RESEARCH REPORT

March 2004

Robust Spatial Analysis of Rare Crimes

Technical Report Submitted to the Mapping and Analysis for Public Safety (MAPS) Program, National Institute of Justice (NIJ)

Avinash Singh Bhati

URBAN INSTITUTE
Justice Policy Center

URBAN INSTITUTE
Justice Policy Center

2100 M Street NW
Washington, DC 20037
www.urban.org

This report was prepared under Grant 2002-IJ-CX-0006. Opinions expressed in this document are those of the authors, and do not necessarily represent the official position or policies of the U.S. Department of Justice, the Urban Institute, its trustees, or its funders.

JPC Publication: CS03 0100

Robust Spatial Analysis of Rare Crimes

Technical Report Submitted to the
Mapping and Analysis for Public Safety Program
National Institute of Justice

Avinash Singh Bhati
abhati@ui.urban.org / (202) 261-5329

Justice Policy Center, The Urban Institute
2100 M Street, N.W., Washington, D.C. 20037
March 2004

Abstract

Research Goals and Objectives: 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.

Research Design and Methodology: 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.

Research Results and Conclusions: 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 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.

Contents

List of Tables

List of Figures

Acknowledgements

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.

All data used in this study were obtained from public sources with the exception of information needed for mapping Chicago's 865 census tracts to the 343 neighborhood clusters defined by the Project on Human Development in Chicago's Neighborhoods (PHDCN). I thank Dr. Robert Sampson from the PHDCN for providing me with this information.

I thank Dr. Amos Golan (Department of Economics, American University), who served as a consultant on this project, for his guidance and extremely valuable contributions that have helped shape the analysis reported here. I thank Dr. Nancy LaVigne, Dr. Caterina Roman, Dr. Dan Mears and Vera Kachnowski from the Urban Institute for helpful comments on an earlier version of this report as well as participants at the XIth Annual International Crime Mapping Research Conference for valuable input. All remaining errors are mine.

Executive Summary

Major advances have been made in testing and estimating regression models in the presence of spatially-correlated errors when the dependent variables are continuous. Doing so when the criterion measures are discrete has proven to be more challenging. Unfortunately, when rates or counts of relatively rare crimes are analyzed at local (intra-city) areal units like neighborhoods and communities, discrete outcomes are the norm rather than the exception. This report describes one approach that allows researchers to model the effects of explanatory variables on discrete outcomes of interest while allowing spatial structure in the residuals.

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.

Chapter 1

Introduction

Multivariate regression analysis is a common technique researchers use to explain observed patterns in outcomes—the dependent variables—with theoretically provided predictors—the independent variables. Despite heavy reliance on this approach by researchers and its ubiquitous use as a tool to inform policy, the approach is built on several strong assumptions which, in practice, may not hold true. One such assumption is that the *unexplained* variation in the outcome, as captured by the regression residuals or errors, should be devoid of any structure. When modeling patterns of crime across geographic space, however, this assumption is very likely to be violated. To the extent that it is, and we proceed as if it were not, the resulting inferences may be misleading and, hence, a poor foundation on which to base policy.

Major advances have been made in estimating regression models in the presence of spatially autocorrelated errors when the dependent variable is continuous (Anselin, 1988; Anselin and Bera, 1998). But doing so when the criterion (or outcome) measure is discrete has proven to be more difficult and is currently an area of active research. This focus is important because when rates or counts of relatively rare crimes are analyzed at local (intra-city) levels of areal aggregation, such as census tracts or neighborhoods, discrete outcomes are usually the norm. This report describes and applies an information-theoretic method that allows for the incorporation of spatially dependent error structures in regression models of discrete outcomes.

1.1. BACKGROUND

The homicide rate is an example of a crime the qualifies as being "rare," or one that is discrete in nature, when it is measured at local (intra-city) levels of aggregation. By discrete we mean that the observed outcomes are typically low non-negative integers values (e.g., $0, 1, 2, \ldots$).

Researchers have attempted to explain observed cross-sectional variations in homicide rates using macro-structural covariates at various levels of areal aggregation. These in-

clude nations (Braithwaite and Braithwaite, 1980), states (Kennedy, Silverman and Forde, 1991), counties (Land, McCall and Cohen, 1990), Metropolitan Statistical Areas (Balkwell, 1990), cities (William and Flewelling, 1988), and neighborhoods (Baller et al., 2001; Kubrin, 2003). Motivated primarily by social disorganization, strain, and social capital theories of crime, researchers typically seek to establish links between the structural, economic and social conditions in areal units with the rates or counts of homicide observed there. See, among others, Land, McCall and Cohen (1990) and Reiss and Roth (1994) for comprehensive reviews.

Additionally, to isolate and identify the macro-processes leading to different *types* of violence, researchers sometimes estimate models with disaggregated homicide rates with varying bases for the disaggregation. These include disaggregation by race (Cubbin, Pickle and Fingerhut, 2000; Parker and McCall, 1999), intimacy (Avakame, 1998), gender within intimate partner homicides (Felson and Messner, 1998) and homicide types (Williams and Flewelling, 1988; Rosenfeld, Bray and Egley, 1999; Kubrin, 2003).

At higher levels of areal aggregation (e.g., counties or states), when the number of homicides is sufficiently large and non-zero events are observed in most of the sampled units, the outcome may be considered continuous, and traditional spatial analytical methods can be, and have been, applied (Messner et al., 1999; Baller et al., 2001).

As the unit of analysis becomes smaller, however, four things can be expected. First, the number of outcomes observed in each sampled unit decreases, thereby discretizing the criterion variable (i.e., the variable approaches a count measure with a highly skewed distribution). Second, the number of units with zero counts increases, thereby inflating the outcome's distribution at zero. Third, differences in the number of outcomes (e.g., homicides) that could have been observed in sampled units, simply because of differences in the populations-at-risk of experiencing the event, become more pronounced. Finally, explanatory macro-characteristics of areal units like neighborhoods, census tracts, etc., may be more volatile over time than those for larger aggregations like counties, states, etc. Therefore, increasing the counts of the rare crimes at neighborhood or local levels by simply counting over extended periods of time may lead to distorted inferences and may mask true data generating processes.

Of course, the measurement problems noted above relate to any rare crimes. In these instances, Poisson-based regression models are usually more appropriate to employ (Osgood, 2000). However, the existing spatial analytical toolkit readily available to researchers is not directly applicable to these types of non-linear models. In order to study the processes that generate spatial distributions of rare crimes like homicide, therefore, researchers often aggregate over larger areal units, across several types of homicides, or over longer time periods and rely on spatial analytical methods developed for continuous criterion measures. Even when the discrete nature of the criterion measure is explicitly recognized, researchers are often forced to rely on a two-stage approach — (1) convert this measure into an *approximately* continuous variable, and (2) apply traditional spatial analytical methods. Inferences derived from these models, however, could be misleading as they rely on ad-hoc transformations based, more often than not, on mathematically convenient assumptions rather than

a conservative incorporation of only the limited knowledge that researchers typically have about the underlying data generating processes.

1.2. AREAL ANALYSIS OF RARE CRIMES

The incorporation of a spatial dimension in applied work is now a fairly routine component of homicide research (Messner and Anselin, 2003). Although the theoretical basis for linking crime to place can be derived from several well-established sociological and ecological perspectives on crime, deviance and victimization, a proliferation of user-friendly software and geocoded crime data has sparked this recent shift in applied work (Anselin et al., 2000). Borrowing insights from applied research in other fields involving spatially "labeled" data, researchers analyzing violent crime are now well aware that the assumption of independence across geographic space is questionable. Irrespective of whether the notion of space is used for exploratory spatial data analysis (ESDA), for studying spatial diffusion processes and spill-over effects, or simply for addressing inefficiencies introduced by error-dependence across space, spatial econometric methods have proven useful in clarifying the links between macro-covariates and homicide rates (Baller et al., 2001; Morenoff, Sampson and Raudenbush, 2001) or in modeling other rare crimes (Smith, Frazee, and Davison, 2000). As such, they are indispensable tools for the applied researcher interested in studying and explaining the spatial patterning of crime.

With the wealth of geocoded data increasingly becoming available at local levels both from census sources and from primary data collection efforts, researchers analyzing homicides as well as other rare crimes are more frequently confronted with the need to apply spatial econometric methods to data that are discrete by nature. This proves challenging for several reasons.

There are, currently, no well-established and universally accepted (or available) estimation methods for such models. For example, to analyze binomial or multinomial choice outcomes researchers have proposed several likelihood-based estimators (Besag, 1974; Case, 1992; McMillen, 1992; Bolduc, Fortin and Gordon, 1997; Heagerty and Lele, 1998) as well as a generalized method of moments estimator, making use of the concept of a generalized residual (Pinske and Slade, 1998). To analyze count data researchers rely mainly on Bayes, Empirical Bayes, or Maximum-Likelihood estimators of the so-called Poisson "auto-models" (Besag, 1974; Clayton and Kaldor, 1987; Cressie and Read, 1989; Cressie and Chan, 1989; Kaiser and Cressie, 1997; Waller et al., 1997). These estimators, though feasible, can be extremely resource-intensive to implement. Simulation based estimators are an attractive alternative (LeSage, 1999) and are an area of active research (although they too can be resource-intensive).

Unfortunately, the small-sample properties of many of these approaches are not well established. When dealing with intra-city areal units, like neighborhoods, one may be restricted to less than a hundred or a few hundred observations that are spatially dependent. Such sample sizes, coupled with highly collinear (ill-conditioned) data, reduce the efficiency of estimates. Moreover, fully-parametric methods that rely on a likelihood function,

including traditional Bayesian methods, need to invoke some form of strong distributional assumption *apriori*. In non-experimental settings, such as the social sciences, true underlying data generating processes are seldom, if ever, known apriori. Hence, model estimates and inferences derived from them can be extremely sensitive to distributional assumptions.

In practice, researchers typically resort to the two-stage approach described above whereby the criterion variable is first converted into an *approximately* continuous measure and then traditional spatial analytical techniques are applied to it. Though usually feasible, it is unclear whether these transformations always yield their desired corrections (Bailey and Gartell, 1995, pg.277). When analyzing rare crimes like homicides, for example, logarithmic, rate, or Freeman-Tukey type transformations may not yield the desired Gaussian criterion measures. This is especially true when rare events like homicides are measured at local levels of areal aggregation and the number of units with zero counts may be large.

Alternatively, proxy measures—estimated using first-stage non-spatial models—are sometimes used to either redefine the criterion measure directly (Morenoff, Sampson and Raudenbush, 2001) or to statistically "control" for spatial dependence in the second stage (Kubrin, 2003). In finite samples, however, there is no guarantee that the first-stage estimates used in such two-step approaches have the assumed desirable properties. These proxy measures must, by definition, inherit any good/bad properties from a mis-specified first stage.

The Generalized Cross Entropy (GCE) method applied in this project is a flexible, semi-parametric method that avoids strong parametric distributional assumptions. Additionally, rather than parameterize the spatial error-dependence in terms of a few autoregressive parameters, the methodology uses knowledge of the relative spatial positioning or contiguity of the sampled units to guide the structure in the errors—each of which is individually estimated. Despite the simultaneous estimation of the realized errors and the expected outcomes in one step, however, the *complexity* of the information-recovery problem does not increase in the sense that the number of unknowns that need to be estimated remains fixed. Finally, the approach is not very resource-intensive to implement as it does not require high dimensional integration of marginal probabilities nor does it require the inversion of the spatial weight matrix.

1.3. OVERVIEW OF THIS REPORT

The main goal of the 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. As described in this chapter, real-world problems 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. Chapter 2 provides a detailed description of the semi-parametric information-theoretic framework that achieves this goal.

To demonstrate the application of this analytical framework to a real-world issue, a second stated goal of this project was to examine and report the extent to which structural

and socio-economic determinants of different kinds of violence (the disaggregated homicide types) may be distinct, and whether these findings persist at different (intra-city) levels of areal aggregation. More specifically, this research aimed to examine the following questions:

- How do the socio-economic and demographic characteristics in an area affect the amount of violence that that community can expect to experience?
- Are the links violence-type-specific?
- Are these links areal aggregation level-specific, i.e., is there evidence of a modified areal unit problem (MAUP)?

Given the explicit spatial nature of the data needed to answer these questions, they need to be examined in the presence of possibly spatially-dependent errors. The information-theoretic approach developed in this project is used for this purpose. A final question this project aims to answer is whether the inferences derived from the analysis would have been *qualitatively* different had the possible spatial structure in the errors been ignored and traditional non-spatial modeling strategies been used. The data that are used to examine all these issues are described in Chapter 3, with findings discussed in Chapter 4.

Chapter 5 discusses implications of this research effort, lists promising extensions of the proposed analytical framework, and enumerates the merits and drawbacks of the methodology.

Chapter 2

Methodology

This chapter describes the information-theoretic method used in this project. It first explains one application of the Generalized Cross Entropy (GCE) framework, introduced in Golan, Judges and Miller (1996), that may be used for modeling count outcomes under ideal and asymptotic conditions and then explains its finite sample version. Next, it describes how the framework may be extended to allow for spatially-dependent and/or heteroskedastic errors and suggests how the appropriateness of various specifications may be gauged. It concludes with a discussion of substantive spatial-dependence and how the current formulations may be extended to model it.

2.1. INTRODUCTION

Consider, as a point of departure, the following basic identity

$$y_n = s_n + e_n \qquad \forall n = 1, 2, \ldots, N \tag{2.1}$$

where y_n is an observed outcome, s_n is the pure emitted signal or the expected outcome, and e_n is a random noise term that distorts the signal into the observed outcome. The observed outcome may be a binary choice, a non-negative integer, or the more traditional continuous outcome. The signal and noise terms are both unknowns and the challenge for researchers is to use all available knowledge to recover information about them.

Knowledge about the process that is available to the researcher includes a set of exogenous predictors, as suggested by theory, as well as the relative spatial positioning of the observations in the sample. For the former, let us assume that a set of K characteristics exists for each of the n sampled units, recorded as x_{kn}, that are hypothesized to influence the signal. For the latter, we may have knowledge about the spatial proximity of observations to each other, i.e., a measure of the spatial structure in the sample. Let this knowledge be available in the form of a full $N \times N$ matrix. For example, we may have actual measures of the Euclidean distance, or some other measures like social or economic distance, between

all units in the sample. Or we may simply have knowledge about the contiguity of these observations. Ignoring knowledge of the spatial structure in the sample when it is available is therefore an inefficient way of recovering information from the sample.

Observed spatial patterns in the outcomes (y_n) result from patterns in the signals (s_n) that are modeled by researchers. Bad, incomplete, or inappropriate model specifications, however, often yield residuals (e_n) with spatial patterns. Excluding important predictors, for example, will result in the spatial pattern in the outcomes being inherited by the regression residuals. Included regressors that are badly measured, or measured at inappropriate areal units, will have the same effect. Irrespective of its cause(s), the resulting spatial dependence in the errors proves to be a "nuisance" (Anselin and Bera, 1998) when the primary goal is to recover the signals accurately. This chapter explains one way to tackle this problem from an information-theoretic perspective building on the Generalized Cross Entropy (GCE) framework introduced in Golan, Judge and Miller (1996).

2.2. THE GENERALIZED CROSS ENTROPY (GCE) APPROACH

2.2.1. Setting up the basic problem

First consider the case where no information is available about the spatial positioning of the sampled units, i.e., all we have available are the observed outcomes and an exogenous set of possible predictors. We wish to use these data to estimate the signal in order to assess how the hypothesized predictors influence it. To do so within the information-theoretic framework, we first need to re-parameterize all unknowns (i.e., the signal and noise terms) into well defined probabilities. For example, we can define each signal as

$$s_n = \mathbf{z}'\mathbf{p}_n = z_1 p_{1n} + z_2 p_{2n} + \cdots + z_L p_{Ln} \qquad \forall n \tag{2.2}$$

where $\mathbf{z} = (z_1, z_2, \ldots, z_L)'$ is a column vector of real numbers that spans the range of possible values that the signal can take (with $z_1 < z_2 < \cdots < z_L$) and $'$ represents a vector or matrix transpose. Knowledge about bounds on the signal is typically available to the researcher based on knowledge of the dependent variable. For example, the range of possibilities of a binary choice outcome is $s_n \in (0, 1)$; that of a count outcome is $s_n \in (0, +T)$; and for traditional unlimited continuous outcomes, it is $s_n \in \pm T$. Here T is a *sufficiently* large number—one that contains the true signal. If its value is unknown, we may assume a value for T large enough so that it contains, at the very least, the observed outcomes in the sample. The points in $\mathbf{z}$ need to be equally spaced unless prior knowledge to the contrary is available to the researcher. This sequence of support points constitutes an L-dimensional *signal support space*. To these support points are applied a set of proper probabilities that, once estimated, will yield the expected outcome of interest—the signal. By *proper* we mean that these probabilities are non-negative, i.e., $p_{ln} > 0 \; \forall n, l$, and sum to unity, i.e., $\sum_l p_{nl} = 1$ $\forall n$.

The re-parameterization of the noise term can be carried out in a similar way. That is,

we can define each error as

$$e_n = \mathbf{v}'\mathbf{w}_n = v_1 w_{1n} + v_2 w_{2n} + \cdots + v_M w_{Mn} \qquad \forall n \tag{2.3}$$

where $\mathbf{v}$ is an M-dimensional *error support space* that is defined as being symmetric about zero and the probabilities applied to it are proper. Choosing bounds on the error support requires sample-specific consideration that will be discussed in more details later. Unless specific knowledge exists to the contrary, the error support points are equally spaced and symmetrical about 0. These definitions allow us to write a re-parameterized version of (2.1) as

$$y_n = \mathbf{z}'\mathbf{p}_n + \mathbf{v}'\mathbf{w}_n \qquad \forall n = 1, 2, \ldots, N \tag{2.4}$$

where no assumptions are made about the probabilities of interest other than that they are proper in the sense described above.

