ELSEVIER

LETTER

Comments on Boone et al., "Validation of a GIS Facilities Database: Quantification and Implications of Error"

To the Editor:

Boone et al. (1) contribute to a much-needed critical view of Geographic Information Systems (GIS) databases commonly used in spatial epidemiology. Their analysis quantified count, attribute and positional error of a commercial database of physical activity facilities in two U.S. communities. Their findings indicate that the commercial database contained appreciable error, but they found no evidence to suggest the built environment–health association would be biased. I would like to comment on a number of methodological issues which may benefit future studies of a similar nature.

First, the authors are not the first to explore such a validation study and claims that "count and attribute errors in databases of community resources …. have not been assessed" are incorrect. Both Patterson et al. (2) and Zandbergen and Green (3) have characterized the errors associated with schools and were published before Boone et al. (1) was accepted. While Boone et al. (1) characterized a larger variety of facilities, schools represented about one third of their sample. Patterson et al. (2) compared two national data sets of schools and found that they both differed substantially from a local and field-validated data set in terms of attributes and location. Combining the two national data sets resulted in a substantial improvement. Zandbergen and Green (2007) characterized the positional error in school locations using multiple geocoding techniques and found substantial disagreement among the techniques as well as substantial errors in the placement of geocoded locations relative to pollution sources, resulting in significant bias in exposure misclassification at short distances. This suggests that any claims regarding a lack of bias should be examined very carefully.

Second, the authors provide a relatively incomplete characterization of the positional error of geocoded locations. It is firmly established that the positional error of street geocoding is not normally distributed, but more closely resembles a log-normal distribution (4–6). As a result, statistics such as mean and standard deviation are not very meaningful, in particular for the relatively low sample sizes employed by Boone et al (1). Instead, the use of statistics such as median or percentiles (75th, 90th, 95th) is recommended, or alternative representations that capture the error distribution more completely (e.g., cumulative distribution functions or quintiles). As a result, the claim by Boone et al. (1) that the "positional error…was larger than distances observed in residential geocoding validation studies" is not supported by the published data. The total sample size for this part of the analysis ($n = 105$) is also quite small to support this claim, in particular because this is further broken down into urban and non-urban facilities as well as into eight different types of facilities. A more complete characterization of the error distribution would have been more useful to support their claims in addition to serving as a more meaningful comparison for other validation studies.

Third, the authors determine agreement of administratively defined neighborhood and nearest streets for field validated (using Global Positioning System [GPS]) and geocoded locations. Census units were used as neighborhoods and both field validated and geocoded locations were spatially matched to "U.S. census geography shapefiles (2005)". This form of point-in-polygon overlay technique is widely employed but is not appropriate to assign census enumeration unit codes (7, 8). While the analysis serves the purpose of demonstrating the effect of positional error on spatial overlay analysis, the use of census geographies is misleading. The U.S. Bureau of the Census discourages this approach to assign census enumeration unit codes (8) and the "gold standard" in this case should be the use of look-up tables provided by the U.S. Bureau of the Census (7). For the analysis of roads, both field-validated and geocoded locations were spatially matched to "ESRI StreetMap USA (2004)" data. These data represent an enhanced version of the U.S. Bureau of the Census TIGER 2000 files, but are of similarly poor positional accuracy. Given the effort involved in collecting differentially corrected GPS locations, the use of such a low-quality road network is surprising and reduces the robustness of the analysis. The "gold standard" in this case should have been a data set of street centerlines from local authorities or similar high quality reference data. The spatial matching of geocoded locations to the nearest street segment is also confounded by the fact that no end off-set was employed in street geocoding. Facilities at the start or end of a specific address range could easily be snapped to the intersecting street segment instead of the street segment along which the locations were geocoded.

