## *Extended Abstract*

### Monte Carlo simulation-based recommendations for reducing the risk of bias in multilevel models introduced by mis-measuring the neighborhood

Claudia Nau, PhD
Johns Hopkins School of Public Health
cnau@jhpsh.edu

**Abstract (150 words)**

This paper develops recommendations on how to minimize the risk and size of bias in multilevel models if the scale of the neighborhood effect is unknown or data is not available to model the neighborhood-scale properly. These recommendations are derived from a comprehensive set of “hybrid” simulation experiments that combine a Monte Carlo simulation approach with census information at the block-, block-group-, and tract-level to generate plausible data scenarios. Results suggest that, all else equal, the larger scale is a safer choice. Furthermore, results caution against the use and interpretation of level- two variances and intra-class-correlation coefficients. Three-level models should be estimated if possible, and modelfit statistics should be used to assess the scale, or, combination of scales, at which neighborhood effects operate. Results also suggest that the scale needs to be grossly mis-measured in order to cause significant bias.

### INTRODUCTION

“Neighborhood” is a fluid concept whose meaning changes depending on the outcome, exposure and social group under consideration (Chaix, Merlo, Evans, Leal, and Havard 2009; Galster 2001; Guest and Lee 1984). Administrative units such as census tracts, block-groups and blocks can therefore hardly fit the demands of the concept. Nevertheless, census units, in particular census tracts, have remained the single most common neighborhood proxy in neighborhood effects research (Lee, Reardon, Firebaugh, Farrell, Matthews, and O'Sullivan 2008; Matthews 2008), mostly because of their availability and extreme practicality. While new sophisticated approaches to operationalizing the neighborhood are being developed, “how much” and under what circumstances the use of census units affects the results of neighborhood effects research remains to be assessed.

Defining the size of the neighborhood is an important conceptual and methodological issue in the measurement of neighborhoods. The geographic scale at which neighborhood effects operate is rarely known and even if it can be hypothesized, data limitations often force researchers to adopt a particular geographic scale for their analysis. In addition, since the ACS has replaced the census SF3 questionnaire large margins of errors of estimates at the block-group-level often force researchers to use tract-level information (Voss 2011). Furthermore, understanding scale sensitivity of neighborhood-effects also provides information on the importance of boundary issues in neighborhood effects research. A low scale sensitivity of results from multilevel models, for example, suggests that boundary issues at a particular scale are unlikely to be critical for the prediction of neighborhood effects: If changing the scale of the neighborhood does matter little for the prediction of neighborhood effects, then, fine-tuning their boundaries to better reflect actual neighborhood characteristics are unlikely to influence the prediction of model parameters either.

Prior research in Geography on the modifiable area unit problem has demonstrated that the impact of scale and boundary issues on statistical analysis can be substantial (for example: Gehlke and Biehl 1934, Openshaw 1978, Arabia 1989, Fotheringham and Wong 1991). Spielman and Yoo (2009) use a simulation study to demonstrate that even the sign of the predictor-outcome relationship can change under extreme conditions. Studies by geographers, however, have not assessed the modifiable area unit in a multilevel framework and the study by Spielman and Yoo, uses fully simulated data and yields no information on whether or not such extreme biases are likely to occur in “every-day” statistical analysis.

My goal is to generate evidence on the kind and magnitude of bias that occur if a neighborhood is assumed to be either bigger or smaller than the scale at which the actual neighborhood effects (or combination of neighborhood effects) are operating. In addition, I will explore the factors that exacerbate or mitigate this kind of bias. The outcome for this simulation study is Body Mass Index (BMI). BMI has been chosen because it is a health outcome that is of major importance to population health and because it has been the subject of numerous contextual studies that have found neighborhood-effects at various geographic scales. It will be shown however, that results from this study can be generalized to all continuous outcomes that are approximately normally distributed.