Next, we wish to use the available data in constraining the values that these probabilities can take. In the current setting, a natural method of introducing exogenous information is through the formulation of moments. Suppose we pre-multiply both sides of (2.4) by x_{kn} and then sum the resulting products over the entire sample. This would yield the following K moment equations[1]

$$\sum_n x_{kn} y_n = \sum_n x_{kn} \mathbf{z}'\mathbf{p}_n + \sum_n x_{kn} \mathbf{v}'\mathbf{w}_n \qquad \forall k = 1, 2, \ldots, K \tag{2.5}$$

along with the N adding-up constraints that we would like to impose on each set of proper probabilities

$$\sum_l p_{ln} = \sum_m w_{mn} = 1 \qquad \forall n = 1, 2, \ldots, N. \tag{2.6}$$

So far we have only re-parameterized the unknowns into proper probabilities and have imposed moment restrictions on them using all the available predictors we have. Under *ideal* experimental conditions, where it may be reasonable to assume that the errors have no structure to them, that the regressors are measured without error, and that the model is well-specified, it may be reasonable to assume that the regressors $\{x_{kn}\}$ are completely uncorrelated with the errors e_n. If so, then we can make the strong orthogonality assumptions of

$$\sum_n x_{kn} e_n = \sum_n x_{kn} \mathbf{v}'\mathbf{w}_n = 0 \qquad \forall k = 1, 2, \ldots, K \tag{2.7}$$

so that the moment constraints of (2.5) are reduced to

$$\sum_n x_{kn} y_n = \sum_n x_{kn} \mathbf{z}'\mathbf{p}_n \qquad \forall k = 1, 2, \ldots, K. \tag{2.8}$$

[1]Note that if we were dealing with multinomial choice outcomes with J choices, then there would be one set of constraints for each of the choices resulting in $K \times J$ moment equations.

2.2.2. Estimation

Having restricted the moment constraints to be pure (noiseless), where the observed (sample) moments are *exactly* matched to the expected (population) moments, we see that the only unknowns remaining in the problem are the probabilities $\mathbf{p}_n$. Now we have a set of K moment equations (2.8) and a set of N adding-up equations (2.6) constraining a set of $N \times L$ probabilities. This is an ill-posed inversion problem (Levine, 1980) with more unknowns than equations linking them and, as such, an infinite number of solutions can satisfy the constraints. How do we select one out of these infinite solutions? Faced with such a problem in statistical mechanics, Jaynes (1957a; 1957b) proposed maximizing the uncertainty implied by the probabilities as a means of selecting an optimal solution. In other words, from all the probability vectors that satisfy the moment and adding-up constraints, Jaynes proposed selecting the one that implies maximum uncertainty. That way the recovered information will be as conservative as the data allow it to be. Put another way, the optimal solution chosen should be the one that only "just" satisfies the constraints required of it.

The next obvious question then is, "How does one measure (quantify) uncertainty?" In the context of a problem in communication theory, Shannon (1948) had defined the uncertainty contained in a message with J mutually exclusive and exhaustive outcomes as $H(\mathbf{p}) = -\sum_j p_j \ln p_j$. This quantity, termed *Information Entropy* by Shannon, is maximized when all possibilities are equally likely (probable), i.e., when $p_j = 1/J \ \forall j$, and is at a minimum of 0 when any one of the possibilities is certain, i.e., when $p_j = 1$ for some j and zero for the rest. Entropy derived from two sources of uncertainty are additive *only* if they are independent sources of uncertainty (Shannon, 1948). In what came to be known as the Maximum Entropy formalism, Edwin Jaynes proposed to use this measure—Entropy—as the criterion function to maximize, subject to all available constraints, in order to derive conservative inferences from a sample.[2]

Applying this principle to our pure moment constraints case, the resulting information-recovery task is formulated as a constrained optimization problem that can be written as

$$\max_{\mathbf{p}} \quad H(\mathbf{p}) = -\mathbf{p}' \ln \mathbf{p} = -\sum_{ln} p_{ln} \ln p_{ln} \tag{2.9}$$

subject to (2.8) and (2.6) where $\mathbf{p} = (\mathbf{p}_1', \mathbf{p}_2', \ldots, \mathbf{p}_N')'$. Since Entropy is additive only for independent sources of uncertainty, an implicit assumption in (2.9) is that the signals are independent across sample units.

If, in addition to the moment constraints, we have additional non-sample information about the signals in the form of prior probabilities $\{p_{nl}^0\}$ that have the same dimension and are defined on the same space as the posteriors $\{p_{nl}\}$, then an equivalent problem is to minimize the informational distance between the prior and the posterior probabilities. Unlike the Maximum Entropy (ME) approach, where we maximize uncertainty implied by the proba-

[2]There is a growing literature dealing with Information and Entropy Econometrics that builds on this view of data analysis. See Golan (2002) and other articles in that special issue of the *Journal of Econometrics* for recent theoretic and applied work in this field.

bilities, in the Minimum Cross Entropy formalism, we minimize the Cross Entropy (CE), or the Kullback-Leibler (KL) informational distance (Kullback, 1959), between the posterior probabilities and their priors. For a message with J mutually exclusive and exhaustive outcomes with prior probabilities p_j^0, the KL informational distance or the CE is defined as $CE = \sum_j p_j \ln(p_j/p_j^0)$. Therefore, given prior probabilities $\mathbf{p}^0$, the resulting constrained optimization problem is to

$$\min_{\mathbf{p}} \quad CE(\mathbf{p};\mathbf{p}^0) = \mathbf{p}' \ln(\mathbf{p}/\mathbf{p}^0) = \sum_{ln} p_{ln} \ln(p_{ln}/p_{ln}^0) \tag{2.10}$$

subject to (2.8) and (2.6).

The ME formulation is a special case of the minimum CE problem when the priors in the latter are forced to be uniform. Therefore, in what follows we will restrict our derivations and explanations to only the minimum CE formulations.

The CE problem is a constrained minimization problem that can be solved analytically using the method of Lagrange. The primal Lagrangian ($\mathcal{L}_{CE}^P$) for this problem is set up as

$$\mathcal{L}_{CE}^P = \sum_n \mathbf{p}_n' \ln(\mathbf{p}_n/\mathbf{p}_n^0) + \sum_k \lambda_k \left\{ \sum_n x_{kn} y_n - \sum_n x_{kn} \mathbf{z}' \mathbf{p}_n \right\} + \sum_n \mu_n \{1 - \mathbf{1}' \mathbf{p}_n\} \tag{2.11}$$

where $\{\lambda_k\}$ and $\{\mu_n\}$ are the sets of Lagrange multipliers corresponding to the imposed constraints.

Solving the first order conditions for this optimization problem analytically we obtain optimal solutions for the probabilities of interest as

$$\hat{p}_{ln} = \frac{p_{ln}^0 \exp(z_l \sum_k x_{kn} \hat{\lambda}_k)}{\sum_l p_{ln}^0 \exp(z_l \sum_k x_{kn} \hat{\lambda}_k)} = \frac{p_{ln}^0 \exp(z_l \mathbf{x}_n' \hat{\boldsymbol{\lambda}})}{\Omega_n} \qquad \forall n, l \tag{2.12}$$

where $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K)'$ are the optimum Lagrange multipliers corresponding to the K data constraints, $\mathbf{x}_n = (x_{n1}, x_{n2}, \ldots, x_{nK})'$ is a vector of K covariates for the nth observation, and the partition function (Ω_n) ensures that the probabilities sum to one. Inserting these optimum solutions back into the primal constrained optimization problem of (2.11) we can derive a dual unconstrained version of the optimization problem, where the dual objective is a function of the Lagrange multipliers, as

$$\mathcal{L}_{CE}^D = \sum_{kn} x_{kn} y_n \lambda_k - \sum_n \ln \Omega_n. \tag{2.13}$$

This unconstrained dual optimization problem typically does not have an analytical solution, but a numerical one can be obtained using optimization techniques available in a variety of software. Once we obtain optimum values for the Lagrange multipliers we can then recover the signals, or expected outcomes, using (2.12) and (2.2).

2.2.3. Noisy moment constraints

In the optimization problem derived above, strong assumptions were made regarding the exact matching of the observed (sample) moments and the expected (population) moments. In most real-world non-experimental settings, e.g., almost all of social science research, the sample moments may not be perfect analogs of the population moments being estimated. Hence, making strong assumptions when they are possibly violated may yield inferior results. In order to impose less restrictive constraints than those implied by (2.8), we need to allow some flexibility in the formulation of the moment constraints. Rather than force $\sum_n x_{kn}e_n = 0\ \forall k$ as in (2.7), one way to allow this flexibility is to require only that the cross products *shrink* to 0 as the sample size increases. Following the traditional consistency requirement, plim $\frac{1}{N}\sum_n x_{kn}e_n \rightarrow 0\ \forall k$, which underlies most likelihood-based estimators, we use N as a shrinkage factor explicitly in our formulation. That is, we replace the pure (noiseless) moment constraints of (2.8) with

$$\sum_n x_{kn}y_n = \sum_n x_{kn}\mathbf{z}'\mathbf{p}_n + \frac{1}{N}\sum_n x_{kn}\mathbf{v}'\mathbf{w}_n \qquad \forall k = 1, 2, \ldots, K. \tag{2.14}$$

This formulation implies that, in any finite sample, we do not force the moments to hold exactly. The amount of flexibility *allowed* in the constraints depends on the sample size and the specification of $\mathbf{v}$.[3] The amount of flexibility *used* by the estimator, however, depends on the observed data. The trivial solution $w_{mn} = 1/M\ \forall n, m$ or $e_n = 0\ \forall n$ is within the allowable solutions for the unknown error terms. That is, if the sample is close to being *ideal*, then it should only help reduce uncertainty about the signal and $e_n \approx 0, \forall n$. If the sample is *imperfect* in any sense, then by allowing some flexibility in the constraints we allow for a more stable solution to the optimization problem. In other words, loosening the constraints does not *force*, but rather *allows*, the solutions to be different from those obtained by exactly matching the observed and expected moments. The solutions will be different to the extent that the observed data do not strictly conform to the assumption of $\sum_n x_{kn}e_n = 0\ \forall k$.

Unlike the pure (noiseless) moments constraints cases of the previous section, the problem is now defined in terms of two sources of uncertainty — relating to $\mathbf{p}_n$ and $\mathbf{w}_n$. Following the Maximum Entropy or Cross Entropy formalisms as described above, we can now set up a more *generalized* information recovery problem. This approach, aptly termed the Generalized Maximum Entropy (GME) or the Generalized Cross Entropy (GCE) method, was introduced by Golan, Judge, and Miller (1996). As before, the GME formulation is but a special case of the GCE when the prior probabilities for $\mathbf{p}_n$ as well as $\mathbf{w}_n$ are forced to be

[3] With a binary choice outcome, for example, where the signal and noise terms are both $\in (0, 1)$, the noise terms must be $\in \pm 1$. Therefore, a simple definition of the noise support space would be to let $\mathbf{v} = (-1, +1)'$. The definition of the error support space will vary by context or application. A more detailed discussion of the definition of support spaces used in this project is provided in Section 2.5.

uniform. The primal constrained optimization problem now is to

$$\min_{\mathbf{p},\mathbf{w}} \quad CE(\mathbf{p},\mathbf{w};\mathbf{p}^0,\mathbf{w}^0) = \mathbf{p}'\ln(\mathbf{p}/\mathbf{p}^0) + \mathbf{w}'\ln(\mathbf{w}/\mathbf{w}^0) \tag{2.15}$$

subject to the flexible moment constraints of (2.14) and the adding-up constraints of (2.6). Note that despite the flexible constraints, the assumption of independence between the signal and noise terms and across sample units must be maintained in order for the Cross Entropy from each of these sources of uncertainty to be additive. Following through with the optimization we obtain

$$\hat{p}_{ln} = \frac{p^0_{ln}\exp(z_l\sum_k x_{kn}\hat{\lambda}_k)}{\sum_l p^0_{ln}\exp(z_l\sum_k x_{kn}\hat{\lambda}_k)} = \frac{p^0_{ln}\exp(z_l\mathbf{x}_n'\hat{\boldsymbol{\lambda}})}{\Omega_n} \qquad \forall n,l \tag{2.16}$$

and

$$\hat{w}_{mn} = \frac{w^0_{mn}\exp(v^*_m\sum_k x_{kn}\hat{\lambda}_k)}{\sum_m w^0_{mn}\exp(v^*_m\sum_k x_{kn}\hat{\lambda}_k)} = \frac{w^0_{mn}\exp(v^*_m\mathbf{x}_n'\hat{\boldsymbol{\lambda}})}{\Psi_n} \qquad \forall n,m \tag{2.17}$$

as the optimum solutions for the probabilities of interest, where $v^*_m = v_m/N$. Once again, these solutions may be used along with the primal Lagrangian function to derive a dual unconstrained optimization problem in the unknown Lagrange multipliers as

$$\mathcal{L}^D_{GCE} = \sum_{kn} x_{kn}y_n\lambda_k - \sum_n \ln\Omega_n - \sum_n \ln\Psi_n \tag{2.18}$$

where Ω_n and Ψ_n are the partition functions for the two sets of probabilities.

2.3. NONSPHERICAL ERRORS

In the preceding sections we described how the noiseless moment constraints can be relaxed for any finite sample and how these constraints, along with the ME/CE principle can be used to recover information from a sample of observed data. The resulting flexibility, however, still relies on the assumption that the errors are uncorrelated or that they are determined independently of one another. To the extent that we believe there to be some structure in the errors, we need to explicitly use this knowledge in recovering information from the sample.

Consider the case where this structure is explicitly known, i.e., let the *known* error covariance matrix be denoted as $\sigma^2\boldsymbol{\Phi}$ where $\boldsymbol{\Phi}$ is a $N \times N$ positive definite matrix. Then, writing the identity (2.1) in matrix notation as

$$\mathbf{y} = \mathbf{s} + \mathbf{e} \tag{2.19}$$

and setting $\mathbf{e} = \boldsymbol{\Phi}\mathbf{u}$ we obtain a new identity

$$\mathbf{y} = \mathbf{s} + \boldsymbol{\Phi}\mathbf{u} \tag{2.20}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$ and $\mathbf{s} = (s_1, s_2, \ldots, s_N)'$ are as defined before. The new set of

errors ($\mathbf{u}$) are now assumed to be completely devoid of structure, although they combine with each other in a systematic way (coded in $\mathbf{\Phi}$) to create signal distortion. Even with this knowledge we are unable to use the exogenous data $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)'$ to create exact moment constraints because

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{s} + \mathbf{X}'\mathbf{\Phi}\mathbf{u} \tag{2.21}$$

and assuming $\mathbf{X}'\mathbf{u} = \mathbf{0}$ does not necessarily imply $\mathbf{X}'\mathbf{\Phi}\mathbf{u} = \mathbf{0}$. Therefore, we typically first transform the problem by pre-multiplying both sides of the equality (2.20) by $\mathbf{\Phi}^{-1}$ to get

$$\mathbf{\Phi}^{-1}\mathbf{y} = \mathbf{\Phi}^{-1}\mathbf{s} + \mathbf{u} \tag{2.22}$$

and then, using the orthogonality assumption of $\mathbf{X}'\mathbf{u} = \mathbf{0}$, we obtain noiseless moment constraints:

$$\mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{y} = \mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{s}. \tag{2.23}$$

If we make the General Linear Model assumption of $\mathbf{s} = \mathbf{X}\boldsymbol{\beta}$, then the above moments yield the Generalized Least Squares estimates of $\boldsymbol{\beta}$ (Judge et al., 1988, pg.330). If $\mathbf{\Phi}^{-1}$ is unknown but can be consistently estimated in the first of a two-stage procedure, then this results in the Feasible or Estimated Generalized Least Squares Estimator. Nothing precludes us from applying this approach in the non-linear case (see Mittelhammer, Judge and Miller [2000], pg.361–368) or the case where the signal is left unspecified as in our formulation. However, there are several practical difficulties in applying this approach. First, the error covariance structure $\mathbf{\Phi}$ is seldom, if ever, known. Second, even if we use a two-stage procedure, there is no guarantee that the estimated $\hat{\mathbf{\Phi}}$ will be a positive definite matrix and therefore invertible. Creative assumptions often need to be made to ensure that it is invertible. The most common approach, of course, is to explicitly parameterize the entire error structure in terms of a few parameters and estimate them simultaneously with the signals. This fully-parametric framework, though well developed for the linear model, is less tractable when estimating non-linear models like binary choices or count outcomes.

In what follows, the GCE framework of the previous section is extended to allow for errors that may be heteroskedastic and/or autocorrelated across space.

2.3.1. GCE with heteroskedastic errors

In the GCE formulation of the last section, fixed weights of $1/N$ were applied to each of the observations in the sample while formulating the flexible moment constraints of (2.14). To allow for heteroskedasticity of an unspecified form we can replace the fixed weights of $1/N$ by an unknown weight π_n that is allowed to vary across the cross-sectional units. The π_n are now an additional set of proper probabilities that need to be estimated. Again, by proper we mean that $\pi_n > 0 \; \forall n$ and $\sum_n \pi_n = 1$.

