Published in final edited form as:

Am J Prev Med. 2009 August ; 37(2): 157–166. doi:10.1016/j.amepre.2009.04.021.

Identifying Clusters of Active Transportation Using Spatial Scan Statistics

Lan Huang, PhD, David G. Stinchcomb, MA, MS, Linda W. Pickle, PhD, Jennifer Dill, PhD, and David Berrigan, PhD, MPH

From the Statistical Research and Applications Branch (Huang, Stinchcomb, Pickle), and the Applied Research Program (Berrigan), Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Maryland; and the College of Urban and Public Affairs (Dill), Portland State University, Portland, Oregon

Abstract

Background—There is an intense interest in the possibility that neighborhood characteristics influence active transportation such as walking or biking. The purpose of this paper is to illustrate how a spatial cluster identification method can evaluate the geographic variation of active transportation and identify neighborhoods with unusually high/low levels of active transportation.

Methods—Self-reported walking/biking prevalence, demographic characteristics, street connectivity variables, and neighborhood socioeconomic data were collected from respondents to the 2001 California Health Interview Survey (CHIS; N=10,688) in Los Angeles County (LAC) and San Diego County (SDC). Spatial scan statistics were used to identify clusters of high or low prevalence (with and without age-adjustment) and the quantity of time spent walking and biking. The data, a subset from the 2001 CHIS, were analyzed in 2007–2008.

Results—Geographic clusters of significantly high or low prevalence of walking and biking were detected in LAC and SDC. Structural variables such as street connectivity and shorter block lengths are consistently associated with higher levels of active transportation, but associations between active transportation and socioeconomic variables at the individual and neighborhood levels are mixed. Only one cluster with less time spent walking and biking among walkers/bikers was detected in LAC, and this was of borderline significance. Age-adjustment affects the clustering pattern of walking/biking prevalence in LAC, but not in SDC.

Conclusions—The use of spatial scan statistics to identify significant clustering of health behaviors such as active transportation adds to the more traditional regression analysis that examines associations between behavior and environmental factors by identifying specific geographic areas with unusual levels of the behavior independent of predefined administrative units.

Introduction

There is a recent and intense interest in the potential impact of neighborhood environments on energy balance–related health behaviors.^{1–4} This interest is due in part to the current epidemic

© 2009 American Journal of Preventive Medicine

Address correspondence and reprint requests to: Lan Huang, PhD, Statistical Research and Applications Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, Room 5043, 6116 Executive Boulevard, Bethesda MD 20892-8317. huangla@mail.nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

of obesity and in part to continued interest in building healthy communities. One focus of this research has been on the relationship between the built environment and physical activity.^{1,5,6} A number of studies have demonstrated associations between diverse measures of urban form and active transportation or leisure activities such as walking and biking. However, the amount of variation in active transportation in these studies has been relatively small. In addition, the results are inconsistent, and the conclusions are based almost entirely on cross-sectional studies.^{6–8} Therefore, there is a significant need for new approaches to identifying the contextual and environmental correlates of active transportation in order to determine if this is a viable avenue for public health interventions and policy development.

Recent work on the environmental correlates of active transportation has largely involved regression analyses of the associations between active transportation and diverse demographic and contextual variables. Such approaches have provided insight into the relative contributions of density, diversity, design, and demography to levels of active transportation.⁹ Nonetheless, it seems possible that such approaches are limited by the ability of simple models to describe the effects of multiple contextual and environmental influences, particularly if these influences vary geographically. Identification of spatial clusters of active transportation may provide a tool to determine whether a set of environmental and demographic variables have been identified that can account for variation in active transportation as well as serve a policy function by characterizing regions lacking environmental characteristics that promote active transportation.

Visual inspection of maps has long been a tool in epidemiology.^{10,11} More recently, over 100 statistical tests¹² have been proposed to identify significant overall clustering or specific clusters of disease cases or other variables of interest in spatial data sets. The aim of the current study is to identify localized clusters of active transport. SaTScan, a likelihood-based spatial scan statistic¹³ is good at detecting these clusters,^{14–18} and it has been used in past studies to search for clusters of breast cancer deaths,¹⁹ activity in birds,²⁰ symptoms reported by patients at hospital emergency rooms,²¹ pharmacy sales, and physician visits.²² Notably, SaTScan has also been extended to identify elliptical as well as circular clusters of disease cases,²³ grade and stage of cancer diagnosis,²⁴ and survival rates.²⁵

In the current article, the application of SaTScan is extended to identify clusters of behavioral data in a large survey where geographic identifiers are available. Specifically, spatial scan statistics are applied to search for clusters of neighborhoods with a high or low prevalence of walking or biking and for clusters associated with a higher or lower duration of walking and biking among walkers/bikers in a large sample of survey respondents in two diverse counties in California. To our knowledge, the current study is the first reported use of spatial scan techniques to search for clusters of active transportation.

Methods

A subset of data were analyzed in 2007–2008 from the 2001 California Health Interview Survey (CHIS). This survey is a large (N=55,428 households) random digit–dial telephone survey in California, administered in seven languages (English, Spanish, Mandarin, Cantonese, Vietnamese, Korean, and Khmer). The CHIS 2001 response rate, based on the American Association for Public Opinion Research equation RR4,²⁶ was 43.3% with a cooperation rate of 63.7% (weighted to account for the sample design; 77.1% unweighted). More than 70% of survey respondents supplied the name of the nearest intersection to their residence (in Los Angeles County [LAC], 8728/12,196=71.5%; in San Diego County [SDC], 1952/2672=73%). These addresses were geocoded to represent the location of each respondent for purposes of this analysis.