In order to assess not only the direction but also the magnitude of the bias that is introduced by misrepresentation of the scale this study uses, what I call, a hybrid simulation experiment. This approach is different from a full simulation study in that it does not generate all data in the simulation process. Instead, this approach uses census-information of Los Angeles County and parameters derived from a multilevel model estimated on the Los Angeles Family and Neighborhood Survey to generate plausible scenarios that are similar to those that researchers are likely to encounter in their own work.

I find that scale effects can introduce substantial downward bias into the estimates of the fixed and random effects at the neighborhood-level. All else equal, modeling a larger scale reduces the risk and size of bias. The magnitude of the bias depends, however strongly on how much the distribution of neighborhood characteristics changes when being aggregated at different scales. The variance components and intra-class-correlation coefficients are hugely influenced by scale-mis-measurement. Their bias can lead to over- or underestimation depending on the type of scenario. I also find that modelfit measures can serve under certain circumstances to detect the “right” scale. While scale issues can introduce considerable bias into the estimates of the fixed and random effects of multilevel models they matter only if the scale is grossly mis-measured. I will briefly describe the simulation approach and data and give an example of the results of the most complex simulations. Detailed recommendations that derive from the results of the full set of simulations are given in the conclusion.

The final paper will motivate and present the simulation scenarios in more detail and use graphic representation methods similar to Figure 1 to summarize the findings from all types of scenarios.

## METHOD

The hybrid simulation approach used for this study grafts the controlled environment of a simulation experiment on "real-time" census information (census boundaries and characteristics at the block-, block-group- and tract-level (Table 1)) from Los Angeles County. As a point of departure I further use simulation parameters borrowed from a multilevel model estimated on data from the Los Angeles Neighborhood and Family Survey to generate realistic starting scenarios (Table 2).

Using a Monte-Carlo simulation framework I will generate sets of individual-level BMI values that each contain a neighborhood effect at the block, block-group and tract-level, or a combination of effects stemming from different contextual levels. On each set of outcomes three to four multilevel models will then be estimated, one assuming the "true" (combination of) scale(s), that is, the scale(s) at which the neighborhood effects were generated, and all possible two-level "false" models that model the neighborhood at a wrong scale.

The "true" model will reproduce the data-generating parameters; the false models will allow assessing how the estimates of

(1) the neighborhood effect (the beta coefficient(s) of the neighborhood effect(s) ),
(2) the neighborhood-level variance
(3) and Intra-Class-Correlation Coefficient

differ from the data-generating parameters when the wrong neighborhood definition is chosen. Three types of scale-variable scenarios will cover a broad array of possible combinations of neighborhood-scales: I build on the simplest scenario where

(A) one neighborhood effect acts at a single scale

then, I built scenarios where

(B) one contextual characteristic (e.g. percent Hispanic population) is operating through different mechanisms at two different scales

and next generate scenarios where

(C) two contextual predictors are operating at two different scales but are measured at only one scale.

The starting scenarios will be manipulated in order to assess how the effect of misrepresenting the neighborhood scale varies by

(i) the strength of the underlying neighborhood effect and
(ii) the distributional qualities of the neighborhood predictor.

Finally, I will assess if
(4) a modelfit measure (the Akaike criterion) can serve to detect the "true model" that adequately represents the neighborhood scale(s).

Results of these scenarios present certain regularities that allow developing a set of recommendations for minimizing the risk and size of bias introduced by scale-mis measurement.

## DATA

The real data-based simulation presented here requires two types of input: the distribution of neighborhood characteristics at the block, block-group and tract-level, as well as model parameters that translate neighborhood differences in the predictor variable into change in the individual level outcome. Neighborhood-level predictors come from the Summary File 1 (SF1) from the 2000 census of Los Angeles County. An initial set of model parameters are drawn from a multilevel model that has been estimated on information from the first wave of the Los Angeles Family and Neighborhood Survey (Table 2).

