Law Enforcement Resource Allocation (LERA) System

Michael Welsman-Dinelle and April Webster

University of British Columbia

ABSTRACT

LERA, the information visualization system presented in this paper, is an interactive, scatterplot visualization system that was designed to support crime analysts in exploring the effects of various law enforcement administration programs and policies on crime rates. Several important features were incorporated in our system to meet this goal. In particular LERA provides the user with the ability to generate a linear regression line for a scatterplot to demonstrate more succinctly the relationship between two variables; the ability to detect and remove outliers; the ability to view several linked scatterplots in a small multiple; focus and context capabilities in which focus is provided by a state average point; and, filtering at the state level to help the user cope with visual clutter. General information visualization design principles were also followed to ensure visual salience.

An initial prototype of LERA was evaluated through usability testing and found to be easy to use and easier to use than Excel, the data analysis tool commonly used by crime analysts for the task of exploring the effects of programs and policies on crime rates. User feedback was reviewed and some of the suggested changes were incorporated into the version of LERA presented in this paper.

1 INTRODUCTION

1.1 Crime Analysis

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 including the introduction of programs aimed at crime reduction 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 how best to allocate limited agency resources.

To answer questions such as these, crime analysts typically have a "statistical tool bag" [4] at their disposal. The types of tools that can be found in a crime analyst's tool bag are typically not very statistically sophisticated. Most law enforcement agencies utilize very basic statistical measures and analysis methods such as frequencies (82%), measures of centrality (i.e., means, medians and modes; 64%), correlation (58%), and cluster analysis (60%) [4]. Fewer agencies still employ standard deviations (49%) and regression analysis (36%) [4]. The most sophisticated type of analysis that crime analysts typically perform in the course of their work is linear regression analysis.

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.

1.2 Tasks

As indicated in the Introduction, this paper describes a system designed to support a crime analyst in the course of his/her work. In particular, we chose to address three different classes of tasks typically required of a crime analyst in determining how different law enforcement administration policies and programs affect crime rates. It is important to note that each of these three tasks is a specific example of one of the three larger classes of tasks we chose to support.

- Task #1: The crime analyst would like to investigate how annual operating budget per resident impacts the different types of crime (a one-to-one comparison between a single type of program and a particular crime rate for each type 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.

1.3 Data

Two different data sets were combined for use with our information visualization system. The first data set contains crime rate data and was obtained from the US Federal Bureau of Investigation (FBI) Uniform Crime Reports [5]. The second data set contains administration and management data and was obtained from the US Department of Justice Law Enforcement Management and Statistics (LEMAS) data [9].

The crime report data 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 nonviolent (burglary, larceny-theft and motor vehicle theft).

LEMAS data 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 either a program or a policy.

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

awebster@cs.ubc.ca, mwelsman@cs.ubc.ca

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 100 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 included in the LEMAS data set.

Our model of the data in LERA is a tabular database with one entry or tuple per law enforcement agency. The schema of the database is the union of the LEMAS schema and the crime report data schema. The data in the database was cleaned by removing agencies that did not have data values for one or more of the values in the combined schema. Handling of incomplete data is something that we could add in a future version of LERA.

2 Related Work

In this section, we present the solutions we explored for the purpose of supporting a crime analyst in the types of tasks described in Section 1.2. We also introduce our chosen information representation solution, the scatterplot, as well as one of the most common techniques for simultaneously displaying multiple scatterplots in a single display, the small multiple. We finish with a description of scagnostics, a measure we use in LERA for ordering the scatterplots in a small multiple to facilitate pattern exploration.

2.1 Solutions Considered

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

2.1.1 Parallel Coordinates

Parallel Coordinates were proposed in [18] 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.

2.1.2 Table Lens

Table Lens was the solution put forth in [11] 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 a scatterplot would be more beneficial to the user than the comparison of distributions in the Table Lens solution.

2.1.3 Graph Drawing Techniques

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.