After exclusion of respondents with missing or invalid data, 8506 respondents from LAC and 1883 respondents from SDC were used in the analysis (Figure 1A and 1D). Because LAC and SDC are not adjacent, the analysis was implemented separately in the two locations. The data from these two counties are the only data from CHIS 2001 with geocodes available. Data were analyzed from both counties in order to demonstrate the performance of the proposed statistical methods on larger (LAC) and smaller (SDC) samples from cities with distinct geographic, cultural, and economic characteristics. A combination of CHIS variables, census data, and street connectivity data generated from a GIS were used to create a working data set for the spatial cluster analysis and for subsequent characterization of active transportation clusters.

California Health Interview Survey Variables

The CHIS 2001 data included in this study were measures of nonleisure-time walking and biking (NLTWB), and multiple relevant demographic variables. NLTWB was measured by asking three short questions: (1) *Over the past 30 days, have you walked or bicycled to or from work, school, or to do errands?*; (2) *How many times per day, per week, or per month did you do this?*; and (3) *And on average, about how many minutes did you walk or ride your bike each time?*. NLTWB was analyzed either as a measure of prevalence such as yes/no (any walking/biking or none) from the answers to the first question, or as a measure of duration, such as minutes per week spent walking/biking, derived from the answers to the second and third questions.

Demographic and SES variables including age, gender, race, education, and income were also extracted from the CHIS survey resource for each respondent, as were self-reported health status, immigration status, and employment status. For self-related health status, an activity-related variable was chosen based on responses to the query: *How much does your health limit you when climbing several flights of stairs?*. Responses were on a three-part scale: *limited a lot, limited a little, not limited at all*. CHIS includes other variables related to diet, tobacco and alcohol use, cancer-screening practices, and healthcare coverage; in the current study, the focus was on variables commonly used in past studies of active transportation.

Contextual Variables

The street connectivity and two density-related variables were analyzed using circular buffers (areas around a point) of radius 0.5 km surrounding each respondent's location (nearest intersection to home). These buffers were defined using the Topologically Integrated Geographic Encoding and Referencing System (TIGER) map files from the 2000 U.S. Census Bureau and implemented with GIS software (ArcView, ESRI, Inc.). Data concerning population and employment density and characteristics of the street network for each buffer were then calculated using information from the U.S. Census at the census tract or census-block (administrative units that are nested within census tracts) level.

Population density within a buffer was generated by downloading U.S. Census data at the census-block level. Each half-kilometer buffer usually overlapped more than one census block. Population density was assumed to be uniform within each census block, and a portion of the population within a census block was assigned to the buffer based on the area of the census block within the buffer. For example, if a buffer covers half of a census block, half of the census block's population is assigned to that buffer, in addition to the population in census blocks that was completely within the buffer. The total population in the buffer was then divided by the area (0.785 square kilometers). Employment density data were generated using data from the metropolitan planning organization for each area—the Southern California Association of Governments (SCAG) for Los Angeles and the San Diego Association of Governments (SANDAG) for San Diego. Each agency provided total employment data by census tract for the year 2000. The method to calculate employment density was identical to that for calculating

population density, except that tracts were used instead of blocks. Therefore, the variances associated with population and employment densities are likely to differ in this study.

Four measures of street connectivity were also extracted: intersection density, street network density, census-block density, and average street block (segments of a street between intersections) length. The street network density is calculated by summing the lengths of all the streets within the buffer (the total street network distance within the buffer, ignoring the number of lanes on a road) and dividing by the area of the buffer (0.785 square kilometers). The portion of a street that continued outside the buffer was not included. Census-block density is the total number of census blocks within a buffer divided by the area of the buffer (0.785 square kilometers). Census-block boundaries generally coincide with streets. If a portion of a census block was outside a buffer, only the area of the census block within the buffer was included. The average street block length is the average length of the street blocks that are completely or partially within the buffer. For street blocks that continue outside the buffer, the entire length of the street block is included in the calculation. Truncating the street block at the buffer boundary would have reduced the length of the block artificially. Six buffers were randomly chosen, including two that overlapped, and connectivity estimates were validated by plotting the buffers with the links, nodes, and blocks and then manually extracting the data.

Spatial Analysis

Spatial clustering of the prevalence of walking or biking in LAC and SDC was separately analyzed using circular spatial scan statistics based on a Bernoulli model¹³ with walkers/bikers as cases and others as controls, with and without age-adjustment.¹⁶ Because SaTScan software allows adjusting for categorical variables with up to only 12 categories with a Bernoulli model, age was categorized into six levels (18–29, 30–39, 40–49, 50–59, 60–69, and ≥ 70 years), which were then used for age-adjustment in the Bernoulli model-based analysis scan. The length of time spent in active transportation by walkers/bikers was also explored using scan statistics based on a normal model (www.satscan.org) for continuous data (no covariate adjustment allowed with this model). The individual characteristics, buffer characteristics, and the street connectivity status within significant clusters were then examined by comparing the factors within the clusters to those outside the clusters using a *t*-test for continuous variables and a chi-square test for categorical variables.

A circular scan statistic searches one or more circles with a series of hierarchically overlapping areas of increasing size around each centroid (in this case the nearest intersection to the respondent's home address). Detection of a significant cluster of high/low prevalence indicates that the people inside the cluster have a significantly higher/lower likelihood of walking/biking for transportation compared with the respondents in other areas. The maximum search window in the scan approach is user-selected. Smaller maximum search windows are useful to identify local clusters in more homogeneous neighborhoods. To select the appropriate maximum search window, the initial maximum search window was set to include 50% of the total number of respondents, and then successively smaller search windows were examined. Large clusters have a better chance of being declared significant ($p < 0.05$) compared with small clusters because of their larger sample size. However, a large cluster may cover a large geographic area that includes local neighborhoods with diverse characteristics of individuals, buffers, and street connectivity. Ultimately, a maximum cluster size was selected that contained 3% of the population, because a smaller cluster size would contain a more homogeneous population, and almost all detected clusters with smaller sizes are within clusters found with a larger maximum size. Selection of maximum search windows generally requires consideration of study aims and some exploratory analysis such as the above.

