Well, it depends on where you're born: A practical application of geographically weighted regression to the study of infant mortality in the U.S.

P. Johnelle Sparks and Corey S. Sparks[1]

Introduction

Infant mortality remains one of the most sensitive measures indicating the general health status of the population (see Clarke et al. 1994; Cramer 1987; Nersesian 1988; Newland 1981; Singh and Yu 1995). Major medical advances in prenatal and early infant care have led to dramatic decreases in infant mortality rates in the United States. This is largely attributable to fewer infants being born low birth weight and reducing complications during pregnancy that could lead to an infant death. Recent reports from the National Center for Health Statistics indicate that the overall infant mortality rate for the U.S. in 2003 was 6.85 deaths per 1,000 live births (Hoyert et al. 2006). Yet, the U.S. still lags behind many industrialized nations with regard to overall infant mortality rates, and variation in rates across places in the United States and between racial/ethnic groups remains (Nersesian 1988). The goal of our research is to identify possible explanations for the variation in infant mortality rates across space.

Social structure theory suggests that the social, racial, economic, and educational distributions of places may impact health outcomes (Bird and Bauman 1995). This theory argues that minority concentration and residential segregation are just two possible components of the social structure that could negatively impact health outcomes at the aggregate level, such as for counties or states. Within the social structure, resources are distributed unequally. Therefore, the social structure can perpetuate inequality in such things as quality employment, affordable and safe housing, incomes, or health care resources to name a few. It is from this theoretical perspective that we test for spatial variation in county infant mortality rates using geographically weighted regression.

The purpose of this paper is to examine the associations between various socioeconomic indicators of inequality and county infant mortality rates across the United State. By using analytical methods that allow the relationships between our socioeconomic indicators and infant mortality to vary across space, we hope to provide a better understanding of the often continuous and spatially varying nature of inequality and health. The ultimate goal of this work is inform policy makers about the distinct regional variations in the relationship between socioeconomic inequality and population health.

Data and methods

Data for this analysis are taken from two sources: Compressed Mortality Files from the National Center for Health Statistics and 2000 U.S. Census of Population and Housing, Summary File 3. Five year sex-race standardized infant mortality rates for the years 1998-2002 serve as the dependent variable in this analysis. Standardization of mortality rates is used in order to facilitate the comparison of rates across groups. It is important in this analysis to standardize mortality rates based on sex and race, because mortality risks vary greatly based on these demographic characteristics during the first year of life. For this analysis we use the sex-race distribution of the 2000 U.S. population as our standard population. Independent variables for this analysis are taken from the Census and serve as county-level economic and social indicators. We include five independent variables in our analyses: percentage of the county population that is rural, percentage of the county population that is black, percentage of the county population that is Hispanic, the percentage of the county population below age 5 that lives in a family that is below the federally

[1] Department of Demography and Organization Studies, University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249.

designated poverty level, and percentage of female headed households in the county. The percentage of the county population that is rural, defined as a population not classified as urban by the Census Bureau, is constructed by dividing the county's rural population by the county total population and multiplying this value by 100. Similarly the percentages of the county population that is black or Hispanic are constructed by taking the total number of black or Hispanic residents per county and dividing those numbers by the total county population in 2000 and multiplying the value by 100. The number of county children below the age of five living below the federally designated poverty threshold is divided by the total county population under the age of five and multiplied by 100 to construct the percentage of the country population under the age of 5 living in poverty. To construct the percentage of female headed households we take the total number of female headed households per county divided by the county population and multiplying this value by 100.

The primary method used in this paper is geographically weighted regression (GWR) (Brunsdon et al. 1998; Fotheringham et al. 2002). The primary benefit of using the GWR approach is that it allows us to visualize how the effects of each covariate in our model vary over geographic space. This approach is opposed to the autoregressive models frequently used in spatial demography that, while controlling for autocorrelation in either the dependent variable or the error structure, still estimates global regression parameters. Since we are primarily interested in how the processes that influence infant mortality rates vary across space, global parameter estimates are not sufficient to increase our understanding of the process.

