Document Title:	Robust Spatial Analysis of Rare Crimes, Executive Summary
Author(s):	Avinash Singh Bhati
Document No.:	205909
Date Received:	June 2004
Award Number:	2002-IJ-CX-0006

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

EXECUTIVE SUMMARY

April 2004

Robust Spatial Analysis of Rare Crimes

Executive Summary submitted to the Mapping and Analysis for Public Safety (MAPS) Program, National Institute of Justice (NIJ)

Avinash Singh Bhati

URBAN INSTITUTE
Justice Policy Center

Robust Spatial Analysis of Rare Crimes

EXECUTIVE SUMMARY*

Avinash Singh Bhati

Justice Policy Center, The Urban Institute
2100 M Street, N.W., Washington, D.C. 20037

March 2004

OVERVIEW

The main goal of this project was to develop an analytical approach that will allow researchers to incorporate spatial error structures in models of rare crimes. In order to examine the causes of violence, researchers are frequently confronted with the need to apply spatial econometric methods to models with discrete outcomes. Appropriate methods for doing so when the outcomes are measured at intra-city areal units are lacking. The aim of this research was to fill that gap.

This research effort developed and applied the framework to a real-world empirical problem. It examined the socio-economic and demographic determinants of disaggregate homicide rates at two different intra-city levels of areal aggregation and compared inferences derived from several sets of models. The analysis was conducted on disaggregated homicide counts (1989-91) recorded in Chicago's census tracts and neighborhood clusters using explanatory factors obtained from census sources.

An extension of the Generalized Cross Entropy (GCE) method was applied to these data in an attempt to utilize their flexibility in allowing error structures across space. In addition, an information-based measure was developed and used in selecting the hypothesized error structure that "best" approximates the true underlying structure.

Findings from this research confirmed that ignoring spatial structures in the regression residuals often leads to severely biased inferences and, hence, a poor foundation on which to base policy. In addition, evidence was found of homicide type-specific and areal units-specific models, highlighting the need for disaggregating violence into distinct types. However, resource deprivation in

*This research was supported by grant # 2002-IJ-CX-0006 from the Mapping and Analysis for Public Safety (MAPS) program of the National Institute of Justice (NIJ), Office of Justice Programs, U.S. Department of Justice. Points of view in this document are those of the author and do not represent the official positions or policies of the U.S. Department of Justice, the Urban Institute, its trustees or its funders.

a community was found to be a reliable and persistent predictor of all types of violence analyzed and at both levels of areal aggregation. Additionally, there was evidence of a spill-over effect of resource deprivation on the amount of violence expected in neighboring areas. This highlights the need for taking seriously the spatial structure in a sample when planning for and implementing policy measures, especially at the intra-city level, where the observational units are spatially linked in meaningful ways.

The GCE approach utilized in this project offers several avenues for future research especially as they relate to the analysis of rare crimes. This includes the possibility of modeling other substantive spatial processes, an improved modeling of underlying population-at-risk instability, modeling mixed processes, and modeling spatio-temporal dynamics.

BACKGROUND

Researchers have attempted to model the observed cross-sectional variations in homicide rates using macro-structural covariates at various levels of areal aggregation. These include studies where, prior to modeling the phenomenon, researchers aggregate homicide counts within countries, states, counties, Metropolitan Statistical Areas, neighborhoods, or census tracts. Typically, researchers also aggregate across various types of homicides when they are interested in modelling violence in general. Alternately, they sometimes model disaggregated homicide rates, with the homicide type or the victim/offender race, gender, etc., forming the bases for disaggregation. Several of the existing studies also use data aggregated over a few to several years, assuming, implicitly if not explicitly, relative stability in the data generating processes over time.

At higher levels of areal aggregation, when the number of homicide counts may be sufficiently large and when the underlying data generating mechanisms may in fact be temporally stable, these aggregations yield criterion measures (dependent variables) that can either be considered continuous or, at the very least, can be satisfactorily transformed into continuous variables. Therefore, the traditional spatial analytical toolkit — commonly labeled "Spatial Econometrics" — that is well developed for the linear model, can be applied directly. At lower levels of areal aggregation, however, several problems preclude a direct application of these methods.

At local (intra-city) levels of areal aggregation, such as neighborhoods, census tracts, blocks, etc., more often than not the count of rare crimes (e.g., homicides) is extremely low. For many of the units the researcher may record no events, yielding a sample with a preponderance of zero counts. In addition, the distributions of the observed outcomes in the sample is typically highly skewed. One could aggregate the events over extended periods of time (such as a decade or two) and hope to obtain sufficiently high counts that would allow the outcome to be treated as continuous. However, macro-characteristics at local levels of areal aggregation are typically more volatile over time than those at higher levels of aggregation. Hence, temporal aggregation over extended periods of time may lead to distorted inferences which could aggravate, rather than mitigate, the problem. Finally, when counts are low, commonly used data-transformation approaches such as Freeman-Tukey, logarithmic, etc., result in transformed variables that do not necessarily yield the continuous, smooth, symmetrical distributions they are supposed to yield. As such, they are neither an optimal nor a guaranteed solution.