Clustering of higher/lower prevalence in active transportation does not imply longer/shorter walking/biking time. Therefore, clustering of the length of time spent walking/biking among

the walkers/bikers was also explored using the spatial scan method based on the normal model (www.satscan.org). The total sample size is smaller in this subanalysis because not all respondents reported any walking/biking for transportation (3573 in LAC, 680 in SDC). A 3% maximum search window size is too small for this case, especially for SDC analysis. Therefore, a minimum window size of two individuals and a maximum size of 10% of total walkers/bikers were used for the analysis of active transportation duration.

Results

Respondents from LAC and SDC have similar characteristics compared to the entire CHIS sample in California, except that the LAC/SDC sample is more racially/ethnically diverse (Table 1). Compared to a U.S nationwide survey, the combined study area and California as a whole are more racially/ethnically diverse, younger, have a lower average income, and have more immigrants, more college graduates, and more residents who did not graduate from high school. LAC has fewer non-Hispanic whites, more immigrants, and lower income and education levels than SDC. These comparisons are not relevant to the implementation of the spatial scan approach, but they may influence interpretation of the findings of such analyses.

Overall, the percentage of walking/biking and the average walking/biking time over all respondents in LAC is higher than it is in SDC (42.0% vs 36.1% and 84 vs 80 minutes per week, respectively). Without age-adjustment, there are more significant clusters detected in LAC than in SDC, perhaps because of greater spatial variation of active transportation or larger sample size in LAC (Figure 1B, 1C, and 1E). Five clusters with more walkers/bikers and two clusters with fewer walkers/bikers are detected in LAC using 3% of the number of total respondents as the maximum search window (Figure 1B). Cluster 5 is significant when looking for clusters of only more walkers/bikers; it is not quite significant ($p=0.066$) when testing for both more and fewer walkers/bikers. All other clusters are very significant ($p<0.01$). The relative risks of being a walker/biker were 1.7, 1.5, 1.6, 1.5, 1.5, 0.4, and 0.6 for Clusters 1 to 7, respectively. Three significant clusters of more walkers/bikers (p -values <0.05) and a marginally significant one with fewer walkers/bikers (p -value=0.06) were detected in SDC (Figure 1E) when searching for clusters of either more or fewer walkers/bikers. The relative risks of being a walker/biker in SDC were 2.2, 2.8, 2.0, and 0.09 for Clusters 1 to 4, respectively. Among walkers/bikers, average walking/biking times per week were 188 minutes ($n=3573$, $SD=319$) in LAC and 173 minutes ($n=680$, $SD=232$) in SDC. One area with lower walking/biking duration was detected in LAC (Figure 1C), but no significant clusters were detected in SDC, suggesting that the duration of activities among those walkers/bikers is spatially homogeneous or that there is a lack of power because of modest sample sizes, particularly in SDC.

To illustrate the capacity of the spatial scan approach to explore contributions of specific variables to clustering, the clustering pattern after adjusting for age was also evaluated. Because younger people tend to have more active transportation than older people, spatial heterogeneity in age distributions alone could result in spatial clustering of active transportation. In LAC, age-adjustment accounts for some but not all of the significant clusters of active transportation (Figure 1F vs 1B). The clusters with high prevalence in SDC were similar with and without age-adjustment (maps after age-adjustment not shown). However, the marginal cluster of low prevalence (No. 4 in Figure 1E) is no longer found after age-adjustment.

Characteristics of Clusters

The patterns of the other covariates in the analysis without age-adjustment (Table 2 and Table 3) were similar to those with age-adjustment (see Appendixes A and B, available online at www.ajpm-online.net). This discussion focuses on cluster properties for clusters identified without age-adjustment, allowing for description of age characteristics inside/outside clusters.

Neighborhood characteristics were consistently different between high- and low-walking clusters in both counties, whereas comparisons of individual characteristics varied geographically (Table 2 and Table 3). Population and employment densities were higher in high-walking clusters and lower in low-walking clusters. Similarly, higher street, block, and intersection densities; shorter block lengths; and the presence of a bus route were associated with more walking and biking for transportation, compared to the rest of each county. Most of these street connectivity differences were highly significant.

People were slightly younger in the high-walking areas and older in the low-walking areas, but gender and BMI did not seem to be related to the pattern of active transportation. The associations between walking/biking and other individual characteristics generally varied among the clusters. In LAC, high-walking Clusters 1, 2, 4, and 5 had lower income and education levels, whereas Cluster 3 residents had higher education and income levels. Clusters 1, 2, and 5 had predominantly Hispanic residents, whereas Clusters 4 and 5 had more black residents; Cluster 3 is predominantly non-Hispanic white compared to areas outside the clusters (Table 2). In SDC, high-walking Clusters 1 and 3 had lower income and education levels; the small Cluster 2 had more highly educated residents but not a significantly different income distribution compared with the rest of the county. Cluster 1 had more Hispanic and black residents; Cluster 3 had more Hispanic residents; and Cluster 2 had more non-Hispanic whites than areas outside the clusters. In contrast, the low-walking clusters in both counties have more non-Hispanic white residents and higher income levels.

Self-reported health status showed mixed associations with walking/biking prevalence. For example, LAC high-walking Clusters 1, 2, 3, and 4 had fewer residents who responded that they were limited a lot in climbing stairs compared those to outside the clusters, but more residents had such limitations in Cluster 5. In Cluster 4, respondents reported the highest health status among all cluster areas and outside cluster areas (with 76.5% without limitation of climbing stairs). Street connectivity is very high in Cluster 4, so despite significantly fewer bus stops, freeways, and bus routes, it still appeared to be a cluster of more walkers/bikers, and their average duration walked/biked is the longest (264 minutes per week).