If the observed sample supports *equal* weighting of the errors, then we should obtain estimates of $\hat{\pi}_n \approx 1/N \; \forall n$. If, on the other hand, the data support *unequal* weighting of the sample errors, then $\hat{\pi}_n$ should be different for some or all n. The resulting GCE estimates will, therefore, be consistent with "optimally re-weighted errors" and can, hence, be con-

sidered heteroskedastic consistent. This approach is similar to the optimal re-weighting of the estimating equations allowed in information-theoretic and empirical likelihood-based methods recently discussed by Imbens, Spady and Johnson (1998).

The resulting heteroskedastic consistent flexible moment constraints can be written as

$$\sum_n x_{kn} y_n = \sum_n x_{kn} \mathbf{z}' \mathbf{p}_n + \sum_n x_{kn} \pi_n \mathbf{v}' \mathbf{w}_n \qquad \forall k = 1, 2, \ldots, K. \tag{2.24}$$

with the added requirement that $\sum_n \pi_n = 1$. These constraints are similar to the flexible constraints of (2.21) with $\mathbf{\Phi}$ defined as a diagonal square matrix with π_n as its nth diagonal element. To proceed, note first that, assuming independence between π_n and w_{mn}, we can define an auxiliary joint probability measure $q_{mn} = w_{mn} \cdot \pi_n$ so that $\pi_n = \sum_m q_{mn}$ $\forall n$ and $w_{mn} = q_{mn} / \sum_m q_{mn}$ $\forall m, n$ are marginal and conditional probabilities (respectively) derivable from q_{mn}. In addition, since $\sum_n \pi_n = 1$ and $\sum_m w_{mn} = 1$ $\forall n$, then, $\sum_{mn} q_{mn} = 1$ *over* all m and n. Finally, given the above definition of q_{mn}, we may specify its associated priors as $q^0_{mn} = \pi^0_n \cdot w^0_{mn}$. Unless, knowledge to the contrary is available to the researcher, these priors are defined as being uniform and therefore $q^0_{mn} = 1/(M \cdot N)$.

Reformulating the task from one of recovering $\mathbf{p}$, $\boldsymbol{\pi}$ and $\mathbf{w}$ into one of recovering $\mathbf{p}$ and $\mathbf{q}$, the resulting constrained optimization problem that incorporates all available knowledge is

$$\min_{\mathbf{p},\mathbf{q}} \quad CE(\mathbf{p}, \mathbf{q}; \mathbf{p}^0, \mathbf{q}^0) = \mathbf{p}' \ln(\mathbf{p}/\mathbf{p}^0) + \mathbf{q}' \ln(\mathbf{q}/\mathbf{q}^0) \tag{2.25}$$

subject to the flexible moment constraints

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{Z}\mathbf{p} + \mathbf{X}'\mathbf{V}\mathbf{q} \tag{2.26}$$

and the adding-up constraints of $\sum_l p_{ln} = 1$ $\forall n$ and $\sum_{mn} q_{mn} = 1$. The matrices of the signal and error supports are defined as

$$\mathbf{Z} = \mathbf{I} \otimes \mathbf{z}' = \begin{pmatrix} \mathbf{z}' & & & \\ & \mathbf{z}' & & \\ & & \ddots & \\ & & & \mathbf{z}' \end{pmatrix} \quad \text{and} \quad \mathbf{V} = \mathbf{I} \otimes \mathbf{v}' = \begin{pmatrix} \mathbf{v}' & & & \\ & \mathbf{v}' & & \\ & & \ddots & \\ & & & \mathbf{v}' \end{pmatrix} \tag{2.27}$$

where $\mathbf{I}$ is an identity matrix and $\otimes$ denotes the Kronecker product.

As before, we obtain optimal solutions for $\mathbf{p}$ and $\mathbf{q}$ by setting up the primal Lagrange function and following through with the optimization. The optimal solution for $\mathbf{p}$ is as given in (2.16) and that for $\mathbf{q}$ is

$$\hat{q}_{mn} = \frac{q^0_{mn} \exp(v_m \sum_k x_{kn} \hat{\lambda}_k)}{\sum_{mn} q^0_{mn} \exp(v_m \sum_k x_{kn} \hat{\lambda}_k)} = \frac{q^0_{mn} \exp(v_m \mathbf{x}'_n \hat{\boldsymbol{\lambda}})}{\Gamma} \qquad \forall n, m. \tag{2.28}$$

Note the distinction between this solution and the optimal solution for $\hat{\mathbf{w}}$ given in (2.17). Unlike the solution for $\hat{\mathbf{w}}$, where the fixed weights $1/N$ were applied directly to the support

space **v**, here the support space for each error is not shrunk directly. Rather, the partition function Γ is defined over the entire sample and, as such, allows for an observation specific rate of shrinkage. Using the obtained solutions back in the primal, as before, we may derive an unconstrained dual optimization problem that can be solved in a variety of software. The heteroskedastic consistent GCE dual problem is

$$\mathcal{L}^D_{GCE} = \sum_{kn} x_{kn} y_n \lambda_k - \sum_n \ln \Omega_n - \ln \Gamma. \tag{2.29}$$

2.3.2. GCE with autocorrelated errors

The heteroskedastic consistent formulation of the GCE problem derived above can be seen as a special case of a more general formulation that not only allows error shrinkage rates to be determined endogenously, but also allows the optimally re-weighted errors to combine with each other in order to create signal distortion. That is, the moment constraints of (2.26) can be seen as a special case of the constraints

$$\sum_n x_{kn} y_n = \sum_n x_{kn} \mathbf{z}' \mathbf{p}_n + \sum_n x_{kn} \sum_j a_{nj} \pi_j \mathbf{v}' \mathbf{w}_j \qquad \forall k = 1, 2, \ldots, K \tag{2.30}$$

or, in matrix notation,

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{Z}\mathbf{p} + \mathbf{X}'\mathbf{A}\mathbf{V}\mathbf{q} \tag{2.31}$$

where **A** is a row-standardized hypothesized error structure matrix such that (2.26) is obtained by setting $\mathbf{A} = \mathbf{I}$.

The resulting optimization problem is very similar to the heteroskedastic case above. The addition of the link matrix only alters the definition of Γ in the optimal solution and, therefore, in the derivation of the dual objective function. Now, the optimal solution for **q** is

$$\hat{q}_{mn} = \frac{q^0_{mn} \exp(v_m \sum_k \tilde{x}_{kn} \hat{\lambda}_k)}{\sum_{mn} q^0_{mn} \exp(v_m \sum_k \tilde{x}_{kn} \hat{\lambda}_k)} = \frac{q^0_{mn} \exp(v_m \tilde{\mathbf{x}}'_n \hat{\boldsymbol{\lambda}})}{\tilde{\Gamma}} \qquad \forall n, m \tag{2.32}$$

where $\tilde{\mathbf{x}}_n$ is the nth row from the matrix $\tilde{\mathbf{X}} = \mathbf{A}'\mathbf{X}$ and $\tilde{\Gamma}$ is the partition function that is based on $\tilde{\mathbf{X}}$. The resulting dual objective function is derived in an identical fashion to the heteroskedastic case, with the exception that Γ is replaced with $\tilde{\Gamma}$.

To the extent that the off-diagonal elements in **A** are allowed to be non-zero, the optimally re-weighted errors are permitted to combine while distorting the signal. The matrix **A** is a row-standardized version of a spatial link matrix, say $\mathbf{A}^*$, which can be specified in a number of different ways. As stated above, if the desire is to allow for only heteroskedasticity, then $\mathbf{A}^* \equiv \mathbf{I}_n$ or

$$a^*_{nj} = \begin{cases} 1 & \forall j = n \\ 0 & \forall j \neq n. \end{cases} \tag{2.33}$$

To code heteroskedasticity as well as local first-order autocorrelation, we can define $\mathbf{A}^* =$

$\mathbf{I} + \mathbf{C}$, where $\mathbf{C}$ is a first order spatial contiguity matrix. That is,

$$a^*_{nj} = \begin{cases} 1 & \forall j = n \\ 1 & \forall j \in J_n \\ 0 & \text{for all other } j \end{cases} \tag{2.34}$$

where $j \in J_n$ is taken to read "all j units within the neighborhood of the nth unit." In order to include distance-based dependence for local neighbors (based on contiguity alone), we can define this matrix as $\mathbf{A}^* = \mathbf{I} + \exp(-\mathbf{D}) \odot \mathbf{C}$ where $\odot$ represents an element-by-element matrix multiplication, $\mathbf{D}$ represents an $N \times N$ matrix of distances between all pairs of data points, and $\mathbf{C}$ is as defined above. Finally, allowing global dependence, albeit with some distance-based decay, can be represented by setting $\mathbf{A}^* = \exp(-\mathbf{D})$. Row-standardizing $\mathbf{A}^*$ generally yields an asymmetric matrix, i.e., $\mathbf{A}' \neq \mathbf{A}$.

2.4. HYPOTHESIS AND SPECIFICATION TESTS

2.4.1. Hypothesis concerning parameters

If a solution exists, then the optimization problems described above yield estimates of the Lagrange multipliers ($\hat{\boldsymbol{\lambda}}$) that can be used to recover the probabilities of interest ($\hat{\mathbf{p}}_n$ and $\hat{\mathbf{w}}_n$), the corresponding signals ($\hat{s}_n$) and realized error terms ($\hat{e}_n$), and ultimately to derive marginal effects ($\partial \hat{s}_n / \partial x_{kn}$) and other inferential quantities as needed. Underlying all these quantities, however, are the set of estimated Lagrange multipliers ($\hat{\boldsymbol{\lambda}}$).

The dual versions of the information-recovery tasks are non-linear unconstrained optimization problems in the K Lagrange multipliers. Therefore, in addition to estimating these parameters, we can use the dual objective function to estimate a covariance matrix for $\hat{\boldsymbol{\lambda}}$. This can be used to conduct hypothesis tests on the Lagrange multipliers of interest or any derivatives thereof. The covariance matrix for the Lagrange multipliers ($\boldsymbol{\Sigma}_\lambda$) is defined as the inverse negative Hessian of the dual objective function evaluated at the optimal values of the Lagrange multipliers.

$$\hat{\boldsymbol{\Sigma}}_\lambda = \left\{ -\frac{\partial^2 \mathcal{L}^D_{GCE}}{\partial \hat{\boldsymbol{\lambda}} \partial \hat{\boldsymbol{\lambda}}'} \right\}^{-1} \tag{2.35}$$

This may either be computed analytically or may be retrieved from the numeric routine that is used to solve the optimization problem. In either case, the square-root of the diagonal elements of this matrix are estimated standard errors for the Lagrange multipliers. This yields a measure of the stability of the Lagrange multipliers and, consequently, a means of performing hypothesis tests on any inferential quantity based on them.

One quantity that is usually of interest is the marginal effect of the independent variables on the signal, i.e., $\partial \hat{s}_n / \partial x_{kn}$. As in all other non-linear models, this quantity must be evaluated at some data point (which is usually the sample means of the predictors). Let $\mathbf{x}_*$ represent a generic point of evaluation and let $\hat{p}_{l*}$ represent the probability computed at that point. Then, in our semi-parametric formulation of the signal, the marginal effects are

computed as

$$\hat{\gamma}_k = \frac{\partial \hat{s}_*}{\partial \hat{x}_{k*}} = \hat{\lambda}_k \Big\{ \sum_l z_l^2 \hat{p}_{l*} - \big(\sum_l z_l \hat{p}_{l*} \big)^2 \Big\} \tag{2.36}$$

or, in matrix notation, as

$$\hat{\boldsymbol{\gamma}} = \frac{\partial \hat{s}_*}{\partial \mathbf{x}_*} = \hat{\boldsymbol{\lambda}} \Big\{ \mathbf{z}^{2\prime} \hat{\mathbf{p}}_* - (\mathbf{z}' \hat{\mathbf{p}}_*)^2 \Big\} \tag{2.37}$$

which are non-linear functions of the underlying Lagrange multipliers (because $\hat{\mathbf{p}}_*$ are functions of $\hat{\boldsymbol{\lambda}}$). In order to convert the estimated covariance matrix of $\hat{\boldsymbol{\lambda}}$ into an estimate of the covariance among $\hat{\boldsymbol{\gamma}}$, we make use of the delta-method (Greene, 2000, pg.357). That is,

$$\hat{\boldsymbol{\Sigma}}_\gamma = \left(\frac{\partial \hat{\boldsymbol{\gamma}}}{\partial \hat{\boldsymbol{\lambda}}'} \right) \hat{\boldsymbol{\Sigma}}_\lambda \left(\frac{\partial \hat{\boldsymbol{\gamma}}}{\partial \hat{\boldsymbol{\lambda}}'} \right)' \tag{2.38}$$

where, given the definition of $\hat{\boldsymbol{\gamma}}$ above, we have

$$\begin{aligned} \frac{\partial \hat{\boldsymbol{\gamma}}}{\partial \hat{\boldsymbol{\lambda}}'} &= \Big\{ \mathbf{z}^{2\prime} \hat{\mathbf{p}}_* - (\mathbf{z}' \hat{\mathbf{p}}_*)^2 \Big\} \cdot \mathbf{I} \\ &\quad + \Big\{ \mathbf{z}^{2\prime} \hat{\mathbf{p}}_* + (\mathbf{z}^{2\prime} \hat{\mathbf{p}}_*)(\mathbf{z}' \hat{\mathbf{p}}_*) - 2(\mathbf{z}' \hat{\mathbf{p}}_*)^3 \Big\} \cdot \hat{\boldsymbol{\lambda}} \mathbf{x}_*'. \end{aligned} \tag{2.39}$$

In addition, making use of the optimized values of the objective functions, an Entropy Ratio (*ER*) test (analogous to a Likelihood Ratio test) can be constructed to test joint hypotheses on the Lagrange multipliers. This statistic is defined as twice the difference between the maximized values of the objective functions of two nested models. Like the Likelihood Ratio statistic, the *ER* statistic has an asymptotic χ^2 distribution with R degrees of freedom, where R corresponds to the number of restrictions (Jaynes, 1979, pg.67). In other words, denoting the optimized values of the objective function for the restricted and unrestricted models as $\tilde{\mathcal{L}}^D_{GCE}$ and $\hat{\mathcal{L}}^D_{GCE}$ respectively, the *ER* statistic, defined as

$$ER_R = 2 \Big\{ \tilde{\mathcal{L}}^D_{GCE} - \hat{\mathcal{L}}^D_{GCE} \Big\} \sim \chi^2_R \tag{2.40}$$

can be used to test the validity of the restrictions jointly.

2.4.2. Hypothesis concerning specification

Although several structures can be hypothesized for the spatial error correlation, in the information-theoretic framework described above, they all lead to non-nested models. Since the correlation in the errors is not imposed by means of explicit data constraints, resulting in corresponding Lagrange multipliers, it cannot be tested using the framework described above. Therefore, we need some other criterion to choose among the alternate error structures.

Once we have estimated a model, or more precisely the Lagrange multipliers, we may use them to recover all the probabilities in the model. In the GCE models with heteroskedas-

ticity and/or autocorrelation, this means both $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$. Using these estimated probabilities we can construct the entropy measures for each of the sources of uncertainty in the model. The Entropy of a probability vector is a measure of the amount of uncertainty it implies. Uncertainty, on the other hand, is an inverse measure of information.[4] The more uncertain we are about the outcome, the less information the model conveys about it. Therefore, we should be able to combine the computed Entropy measures from these distinct sources of uncertainty for a pair of models to yield a criterion for selecting between them. Additionally, we should be able to use these measures to choose among several non-nested models. Below we describe one way of doing so.

Consider a pair of models $\mathcal{M}_0$ and $\mathcal{M}_1$ where $\mathcal{M}_0$ is encompassed by $\mathcal{M}_1$. That is, even though $\mathcal{M}_0$ is not nested within $\mathcal{M}_1$ parametrically, $\mathcal{M}_1$ allows within it all the specifications in $\mathcal{M}_0$. For example, a model with $\mathbf{A}^* = \mathbf{I}$ is encompassed by the model that defines $\mathbf{A}^* = \mathbf{I} + \mathbf{C}$ if everything else in the model remains fixed. That is, a model that permits *only* heteroskedasticity is encompassed within a model that permits heteroskedasticity *and* first-order spatial error-correlation. In fact, all the specifications of $\mathbf{A}^*$ described above (except when $\mathbf{A}^* = \mathbf{0}$) encompass the heteroskedastic consistent model where $\mathbf{A}^* = \mathbf{I}$. In a similar manner, the heteroskedastic consistent model encompasses the ME/CE models which may be obtained by setting $\mathbf{A}^* = \mathbf{0}$. Therefore, logically, all models that permit heteroskedasticity as well as autocorrelation encompass the pure ME/CE models.

The GCE models described above all have one thing in common; they are all attempts at capturing the structure in the errors non-parametrically. If the structure hypothesized in $\mathbf{A}$ is a good approximation of reality, then this must help us gain information about the error structure *without* giving up too much information about the signals. Of course, if it helps us gain information about the signals, all the better. Therefore, if we define entropy measures $H(\hat{\mathbf{p}}_0)$ and $H(\hat{\mathbf{q}}_0)$ as quantifying our uncertainty about the signal and noise terms in model $\mathcal{M}_0$, and $H(\hat{\mathbf{p}}_1)$ and $H(\hat{\mathbf{q}}_1)$ as quantifying our uncertainty about the corresponding measures in model $\mathcal{M}_1$, then we may define the relative *gain* in error information as

$$\mathcal{H}_e = H(\hat{\mathbf{q}}_0)/H(\hat{\mathbf{q}}_1) \tag{2.41}$$

and the relative *loss* in signal information as

$$\mathcal{H}_s = H(\hat{\mathbf{p}}_1)/H(\hat{\mathbf{p}}_0). \tag{2.42}$$

These quantities may be defined equivalently in terms of the Normalized Entropy measures $S(\hat{\mathbf{p}}) = H(\hat{\mathbf{p}})/H(\hat{\mathbf{p}}^0)$ (Golan et al., 1996, pg.27). Since the prior probabilities in each pair of competing models is the same, the normalizing constants cancel each other out, and we obtain the definitions given above.