The simulation study relies on a 10% random sample of all Los Angeles County tracts that have at least one inhabitant. Table 1 shows the number of census units at each level of all LA County and the 10% sample used in this dissertation. I will use the percentage of Hispanic population, percentage of African American population and percentage of vacant units per census unit as neighborhood-level predictors.[1] Table 1 compares their means and standard deviations at each of the three census levels in all of L.A. County and the 10% sample used for the simulations.

For the present analysis, these variables are not only of interest because they are three common neighborhood-level predictors, but also because they constitute three different simulation scenarios. Note that each predictor differs in how its distribution changes across scales. Misrepresenting the neighborhood scale would matter less if a predictor had the exact same distribution at each level. If, however, a neighborhood characteristic had little between-unit variance at one scale but high between-neighborhood variance at another, choosing the “wrong” neighborhood scale would cause a larger distortion in the x-y relationships. Percent Hispanic for example, changes its variance little when aggregated at different levels while percent vacant units reduces its variance by almost one third when being aggregated at the tract-level instead of the block-level. Mis-measurement is therefore more likely to affect models that mis-represent the scale of the effect of percent vacant units compared to the effect of percent Hispanic population.

## RESULTS

Table 3 presents as example the results of the most complex scenarios where one of the three neighborhood predictors is operating at the block- and another at the tract-level. Both are measured in turn with one true model (a three-level model) and two false models, a block-level and a tract-level model. Results are expressed as percent of the simulation parameters to facilitate comparisons. The beta-coefficients of the false block-level models can be found in Panel A and those yielded by the false tract-level models in Panel B. Each panel is made up of two leading columns that contain the results of the true three-level model (which yield estimates that account for 100% of the simulation parameter) and a set of 3x3 rectangular sub-panels. The off-diagonal sub-panels contain the results of the six pairs of predictors. The first row of sub-panels in Panel A contains all results of block-level models that have been estimated on a true block-level effect of the predictor Hispanic and one other neighborhood predictor acting at the tract-level[2]. The second row of sub-panels lists the estimates of models on scenarios that contain a block-level effect of African-American and a tract-level effect of one of the other two predictors. Analogously, the first column of sub-panels contains the results of block-level models estimated on outcomes that contain true tract-level effects of the percent Hispanic in addition to

---

[1] From here on I will refer to these predictors only as, for example, “the percent Hispanic” instead of the percent Hispanic population.

[2] Panel A: sub-panel in row 1, column 2: contains results of block-level models estimated on BMIs containing a block-level effect of percent Hispanic and a tract-level effect of percent African American; Row 1, Column 3: results of block level models on BMIs containing a block-level effect of percent Hispanic and a tract-level effect percent vacant units etc…

another variable operating at the block-level. The second column of sub-panels then presents the results of block models estimated on outcomes containing a tract-level effect of the percent African American plus one of the other two neighborhood predictors operating at the block-level. In each sub-panel the beta-coefficient of the block-level effect is listed first and the coefficient of the tract-level predictor second. So for example, the second panel in the first row contains the results of the model that measures both the true block-level effect of percent Hispanic and the true tract-level effect of percent African American at the block-level. Panel B is organized in an analogous fashion but contains the results of all tract-level models on the same scenarios. To facilitate comparisons the columns containing the neighborhood effects that are falsely measured are labeled as such and their background is greyed.

