Law Enforcement Resource Allocation (LERA) Visualization System

April Webster (awebster@cs.ubc.ca)
Michael Welsman-Dinelle (mwelsman@cs.ubc.ca)

Problem Description

Hundreds of law enforcement agencies exist in the United States. The primary role of these agencies is to deal with local criminal activity. They do so through a wide range of policy and management approaches and by applying different types of technology to the job. Crime is easy to track as an isolated phenomenon, but it is much more difficult to assess the real impact of different policy decisions. Does technology correlate with higher officer performance? Do anti-drug programs help to reduce youth crime? These are the kinds of questions administrators and crime analysts must answer when deciding the best way to allocate limited agency resources.

Our main goal is to enable crime analysts to answer these questions by bringing together both crime data and crime enforcement policies into a single visualization system. Our system will allow analysts to isolate the correlations between policies and crime rates. It is thought that this information, along with domain knowledge, will allow analysts to come up with new insights into the interaction between criminal activity and law enforcement agency management policies.

Dataset

Two different data sets will be combined for use with our information visualization system:

(1) US Federal Bureau of Investigation (FBI) *Uniform Crime Reports* (http://www.fbi.gov/ucr/ucr.htm)

(2) US Department of Justice *Law Enforcement Management and Statistics* (LEMAS) data (http://www.ojp.usdoj.gov/bjs/lawenf.htm).

The crime report data (1) is publicly available as comma-separated value (CSV) on the US FBI website. This data is numerical and is provided on an annual basis from 1985 to 2005. It is comprised of seven types of crime categorized

either as violent (murder, forcible rape, robbery, aggravated assault) or non-violent (burglary, larceny-theft and motor vehicle theft).

LEMAS data (2) is provided in HTML format or in original data files and is only publicly available online for the year 2000. It is comprised of specialized units (a Boolean value indicating whether or not the agency operates each of eleven types of special units, such as juvenile crime, drug education in schools, drunk drivers, cybercrime child abuse, etc), investment in technology (digital imaging, video cameras, and computers), training (hours in academy, field) and budgets. For the remainder of this document, we will refer to any LEMAS data variable as a “program”.

Both datasets are indexed by US local law enforcement agency, of which there are over 770 nationwide. This allows us to merge the two data sets to look at correlations between crime rates and how agencies are structured or how they spend their budgets on various programs. Unfortunately, LEMAS data is not collected annually so we are limited to the year 2000 and cannot present changes in the data over time.

The merged data set contains approximately 20 fields indexed by law enforcement agency. There are 0-100 (on average approximately 30) agencies per state. Only agencies with 100 or more sworn officers are included in the LEMAS data set. This is why Wyoming, for example, has no entries in the LEMAS data. The LEMAS dataset is the limiting dataset in terms of the represented agencies. As a result, we will consider only those agencies that are inlcuded in the LEMAS data set.

Tasks

The following are three tasks that a crime analyst involved in resource allocation would need to perform in the course of their work. It is important to note that each of these is representative of a larger class of tasks.

Task #1: The analyst would like to investigate how field training (number of hours of training) impacts the different types of crime.

Task #2: The analyst would like to determine if juvenile crime units have an effect on motor vehicle theft. Do drug education programs have an effect on motor vehicle theft? What is the effect of having both a juvenile crime unit and a drug education program on the incidence of motor vehicle theft? Which program is more effective at reducing motor vehicle theft (a juvenile crime unit only, a drug education program only, or both a juvenile crime unit and a drug education program)?

Task #3: The analyst would like to establish which programs have been the most effective in reducing the violent crime rate.

Proposed Information Visualization Solution

Searching for relationships between different variables in a large dataset can be a time consuming and frustrating task for a crime analyst. In many cases this task is performed using a statistics program (e.g., SAS, SPSS, or R) or data analysis program (e.g., Microsoft Excel) that does not allow the analyst to interactively explore the data to find more interesting trends and correlations that may not be detectable solely by examining the results of statistical analysis.

Previous Work

We considered four different existing tools for interactively visualizing correlation, each of which have been presented in the literature. In particular, we looked at Parallel Coordinates, Table Lens, general graph drawing techniques, and scatterplots. In our investigation of these potential solutions and their suitability for our particular problem, we eliminated all but the scatterplot.

