Application of Geographic Perturbation Methods to Residential Locations in the Oregon Household Activity Survey

Proof of Concept

Kelly J. Clifton and Steven R. Gehrke

Travel demand models have advanced from zone-based methods to favor activity-based approaches that require more disaggregate data sources. Household travel surveys gather disaggregate data that may be used to inform advanced travel demand models better and also to improve the understanding of how nonmotorized travel is influenced by a household's surrounding built environment. However, the release of these disaggregate data is often limited by a confidentiality pledge between the household participant and survey administrator. Concerns about the disclosure risk of survey respondents to household travel surveys must be addressed before these household-level data may be released at their disaggregate geography. In an effort to honor this confidentiality pledge and facilitate the dissemination of valuable travel survey data, this research (*a*) reviews geographical perturbation methods that seek to protect respondent confidentiality; (*b*) outlines a procedure for implementing one promising practice, referred to as the "doughnut masking technique"; and (*c*) demonstrates a proof of concept for this technique on 10 respondents to a household activity travel survey in the Portland, Oregon, metropolitan region. To examine the balance between limiting disclosure risk and preserving data utility, four trials were conducted and measures of household anonymity and built environment variation were analyzed for the relocated household in relation to its actual location. Results of this demonstration revealed that increases in the potential displacement distance of a geographically perturbed household generally reduced disclosure risk and also limited data utility.

A household travel survey provides a rich amount of disaggregate data to researchers, but the release of these valuable data is often limited by a confidentiality agreement between the survey respondent and administrator. The inability to release such data has been an obstacle to an improved understanding of linkages between travel behavior and the built environment because these disaggregate data are most appropriate for study as they avert the problem of an ecological fallacy (*1*). Accordingly, previous studies examining the relationship of household travel patterns to the built environment have questioned past inferential comparisons of the disaggregate housing unit to a more aggregate neighborhood depiction (*2, 3*) and have confirmed the inherent risk of aggregating household data into the zones traditionally exhibited in four-step travel demand models (*4*). To avoid this modeling pitfall, travel behavior research has continued to advance in the direction of using activity-based travel demand models that rely extensively on disaggregate built environment and socioeconomic measures to properly capture their effects on observed household travel (*5, 6*). Understanding these connections has become ever more important at the same time regional travel demand models increasingly include nonmotorized travel (*7*).

Although the use of disaggregate data pertaining to the household and its surrounding built environment will ultimately improve travel demand modeling, the release of data that may be geographically referenced to an individual household raises important concerns about the preservation of survey respondent confidentiality. Although largely absent in travel behavior research, the negotiation of this complication, centered on protecting respondent privacy while ensuring the possibility of valid geographical analyses, has received recent attention in public health research (*8*), which has sought to address this conflict by applying geographic perturbation methods (*9*). The impetus behind using geographic perturbation approaches has been to spatially modify the precise household location to an extent sufficient for these sensitive data to be made available to external users (*10*). Applying a geographic perturbation method seeks to minimize the disclosure risk or breach in confidentiality that may permit a user of the spatial data source to discern either the identity of a respondent or any associated attribute observed under a confidentiality pledge (*11*).

These disclosure risks must be weighted by the data custodian against the complementary concept of data utility, which defines the retained value of the geographically altered data source to the external user. For these data to be disseminated to a wider audience, the introduction of spatial error must be minimized by the data custodian. This notion of spatial error minimization is important especially to travel demand modelers interested in the relationships between nonmotorized travel and the built environment. Walk trips often occur over short distances (e.g., ½ mi). As such, the geographic perturbation of a trip origin (e.g., household) beyond this threshold distance may significantly limit an improved understanding of this small-scale connection between nonmotorized travel and the built environment. In general, decreasing the level of disclosure risk by applying a more stringent geographic perturbation method will also decrease the accuracy of inferences obtainable from the perturbed data set (*12*).

Department of Civil and Environmental Engineering, Portland State University, P.O. Box 751, Portland, OR 97207-0751. Corresponding author: K. J. Clifton, kclifton@pdx.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2354, Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 40–50.
DOI: 10.3141/2354-05

Although this trade-off between disclosure risk and data utility has been recognized, no consensus on a geographic perturbation method that does not affectedly limit any spatial analysis to visualize and disseminate confidential data has been reached (*13*). In contribution to this methodological debate, this research provides an overview of geographic perturbation methods in practice and details an approach toward using one promising technique to enable a wider dissemination of household travel survey data. The methodological overview is then followed by a proof of concept of this innovative technique on a limited number of respondents to the 2010 Oregon Household Activity Survey in the Portland metropolitan region, chosen according to the diversity in surrounding built environments, and concludes with a discussion of potential directions for future applications of this technique.

METHODOLOGICAL BACKGROUND

Traditionally, the most common geographic perturbation method has been to conceal individual records through a process of aggregation through either zonal or point aggregation (*10*). The former technique enumerates all households in a predefined administrative entity; whereas, the latter technique allocates many households to a single location (*9*). The use of zonal aggregation has been a longstanding practice in conventional travel demand modeling and has consequently resulted in introducing biases related to ecological correlation, limited transferability, data inefficiency, policy insensitivity, and a lack of behavioral fidelity (*14*). Point aggregation has often served as a counterpart to zonal aggregation through the practice of trip assignment linking geometric centers of traffic analysis zones (*15*). The pervasiveness of this geographic perturbation method has been attributed to the ease for the data custodian, who may be technically or practically limited to conduct this approach, but its application certainly limits the conceivable utility of household travel survey data (*13*).

A second set of geographic perturbation methods reflects affine point transformations in which household data points may be deterministically repositioned to a new set of geographic locations (*16*). One specific affine point transformation technique, translation, laterally shifts household locations a specified distance and direction from their original location while preserving geographic scale (*17*). A second affine point transformation technique alters each household location by multiplying each geographic coordinate of the original household location by a scaling constant (*10*). This multiplying of a constant differentiates the scale change affine point transformation technique from the translation technique in which a constant is added to each geographic coordinate. Another affine point transformation technique rotates the location of all households in a data set by a fixed angle selected by the data custodian (*9*). Alternatively, the data custodian may choose to translate the original household location before the rotation has been applied so as to pivot the perturbed location from some arbitrary position. Correspondingly, an affine point transformation approach may be performed that uses any combination of the three techniques of translation, scaling change, or rotation to mask the original household location of the survey respondent.