All else equal, tract models introduce much less bias to the mis-measured beta-coefficients than do the block-level models. Thus, for example, in the two tract-level models on the outcomes containing the block-level effect of percent Hispanic at too big a scale the false beta coefficients reproduces the true effect by between 89% to 100% (Table 3, Panel B).[3] In comparison, true tract-level effects of percent Hispanic (first column of graphs), if measured falsely with a model that is too small (Panel A), yield beta- coefficients that capture only between 77-82% of the true effect. Similarly, the true effect of the predictor, percent African American population, when captured with a model that assumes the scale to be too big reproduces the true effect almost exactly (i.e., close to 100 percent) or is just slightly over-estimated for the two smallest effect sizes (Table 3, Panel B, row two of sub-panels). The block-level models, in comparison, under-estimate the true effect by 32 to 20% (Table 3, Panel B, second column of sub-panels, false coefficients for African American). As expected, the predictor that is most susceptible to mis-modeling of the scale is the percent vacant units. This predictor's neighborhood effects are downward biased by 47% to 51% if they act at the block-level but are measured with a tract-level model (Table 3, Panel B, last row of sub-panels).The tract-level effect that is falsely measured with a block-level model however, is biased downward by 89% to 86% (Table 3, Panel A, last column of sub-panels). The size of bias are strikingly consistent along each column in Panel A and along each row in Panel B. This similarity suggests that the bias introduced to the mis-measured effect does not depend on the combination of falsely and truly measured predictors in the model.

Next, I will consider the truly measured neighborhood effects that are contained in each model along with the mis-measured neighborhood effect. (In Table 3) these are all effects that are not shaded grey). In Panel B the truly measured effect-sizes vary between 100% (tract-level effect of vacant units) and 112%.[4] Since absolute effect sizes are small these differences would not have introduced any changes to the interpretation of results.

Figure 1 presents the ICCs of both block- and tract-level. Each panel presents the ICCs of the block and the tract model for a particular predictor pair operating at the tract and the block-level. The upper dashed grey line presents the ICC of the true three-level model for two neighborhoods, the lower dash-dot line traces the ICCs of a single neighborhood level only. The closer the symbols fall to the upper dashed grey line the closer would be the researcher's

---

[3] The strongest (or weakest bias) are observed in most scenarios for the smallest effect size (b*0.5= 0.009). While none of the coefficients for this effect size is hugely different from the estimates of scenarios containing bigger effects it has to be noted that this is likely due either the sample size or the number of simulation runs being too small to reproduce such a small effect more reliably. In what follows I will therefore discuss the results for the bigger effect sizes and mention the result for b*0.5 separately, should they differ.

[4] Please note that the absolute difference of 3.5% over estimation accounts for only 0.001 point difference in BMI even for the biggest effect size.

conclusion to the “true” scenario. In all scenarios the block-level ICCs would lead to conclusions similar to those of the true models. In tract-level models, however the researchers would conclude that the between-neighborhood variance is only half as big as it is at both scales. The ICCs increase with effect size and more so if the mis-measured predictor has a high variance (e.g. Percent Hispanic). Overall, the bias in the ICC is substantial and its magnitude and direction depends on the type of mis-measurement and the size of the effect.

## CONCLUSION

Based on the full set of simulation experiments that will be presented in the final paper, I am making the following recommendations for neighborhood effects researchers:

1. If in doubt use a bigger scale but expect underestimation of neighborhood effects (beta-coefficients of neighborhood predictors). In my work the census tract proved to be the most robust level of measurement.

2. If possible, always examine how the mean and variance of the predictors changes when aggregated at different scales. The size of the bias introduced by scale mis-measurement can be minimal for predictors whose distribution varies little (and conversely important if the distribution of predictors changes drastically).

3. If there is doubt as to whether the same effect is operating at multiple scales then:
   a. Estimate three level model, and use the AIC to compare two- and three- level models.
   b. If estimating a three level model is not possible, use the larger scale model, and expect underestimation of the total neighborhood effect.

4. If there is some certainty on the scale of some effects but not on others then note that the truly measured effects are only affected little by the mis-measurement of other predictors in the models.

5. Do not interpret the neighborhood level variances and ICC.

6. Fine-tuning neighborhood boundaries is unlikely to improve the performance of multilevel models in predicting neighborhood effects since their results are relatively robust to moderate changes of scale (for example, block-group and tracts yielded comparable results).