To see if the flexibility provided in a given model $\mathcal{M}_1$ over that provided in model $\mathcal{M}_0$ is worthwhile, we can compare these two measures. To do so, we define a composite ratio of these measures as

$$\mathcal{H}_* = \mathcal{H}_e/\mathcal{H}_s \tag{2.43}$$

[4] See, among others, Soofi (1994) for a general discussion about the concept of information.

which can be used to gauge the relative efficiency of $\mathcal{M}_1$ over $\mathcal{M}_0$. If $\mathcal{H}_* < 1$ then the gains made by increasing knowledge about the error structure are too costly. On the other hand, if $\mathcal{H}_* > 1$ then the gains are worthwhile. Since all definitions in the two models are identical with the exception of $\mathbf{A}$, these computations can be taken to mean the following: If one has to give up too much information on the signals in order to gain information about the errors, then the data are clearly not supporting that error structure. If, on the other hand, the data do support the hypothesized error linkages, then the gains in information about the errors should far outweigh the losses we incur in terms of the signals.

Finally, if we have several models that encompass the same underlying model, e.g., $\mathbf{A} = \mathbf{I}$ is encompassed by all models with $\mathbf{A} \neq \mathbf{I}$ other than $\mathbf{A} = \mathbf{0}$, then we can compare the composite relative efficiency measure across all models that are worthwhile (i.e., $\mathcal{H}_* > 1$) and select the one that offers the highest gains. This model can be viewed as the one that is favored by the sample as being the closest to the underlying data generating process among all competing models and, in that sense, should be considered the "best" model.

2.5. SPECIFYING THE SUPPORT SPACE

So far the flexibility allowed in the moment constraints and the estimation implications derived thereof have been discussed for abstractly defined signal and noise supports. If natural bounds exist for these unknowns, then that knowledge may be used directly in specifying the supports. If they do not exist, then specifying the supports requires careful sample-specific considerations. Below we discuss the specification for two examples — binary choice and count outcomes — that are relevant to this project.

2.5.1. Binary choice outcomes

In the case of binary choices, there exist natural bounds for both the signal as well as the noise terms: the observed and expected outcomes in this case can only exist between 0 and 1. This means the signals are naturally bounded by 0 and 1, i.e., $z_l \in (0, 1)$. A simple specification would be $\mathbf{z} = (0, 1)'$. Now, if we observe an outcome (i.e., $y_n = 1$) but predict it as being nearly impossible (i.e., $\hat{s}_n \approx 0$), then the error can be as high as $+1$. Or, if the binary choice is not observed (i.e., $y_n = 0$) but we predict it with near certainty (i.e., $\hat{s}_n \approx 1$), then the error can be as low as -1. In other words, the errors are also naturally bounded between ± 1, i.e., $v_m \in \pm 1$ and a simple specification would be $\mathbf{v} = (-1, +1)'$.

If we specify the support spaces as described above and create noiseless moment constraints of (2.8), then the resulting Maximum Entropy solutions are identical to the Logit parameters. In fact, under this specification the Maximum Entropy dual objective function turns out to be identical to the Logit log-likelihood function. As such, all inferences derived from it, including the parameter estimates and their covariance matrix, are identical to those that would be recovered from the Logit model. The GME/GCE specification results in loosened constraints and, in finite samples, yields superior (i.e., more stable) parameter estimates. The asymptotic equivalence of the GME/GCE model to the Maximum Likelihood Logit model is explicitly demonstrated in Golan, Judge, and Perloff (1996).

2.5.2. Count outcomes

Count outcomes can be thought of as a summation over a large but finite sequence of independent and identical binary choices. That is the motivation underlying a Binomial distribution and the Poisson distribution is, in fact, obtained at the limit when the number of binary choices in the sequence approach ∞. Suppose, then, that we define the count outcome as T times the signal and noise terms in each of the T individual binary choices. That is, we let

$$\mathbf{y} = T\{\mathbf{s} + \mathbf{e}\} \tag{2.44}$$

where the underlying signal and noise supports ($\mathbf{z}$ and $\mathbf{v}$) are as defined for the binary choice outcome. The resulting expected count would now be $\in (0, T)$. However, since there are T different sources of signal and noise uncertainty for each sampled unit, the entropy function must be appropriately scaled. The flexible moment constraints for the count outcome case therefore get re-defined as

$$\sum_n x_{kn} y_n = T\Big\{ \sum_n x_{kn} \mathbf{z}' \mathbf{p}_n + \frac{1}{N} \sum_n x_{kn} \mathbf{v}' \mathbf{w}_n \Big\} \qquad \forall k = 1, 2, \ldots, K \tag{2.45}$$

and the objective function is accordingly re-defined as

$$\min_{\mathbf{p},\mathbf{w}} \quad CE(\mathbf{p}, \mathbf{w}; \mathbf{p}^0, \mathbf{w}^0) = T\Big\{ \mathbf{p}' \ln(\mathbf{p}/\mathbf{p}^0) + \mathbf{w}' \ln(\mathbf{w}/\mathbf{w}^0) \Big\}. \tag{2.46}$$

All derivations and extensions (for heteroskedastic and/or auto-correlated errors) from the previous sections follow exactly as explained in Section 2.3 with a scaling factors T appropriately included. The final dual objective function for a GCE model for count outcomes with potentially heteroskedastic and spatially auto-correlated errors is defined as

$$\mathcal{L}^D_{GCE} = \sum_{kn} x_{kn} y_n \lambda_k - T \sum_n \ln \Omega_n - T \ln \tilde{\Gamma} \tag{2.47}$$

where all terms have been defined before. The only remaining issue is a choice for the value of T. Here we use knowledge of the empirical distribution of y_n as a guiding factor. If, for example, we observe counts only as high as 10, then it is unlikely that the underlying expected count may be much higher than that. Therefore, as a conservative rule of thumb, we use $3 \times \max(y)$ in any sample as the fixed value for T.

2.6. DISCUSSION

Several special considerations must be kept in mind for the count outcome models. First is the issue of "population-at-risk." Clearly, allowing T to vary across observations is a trivial extension in the noiseless moment constraints setting. It complicates issues considerably, however, when dealing with the heteroskedastic and/or autocorrelated errors case. It is simpler to use the population-at-risk (or its natural log) explicitly as an additional predictor

in the regressions. This is the approach we use in our analysis. Our ongoing research considers extending the model to the case where the realm of possibility for the expected outcome in each sampled unit is different *and* explicitly known. The extension is discussed more in Chapter 5.

A second important consideration pertains to over-dispersion. Typically, the Poisson assumption is a very restrictive one, as its first and second moments are equal. To allow for some flexibility, researchers rely on some variant of the traditional Poisson model that permits over-dispersion. Several such variants are available, each resulting from different assumptions made about an over-dispersing random variable (Cameron and Trivedi, 1986). In the GCE specification described above, rather than parameterize the heteroskedasticity *indirectly* through an additional variable for which additional (often mathematically convenient) distributional assumptions must be made, we allow for heteroskedasticity *directly* by allowing an endogenous optimal re-weighting of the errors.

A final consideration in count outcome models is the over-representation of zero outcomes in the sample. In such settings, it may be reasonable to model the choice between no event and some event as a different mechanism from that yielding the number of events. Extending the GCE to extract the components of these mixed processes is also part of our ongoing research and is not discussed here.

The main emphasis in this project is on utilizing the flexibility of the GCE to deal with error-correlation across space, i.e., to treat spatial dependence as a "nuisance." Other forms of spatial structures may also exist in the data. If the spatial relationships are theorized to be of a *substantive* nature, then they should be modeled as such.

There are two forms of substantive spatial processes that can be modeled. First, one may hypothesize that the signal (or expected outcome) is directly related to the observed explanatory factors in neighboring areas, i.e., where $s_n = f(\mathbf{x}'_n, \mathbf{x}'_{j\in J_n})$. Modeling substantive spatial dependence of this type is easily permitted in the current model, as explained in the preceding sections, by including $\mathbf{WX}$ (or a subset thereof) as additional variables in the design matrix constraining the probabilities of interest. Here $\mathbf{W}$ is a spatial weight matrix rather than the link matrix $\mathbf{A}$. It is typically row standardized and has $w_{nj} = 0\ \forall n = j$ so that the nth row of $\mathbf{WX}$ (i.e., $\sum_j w_{nj}x_{kj}\ \forall j \in J_n$) is, in effect, a spatially weighted average of relevant neighboring area predictors. In the next chapter, we explicitly estimate a set of such models that allow the spatial lag of a predictor variable to enter the hypothesized set of regressors and discuss findings.

Second, the signal may be hypothesized to depend on the observed outcomes in neighboring areas, i.e., where $s_n = f(\mathbf{x}'_n, \mathbf{x}'_{j\in J_n}, y_{j\in J_n})$. Modeling and interpreting substantive spatial effect of this kind are more difficult. From an estimation point of view, including $\mathbf{Wy}$ with the regressors constitutes a violation of the orthogonality assumption because of its endogeneity. Unlike in time-series analysis, where the time-lag of the dependent variable is treated as *pre-determined* and therefore uncorrelated with the current period noise, the spatial-lag term is endogenous and therefore correlated with local area errors (Anselin and Bera, 1998). Therefore, even asymptotically, one should not expect $\mathbf{X}'\mathbf{e}$ to vanish. More importantly, however, the interpretation of a significant coefficient on $\mathbf{Wy}$ in

single cross-sections (such as the analysis performed in this study) is not clear (Anselin, 2002). A promising specification that the GCE framework may permit is one where *signal*-autocorrelation is modeled directly rather than via a functional dependence on **Wy**. This avenue of research has the potential for allowing a *simultaneous* approach to modeling substantive spatial dependence.

Chapter 3

The Data

For this project we analyzed data obtained from public sources about violence in the city of Chicago. These data were obtained at the lowest unit of analysis — the census tract (CT) — and were later aggregated up to the neighborhood cluster (NC) level. The neighborhood clusters are defined by the Project on Human Development in Chicago's Neighborhoods (PHDCN) as combinations of Chicago's 865 census tracts that are "geographically contiguous and socially similar" (Morenoff, Sampson and Raudenbush, 2001). The mapping of census tracts to neighborhood clusters was obtained directly from Dr. Robert Sampson of the PHDCN.

Using 1990 geographic definitions, there were 865 census tracts with an average population of approximately 3,200 people. The aggregated 343 neighborhood clusters had an average population of roughly 8,000 people. The project aimed at examining the extent to which socio-economic and demographic characteristics of these 865 census tracts, and alternately the 343 neighborhood clusters, affected the amount of violence experienced by these communities. Moreover, the project aimed at comparing inferences derived from the two units of analysis and for the various types of homicides that were analyzed. The disaggregated homicides types analyzed in this project are described in the next section, followed by a description of the independent variables used in the project.

3.1. DISAGGREGATED HOMICIDE COUNTS

Data on the dependent variable, the Disaggregated Homicide Victimization Counts over the 3-year period (1989–91), were obtained from *ICPSR 6399: Homicides in Chicago, 1965-1995 (Part 1, Victim Level File)*. That data file is a compilation of all homicides reported to the police between 1965 and 1995 (Block and Block, 1998). This file contains detailed information on victim, offender, and the circumstances of each homicide as well as the offense date. Additionally, it contains a variable that indicates the "type" of homicide. This variable, SYNDROME, distinguishes between the various homicide sub-types that were analyzed in this project. The original coding in the data contains 10 different categories which

include gang-related (01), sexual assault (02), instrumental (03), spousal attack (04), child abuse (05), other family expressive (06), other known expressive (07), stranger expressive (08), other (09) and mystery (10). We re-combined these into six categories by collapsing values 04, 05 and 06 into a generic 'Family' related expressive category and 02, 09 and 10 into the 'Other' category.[1] All analysis, therefore, is performed on the 6 disaggregated homicide types that are classified as being gang-related (**GNG**), instrumental (**INS**), family related expressive (**FAM**), known person expressive (**KNO**), stranger expressive (**STR**), and other (**OTH**), in addition to a model estimated for all homicides combined (**ALL**).

Each victimization in this file is flagged by the location where the victim's body was found. In the public release version of the data, this information is provided only by a census tract number. Using this information, along with the re-coded homicide-types, raw counts were computed at the census tract levels for a 3-year time frame spanning the years 1989, 1990 and 1991. Raw counts were then aggregated up to the neighborhood cluster level. Figure 3.1 shows the distribution of the disaggregated homicide counts measured at the census tract level, whereas Figure 3.2 shows the same at the neighborhood cluster level. It is clear from these figures that the distributions of the criterion measures are highly skewed and that large numbers of areal units have zero counts. In fact, the number of neighborhood clusters with no reported victims range from a low of about 40% (KNO) to a high of about 63% (STR) of the sample. Similarly, the number of census tracts with no reported homicide victims range from a low of 63% (KNO) to a high of 80% (STR) of the sample.

Next, in order to assess whether the criterion measures are randomly distributed across space, they were mapped for visual analysis. Though not a formal procedure, we also computed simple linear regression coefficients for each of these crime counts regressed on their first-order spatial lags (using the queen contiguity criterion) to get an idea of the extent and direction of spatial auto-correlation. The models were specified as $y_n = \alpha + \beta\{\mathbf{Wy}\}_n + e_n$. Table 3.1, where these coefficients ($\hat{\alpha}$ and $\hat{\beta}$) are displayed, along with visual inspection of the maps in Figure 3.3 and Figure 3.4, indicates that the dependent variables being analyzed are not randomly distributed across space and, in all cases, they appear to be positively auto-correlated.

3.2. INDEPENDENT VARIABLES

The structural, social, and economic indicators used to model these disaggregate homicide counts were obtained from the Neighborhood Change Database (NCDB) maintained by The Urban Institute. The NCDB contains social, demographic, economic, and housing data on census tracts in the United States for 1970, 1980, 1990, and 2000. Data in the NCDB are based on information gathered by the U.S. Bureau of the Census in its decennial censuses. The Bureau makes census tract data available to the public in both printed and machine-readable formats. The NCDB contains this public information in one database.

[1] The 'Other' category was intended to be a "catch-all" category for homicides that were not classified in any of the remaining types. Since the number of sexual assault homicides were very small, they were included in this category.

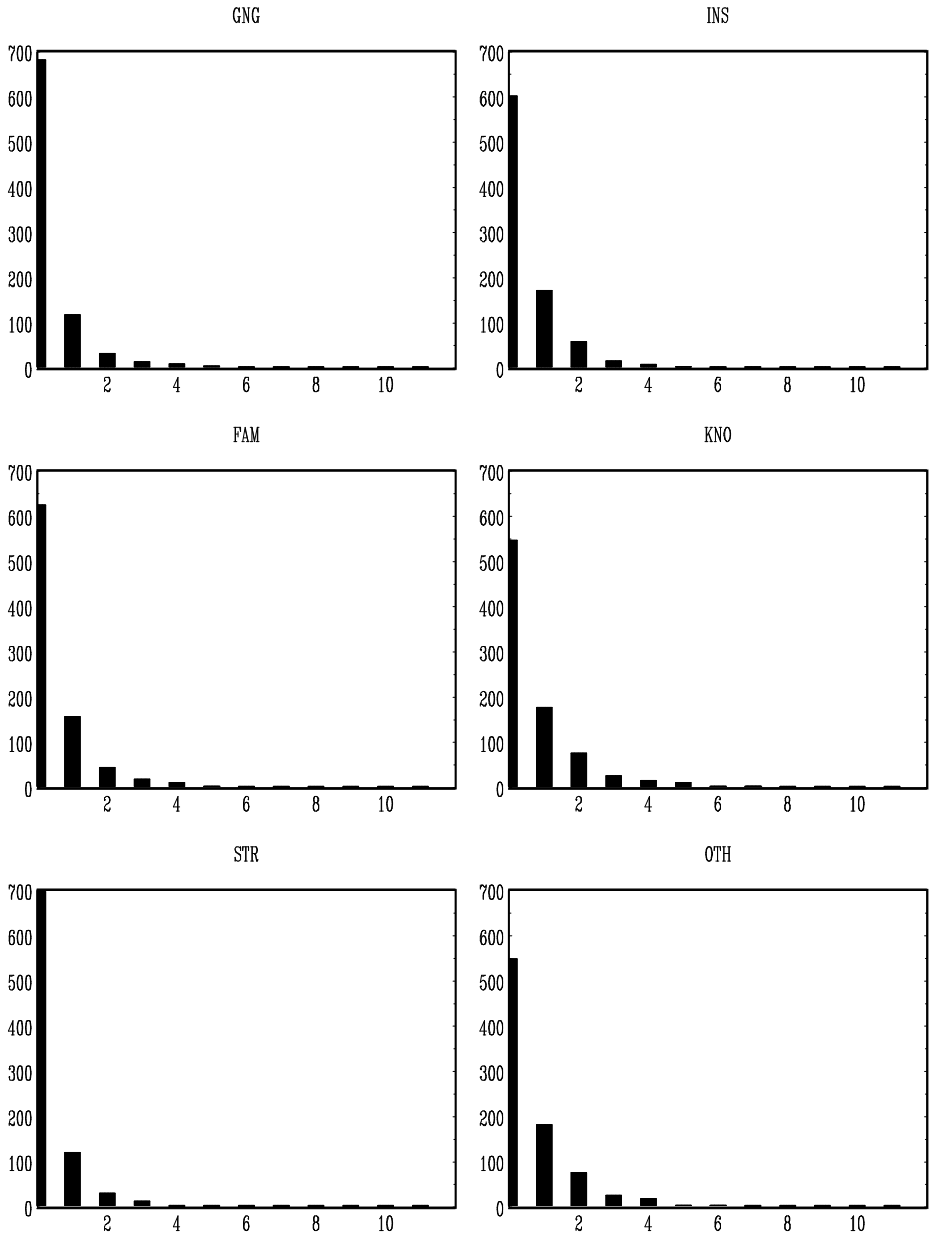

Figure 3.1: Frequency distribution of disaggregated homicide counts in Chicago's 865 Census Tracts (1989–91)