Applying an affine point transformation provides a more disaggregate household representation, thus improving the utility of the data set to the user interested in activity-based travel demand models or nonmotorized transportation. Moreover, the use of an affine point transformation preserves household representation at a scale necessary to avoid any ecological fallacy when corresponding disaggregate built environment measures are used. Nevertheless, although affine point transformation may potentially increase the utility of travel survey data for the user, applying this method may also fail to mitigate the risk of disclosing the true location of the household to a data intruder. The reason for this increased risk is that all perturbed household locations are deterministically repositioned through the use of a constant specified by the data custodian, which if identified by a data intruder would allow the reverse geocoding of the perturbed location back to its original location.

This disclosure risk has been intuitively referred to as "identity disclosure" and refers to this ability of a data intruder to directly associate a household to a record in a publicly released data set. According to the hierarchical risk framework, the highest level of disclosure risk, called a first-tier risk, is the realization of this direct one-to-one correspondence between the perturbed and actual household location without the use of any supplemental external data source (*18*). Having established this link, and given the ease of access to readily available secondary data sources, a data intruder may then hypothetically associate the household's identity to other characteristics unique to this household. This potential to release additional sensitive household data, recognized as attribute disclosure, is a second-tier risk (*19*).

To provide added protection against these forms of disclosure risk, the data custodian may opt to perform a geographic perturbation method by randomly selecting the translation distance, scaling distortion, and rotation direction for each spatially relocated household (*9*). This perturbation approach, which has also been referred to as "jittering" (*20*), differs from the affine point transformation technique by imposing a heuristic that each household in the travel survey data set must be uniquely and randomly displaced from its original spatial position as opposed to each household being displaced identically and deterministically (*10*). Kwan et al. illustrated three random spatial perturbation techniques in which the perturbed household location has been randomly placed (*a*) within a polygon whose radius or perimeter has been defined relative to the original household location, (*b*) along a line feature such as the radius of a circle chosen by the data custodian with a center representative of the original location, or (*c*) within a circle where the original household location defined the circle's center (*17*).

In addition, these random spatial perturbation techniques have accounted for the population density characterizing the household's neighborhood to inform the random translation of the perturbed location (*21*). This added consideration displaces households residing in a neighborhood with a lower population density by greater distances than survey respondents who reside in neighborhoods characterized by a higher population density because the former household has a greater risk of first-tier disclosure (*10*). Therefore, a random geographic perturbation technique incorporating a level of displacement dependent on the residential density of the neighborhood would appear to have a greater prospect for reducing disclosure risk than an affine point transformation. Analogous to an affine point transformation, a random geographic perturbation technique would also have an enriched potential for improving data utility over an aggregation method. Nonetheless, one shortcoming of this improved technique is the potential for a perturbed household location to be randomly repositioned close to the original location, which compromises the intention to negate the likelihood of a one-to-one correspondence.

To reduce the potential for identity disclosure, an adaptive random geographic perturbation technique has recently been proposed in which the relocated household must be displaced from the actual household location by a minimum distance assigned by the data custodian (*22, 23*). This random geographic perturbation technique generates a torus, or doughnut-shaped space, around the original household location that represents the potential area in which a household may be repositioned. This doughnut masking technique has shown initial promise toward improving the risk of disclosure with a negligible effect on the specificity and sensitivity of detecting clustering patterns or trends associated with the utility of the data source (*22*). As with the aforementioned random geographic perturbation techniques, precaution must be practiced by the data custodian when defining the radii associated with the inner and outer rings of the doughnut because the greater the distance between the actual household location and the perturbed location, the greater the spatial arrangement attributed to the original data may deviate (*17*).

METHODOLOGICAL APPROACH

Hampton et al. described two strategies for applying a doughnut masking technique to a household data source that either (*a*) allows the geographically perturbed household to be relocated outside its original zonal boundary or (*b*) restricts the geographically perturbed household to remain within its original neighborhood (*22*). The former approach may be more suitable for use with the disaggregate data synonymous with activity-based travel demand models and their adherence to comparable geographic scales, whereas the latter seems more suitable for use in a four-step model reliant on aggregate data sources. As such, the selection and subsequent construction of inner and outer rings of a doughnut masking technique appropriate for an analysis of disaggregate built environment measures will be discussed.

In this methodological approach, the outer ring of the doughnut represents the maximum distance a household may be displaced from its original geographic location and may be thought of as a strategy to limit spatial error introduced to the perturbed data set. By constraining the maximum distance a household may be repositioned, the data custodian has attempted to minimize any variation introduced to the true relationship of a built environment measure with the survey respondent's actual household location when compared with the relationship of the same measure with the geographically perturbed household location At a disaggregate scale, the built environment influences household travel behavior, particularly the choice of nonmotorized modes. Thus, a reduction in the possible displacement distance of a household repositioned by the application of a geographic perturbation method must be sought by the data custodian.

The connection between the built environment and nonmotorized travel has commonly been measured through the use of straight line buffers extending from the household's physical location (*24*) with ½-mi and 1-mi buffers often applied (*25, 26*). In accordance, these two distances were explored as radii values for the doughnut's outer ring (r_{or}).

In complement to the outer ring and its relationship with the maximization of data utility, the intention of the inner ring of the doughnut masking technique is to minimize the disclosure risk associated with a data intruder properly identifying a household respondent from a perturbed data set. Although the objective of disclosing perturbed household-specific data is to ensure that the data are sufficiently anonymous, the decision concerning an acceptable level of anonymity is often left to the data custodian to make (*27*). The following adaptation of a formula used by Allshouse et al. allows the data custodian to select a radius for the inner ring, r_{ir}, relative to the geographic area of the underlying neighborhood (e.g., census tract) that assumes a homogeneous residential density (*23*).

$$r_{\text{ir}} = \sqrt{\left(\left(\frac{\text{area}}{\pi}\right)\times k\right)} \tag{1}$$

In Equation 1, the inner ring radius for a household in a particular neighborhood is equal to the square root of the product of the quotient of the neighborhood's area divided by π, which is multiplied by the number of occupied households in the neighborhood selected by the data custodian as providing a sufficient level of anonymity (*k*). In this study, two separate inner ring perimeters were selected, which reflect 0.5% and 1% of the total number of occupied households in the neighborhood, respectively. The combination of these two inner ring radii with the aforementioned outer ring radii defined the four doughnut size trials explored in this study.