TABLES

Table 1: Means and standard deviations for the percent Hispanic population, African American population, percent vacant units for the block, block-group and tract-level of LA county and the 10% sample

| | Sample | | | LA County | | |
|---|---|---|---|---|---|---|
| | Block Level | Block-Group Level | Tract Level | Block Level | Block-Group Level | Tract Level |
| Hispanic | 37.43 | 43.135 | 44.152 | 29.02 | 41.828 | 43.469 |
| | (32.899) | (31.668) | (30.566) | (31.748) | (29.861) | (29.492) |
| African American | 9.324 | 10.438 | 9.502 | 7.35 | 10.204 | 9.607 |
| | (19.323) | (18.515) | (15.792) | (17.286) | (17.571) | (15.685) |
| Vacant Units | 3.955 | 4.411 | 4.533 | 3.38 | 4.156 | 4.362 |
| | (6.725) | (3.767) | (3.643) | (8.209) | (4.156) | (4.09) |
| n of neighborhoods | 6744 | 614 | 204 | 89614 | 6351 | 2054 |

Table 2: Hierarchical Linear Model on sample of randomly selected adults of the Los Angeles Family and Neighborhood Survey (listwise deleted)

| | Coefficient (std err) | t-values |
|---|---|---|
| Intercept | 24.931 (0.628) | 40.32 |
| Age in years | 0.018 (0.005) | 3.73 |
| Female | -0.927 (0.155) | -6 |
| African American (ref White) | 0.766 (0.289) | 2.65 |
| Latino | 0.918 (0.235) | 3.9 |
| Asian | 0.73 (0.649) | 1.13 |
| Other Race | -1.56 (0.63) | -2.47 |
| Living w partner | 0.519 (0.167) | 3.11 |
| Log family income | 0.011 (0.041) | 0.28 |
| High School degree | 0.257 (0.217) | 1.18 |
| *Neighborhood Predictor* | | |
| Percent Hisp | (0.017) 0.005 | 3.85 |
| *Level 1 error Variance* | *13.75* | |
| *Level 2 error Variance* | *0.48* | |
| *N individuals* | *2337* | |
| *N tracts* | *90* | |

Table 3: Relative difference of beta-coefficients and simulation parameters, in percent.

| PANEL A BLOCK MODELS | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| True Eff. | | | Tract Hispanic | | Tract Af.-Am. | | Tract Vacant | |
| | *b blk* | *b Tr* | *True* | *False* | *True* | *False* | *True* | *False* |
| | | | | | *Bk. Hisp* | *Tr. A.-A.* | *Bk. Hisp* | *Tr. Vac* |
| Block | *100* | *100* | | | 100 | 77.8 | 100.0 | 11.1 |
| Hisp | *100* | *100* | | | 105.9 | 82.4 | 100.0 | 17.6 |
| | *100* | *100* | | | 103.8 | 76.9 | 100.0 | 15.4 |
| | *100* | *100* | | | 102.9 | 77.1 | 100.0 | 14.3 |
| | | | *Bk A.-A.* | *Tr Hisp* | | | *Bk A.-A.* | *Tr. Vac* |
| Block | *100* | *100* | 100 | 77.8 | | | 100.0 | 11.1 |
| Af.- Am. | *100* | *100* | 111.8 | 82.4 | | | 105.9 | 11.8 |
| | *100* | *100* | 107.7 | 80.8 | | | 103.8 | 15.4 |
| | *100* | *100* | 108.6 | 80.0 | | | 102.9 | 14.3 |
| | | | *Bk Vac* | *Tr Hisp* | *Bk A.-A.* | *Tr. A.-A.* | | |
| Block | *100* | *100* | 100.0 | 77.8 | 100.0 | 77.8 | | |
| Vac | *100* | *100* | 105.9 | 82.4 | 111.8 | 82.4 | | |
| | *100* | *100* | 103.8 | 80.8 | 107.7 | 76.9 | | |
| | *100* | *100* | 105.7 | 80.0 | 108.6 | 77.1 | | |