No significant clusters of longer or shorter walking/biking times were found in either county, independent of the prevalence of active transportation, but one area in LAC had a shorter walking/biking duration that was of borderline significance (Figure 1C). This area had residents with more education (86% vs 59% college education), higher income (75% vs 44% with income more than 300% of poverty level), fewer immigrants (23% vs 29%), and more non-Hispanic whites (75% vs 39%).

Although overall the length of time spent in active transportation did not cluster significantly in the study counties, the average walking/biking time did vary among the high- and low-prevalence clusters. The average time spent walking/biking among walkers/bikers in LAC Cluster 3 is lower than the average time spent outside the clusters (181 vs 192 minutes), whereas the time spent is longer inside than outside the other clusters (239, 193, 264, and 269 minutes for Clusters 1, 2, 4, and 5, respectively). Even though more people were found to walk/bike in SDC Clusters 2 and 3, their time spent is not long compared with those people who do walk/bike outside the clusters (118 and 158 vs 170 minutes per week). However, SDC high-walking Cluster 1 had a longer time spent walking/biking than the area outside clusters (263 minutes in Cluster 1 vs 170 minutes outside SDC clusters).

Discussion

The identification of disease clusters has been a prominent tool for health researchers at least since Snow's analysis of cholera.¹¹ Here, this approach is extended to the analysis of active

transportation, a potential contributor to public health via the beneficial health effects of physical activity. Studies using traditional regression models have suggested that multiple factors may influence walking/biking. For example, transportation walking appears to be influenced by age, gender, auto ownership, and SES,^{9,27} as well as diverse environmental variables.^{28,29} In the current paper, it is demonstrated that spatial scan techniques can be used to find neighborhood scale clusters of frequency and duration of walking and biking in two California counties. This demonstration could help to identify unexplored correlates of walking through more detailed analysis of identified neighborhoods and to guide allocation of resources to specific neighborhoods.

Note that cluster detection and regression models are part of a logical sequence of analyses aimed at understanding the determinants of geographic variation in behavior or disease. The spatial scan statistics search for local clusters with high/low active transportation. The comparison here of the factors in/outside a particular cluster is the first step in an attempt to identify factors associated with active transportation. A traditional regression model is the next step in assessing the significance of these and other factors simultaneously. Although this type of model is preferable to the simple univariate comparison tests reported here, some effects in such a model may turn out to be nonsignificant if they have an impact on active transportation selectively in local areas rather than at a countywide level.

Application of the spatial scan method to CHIS data in LAC and SDC identified a number of significant clusters of high and low prevalence of walking or biking for transportation. The current examination of neighborhood characteristics supports previous studies²⁷ suggesting that there are multiple factors that have an impact on physical activity at a very local level. There was no clear profile of a “high walking” place. High-prevalence clusters were consistently found to have significantly higher population and employment density and street connectivity, but the associations between more frequent walking/biking and race/ethnicity, population density, income, and even age varied across the significant clusters. The clusters of high walking prevalence identified in LAC include two communities on the ocean (Long Beach and Santa Monica) and three communities in the downtown and central regions of the city. This finding also supports the argument that walking/biking for transportation is influenced by a range of different demographic, economic, and environmental variables.

Many past studies of active transportation and the environment rely on data aggregated to administrative areas, such as counties, census tracts, or ZIP codes.^{3,6,9} A notable feature of the spatial scan approach is that it uses information about individuals to identify spatial variation in a disease or behavior rather than identifying administrative units differing in some population characteristic. Nevertheless, aggregate data from administrative units containing respondents will often be required for multilevel analyses, and selection of the appropriate unit for such data is critically important, as the degree of clustering and conclusions drawn can differ by geographic scale.³⁰

Conventional power analyses are not possible for spatial scan approaches. Many analytical and simulation studies indicate that the power of detecting clusters varies by geographic features of the study region, the definition of the geographic unit, the strength of the clustering pattern as measured by disparity in relative risks inside/outside clusters, the geographic size and the population of the clusters, and the sample size (total number of events in the whole study region).^{14–18} There is no standard way to obtain a condition (such as a sample size) that achieves a pre-specified power level for this kind of analysis. However, it has been found¹⁸ that SaTScan has power greater than 90% to detect clusters with a relative risk of 1.5 when the sample size was 2500 (smaller sizes were not tested). In the current study, the sample size was much smaller in SDC (680) than in LAC (about 3600), which may be one of the reasons that

more clusters were detected in LAC compared to SDC and that no cluster was found to be related to active transportation duration in SDC.

In contrast to other cluster-detection techniques (e.g., Getis-Ord Gi*,³¹ AMEOBA,³² ULS³³), SaTScan is the only one that has the ability to adjust for covariates. To illustrate this function, an adjustment was made for age in the analysis of spatial clusters of active transportation in LAC and SDC. If a covariate of interest is not significantly different inside and outside a cluster, then adjustment will not influence the detection of the clusters. However, adjustment for covariates that do differ can alter the significance of a cluster or eliminate it, as it did for the LAC study in the current analysis, depending on the correlations among trait values of respondents constituting the cluster.