2.1.4 Map-based Representation

Finally, we also considered using a map to demonstrate the effect of an independent variable on a dependent one. However, a map is more suited to showing how a variable of interest varies spatially, not to show the relationship between two non-spatial variables. Moreover, we wanted to show a kind of abstraction of the information that would not be bogged down by geography. And, finally mapping the crime rate and LEMAS data was simply not possible as the data was not spatially coded (i.e., latitude and longitude was not available).

2.2 Scatterplot

The scatterplot is the tool most commonly used by crime analysts to conduct regression analysis to determine the correlation between two variables [4]. However, the tool that is currently available and most commonly used by crime analysts, namely data analysis programs such as Microsoft Excel [4], are neither interactive nor flexible. Obviously, an interactive information visualization system which provides the same functionality and more is needed. In LERA, we attempt to provide such a system.

2.3 Small multiples

The small multiple is a technique introduced by Bertin in 1967 and later popularized by Tufte in 1983 [10]. It is defined by MacEachren et. al. as a "set of juxtaposed data representations that together support understanding of multivariate information" [10]. The contribution of the small multiple is its support for answering the single question that is "at the heart of quantitative reasoning," [15] namely "compared to what?" [15]. A small multiple of scatterplots is useful in the problem domain presented herein as it facilitates the comparison of the effect of a single independent variable on several dependent variables and/or the effect of several independent variables on a single dependent variable. As well, allowing the user to reorder the scatterplots in a small multiple can facilitate the exploration of patterns in the relationships between different variables [10].

2.4 Scagnostics

A scagnostic is essentially a cognostic (or computer guiding diagnostic) [13] for ordering scatterplots in a scatterplot matrix, "a (usually) symmetric matrix of pairwise scatterplots" [18]. It was introduced by Tukey in the 1980s [13]. Scagnostics can also be applied to order the scatterplots in a small multiple [18]. Using scagnostics to order scatterplots helps alleviate the problem of easily being able to explore a large number of scatterplots when the number of scatterplots grows cumbersome [19].

3 Description of Solution

In LERA we provide an interactive scatterplot information visualization system that allows a user to quickly and effectively explore a large data set to discover possible correlations between variables. In the sections that follow, we describe the general design choices we made in developing LERA as well as the specific features that were implemented to support data exploration in the domain of crime analysis.

3.1 General design

The choices made in visually representing data can and frequently do have a major effect on how that information is perceived and interpreted by the viewer. It is important to carefully consider how information should best be encoded as poorly chosen encodings can be both confusing and misleading [12].

3.1.1 Visual encodings

In LERA, we encode four different variables in a single data point: X- and Y-values which together are represented as a point's position in a scatterplot, colour and size. X- and Y-values encode any of the numeric variables in LERA's data set and represent the most important characteristics being investigated in a scatterplot. In particular, when determining the relationship between two variables, the effect of the independent variable (X) on the dependent variable (Y) is the point of interest, and the primary task being supported by LERA.

Colour was chosen to encode a third categorical (in most cases a Boolean) variable for each data point. The use of colour to encode a categorical feature of the data can help the user discover if differences exist in the relationship between X and Y for each of the different categories encoded by colour.

Size encodes the fourth and final variable. Like X and Y, size is used to represent the value of a numeric variable. However, unlike X and Y, the size of a data point will not directly correspond to the relative value of the variable it encodes for a particular law enforcement agency. Instead, the size of a point will be one of 5 possible sizes. The five possible size groups or bins are determined by taking the difference between the smallest and largest values of a numeric variable and dividing up this difference into five equally-sized bins, each of which is appointed a particular size. This choice was made because the human eye cannot easily detect differences in size particularly when the sizes of two objects being compared are small and it also underestimates the size of larger symbols [14].

