Measures of Sprawl Applied to Urban and Rural Areas, and Cardiovascular Disease Mortality: A Bayesian Spatial Analysis

Gerald Shoultz, ASPH/CDC Fellow, National Center for Health Statistics

BACKGROUND

The health impacts and consequences of sprawl are numerous. Examples of impacts include: Greater difficulty in finding safe places for walking and bicycling, greater problems with mobility for the elderly, the poor and children, and increased air pollution due to greater automobile emissions. Direct health consequences include increased incidence of asthma and chronic obstructive pulmonary disease following from increased air pollution, greater obesity due to decreased opportunities to walk to various points, and increased rates of cardiovascular disease due to increased obesity.

McCann and Ewing (2003), in a study of over 400 urban counties, found a positive relationship between sprawl and obesity. This paper modifies and extends the procedures of McCann and Ewing (2003) and Ewing et al (2002) in at least four ways. *First*, using only US Census data a modified form of Ewing et al's index of sprawl is developed; their study used other sources in addition to census data. *Second*, this index, unlike both Ewing et al and McCann and Ewing, is then calculated for ALL counties or US census equivalents in the United States—urban, suburban, exurban and rural. *Third*, since the entire United States is used as opposed to isolated urban/suburban areas it is possible to account for geographic adjacencies/relationships between counties; the model used here will do so. *Fourth*, in addition to county-level measures of sprawl and socioeconomic deprivation for 1990 measures of changes in both county-level sprawl and county-level socioeconomic deprivation will be developed and used in the model. A model for the relationship between our county-level measure of sprawl and Cardiovascular disease (CVD) mortality for White males age 35 and over for the years 1999-2001 will be presented; this model will control for age, Lung Cancer mortality (as a proxy for smoking) and county-level socioeconomic status. The model will be used to test the hypothesis that counties with a greater level of sprawl as defined by our index will have greater risk of CVD mortality.

Most measures of sprawl have land in or near urban areas in mind. But development of measures of sprawl outside the urban areas is like the development itself-sparse. Often ignored in such studies is the growth of exurbia. Because of this limitation information on the health impacts of sprawl on such areas are also lacking, and conclusions based on studies focused on urban and suburban areas cannot always be applied to exurban and rural areas. The *Merriam-Webster Online Dictionary* (2003) defines exurb as "a region or settlement that lies outside a city and usually beyond its suburbs and that often is inhabited chiefly by well-to-do families"(1). Exurbia, loosely defined, is the area generally geographically adjacent to, but not far from, both the urban and suburban fringe. Benefits such as lower taxes, cheaper land, less costly services, lower crime rates, and numerous other benefits similar to those found in small towns, coupled with current technology enabling businesses to operate profitably outside the city, have slowly but measurably drawn families and employers to the suburbs (*The move to Exurbia . . . and beyond* 2002 July 17). But with the growth of exurbia comes additional sprawl, additional reliance upon the car, and all the health problems that go with both.

Previous studies tend to focus on some form of density to measure sprawl. This, as Ewing et al (2002) succinctly put it "flies in the face of both the technical literature and popular conceptions of sprawl."(p. 10). The metropolitan areas in their study were home to over 150 milling people. Using 1990 and 2000 Census data, American Housing data for various years in the 1990's, the 1990 Census Transportation Planning Package, and data from the Census TIGER files Ewing et al. reduced 22 variables to four by means of Principal Components Analysis. These four factors, according to Ewing et al., represented "density, land use mix, degree of centering, and street accessibility" (p. 16). From the four factors an overall index was derived and compared to such transportation and travel outcomes as "distance driven per person per day . . . , average vehicle ownership per household, . . . , percent of commuters taking transit to work, percent of commuters walking to work, . . . , [and] ozone pollution levels."(p. 17). Controlling variables included per capita income, percent of the population between age 20-64 years of age, average size of the household, and the population of the metropolitan area. Response variables included distance driven per person per day, average commute times, and ozone pollution levels as outcome variables. They concluded that "for most travel and transportation outcomes, sprawling regions perform less well than compact ones." (p. 4).