The GWR model takes the traditional OLS model and extends the framework by allowing local, rather than global parameter estimates of β_i, this is rewritten as:

$$y_i=\beta_0(u_i,v_j)+\sum_k \beta_k(u_i,v_j)x_{ik}+\epsilon_i$$

, where now each β_i is estimated at the location u_i,v_j where i and j are the coordinates or geographic location of the observation i. $\beta_i(u_i,v_j)$ is the local realization of the continuous β function at point i. This constructs a trend surface of parameter values for each independent variable and the model intercept. Note that the basic OLS regression model above is just a special case of the GWR model where the coefficients are constant over space. The parameters in the GWR are estimated by weighted least squares. The weighting matrix is a diagonal matrix, with each diagonal element being a function of the location of the observation. If W_i is the weighting matrix at location i, then the parameter estimate at that location would be specified as:

$$\hat{\beta}_{\mathbf{i}}=\left[\mathbf{X'W_iX}\right]^{-1}\mathbf{X'W_iY}$$

The role of the weight matrix is to give more value to observations that are close to i, as it is assumed that observations that are close will influence each other more than those that are far away. The form of this weight matrix can vary (as far as how it is calculated). Typically a distance based weight is used, and an observation has 0 weight if it is beyond a distance d_{ij}. This distance weighting is used for defining neighborhoods on which to base the estimates of β_i. If an observation is within the distance d, then it is included in the analysis for location i. We employ the kernel method of weighting proposed by Brunsdon et al. and Fotheringham et al., in which a global kernel bandwidth is estimated via cross-validation of the observed and predicted mortality rates. This bandwidth parameter defines the "neighborhood" or the number of observations in the data that will be used to estimate the regression parameters at that specific i,j location. Model estimation is done in R version 2.7.2 using the spgwr library and we test the improvement in model performance using the global F-tests and the tests for geographic variation in the regression parameters derived in Leung et al. (2000).

Preliminary results

The results of our initial OLS regression models are provided in Table 1. They suggest that as the percentage of the county population that is rural, the percentage of the county population that is black and the percentage the county population that is Hispanic increase, these indicate a decrease in the county infant mortality rate, while as the percentage of the county population of children in poverty and the percentage the county population of households with a female head increase, that these factors tend to indicate increase county infant mortality rates.

Table 1. Results from Ordinary Least Squares (OLS) Regression Model for US County Infant Mortality Rates, 1998-2002.

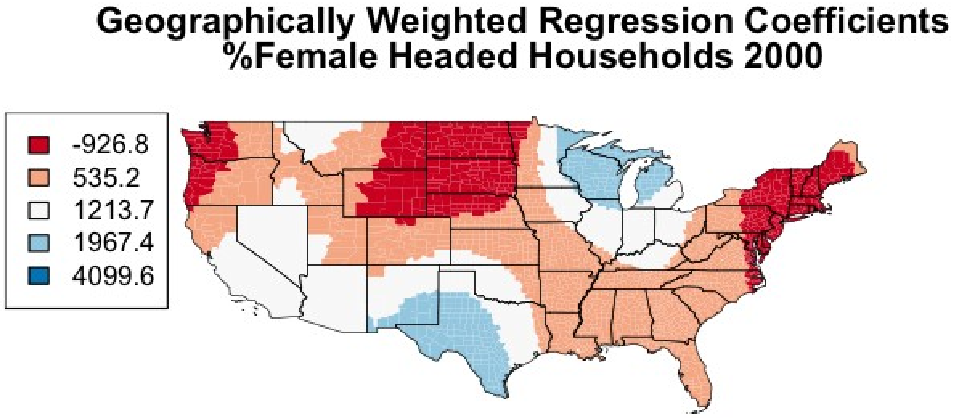
Variable	Estimate	t	Prob(>|t|)
Intercept	383.00	5.69	<.0001
% Rural	-153.04	33.97	<.0001
% Black	-205.74	80.02	.0102
% Hispanic	-325.39	80.22	<.0001
% Children in Poverty	982.25	213.51	<.0001
% of Female Household Heads	1129.82	266.99	<.0001