The problems noted above have, of course, long been recognized by researchers and there exist a multitude of models and methods that are more appropriate for use when the criterion measure is discrete. But what is problematic in these approaches is the incorporation of the spatial structure in the sample. With the wealth of geocoded data that are increasingly becoming available at local levels both from census sources and from primary data collection efforts, researchers analyzing homicides or other rare crimes are more frequently confronted with the need to apply spatial econometric methods to models with discrete outcomes.

GOALS OF THE PROJECT

The main goal of this project was to develop an analytical framework that can be used for robust analysis of rare crimes that are typically observed at local (intra-city) levels of areal aggregation. The need for such methods is pressing; common real-world data and sample features such as discrete outcomes, finite samples, ill-conditioned data, spatial clustering, ill-measured regressors, etc., all preclude a simple adoption of the standard Ordinary Least Squares (OLS) framework with its associated spatial-analytical toolkit.

As a means of applying this method to a real-world empirical problem, a second goal of this project was to assess the impacts of socio-economic and demographic characteristics of a community that are commonly theorized to affect the amount of violence it can expect to experience; to assess whether these effects are persistent across different kinds of violence (as measured by disaggregated homicide rates); and to assess if the findings hold across different units of areal aggregation. Therefore, an implicit goal was to compare inferences across models that do and do not treat each of the disaggregated homicide rates as having distinct data generating processes, as well as across models that do and do not allow for structures in the regression residuals.

Observed spatial patterns in the outcome can result from several forms of spatial processes. In this project, the aim was to utilize the flexibility of the information-theoretic framework in order to allow spatial structures in the regression residuals. Therefore, models with substantive spatial processes are not included here. This project does, however, examine the impacts of neighboring-area predictors on a local area's criterion of interest. In other words, in addition to modeling an area's homicide rates on its "own" level of resource deprivation, for example, this project also examines the extent to which it may be affected by "cross" or neighboring-area levels of resource deprivation.

METHODOLOGY

This project utilizes the flexibility of the Generalized Maximum Entropy (GME) and Generalized Cross Entropy (GCE) methods that are semi-parametric, information-theoretic approaches to deriving inferences from a sample. The flexibility they afford over the more traditional Maximum Likelihood methods is what allows for an easy incorporation of several forms of error structures. This includes cross-sectional models with heteroskedastic errors, models with spatially autocorrelated errors, or both. In addition, the form of error-correlation can be specified as being local, global, or global with a distance decay. The framework builds on an information-theoretic perspective of data analysis — that the sample conveys "information" about the phenomenon of interest and the aim of the researcher is to utilize all available knowledge in recovering this information in a conservative manner. Therefore, the observed data may be thought of as reducing uncertainty about the outcomes of interest as well as the errors. Building on this *uncertainty-reducing* role of the data, this project also derives a means of gauging the appropriateness of various hypothesized error structures.

The GME/GCE framework utilized in this project avoids strong distributional assumptions and models the error structures non-parametrically. Therefore, it avoids increasing the *complexity* of the information recovery task (i.e., the total number of parameters to be estimated in spatial or the non-spatial models are the same). The approach is not very resource-intensive as it does not require integration of high dimensional probabilities nor does it require the inversion of a spatial weight matrix.

The main drawback of this analytical strategy is that currently it is not available in standard software and therefore requires specialized manual programming. However, the manual programming that is needed to estimate spatial and non-spatial models of count outcomes can be done in standard and readily-available programming languages like SAS, GAUSS, etc. In addition, the ETS module of SAS is in the process of introducing a specialized procedure that is designed for the estimation of discrete outcomes with the GME/GCE framework used in this project. As such, introducing spatial-econometric capabilities to that module is possible but must await more complete and comprehensive testing of the extensions developed here.

DATA

This project analyzes homicide counts across Chicago's census tracts (CT) and alternately its neighborhood clusters (NC). The 343 neighborhood clusters in Chicago are defined by the Project on Human Development in Chicago's Neighborhoods (PHDCN) as clusters of its 865 census tracts. The mapping of the CTs to the relevant NCs was obtained from staff at the PHDCN and is used with their permission. All other data used in this project were obtained from public sources. All raw data were obtained at the CT level and then aggregated up to the NC level.

The counts of disaggregated homicide rates (1989–91) were the dependent variable in the analysis and were obtained from *ICPSR Study 6399: Homicides in Chicago, 1965–1995 (Part 1, Victim Level File)*. This data file contains detailed information on victim, offender, and circumstances of each of the homicides reported to the Chicago police between 1965 and 1995. It includes a variable SYNDROME that was used to classify the homicides into the six categories used in this project. These include homicides that were categorized as being gang related (GNG), instrumental (INS), family related expressive (FAM), known person expressive (KNO), stranger expressive (STR), and other (OTH).