Using Ewing et al's study as a starting point, three categories of indicators of sprawl are derived: (1) Housing Density, (2) Centers Concentration and Clustering, and (3) Road Accessibility. These categories are intended to account for elements of urban and rural areas. From these categories a single measure of sprawl for 1990 and a single measure of the change in sprawl from 1990 to 2000 for each of 3137 counties or combinations of counties or their equivalents in the United States is obtained. These indices are used to model the relationship between sprawl and Cardiovascular disease (CVD) mortality for white males age 35 and over for the years 1999 to 2001.

FORMATION OF THE MODEL

Three variables are controlled for in our model: Age (with a cubic polynomial), Lung Cancer Mortality (as a proxy variable for smoking), and county-level socioeconomic deprivation for 1990 (via index of Singh). Furthermore, spatial autocorrelation is accounted for via the Conditional Autoregression (CAR) model of Besag et al (1991).

The log-linear model fitted is:

$$O_i \sim Poisson(\mu_i)$$

$$\ln(\mu_i) = \ln(E_i) + \alpha + \beta_X X_i + \beta_{age} age_i + \beta_{age^2} age_i^2 + \beta_{age^3} age_i^3 + \beta_{Sprawl} sprawl_i + \beta_{\Delta Sprawl} \Delta sprawl_i + \beta_{DepInd} Depind_i + \beta_{\Delta DepInd} \Delta Depind_i + \beta_{LungC} LungC_i + s[county_i] + v_i$$

Where

i	=	County/age-group combination (i.e., Montgomery County white males age 35-45)
O_i	=	Observed number of deaths in county/age-group i
E_i	=	Expected number of deaths in county/age-group i (based on county and US populations and US mortality for the given age group)
α	=	Overall parameter
X_i	=	Density variable in question (1990 Housing Density, Change in Housing Density from 1990 to 2000, or percent change in Housing Density from 1990 to 2000).
Age_i	=	age group (- 2.5 for 35 to 44 year olds, -1.5 for 45 to 54 year olds, . . . , 2.5 for those 85+)
$(Age_i)^2$	=	Square of Age_i
$(Age_i)^3$	=	Cube of Age_i
$Sprawl_i$	=	County-level sprawl index for 1990
$\Delta Sprawl_i$	=	Change in county-level sprawl index from 1990 to 2000
$DepInd_i$	=	County-level deprivation index for 1990
$\Delta DepInd_i$	=	Change in county-level deprivation index from 1990 to 2000
$LungC_i$	=	County-level Lung Cancer Mortality for 1999-2001
$s[county_i]$	=	Structured variation based on adjacency of counties (spatial autocorrelation), using the Conditional Autoregression (CAR) Model of Besag et al (1991); the precision (reciprocal of variance) is considered to have a specified distribution
V_i	=	unstructured variation, $v[i] \sim N\left(0, \sigma_\tau^2\right)$, $1/\sigma_\tau^2 \sim \textit{Specified Distribution}$

A Hierarchical Bayesian model of Poisson regression is used to model the relationship between CVD mortality and the other variables. The precision (reciprocal of the variance) for the structured variation and unstructured variation are both random. Four distributions will be considered: Three are Gamma(α, β) distributions with parameters α and β (Gamma(0.5,1/0.0005), Gamma(0.001, 1/0.001), and Gamma(1,1)) and the fourth is a Uniform distribution with endpoints 0 and 1000. For each distribution Markov Chain Monte Carlo (MCMC) methods will be used to obtain parameter estimates, and the Uniform(0,1000) distribution will be used to obtain estimates of relative risk of CVD mortality. A minimum number of iterations for each of three chains will be conducted as a “burn-in” to achieve convergence, and then an additional number of iterations per chain will be done to achieve the actual estimates. Maps of relative risk estimates for CVD mortality by age and county will be obtained. These maps will be used to discuss geographic trends for CVD mortality.