Title

Policy and programmatic importance of spatial alignment of data sources.

Permalink
https://escholarship.org/uc/item/382768fw

Journal
American journal of public health, 96(3)

ISSN
0090-0036

Authors
Ong, Paul
Graham, Matthew
Houston, Douglas

Publication Date
2006-03-31

Peer reviewed

Policy and Programmatic Importance of Spatial Alignment of Data Sources

Geographic information systems have proven instrumental in assessing environmental impacts on individual and community health, but numerous methodological challenges are associated with analyses of highly localized phenomena in which spatially misaligned data are used.

In a case study based on child care facility and traffic data for the Los Angeles metropolitan area, we assessed the extent of facility misclassification with spatially unreconciled data from 3 different governmental agencies in an attempt to identify child care centers in which young children are at risk from high concentrations of toxic vehicle-exhaust pollutants. Relative to geographically corrected data, unreconciled information produced a modest bias in terms of aggregated number of facilities at risk and a substantial number of false positives and negatives. (Am J Public Health. 2006;96:499–504. doi:10.2105/AJPH.2005.071373)

Paul Ong, PhD, Matthew Graham, MA, and Douglas Houston, MA

GEOGRAPHIC INFORMATION systems (GIS) have proven instrumental in assessing environmental impacts on individual and community health.[1–13] Recent studies have begun to systematically address technological limitations associated with GIS by enhancing accuracy of positional, attribute, and temporal data; by tracking demographics and disease as geographic boundaries change over time; by identifying the best household and area measures of socioeconomic status; and by determining appropriate scales for studying links between environmental exposures and health outcomes.[14–25] Improved data and advances in techniques have enabled epidemiological and atmospheric researchers to apply GIS to highly localized problems, but such analyses present numerous methodological challenges, especially when data of different pedigrees are not collocated at small geographic scales.

We assessed the impact of using geographically unreconciled traffic volume data along with census-based street data in a case study of child care centers whose locations near major roadways could put young children at risk from high concentrations of toxic vehicle exhaust pollutants. Recent epidemiological evidence indicates a heightened prevalence of respiratory morbidity and mortality among people living near high-traffic roadways, and childhood cancer, brain cancer, leukemia, and preterm and low-weight births have been positively associated with traffic density among those living near such roadways.[26–31] Although other environmental risk factors may be present in high-traffic areas, air pollution studies point to the significance of high concentrations of vehicle-generated pollutants such as carbon monoxide and ultrafine particles. Typically, pollutants decline exponentially to near background levels within as little as 150 m of major roadways, with the greatest decrease occurring within 50 m.[32–34]

Because dispersed monitoring stations are insufficient to determine pollutant concentrations at nonadjacent locations, and given the expense of directly measuring pollutants at multiple sites, researchers conducting epidemiological and distributional studies have used traffic volume line data and census-based line data to approximate exposure to vehicle-related pollutants.[30,35–38] However, this method can result in exposure misclassifications if these data sets are not precisely "aligned" with each other in GIS analyses. (Such discrepancies are not uncommon in health-related research, especially when data from different sources are used. Detailed statistics on misalignments in this study are described later.) Such geographic misalignments can result from the underlying data source, data cleaning processes, or the original intended scale of the data. We examined the effects of reassigning attribute data from 1 geographic data source to census-based line data on estimated exposure levels of facilities geocoded via census-based line data. A more general issue not broached in this article is the question of "georeferencing," or determination of spatial accuracy relative to the earth.

Previous studies have addressed misalignment problems associated with geographic data sets in different ways. One approach is to increase buffer areas beyond the ideal criterion distance to avoid false negatives, but this method can produce false positives.[37,38] Wilhelm and Ritz addressed such misalignments by transferring traffic count values from the original traffic line geography to census-based line segments—a method similar to that described here—but only for a select set of neighborhoods.[30] Green et al.[35] and Houston et al.[37] did not correct misalignments but assumed that discrepancies between traffic volume and census-based geographies are randomly distributed and do not produce spatial biases.[35] Although the value of spatially aligned data was recognized in these studies, none of them included systematic comparisons of results from reconciled and unreconciled data sets.