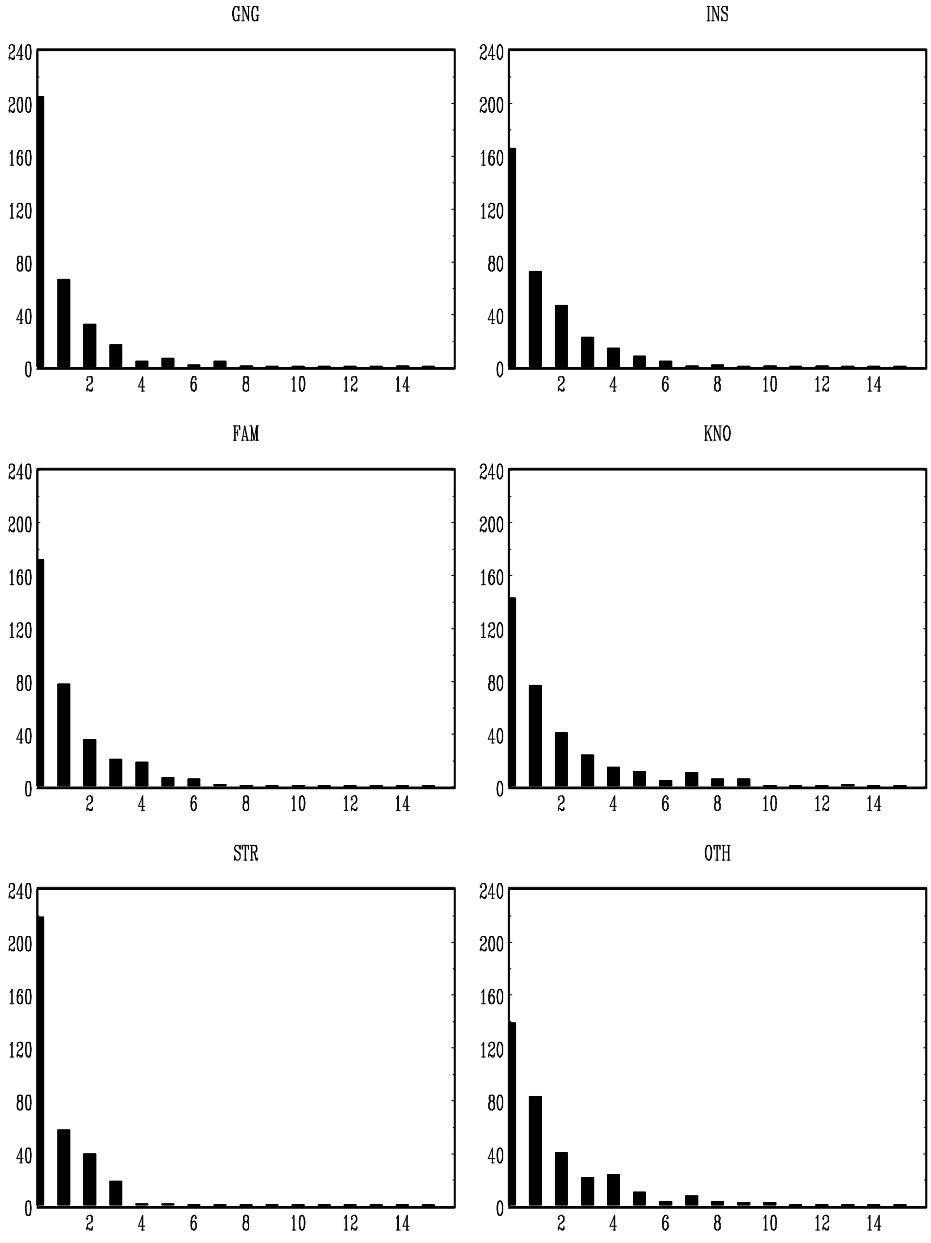

Figure 3.2: Frequency distribution of disaggregated homicide counts in Chicago's 343 Neighborhood Clusters (1989–91)

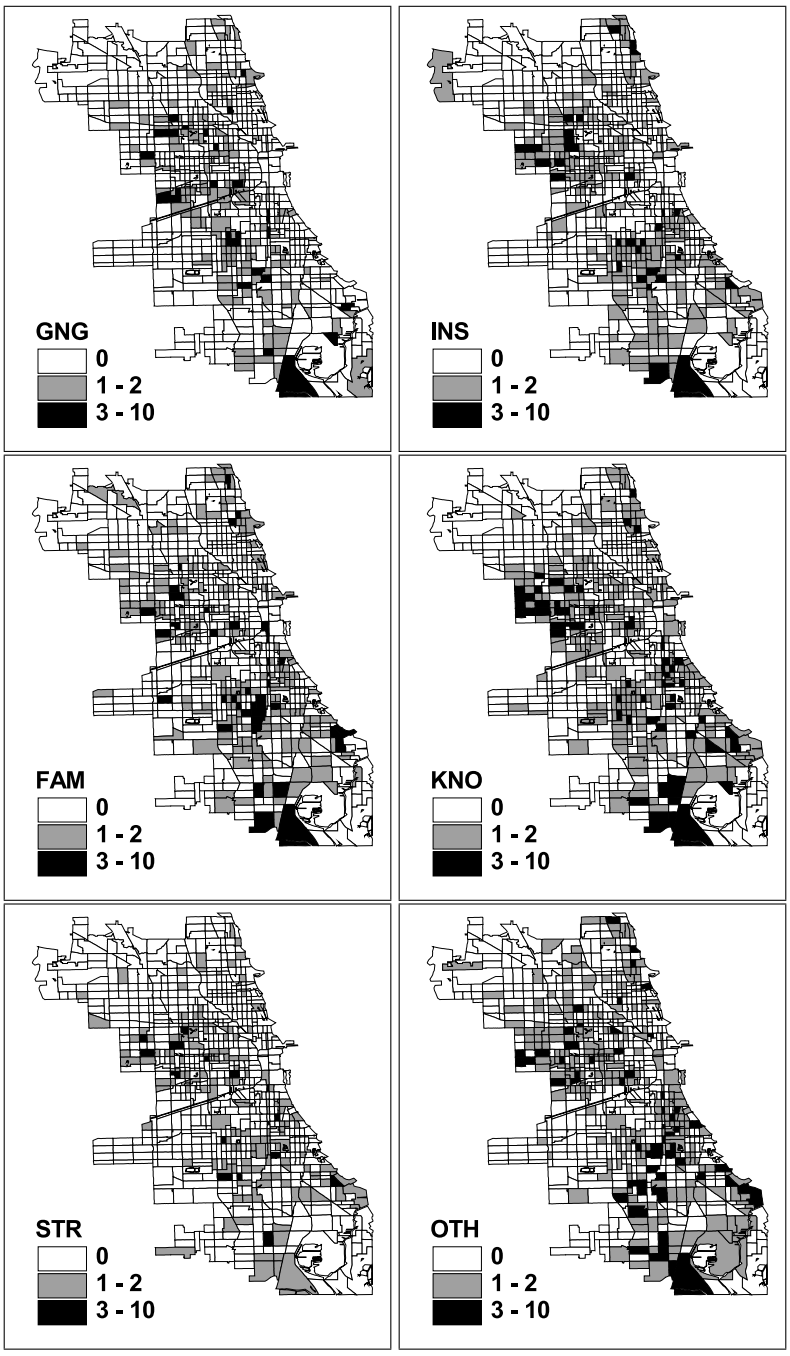

Figure 3.3: Geographic distribution of disaggregated homicide counts in Chicago's 865 Census Tracts (1989–91)

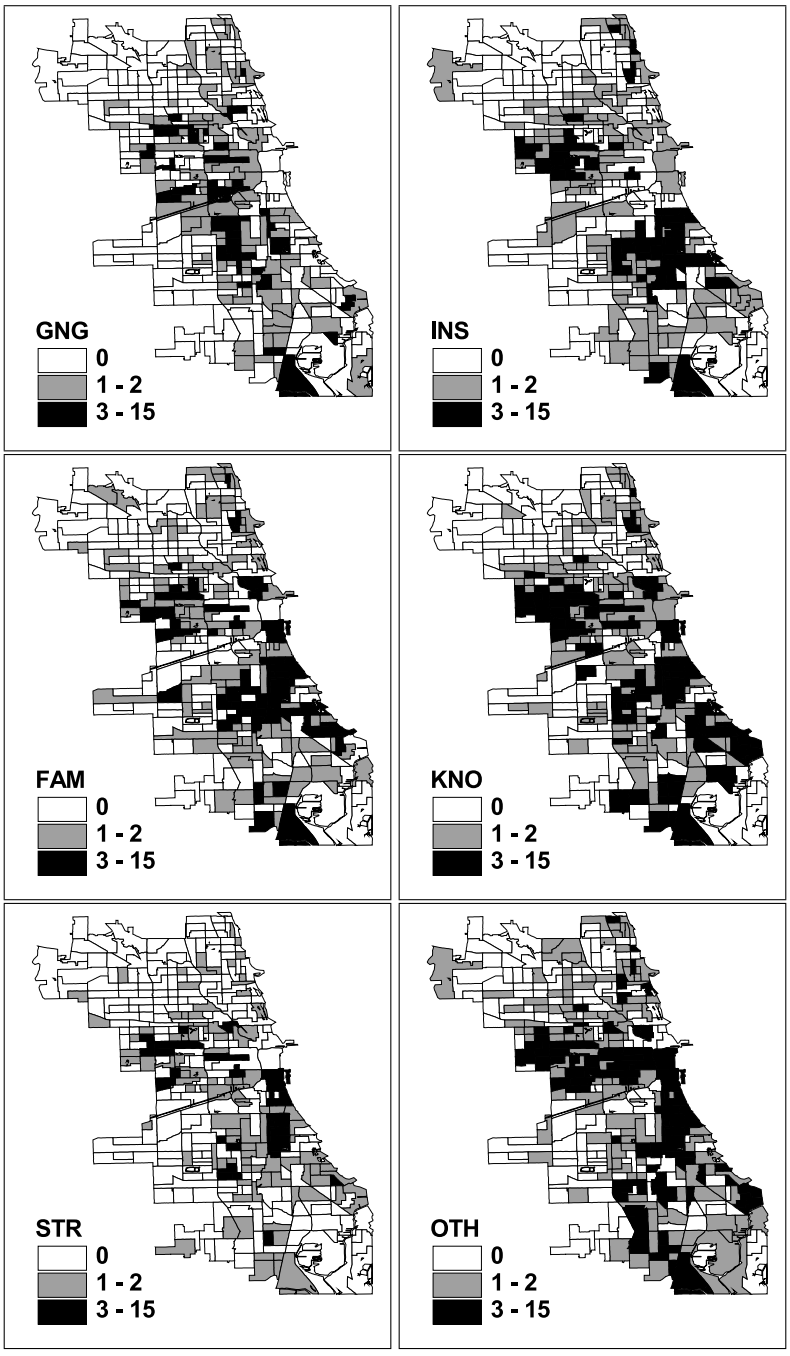

Figure 3.4: Geographic distribution of disaggregated homicide counts in Chicago's 343 Neighborhood Clusters (1989–91)

Table 3.1: OLS regression coefficients of disaggregated homicide counts (1989–91) regressed on their first-order spatial lag (using a queen contiguity criterion) and an intercept

	ALL	GNG	INS	FAM	KNO	STR	OTH
Unit of Analysis: Neighborhood Cluster (N = 343)							
Intercept	0.29	0.20*	0.21*	0.40**	0.22	0.28**	0.29*
	(0.390)	(0.114)	(0.120)	(0.123)	(0.154)	(0.082)	(0.158)
Wy (Spatial Lag of **y**)	0.93**	0.79**	0.81**	0.63**	0.88**	0.59**	0.80**
	(0.054)	(0.092)	(0.072)	(0.079)	(0.065)	(0.083)	(0.071)
Unit of Analysis: Census tract (N = 865)							
Intercept	0.53**	0.12**	0.19**	0.20**	0.21**	0.17**	0.32**
	(0.142)	(0.035)	(0.041)	(0.043)	(0.053)	(0.029)	(0.054)
Wy (Spatial Lag of **y**)	0.73**	0.66**	0.59**	0.53**	0.69**	0.35**	0.50**
	(0.047)	(0.058)	(0.055)	(0.063)	(0.052)	(0.067)	(0.057)

** $p < 0.05$; * $p < 0.1$ using conventional t-tests
Unstandardized linear regression coefficients with asymptotic standard errors in parenthesis

For this analysis we constructed 9 measures of the structural, social and economic conditions within the areal unit being analyzed. The definition of the measures was kept constant across the two levels of areal aggregation. Data were collected at the CT level and later aggregated to the NC level of analysis. Data not available or computable at the census tract level were imputed from the relevant neighborhood cluster level. That is, the value for a given characteristic at the neighborhood level was used as the assigned value for a census tract missing that information. A description of each of these measures, along with the source of the measure and the number of census tracts missing that measure (in parenthesis), is provided below.

SHRBLK *Proportion of the neighborhood population that is black*. This variable, quantifying racial makeup of the community, was obtained directly from the NCDB (# missing = 14).

SHRHSP *Proportion of neighborhood population that is Hispanic*. This variable, also quantifying the racial makeup of the community, was directly obtained from the NCDB (# missing = 14).

PNFH *Proportion of neighborhood households that are non-family*. This variable was included to measure social disorganization and family structure and was directly obtained from the NCDB (# missing = 14).

FEMH *Proportion of neighborhood households with children that are headed by females*. This variable was also included to measure social disorganization and family structures and was also directly obtained from NCDB (# missing = 25).

YMEN *Young men aged 15-25 as a proportion of areas total population*. This variable was computed by dividing the count of relevant young men by the total area population. It was included to measure the prevalence in the population of persons who are typically thought of as committing violent crimes (# missing = 14).

UNEMP *The neighborhood unemployment rate* as defined by the Bureau of Census. This variable includes persons over the age of 16 who were in the civilian labor force, but were unemployed. The variable was obtained directly from the NCDB and was included to assess the effects of one form of economic deprivation on violence (# missing = 17).

POVRT *Proportion of neighborhood population below the poverty line* (as defined for 1989). This variables was also included as another measure of economic deprivation and was obtained directly from the NCDB (# missing = 14).

RESST *Residential stability in the area*. This variable was computed as the proportion of owner occupied housing units where the head of household has lived for at least 5 years (# missing = 14).

LPOP *The natural log of the total residential population* of the areal unit. This variable was included as a control for variations in population-at-risk among the areal units (# missing = 14).

Univariate characteristics and bivariate correlations of the resulting measures reveal that the data are highly collinear (see Table 3.2). In order to mitigate the ill-effects of collinearity, we used a data-reducing technique common in this literature. We first assessed the extent to which these data elements could be collapsed into underlying latent constructs using factor analysis on all the included variables (except LPOP). Then, having decided on a set of covariates that constitute meaningful latent constructs, we used a confirmatory factor analysis to create the latent constructs. This procedure was repeated at both the NC as well as the CT levels of analysis. Similar results were obtained at both levels. Using a retention criterion of a minimum Eigenvalue of 1, we obtained three significant latent constructs. The first one indicated high (larger than 0.5) factor loadings of SHRBLK, FEMH, UNEMP, and POVRT resulting in a measure that captures underlying resource deprivation. The second construct indicated high loadings of SHRSHP and YMEN not yielding any clear meaning to the underlying latent construct; and the third factor indicated high loadings of PNFH and RESST, once again not yielding any clear meaning for the underlying latent construct. Therefore, for purposes of the regression analysis, we computed a single resource deprivation index (RESDEP) using the four covariates that loaded high on it and used the remaining covariates in their manifest forms. In all the regression models, therefore, the basic set of explanatory measures used are RESDEP, SHRHSP, PNFH, YMEN, RESST and LPOP (in addition to an intercept term).

Despite the reduction in the dimensionality of the regressors, collinearity among the predictors still remains. Using the design matrix condition number (Belsley, 1991), defined

as $\kappa(\mathbf{X}'\mathbf{X}) = \eta_+/\eta_-$ where $\eta_{(+,-)}$ are the highest and lowest singular values of the data matrix $\mathbf{X}$ (with columns scaled to unit length), we assessed this collinearity to still be serious. At the NC level the data matrix has a condition number of 218.7 whereas at the CT level it is 176.3.[2]

In the next chapter, we provide results of the various models analyzed using the data described above.