After establishing the proper bounds, the data custodian then must randomly reposition the household into the doughnut circling the original location. Rushton et al. provide a methodological framework, adjusted for this random geographic perturbation process, which enables the data custodian to convert the displacement distance from customary units of feet and miles to latitudinal and longitudinal decimal degrees, DD, through the use of the following two equations (*10*):

$$\Delta\text{DD}_y = \frac{v_a}{(69.17\times r_{\text{or}})} \tag{2}$$

$$\Delta\text{DD}_x = \frac{\left(\dfrac{v_b}{69.17\times r_{\text{or}}}\right)}{\left(\cos\left(\dfrac{\text{lat}_1}{180}\right)\times\pi\right)} \tag{3}$$

In Equation 2, the change in decimal degrees for the *y*-coordinate is represented as ΔDD_y and the parameter v_a represents a random value between $-r_{\text{or}}$ and $-r_{\text{ir}}$ or r_{ir} and r_{or}, while the constant, 69.17, represents a scalar factor to compensate for the east–west narrowing of the distance between two meridians as they near either pole (*10*). Equation 3, which converts a customary unit distance into the decimal degrees required for altering the longitudinal measure, uses the same scalar factor as well as the parameter v_b to represent a unique random value in the same range as v_a and the latitudinal coordinate associated with the actual household location, lat_1, to calculate the change in decimal degrees of the *x*-coordinate, ΔDD_x. The conditions of Formula 4 and Formula 5 must additionally be met for the data custodian to guarantee the perturbed location will be in the doughnut:

$$r_{\text{ir}}^2 \le v_a^2 + v_b^2 \tag{4}$$

$$r_{\text{or}}^2 \ge v_a^2 + v_b^2 \tag{5}$$

Once the displacement distance has been calculated in decimal degrees, Equation 6 and Equation 7 are then calculated to define the new latitude and longitude coordinates of the household geographically perturbed by applying the doughnut masking technique.

$$\text{lat}_2 = \text{lat}_1 + \Delta\text{DD}_y \tag{6}$$

$$\text{long}_2 = \text{long}_1 + \Delta\text{DD}_x \tag{7}$$

EMPIRICAL APPLICATION AND PROOF OF CONCEPT

To understand better the potential effect of applying a geographic perturbation method to travel survey data as it relates to preserving data utility and respondent confidentiality, this research tested the doughnut masking technique on 10 Portland area respondents of the 2010 Oregon Household Activity Survey as a proof of concept. Each respondent belongs to a household in a unique neighborhood, defined here as a 2010 U.S. census tract, in the metropolitan region selected by the research team on the basis of its spatial location in the three-county study area and variation in population density (Figure 1). The built environment measures examined for each of the selected neighborhoods and the magnitude of these measures as they relate to the randomly selected household in each neighborhood are provided in Table 1.

These select households were then geographically perturbed through a simulation process performed for each of the four combinations of doughnut sizes described in the methodological approach. The first trial randomly perturbed the original household location in each neighborhood 100 times within a doughnut-shaped area with an outer ring radius extending ½ mi from the original location and an inner ring with a radial distance associated with the masking of the household by 0.5% of all occupied households in the neighborhood. The second trial maintained the same outer ring radius, but extended the inner ring radius unique to each neighborhood from a distance relative to 0.5% of the occupied households in the neighborhood to a distance related to 1% of the occupied households. The third and fourth trials extended the outer ring to a distance of 1 mi with the third trial having an inner ring radius equivalent to the first trial and the fourth trial having an inner ring radius equivalent to the second trial. In total, each household location was geographically perturbed 400 times during four trials in an effort to examine the effect applying a doughnut masking technique had on data utility and disclosure risk. Figure 2 illustrates this simulation process for one household being geographically perturbed in each of these four trials.

Analysis of Data Utility

After the doughnut masking technique was conducted on one household in each of the 10 distinct neighborhoods, the percent root mean square