Fourth, the criterion measure for the location of facilities is the "obtained curbside GPS" reading. As the authors themselves acknowledge, some of the facilities are quite large and it is debatable what the criterion location should be. Several other studies on geocoding quality using relatively small sample sizes have employed digital orthophotography to identify building locations (2–4, 9–11) and it is

360 Park Avenue South, New York, NY 10010

1047-2797/08/$–see front matter
doi:10.1016/j.annepidem.2008.06.003

AEP Vol. 18, No. 10
October 2008: 823–824

unclear why the curbside location would be more meaningful. Presumably, for the analysis of spatial accessibility, the location of the entrance to the facility may be of most interest, but this will vary with the specific objective of the subsequent analysis. In any case, I question the claim that the facilities were mapped to within an accuracy of 1 to 2 meters because: (a) no evidence is presented to support this statement (other than a reference to the equipment used); (b) the statement lacks an accuracy metric (e.g., 68th percentile or Root Mean Square Error [RMSE]); and (c) the criterion location is ambiguously defined.

Finally, the authors claim that the errors found are not likely to result in systematic misclassification, but these claims are not supported by the findings. Most types of errors characterized were larger in non-urban areas relative to urban areas and the commercial database consistently underestimated facility counts relative to field census counts. Both these results warrant a closer examination of potential bias introduced in subsequent analyses. Several other potential sources of error and bias, including the direction of positional error or selection bias in geocoding match rates by type of facility or by neighborhood, were not investigated. Sample sizes were also too small to test differences in error between different types of facilities. Most importantly, no specific scenario was analyzed which would allow for an evaluation of potential biases. Presumably, the database will be used to generate individual-level metrics of access to physical activity facilities. However, without outlining the specifics of this analysis and carrying out some form of error propagation study, the magnitude of the effects of the errors and the presence of any potential biases from using the commercial database are unknown.

Any study employing spatial data in a GIS environment should determine the "fitness for use" of the data prior to any spatial analysis. The publication of Boone et al. (1) is among a handful of studies in spatial epidemiology that attempt to do this in a consistent manner, but the claim that "bias to associations with health-related outcomes is probably small and toward the null" is not supported by the analysis results presented to date.

Methodological limitations aside, Boone et al. (1) present a meaningful addition to the literature on the quality of spatial databases of non-residential locations, but future efforts in this area should consider refinements to the methodology. As the scale of spatial epidemiological analysis becomes finer, the demands on the accuracy of spatial databases and on the robustness of spatial data processing and analysis become greater. While the availability of commercial databases and the ease of use of GIS analysis tools present great opportunities for spatial epidemiology, the burden remains on the researcher to ensure such databases and tools are sufficiently robust to provide meaningful insights into the relationships between environmental factors and health-related outcomes.

P.A. Zandbergen, PhD
Department of Geography
University of New Mexico
Albuquerque, NM

REFERENCES

1. Boone JE, Gordon-Larsen P, Stewart JD, Popkin BM. Validation of a GIS facilities database: quantification and implications of error. Ann Epidemiol. 2008;18:371–377.
2. Patterson L, Urban M, Myers A, Bhaduri B, Bright E, Coleman P. Assessing spatial and attribute errors in large national datasets for population distribution models: a case study of Philadelphia county schools. GeoJournal. 2007;69:93–102.
3. Zandbergen PA, Green JW. Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. Environ Health Perspect. 2007;115:1363–1370.
4. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. Int J Health Geogr. 2003;2:10.
5. Karimi HA, Durcik M. Evaluation of uncertainties associated with geocoding techniques. Computer-Aided Civil Infrastructure Engineering. 2004; 19:170–185.
6. Zandbergen PA. Positional accuracy of spatial data: non-normal distributions and a critique of the National Standard for Spatial Data Accuracy. Transactions in GIS. 2008;12(1):103–130.
7. Rushton G, Armstrong MP, Gittler J, Greene B, Pavlik CE, West MW, et al. Geocoding in cancer research: a review. Am J Prev Med. 2006; 30(2S):S16–S24.
8. U.S. Bureau of the Census. Cartographic boundary files. Washington (DC): U.S. Bureau of the Census, Geography Division; 2004.
9. Schootman M, Sterling DA, Struthers J, Yan Y, Laboube T, Emo B, et al. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. Ann Epidemiol. 2007;17:464–470.
10. Strickland MJ, Siffel C, Gardner BR, Berzen AK, Correa A. Quantifying geocode location error using GIS methods. Environ Health. 2007;6:10.
11. Zimmerman DL, Fang X, Mazumdar S, Rushton G. Modeling the probability distribution of positional errors incurred by residential address geocoding. Int J Health Geogr. 2007;6:1.