We evaluated the impact of reassigning traffic counts to a census-based geography for Los Angeles County, which is home to 9.5 million people and covers approximately 12 300 km^2. Our evaluation took the form of a case study designed to identify licensed child care facilities close to major roadways with high traffic volumes. Assessments were made at both the policy level ("What is the prevalence of the problem?") and the programmatic level ("Which facilities are affected?"). Results suggested that use of reconciled data provided valuable methodological enhancements in terms of identification of "at-risk" centers.

METHODS

Data

Traffic volume data and associated street geography data were obtained from the California Department of Transportation (CalTrans). We gathered child care facility information from California's Department of Social Services, and we geocoded addresses to the Topologically Integrated Geographic Encoding and Referencing (TIGER) street file, a standard geographical reference widely used by public health researchers and social scientists. (We used CalTrans traffic volume data from various years depending on when traffic for each road segment was recorded. Child care facility and TIGER data were from 2000.) We transformed geographic data into a common geographic projection, Universal Transverse Mercator, so that we could construct geographic overlays and make consistent distance calculations. Two collective data sets were assembled: one overlaying the geocoded child care facilities on the original traffic data "as is"—without reconciling spatial misalignments—and one overlaying the facility points onto traffic data assigned to the common TIGER street file, thus eliminating spatial misalignments.

CalTrans Data

We report traffic counts for segments of CalTrans' roadway network, which had its origins in US Geological Service Digital Line Graph transportation data and has been used in previous studies focusing on traffic distribution and impacts.[27,30,35,37,38] CalTrans used a tolerance (or geographic error) of 10 m when transforming original US Geological Service coordinates into the roadway data used in this study.[39] The roadway network does not include local roads, but this exclusion has only limited effects on estimations of potential risk from mobile-source air pollutants, because local roads carry light traffic. With this restriction, the data set included 9230 mi (14 855 km) of roads divided into 37 403 segments. On the basis of seasonal fluctuations, weekly variations, and other variables, CalTrans adjusts counts to estimate annual average daily traffic (AADT), representing the total annual volume of vehicles divided by 365 days. Unfortunately, these data do not include adequate segment-level information to geocode address locations.

TIGER Data

The US Census Bureau's TIGER street file is readily available (at http://www.census.gov/geo/www/tiger/index.html), inexpensive, and consistent with census tabulation geographies (e.g., tracts). As a result, the file has been widely used by researchers, particularly in assigning geocoded information to census polygons (land areas, ranging in size from blocks and tracts to counties and states, defined by the US Census Bureau for the purpose of data collection or data reporting). This internal consistency between geocoded addresses and census polygons allows researchers to use socioeconomic contextual information derived from the decennial census. TIGER objects are not necessarily accurate in relationship to their global position, and disparities vary according to geographic location.[40] Moreover, there is no guarantee that addresses geocoded with TIGER streets are properly spatially referenced to the CalTrans road network, a problem reported in other studies.[25,30,36]

Child Care Facility Data

Data on licensed Los Angeles child care facilities, obtained from the California Department of Social Services, included the addresses of 3430 licensed child care centers and 3399 large family care homes (those licensed to care for more than 8 children). We geocoded facility addresses (approximate match rate: 95%) to TIGER streets using a 10-m offset from roadway center lines to provide standard coordinates. In some studies, the positional accuracy, or correspondence to "real-world" locations, of geocoded addresses can have significant analytic impacts; however, we were concerned only with the internal consistency of the relative alignments of GIS data sets to each other. Other statistical implications of geocoding (and the use of offsets) are well known but were not within the scope of this study.[19,21,25,41,42]

