

Statistical Methods in Criminological Sciences using Systat

Introduction

Criminologists and criminal justice researchers depend on analytic methods that they import from other disciplines. As in several other fields of social research (for example, political science and sociology), the analytic methods of criminology and criminal justice originated in statistics, econometrics, epidemiology, and psychometrics. Developments in these areas are occurring at a rapid pace, and a set of papers that concentrated on promising analytical techniques would soon be amusingly obsolete. However, the range of analytic methods in criminology and criminal justice will continue to expand during the 21st century.

Why analytical techniques become obsolete is a question to ponder. More than two decades ago, Hubert Blalock suggested that the most important challenge to empirical research in sociology was not to develop more sophisticated analytical methods. Instead, according to Blalock, the key element in advancing knowledge about society was a better understanding of data and measurement.

This observation is equally true for inquiry in crime and justice today. Unlike statistical techniques, criminologists and criminal justice researchers play a major role in controlling and shaping the data they use. The form and content of data collection can greatly expand or limit the range of questions that scholars might address.

However the purpose of the note is not to go into the intricacies of the above issues. The purpose was only to convey that data analysis heavily depends on measurement decisions.

Criminologists and criminal justice researchers use statistical analysis on different types of data. **Systat** provides a bouquet of data analytical methods such as: Interpreting Data Distributions - Graphical Techniques, Measures of Central Tendency, Measures of Dispersion, Exploratory Data Analysis, Point Estimation and Confidence Intervals, From Estimation to Hypothesis Testing, Data Analysis With Categorical Variables, Bivariate Correlation and Regression, Multiple Regression and Partial Correlation, Nonparametric or Distribution-Free Statistical

Tests, Regression Analysis with a Dichotomous Dependent Variable - Logit and Probit Models and many more.

Data Clustering and Analytical Issues

One of the most notable characteristics of crime is that it clusters. Criminal acts do not extend evenly over space, and they are not constant over time. The first criminologists noticed these variations, and patterns in space and time were a major concern of the discipline from its beginning. Quetelet and other 19th-century statisticians closely studied differences in crime across communities. In 1837, Poisson derived his famous count distribution in a time-series study of criminal convictions.

At the beginning of the 21st century, clustering and its implications still play a central role in the study of crime. More generally, however, clustering occurs in both temporal and cross-sectional data and in both individual and aggregate analyses. The two basic forms of analysis and two basic data structures create four possible combinations: individual temporal, aggregate cross-sectional, individual cross-sectional, and aggregate temporal. Although each combination poses special problems of its own, all four generate similar clustering issues. Cluster effects will likely continue to challenge and fascinate criminal justice scholars well into the future.

Currently, the best understood clustering issues involve aggregate temporal analyses, such as trends in drug use or the fear of crime. Here, clustering arises because observations that are close in time tend to be more similar to each other than to observations in the distant past or future. The autocorrelation that this clustering generates is the subject of a large and ever-growing statistical literature from which criminologists often draw.

Systat Time Series implements a wide variety of time series models, including linear and nonlinear filtering, Fourier analysis, seasonal decomposition, nonseasonal and seasonal exponential smoothing, and the Box-Jenkins approach to nonseasonal and seasonal ARIMA. The general strategy for time series analysis is to:

- Plot the series using **T-plot**, **ACF**, **PACF**, or **CCF**.
- Transform the data to stabilize the variance across time or to make the series stationary using **Transform**.

- Smooth the series using moving averages, running medians, or general linear filters using **LOWESS** or **Exponential** smoothing.
- Fit your model using **ARIMA**.
- Examine the results by plotting the smoothed or forecasted results.

Before performing a particular time series, you can specify how missing values should be handled.

Clustering occurs in aggregate cross-sectional studies in "contextual" or "multilevel" analyses. Examples include studies of the effects of neighborhood conditions on victimization risks and sentencing outcomes for defendants in different court systems. Here, researchers examine both individual effects and the aggregate effects of the clusters. Statistical models for this situation were the subject of much attention in the 1990s.

Systat Mixed regression estimates models containing combinations of fixed and random effects for response data having a normal distribution. **Mixed models**, or **multilevel models**, have also been referred to as "hierarchical linear models", "random coefficient models", and "variance component models".

