1. Full Title:

Comparing Methods of Incorporating Spatial Correlation in Models of the Association Between Hospitalized Myocardial Infarction and the Burden of Socioeconomic Status in Communities

b. Abbreviated Title (Length 26 characters):

Modeling Neighborhood SES

2. Writing Group: Kuo-Ping Li, Gerardo Heiss, Kathryn Rose, C. Suchindran, Eric Whitsel, Joy Wood. We welcome others who are interested.

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. __X___ [please confirm with your initials electronically or in writing]

First author: Kuo-Ping Li, PhD
Address: Dept. of Biostatistics
CB# 7420, McGavran-Greenberg Hall
University of North Carolina
Chapel Hill, NC 27599-7420

Phone: 919-843-6296 Fax: 919-843-4630

Email: kpli@email.unc.edu

Corresponding/senior author (if different from first author correspondence will be sent to both the first author & the corresponding author):

Kathryn Rose
Cardiovascular Disease Program
137 E Franklin St, Suite 306
Chapel Hill NC 27514

E-mail: kathryn_rose@unc.edu
3. **Timeline:**

Analyses to begin in Summer 2006. Draft of manuscript is expected during Fall 2006.

5. **Rationale:**

Regression models are widely used in public health studies to relate explanatory variables to the outcomes. The data used in regression analysis are often (perhaps too often) assumed to be independently distributed. However, this assumption can be an oversimplification. This is especially true for spatial data as it is well known that observations that are close to each other are more similar than those that are far apart [1]. Thus, models that consider spatial correlation are more desirable for spatial data.

In general, spatial models offer two types of benefit. The first considers the loss of information introduced by spatial correlation relative to independent samples of the same size. In this situation, incorporating the correlation into the model improves the accuracy of statistical inference. Another benefit of spatial models derives from the view that spatial correlation itself is an extra source of information. For example, when analyzing incidence of a rare disease, spatial correlation can be used to stabilize low counts in a region by “borrowing strength” from nearby regions.

In ARIC ancillary study # 2004.05 (The Burden of SES in Communities), addresses abstracted from hospital records and death certificates for ARIC surveillance events have been geocoded so that neighborhood (census tract level) socioeconomic characteristics can be linked with event records. This allows an examination of the association between neighborhood socioeconomic characteristics and MI occurrence, treatment, and survival. Through georeferencing it becomes possible to incorporate spatial correlation into the analysis. One of the aims of this study is to examine whether the variation in socioeconomic characteristics of different spatial regions contributes to the observed variation in the rates of hospitalized MI events in these regions. As mentioned above, including spatial correlations in the model can improve the model accuracy as well as stabilize the observed event counts which can be quite low in many regions.

Some methodological issues have to be considered in order to treat the spatial correlation properly. Three types of spatial models are commonly used, i.e., (1) fixed effect general linear models with spatial covariance structures (spatial GLM) [2-4], (2) mixed effect models with spatial random effects (spatial GLMM) [5], and (3) computational Bayesian spatial models [6-7]. In general, the choice of spatial model can greatly affect the results of statistical estimation. The first purpose of this manuscript is to compare these three modeling approaches and examine their strengths and shortcomings in the context of ARIC community surveillance.

Second, all the spatial models mentioned above rely on user-specified “neighbor structure”, that is, a description of geographical relationships of spatial regions under consideration. The neighbor structure can be defined based on region-to-region distances, the adjacency of regions, or some weighted combination of these (for example, weighted
by population size). The spatial units considered in our models are the census tracts in the four ARIC community surveillance study sites. These census tracts vary widely in several ways: land area, number of neighbors within a given distance, number of nearest neighbors, etc. For example, the land area of the 75 census tracts in Forsyth County, NC varies from 0.8 to 78 km², and the number of neighboring tracts within a radius of 10 km varies from 3 to 45. There is less variation in the number of tracts adjacent to a tract, which ranges from 2 to 8. Moreover, these differently sized tracts are not randomly located but are highly clustered—tracts in downtown areas tend to be small and to have many neighbors; suburban and rural tracts are large and have fewer neighbors. This implies that the neighbor structure based on tract-tract distances would be very different from neighbor structure based on adjacency of tracts. There is currently no consensus on the best approach to construct the neighbor structure in a given context. Thus, the second purpose of this manuscript is to construct the expressions of distance- and adjacency-based neighbor structures of ARIC study sites, compare the results of the two approaches, and to discuss their uses for spatial analysis in the ARIC study.