Data Reconciliation

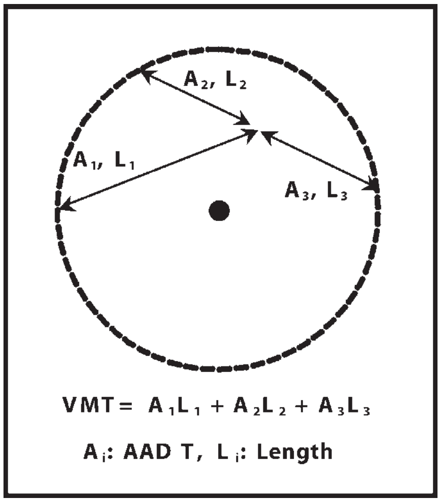
As mentioned earlier, the CalTrans network includes heavily traveled roadways and contains about one third of the TIGER network road segments (89 813 of 267 579 overall segments and 13 772 of a total of 43 401 km). In terms of segments included in both data sets, CalTrans street segments are generally not colinear with or well referenced to TIGER street segments. (An analysis of major thoroughfares [AADT >24 000] showed that the average geographic discrepancy between 2 street segments was 13.3 m, with a standard deviation of 19.5 m. Differences ranged from less than 1 m to more than 400 m.) One of the consequences of using the 2 layers "as is" is that child care facilities geocoded with TIGER geography are not accurately referenced to CalTrans segments.[35,36] For convenience, we refer to the collection of data assembled through overlaying the CalTrans road network and the child care data as the "unreconciled data set."

We created a "reconciled data set" by reassigning the traffic volume data from the CalTrans network to corresponding TIGER street segments. This allowed both the geocoded child care facilities and the traffic data to be spatially referenced to a common street layer. Processing consisted of multiple steps to reconcile these 2 layers. The first, automated step matched streets on the basis of proximity and associated name (or route number). Because of the automated nature of this process, secondary errors resulted from (1) miscoded street names, either in TIGER or in CalTrans; (2) large displacements between matching streets; and (3) similarly named segments being too close in proximity. Two more rounds of visual comparisons were performed to correct manually these mismatching errors. Corrections consisted of reviewing the 2 street layers in ArcMap 8.3 (ESRI, Redlands, Calif), identifying TIGER streets that had not yet been matched or identifying secondary problems, and manually correcting the linking table between CalTrans and TIGER. The time required for this processing was approximately 130 person-hours. A similar matching process has been described at greater length by Wu et al.[25]

Analytic Method

Test criteria. We assessed the impact of spatial discrepancies between CalTrans traffic data and reconciled TIGER data by comparing the risk status of child care facilities classified according to close proximity to heavy traffic. We determined risk status using 2 tests based on the different roadway networks. A facility is deemed to be "at risk" when the volume of traffic within a given criterion distance is above a certain threshold. However, operationalizing such criteria is problematic, because the existing literature does not provide a standard for classifying facilities. Nonetheless, previous studies have identified 2 key parameters: volume of traffic and distance to roadways. Volume of traffic and risk from pollutants have been shown to be positively correlated, as have proximity and risk. We developed 2 traffic volume thresholds and 3 distance categories to examine how different criteria interact with geographic discrepancies to affect risk classifications of child care facilities.

We used vehicle miles traveled (VMT)—defined as the product of AADT and the length of each street segment—as our measure of traffic volume. Aggregate VMT for a site is the sum of VMT for each segment within a circular buffer; in this study, buffers had a radius of 50 m, 100 m, or 150 m (Figure 1). Only portions of street segments falling within these buffers were used to aggregate VMT for each child care center. Two thresholds were used to classify whether a facility was in dangerously close proximity to heavy traffic: 4500 VMT as the moderate-risk threshold and 9000 VMT as the high-risk threshold. These thresholds were roughly equivalent to the maximum AADT thresholds of 25 000 and 50 000 used in other studies.[26,31,35,36] We used the following formula to classify the risk status of a given (here *i*th) facility:



Note. AADT = annual average daily traffic.