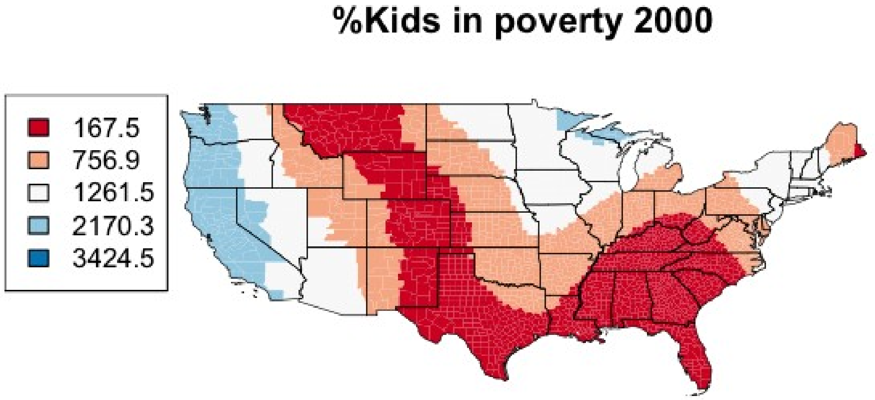
While these effects provide a baseline estimate of the effects of our independent variables, we are more concerned with testing the hypothesis that these parameters vary across space. Table 2 presents the results of the GWR, indicating the minimum, median, first and third quartiles, and maximum values for the estimated regression coefficients. The OLS estimates are also given for comparison.

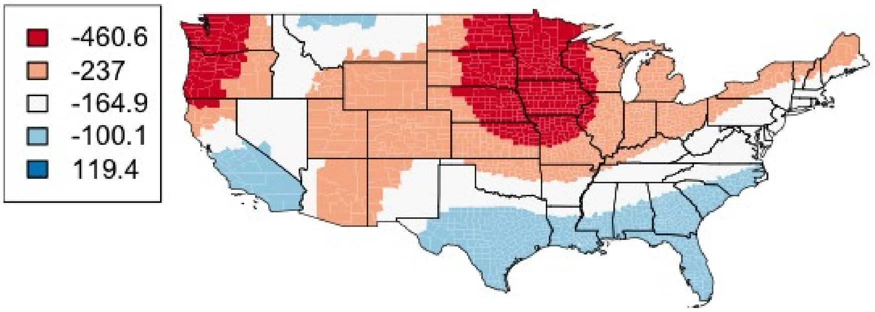
Table 2 Distribution of GWR Coefficients for US County Infant Mortality Rates, 1998-2002

Variable	Minimum	1st Quartile	Median	3rd Quartile	Maximum	OLS Estimate
Intercept	-453.2	335.9	446.2	503.0	782.3	383.00
% Rural	-460.6	-208.7	-162.5	-113.9	119.4	-153.04
% Black	-1762.0	-322.9	-191.2	-77.1	1373.0	-205.74
% Hispanic	-1085.0	-408.5	-262.7	.137.6	426.0	-325.39
% Children in Poverty	167.5	604.8	867.6	1300.0	3425.0	982.25
% of Female Household Heads	-926.8	771.2	1010.0	1364.0	4100.0	1129.82