Moreover, it is difficult for a user to distinguish differences in the size of data points when they are not within close proximity of each other on the scatterplot. Another problem that could surface if the exact value of a data point was encoded by size is occlusion.
This particular characteristic of the size encoding makes it difficult for the user to make precise and meaningful judgments about the differences between two points based solely on their size. However, size encoding will and can be useful for making judgments about the relative differences in the data points. It is for this reason that the size is not provided in the mouseover textbox for a point: size is not on par with X, Y and colour encodings. It provides some visual cues, but these are not as important as those provided by X, Y and colour.

In summary, for each law enforcement agency, a total of four features or variables can be represented in LERA.

3.1.2 Visual salience

Visual salience is the perceived importance of an object by the visual system. It is important to take this into consideration when designing an interface. According to Tufte [15], background elements such as the axes, the horizontal bars and the background colour of a scatterplot should have lower salience. And, foreground elements such as data points, regression lines and the font on mouseover textboxes and menubars should have high salience. As such, we chose to encode the background colour of the scatterplot display with a light gray colour, and the horizontal lines and axes in white to ensure that they didn't distract the user's focus from the more important features in the scatterplot display. Data points, regression lines and menubar fonts were encoded with colour to ensure that they would be easily distinguishable from the background elements and capture the user's attention.

3.1.3 Plotting symbols

Two different plotting symbols were chosen for representing points in LERA, a cross to represent individual law enforcement agencies, and a filled circle to represent state aggregates (i.e., the average data point across all law enforcement agencies in a state). The two default symbols were chosen from the set of symbols recommended by Cleveland in [3] for data sets where there may be overlap. Crosses minimize occlusion and circles are easy to tell apart from crosses; since there are fewer aggregates than agencies (i.e., 770 agencies and 52 states in the US), we chose to use circles for aggregates to minimize occlusion problems as much as possible.

3.1.4 Colour

As indicated in the section on Visual encodings, colour is used to encode categorical variables. The accepted practice for distinguishing the different values of a categorical variable with colour is to use hue, "the perceptual dimension of color that we associate with color names, like red, green, blue, and yellow" [1]. In choosing the two different colours schemes to be used in LERA, one colour scheme for Boolean variables and one for categorical variables, this principle was heeded.

For Boolean variables two unique hues, red and blue were chosen. Since each of the two possible values of a Boolean variable is considered to be equally important, the saturation level and luminance of the two hues were kept the same: both colours have low saturation and have low luminance [1]. The choice of red and blue was made so as to be distinguishable by color-blind individuals. And, desaturated and lower luminance levels of red and blue were selected based on feedback obtained from participants in the usability test which indicated that the higher luminance, highly saturated colours against a dark background were hard on the eyes and would be difficult to look at for long periods of time. An example of the initial colours chosen for the first LERA prototype which has used in the usability test can be seen in Figure 1.

Figure 1. First LERA prototype that was used in usability testing. The colour scheme and display characteristics were changed in the subsequent version based on user feedback gathered during usability testing.

The categorical colour scheme which can be seen in Figure 3, consists of six unique hues. Only six different hues were chosen as the human eye can only easily distinguish a small number of different colours [17]. The source of the six hues was [2] and not Ware's maximally discriminable set of colours as this was the source of the original colour schemes used in the first prototype of

LERA that were found to be difficult to look at by the participants. As two of the hues are red and green, saturation and luminance was used to help distinguish between them: the green hue has a higher saturation and luminance than the red hue. Otherwise, the other hues are low saturation and low luminance for the same reasons presented in the Boolean colour scheme.

It is important to point out that the two colour schemes used in LERA were tested for the different types of color-blindness: deuteranope (a form of red/green color deficit), protanope (another form of red/green color deficit) and tritanope (a rare form blue/yellow color blindness) [16].

3.2 Statistical analysis

Regression analysis is statistical method that is used to quantify the type and strength of a relationship between a dependent variable and one or more independent variables. In essence, it provides a measure of the degree of the effect an independent variable has on a dependent variable. In LERA the most basic form of regression, namely linear regression, is provided to the user (i.e., the user can add a linear regression line to any scatterplot).

