Quantifying the Loss of Information Due to Geomasking in Health Survey Data

All research done on human subjects must protect the anonymity of our subjects, but their residential locations will give away their identity. Therefore, social science research and specifically health survey research takes great pains to mask, aggregate, or otherwise obfuscate the exact locations of subjects. We do this in published research, but we also mask geographic data during data sharing or dissemination efforts. I argue that geomasking does harm to the accuracy and validity of our research results. We need to come up with better strategies and tools for balancing the risk of identity disclosures against potential loss of information due to geomasking. My paper is a first step in this direction.

I will compare three geomasking techniques using data on self-rated health, smoking, and body mass index (BMI). I use data from the 2008-2013 waves of the Survey of Health of Wisconsin (SHOW). SHOW is uniquely positioned to address this question because it is an in-person survey that uses a stratified a random sample.

Background on Geomasking
Previous literature suggests, “. . . a rather consistent trade-off between protection of geoprivacy and accuracy of analytical results (Kwan et al, 2004, 26).” This is because random perturbations are introducing measurement error into the outcome variable and/or covariates. “The practical implication of this result is that second-moment analyses of randomly perturbed data are likely to be conservative, in the sense that they are likely to underestimate the true extent of the spatial heterogeneity or clustering (Diggle, 1993, 99).” Street networks have not been widely used to create geomasks, and they may provide a useful compromise between anonymity and accuracy.

Data and Methods
SHOW (Nieto et al, 2010) collects health information through in-person interviews, physical measurements, and bio-samples. The sampling procedure uses block groups stratified by
income and randomly selects addresses within the block groups. This sampling method helps ensure that there are enough subjects located within 1000 meters of each other to conduct the present analysis. Simple random sampling from across the state would probably make the current analysis impossible due to sparse data points. All cases in SHOW (N = 3384) were eligible for analysis. Cases were excluded if they had missing data or problematic geocoding—leaving a final sample of 2,619 individuals from 2008–2013. (Some households have more than one survey subject.)

For each of the 3 variables of interest, I ran a regression to control for age, gender, race, income, and education; saved the residuals; and used the residuals in the analysis of spatial patterns. This approach is analogous to using instrumental variables, and I call the residuals “filtered outcomes.” (Future research needs to be done on validating this method or finding alternative means of “controlling” for demographics.)

I applied three geomasking techniques. First, I selected a random location from the subject’s census tract. Second, I selected a random location within a 5km radius. Third, I added a novel approach of selecting a random location along the subject’s road segment.

**Measuring Spatial Autocorrelation**

Spatial autocorrelation can be measured at a global level with Moran’s I or Getis’s G, however, I am interested in how much the geomasking technique may change patterns at different distances—not just global autocorrelation. To accomplish this I made variograms of the filtered outcomes under alternative geomasks.

For the data with the true household location I created a bootstrap procedure to find out the expected bounds of spatial covariance in this data set. I selected 900 points with replacement and computed 1000 bootstrap draws for each filtered outcome. The confidence intervals in Figure 1 are based on the bootstrap procedure.

**Results**

Figure 1 demonstrates this method applied to BMI (conclusions are very similar for smoking and self-rated health). This image shows the covariance at different distance—in a situation with no spatial patterning, the expected covariance is 1. Covariances below 1 indicate autocorrelation, and covariances above 1 are either dispersal patterns or unexplained heterogeneity. The outer lines are based on the bootstrap draws from the actual locations. The solid red line represents the 95% confidence interval, and the dotted black line represents the 85% CI. The symbols are the observed covariances from the data under the 3 alternative geomasks. When symbols fall outside the lines, we have evidence that the geomasking technique has introduced “significant” changes in the spatial pattern at the given distance class. Random perturbations significantly change the patterns at 600, 800, 1800, and 2000 meters. Randomizing along a road or picking a random location in a census tract each have one significant point, but these may be due to oddities in the real data (i.e. the solid red lines constrict noticeably at approximately 500 and 2000 meters.)
Discussion

The results suggest that randomly jittering locations within a 5km radius will significantly change the results. Geomasking along a road segment may be a good approach because it does not significantly change the pattern as much as the other methods. Additionally, many more “+” symbols are below the line indicating covariance of 1, thus showing that BMI tends to be autocorrelated at distances below 3km, which matches the pattern found when the true locations are used. Perhaps this method alleviates some of the challenges described by Diggle. Granted the census tract method also shows almost no significant changes in the spatial pattern, but this method erases most of the autocorrelation below 3km (i.e., all of the triangles are above the line at covariance of 1).

One question then becomes, “is giving away a subject’s road segment too much of a risk for re-identifying them?” Almost certainly the answer depends on where the subject lives and how much other personal information is combined with the spatial attributes. But this question at least moves us away from a false dichotomy between anonymity and accuracy, and toward a tractable problem.

This is an important topic because spatial aspects of health research are hindered because data is aggregated to census tracts, zip codes or hospital catchment areas. Requiring collaborators and outside researchers to come use our “cold rooms” is impractical in most cases. This paper suggests there may be workable compromises. Further research is needed on a) an appropriate way to “control” for demographic variables in this context, b) a further development of the bootstrap methods within the variogram procedure to firm up the conclusions based on the confidence intervals, c) replication with other datasets, variables, and time frames, and d) perhaps a method for turning the complicated information in Figure 1 into a single indicator measure for each geomasking technique.
References