There are several limitations to this study. First, estimates of frequency and duration of walking and biking for transportation are based on self-reports. Such reports are likely to overestimate amount of walking.³⁴ Second, data are available for locations of current residences without regard to how long the respondents had lived there or where they lived previously, which may have influenced their inclination to walk or bike. Data on important characteristics of the built environment that may influence residents to walk or bike for transportation, such as neighborhood land use mix,^{27,35} are also lacking. Third, this approach identifies circular clusters, but real “neighborhoods” or walking domains are likely to have more complex shapes.^{23,33,36} Fourth, different buffer sizes or shapes could give different neighborhood characteristics, which may lead to a slightly different summary of some of the variables in Table 2 and Table 3. Caution is required in the interpretation of the neighborhood level effects. It is well recognized that associations based on aggregated characteristics may not hold for individuals, but because the aggregation is over a small buffer area and is meant to describe conditions in the neighborhood rather than for the individuals who live there, this is likely only a minor problem. Fifth, the selection of the maximum window size in this study is subjective. Therefore, this kind of analysis provides only clues to the local relationship between active transportation and selected covariates. Other study designs are required to confirm these associations.

Despite these limitations, the spatial scan method is a promising approach to learning more about geographic variation in the prevalence of health behaviors. The method requires geographic identifiers at the individual level. Choice of a maximum window size for the cluster scan will be a compromise between including a population large enough to provide sufficient statistical power to detect significant clusters and small enough to identify local neighborhoods of interest consistent with the goals of the study. The spatial scan approach has the potential to be part of a toolkit for researchers to find new correlates of active transportation and other health behaviors and a method for policymakers to identify areas to serve as positive examples of walker/biker-friendly neighborhoods or neighborhoods that might provide opportunities for intervention.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Bill Davis for advice on weighting of the CHIS data set, Robert Adamski for data extraction efforts, and several anonymous reviewers for helpful comments. We also acknowledge the CHIS team and CHIS respondents for creating a fascinating data resource. No financial disclosures were reported by the authors of this paper.

References

1. Sallis JF, Bauman A, Pratt M. Environmental and policy interventions to promote physical activity. *Am J Prev Med* 1998;15(4):379–397. [PubMed: 9838979]
2. Berrigan D, Troiano RP. The association between urban form and physical activity in U.S. adults. *Am J Prev Med* 2002;23(2S):74–79. [PubMed: 12133740]
3. Ewing R, Schmid T, Killingsworth R, Zlot A, Raudenbush S. Relationship between urban sprawl and physical activity, obesity, and morbidity. *Am J Health Promot* 2003;18(1):47–57. [PubMed: 13677962]
4. Gordon-Larsen P, Nelson MC, Page P, Popkin BM. Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics* 2006;117(2):417–424. [PubMed: 16452361]
5. Swinburn B, Egger G, Raza F. Dissecting obesogenic environments: the development and application of a framework for identifying and prioritizing environmental interventions for obesity. *Prev Med* 1999;29(6 Pt 1):563–570. [PubMed: 10600438]
6. Wendel-Vos W, Droomers M, Kremers S, Brug J, van Lenthe F. Potential environmental determinants of physical activity in adults: a systematic review. *Obes Rev* 2007;8(5):425–440. [PubMed: 17716300]
7. IOM. Examining the Evidence—special report 282. Washington DC: National Academies Press; 2005. Does the built environment influence physical activity?.
8. Papas MA, Alberg AJ, Ewing R, Helzlsouer KJ, Gary TL, Klassen AC. The built environment and obesity. *Epidemiol Rev* 2007;29:129–143. [PubMed: 17533172]
9. Frank, L.; Engelke, P.; Schmid, T. Health and community design: the impact of the built environment on physical activity. Washington DC: Island Press; 2003.
10. Koch, T. Cartographies of disease: maps, mapping, and medicine. Redlands CA: ESRI Press; 2005.
11. Snow, J. On the mode of communication of cholera. London: John Churchill; 1855.
12. Kulldorff M. Tests of spatial randomness adjusted for an inhomogeneity: a general framework. *J Am Stat Assoc* 2006;101(475):1289–1305.
13. Kulldorff M. A spatial scan statistic. *Commun Stat Theory Methods* 1997;26(6):1481–1496.
14. Song C, Kulldorff M. Power evaluation of disease clustering tests. *Int J Health Geogr* 2003;2(1):9. [PubMed: 14687424]
15. Kulldorff M, Tango T, Park P. Power comparisons for disease clustering tests. *Comput Stat Data Anal* 2003;42(4):665–684.
16. Waller LA, Hill EG, Rudd RA. The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. *Stat Med* 2006;25(5):853–865. [PubMed: 16453372]
17. Kulldorff M, Song C, Gregorio D, Samociuk H, DeChello L. Cancer map patterns: are they random or not? *Am J Prev Med* 2006;30(2):S37–S49. [PubMed: 16458789]
18. Huang L, Pickle LW, Das B. Evaluating spatial methods for investigating global clustering and cluster detection for cancer cases. *Stat Med* 2008;27:5111–5142. [PubMed: 18712778]
19. Kulldorff M, Feuer EJ, Miller BA, Freedman LS. Breast cancer clusters in the northeast United States: a geographic analysis. *Am J Epidemiol* 1997;146(2):161–170. [PubMed: 9230778]
20. Mostashari F, Kulldorff M, Hartman JJ, Miller JR, Kulasekera V. Dead bird clusters as an early warning system for West Nile virus activity. *Emerg Infect Dis* 2003;9(6):641–646. [PubMed: 12781002]
21. Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. *Emerg Infect Dis* 2004;10(5):858–864. [PubMed: 15200820]
22. Kulldorff M, Mostashari F, Duczmal L, Yih KW, Kleinman K, Platt R. Multivariate scan statistics for disease surveillance. *Stat Med* 2007;26(8):1824–1833. [PubMed: 17216592]
23. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. *Stat Med* 2006;25(22):3929–3943. [PubMed: 16435334]
24. Klassen AC, Kulldorff M, Curriero F. Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. *Int J Health Geogr* 2005;4(1):1. [PubMed: 15649329]