3.2.1 Regression lines

A regression line represents the relationship between a dependent variable and an independent variable. The slope and y-intercept of a regression line are calculated using the data points in a scatterplot. The formulae used to determine slope and y-intercept are shown in Equation (1).

$$m = \frac{n\sum(xy) - \sum x \sum y}{n\sum(x^2) - (\sum x)^2} \qquad b = \frac{n\sum y - m\sum x}{n}$$

Equation 1. Formulae for the slope (m) and y-intercept (b) of a linear regression line. x and y are the x- and y-values of the points in a scatterplot.

Regression lines are typically shown with their regression equations and R-squared value, a characteristic that is respected in LERA. The R-squared value of a regression line is a measure of how well the regression line fits the data. As such, R-squared is necessary for quantitatively assessing the effect of an independent variable on a dependent one. An R-squared value ranges between 0 and 1.0, where 0 indicates that there is no relationship at all and 1.0 indicates a perfect fit. The formula for R (i.e., the root of R-squared) is presented in Equation (2).

$$r = \frac{n\sum(xy) - \sum x \sum y}{\sqrt{\left[n\sum(x^2) - (\sum x)^2\right]\left[n\sum(y^2) - (\sum y)^2\right]}}$$

Equation 2. Formula for R-squared. x and y are the x- and y-values of the points in a scatterplot.

3.2.2 Outlier detection and removal

In LERA, we support both the manual removal and automatic detection and mass removal of outliers. Manual outlier removal can be done by the user via a quasimode of Shift + click. Automatic detection of outliers employs a statistical method using point residuals. A residual is the vertical distance of a point from the regression line or in other terms the actual y-value minus the observed y-value of a point. The user can choose to have outliers automatically identified using a menu item on the menubar. This particular function requires the user to specify the percentage of outliers they want to select. So, if the user specifies one percent, the one percent of data points with the largest residuals (i.e., they are the one percent of points the farthest away from the regression line) would be highlighted by a circular transparent background blur patch which appears behind each point. This background blur patch is initially exaggerated in size to draw the user's attention and then through the use of animation it shrinks quickly to a smaller, more appropriate size (see Section 3.31 for a more comprehensive explanation of animation). Once outliers have been highlighted, the user can choose to remove them, in which case a simple button click will suffice. Figure 2 shows the LERA interface after the user has chosen to select two percent of the outliers.

As the user may later to decide that s/he wants to work with the entire dataset after removing outliers (as suggested in our usability testing), we have provided the user with the ability to restore all points in the original dataset (i.e., the user simply needs to click a button on the interface).

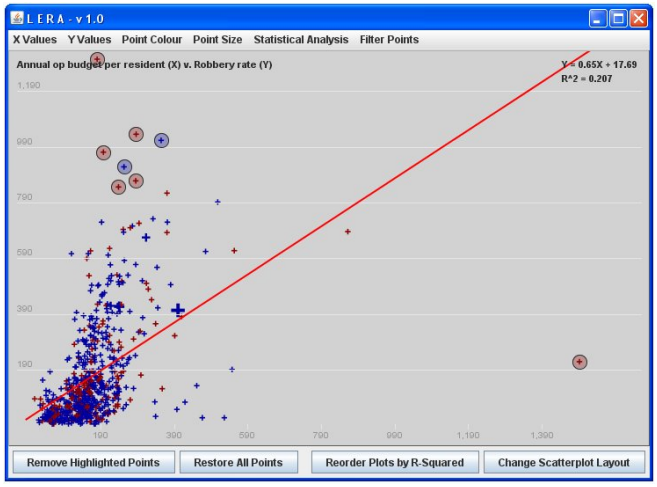


Figure 2. Automatic outlier detection (user specified 1% of outliers, the 1% with the largest residuals, to be removed). The detected outliers are indicated by a circular transparent background blur patch which appears behind each outlier.