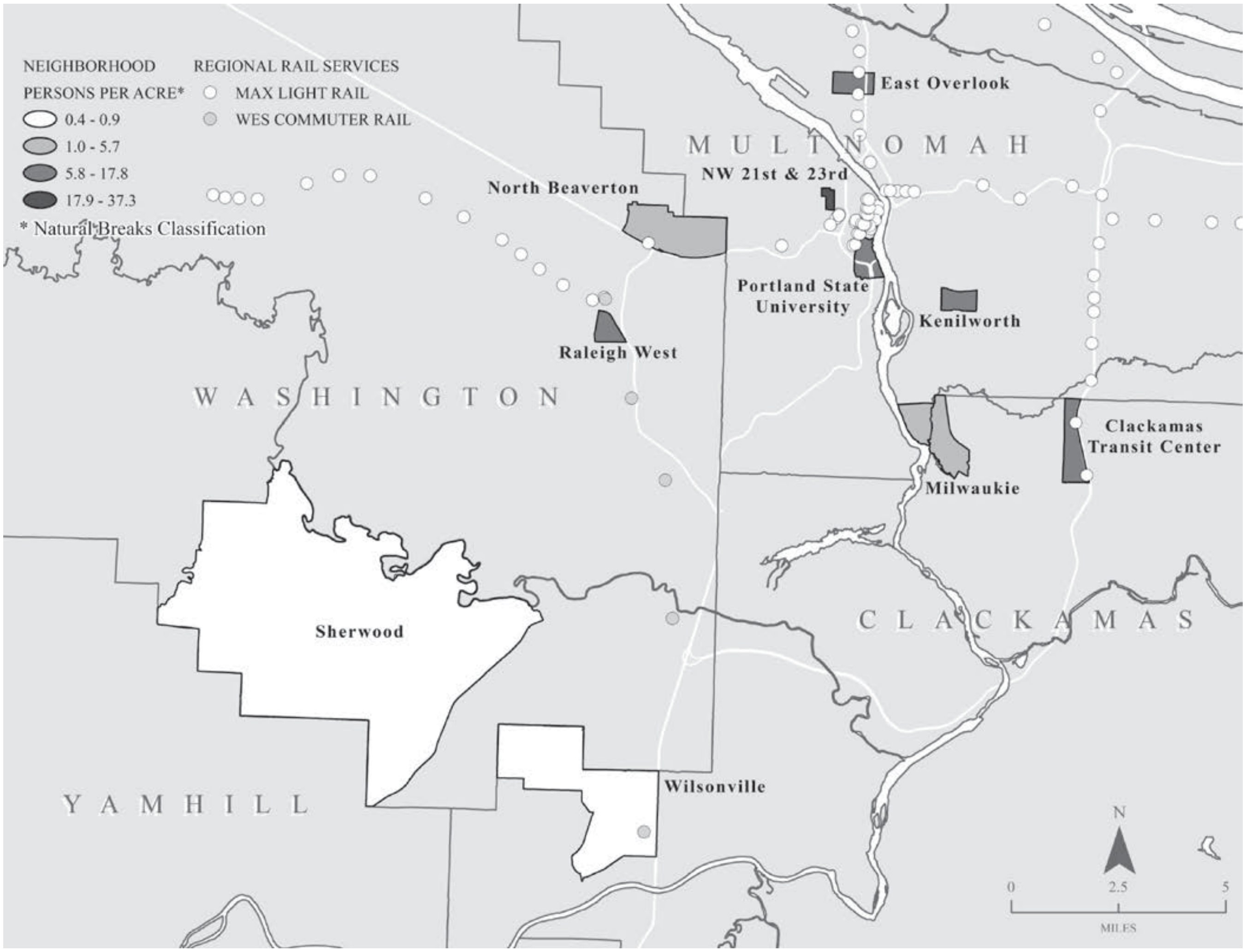
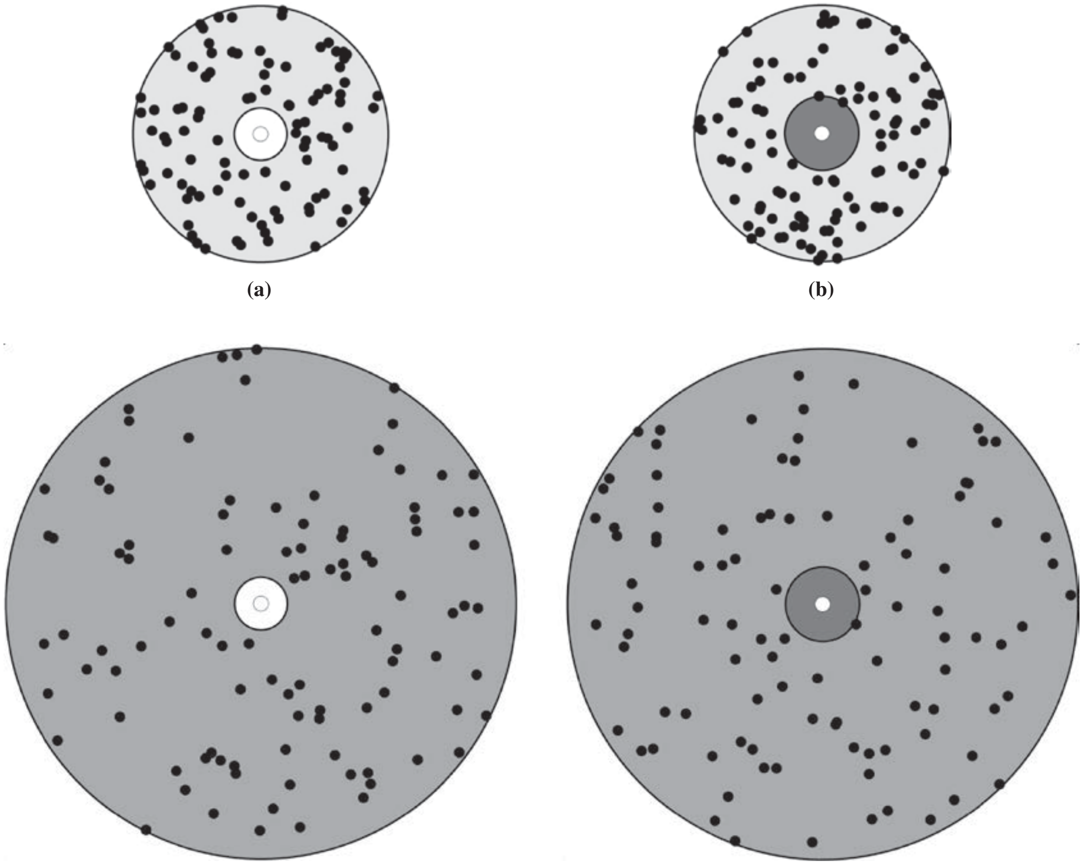


FIGURE 1 Population density of 10 neighborhoods in the Portland metropolitan region.

TABLE 1 Built Environment Measures for Selected Household from 10 Portland Neighborhoods

	Built Environment Variable							
Portland Neighborhood	Persons per Acre (½-mi buffer)	Jobs per Acre (½-mi buffer)	Jobs per Person (½-mi buffer)	Intersections per Acre (½-mi buffer)	Distance to Nearest Park (ft)	Distance to Nearest Grocery (ft)	Distance to Nearest Rail Station (ft)	Distance to Nearest Bus Stop (ft)
Milwaukie	3.90	4.49	1.15	0.23	1,163	2,682	9,067	528
	5.96	4.37	0.81	0.26	537	2,910	16,026	314
Clackamas Transit Center	9.69	10.88	1.12	0.30	822	1,366	11,106	622
	10.68	6.65	0.64	0.23	935	1,843	2,577	303
Wilsonville	0.86	1.14	1.32	0.06	385	5,682	5,987	11,036
	4.01	0.74	0.17	0.18	433	7,374	5,886	827
Kenilworth	17.80	3.39	0.19	0.64	875	1,312	14,127	538
	14.84	4.65	0.34	0.50	938	709	10,650	219
East Overlook	11.81	1.36	0.11	0.71	372	2,152	3,366	669
	10.86	2.87	0.27	0.54	491	1,762	1,796	339
Northwest 21st and 23rd Avenue	37.29	49.56	1.33	0.97	651	679	6,527	310
	26.81	25.61	0.96	0.79	476	455	1,167	141
Portland State University	11.66	50.16	4.30	0.87	629	2,712	12,839	258
	14.35	69.22	4.97	0.71	200	3,477	586	112
North Beaverton	5.73	5.49	0.96	0.22	527	2,909	4,911	566
	5.93	3.99	0.81	0.22	341	3,479	3,825	404
Raleigh West	11.25	3.27	0.29	0.45	450	2,575	9,564	424
	11.54	5.20	0.55	0.39	435	1,501	3,947	235
Sherwood	0.40	0.05	0.14	0.02	421	12,642	3,968	12,148
	2.37	0.22	0.42	0.07	2,646	6,191	31,516	10,671

NOTE: Data in rows for each Portland neighborhood represent unit of analysis: Row 1 = U.S. census tract; Row 2 = OHAS household.