Table 3.2: Bivariate correlation coefficients and descriptive statistics for the areal macro-characteristics used in the analysis

	SHRBLK (1)	SHRHSP (2)	FEMH (3)	PNFH (4)	YMEN (5)	UNEMP (6)	POVRT (7)	RESST (8)	LPOP (9)
Unit of analysis: Neighborhood cluster (N = 343)									
(1)	1	−0.53**	0.88**	−0.23**	0.10*	0.71**	0.55**	0.19**	−0.16**
(2)	−0.53**	1	−0.32**	−0.19**	0.43**	−0.16**	0.04	0.15**	−0.07
(3)	0.88**	−0.32**	1	−0.15**	0.16**	0.84**	0.78**	0.31**	−0.23**
(4)	−0.23**	−0.19**	−0.15**	1	−0.29**	−0.31**	−0.13**	−0.04	0.27**
(5)	0.10*	0.43**	0.16**	−0.29**	1	0.24**	0.28**	0.11*	−0.14**
(6)	0.71**	−0.16**	0.84**	−0.31**	0.24**	1	0.88**	0.32**	−0.27**
(7)	0.55**	0.04	0.78**	−0.13**	0.28**	0.88**	1	0.42**	−0.23**
(8)	0.19**	0.15**	0.31**	−0.04	0.11**	0.32**	0.42**	1	−0.15**
(9)	−0.16**	−0.07	−0.23**	0.27**	−0.14**	−0.27**	−0.23**	−0.15**	1
Mean	0.42	0.20	0.40	0.32	0.08	0.14	0.23	0.14	8.94
S.D	0.43	0.26	0.23	0.15	0.02	0.10	0.17	0.04	0.37
Unit of analysis: Census tract (N = 865)									
(1)	1.00	−0.54**	0.83**	−0.16**	−0.02	0.66**	0.56**	0.13**	−0.11**
(2)	−0.54**	1.00	−0.31**	−0.21**	0.22**	−0.17**	−0.03	0.07**	−0.02
(3)	0.83**	−0.31**	1.00	−0.09**	0.03	0.74**	0.73**	0.17**	−0.26**
(4)	−0.16**	−0.21**	−0.09**	1.00	−0.06*	−0.21**	−0.09**	0.10**	−0.17**
(5)	−0.02	0.22**	0.03	−0.06*	1.00	0.08**	0.10**	−0.02	−0.01
(6)	0.66**	−0.17**	0.74**	−0.21**	0.08**	1.00	0.81**	0.15**	−0.28**
(7)	0.56**	−0.03	0.73**	−0.09**	0.10**	0.81**	1.00	0.26**	−0.34**
(8)	0.13**	0.07**	0.17**	0.10**	−0.02	0.15**	0.26**	1.00	−0.08**
(9)	−0.11**	−0.02	−0.26**	−0.17**	−0.01	−0.28**	−0.34**	−0.08**	1.00
Mean	0.43	0.19	0.42	0.34	0.08	0.15	0.26	0.14	7.75
S.D	0.44	0.26	0.25	0.19	0.04	0.12	0.20	0.08	0.97

** $p < 0.05$; * $p < 0.1$.

[2] A condition number of 1 implies a perfectly orthogonal design matrix. Condition numbers as low as 30 can indicate potentially damaging multicollinearity whereas condition numbers as high as 900, that are not uncommon in real-world design matrices, can result in degrading multicollinearity.

Chapter 4

Findings

This chapter presents the findings from applying the methods explained in Chapter 2 to the data described in Chapter 3. Baseline models are first estimated using the Maximum Likelihood framework for comparison purposes. Next, using the neighborhood cluster (NC) level model for the total homicide counts (ALL) as an example, the workings of the GCE estimator are presented and discussed for various hypothesized error-structures. Models for the disaggregated homicides are then presented and findings are discussed only for the best models (as gauged by $\mathcal{H}_*$). In order to assess whether and to what extent spatially-lagged exogenous predictors may influence the criterion measures, models that include a spatially-lagged variable are presented and discussed in light of the spatial spill-over effects they imply. The chapter concludes with a summary discussion of the findings.

4.1. THE BASELINE COUNT OUTCOME MODELS

Table 4.1 presents estimates of a basic set of models estimating the effects of these predictors on the count of the various disaggregated homicide types. The models are estimated using the traditional Poisson regression framework available in most statistical software.

There are several points worth noting here. First, resource deprivation (RESDEP) is a strong and persistent predictor of all the types of homicides analyzed at both the NC and CT levels. In a similar manner, the log of the total population (LPOP) is a significant predictor of the total count of homicides as well as all disaggregate homicide types at both levels of analysis. Additionally, the estimated coefficient on LPOP in most of the models is close to 1 indicating the possible appropriateness of a rate transformation. Modeling a rate-transformed count outcome as a Poisson process is the same as modeling the original count outcome with the coefficient on the log of the rate divisor set equal to 1.

The similarity across the models, however, ends there. The Poisson models show evidence of distinct homicide-type and analysis-level specific processes.

At the NC level of analysis, increases in the percent of Hispanics in the total population are positively associated with only the total (ALL), gang related (GNG), stranger

Table 4.1: Maximum Likelihood coefficient estimates of baseline Poisson regressions with disaggregated homicides (1989–91) modeled on area macro-characteristics

	ALL	GNG	INS	FAM	KNO	STR	OTH
Unit of Analysis: Neighborhood Cluster (N = 343)							
INTERCEPT	−7.79**	−9.34**	−8.69**	−10.26**	−10.48**	−8.86**	−8.68**
	(0.611)	(1.441)	(1.311)	(1.406)	(1.120)	(1.740)	(1.134)
RESDEP	0.85**	0.79**	0.80**	0.85**	0.87**	0.93**	0.86**
	(0.029)	(0.077)	(0.061)	(0.063)	(0.051)	(0.081)	(0.052)
SHRHSP	0.97**	2.53**	0.08	0.42	0.11	0.97**	0.46*
	(0.122)	(0.264)	(0.283)	(0.301)	(0.241)	(0.361)	(0.244)
PNFH	0.37**	−1.47**	0.37	0.47	0.16	0.14	1.17**
	(0.183)	(0.559)	(0.382)	(0.389)	(0.328)	(0.537)	(0.321)
YMEN	1.10	0.65	6.64**	−4.89	7.10**	−3.23	2.25
	(1.365)	(3.300)	(2.797)	(3.165)	(2.351)	(4.088)	(2.506)
RESST	−0.03	1.38	−1.46	−1.09	−0.07	0.49	0.42
	(0.578)	(1.394)	(1.277)	(1.316)	(1.055)	(1.623)	(1.057)
LPOP	0.98**	0.95**	0.91**	1.16**	1.12**	0.90**	0.91**
	(0.067)	(0.161)	(0.143)	(0.151)	(0.122)	(0.189)	(0.123)
Unit of Analysis: Census Tract (N = 865)							
INTERCEPT	−7.08**	−9.34**	−8.15**	−10.93**	−8.42**	−8.82**	−7.16**
	(0.310)	(0.773)	(0.658)	(0.736)	(0.559)	(0.877)	(0.547)
RESDEP	0.87**	0.82**	0.86**	0.89**	0.94**	0.97**	0.84**
	(0.029)	(0.081)	(0.062)	(0.065)	(0.052)	(0.081)	(0.051)
SHRHSP	0.99**	2.58**	0.29	0.31	0.41*	0.87**	0.42*
	(0.111)	(0.244)	(0.263)	(0.278)	(0.221)	(0.332)	(0.219)
PNFH	0.24	−1.00**	0.27	0.49	0.59**	0.17	0.60**
	(0.162)	(0.485)	(0.336)	(0.353)	(0.281)	(0.463)	(0.281)
YMEN	0.41	1.57	0.69	−1.07	1.39	−0.54	0.91
	(0.763)	(1.966)	(1.554)	(1.909)	(1.255)	(2.298)	(1.308)
RESST	0.08	0.99	−2.21**	0.32	−0.18	0.54	0.64
	(0.439)	(1.148)	(0.992)	(1.033)	(0.781)	(1.190)	(0.727)
LPOP	0.91**	0.93**	0.91**	1.20**	0.92**	0.87**	0.76**
	(0.035)	(0.087)	(0.075)	(0.082)	(0.062)	(0.098)	(0.061)

** $p < 0.05$; * $p < 0.1$
Unstandardized coefficients with asymptotic standard errors in parenthesis

related (STR), and other types (OTH) of homicides. Similarly, increases in the proportion of NC households that are non-family is positively associated with other types (OTH) of homicides but negatively associated with gang related (GNG) homicides and not the rest. Increases in youthfulness of the underlying population, as measured by the proportion of young males in the underlying population (YMEN), are somewhat surprisingly associated only with instrumental (INS) and known expressive (KNO), but not other types of homicides. Residential stability (RESST) is not significantly associated with any of the homicide sub-types analyzed nor with the count of all homicides.

Similar findings were observed at the CT level, although there are some important distinctions. Unlike the NC level, the proportion of CT households that are non-family is no longer significantly associated with the total count of ALL homicides. For the instrumental homicides (INS), youthfulness of the underlying population (YMEN) is no longer significantly associated with the criterion measure, whereas increases in residential stability (RESST) are now negatively associated with instrumental homicide counts. In a similar manner, YMEN is no longer a significant determinant of known person related expressive homicides (KNO) while the proportion of CT households that are non-family (PNFH) and the proportion of CT population that is Hispanic (SHRHSP) are now both significantly and positively related to the criterion measure.

When compared to the total homicide counts (ALL) models at both the NC and CT levels, it is evident from these basic models that ignoring differences in homicide types can severely distort inferences. Moreover, there is evidence of the modifiable areal unit problem (MAUP), that is, the processes that operate at the NC level may, in some instances, be *qualitatively* different from the processes that operate at the CT level.

4.2. STRUCTURED ERROR MODELS

In order to avoid making the strong distributional assumptions of the Poisson models, as well as to allow for potentially heteroskedastic and/or autocorrelated errors, we next turn to the GCE framework described in Chapter 2. In all the GCE models estimated below, we retain the same definitions of the signal and error supports. That is, we define each signal as a sum of T binary choices with each choice defined over the support $\mathbf{z} = (0, 1)$ with uniform prior probabilities. Also, we define the error associated with each binary choice with a support of $\mathbf{v} = (-1, +1)$ also with uniform prior probabilities. Finally, for each sample, we define the number of binary choice being summed (T) as $3 \times \max(y)$. With these specifications, the resulting Lagrange multipliers are not directly comparable with the Poisson regression coefficients. Therefore it is more appropriate to display and discuss the marginal effects.

4.2.1. An example: All homicides at the NC level

In order to explain the workings of the GCE estimator, we first provide detailed analysis for the regression model of ALL homicides when analyzed at the NC level (see Table 4.2). Five

Table 4.2: GCE estimates of the Lagrange Multipliers and relevant marginal effects for a NC level model of ALL homicides (1989–91) using various error-structure specifications

	Model I		Model II[a]		Model III[b]		Model IV[b]		Model V[b]	
	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\lambda}$	$\hat{\gamma}$
INTERCEPT	−13.31**	−70.21**	−12.17**	−67.27**	−11.85**	−66.95**	−12.49**	−69.13**	−12.59**	−69.66**
	(0.643)	(3.454)	(0.607)	(3.401)	(0.603)	(3.466)	(0.620)	(3.461)	(0.623)	(3.482)
RESDEP	0.95**	5.00**	0.86**	4.76**	0.84**	4.75**	0.87**	4.82**	0.89**	4.90**
	(0.031)	(0.165)	(0.029)	(0.161)	(0.029)	(0.163)	(0.029)	(0.163)	(0.030)	(0.164)
SHRHSP	1.06**	5.56**	0.90**	4.97**	0.88**	4.99**	0.92**	5.08**	0.96**	5.32**
	(0.127)	(0.658)	(0.120)	(0.655)	(0.119)	(0.665)	(0.121)	(0.662)	(0.122)	(0.665)
PNFH	0.40**	2.13**	0.31*	1.72*	0.16	0.93	0.30*	1.67*	0.34*	1.85*
	(0.192)	(1.009)	(0.180)	(0.994)	(0.179)	(1.011)	(0.181)	(1.000)	(0.183)	(1.012)
YMEN	1.05	5.54	0.46	2.55	0.08	0.43	0.43	2.37	0.60	3.30
	(1.454)	(7.663)	(1.381)	(7.632)	(1.399)	(7.903)	(1.404)	(7.772)	(1.415)	(7.826)
RESST	0.08	0.41	0.21	1.14	0.40	2.26	0.15	0.81	0.14	0.77
	(0.605)	(3.188)	(0.576)	(3.183)	(0.574)	(3.242)	(0.584)	(3.236)	(0.587)	(3.248)
LPOP	1.09**	5.72**	0.98**	5.42**	0.95**	5.39**	1.02**	5.63**	1.02**	5.66**
	(0.070)	(0.369)	(0.066)	(0.364)	(0.066)	(0.372)	(0.068)	(0.371)	(0.068)	(0.373)
Model Diagnostics										
$\mathcal{L}^{D}_{GCE}$	16916.0		16644.4		16595.5		16648.0		16661.2	
Pseudo R^2	66.4		68.3		68.1		68.2		67.8	
$\mathcal{H}_e$	–		1.055		1.020		0.9995		0.996	
$\mathcal{H}_s$	–		1.166		1.195		0.9996		0.978	
$\mathcal{H}_*$	–		1.105		1.172		1.0002		0.983	

** $p < 0.05$; * $p < 0.1$;
Unstandardized Lagrange multipliers ($\hat{\lambda}$) and associated marginal effects ($\hat{\gamma}$) with asymptotic standard errors in parenthesis
Model I: $\mathbf{A}^* = \mathbf{0}$; Model II: $\mathbf{A}^* = \mathbf{I}$; Model III: $\mathbf{A}^* = \mathbf{I} + \mathbf{C}$; Model IV: $\mathbf{A}^* = \mathbf{I} + \exp(-\mathbf{D}) \odot \mathbf{C}$; Model V: $\mathbf{A}^* = \exp(-\mathbf{D})$
[a] for computing $\mathcal{H}_*$ encompassed model is I; [b] for computing $\mathcal{H}_*$ encompassed model is II

alternate models, corresponding to five types of error structures, were estimated. Model I most closely corresponds to the baseline Poisson regression specification as $\mathbf{A}^* = \mathbf{0}$. Model II allows for only heteroskedasticity. Models III, IV and V allow for heteroskedasticity and, respectively, first-order local error-correlation, first-order local error-correlation with distance-based decay, and global error-correlation with distance based decay. For each of these specifications, Table 4.2 displays the Lagrange multipliers as well as the computed marginal effects with associated asymptotic standard errors.

Visual comparison across the specifications illustrates the difference is parameter values that emerges as a result of the various forms of flexibility afforded to the basic model. The Lagrange multipliers as well as the marginal effects for Models II — V are invariably smaller in absolute value than those for model I. Also, the estimated standard errors for these parameters are invariably lower for models II — V than for model I. That is, affording the model some flexibility does seem to result in more conservative but more stable parameter estimates.

Although there are no sign changes across the various specifications, the reliability of predictors sometimes changes considerably across the specifications. For example, the proportion of NC households that are non-family (PNFH) would seem to have a significant positive association with all homicide counts (ALL) if one did not permit error structures. On the other hand, allowing heteroskedasticity along with first order local error-correlation (Model III) renders that predictor insignificant.

In order to select from the various specification we computed the composite relative information gain/loss measures for each of the models. These measures are displayed in the lower panel of Table 4.2. They are not applicable for Model I as it is never the alternate model. Model II is assessed against the null of Model I. Clearly, allowing heteroskedasticity alone is a desirable form of flexibility as $\mathcal{H}_* = 1.105 > 1$. Next Models III, IV and V are compared against Model II. Here we find that Models III and IV are *more* desirable than Model II ($\mathcal{H}_* > 1$) but Model V is less desirable ($\mathcal{H}_* < 1$). Between Models III and IV, Model III is clearly more desirable, as the gain in information about the noise component outweighs the loss in information about the signal by a much larger proportion. Therefore, it seems that among all the models tried here, a first-order local error-correlation structure is the closest approximation to the underlying data generating process.

Although this is not always to be expected, in the sample analyzed above, the Pseudo R^2 measure (defined here as the proportion of observed variance in the criterion measure explained by the predictors) increases relative to Model I where error structure was not permitted. This is somewhat surprising given that likelihood-based methods are designed for optimal predictive-accuracy *within* the sample being analyzed.

4.2.2. Disaggregated homicide models

We next estimated the models for each of the disaggregated homicides. Using the $\mathcal{H}_*$ criterion as a guide, we selected one of the several models as being the best. In all cases, Model III appeared to be the error specification closest to the underlying data generating process. Marginal effects implied by these models are presented in Table 4.3.