FIGURE 1—Calculating total vehicle miles traveled (VMT).

$$(1)\quad \text{At Risk}_{i,j,k,m} = \begin{cases} 0, \ \text{VMT}_{i,j,k} < \text{Threshold}_m \\ 1, \ \text{VMT}_{i,j,k} \geq \text{Threshold}_m \end{cases}$$

$$\text{Concentric Circle: } j \in \begin{Bmatrix} 0\text{-}150\text{m;} \\ 0\text{-}100\text{m;} \\ 0\text{-}50\text{m} \end{Bmatrix}$$

$$\text{Data Set: } k \in \begin{Bmatrix} \text{reconciled;} \\ \text{unreconciled} \end{Bmatrix}$$

$$\text{Threshold: } m \in \begin{Bmatrix} 4500 \text{ VMT;} \\ 9000 \text{ VMT} \end{Bmatrix}$$

Two tests, one based on CalTrans traffic data and one based on reconciled TIGER data, were used to classify facility risk for each combination of threshold and buffer distance. We expected the number of facilities classified as at risk to decrease with increasing threshold and with decreasing buffer distance.

Spatial misalignments. Spatial misalignments between CalTrans and TIGER can produce 2 opposite classification errors: a positive classification by the CalTrans test but a negative classification by TIGER, as well as the converse case. Of course, a facility might be correctly classified even when the 2 networks are misaligned if both CalTrans and TIGER tests measure traffic volume either above or below the criterion threshold. We defined 4 classification types according to the scenarios described and the ability of the unreconciled CalTrans data to correctly classify child care facilities at risk: *consistent positive*, *false positive*, *false negative*, and *consistent negative*. These classifications are described in Table 1.

We evaluated the effects of misclassification resulting from the use of unreconciled data in 2 ways. For policy considerations, prevalence of risk is central, so the key statistic is the ratio of at-risk facilities to all facilities. Both the reconciled and unreconciled data sets involved the same denominator (all geocoded child care facilities) but different numerators (facilities classified as at risk according to each test). If spatial misalignments were randomly distributed with an expected offset distance of zero, false positives and false negatives would net out, and use of unreconciled data would not affect calculated prevalences. Without evidence to support such an assumption,

TABLE 1—Classifications of At-Risk Status of Child Care Facilities: Los Angeles Metropolitan Area, 2000

Interpretation	Aggregate VMT: Unreconciled Data	Aggregate VMT: Reconciled Data
Consistent positive: at-risk facilities correctly classified by both data sets	Above threshold	Above threshold
False positive: not-at-risk facilities misclassified as at risk by unreconciled traffic network	Above threshold	Below threshold
False negative: at-risk facilities misclassified as not at risk by unreconciled traffic network	Below threshold	Above threshold
Consistent negative: not-at-risk facilities correctly classified by both data sets	Below threshold	Below threshold

Note. VMT = vehicle miles traveled.

TABLE 2—At-Risk Childcare Facilities, by Threshold and Area of Interest (n = 6829: Los Angeles Metropolian Area, 2000

	Above Moderate Threshold[a]			Above High Threshold[b]		
Radius of Circle, m	Reconciled, No.	Unreconciled, No.	Relative Overcount, No. (%)	Reconciled, No.	Unreconciled, No.	Relative Overcount, No. (%)
150	2202	2223	21 (0.95)	564	591	27 (4.79)
100	653	668	15 (2.3)	145	148	3 (2.07)
50	32	29	−3 (−9.38)	10	13	3 (30.0)

[a] 4500 vehicle miles traveled.
[b] 9000 vehicle miles traveled.

however, the impact of misalignments must be determined empirically.

From a programmatic perspective, the absolute and relative numbers of false positives and false negatives are the important statistics. Limited intervention budgets necessitate identifying the facilities truly at risk, and an excessive number of false positives would divert resources from programmatic objectives. Alternatively, false negatives would impose intervention costs on sites not actually at risk. Unlike aggregating data to estimate prevalence of risk for policy considerations, the costs associated with false positives and false negatives are cumulative rather than offsetting. Consequently, use of unreconciled data may have greater effects at the programmatic level than at the policy level.