(a)

(b)

(c)

(d)

FIGURE 2 Four simulation trials of the doughnut-masking technique for one selected household: (*a*) Trial 1: 0.5% inner ring, ½-mi outer ring; (*b*) Trial 2: 1% inner ring, ½-mi outer ring; (*c*) Trial 3: 0.5% inner ring, 1-mi outer ring; and (*d*) Trial 4: 1% inner ring, 1-mi outer ring.

error (RMSE) was calculated for each built environment measure as a representation of the variation introduced by geographically perturbing the original household location. An evaluation of the RMSE in the four trials informs the effect of geographically perturbing a household with the doughnut masking technique in relation to preserving the utility of the original data set. Although there is a conceptual understanding among travel demand modelers that a geographically perturbed data set reduces the danger of a first-tier disclosure risk, there has been little objective research to define an acceptable tolerance for introduced spatial error. By geographically perturbing a household location in this empirical application, spatial error has been added to the resulting data set. This proof of concept analyzes the extent of this spatial error, or variation, in eight built environment measures introduced by applying the doughnut masking technique. The level of introduced spatial error ultimately informs the ability of the data custodian intent on improving the understanding of nonmotorized travel to disseminate disaggregate household data to a wider audience.

Table 2 denotes the percent RMSE results for the two trials, with a ½-mi outer ring to examine the variation in built environment measures introduced by geographically perturbing the original household. Likewise, Table 3 compares the variation of these built environment measures on the perturbed household locations repositioned within a 1-mi outer ring from the actual association of these measures on the household, but with an enlarged inner ring radius that potentially relocates the household a greater distance from its true location.

In general, a cross-table examination found that the RMSE values linked to the built environment measures for the two trials with a ½-mi outer ring were lower than the RMSE values for built environment measures for the two trials with a 1-mi outer ring radius. Intuitively, one may have hypothesized that an increase in the maximum allowable displacement distance, as in Trials 3 and 4, would increase the level of spatial error introduced to the data set. An examination into this expected trend revealed that the households in the most densely populated neighborhoods in Multnomah and Washington Counties, Northwest 21st and 23rd Avenue and Raleigh West, had a greater RMSE value for all built environment measures in the trials with a greater outer ring radius. Comparatively, the households of Wilsonville and Sherwood showed an increase in spatial error to be consistent with an increase in the tolerable displacement distance. These two neighborhoods are most reflective of a traditional suburban setting with lower residential and employment densities and greater distances to transit. However, instances of inconsistency in the hypothesized trend were found in the variation for specific built environment measures in the remaining six neighborhoods.

The percentage of introduced spatial error for built environment measures varied across urban and suburban contexts on the basis of the classification of the measure. Examining the percent RMSE for the two census tracts farthest from Portland's central business district, the two built environment measures related to employment and the distance to nearest park measure exhibited a high percentage of spatial error during the four trials. As for the two households closest to Portland's city center, the built environment measures with a consistently high percent RMSE included distance to nearest bus stop and nearest park. Overall, with the exception of the nearest-to-park built environment measure, those measures related to destination accessibility and distance to transit exhibited lower spatial variation in the more suburban contexts with lower residential densities, whereas the employment density and jobs–housing balance diversity measures exhibited the most variation in these suburban contexts. For the households in the six neighborhoods with the highest residential density, the four distance-related measures generally had the highest percent spatial error introduced by geographically perturbing the household location. This trend held true, especially for those four households located in higher density neighborhoods outside the Portland city center in which the only measure with a percent RMSE greater than 100 was the jobs–housing balance measure for the fourth trial.

Comparing the third with the first trial, in which the outer ring radius was extended while maintaining the smaller inner ring radius, the analysis of spatial error revealed three built environment measures that consistently increased in variation across all neighborhoods. One of the four density-related measures of the built environment, intersections per acre, was found to have a percent RMSE value consistently lower for the smaller doughnut size; whereas two distance-related built environment measures, feet to nearest grocery store and rail station, had lower RMSE values associated with the first trial characterized by the shortest inner ring and outer ring radius combination. This behavioral stability in the spatial error of these built environment measures as well as the performance of the residential density measure noted in Table 4 make these suitable candidates for future applications of the doughnut masking technique seeking to link the outer ring radius to variation in a specific built environment measure.

Analysis of Disclosure Risk

As a supplement to the examination above into data utility preservation, a comparison of the disclosure risk associated with each of the four trials was conducted. Whereas the size of the doughnut's outer ring was fluctuated to examine built environment variation in the context of conserving data utility, the size of the doughnut's inner ring may be increased or decreased depending on the anonymity level deemed necessary by the data custodian. Similar to the discussion about consensus on a geographic perturbation method most appropriate for application to the data set, there have also been a variety of statistical approaches tested to quantify the risk of disclosure attributed to a geographically perturbed household (*8*). Sweeney proposed one such measurement for determining a one-to-one correspondence between the geographically perturbed and original household location referred to as the *k* anonymity requirement (*27*). When disclosure risk is examined, the concept of *k* anonymity may be understood in regard to *k* representing the number of households from which the original household cannot be reversely identified (*21*). An estimation of this *k* anonymity statistic may be calculated by using Equation 8 (*23*):

$$\hat{k} = \pi \times d^2 \times \left(\frac{N}{\text{area}}\right) \tag{8}$$

The estimated *k* anonymity statistic in Equation 8 is equal to the product of π and the squared distance between the original household location and the geographically perturbed location multiplied by the number of occupied households in the household's neighborhood (*N*) divided by the land area of the neighborhood encompassing the household. Akin to the previous analysis of built environment variation attributed to a resizing of the outer ring of the doughnut, the estimated *k* anonymity statistic was calculated for the selected households in each of the 10 neighborhoods during the four trials. Table 4 reports the median estimated *k* anonymity statistic as well as the minimum and maximum estimated *k* anonymity statistics found in each neighborhood in the 100-simulation sample for each trial.