Table 4.3: GCE marginal effect estimates of area macro-characteristics on disaggregated homicides (1989–91)

	ALL	GNG	INS	FAM	KNO	STR	OTH
Unit of Analysis: Neighborhood Cluster (N = 343)							
INTERCEPT	−66.95**	−11.14**	−13.66**	−14.57**	−23.40**	−7.50**	−19.13**
	(3.466)	(1.418)	(1.631)	(1.620)	(2.003)	(1.246)	(1.914)
RESDEP	4.75**	0.59**	0.93**	0.96**	1.51**	0.60**	1.38**
	(0.163)	(0.067)	(0.074)	(0.073)	(0.093)	(0.056)	(0.087)
SHRHSP	4.99**	2.10**	0.06	0.50	0.12	0.58**	0.67**
	(0.665)	(0.246)	(0.334)	(0.332)	(0.404)	(0.247)	(0.392)
PNFH	0.93	−1.67**	0.01	0.17	−0.22	−0.26	1.35**
	(1.011)	(0.456)	(0.474)	(0.456)	(0.573)	(0.374)	(0.543)
YMEN	0.43	0.01	6.51**	−7.62**	10.49**	−3.92	0.82
	(7.903)	(3.352)	(3.644)	(3.803)	(4.331)	(3.036)	(4.389)
RESST	2.26	2.29**	−0.82	−0.35	0.96	1.06	1.98
	(3.242)	(1.337)	(1.575)	(1.528)	(1.847)	(1.160)	(1.787)
LPOP	5.39**	0.76**	0.97**	1.22**	1.84**	0.52**	1.36**
	(0.372)	(0.152)	(0.175)	(0.172)	(0.213)	(0.133)	(0.205)
Unit of Analysis: Census Tract (N = 865)							
INTERCEPT	−24.78**	−4.45**	−5.31**	−6.15**	−7.93**	−3.08**	−6.73**
	(0.702)	(0.308)	(0.320)	(0.353)	(0.403)	(0.252)	(0.361)
RESDEP	1.95**	0.27**	0.41**	0.40**	0.65**	0.26**	0.54**
	(0.065)	(0.028)	(0.029)	(0.031)	(0.037)	(0.023)	(0.033)
SHRHSP	2.12**	0.90**	0.15	0.16	0.28**	0.22**	0.27**
	(0.240)	(0.090)	(0.122)	(0.121)	(0.147)	(0.088)	(0.139)
PNFH	0.38	−0.30**	0.07	0.16	0.32**	0.01	0.32**
	(0.350)	(0.155)	(0.159)	(0.157)	(0.191)	(0.123)	(0.182)
YMEN	0.42	0.42	0.19	−0.66	0.78	−0.26	0.44
	(1.733)	(0.729)	(0.776)	(0.877)	(0.906)	(0.641)	(0.883)
RESST	0.56	0.45	−0.85**	0.28	0.05	0.21	0.51
	(0.955)	(0.401)	(0.463)	(0.455)	(0.528)	(0.318)	(0.470)
LPOP	1.99**	0.31**	0.42**	0.51**	0.62**	0.23**	0.49**
	(0.070)	(0.029)	(0.032)	(0.033)	(0.040)	(0.025)	(0.036)

** $p < 0.05$; * $p < 0.1$
Unstandardized marginal effects with asymptotic standard errors in parenthesis

Rather than examine these final models in isolation, we discuss our findings only in relation to the baseline Poisson models. That is, had we ignored the possible structure in the error, how different qualitatively would our inferences have been?

Three sets of findings are worth highlighting. First, at the NC level some predictors are now significantly related to the criterion measures while under the baseline Poisson model they were not (e.g., RESST in the GNG model and YMEN in the FAM model). Alternately, while PNFH was significantly related to the total count of homicides (ALL), its no longer a significant predictor of it under the GCE specification.

Second, these qualitative differences are *not* observed at the CT level of analysis. Therefore, allowing error structure flexibility further highlights the MAUP resulting in more differences in the macro-processes operating at the two different levels.

Finally, the fact that qualitative differences are not obtained between the baseline Poisson models and those obtained by the GCE setting at the CT level could be because the amount of error correlation at the CT level is lower than at the NC level. One finding that lends support to this interpretation is the absolute size of the regression coefficients presented in Table 3.1. There, it appears that the amount of spatial-correlation among the criterion measures is stronger (larger coefficients) at the NC level than at the CT level. Hence, even though correcting for error correlation may yield quantitative changes in inferences derived, qualitative findings could remain largely unaltered.

4.2.3. Spatially-lagged regressor models

In order to assess whether spatial lags of the most important predictor—resource deprivation (RESDEP)—may influence the criterion measures directly, we re-estimated the models with a spatial lag of the resource deprivation index included as one of the predictors. In most of the models, we find that the spatial lag term is highly significant, albeit with a smaller coefficient implying a distance decay effect.

Table 4.4 displays the marginal effects of the "own" and "cross" areal unit effects of resource deprivation (RESDEP) on the various disaggregated homicide types analyzed. Once again the findings presented here are for models deemed the "best" using the $\mathcal{H}_*$ criterion described before. Of course, changing the model specification may change inferences regarding all the remaining independent variables. However, for purposes of discussion, only the marginal effects of RESDEP and its spatial lag term are displayed in Table 4.4.

Four important points are worth highlighting here. First, resource deprivation of neighboring area has a positive effect on almost all types of homicides observed in the central area. That is, increases in the resource deprivation of surrounding areas are associated with an increase in the amount of violence one can expect in central units, even if the extent of resource deprivation in the central unit remains unaltered. This also implies that an increase in the resource deprivation in one area spreads out to its neighboring areas. Therefore, changes in resource deprivation have a *spill-over* effect on the amount of violence in neighboring areas.

Second, there is evidence of distance decay. That is, the effects of changes in neighboring area resource deprivation (i.e., the "cross" effects) are usually weaker than the effects

Table 4.4: GCE marginal effect estimates of RESDEP and its spatial lag on disaggregated homicides (1989–91)

	ALL	GNG	INS	FAM	KNO	STR	OTH
Unit of Analysis: Neighborhood Cluster (N = 343)							
RESDEP	3.91**	0.50**	0.73**	0.77**	1.29**	0.52**	1.29**
	(0.222)	(0.098)	(0.104)	(0.102)	(0.125)	(0.078)	(0.118)
RESDEP_L	1.45**	0.16**	0.35**	0.32**	0.39**	0.14**	0.17
	(0.245)	(0.114)	(0.116)	(0.114)	(0.136)	(0.088)	(0.132)
Unit of Analysis: Census tract (N = 865)							
RESDEP	1.45**	0.16**	0.29**	0.28**	0.51**	0.25**	0.43**
	(0.098)	(0.043)	(0.046)	(0.046)	(0.054)	(0.033)	(0.051)
RESDEP_L	0.87**	0.18**	0.21**	0.20**	0.25**	0.03	0.21**
	(0.111)	(0.049)	(0.053)	(0.052)	(0.061)	(0.038)	(0.059)

$^{**}p < 0.05$; $^{*}p < 0.1$
Unstandardized marginal effects with asymptotic standard errors in parenthesis

of an area's resource deprivation on its level of violence (i.e., the "own" effect).

Third, as with the other models, there is some evidence of differences in the processes linking disaggregate homicide types to resource deprivation. For instance, the spatial lag of RESDEP does not seem to be associated with local area stranger related homicides (STR) at the CT level, but it is a significant predictor of STR at the NC level. The reverse is obtained for the models relating spatially lagged RESDEP to the other homicides (OTH).

Finally, the differences between the "cross" and "own" effects are usually more pronounced at the NC level than at the CT level. One may speculate that this finding implies that the spatial spill-over at the CT level extends beyond the first order-contiguous neighbors. This is, however, pure speculation. Additional analysis is needed to fully support this view.

4.3. SUMMARY OF FINDINGS

The findings discussed in this chapter may be summarized as follows:

1. Whether or not we allow for spatial structure in the errors, we find some evidence of distinct homicide-type and analysis-level specific macro-processes. This finding is consistent with other studies that have recently reported similar differences (Kubrin, 2003). On the other hand, we also find evidence that resource deprivation is a strong, reliable and persistent predictor of all the homicide-types analyzed and at all levels of analysis. This finding is also consistent with prior research.

2. Extending traditional Poisson regression models to allow for auto-correlated struc-

tures in the errors yields two important findings. First, at the NC level, the differences in inferences regarding homicide-type specific macro-processes becomes more pronounced. Second, this finding is not replicated at the CT level. Given that the rough measure of auto-correlation in the outcomes used in this study suggest stronger spatial auto-correlation at the NC level than at the CT level of analysis, 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 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: If we assume away structure in the errors, then we are assuming *more* than we know. To the extent that this assumption is not supported by the data, we are probably deriving misleading and biased inferences from the data. Allowing flexibility in the moments simply means we let the data decide whether or not to use the flexibility. If the hypothesized error structure 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 we 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. Finally, there seems to be evidence of spill-over effects of the resource deprivation measure. For convenience we 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 provide better fit and more meaning in some contexts. 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.

These findings are further discussed in the next chapter in light of their practical implications for modeling rare crimes that are not randomly distributed across geographic space.

Chapter 5

Conclusions

In this chapter we discuss conclusions resulting from this research effort and enumerate some promising future avenues of research. The research effort described in the preceding chapters has important substantive, methodological and practical implications.

5.1. IMPLICATIONS

5.1.1. Substantive implications

We conclude from this analysis that ignoring knowledge of the spatial positioning of sample units can yield misleading inferences. This research effort confirms that some predictors would have been erroneously deemed irrelevant and some would have been erroneously deemed significant had spatial error correlation not been allowed. In addition, unlike the linear models case, we find changes in parameter and marginal effect estimates to be quite substantial across various hypothesized error structures. This implies that, unlike linear models, where ignoring error correlation does not bias results but leads only to inefficiency, in non-linear models for count outcomes, the parameters may in fact be substantially biased in addition to being inefficient.

On the other hand, we also find that ignoring spatial spill-over effects of predictors in one location on the criterion measures in neighboring locations can not only result in models with poorer fit, but can also lead researchers to underestimate or underpredict the overall (system-wide) effects of policy measures. For example, if policy measures are aimed at reducing resource deprivation in order to reduce violence, then changing resource deprivation levels in one area should have an effect on the levels of violence in neighboring areas as well. Hence, policy planning must take into account anticipated benefits that accrue not only from "own" area effects but also any "cross" area effects that may exist.

5.1.2. Methodological implications

From a methodological point of view, the GCE approach seems to offer a variety of desirable benefits over fully parametric likelihood based methods. Most importantly, it allows us to model heteroskedastic and auto-correlated error structures without making strong distributional assumptions. In small, finite, non-experimental data sets, it uses more flexible constraints and yields more stable/reliable solutions. As found in this analysis, some GCE models that allow error flexibility even offer higher in-sample predictive powers than those that do not permit error structures.

As described in this report, modeling autocorrelated error structures within the GCE framework does *not* entail an increase in the parameter space. That is, even though the errors are allowed to be heteroskedastic and autocorrelated, these structures are allowed non-parametrically. Additionally, in the error-correlation case, the dual objective function is not defined in terms of the full $N \times N$ spatial link matrix ($\mathbf{A}$). It is defined in terms of $\tilde{\mathbf{X}} = \mathbf{A}'\mathbf{X}$. Therefore, this matrix can be computed once and for all outside a numeric optimization routine (see the SAS program listed in Appendix B). This reduces the memory requirement and increases the efficiency of the optimization problem immensely.

5.1.3. Practical implications

The methodology applied in this project and described in this report is fast becoming available as part of conventional econometric software. The next release of SAS is slated to have a procedure explicitly dedicated to entropy based model estimation (PROC ENTROPY). As yet, this procedure does not have the explicit capability to model spatially correlated error structures. Future releases should provide enhanced capabilities.

As of now, spatial and non-spatial count outcome models may be modeled using manually programmed statements as shown in Appendix B. An outstanding and complicating issue is how one computes and accesses the spatial weight matrix. For this research, we computed the weight matrix using SpaceStat© and then imported it into SAS. Given the GIS capabilities of SAS, however, it is conceivable that SAS will be able to perform these computations efficiently in the near future.

Currently, a second complication is the size of this weight matrix. In the sample SAS code provided in Appendix B, the columns of the weight matrix are read in as N columns of a SAS data set. This is, of course, an immense waste of resources as the sparseness of most weight matrices is not utilized here. One promising feature that was not implemented in this project is to read a sparse contiguity or weight matrix as a set of variables (defined, in SAS, as an ARRAY) and perform the relevant computations on these variables. This is feasible, but promises to be a complicated programming task. It may also be possible to read a sparse weight matrix file directly into the Matrix module of SAS and perform the needed computations there. This too seems feasible but was not attempted here.

Finally, it is possible to perform the non-linear numeric optimization in other software such as GAUSS©. Since SpaceStat is able to read/write matrices to GAUSS format, it may be feasible to simple read the weight matrices directly into a GAUSS program that uses an

optimization module within GAUSS to do the analysis.

5.2. FUTURE RESEARCH

5.2.1. More flexibility

The flexibility of the GCE method was used for a very narrow purpose in this project; to allow for spatial error correlation in count outcome models. However, the GCE framework allows for a lot more flexibility than that. Future research may utilize this flexibility to, for example, gauge the effects of increasing the density of the support spaces in derived inferences. In this project we defined $L = M = 2$. By increasing $L > 2$ or $M > 2$, we should be able to recover higher moments of each and every signal and noise term. This may yield increased clarity and precision.

In addition, future research may utilize the flexibility of the GCE to allow for a mixture of binary (Yes/No) process with the count process. Such a setting would allow researchers to model the so-called zero-inflated count outcome models. In this research this issue was largely ignored with the aim of isolating and addressing the problem of error-correlation. With a large number of units yielding no homicide victims, however, especially at the CT level of analysis, models that permit this flexibility while allowing error-correlation may yield clearer insights into the underlying data generating processes.

Finally, when modeling true binomial counts where the maximum number of observable events are finite, known explicitly, and vary over the areal units, values of a *variable* T_n rather than a *fixed* T are available to the research. As such, the information-recovery problem should utilize this knowledge by re-difining moment constraints accordingly. Future research may extend the current formulation to allow that flexibility.

5.2.2. Endogenous and simultaneous processes

In this project we modeled the criterion measure on a set of exogenous predictors. This meant an ability to assume away the problem with endogenous regressors. In reality, of course, data are generated from more complex processes where some or several of the predictors may be endogenous. This commonly occurs, for example, in models of substantive spatial process where the outcomes in neighboring areas are theorized to influence the expected outcomes in the central areas. In addition, other neighborhood characteristics that are typically used in modeling areal data may also be endogenous (Dietz, 2002). In such settings, we obtain single equation models with potentially endogenous regressors and some form of an instrumental variable approach is required. Extending the GCE to model count outcomes with endogenous regressors in an instrumental variables framework is part of ongoing research. This avenue of research can also be extended to include simultaneous equation models for count outcomes where errors may be correlated within equation (across space) and across equation (within observations). In a similar manner, when repeated observations may be available for the same set of areal units over time, the GCE method should extend easily to provide a robust setting for assessing spatial and spatio-temporal dynamic.

5.2.3. More tests

Finally, future research may evaluate the predictive accuracy of the models either in repeated samples and/or in fresh samples. Assessing the ability of an estimator to yield accurate expectations when the true underlying data generating process is known to the researcher (such as in Monte Carlo experiments) is an ideal means of comparing competing estimators in their ability to properly recover the data generating process. In prior applications, GME and GCE estimators have been shown to have superior properties with such simulated data, especially when the sample sizes are small.

The increased stability of the GCE estimators in all finite samples seems to also suggest that they should provide superior out-of-sample predictions. Assessing the performance of models in fresh samples will provide a means of assessing whether or not the increased stability of estimated parameters translates into increased predictive powers of future events—an essential component of any modeling exercise if it is to have policy relevance.

Exploring the aforementioned extensions and performing detailed diagnostic testing are part of ongoing research. We believe, the findings from this project suggest that the GCE framework is well suited to incorporate the more realistic but more complex processes noted above with minimal reliance on strong distributional assumptions. Accordingly, it provides a more conservative but reliable analytical strategy for informing policy.

References

Anselin, Luc (1988). *Spatial Econometrics: Methods and Models*. Dordretch, Netherlands: Kulwer Academic Publishers.

Anselin, Luc (2002). "Under the hood. Issues in the specification and interpretation of spatial regression models" *Aggrecultural Economics* (forthcoming). Downloaded on May 15, 2003 from http://agec221.agecon.uiuc.edu/users/anselin/papers.html

Anselin, Luc and Anil K. Bera (1998). "Spatial dependence in linear regression models with an introduction to spatial econometrics." in Amman Ullah and David A. Giles (eds.), *Handbook of Applied Economic Statistics*, pg. 237–289. New York, NY: Marcel Dekker.

Anselin, Luc, Jacqueline Cohen, David Cook, Wilper Gorr and George Tita (2000). "Spatial Analysis of Crime." in David Duffee (ed.), *Criminal Justice 2000, Volume 4: Measurement and Analysis of Crime and Justice*, pg. 213–262 Washington, DC: National Institute of Justice (NCJ 182411).

Avakame, Edem F. (1998). "How different is violence in the home? An examination of some correlates of stranger and intimate homicide." *Criminology*. 36(3):601–632.

Bailey, Trevor C. and Anthony C. Gartell (1995). *Interactive spatial data analysis*. Essex, UK: Addison Wesley Longman Limited.

Balkwell, James W. (1990). "Ethnic inequality and the rate of homicide." *Social Forces*. 69:53–70.

Baller, Robert D., Luc Anselin, Steven F. Messner, Glenn Deane, and Darnell F. Hawkins (2001). "Structural covariates of U.S. County homicide rates: Incorporating spatial effects." *Criminology*. 39(3):561–590.

Belsley, David A. (1991). *Conditioining diagnostics: Collinearity and weak data in regressions*. NY: John Wiley and Sons.

Besag, Julian (1974). "Spatial interaction and the statistical analysis of lattice systems." *Journal of the Royal Statistical Society, Series B (Methodology)*. 36)2):192–236.

Block, Carolyn Rebecca and Richard L. Block (1998). *Homicides in Chicago: 1965—1995, Part I (Victim-Level Data), Codebook*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Braithwaite, John and Valerie Braithwaite (1980). "The effect of income inequality and social democracy on homicide." *British Journal of Criminology*. 20:45–53.

Bolduc, Denis, Bernard Fortin, and Stephen Gordon (1997). "Multinomial Probit estimation of spatially interdependent choices: An empirical comparison of two techniques." *International Regional Science Review*. 20(1&2):77-101.

Cameron, Colin A. and Pravin K. Trivedi (1986). "Econometric models based on count data: Comparisons and applications of estimators and tests." *Journal of applied econometrics*. 1:29–54.

Case, Anne (1992). "Neighborhood influence and technological change." *Regional Science and Urban Economics*. 22:491–508.

Clayton, David, and John Kaldor (1987). "Empirical Bayes estimates of age-standardized relative risk for use in disease mapping." *Biometrics*. 43:671–781.