| PANEL B TRACT MODELS | | | Tract Hispanic | | Tract Af.-Am | | Tract Vacant | |
|---|---|---|---|---|---|---|---|---|
| | *b blk* | *b Tr* | *False* | *True* | *False* | *True* | *False* | *True* |
| | | | | | *Bk. Hisp* | *Tr. A.-A.* | *Bk. Hisp* | *Tr. Vac* |
| Block | *100* | *100* | | | 100 | 100 | 88.9 | 88.9 |
| Hisp | *100* | *100* | | | 100 | 100 | 100 | 100 |
| | *100* | *100* | | | 96.2 | 100.0 | 96.2 | 100.0 |
| | *100* | *100* | | | 97.1 | 100.0 | 97.1 | 100.0 |
| | | | *Bk A.-A.* | *Tr Hisp* | | | *Bk A.-A.* | *Tr. Vac* |
| Block | *100* | *100* | 100 | 100 | | | 100 | 100 |
| Af. | *100* | *100* | 105.9 | 105.9 | | | 105.9 | 100.0 |
| Am | *100* | *100* | 100.0 | 100.0 | | | 100.0 | 103.8 |
| | *100* | *100* | 100.0 | 100.0 | | | 100.0 | 100.0 |
| | | | *Bk Vac* | *Tr Hisp* | *Bk Vac* | *Tr. A.-A.* | | |
| Block | *100* | *100* | 55.6 | 100.0 | 55.6 | 100.0 | | |
| Vac | *100* | *100* | 52.9 | 105.9 | 58.8 | 105.9 | | |
| | *100* | *100* | 57.7 | 100.0 | 53.8 | 103.8 | | |
| | *100* | *100* | 60.0 | 100.0 | 54.3 | 102.9 | | |