25. Huang L, Kulldorff M, Gregorio D. A spatial scan statistic for survival data. *Biometrics* 2007;63(1): 109–118. [PubMed: 17447935]
26. American Association for Public Opinion Research. Standard definitions: final dispositions of case codes and outcome rates for surveys. Vol. 3rd Ed.. Lenexa KS: American Association for Public Opinion Research; 2004.
27. Saelens BE, Sallis JF, Frank LD. Environmental correlates of walking and cycling: findings from the transportation, urban design, and planning literatures. *Ann Behav Med* 2003;25(2):80–91. [PubMed: 12704009]
28. Lee C, Moudon AV. Correlates of walking for transportation or recreation purposes. *J Phys Act Health* 2006;3(1):S77–S98.S
29. Pikora TJ, Giles-Corti B, Knuiman MW, Bull FC, Jamrozik K, Donovan RJ. Neighborhood environmental factors correlated with walking near home: using SPACES. *Med Sci Sports Exerc* 2006;38(4):708–714. [PubMed: 16679987]
30. Waller, LA.; Gotway, CA. *Applied spatial statistics for public health data*. Hoboken NJ: John Wiley & Sons; 2004.
31. Aldstadt J, Getis A. Using AMOEBA to create a spatial weighted matrix and identify spatial clusters. *Geogr Anal* 2006;38:327–343.
32. Ord JK, Getis A. Local spatial autocorrelation statistics: distribution issues and an application. *Geogr Anal* 1995;27(4):286–306.
33. Patil GP, Taillie C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environ Ecol Stat* 2004;11:183–197.
34. Troiano RP, Berrigan D, Dodd KW, Masse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. *Med Sci Sports Exerc* 2008;40(1):181–188. [PubMed: 18091006]
35. Oliver LN, Schuurman N, Hall AW. Comparing circular and network buffers to examine the influence of land use on walking for leisure and errands. *Int J Health Geogr* 2007;6:41. [PubMed: 17883870]
36. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. *Int J Health Geogr* 2005;4:11. [PubMed: 15904524]

Figure 1.

(A) Nearest intersection to respondents' residences in Los Angeles County; (B) Clusters of high/low walking/biking prevalence in Los Angeles County, without age-adjustment; (C) Cluster of short walking/biking duration among walkers/bikers in Los Angeles County without covariate adjustment; note that no clusters of long/short walking/biking duration among walkers/bikers were detected in San Diego County; (D) Nearest intersection to respondents' residences in San Diego County; (E) Clusters of high/low walking/biking prevalence in San Diego County without age-adjustment; (F) Clusters of high/low walking/biking prevalence in Los Angeles County with age-adjustment; note that the clusters of high walking/biking prevalence in San Diego County is the same as those detected in Los Angeles County, so the map for San Diego County with age-adjustment is not provided

Table 1
Demographics of subject counties (based on respondents only), California and the U.S., % unless otherwise indicated

Variable	Los Angeles County	San Diego County	Combined*	California*	U.S.**
Area (square km)	12,308	11,721	14,039	410,000	9,631,000
Population (10,000)	966.2	281.3	1247.5	3440	28500
Sample size (n)	8,547	1,891	10,438	56,270	69,244
Gender					
Male	50.7 (0.4)	50.3 (0.7)	50.6 (0.3)	48.9 (0.0)	47.9 (0.1)
Female	49.3 (0.4)	49.7 (0.7)	49.4 (0.7)	51.1 (0.0)	52.1 (0.1)
Race/ethnicity					
Non-Hispanic white	39.4 (0.4)	61.2 (0.8)	44.6 (0.3)	50.2 (0.0)	73.6 (0.4)
Non-Hispanic black	9.9 (0.2)	5.1 (0.5)	8.8 (0.2)	5.9 (0.0)	11.2 (0.3)
Hispanic	38.2 (0.4)	22.8 (0.8)	34.5 (0.4)	29.3 (0.0)	10.6 (0.2)
Other	12.5 (0.3)	10.8 (0.6)	12.1 (0.2)	14.5 (0.0)	4.6 (0.2)
Age (years)					
18–39	47.6 (0.4)	47.3 (0.8)	47.5 (0.4)	45.7 (0.0)	39.6 (0.3)
40–59	33.7 (0.4)	32.8 (0.7)	33.5 (0.4)	34.9 (0.1)	38.1 (0.3)
60+	18.8 (0.3)	19.8 (0.8)	19.0 (0.3)	19.4 (0.1)	22.3 (0.2)
Income (% of poverty level)					
<100	17.2 (0.5)	10.6 (1.0)	15.6 (0.4)	15.7 (0.2)	10.0 (0.2)
100–200	21.7 (0.6)	21.0 (1.2)	21.5 (0.6)	20.4 (0.3)	16.8 (0.3)
201–300	14.0 (0.5)	14.2 (1.0)	14.0 (0.5)	14.2 (0.2)	17.2 (0.3)
>300	47.2 (0.7)	54.2 (1.5)	48.8 (0.7)	49.7 (0.2)	56.0 (0.4)
Education					
<High school	20.9 (0.7)	14.6 (1.1)	20.9 (0.5)	21.4 (0.1)	16.7 (0.2)
High school graduate	23.2 (0.5)	25.0 (1.2)	23.2 (0.5)	23.3 (0.1)	30.5 (0.3)
>High school	55.9 (0.7)	60.4 (1.3)	55.9 (0.6)	55.3 (0.2)	52.8 (0.4)
Immigration status					
Born in U.S.	61.1 (0.7)	77.0 (1.1)	64.9 (0.6)	66.7 (0.2)	85.0 (0.3)
Not born in U.S.	38.9 (0.7)	23.0 (1.1)	35.1 (0.6)	33.3 (0.2)	15.0 (0.3)
Years lived in U.S. if not born in U.S.					