Cressie, Noel and Ngai H. Chan (1989). "Spatial modeling of regional variables." *Journal of the American Statistical Association*. 84(406)393–401.

Cressie, Noel, and Timothy R. C. Read (1989). "Spatial data analysis of regional counts." *Biometrical Journal*. 31(6):699–719.

Cubbin, Catherine, Linda Williams Pickle, and Lois Fingerhut (2000). "Social Context and geographic patterns of homicide among U.S. Black and White males." *American Journal of Public Health*. 90(4):579–587.

Dietz, Robert D. (2002). "The estimation of neighborhood effect in the social sciences: An interdisciplinary approach." *Social Science Research*. 31:539–575.

Felson, Richard B. and Steven F. Messner (1998). "Disentangling the effects of gender and intimacy on victim precipitation in homicide." *Criminology*. 36(2):405–423.

Golan, Amos (2002). "Information and Entropy econometrics — Editor's view." *Journal of Econometrics*. 107(1&2):1–15.

Golan, Amos, George Judge and Douglas Miller (1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Chichester, England: John Wiley & Sons.

Golan, Amos, George Judge and Jeffrey M. Perloff (1996). "A maximum entropy approach to recovering information from multinomial response data." *Journal of the American Statistical Association*. 91(434):841–853.

Greene, William A. (2000). *Econometric Analysis* (IV ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Heagerty, Patrick J. and Subhash R. Lele (1998). "A composite likelihood approach to binary spatial data." *Journal of the American Statistical Association*. 93(443):1099–1111.

Imbens, Guido W., Richard H. Spady, and Phillip Johnson (1998). "Information-theoretic approaches to inference in moment condition models." *Econometrica*. 66(2):333–357.

Jaynes, Edwin T. (1957a). "Information theory and statistical mechanics." *Physics Review*. 106:620–630.

Jaynes, Edwin T. (1957b). "Information theory and statistical mechanics II." *Physics Review*. 108:171–190.

Jaynes, Edwin T. (1979). "Where do we stand on Maximum Entropy?" pp.15–118 in Raphael D. Levin and Myron Tribus (eds.) *The Maximum Entropy formalism* Cambridge MA: The MIT Press.

Judge, George G., R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee (1988). *Introduction to the theory and practice of econometrics*, 2nd ed. NY: John Wiley and Sons.

Kaiser, Mark S., and Noel Cressie (1997). "Modeling Poisson variables with positive spatial dependence." *Statistics and Probability Letters*. 35:423–432.

Kennedy, Leslie W., Robert A. Silverman, and David R. Forde (1991). "Homicide in urban Canada: Testing the impact of income inequality and social disorganization." *Canadian Journal of Sociology*. 16:397–410.

Kubrin, Charis E. (2003). "Structural covariates of homicide rates: Does type of homicide matter?" *Journal of Research in Crime and Delinquency*. 40(2):139–170.

Kullback, J. (1959). *Information Theory and Statistics*. New York, NY: John Wiley.

Land, Kenneth C., Patricia L. McCall, and Lawerence E. Cohen (1990). "Structural covariates of homicides rates: Are there any invariances across time and social space?" *American Journal of Sociology*. 95(4):922-963.

LeSage, James P. (1999). *Spatial Econometrics*. Last accessed on May 15, 2003 at http://www.rri.wvu.edu/WebBook/LeSage/spatial/spatial.html

Levine R. D. (1980). "An Information theoretic approach to inversion problems." *Journal of Physics A*. 13:91–108.

Mittelhammer, Ron C., George G. Judge, and Douglas J. Miller (2000). *Econometric foundations*. Cambridge UK: Cambridge University Press.

McMillen, Daniel P. (1992). "Probit with spatial autocorrelation." *Journal of Regional Science*. 32(3):335–348.

Morenoff, Jeffrey D., Robert J. Sampson, and Stephen W. Raudenbush (2001). "Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence." *Criminology*. 39(3)517–559.

Messner, Steven F., Luc Anselin, Robert D. Baller, Darnell Hawkins, Glenn Deane, and S. Tolnay (1999). "The spatial patterning of county homicide rates: An application of exploratory spatial data analysis." *Journal of Quantitative Criminology*. 15(4):423–450.

Messner, Steven F. and Luc Anselin (2003). "Spatial analysis of homicides with areal data." (forthcoming) in M. Goodchild and D. Janelle (eds.) *Spatially Integrated Social Sciences*. NY: Oxford University Press. Downloaded on May 15, 2003 from http://agec221.agecon.uiuc.edu/users/anselin/papers.html

Osgood, Wayne D. (2000). "Poisson-based regression analysis of aggregate crime rates." *Journal of Quantitative Criminology*. 16(1):21–43.

Parker, Karen F., and Patricia L. McCall (1999). "Structural conditions and racial homicide patterns: A look at the multiple disadvantages in urban areas." *Criminology*. 37(3):447–477.

Pinske, Joris and Margaret E. Slade (1998). "Contracting in space: An application of spatial statistics to discrete-choice models." *Journal of Econometrics*. 85:125–154.

Reiss, A and Jeffrey Roth (1994) *Understanding and preventing violence*. Washington, DC: National Academies Press.

Rosenfeld, Richard, Timothy M. Bray, and Arlen Egley (1999). "Facilitating violence: A comparison of gang-motivated, gang-affiliated, and non-gang youth homicides." *Journal of Quantitative Criminology*. 15:495–516.

Shannon, Claude (1948). "A mathematical theory of communication." *Bell System Technical Journal*. 27:379–423.

Smith, William R., Sharon Glave Frazee, and Elizabeth L. Davison (2000). "Furthering the integration of routine activity and social disorganization theories: Small units of analysis and the study of street robbery as a diffusion process." *Criminology*. 38:489–523.

Soofi, Ehsan S. (1994). "Capturing the intangible concept of information." *Journal of the American Statistical Association*. 89:1243–1254.

Waller, Lance A., Bradley P. Carlin, Hong Xia, and Elan E. Gelfand (1997). "Hierarchical spatio-temporal mapping of disease rates." *Journal of the American Statistical Association*. 92(438):607–617.

Williams, Kirk R., and Robert L. Flewelling (1988). "The social production of criminal homicide: A comparative study of disaggregated rates in American cities." *American Sociological Review*. 53(3)421–431.

Appendix A

Common Acronyms

List of technical acronyms used for the methods developed in Chapter 2.

ML	Maximum Likelihood
OLS	Ordinary Least Squares
ME	Maximum Entropy
CE	Cross Entropy
GME	Generalized Maximum Entropy
GCE	Generalized Cross Entropy
SMA	Spatial Moving Average
SAR	Spatial Auto-Regressive

List of acronyms referring to variables and aggregation levels. Refer to Chapter 3 for their definition and computation.

NC	Neighborhood Cluster
CT	Census tract
MAUP	Modifiable Areal Unit Problem
GNG	Gang related homicides
INS	Instrumental homicides
FAM	Family related homicides
KNO	Known person homicide
STR	Stranger related homicide
OTH	Other non-classified homicides
ALL	All homicides
RESDEP	Resource Deprevation
SHRHSP	Share of residential population that is hispanic
PNFH	Proportion of non-family households
YMEN	Proportion of residential population that is young and male
RESST	Residential Stability
LPOP	Log of the residential population

Appendix B

SAS Code

In this appendix we list a sample SAS Macro that may be used to estimate spatial and non-spatial count outcome models in the GCE framework. The macro presented below (called SEC) requires the following as inputs.

1. `DS`, a SAS data set containing the independent and dependent variables;
2. `CS`, a SAS data set containing *only N* columns corresponding to a contiguity matrix (sorted in the order of observations in `DS`);
3. `AS`, a SAS data set containing *only N* columns corresponding to a distance matrix (sorted in the order of observations in `DS`);
4. `DEP`, the name of a single dependent variable;
5. `IND`, a list of names that constitute the independent variables;
6. `INDL`, a list of names that constitute the independent variables that are modeled as having a spatial spill-over effect (if no variables is to be included then this must be set equal to a missing string, i.e., define `INDL` as `""`),
7. `DDD1`, the hypothesized spatial structure in the errors; and
8. `DDD0`, the null against which to compare this hypothesized structure (for constructing $\mathcal{H}_*$).

For specifying the spatial structures, the following codes are needed. `D0` – No error structure (corresponding to Model I in Table 4.2); `DH` – Only heteroskedastic errors (Model II in Table 4.2); `DH1C` – Heteroskedasticity and first-order local autocorrelation (Model III in Table 4.2); `DH1D` – Heteroskedasticity and first-order local autocorrelation with distance decay (Model IV in Table 4.2); and `DHGD` – Heteroskedasticity and global autocorrelation with distance decay (Model V in Table 4.2).

The SAS macro `SEC`, with explanatory comments, is listed below.

```
%macro sec(ds,cs,as,dep,ind,indl,ddd1,ddd0);

/* reading independent variables into a temporary data set */
data x(keep=&IND);
retain &IND;
set &ds; run;

/* reading independent variable for
   spatial-lagging into a temporary data set */
%if &indl ^= "" %then %do;
data xl(keep=&INDl);
retain &INDl;
set &ds; run;
%end;

/* reading dependent variable into a temporary data set */
data y(keep=&dep);
set &ds; run;

/* starting the IML procedure */
proc iml;

/* reading the SAS data sets into Matrices */
use y;
  read all var{&dep} into y;
  close y;
use x;
  read all into x0;
  close x;
%if &indl. ^= "" %then %do;
use xl;
  read all into x0l;
  close xl;
%end;
use &cs;
  read all into CN;
  close &cs;
use &as;
  read all into DD;
  close &as;

reset noname;

/* creating a variable name string for printing */
VNM = {INTERCEPT &ind.};
vnm = vnm`;

%if $indl ^= "" %then %do;
VNM = {INTERCEPT &ind};
VNML = {&indl};
vnm = vnm`//("SL_"+VNML)`;
%end;
```

```
/* initializing quantities */
t = 3*y[<>];
n=nrow(y);

x=J(n,1,1)||x0;

/* if applicable, computing spatial lag of relevant predictors */
%if &indl. ^= "" %then %do;
x0ll = (CN/(CN[,+]*J(ncol(CN),1,1)‘))*x0l;
x=J(n,1,1)||x0||x0ll;
%end;

k=ncol(x);

/* defining support space and priors */
z={0,1};
l = nrow(z);
p0 = J(n,l,(1/l));
v={-1,1};
m=nrow(v);
q0=J(m,1,(1/(m*n)));

/* using data to create the A matrix */

D0 = 0; /* i.i.d. case */
DH = i(n); /* no connectivity (only het no aut)*/

DH1C_ = i(n)+CN; /* only contiguity (het + local aut)*/
DH1C = DH1C_/(DH1C_[,+]*J(ncol(DH1C_),1,1)‘);

DH1D_ = i(n)+exp(-DD)#CN; /* distance decay exponential (only first order) */
DH1D = DH1D_/(DH1D_[,+]*J(ncol(DH1D_),1,1)‘);

DHGD_ = exp(-DD); /* distance decay exponential (het + global auto) */
DHGD = DHGD_/(DHGD_[,+]*J(ncol(DHGD_),1,1)‘);

/* computing the x_tilda (called xx here) matrix */
d = &ddd1.;
xx = d‘*x;

/* defining the objective function for numeric optimization */
start serc(bb) global(x,y,t,z,p0,v,q0,k,l,m,n,xx);
   b = bb‘;
   llf = b‘*x‘*y
           - t*J(n,1,1)‘*log((exp(x*b*z‘)#p0)*J(l,1,1))
           - t*log(J(n,1,1)‘*(exp(xx*b*v‘)*q0));
    return(llf);
finish serc;

/* defning the analytical gradients for the numeric optimization */
start g_sec(bb) global(x,y,t,z,p0,v,q0,k,l,m,n,xx);
```

```
   b = bb';
   omega = (exp(x*b*z')#p0)*J(1,1,1);
   p = (p0#exp(x*b*z'))/(J(1,1,1)'@omega);
   psi = J(n,1,1)'*(exp(xx*b*v')*q0);
   q = (q0'#exp(xx*b*v'))/(J(n,m,1)@psi);
   e = q*v*t;
   s = (p*z)*t;
   gr = (x'*(y-s)-xx'*e)';
   return(gr);
finish g_sec;

/* starting values for K lagrange multipliers set to 0 */
x0 = J(k,1,0);

/* option vector specifying maximization */
optn = {1 0};

/* calling the numeric optimization routine */
CALL NLPNRA(rc,xres_,"serc",x0,optn,,,,,"g_sec");

/* computing the hessian using finite difference methods */
CALL NLPFDD(ff,gg,hh,"serc",xres_,,"g_sec");

/* computing standard errors and test statistics for the parameters */
cov = inv(-hh);
ase = sqrt(vecdiag(cov));
bhat = xres_';
wald = (bhat/ase)##2;
pval = 1-probchi(wald,1);

/* recovering probabilities and signal based on Lagrange multipliers */
omega = (exp(x*bhat*z')#p0)*J(1,1,1);
p = (p0#exp(x*bhat*z'))/(J(1,1,1)'@omega);
psi = J(n,1,1)'*(exp((xx*bhat)*v')*q0);
q = (q0'#exp((xx*bhat)*v'))/(J(n,m,1)@psi);
s = (p*z)*t;

/* setting the evaluation point at sample mean */
xb = x[:,]';
pb = p[:,]';

/* computing marginal effects and associated standard errors */
gam = t*bhat*((z'##2)*pb-(z'*pb)##2);
d_gam_d_bhat = t*(((z'##2)*pb-(z'*pb)##2)*i(k)
                 + (((z'##2)*pb)+((z'##2)*pb)*(z'*pb)
                      -2*(z'*pb)##3)*bhat*xb');
cov_gam = d_gam_d_bhat*cov*d_gam_d_bhat';
ase_gam = sqrt(vecdiag(cov_gam));
wald_gam = (gam/ase_gam)##2;
pval_gam = 1-probchi(wald_gam,1);

/* priting GCE parameter estimates */
```

```
print "GCE Results for Model with DEP=&DEP. and A=&ddd1.";

print bhat[format=9.4 rowname=vnm colname="LAMBDA"]
      ase[format=9.4 colname="ASE"]
      wald[format=6.2 colname="WALD"]
      pval[format=6.2 colname="PVAL"]
      gam[format=12.4 colname="GAMMA"]
      ase_gam[format=9.4 colname="ASE"]
      wald_gam[format=6.2 colname="WALD"]
      pval_gam[format=6.2 colname="PVAL"];

/* model goodness of fit diagnostics */
vy=((y-y[:])`*(y-y[:]))/n;
sse = (y-s)`*(y-s);
psr = 100*(1-(sse/(n*vy)));

/* priting diagnostics */
print "Basic Model Diagnostics";
r = {"Number of Obs Used" "Optimum Value of Objective Function"
      "Sum of Squared Errors" "Pseudo R-Sq"};
diag = n//ff//sse//psr;
print diag[format=12.1 rowname=r];

/* computing entropy for signal and noise terms (alternate model) */
h_p1 = -J(n,1,1)`*(p#log(p))*J(l,1,1);
h_q1 = -J(n,1,1)`*(q#log(q))*J(m,1,1);

/* some re-definitions for computing parameter estimates for the null model */
d = &ddd0.;
xx = d`*x;
CALL NLPNRA(rc,xres_n,"serc",x0,optn,,,,,"g_sec");
bhat_n = xres_n`;

/* recovering probabilities and signal based on Lagrange multipliers */
omega_n = (exp(x*bhat_n*z`)#p0)*J(l,1,1);
p_n = (p0#exp(x*bhat_n*z`))/(J(l,1,1)`@omega_n);
psi_n = J(n,1,1)`*(exp((xx*bhat_n)*v`)*q0);
q_n = (q0`#exp((xx*bhat_n)*v`))/(J(n,m,1)@psi_n);

/* computing entropy for signal and noise terms (null model) */
h_p0 = -J(n,1,1)`*(p_n#log(p_n))*J(l,1,1);
h_w0 = -J(n,1,1)`*(q_n#log(q_n))*J(m,1,1);

/* computing net information gain/loss measures */
model_e = (h_p1 || h_w1) // (h_p0 || h_w0) ;
rr = {"Alternate" "Null"};
cc = {"H(p)" "H(q)"};
s_l = h_p1 / h_p0;
n_g = h_w0 / h_w1;
netg = n_g / s_l;
lossgain = s_l // n_g // netg;
rrr = {"Signal Loss" "Noise Gain" "Net Gain"};
```

```
/* printing net information gain/loss measures */
print "Assessing Alternate: &ddd1. against Null: &ddd0.";
print model_e[format=12.4 rowname=rr colname=cc];
print lossgain[format=8.4 rowname=rrr];

quit;

%mend sec;


/* defining a macro that calls SEC recursively for several alternate and
   null error structures with a fixed substantive model */

%macro models(ds,cs,as,dep,ind,indl);
%sec(&ds,&cs,&as,&DEP,&IND,&INDL,D0,D0);
%sec(&ds,&cs,&as,&DEP,&IND,&INDL,DH,D0);
%sec(&ds,&cs,&as,&DEP,&IND,&INDL,DH1C,DH);
%sec(&ds,&cs,&as,&DEP,&IND,&INDL,DH1D,DH);
%sec(&ds,&cs,&as,&DEP,&IND,&INDL,DHGD,DH);
%mend models;

/* defning some strings of variable names */
%let ind=RESDEP SHRHSP PNFH YMEN RESST LPOP;
%let indl=RESDEP;

/* calling the macro MODELS that will produce estimates of the five
   Census tract level models presented in Table 4.2 of this report */
%models(CTP,CONT,DIST,ALL,&IND,&INDL);
```