5. Main Hypothesis/Study Questions:

We will consider the following three types of models:
1. Spatial GLM,
2. Spatial GLMM,
3. Computational Bayesian spatial models.

Two ways of constructing the neighbor structure will be used to describe the spatial relationship of tracts:
1. Based on tract-tract distance,
2. Based on adjacency of tracts.

We will attempt to answer the following questions:
1. Do different model-neighbor structure combinations give different estimates of model parameters? If so, what factors contribute to the differences?
2. Within the context of ARIC community surveillance, what would be the strengths and shortcomings of each modeling approach?

Data:

1. The outcomes: aggregated hospitalized MI event counts for the period 1999-2001 of census tracts within the ARIC study sites, according to the patients’ places of residence, sex, race (black, white), and 5-year age groups. Expected event counts are also calculated using internal standardization.
2. Covariates: Census tract-level socioeconomic status (SES) data from the year 2000, including median household income, sex ratio of population, ratio of black population to white population, and age distribution.
3. Spatial information of census tracts: the coordinate (latitude, longitude) of census tract centroids
**Exclusions:**

We will include hospitalized, definite or probable MI surveillance events. We will limit cases to those occurring in 1993 and later, as addresses were not abstracted prior to this time. We will exclude ungeocodable events, geocoded events outside ARIC community boundaries, as well as Blacks in Washington County and Minneapolis.

**Analyses:**

We will use Poisson regression as the common form to model the observed incident counts in terms of the expected counts, SES covariates, and tracts’ neighbor structures. For spatial GLM, we will construct a model in which the outcomes are treated as independent, and all spatial variations on the outcomes are described by the fixed, location-dependent covariate effect and spatially correlated residues. For spatial GLMM, we will construct a model similar to spatial GLM but with spatial random effects with non-spatial residues. For the computational Bayesian model, we will construct a Bayesian counterpart of the spatial GLMM, and use computational Bayesian techniques to estimate the model parameters. In all three models, we will use conditional autoregressive (CAR) models to describe the spatial correlation.

The neighbor structure has to be specified for each CAR model as follows: For the distance-based neighbor structure, we will designate all tracts whose centroids are within a given radius from the centroid of the tract under consideration as its neighbors. Different radii (2 km, 5 km, 10 km) will be used for comparison. For the adjacency-based neighbor structure, we will include tracts that share common boundaries with the tract under consideration as its neighbors.

For computation, suitable matrix representations are needed for the neighbor structures. The GeoBUGS software package will be used to obtain the matrix representation of the adjacency-based neighbor structure. The process of constructing a distance-based neighbor structure is more complicated, and currently there is no convenient software for doing this. We will write our own software to obtain the matrix representation of the distance-based neighbor structure.

The GLM and GLMM will be evaluated by general-purpose statistical packages such as SAS and R. For the Bayesian model, WinBUGS will be used to carry out the numerical computation.

To answer our study questions, the estimated model parameters and their confidence intervals (GLM and GLMM) and credible sets (Bayesian models) will be examined. The residues will also be checked. Criteria such as AIC and DIC will be used to compare different models.


7.a. Will the data be used for non-CVD analysis in this manuscript?  
____ Yes  
__x__ No
b. If Yes, is the author aware that the file ICTDER02 must be used to exclude persons with a value RES_OTH = “CVD Research” for non-DNA analysis, and for DNA analysis RES_DNA = “CVD Research” would be used?  

Yes  ____ No

(This file ICTDER02 has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript?  ____ Yes  

_x___ No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER02 must be used to exclude those with value RES_DNA = “No use/storage DNA”?  

n/a  ____ Yes  ____ No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at:  http://www.cscc.unc.edu/ARIC/search.php

__X___ Yes  _______ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

MS 1102, MS1103

11. a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data?  

_x__ Yes  ____ No

11.b. If yes, is the proposal  

__X__ A. primarily the result of an ancillary study (list number*)

AS 2004.05

___  B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* _________  _________  _________)

*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.
References