TABLE 2 Percent Root Mean Square Error of Built Environment Measures for Trials with 0.5% Inner Ring Radius

Portland Neighborhood	Trial: Outer Ring Radius (mi)	Persons per Acre (½-mi buffer) (%)	Jobs per Acre (½-mi buffer) (%)	Jobs per Person (½-mi buffer) (%)	Intersections per Acre (½-mi buffer) (%)	Distance to Nearest Park (ft) (%)	Distance to Nearest Grocery (ft) (%)	Distance to Nearest Rail Station (ft) (%)	Distance to Nearest Bus Stop (ft) (%)
Milwaukie	Trial 1: 0.5	59	**68**	**113**	49	154	49	10	48
	Trial 3: 1.0	64	**56**	**86**	51	418	77	20	49
Clackamas Transit Center	Trial 1: 0.5	**33**	29	37	18	**189**	32	59	94
	Trial 3: 1.0	**25**	73	73	26	**183**	42	95	154
Wilsonville	Trial 1: 0.5	57	185	313	44	1,196	18	21	12
	Trial 3: 1.0	76	559	16,148	61	1,635	34	39	22
Kenilworth	Trial 1: 0.5	26	34	**73**	10	**74**	144	15	**118**
	Trial 3: 1.0	26	46	**70**	23	**62**	177	29	**114**
East Overlook	Trial 1: 0.5	9	27	24	12	**54**	61	77	60
	Trial 3: 1.0	20	41	82	20	**53**	104	149	57
Northwest 21st and 23rd Avenue	Trial 1: 0.5	20	50	49	20	1,682	170	56	437
	Trial 3: 1.0	56	104	260	31	2,421	333	94	685
Portland State University	Trial 1: 0.5	75	79	90	25	282	30	249	**128**
	Trial 3: 1.0	87	101	90	38	328	49	476	**116**
North Beaverton	Trial 1: 0.5	23	**714**	**1,255**	12	57	31	41	58
	Trial 3: 1.0	29	**637**	**1,153**	19	87	50	74	73
Raleigh West	Trial 1: 0.5	46	33	46	17	61	73	52	181
	Trial 3: 1.0	70	46	61	36	91	80	91	236
Sherwood	Trial 1: 0.5	96	365	373	97	99	25	6	29
	Trial 3: 1.0	136	2,552	993	125	149	45	11	51

NOTE: Boldface = instances of greater percent RMSE for respective built environment measure in trial with 1-mi outer ring radius when compared with trial in ½-mi outer ring radius.

TABLE 3 Percent Root Mean Square Error of Built Environment Measures for Trials with 1% Inner Ring Radius

Portland Neighborhood	Trial: Outer Ring Radius (mi)	Persons per Acre (½-mi buffer) (%)	Jobs per Acre (½-mi buffer) (%)	Jobs per Person (½-mi buffer) (%)	Intersections per Acre (½-mi buffer) (%)	Distance to Nearest Park (ft) (%)	Distance to Nearest Grocery (ft) (%)	Distance to Nearest Rail Station (ft) (%)	Distance to Nearest Bus Stop (ft) (%)
Milwaukie	Trial 2: 0.5	58	**67**	110	48	161	44	9	46
	Trial 4: 1.0	60	**60**	112	51	350	68	17	49
Clackamas Transit Center	Trial 2: 0.5	33	32	40	18	170	37	**59**	**106**
	Trial 4: 1.0	60	60	112	51	350	68	**17**	**49**
Wilsonville	Trial 2: 0.5	52	217	374	42	1,104	17	21	13
	Trial 4: 1.0	82	493	2,045	66	1,801	36	41	22
Kenilworth	Trial 2: 0.5	26	37	65	12	**71**	162	15	**125**
	Trial 4: 1.0	30	43	75	25	**63**	198	28	**115**
East Overlook	Trial 2: 0.5	8	29	26	12	48	67	65	49
	Trial 4: 1.0	20	48	111	22	53	107	130	58
Northwest 21st and 23rd Avenue	Trial 2: 0.5	24	43	42	20	1,703	177	61	427
	Trial 4: 1.0	58	105	208	32	1,939	336	90	538
Portland State University	Trial 2: 0.5	71	**101**	86	21	228	29	262	**134**
	Trial 4: 1.0	101	**88**	101	42	361	54	461	**116**
North Beaverton	Trial 2: 0.5	25	**791**	**1,435**	14	60	35	39	57
	Trial 4: 1.0	28	**704**	**1,286**	21	84	39	68	58
Raleigh West	Trial 2: 0.5	59	38	52	13	58	72	61	149
	Trial 4: 1.0	73	51	71	41	88	97	101	340
Sherwood	Trial 2: 0.5	98	345	470	101	108	26	6	31
	Trial 4: 1.0	127	2,032	782	113	163	44	10	51

NOTE: Boldface = instances of greater percent RMSE for respective built environment measure in trial with 1-mi outer ring radius when compared with trial in ½-mi outer ring radius.

TABLE 4 ***K*** **Anonymity Statistics for Four Simulation Trials**

	Trial Number and Description				
Portland Neighborhood	*K* Anonymity Statistic Measure	Trial 1: ½% Inner Ring, ½-mi Outer Ring	Trial 2: 1% Inner Ring, ½-mi Outer Ring	Trial 3: ½% Inner Ring, 1-mi Outer Ring	Trial 4: 1% Inner Ring, 1-mi Outer Ring
Milwaukie	Minimum	33	29	15	38
	Median	592	496	2,435	2,035
	Maximum	986	960	3,932	3,941
Clackamas Transit Center	Minimum	19	37	86	158
	Median	1,065	1,070	4,339	4,466
	Maximum	2,221	2,199	8,916	8,892
Wilsonville	Minimum	14	16	20	18
	Median	92	104	365	359
	Maximum	185	187	742	733
Kenilworth	Minimum	15	48	41	25
	Median	1,980	1,954	6,783	8,434
	Maximum	4,067	4,018	16,280	16,073
East Overlook	Minimum	10	23	31	107
	Median	1,460	1,637	5,160	5,081
	Maximum	2,745	2,699	10,488	10,986
Northwest 21st and 23rd Avenue	Minimum	41	143	415	1,191
	Median	7,307	7,750	27,375	28,698
	Maximum	13,477	13,176	53,499	54,010
Portland State University	Minimum	74	90	84	145
	Median	1,725	1,739	7,365	6,984
	Maximum	3,710	3,691	14,573	14,764
North Beaverton	Minimum	69	41	88	47
	Median	726	818	2,671	2,736
	Maximum	1,375	1,384	5,447	5,420
Raleigh West	Minimum	13	39	256	275
	Median	1,293	1,126	5,173	5,725
	Maximum	2,454	2,442	9,793	9,799
Sherwood	Minimum	13	27	18	33
	Median	37	45	148	127
	Maximum	62	62	251	250