RESULTS

The results of the analysis of risk classification and prevalence rates are presented in Table 2. As expected, the number of at-risk facilities decreased when traffic thresholds were higher and buffer distances were smaller, regardless of whether reconciled or unreconciled roadway data were used. It should be noted that the relationship between risk classification and buffer area (αr^2) was nonlinear, with the aggregate number of at-risk facilities increasing at a greater rate than the buffer area. Also, doubling the VMT threshold from moderate (4500 VMT) to high (9000 VMT) dramatically reduced the number of at-risk facilities in most cases. Although these patterns are not central to our focus, they strongly suggest that classification criteria have a significant impact on the reported magnitude of the risk associated with close proximity to high levels of traffic.

Policy Concerns

More germane to our focus is a comparison between the relative and aggregate number of at-risk facilities calculated via the 2 street geographies. For most permutations of threshold and buffer size, the reconciled data identified fewer at-risk sites than did the unreconciled data, the exception being the moderate threshold at 50 m. As a percentage of facilities identified by reconciled data, the relative number of discrepancies tended to decrease with increasing buffer size; here the exception was the high threshold for 100 m. This pattern suggests that the offset error between the reconciled and unreconciled data sets was not randomly distributed; otherwise, the relative discrepancies (which ranged from 1% to 30%) would have been close to zero in all cases.

The most restrictive buffer (50 m) tended to result in the highest relative errors (an undercount of 9% with the moderate threshold and an overcount of 30% with the high threshold), but such a small buffer is useful only in identifying facilities in immediate proximity to the largest pollution sources (freeways and major thoroughfares). Larger buffers tend to capture more of the at-risk facilities from moderately large or multiple sources. Also, for these larger buffers, the reconciled data showed improvement over the unreconciled data, although this improvement was not as large as that associated with the 50-m buffer.

In policy terms, these findings demonstrate that errors due to spatial misalignment are not randomly distributed; consequently, estimation of the number of child care facilities at risk is biased when unreconciled data are used at the specified buffer distances. With a 150-m buffer, the distance at which vehicle-related air pollutants drop close to "background" concentrations, the bias was small, and the unreconciled data may have been sufficiently accurate to gauge the overall magnitude of the problem. This buffer size, however, identifies facilities that are at a low level of exposure. The finding of the nonlinear increase in the count with greater criterion distance indicates that padding the buffer area to account for spatial misalignment could artificially produce a substantial overcount.

Programmatic Concerns

The problem of discrepancies in classification and the improvements shown with reconciled data are even greater when we examine the number of false positives and false negatives, which are reported in Table 3 along with consistent positives. Absolute values are presented in the top panel, and the bottom panel reports classification rates as percentages of the total number of at-risk facilities for each combination of threshold and radius identified in the reconciled data set.

For the least restrictive risk criteria (150 m, moderate threshold), 12.5% (5.8% as false negative and 6.7% as false positive) of 2202 at-risk facilities (identified through reconciled data) were misclassified when unreconciled data were used. For the 100-m buffer, the resulting misclassifications were 25% with the moderate threshold and 48% with the high threshold. For the smallest buffer, the corresponding misclassification rates were 97% and 110%. Overall, the absolute number of misclassified facilities increased with increasing circle size; however, the misclassification rate (as a percentage of correctly identified at-risk facilities) generally *decreased* as circle size increased. Essentially, although the unreconciled data involved fewer total misses as

TABLE 3—Classification of At-Risk Status of Child Care Facilities: Los Angeles Metropolitan Area, 2000

	Above Moderate Threshold			Above High Threshold		
Radius of Circle, m	Consistent Positive	False Negative	False Positive	Consistent Positive	False Negative	False Positive
Absolute no.						
150	2075	127	148	512	52	79
100	578	75	90	112	33	36
50	15	17	14	6	4	7
Classification rate, %[a]						
150	94.2	5.8	6.7	90.8	9.2	14.0
100	88.5	11.5	13.8	77.2	22.8	24.8
50	46.9	53.1	43.8	60.0	40.0	70.0

[a]Percentage of total facilities classified as at risk according to the reconciled data set for each threshold and radius.

buffer size decreased, the odds of facilities being incorrectly identified increased.