In addition to information about the homicide type, this file also contains information about the geographic location of the homicide (where the body was found). In the public release version of this file, this information is only recorded as the census tract number where the homicide occurred.

This variable was used, along with the above mentioned homicide type categories, to create counts of the number of homicides observed in each of the 865 census tracts in Chicago between the years 1989-91. For the NC level analysis, they were further aggregated up to the NC level.

As one would suspect, the distribution of these disaggregated homicide rates was extremely skewed, and there were large numbers of census tracts as well as neighborhood clusters that had no homicides reported during the period being studied. In fact, the number of neighborhood clusters with no reported homicides ranged from a low of about 40% (KNO) to a high of 63% (STR) of the sample. Similarly, the number of census tracts with no homicides reported ranged from a low of 63% (KNO) to a high of 80% (STR) of the sample. In addition, visual inspection of the maps plotting the counts of homicides at the neighborhood cluster as well as the census tracts level conveyed the impression of strong clustering of outcomes across space. Though not a formal test, in order to gauge the extent and direction of spatial autocorrelation in the outcomes, simple Ordinary Least Squares (OLS) regressions were estimated for each of the dependent variables with their spatial lags as independent variables. The results from this analysis confirmed that the outcomes were in fact positively correlated across space. Additionally, this analysis suggests that the autocorrelation of the outcomes is generally stronger at the NC level than at the CT level of analysis.

The independent variables used in the analysis were also initially obtained at the census tract level and were then aggregated up to the neighborhood cluster level. All of these variables were obtained from census sources for the year 1990 (or as close as possible to it). Some census tracts had missing information on some or several predictors. In order to concentrate on the main goal of modeling spatial error-correlation, this project used simple mean imputations to replace missing values at the census tract level. That is, missing values for an independent variable in a given census tract was set equal to the mean of the non-missing values for all census tracts in the same neighborhood cluster as the census tract missing the desired information. This resulted in a sample with no missing information at the census tracts. Therefore, when aggregating to the neighborhood cluster level, no missing data imputations needed to be performed.

The independent variables used in this study were constructed in order to quantify the most commonly cited predictors of violence in this literature: social disorganization, socio-economic deprivation, demographic composition, and residential stability. Nine data elements were initially gathered and analyzed for the presence of meaningful underlying latent constructs. At both the NC and the CT levels, this exploratory analysis yielded a resource deprivation index that was then computed and used as a stand-alone variable. This data-reduction approach yielded a set of six regressors that were used in all final models. The six predictors were: resource deprivation (RESDEP), share of the area's population that was Hispanic (SHRHSP), proportion of all households in the area that were non-family (PNFH), proportion of the area's population who were young men between the ages of 15-25 (YMEN), residential stability (RESST), and the natural log of the area's total population (LPOP). These measures are described in more detail in the technical report. Despite the reduction in the dimension of the correlated data, the resulting measures still showed an amount of collinearity that is cause for concern. We were unable to create more meaningful latent constructs from the remaining data elements, however, so the analysis was finally performed on all six measures listed above.

FINDINGS

Baseline models were estimated first in order to later compare them with inferences derived from the GME/GCE models. Next, models were estimated in the GME/GCE framework for all the disaggregated homicide types and, for each type, several error structures were modeled. Each of these was gauged against the others using an information-based measure in order to assess the appropriateness of the underlying error structure. Final inferences were derived and reported from the models deemed the "best" using this criterion. In order to allow for there to be some spill-over effects of the strongest and most reliable predictors, the models were re-estimated with spatial-lags of these predictors included in the set of regressors. Once again, all forms of error structures were allowed and inferences were based only on those that were deemed the closest to the underlying process. The main findings from this set of analysis can be summarized as follows:

1. Whether or not we allow for spatial structure in the errors, there is some evidence of distinct homicide-type- and analysis-level-specific macro-processes. On the other hand, there is also evidence that resource deprivation is a strong, reliable and persistent predictor of all the homicide-types analyzed and both levels of analysis. These findings are consistent with prior research.

2. Extending traditional Poisson regression models to allow for autocorrelated structures in the errors yields some important findings. At the NC level, the differences in inferences regarding homicide-type-specific macro-processes become more pronounced. However, this finding is not replicated at the CT level. Coupled with the finding that the spatial autocorrelation in the outcomes is generally stronger at the NC level than at the CT level, this finding suggests that allowing spatial structure in the errors helps *clarify* the underlying macro-processes when the flexibility is desired but does not contaminate inferences when it is unnecessary.