Results of this analysis pointed to an overall trend in which the minimum estimated *k* anonymity increased for a majority of households when the inner ring radius was extended while preserving an identical outer ring radius. This finding supported the hypothesis that a larger inner ring radius would decrease the potential for a first-tiered identity disclosure. Yet, there were five instances in which the increase in the inner ring radius from a distance reflective of 0.5% of occupied households in the neighborhood to a distance with an estimated *k* anonymity of 1% of occupied households resulted in a lower minimum estimated *k* anonymity statistic. The minimum estimated that *k* anonymity statistic for the Milwaukie household was greater in the first trial defined by the shorter inner ring, as was the case for the household in the North Beaverton neighborhood where the minimum estimated *k* anonymity statistic was 69 for the 0.5% inner ring and 41 for the 1% inner ring tested in the second trial. The reporting of a minimum estimated *k* anonymity statistic highlights the debate concerning the decision of an acceptable level of disclosure risk. For example, the minimum estimated *k* anonymity statistic of the 100 simulations related to the first trial for the East Overlook household indicated that there were only 10 households a data intruder must choose between to make a one-to-one correspondence between the original and geographically perturbed household.

A comparison of the minimum estimated *k* anonymity statistic for the first pair of trials with the second pair of trials shows that an increase in the outer ring radius led to an increase in the minimum estimated disclosure risk. Intuitively, one may hypothesize that an increase in the outer ring radius when the inner ring radius is preserved would increase the median estimated *k* anonymity statistic calculated for each household during 100 simulations. This finding is attributable to the indirect association between the chosen inner ring radii and selected outer ring radii in this four-trial proof of concept.

DISCUSSION OF RESULTS

Central to this empirical application has been the exploration of an appropriate balance concerning a minimization of disclosure risk, estimated by *k* anonymity, and maximization of data utility, measured by variation in the built environment introduced by geographically perturbing a household location. Application of the doughnut masking technique has shown promise as an approach in negotiating this balance, which has been largely overlooked by the travel survey community. However, as is often the case in proof of concept contributions, this study has also raised questions to be addressed by future applications.

In relation to the minimization of disclosure risk, the concept of *k* anonymity has gained traction in public health research although its operationalization has been diverse. The estimated *k* anonymity statistic provides the data custodian with an initial idea of the confidentiality provided by the chosen inner ring radius, yet there is importance in realizing that this statistic will most likely vary from household to household in the perturbed data set (*21*). Also, *k* anonymity, in this application, represents an estimate rather than actual count of households that protect a respondent from a first-tier identity disclosure. Moreover, estimated *k* anonymity is a density-based measure, which emphasizes the differences inherent in preventing a first-tiered disclosure breach in an urban context versus a setting marked by a lower residential density. To illustrate this point, a student household in the urban neighborhood of Portland State University (PSU) may not need to be relocated as much as a household in a suburban context would need to be relocated because there may be plenty of households with similar attributes located near their multilevel dormitory unit to mask the household's unique characteristics.

Also, trial selection in this application was performed with the knowledge that the inner ring linked to reducing disclosure risk would completely reside in the data utility–specific outer ring. As such, attention must be given to a scenario in which a household in a low-density residential neighborhood requires a threshold of anonymity that extends the disclosure risk-specific inner ring radius beyond the 1-mi radius related to a common walk trip distance that was selected to limit built environment variation. This scenario underscores a discussion on the selection of an appropriate outer ring radius for a neighborhood with a lower residential density. Because households in more suburban neighborhoods must be repositioned a greater minimum distance than their urban counterparts to preserve survey participant anonymity, consideration should be given to the possibility of relating the doughnut's outer ring radius to a specific built environment measure. For instance, a household in the more suburban context such as Sherwood may have its outer ring radius linked to employment density because this measure exhibited the greatest variability for this household in the third trial. Meanwhile, the household in the PSU neighborhood may have its outer ring radius linked to the rail distance measure because the spatial error related to this transit accessibility variable is greater than other measures in this third trial.

Also, although the built environment measures used in this research spanned the 5D classification scheme, there are countless measures to explore in future applications of the doughnut masking technique such as additional measures of land use mix or distance to transit (*28*). The variables specific to this study highlight the importance of disaggregate measures to an improved understanding of nonmotorized travel. Accordingly, disaggregate measures should be used exclusively in any future application of the doughnut masking technique. The use of these measures will address ecological fallacy concerns that have clouded past zonal-based travel demand models.

Moreover, the application of the doughnut masking technique in this study has been based solely on the reduction of first-tier disclosure risk. Although the importance of lessening the potential for this highest breach in respondent confidentiality is well understood, steps must also be taken to reduce the likelihood of a second- or third-tier disclosure (*19*). One additional step to decrease the possibility of a lower-tiered disclosure breach for a travel survey respondent may be to geographically perturb the activity locations of household members (e.g., workplace). Information on a household member's workplace may be the additional piece of knowledge a data intruder needs to properly identify a household location. For instance, a household member may have an occupation supported only by one regional employer (e.g., university professor). Similarly, a household member may shop at a grocery store that is atypical from the shopping destination of other households in the neighborhood. In such circumstances, a data custodian may perturb the activity locations of the household member in addition to the residence. Furthermore, the data custodian may give consideration to the geographic offsetting of travel routes, which may necessitate other perturbation methods or spatial analyses aside from measuring built environment variation to calculate appropriate displacement bounds.

These illustrations reflect challenges in using the doughnut masking technique as a postprocessing strategy. As the manner in which researchers balance disclosure risk and data utility evolves, attention should be given to the possibility of integrating the doughnut masking technique during survey design. Specifically, by establishing algorithms for geographically perturbing household locations within the framework of an automated process, the doughnut masking of a survey respondent may be accomplished during the data collection process. In this survey design, a household susceptible to a first-tier risk may be immediately detected by the algorithm in the automated process as a candidate for geographic perturbation when recorded by the survey administrator and, subsequently, its location may be instantaneously perturbed. Thus, future thought should be given to the necessity of geographically perturbing disaggregate household data when the sample design for a travel survey is devised, because doing so would most likely quicken the ability of agencies to release these valuable data to the public by reducing postprocessing time while upholding the agreed confidentiality pledge.

Ultimately, an improved understanding of the spatial error and respondent anonymity trade-off will better inform the selection of an appropriate survey design when household travel survey data are collected for a wider dissemination. This proof of concept examining the application of the doughnut masking technique to household travel survey data has highlighted one promising approach for travel demand researchers to consider and tabled a greater discussion concerning the proper balance between survey data utility and respondent confidentiality faced by a data hungry profession.