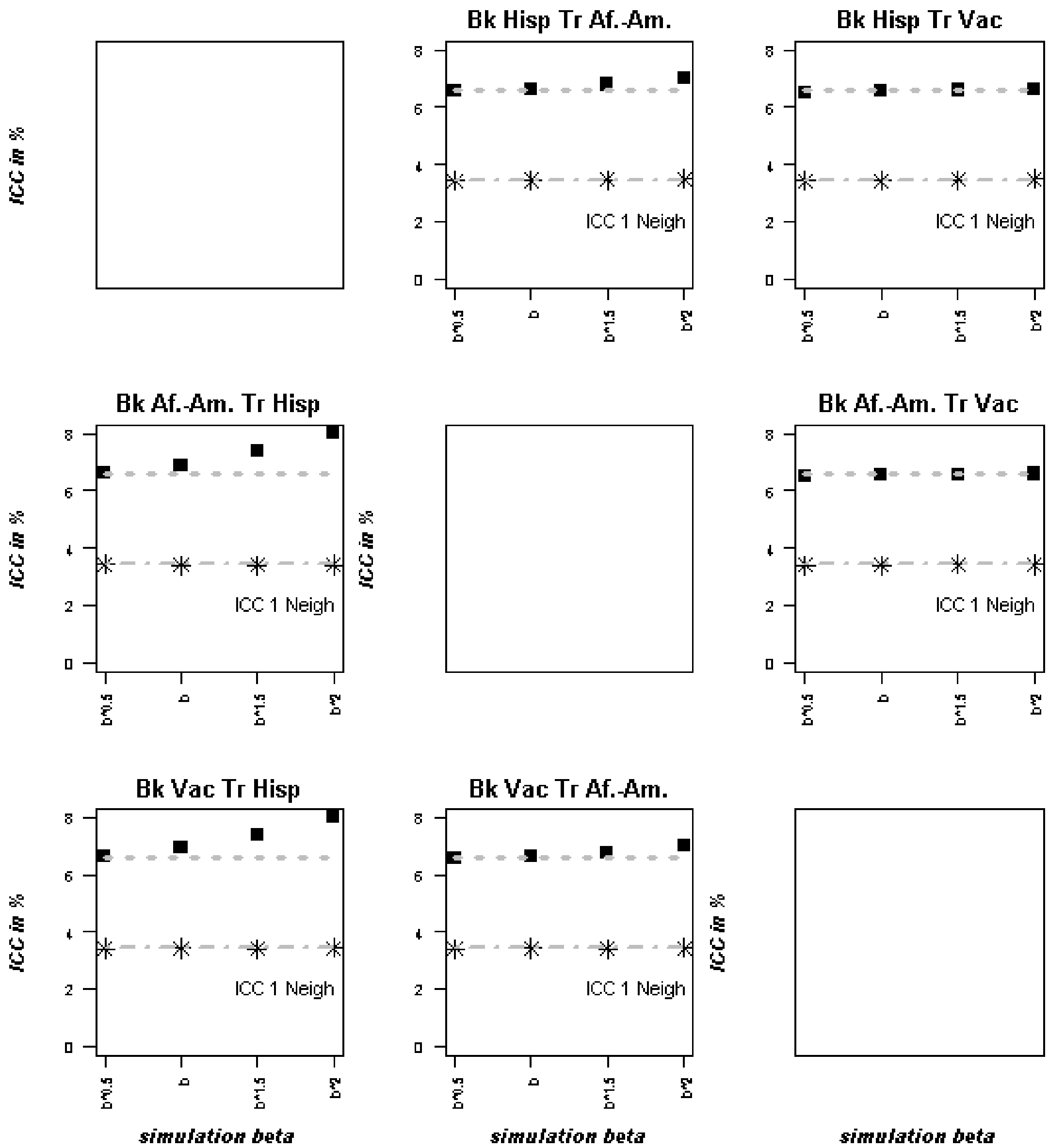


Figure 1: Monte Carlo estimates of ICC of block and tract models on six combinations of neighborhood predictors operating at the block and the tract level.

## REFERENCES


Arbia, G. 1989. *Spatial data configuration in statistical analysis of regional economic and related problems*, vol. 14: Springer.

Chaix, B, J Merlo, D Evans, C Leal, and S Havard. 2009. "Neighbourhoods in eco-epidemiologic research: delimiting personal exposure areas. A response to Riva, Gauvin, Apparicio and Brodeur." *Social Science & Medicine* 69:1306-1310.

Fotheringham, A.S. and D.W.S. Wong. 1991. "The modifiable areal unit problem in multivariate statistical analysis." *Environment and Planning A* 23:1025-1044.

Galster, G. 2001. "On the nature of neighbourhood." *Urban studies* 38:2111-2124.

Gehlke, CE and K. Biehl. 1934. "Certain effects of grouping upon the size of the correlation coefficient in census tract material." *Journal of the American Statistical Association* 29:169-170.

Guest, AM and BA Lee. 1984. "How urbanites define their neighborhoods." *Population & Environment* 7:32-56.

Lee, B.A., S.F. Reardon, G. Firebaugh, C.R. Farrell, S.A. Matthews, and D. O'Sullivan. 2008a. "Beyond the census tract: Patterns and determinants of racial segregation at multiple geographic scales." *American Sociological Review* 73:766-791.

Matthews, SA. 2008. "The Salience of Neighborhood Some Lessons from Sociology." *American Journal of Preventive Medicine* 34:257-259.

Openshaw, S. 1978. "An empirical study of some zone-design criteria." *Environment and Planning A* 10:781-794.

Spielman, SE and E Yoo. 2009. "The spatial dimensions of neighborhood effects." *Social Science & Medicine* 68:1098-1105.

Voss. P. 2011. “Spatial Statistical Methods”. Presentation at the Santa Barbara Specialist Meeting: Future directions in Spatial Demography. Santa Barbara. 2011.