The results for the disaggregated statistics on classification type revealed that spatial misalignment creates significant discrepancies in risk classification. The problem becomes increasingly severe as criteria become more restrictive. This trend is capped by the case of the high threshold at 50 m, for which the unreconciled data incorrectly identified more sites than they correctly identified. Such instances in which a given list of at-risk sites can be trusted less than half of the time are a significant problem from a programmatic perspective. Even in cases in which the problem of discrepancies is not as large, they still can represent an economically inefficient increase in allocating scarce funds, weakening the effectiveness of targeting limited resources. In fact, considering the case of the moderate threshold for 150 m (a distance at which pollutants are still considered a risk), in which "only" 12.5% of facilities were misclassified, 275 child care sites were either missed or incorrectly identified as at risk, a significant number of locations when limited resources for enforcement and remediation are available. Clearly, the reconciled data set makes its greatest contribution at the programmatic level.

DISCUSSION

With the development of less costly and more user-friendly software, decreased costs of computing power, and increased availability of geographically coded data, GIS is proving to be a useful tool in studying the potential health effects of spatially localized environmental hazards. Such trends will continue into the future, encouraging and facilitating more spatially oriented analyses. An important task for improving the usefulness of this approach is identifying problems associated with merging data. One of the strengths of the technology is its capacity to overlay and analyze information from disparate sources, but this is also a potential weakness in that there is no assurance that data sets are properly aligned. This problem is not new to the broader GIS field, but it is worth exploring in the context of public health research.

Our findings reveal some serious discrepancies when data sets are not spatially aligned and suggest that use of unreconciled data has policy and programmatic implications. Of course, it is impossible to say whether our results can be generalized to other data sets and spatial analyses. Moreover, the Los Angeles metropolitan area may have unique land-use and siting rules that affect the number of at-risk child care facilities. Despite these limitations, the discrepancies between the facilities identified by the reconciled and unreconciled data sets are sufficiently large that our findings should raise a red flag for all public health researchers using GIS.

Explicitly, we found that by reconciling traffic volume data with other geographical data on siting of child care centers in Los Angeles County, we could improve estimations of the centers at risk from mobile source pollutants. This data reconciliation altered the results from both the policy and the programmatic perspective. In terms of policy, we found that the actual numbers of sites considered at risk according to our measures were marginally lower than revealed in a similar analysis involving unreconciled data. From a programmatic perspective, we found that using unreconciled data produced a dramatic miscount of those sites incorrectly classified as at risk as well as those misclassified as safe.

We have several recommendations. At a minimum, researchers should assess how well various GIS data sets are spatially referenced to each other. This assessment would include evaluating data from the same agency but for different time periods. Even TIGER geography changes with time as errors are identified and corrected. If data sets are not corrected, it is important to determine whether they should be geographically reconciled. Unfortunately, there is no simple rule for determining when it is necessary to absorb the costly task of eliminating spatiotemporal discrepancies. This issue must be considered on a case-by-case basis. (As one reviewer noted, a potentially important empirical question with policy implications is whether the magnitude of the spatial discrepancy between 2 data sets varies systematically across neighborhoods according to socioeconomic status. If such a pattern exists, any analysis of socioeconomic status disparities in exposure to air pollutants involving unreconciled data would produce systematically biased results. The direction of the bias is an empirical issue that requires additional research.)

Our findings do point to 1 guideline: scale matters. The more localized the effect ("pollution footprint"), the more likely it is that an analysis will benefit from such reconciliation. Regardless of the decision, it is important for researchers to explicitly discuss spatial referencing issues related to the data sets they are using, which will provide readers with a sense of any potential limitations of the findings produced. Although it is important for individual researchers to seriously consider these issues, there is also a need for the field as a whole to develop and adopt standards for geographical data. High-quality GIS data represent a collective good that would enhance future public health research. ■

About the Authors

Paul Ong is with the School of Public Affairs and the Ralph and Goldy Lewis Center for Regional Policy Studies, University of California, Los Angeles. Matthew Graham is with Abt Associates, Cambridge, Mass. Douglas Houston is a doctoral student in urban planning at the University of California, Los Angeles.