Parallel Coordinates were proposed in [5] as a technique for extending the scatterplot beyond three dimensions. We determined that this technique would not be an appropriate solution because for our first task we only need to consider a couple of dimensions. For our second task, we could use Parallel Coordinates to compare a single program with multiple crimes types. Unfortunately, this would require that we repeat the program of interest on alternating axes and this would not be a good use of screen space. As this technique would be unsuitable for two of our three tasks, we felt that we could find a better solution that would work for all three of our task classes.

Table Lens was the solution put forth in [4] for supporting visualization of very large data tables using a fisheye technique to allow users to focus in on label information (such as the numerical value associated with a particular bar in a bar chart). This technique was removed from the list of possible solutions because in our proposed tasks there is no need to be able to focus on the detailed numerical information for particular law enforcement agencies. We are interested in communicating trends and patterns, not the detailed information. We also felt that the spatial cue provided by the scatterplot would be more beneficial to the user than the comparison of distributions in the Table Lens solution.

Similarly, we decided that graph drawing techniques would not be a fitting solution as there is no compelling information that we could use to connect local agencies by edges. Due to the large number of data points, the resulting graph would be very cluttered and difficult to interpret.

Finally, we were left with the scatterplot technique. This is the tool commonly used by most crime analysts to answer the questions proposed by our representative task set. However, the tools that are currently available to crime

analysts are neither interactive nor flexible. Obviously, an information visualization system is needed.

Solution Details

Although an analyst may create a scatterplot using a statistics package, each individual scatterplot must be manually produced for every possible pairwise combination of variables of interest. Similarly, these systems do not allow the user to interact with the scatterplot to change the correlation variables, nor do they easily allow the user to change the input data or remove outliers.

Our system proposes a more thoughtful and flexible approach to the scatterplot. In particular, we plan to implement an interactive scatterplot system that allows the user to quickly and effectively explore a large data set to discover possible correlations between variables. Although we intend to provide a different configuration of scatterplots for each of our three tasks (as detailed below), there are some general characteristics that will be shared. In particular, we aim to create a scatterplot system that incorporates the following features:

1. Use of filtering to select one or more states
2. Ability to easily remove outliers, manually and automatically
3. Ability to add regression curves

The specific features and/or layout of scatterplots in our system are dependent on the type of task.

The first task involves a one-to-one comparison between a single type of program and each of the seven different types of crime. Our proposed solution is a single interactive scatterplot in which the relationship between training requirements and each type of crime is displayed. Type of crime will either have a binary colour encoding to distinguish between violent and non-violent crime (if the user has selected this option), or a seven-way colour encoding to distinguish between the individual types of crime. Each respective colour encoding will be automatically incorporated into the visualization and will represent an optimal encoding based on Tufte's principals.

The second task involves a many-to-one comparison between different programs and a single type of crime. Our proposed solution is a small multiple view [3] of scatterplots. One potentially novel feature that we plan to incorporate is simultaneous interactive marking between scatterplot small multiple views, as demonstrated in Figure 2. This will allow the user to outline the perceived shape of a distribution of points on one scatterplot and, have this outline be simultaneously drawn on the other scatterplots. The purpose of marking is to allow the analyst to more effectively compare the distribution of points between the different scatterplots.

The third task involves a many-to-many comparison between a type of program and a type of crime. Again, our proposed solution is a small multiple view of scatterplots. We intend to investigate a number of different possibilities for the layout of small multiples including a scagnostic SPLOM layout (or some other correlation-related diagnostic) [6] or a Trellis layout [1].

Scenario of Use

The crime analyst for the state of California has been asked by his supervisor to determine if and how the level of field training impacts burglary for the state of California. To answer this question, the analyst inputs "field training" as the independent variable and "burglary" as the dependent variable into the LERA interface. A scatterplot is automatically generated by LERA. As the analyst is only interested in the state of California, he uses the filter to select this state. Finally, the analyst chooses the option of having a regression line automatically fit to the scatterplot. He interprets the resulting scatterplot and reports the information to his supervisor. (The resulting scatterplot is presented in Figure 1).

Personal Expertise

We have limited domain expertise, but if we do decide to include an evaluation component we would like to consult a domain expert. April has utilized Excel to conduct regression analysis on a number of different data sets in a variety of domains.