Variable	Los Angeles County [*]	San Diego County [*]	Combined [*]	California [*]	U.S. ^{**}
1	2.0 (0.4)	5.6 (1.8)	2.5 (0.4)	2.6 (0.2)	1.7 (0.2)
2-5	5.9 (0.6)	7.5 (1.5)	6.2 (0.6)	8.5 (0.3)	16.7 (0.6)
6-10	10.9 (0.7)	14.6 (2.4)	11.5 (0.7)	14.2 (0.5)	13.6 (0.4)
11-15	22.3 (1.2)	23.1 (3.3)	22.5 (1.0)	21.1 (0.5)	14.3 (0.4)
>15	58.8 (1.2)	49.2 (2.9)	57.3 (1.1)	52.7 (0.7)	53.7 (0.7)

^{*} Data for Los Angeles County, San Diego County, and California were obtained from the 2001 California Health Interview Survey. All results are weighted according to survey design characteristics. Values in parentheses are 1 SE.

^{**} Data for the entire U.S. are from the 2001 National Health Interview Survey.

Table 2

Individual and contextual features associated with clusters of high and low walking prevalence in Los Angeles County, % unless otherwise indicated

	Not in clusters			High-prevalence clusters			Low-prevalence clusters		
	1	2	3	4	5	6	7		
n	1709	235	248	149	218	125	167	255	
Active transportation									
Those using active transportation	40.1	70.2 ⁺	61.7 ⁺	65.8 ⁺	60.6 ⁺	64.0 ⁺	18.6 ⁺	25.5 ⁺	
Average walking/biking time (min)									
Among all	77	168 ⁺	120 ^{**}	120	160 [*]	173 ^{***}	25 ⁺	64	
Among walkers/bikers only	192	239	193	181	264	269	133 ^{**}	251	
INDIVIDUAL CHARACTERISTICS									
Age (years)	47	40 ⁺	41 ⁺	43 ^{**}	41 ⁺	44	47	49 ^{**}	
BMI	26.4	27.5 ^{**}	24.9 ⁺	23.7 ⁺	26.9	28.4 ^{**}	27.1	27.5 ^{***}	
Gender									
Male	42.8	40.9	47.6	43.0	48.2	40.0	48.5	44.3	
Female	57.2	59.2	52.4	57.1	51.8	60.0	51.5	55.7	
Race/ethnicity									
Non-Hispanic white	48.6	10.2	44.4	75.8	43.6	1.6	68.9	58.8	
Non-Hispanic black	10.2	4.7	5.7	4.0	13.8	36.0	5.4	4.3	
Hispanic	27.7	71.5	30.7	8.7	29.8	60.8	18.0	26.7	
Other	13.5	13.6	19.4	11.4	12.8	1.6	7.8	10.2	
Education									
<High school	12.4	44.3	16.1	2.0	16.5	45.6	4.8	8.2	
High school	23.7	21.7	17.7	11.4	22.0	36.8	25.8	30.2	
>High school	63.9	34.0	66.1	86.6	61.5	17.6	69.5	61.6	
Income (% of poverty level)									
<100	12.0	45.5	17.3	6.7	21.6	39.2	3.6	7.5	
100–200	18.5	26.0	22.6	10.1	21.6	36.8	13.2	18.8	
201–300	14.4	12.3	14.1	6.0	15.6	9.6	12.0	14.9	
>300	55.1	16.2	46.0	77.2	41.3	14.4	71.3	58.8	

	Not in clusters	High-prevalence clusters							Low-prevalence clusters						
		1	2	3	4	5	6	7							
Health status		+		+					**		+			**	
Limited a lot	11.2	10.2	10.2	8.1	6.0				16.8		11.4			10.2	
Limited a little	26.1	38.3	38.3	22.6	17.5				36.8		22.2			30.2	
Not limited	62.3	51.5	51.4	69.0	76.5				47.2		66.5			58.4	
Other	0.4	0.0	0.0	0.4	0.0				0.0		0.0			1.2	
Born in U.S.		+	+	***					+					+	
No	30.2	69.8	51.2	17.5	30.3				58.4		10.8			22.0	
Yes	69.8	30.2	48.8	82.6	69.7				41.6		89.2			78.0	
Years lived in U.S. (not born in U.S.)	20	16 ⁺	15 ⁺	26 ⁺	16				15 ⁺		20			22	
Employment status			+		+				+		+			**	
Working at job/business	61.1	54.5	69.7	64.4	55.5				48.0		61.7			61.2	
With a job but not at work	3.3	3.0	2.0	4.7	4.1				0.0		5.4			2.0	
Looking for work	3.8	6.0	7.3	4.7	9.6				6.4		1.8			1.2	
Not working at job/business	31.8	36.6	21.0	26.2	30.7				45.6		31.1			35.7	
BUFFER CHARACTERISTICS															
Population density	4308	9720 ⁺	10272 ⁺	5205 ^{***}	9852 ⁺				6908 ⁺		1682 ⁺			2543 ⁺	
Employment density	1479	6980 ⁺	3463 ⁺	3029 ⁺	1833 ⁺				1092 ⁺		201 ⁺			738 ⁺	
Bus stop in buffer		+	+	*	+				+		+			+	
No	35.3	0.4	2.4	44.3	83.0				0.0		100.0			89.0	
Yes	64.7	99.6	97.6	55.7	17.0				100.0		0.0			11.0	
Freeway		+	*	+	+				+		+			+	
No	79.5	56.6	74.6	92.6	96.8				80.0		92.2			69.0	
Yes	20.5	43.4	25.4	7.4	3.2				20.0		7.8			31.0	
Bus route		+	+		+				+		+			+	
No	27.9	0.0	2.4	33.6	81.2				0.0		100.0			89.0	
Yes	72.1	100.0	97.6	66.4	18.1				100.0		0.0			31.0	
Street connectivity															
Street density index	11.8	13.6 ⁺	13.6 ⁺	13.9 ⁺	17.1 ⁺				13.5 ⁺		8.7 ⁺			10.7 ⁺	