3.3 Small multiples

LERA allows the user to view multiple scatterplots at the same time in a single screen through the use of small multiples. This feature facilitates the comparison of bivariate scatterplots and supports the user in trend and pattern detection. In particular, it allows the user to quickly and easily compare the effect of a single law enforcement administration program or policy on several different dependent variables in a single view to allow the user to determine how the program or policy impacts a variety of different crimes. Conversely, it also lets the user compare the effect of several different law enforcement administration programs or policies on a single type of crime to determine which of these programs/policies can better reduce the incidence of that type of crime.

Data exploration is further supported through the use of linked mouseover and highlighting and, through the reordering of scatterplots using scagnostics.

3.3.1 Linked mouseover and highlighting

We provide in LERA two different types of linking between the scatterplots in a small multiple. The first is linking mouseover information in an active scatterplot to all other scatterplots in the small multiple. This feature allows the user to see how a single law enforcement agency fits into the distribution of data points in each scatterplot in the small multiple. As illustrated in Figure 3, linked

mouseover also provides the user with a wealth of information for the highlighted law enforcement agency as each scatterplot can encode three variables, the value of the dependent variable encoded by the Y-value of the point, the value of the independent variable encoded by the X-value of the point and the value for a categorical or Boolean variable encoded by colour. All three of these pieces of information are displayed in the mouseover text box as well as the name of the law enforcement agency the point represents.

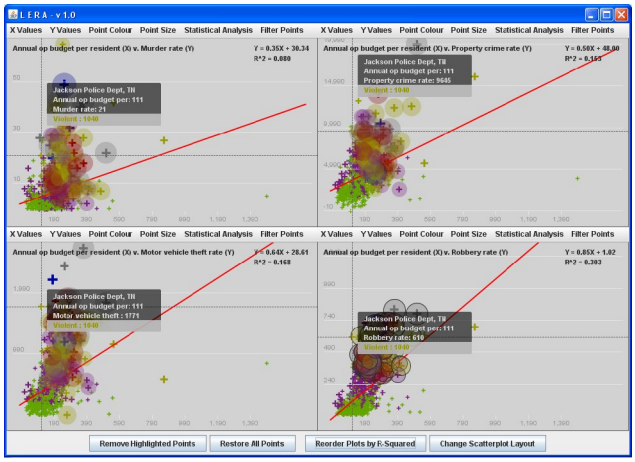


Figure 3. Linked mouseover in a small multiple of scatterplots. Linking is achieved through the use of crosshairs and mouseover textboxes containing the values of variables encoded in the display.

Linked highlighting takes linking of information in the multiple scatterplots one step further by giving the user the ability to select a larger number (or cluster) of data points, each representing a single law enforcement agency, through the use of a lasso. The lasso is encoded in a quasimode, and is activated by clicking and dragging the mouse around the desired set of points. Upon selection of a set of points using the lasso, animation is used as a visual cue to facilitate the user's perception of the change in the display triggered by the act of lassoing. In particular, behind each of the points selected by the lasso a circular transparent background blur patch appears which initially is exaggerated in size to draw the user's attention and then through the use of animation shrinks to smaller, more appropriate size. While it is impossible to provide an example of the animation in this paper, the result is shown in Figure 4. Animation was chosen to facilitate the user's perception of change in the display, and prevent change blindness, because it has been shown to be well suited to peripheral vision [8]. In particular, when attention is focused on a selection in the active scatterplot, animated change in one's peripheral vision provides a very brief high contrast visual cue. Any other cue cannot be as effective since the change in display occurs in the periphery of one's visual range. If we simply highlighted the corresponding lassoed points in the other scatterplots, the user would be much more likely to miss the change.