These models require a data structure in which observations having a common characteristic can be classified into identifiable groups, known as level-2 units, resulting in nesting of the observations within the level-2 units. Mixed regression uses random effects to account for dependencies in the data due to this nesting structure, allowing simultaneous analysis of individuals and the groups to which the individuals belong. For an individual level-2 unit i , the model for mixed regression is:

$$\mathbf{y}_i = \mathbf{W}_i\boldsymbol{\alpha} + \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$$

Where \mathbf{y} is the dependent variable, \mathbf{W} is a design matrix for fixed effects, $\boldsymbol{\alpha}$ is a vector of fixed regression parameters, \mathbf{X} is design matrix for random effects, $\boldsymbol{\beta}$ is a vector of effects specific to unit i , and $\boldsymbol{\varepsilon}$ is a vector of residuals. Models without random effects parallel standard regression models, but use marginal maximum likelihood to derive the parameter estimates instead of least-squares techniques.

Researchers often use mixed regression for the analysis of both **clustered** and **longitudinal data**. In clustered data, observations from different subjects are nested within a larger group, such as students within schools; random effects represent differences between the clusters. In contrast, for longitudinal data, observations are nested within each subject. In this case, the individual can be viewed as the "cluster", so random effects represent differences between subjects.

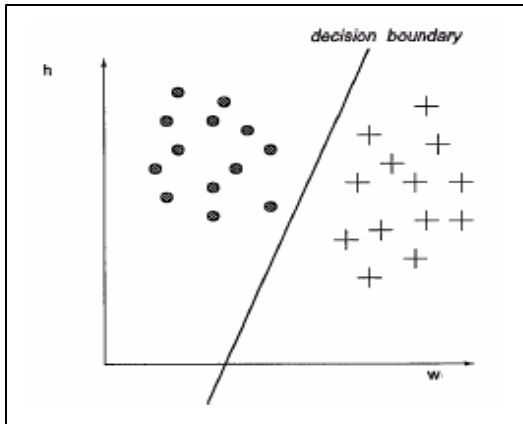
Spatial Data Analysis Tools

The spatial concentration of crime in hot spots leads naturally to their representation on crime maps. Maps of crime incidents permit rapid identification of the geographic location of crime hot spots, but by themselves they contribute little to understanding why crime is concentrated in certain locations. A crucial aspect of pattern recognition techniques such as hot spot analysis is the determination of the extent to which patterns on the map reflect "true" clusters or outliers or whether they are spurious. As is well known, simple visual interpretation of the map is inadequate in this respect because the human mind is conditioned to find meaning and identify patterns and clusters, even when the data represented may be purely random. The use of sound cartographic principles alone does not ensure that a proper interpretation is obtained. What is needed is a careful structuring of the visualization strategy while supplementing the visual aspects with quantitative information.

Spatial statistics involve a variety of methods for analyzing spatially distributed data. **Systat Spatial Statistics** covers two principal areas: fixed-point methods (kriging and Gaussian simulation) and random-point methods (nearest-neighbor distances, polygon area/volumes, quadrat procedures).

Spatial statistics compute a variety of statistics on a 2-D or 3-D spatially oriented data set. Variograms assist in the identification of spatial models. Kriging offers 2-D or 3-D kriging methods for spatial prediction. Simulation realizes a spatial model using Monte Carlo methods. Finally, a variety of point-based statistics are produced, including areas (volumes) or Voronoi polygons, nearest-neighbor distances, counts of polygon facets, and quadrat

counts. Graphs are automatically plotted and summary statistics are printed for many of these statistics.



Discriminant functions are the basis for the majority of pattern recognition techniques. The mathematical definition of such a decision boundary is a "**discriminating function**". It is a function that maps our input features onto a classification space—in the example above, by defining a plane that would separate the two clusters. Pattern

classification techniques fall into two broad categories: numeric and non-numeric. Numeric techniques include deterministic and statistical measures, which can be considered as measures made on the geometric pattern space. Non-numeric techniques are those that take us into the domain of symbolic processing, e.g. methods such as fuzzy sets. **Systat Discriminant Analysis performs linear and quadratic discriminant analysis, providing linear or quadratic functions of the variables that "best" separate cases into two or more predefined groups.**

Summary

The above paragraphs just mention a bird's eye view of methods available in Systat. But Systat provides a powerful statistical and graphical analysis system in a graphical environment using descriptive menus and simple dialog boxes. Simply pointing and clicking the mouse can accomplish most tasks. Systat's command language provides functionality not available in the dialog box interface. Matrix procedure allows you to use matrix algebra to specify statistical analyses and perform data management tasks. Systat can open data files saved in various formats including ArcView (*.SHP). Systat MAP produces maps in oblique gnomonic, oblique stereographic, Mercator conformal, oblique orthographic, Lambert, Robinson, sinusoidal, Miller, and Peters projections. Options are available for filling map polygons with colors, shading, or patterns to indicate the values of a variable (for example, average crime within each district). It is also possible to include the value of a variable within a polygon by using contours or icons.