	Not in clusters	High-prevalence clusters					Low-prevalence clusters	
		1	2	3	4	5	6	7
Block density index	35.7	52.5 ⁺	43.6 ⁺	46.5 ⁺	84.0 ⁺	42.6 ⁺	21.3 ⁺	30.2 ⁺
Intersection density index	45.0	59.7 ⁺	56.8 ⁺	57.6 ⁺	97.7 ⁺	51.6	32.7 ⁺	41.9 ⁺
Average block length (km)	0.171	0.138 ⁺	0.148 ⁺	0.154 ⁺	0.111 ⁺	0.168 ⁺	0.217 ⁺	0.164 [*]

Note: The numbers are average values of the corresponding continuous variables, and percentages in each level of the categoric variables in/out of clusters. The *p*-values are obtained by *t*-test for continuous variables such as minutes, age, BMI, and street density; by chi-square test for categoric variables such as gender, health status, and bus routes. The contrasts are among values that are in cluster(s) vs not in cluster(s).

* *p*<0.1
** *p*<0.05
*** *p*<0.01
+ *p*≤0.001

Table 3

Individual and contextual features associated with clusters of high and low walking prevalence in San Diego County, % unless otherwise indicated

	Not in clusters	High-prevalence clusters				Low-prevalence clusters
		1	2	3	4	
<i>n</i>	1751	43	13	44	32	
Active transportation						
Use active transportation	34.4	76.7 ⁺	100.0 ⁺	70.5 ⁺	3.1 ⁺	
Average walking/biking time (min)						
Among all	58	202 ^{***}	118 [*]	111 ^{***}	7 ⁺	
Among walkers/bikers only	170	263 [*]	118 [*]	158	210	
INDIVIDUAL CHARACTERISTICS						
Age (years)	48	49	42	44	53	
BMI	26.2	26.7	25.1	26.4	25.7	
Gender						
Male	43.1	46.5	61.5	43.2	25.0	
Female	56.9	53.5	38.5	56.8	75.0	
Race/ethnicity						
Non-Hispanic white	68.5	51.2	84.6	38.6	93.8	
Non-Hispanic black	4.5	11.6	0.0	0.0	0.0	
Hispanic	17.5	32.6	7.7	50.0	3.1	
Other	9.5	4.7	7.7	11.4	3.1	
Education						
<High school	8.5	14.0	0.0	22.7	0.0	
High school	24.4	30.2	15.4	27.3	6.3	
>High school	67.1	55.8	84.6	50.0	93.8	
Income (% of poverty level)						
		***		**	**	

	Not in clusters	High-prevalence clusters				Low-prevalence clusters
		1	2	3	4	
<100	7.9	23.3	15.9	20.5	0.0	0.0
100–200	18.7	18.6	0.0	29.6	6.3	6.3
201–300	14.1	16.3	7.7	20.5	6.3	6.3
>300	59.3	41.9	76.9	29.6	67.5	67.5
Health status						
Limited a lot	9.4	14.0	7.7	13.6	9.4	9.4
Limited a little	26.3	18.6	23.1	36.4	21.9	21.9
Not limited	63.8	67.4	69.2	50.0	68.8	68.8
Other	0.5	0.0	0.0	0.0	0.0	0.0
Born in U.S.						
No	18.2	27.9	0.0	54.6	6.3	6.3
Yes	81.8	72.1	100.0	45.5	93.8	93.8
Years lived in U.S. (not born in U.S.)	21	18	NA	18	49	49
Employment status						
Working at job/business	62.3	53.5	76.9	50.0	50.0	50.0
With a job but not at work	2.9	2.3	0.0	2.3	3.1	3.1
Looking for work	2.3	4.7	0.0	6.8	3.1	3.1
Not working at job/business	32.6	39.5	23.1	40.9	43.8	43.8
BUFFER CHARACTERISTICS						
Population density (<i>n</i>)	2860	4408 ⁺	4112 ⁺	8698 ⁺	1518 ⁺	1518 ⁺
Employment density (<i>n</i>)	1110	1388 ^{***}	2362 ⁺	3179 ⁺	543 ⁺	543 ⁺
Bus stop in buffer						
No	29.1	2.3	0.0	0.0	59.4	59.4
Yes	70.9	97.7	100.0	100.0	40.6	40.6

	High-prevalence clusters				Not in clusters	Low-prevalence clusters
	1	2	3	4		
Freeway	+		+	**		
No	77.1	44.2	61.5	36.4	100.0	100.0
Yes	22.9	55.8	38.5	63.6	0.0	0.0
Bus route	***	*	***	***	***	***
No	23.1	2.3	0.0	0.0	50.0	50.0
Yes	76.9	97.7	100.0	100.0	50.0	50.0
Street connectivity						
Street density index	10.0	14.5 ⁺	14.2 ⁺	14.1 ⁺	9.3 ^{**}	9.3 ^{**}
Block density index	27.2	82.1 ⁺	54.8 ⁺	47.9 ⁺	24.2	24.2
Intersection density index	38.2	63.8 ⁺	63.4 ⁺	54.1 ⁺	38.1	38.1
Average block length (km)	0.184	0.133 ⁺	0.178 ^{**}	0.147 ⁺	0.157 ^{***}	0.157 ^{***}

Note: The numbers are average values of the corresponding continuous variables, and percentages in each level of the categorical variables in/out of clusters. The *p*-values are obtained by *t*-test for continuous variables such as minutes, age, BMI, and street density; by chi-square test for categorical variables such as gender, health status, and bus route. The contrasts are among values that are in cluster(s) vs not in cluster(s).

* $p < 0.1$

**

$p < 0.05$

$p < 0.01$

⁺ $p \leq 0.001$