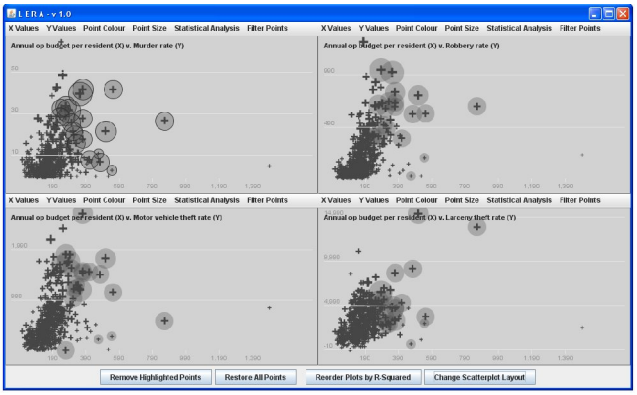


Figure 4. Linked highlighting in a small multiple of scatterplots. Linking is achieved through the use of animated circular transparent background blur patches.

3.3.2 Reordering by scagnostics

A final feature that we provide in the small multiple of scatterplots display is the ability to reorder scatterplots using scagnostics. The effectiveness of small multiples decreases as the number of scatterplots (i.e., the number of variables that are represented) is increased [19]. To help the user quickly and easily identify the scatterplots with the most significant relationships, we allow the user to reorder the scatterplots in a small multiple according to the R-squared value [13] of the scatterplot. As we felt that it would be disorienting to users if LERA dynamically reordered scatterplots when the R-squared value of one of the scatterplots in the small multiple changed (due to the removal of points), we chose to map this functionality to a button. The resulting order of scatterplots will be in decreasing order of R-squared values from left to right, top to bottom starting in the upper left-hand corner of the small multiple grid. In this configuration, the user will be able to focus their attention on the scatterplots where the effect of the independent variable on the dependent variable is strongest.

3.4 Focus and context

Focus and context capability is provided in LERA. Context is provided for one or more states by representing each state by its mean data point or "aggregate" (i.e., the average data point across all agencies in a state). Focus is provided by plotting individual law enforcement agencies for a state(s) of interest. The shape of a symbol used to represent a data point is used to encode a state aggregate. In particular, as pointed out in Section 3.1.3, a filled circle is used to encode an aggregate point. We chose this encoding because we could not use the other channels typically used for encoding information such as colour and size which had already been used to encode other information. Also, if we had chosen to use a larger symbol size instead of a different shape, we would have increased the chances of occlusion. An example of an aggregate point is provided in Figure 5 in which context is provided for the state of California and focus is provided for all other states.

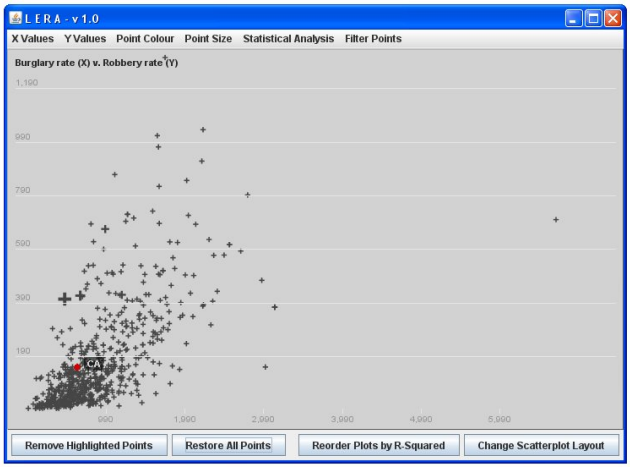


Figure 5. An example of focus and context in LERA. An aggregate point provides context for the state of California (a red circular data point with a textbox indicating the abbreviated state name, CA). All other states are represented in focus by all of the law enforcement agencies contained within their boundaries (the grey crosses).

3.5 Filtering

A scatterplot of the full data set displays hundreds of points at once. This can cause overwhelming visual clutter and in some cases can cause points to be occluded. Furthermore, there are problems with colour encoding in cases where the number of categories to be shown is too large. This is particularly true in the case where colour is mapped to the agency state field since there are over fifty states and districts in the United States.