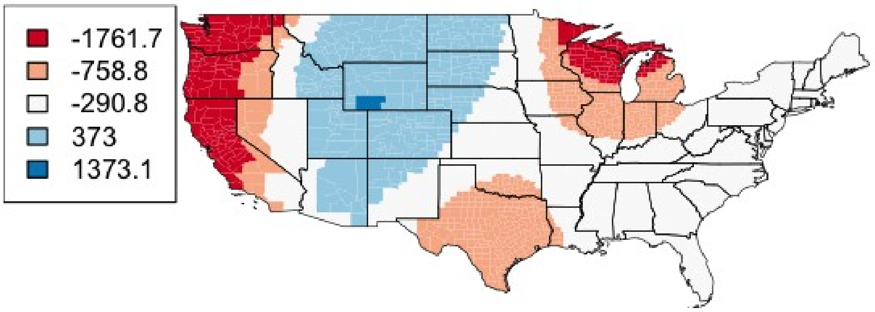
While inspection of the table of coefficients reveals significant variability in the regression coefficients across space, and a visual inspection of the regression parameters better illustrates how the relationships between the independent variables and county infant mortality rates vary across space. The following figures provide an illustration of the geographic variation in regression parameters. The ultimate goal of this work is inform policy makers about the distinct regional variations in the relationship between socioeconomic inequality and population health using county infant mortality rates as an example.


Geographically Weighted Regression Coefficients
%Female Headed Households 2000
-926.8
535.2
1213.7
1967.4
4099.6

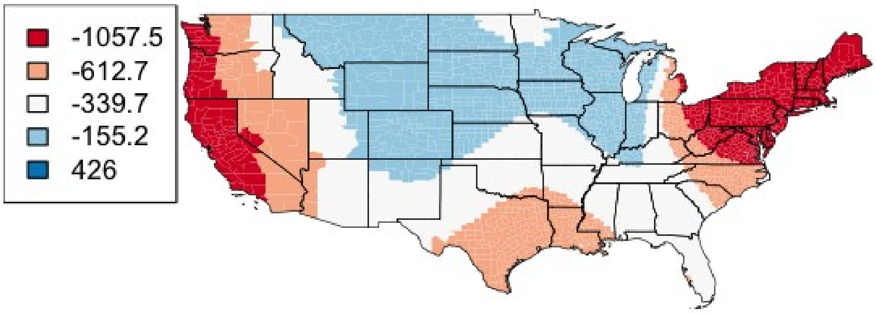




%Rural population 2000



%Black 2000



%Hispanic 2000

References

Bird, S.T. and K.E. Bauman. 1995. “The Relationship between Structural and Health Services Variables and State-Level Infant Mortality in the United States.” *American Journal of Public Health* 85(1): 26-29.

Brunsdon, C., S. Fotheringham, and M. Charlton. 1998. “Geographically Weighted Regression: Modeling Spatial Non-Stationarity.” *Journal of the Royal Statistical Society Series D-The Statistician* 47: 431-443.

Clarke, L.L., F.L. Farmer, and M.K. Miller. 1994. “Structural Determinants of Infant Mortality in Metropolitan and Nonmetropolitan America.” *Rural Sociology* 59(1):84-99.

Cramer, J.C. 1987. “Social Factors and Infant Mortality: Identifying High-Risk Groups and Proximate Causes.” *Demography* 24(3):299-322.

Fotheringham, A., C. Brunsdon, and M. Charlton. (eds.) 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. New York: Wiley.

Hoyert, D.L., M.P. Heron, S.L. Murphy, and H.C. Kung. 2006. “Deaths: Final Data for 2003.” *National Vital Statistics Report* 54(13): 1-120.

Leung, Y., C. Mei, and, W. Zhang. 2002 Statistical tests for spatial nonstationarity based on the geographically weighted regression model. *Environment and Planning A* 32: 9-32.

Nersesian, W.S. 1988. “Infant Mortality in Socially Vulnerable Populations.” *Annual Review of Public Health* 9:361-377.

Newland, K. 1981. *Infant Mortality and the Health of Societies*. Washington, DC: Worldwatch Institute.

R Development Core Team (2008) R: A Language and Environment for Statistical Computing. 2.7.2 ed. Vienna, Austria, R Foundation for Statistical Computing.

Singh, G.K. and S.M. Yu. 1995. “Infant Mortality in the United States: Trends, Differentials, and Projections, 1950 through 2010.” *Health Services Research* 85(7):957-964.