Geographical Information Systems.* Prentice Hall, New York, 2002.
21. Root, E. D. Moving Neighborhoods and Health Research Forward: Using Geographic Methods to Examine the Role of Spatial Scale in Neighborhood Effects on Health. *Annals of the Association of American Geographers,* Vol. 102, No. 5, 2012, pp. 986–995.
22. Duncan, M. J., E. Winkler, T. Sugiyama, E. Cerin, L. duToit, E. Leslie, and N. Owen. Relationships of Land Use Mix with Walking for Transport: Do Land Uses and Geographical Scale Matter? *Journal of Urban Health: Bulletin of the New York Academy of Medicine,* Vol. 87, No. 5, 2010, pp. 782–795.
23. Oliver, L. *Shifting Boundaries, Shifting Results: The Modifiable Areal Unit Problem.* Department of Geography, University of British Columbia, Canada, 2001. http://www.geog.ubc.ca/courses/geog516/talks_2001/scale_maup.html.
24. *Frequently Asked Questions: How Does the ABS Define Urban and Rural?* Australian Bureau of Statistics, Sydney, New South Wales, Australia.

http://www.abs.gov.au/websitedbs/D3310114.nsf/home/frequently+asked+questions#Anchor72014.
25. Frank, L. D., J. F. Sallis, T. L. Conway, J. E. Chapman, B. E. Saelens, and W. Bachman. Many Pathways from Land Use to Health: Associations Between Neighborhood Walkability and Active Transportation, Body Mass Index, and Air Quality. *Journal of the American Planning Association,* Vol. 72, No. 1, 2006, pp. 75–87.
26. Ewing, R., and R. Cervero. Travel and the Built Environment. *Journal of the American Planning Association,* Vol. 76, No. 3, 2010, pp. 265–294.
27. Boarnet, M. G., and S. Sarmiento. Can Land-Use Policy Really Affect Travel Behaviour? A Study of the Link Between Non-Work Travel and Land-Use Characteristics. *Urban Studies,* Vol. 35, No. 7, 1998, pp. 1155–1169.
28. *Topographic Data Dictionary v8.5.* Land and Property Information, Sydney, New South Wales, Australia, 2013.
29. Zappone, C. Supermarket Duopoly Blamed for Soaring Food Prices. *The Sydney Morning Herald,* 2009.
30. *Walk Score Methodology.* Walkscore.com. http://www.walkscore.com/methodology.shtml. Accessed Feb. 2014.
31. Shields, M. A., S. W. Price, and M. Wooden. Life Satisfaction and the Economic and Social Characteristics of Neighbourhoods. *Journal of Population Economics,* Vol. 22, No. 2, 2009, pp. 421–443.
32. Wooden, M., S. Freidin, and N. Watson. The Household, Income and Labour Dynamics in Australia (HILDA) Survey: Wave 1. *Australian Economic Review,* Vol. 35, No. 3, 2002, pp. 339–348.
33. Summerfield, M., S. Freidin, M. Hahn, N. Li, N. Macalalad, L. Mundy, N. Watson, R. Wilkins, and M. Wooden. *HILDA User Manual–Release 12.* Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Victoria, Australia, 2013.
34. Badland, H., G. Turrell, and B. Giles-Corti. Who Does Well Where? Exploring How Self-Rated Health Differs Across Diverse People and Neighborhoods. *Health & Place,* Vol. 22, 2013, pp. 82–89.
35. TableBuilder. In *Census of Population and Housing,* Australian Bureau of Statistics, Sydney, New South Wales, Australia, 2011.
36. Pink, B. *Socio-Economic Indexes for Areas (SEIFA).* Technical Paper 2033.0.55.001. Commonwealth of Australia, Canberra, Australian Capital Territory 2011.
37. Shields, M., and M. Wooden. *Investigating the Role of Neighbourhood Characteristics in Determining Life Satisfaction.* University of Melbourne, Victoria, Australia, 2003.
38. Berke, E. M., T. D. Koepsell, A. V. Moudon, R. E. Hoskins, and E. B. Larson. Association of the Built Environment with Physical Activity and Obesity in Older Persons. *American Journal of Public Health,* Vol. 97, No. 3, 2007, pp. 486–492.
39. *Health Economic Assessment Tools (HEAT) for Walking and for Cycling: Economic Assessment of Transport Infrastructure and Policies.* World Health Organization, Geneva, Switzerland, 2011.
40. Owen, N., J. Salmon, M. J. Koohsari, G. Turrell, and B. Giles-Corti. Sedentary Behaviour and Health: Mapping Environmental and Social Contexts to Underpin Chronic Disease Prevention. *British Journal of Sports Medicine,* Vol. 48, No. 3, 2014, pp. 174–177.
41. Feng, J., T. A. Glass, F. C. Curriero, W. F. Stewart, and B. S. Schwartz. The Built Environment and Obesity: A Systematic Review of the Epidemiologic Evidence. *Health & Place,* Vol. 16, No. 2, 2010, pp. 175–190.
42. Pink, B. Main Structure and Greater Capital City Statistical Areas. In *Australian Statistical Geography Standard,* Vol. 1, Australian Bureau of Statistics, Sydney, New South Wales, Australia, 2011.
43. Kwan, M.-P. How GIS Can Help Address the Uncertain Geographic Context Problem in Social Science Research. *Annals of GIS,* Vol. 18, No. 4, 2012, pp. 245–255.
44. The Uncertain Geographic Context Problem. *Annals of the Association of American Geographers,* Vol. 102, No. 5, 2012, pp. 958–968.
45. Learnihan, V., K. P. Van Niel, B. Giles-Corti, and M. Knuiman. Effect of Scale on the Links Between Walking and Urban Design. *Geographical Research,* Vol. 49, No. 2, 2011, pp. 183–191.
46. Thornton, L., J. Pearce, and A. Kavanagh. Using Geographic Information Systems (GIS) to Assess the Role of the Built Environment in Influencing Obesity: A Glossary. *International Journal of Behavioral Nutrition and Physical Activity,* Vol. 8, No. 1, 2011, p. 71.
47. Morris, J. N., and A. E. Hardman. Walking to Health. *Sports Medicine,* Vol. 23, No. 5, 1997, pp. 306–332.
48. Brown, B. B., I. Yamada, K. R. Smith, C. D. Zick, L. Kowaleski-Jones, and J. X. Fan. Mixed Land Use and Walkability: Variations in Land Use Measures and Relationships with BMI, Overweight, and Obesity. *Health & Place,* Vol. 15, No. 4, 2009, pp. 1130–1141.
49. Cervero, R., and K. Kockelman. Travel Demand and the 3Ds: Density, Diversity, and Design. *Transportation Research Part D: Transport and Environment,* Vol. 2, No. 3, 1997, pp. 199–219.
50. Daniels, R., and C. Mulley. Explaining Walking Distance to Public Transport: The Dominance of Public Transport Supply. *World,* Vol. 28, 2011, p. 30.

The Standing Committee on Environmental Justice in Transportation peer-reviewed this paper.