3. Allowing error-structures in the models almost always yields more conservative (smaller in absolute value) but more stable (smaller standard errors) marginal effects. This is consistent with the following view of information recovery: assuming away spatial structure in the errors means the researcher may be assuming *more* than the data support. To the extent that this assumption is not supported by the data, the analysis may yield misleading inferences. Allowing some flexibility (such as in the GME/GCE approach) simply means that the sample at hand decides whether or not to use the flexibility. If the error structure hypothesized is present in the underlying data generating process, the model utilizes this flexibility and yields more conservative and more stable estimates.

4. Of all the type of structures that were permitted in the models, the data seem to favor the local first-order spatial error-correlation structure. This structure is most similar to a Spatial Moving Average (SMA) process in the errors. On the other hand, a global error-correlation structure with distance based decay would be similar to the Spatial Autoregressive (SAR) structure in the errors. The samples used in this analysis seem to favor the SMA process over the SAR.

5. There seems to be evidence of spill-over effects of the resource deprivation measure. For convenience this research used a simple SAR process with first-order spatial contiguity to

model this spill-over. Other processes may, of course, be very possible. Defining contiguity using distance bands or a fixed number of neighbors may, in some contexts, provide better fit and more meaning. Similarly, the spill-over effects may be facilitated via socio-economic distance rather than purely geographic distance. Such considerations may further allow interesting insights into distinct homicide-type-specific macro-processes.

CONCLUSIONS

The analysis conducted in this study suggests several implications for future analysis of homicide rates as well as other rare crimes.

Substantively, this analysis concludes that ignoring spatial error-correlation in models of count outcomes often yields misleading inferences. This research effort confirms that some predictors would have been erroneously deemed irrelevant and some would have been erroneously deemed relevant had the spatial structure in the errors not been allowed. Although this is a mere confirmation of what is observed in linear models that ignore spatial error correlation structures, this analysis finds that the extent of bias can be considerable in these non-linear models.

On the other hand, this research effort finds strong evidence in favor of a stable predictor like resource deprivation, which is a reliable predictor for all homicide types and at both levels of analysis conducted here. In addition, this research effort also finds a reliable, though distance-decayed, spill-over effect of resource deprivation in neighboring areas on the expected violence in the central unit. Hence, it suggests a careful consideration of the impacts of policy measures that may, for example, target resource deprivation as a means of alleviating the problem of violence. Any such policy initiatives should anticipate and account for potential benefits that not only accrue from direct "own" effects but also from indirect "cross" effects that may exist. Therefore, the impact of a city-wide policy initiatives targeted at improving resource deprivation, for example, can have an aggregate benefit larger than the sum of its benefits on each areal unit individually. In this research project, spill-over effects analyzed were found to be positive. However, the effects would be reversed had a negative spill-over effect been found. Then, the overall benefit from a city-wide initiative would be dampened. Therefore, this analysis suggests careful consideration of the spill-over effects of intervention and other policy initiatives when they are aimed at affecting outcomes across areal units that are spatially linked in some meaningful manner.

From a methodological point of view, the GME/GCE framework offers a variety of desirable benefits over fully-parametric likelihood-based methods. The most important benefit, owing to its flexibility, is the ease with which the GME/GCE framework incorporates spatial heteroskedasticity as well as autocorrelation. Although this is not always to be expected, in some of the models, the GME/GCE estimator even yielded in-sample predictive accuracy better than the Maximum Likelihood estimators.

Practically, the implementation of the GME/GCE framework currently requires manual programming in some software that allows matrix manipulation and that contains some non-linear optimization routines. The IML procedure of SAS (that was used in the project) as well as specialized modules in GAUSS are two commonly used platforms that provide these features. In terms of computer processing time, the GME/GCE solutions are not much slower to obtain than traditional non-spatial Maximum Likelihood methods.

The current research effort offers some promising avenues for future research. In this project, the spatial structure in the sample was used to model spatial error correlation. In addition, some limited use was made of the spatial structure in modeling the spill-over effect of resource deprivation on the outcomes of interest. An important type of spatial effect is where the outcome in the central areal unit is causally linked to the outcomes in neighboring areas. This would suggest a form of diffusion process. Establishing the existence of such processes using single cross-sections of data are difficult, if not impossible. However, extending the GME/GCE framework to model other forms of substantive spatial processes, such as the so-called *simultaneous* models, is a promising area of future research. In addition, extending the GME/GCE to allow for both spatial, temporal, or spatio-temporal processes offers, given its flexibility, many possibilities for additional research. Other areas of research in which the GME/GCE framework could be used include the incorporation of a population-at-risk correction and the extraction of mixed processes, such as the zero-inflated Poisson models. The ability to estimate these models while utilizing all the flexibility of the GME/GCE framework to model error structures promises to allow robust estimation of models of rare crimes at local levels of areal aggregation.

— ◆◆◆ —