To deal with these problems, LERA can filter points and point aggregates by state. Through a standard menu interface the user can choose which US states to display. This allows for the construction of simple plots with a manageable number of points that describe relationships within a well-defined geographic area. When used in conjunction with small multiple scatterplot views, filtering facilitates comparison between geographic areas. Plots with a limited number of states also allow for highly distinguishable colour encoding by state.

4 Medium-Level Implementation

LERA is a Java application deployed as an executable JAR file. The LERA Java source consists of approximately 35 classes and 3,000 lines of code built on top of the Prefuse visualization toolkit [7]. LERA can be broken down into three parts: data and configuration, modelling, and display.

LERA data was manually compiled into one large CSV file from crime report and LEMAS data available online. Prefuse provides functionality for reading CSV files and for compiling them into tables which can then be manipulated using a basic database-like framework. LERA also includes configuration information used to determine the type of each data field (e.g. categorical or numerical) as well as additional details such as data field categories and US state names and abbreviations. This information is used to build menus for X, Y, colour, and size mappings.

At the core of LERA is a subclass of the Prefuse Display class that defines a scatterplot based on the data table described earlier. This scatterplot display uses standard Prefuse actions to handle basic point characteristics (X location, Y location, colour, size), label drawing, and highlight or selection animations. Filtering was accomplished by creating custom predicates to indicate whether or not a scatterplot point should be displayed given the current state. More complicated functionality such as custom labels and crosshair display, lasso drawing and selection, highlighting, aggregate points, and linear regression plots were implemented by drawing overlays on top of the scatterplot. These functions were either not implemented in the desired way in the Prefuse toolkit, did not exist at all, or were not well-documented.

Coordination of multiple scatterplot views is handled by the MultiPlot class, which automatically generates layouts for an arbitrary number of scatterplots and communicates events between scatterplots and overlays. When the user selects a group of points in a multiple scatterplot view in LERA, this generates a call to the MultiPlot class which then propagates highlight calls to each scatterplot, causing the corresponding points to be highlighted in all other plots.

Finally, a number of methods were implemented in order to support statistical analysis. We discovered several statistical packages but these were both overly complicated and lacked the full set of specific features set we needed. As a result, we implemented standard methods for finding a line of best fit (linear regression line), R-squared values, and point residual values.

5 Results

5.1 Screenshots

See Appendix A.1. Please note that some of the screenshots shown in the body of this paper are repeated in better quality representations in the Appendix.

5.2 Scenario of Use

The crime analyst for the state of California has been asked by his supervisor to determine if and how the annual operating budget impacts burglary for the state of California. To answer this question, the analyst inputs “annual operating budget” as the independent variable and “burglary crime rate” 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 and then to have the two percent of data points with the largest residuals identified as outliers, which he subsequently decides to remove automatically using the system. He interprets the resulting scatterplot, shown in Figure 6, which includes an R-squared value and regression line equation. He reports the information to his supervisor.

6 Usability Testing

We decided to conduct a comparative evaluation of a prototype of LERA against Microsoft Excel to determine LERA’s relative ease of use, if applicable. Excel was chosen as the application against which to compare LERA as it is the standard tool used by crime analysts [4].

6.1.1 Goals, Approach and Rationale

The goal of our user study was to determine if LERA is easier to use than Excel for performing the tasks typically required of a crime analyst. Our proposed approach was to evaluate the two applications (LERA and Excel) independently using think-aloud observation. This approach was chosen as it is known to provide rich, qualitative information about the users' experiences in using an application. We also made use of questionnaires to help us quantify which of the two applications was easier to use both on a task-by-task basis and overall. And, we timed the users for each task to

provide some additional quantitative information to facilitate our comparison of the relative usability of the two applications. It is also important to note that we changed the order in which the two applications were tested by the participants (i.e., Excel then LERA versus LERA then Excel). Of the two experts, one began with LERA and the other with Excel. Similarly, the two novices differed in the application they used first for testing.