Proposed Implementation Approach

The programming language that we will use to develop our system will be Java. We may also use Prefuse which provides support for scatterplots. And, we hope to be able to find a statistics toolkit for use in outlier removal and regression curve generation.

The following is a list of the main components of our system ordered from most to least important. The earlier elements in the list should generally be implemented first:

1. Data Management
    - Generate Java objects or an array-based representation of the data from the data files.
    - Priority: critical

2. Outlier removal
    - Provide the user with both automatic and manual methods of removing outliers

- Automatic method would add power to the system by providing more support for statistical analysis
- Manual method would give the user additional flexibility in outlier removal
- Sophistication of feature implementation is contingent on the ability to find a statistics toolkit
- Priority: high

3. Regression curves
    - Addition of regression lines to a scatterplot is a typical
    - Sophistication of feature implementation is contingent on the ability to find a statistics toolkit
    - Priority: high

4. Ordering of small multiples
    - Utililize statistical methods to find good orderings
    - Sophistication of feature implementation is contingent on the ability to find a statistics toolkit
    - Priority: high

5. Aggregation/multilevel
    - An information visualization feature
    - A focus and context feature (context provided for some states using an appropriate measure of centrality such as mean or median; focus provided by plotting individual law enforcement agencies for a state(s) of interest)
    - Priority: high

6. Marking
    - A graphics feature
    - Priority: medium

7. Blurring
    - An image processing problem
    - It isn't clear that this will be an absolute slam dunk as the same basic functionality is provided by marking
    - Would "do the obvious" – apply Gaussian technique to colour to get colour blurring
    - Priority: low

Other Project Work

Ideally, we will also complete a limited study to assess the usability of the system that we have created. This should be done by allowing a domain expert to use our tool and compare it to existing tools.

Milestones

Date	Milestone
Nov 2, 2007	**Phase 0**: • Download & prepare all relevant data • Search for scatterplot & statistics toolkits • Review references
Nov 9, 2007	**Phase 1**: • Implementation of single scatterplot • Determine scope of statistical analysis
Nov 14, 2007	Project Update
Nov 23, 2007	**Phase 2 – Part 1**: • Implement statistical methods (outlier removal, regression lines, etc) • Determine scope of evaluation component.
Nov 30, 2007	**Phase 2 – Part 2**: • Implementation of small multiples • Begin preliminary evaluation
Dec 7, 2007	**Phase 3**: • System evaluation • Implementation of optional features • Stable version of LERA • Draft report
Dec 12, 2007	Final Presentation
Dec 14, 2007	Final Report