REFERENCES

1. Handy, S., M. Boarnet, R. Ewing, and R. Killingsworth. How the Built Environment Affects Physical Activity: Views from Urban Planning. *American Journal of Preventive Medicine,* Vol. 23, 2S, 2002, pp. 64–73.
2. Goulias, K.G., and T.-G. Kim. Multilevel Analysis of Activity and Travel Patterns: Accounting for Person- and Household-Specific Observed and Unobserved Effects Simultaneously. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1752,* TRB, National Research Council, Washington, D.C., 2001, pp. 23–31.
3. Bhat, C., and H. Zhao. The Spatial Analysis of Activity Stop Generation. *Transportation Research Part B,* Vol. 36, 2002, pp. 557–575.
4. Chikaraishi, M., A. Fujiwara, J. Zhang, and K.W. Axhausen. Exploring Variation Properties of Departure Time Choice Behavior by Using Multilevel Analysis Approach. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2134,* Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 10–20.
5. Badoe, D., and E. Miller. Transportation–Land Use Interaction: Empirical Findings in North America, and Their Implications for Modeling. *Transportation Research Part D,* Vol. 5, 2000, pp. 235–263.
6. Davidson, W., R. Donnelly, P. Vovsha, J. Freedman, S. Ruegg, J. Hicks, J. Castiglione, and R. Picado. Synthesis of First Practices and Operational Research Approaches in Activity-Based Travel Demand Modeling. *Transportation Research Part A,* Vol. 41, 2007, pp. 464–488.
7. Rodriguez, D., and J. Joo. The Relationship between Non-Motorized Mode Choice and the Local Physical Environment. *Transportation Research Part D,* Vol. 9, 2004, pp. 151–173.

8. VanWey, L., R. Rindfuss, M. Gutmann, B. Entwisle, and D. Balk. Confidentiality and Spatially Explicit Data: Concerns and Challenges. *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 102, No. 43, 2005, pp. 15337–15342.
9. Armstrong, M., G. Rushton, and D. Zimmerman. Geographically Masking Health Data to Preserve Confidentiality. *Statistics in Medicine,* Vol. 18, 1999, pp. 497–525.
10. Rushton, G., M. Armstrong, J. Gittler, B. Greene, C. Pavlik, M. West, and D. Zimmerman. *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice.* CRC Press, Boca Raton, Fla., 2008.
11. Gutmann, M., K. Witkowski, C. Colyer, J. O'Rourke, and J. McNalley. Providing Spatial Data for Secondary Analysis: Issues and Current Practices Relating to Confidentiality. *Population Research and Policy Review,* Vol. 27, 2008, pp. 639–665.
12. Karr, A., C. Kohnen, A. Oganian, J. Reiter, and A. Sanil. A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. *The American Statistician,* Vol. 60, No. 3, 2006, pp. 224–232.
13. Curtis, A., J. Mills, L. Agustin, and M. Cockburn. Confidentiality Risks in Fine Scale Aggregations of Health Data. *Computers, Environment and Urban Studies,* Vol. 35, 2011, pp. 57–64.
14. Kitamura, R., C. Chen, R. Pendyala, and R. Narayanan. Micro-Simulation of Daily Activity Travel Patterns for Travel Demand Forecasting. *Transportation,* Vol. 27, 2000, pp. 25–51.
15. Chang, K., Z. Khatib, and Y. Ou. Effects of Zoning Structure and Network Detail on Traffic Demand Modeling. *Environment and Planning B,* Vol. 29, 2002, pp. 37–52.
16. Domingo-Ferrer, J., and V. Torra. Disclosure Risk Assessment in Statistical Data Protection. *Journal of Computational and Applied Mathematics,* Vol. 164–165, 2004, pp. 285–293.
17. Kwan, M., I. Casa, and B. Schmitz. Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks? *Cartographica,* Vol. 39, No. 2, 2004, pp. 15–28.
18. Clifton, K. J., and N. Noyan. Framework for Applying Data Masking and Geoperturbation Methods to Household Travel Survey Data Sets. Presented at 91st Annual Meeting of the Transportation Research Board, Washington, D.C., 2012.
19. Duncan, G., and D. Lambert. The Risk of Disclosure for Microdata. *Journal of Business and Economic Statistics,* Vol. 7, No. 2, 1989, pp. 207–217.
20. French, J., and M. Wand. Additive Models for Cancer Mapping with Incomplete Covariates. *Biostatistics,* Vol. 5, No. 2, 2004, pp. 171–191.
21. Cassa, C., S. Grannis, J. Overhage, and K. Mandl. A Context-Sensitive Approach to Anonymizing Spatial Surveillance Data: Impact on Outbreak Detection. *Journal of the American Medical Informatics Association,* Vol. 13, No. 2, 2006, pp. 160–165.
22. Hampton, K., M. Fitch, W. Allshouse, I. Doherty, D. Gesink, P. Leone, M. Serre, and W. Miller. Mapping Health Data: Improved Privacy Protection with Doughnut Method Geomasking. *American Journal of Epidemiology,* Vol. 172, No. 9, 2010, pp. 1062–1069.
23. Allshouse, W., M. Fitch, K. Hampton, D. Gesink, I. Doherty, P. Leone, M. Serre, and W. Miller. Geomasking Sensitive Health Data and Privacy Protection: An Evaluation Using an E911 Database. *Geocarto International,* Vol. 25, No. 6, 2010, pp. 443–452.
24. Clifton, K., A. Smith, and D. Rodriguez. The Development and Testing of an Audit for the Pedestrian Environment. *Landscape and Urban Planning,* Vol. 80, 2007, pp. 95–110.
25. Krizek, K. Operationalizing Neighborhood Accessibility for Land Use–Travel Behavior Research and Regional Modeling. *Journal of Planning Education and Research,* Vol. 22, 2003, pp. 270–287.
26. McGinn, A., K. Evenson, A. Herring, S. Huston, and D. Rodriguez. Exploring Associations between Physical Activity and Perceived and Objective Measures of the Built Environment. *Journal of Urban Health,* Vol. 84, No. 2, 2007, pp. 162–184.
27. Sweeney, L. *K* Anonymity: A Model for Predicting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* Vol. 10, No. 5, 2002, pp. 557–570.
28. Ewing, R., and R. Cervero. Travel and the Built Environment. *Journal of the American Planning Association,* Vol. 76, No. 3, 2010, pp. 265–294.

The Travel Survey Methods Committee peer-reviewed this paper.