Requests for reprints should be sent to Paul Ong, PhD, UCLA School of Public Affairs, 405 Hilgard Ave, Los Angeles, CA 90095 (e-mail: pmong@ucla.edu).

This article was accepted September 24, 2005.

Contributors

P. Ong originated and designed the study, directed the data analysis, and contributed to the interpretation of the data and the writing of the article. M. Graham performed the data analysis and contributed to interpretation of the data and to the writing of the article. D. Houston contributed to the study design, the data and interpretation, and the writing of the article.

Acknowledgments

We wish to thank Jun Wu and 3 anonymous reviewers for their helpful suggestions and the Ralph and Goldy Lewis Center for Regional Policy Studies at the University of California, Los Angeles, which supported this research.

References

1. Acevedo-Garcia D. Zip code-level risk factors for tuberculosis: neighborhood environment and residential segregation in New Jersey, 1985–1992. *Am J Public Health.* 2001;91:734–741.

2. Arno PS, Gourevitch MN, Drucker E, et al. Analysis of a population-based *Pneumocystis carinii* pneumonia index as an outcome measure of access and quality of care for the treatment of HIV disease. *Am J Public Health.* 2002;92:395–398.

3. Cervero R, Duncan M. Walking, bicycling, and urban landscapes: evidence from the San Francisco Bay Area. *Am J Public Health.* 2003;93:1478–1483.

4. Cohen D, Spear S, Scribner R, Kissinger P, Mason K, Wildgen J. "Broken windows" and the risk of gonorrhea. *Am J Public Health.* 2000;90:230–236.

5. Curriero FC, Patz JA, Rose JB, Lele S. The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948–1994. *Am J Public Health.* 2001;91:1194–1199.

6. Greenberg M, Mayer H, Miller KT, Hordon R, Knee D. Reestablishing public health and land use planning to protect public water supplies. *Am J Public Health.* 2003;93:1522–1526.

7. Holcomb CA, Lin M-C. Geographic variation in the prevalence of macular disease among elderly Medicare beneficiaries in Kansas. *Am J Public Health.* 2005;95:75–77.

8. James RC, Mustard CA. Geographic location of commercial plasma donation clinics in the United States, 1980–1995. *Am J Public Health.* 2004;94:1224–1229.

9. Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures—the Public Health Disparities Geocoding Project. *Am J Public Health.* 2003;93:1655–1671.

10. Lee RE, Cubbin C. Neighborhood context and youth cardiovascular health behaviors. *Am J Public Health.* 2002;92:428–436.

11. Maantay J. Zoning, equity, and public health. *Am J Public Health.* 2001;91:1033–1041.

12. Oyana TJ, Rogerson P, Lwebuga-Mukasa JS. Geographic clustering of adult asthma hospitalization and residential exposure to pollution at a United States–Canada border crossing. *Am J Public Health.* 2004;94:1250–1257.

13. Pearl M, Braveman P, Abrams B. The relationship of neighborhood socioeconomic characteristics to birthweight among 5 ethnic groups in California. *Am J Public Health.* 2001;91:1808–1814.

14. Ali M, Park J-K, Thiem V, Canh D, Emch M, Clemens J. Neighborhood size and local geographic variation of health and social determinants. *Int J Health Geogr.* 2005;4:12.

15. Carretta HJ, Mick SS. Geocoding public health data [letter]. *Am J Public Health.* 2003;93:699.

16. Diez Roux AV. Investigating neighborhood and area effects on health. *Am J Public Health.* 2001;91:1783–1789.

17. Dudley G. Scale, aggregation, and the modifiable area unit problem. *Operational Geographer.* 1991;9:28–33.

18. Elliott P, Wartenberg D. Spatial epidemiology: current approaches and future challenges. *Environ Health Perspect.* 2004;112:998–1006.

19. Krieger N, Waterman P, Chen JT, Soobader M-J, Subramanian SV, Carson R. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas—the Public Health Disparities Geocoding Project. *Am J Public Health.* 2002;92:1100–1102.

20. Krieger N, Waterman P, Chen JT, Soobader M-J, Subramanian SV, Carson R. Krieger et al. respond. *Am J Public Health.* 2003;93:699–700.

21. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *Am J Public Health.* 2001;91:1114–1116.

22. Nuckols JR, Ward MH, Jarup L. Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environ Health Perspect.* 2004;112:1007–1015.

23. Openshaw S, Taylor P. The modifiable area unit problem. In: Wrigley N, Bennett RJ, eds. *Quantitative Geography: A British View.* London, England: Routledge & Kegan Paul; 1981:60–69.

24. Soobader M, LeClere FB, Hadden W, Maury B. Using aggregate geographic data to proxy individual socioeconomic status: does size matter? *Am J Public Health.* 2001;91:632–636.

25. Wu J, Funk TH, Lurmann F, Winer A. Improving spatial accuracy of roadway networks and geocoded addresses. *Transactions in GIS.* In press.

26. Edwards J, Walters S, Griffiths RK. Hospital admissions for asthma in preschool children—relationship to major roads in Birmingham, United Kingdom. *Arch Environ Health.* 1994;49:223–227.

27. English P, Neutra R, Scalf R, Sullivan M, Waller L, Zhu L. Examining associations between childhood asthma and traffic flow using a geographic information system. *Environ Health Perspect.* 1999;107:761–767.

28. Garshick E, Laden F, Hart JE, Caron A. Residence near a major road and respiratory symptoms in US veterans. *Epidemiology.* 2003;14:728–736.

29. Pearson RL, Wachtel H, Ebi KL. Distance-weighted traffic density in proximity to a home is a risk factor for leukemia and other childhood cancers. *J Air Waste Manage Assoc.* 2000;50:175–180.

30. Wilhelm M, Ritz B. Residential proximity to traffic and adverse birth outcomes in Los Angeles County, California, 1994–1996. *Environ Health Perspect.* 2003;111:207–216.

31. Wjst M, Reitmeir P, Dold S, et al. Road traffic and adverse effects on respiratory health in children. *BMJ.* 1993;307:596–600.

32. Hitchins J, Morawska L, Wolff R, Gilbert D. Concentrations of submicrometre particles from vehicle emissions near a major road. *Atmospheric Environment.* 2000;34:51–59.

33. Zhu YF, Hinds WC, Kim S, Shen S, Sioutas C. Study of ultrafine particles near a major highway with heavy-duty diesel traffic. *Atmospheric Environment.* 2002;36:4323–4335.

34. Zhu YF, Hinds WC, Kim S, Sioutas C. Concentration and size distribution of ultrafine particles near a major highway. *J Air Waste Manage Assoc.* 2002;52:1032–1042.

35. Houston D, Ong PM, Wu J, Winer A. Proximity of licensed childcare to near-roadway vehicle pollution. *Am J Public Health.* In press.

36. Green RS, Smorodinsky S, Kim JJ, McLaughlin R, Ostro B. Proximity of California public schools to busy roads. *Environ Health Perspect.* 2004;112:61–66.

37. Gunier RB, Hertz A, Von Behren J, Reynolds P. Traffic density in California: socioeconomic and ethnic differences among potentially exposed children. *J Expo Anal Environ Epidemiol.* 2003;13:240–246.

38. Houston D, Wu J, Ong P, Winer A. Structural disparities of urban traffic in Southern California: implications for vehicle-related air pollution exposure in minority and high-poverty neighborhoods. *J Urban Aff.* 2004;26:565–592.

39. California Office of Geographic Information Systems. Functionally classified roads metadata. Available at: http://www.dot.ca.gov/hq/tsip/TSIPGSC/library/libdatalist.htm. Accessed March 11, 2005.

40. US Census Bureau. TIGER FAQ question 22. Available at: http://www.census.gov/cgi-bin/geo/tigerfaq?Q22. Accessed February 25, 2005.

41. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. *Int J Health Geogr.* 2003;2:10.

42. Whitsel EA, Rose KM, Wood JL, Henley AC, Liao DP, Heiss G. Accuracy and repeatability of commercial geocoding. *Am J Epidemiol.* 2004;160:1023–1029.