Illustrations

Figure 1: Basic interface with one scatterplot

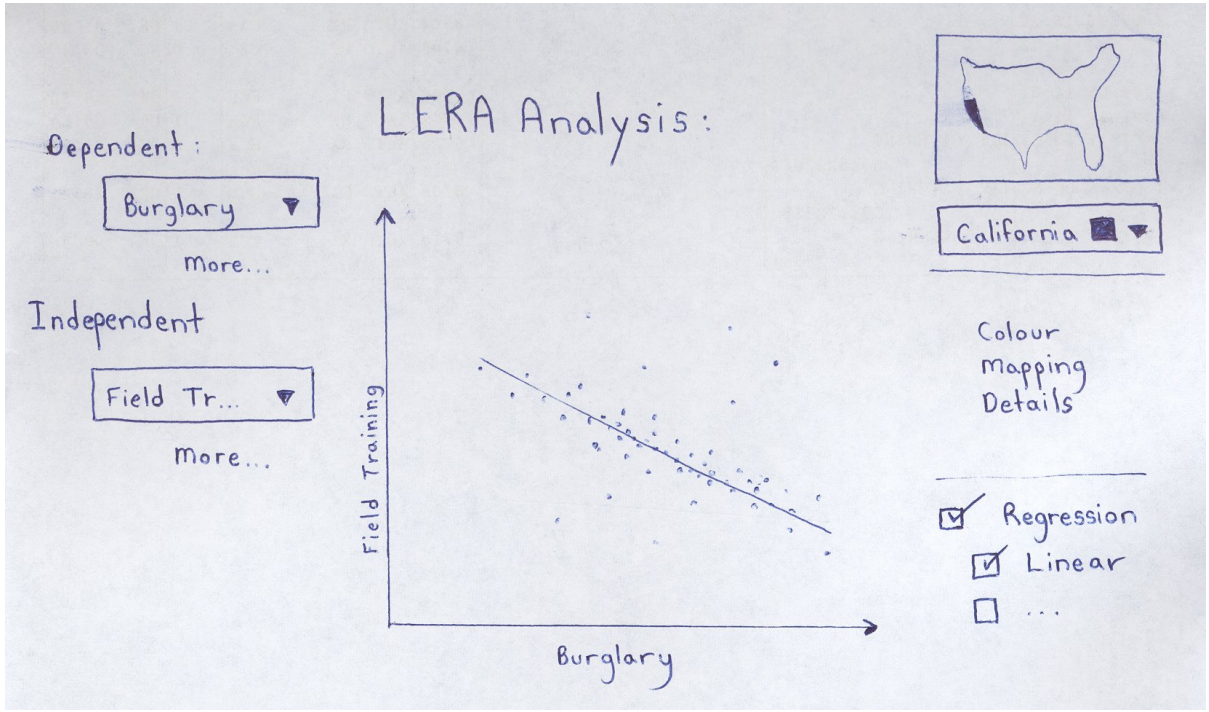
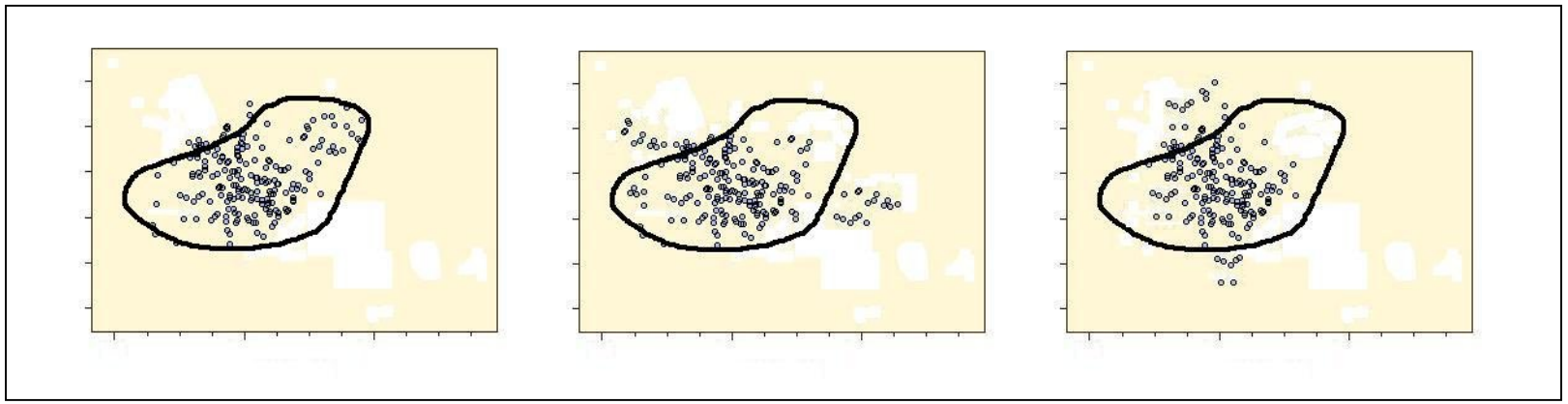


Figure 2: Marking example



References

[1] R. A. Becker, W. S., Cleveland, and M. J. Shyu: The Visual Design and Control of Trellis Display. Journal of Computational and Statistical Graphics 1996(5): 123-155.

[2] Yves Chiricota, Fabien Jourdan, and Guy Melancon: Metric-Based Network Exploration and Multiscale Scatterplot. Proc Info Vis 2004: 135-142.

[3] Alan MacEachren, Xiping Dai, Frank Hardisty, Diansheng Guo, and Gene Lengerich: Exploring High-D Spaces with Multiform Matrices and Small Multiples. Proc Info Vis 2003: 31-38.

[4] Ramana. Rao and Stuart K. Card: The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. ACM SIGCHI 1994: 318-322.

[5] Edward J. Wegman: Hyperdimensional Data Analysis Using Parallel Coordinates. Journal of the American Statistical Association 1990 (85): 664-675.

[6] Leland Wilkinson, Anushka Anand and Robert Grossman: Graph-Theoretic Scagnostics. Proc Info Vis 2005: 157-164.