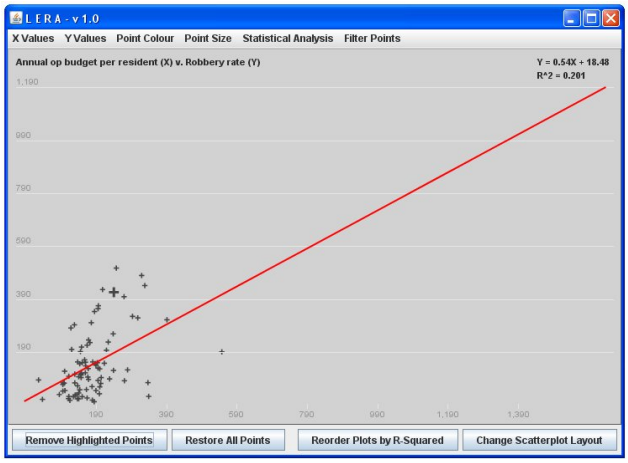


Figure 6. Scatterplot generated by a crime analyst to investigate the effect of annual operating budget per resident on robbery rate in the state of California.

6.1.2 Participant Pool

Our participant pool consisted of one pilot and four participants. Two of the participants are experts in Excel and regression analysis and two are novices. One of the two experts is an expert statistician working in industry.

6.1.3 Results

During the evaluation which was done on an earlier prototype of LERA (than is presented in this paper), each participant (or user) was given a set of six tasks (provided in Appendix A.2) to complete using each application. A screenshot of the LERA prototype that was used for usability testing can be seen in Figure 1. After each task, the participant was asked to rate how easy it was for them to complete the task using the particular application.

Table 1 shows the ease of use ratings, a value from one to ten where a rating of one indicates that the task was very difficult to complete using the given application and a rating of 10 indicates that it was very easy. As can easily be seen, for each of the four participants, it was as easy or easier to complete each of the six tasks using LERA. However, the results in Task #6 may be somewhat misleading and require further explanation as it was impossible to complete this particular task – that of comparing a cluster of points (i.e., law enforcement agencies) in one scatterplot to its distribution to three other scatterplots in a small multiple – using Excel. This particular task was included to obtain user feedback about the linked highlighting feature in LERA. And, it is also important to point out that Task #3 (i.e., creating an aggregate point for California) was found to be quite difficult by one of the participants (as well as the pilot) who did not complete the task.

	Task #1		Task #2		Task #3	
User	**Excel**	**LERA**	**Excel**	**LERA**	**Excel**	**LERA**
1	8	**10**	7	**10**	8	**9**
2	6	**10**	7	**10**	3	**10**
3	2	**10**	9	**10**	1	**10**
4	8	**10**	10	**10**	10	**10**

	Task #4		Task #5		Task #6	
User	**Excel**	**LERA**	**Excel**	**LERA**	**Excel**	**LERA**
1	5	**10**	6	**9**	1	**8**
2	2	**9**	3	**7**	1	**10**
3	2	**3**	4	**10**	1	**10**
4	7	**10**	1	**10**	1	**10**

Figure 7. Table 1: Ease of use ratings for each of the six tasks performed by the four participants of the user study. A rating of 1 indicates that the task was very difficult and a rating of 10 indicates that it was very easy.

Similarly, Figure 7 demonstrates the improvement in performance time (i.e., relative percent decrease in completion time) experienced by each participant by using LERA to complete the given tasks as opposed to Excel. As can be seen in Figure 5, only four of the six tasks are shown. This is due to the fact that one of the novice users (User #3 in particular) could not complete Task #3 (i.e., creating an aggregate for the state of California) using Excel and that Task #6 was impossible to complete using Excel.

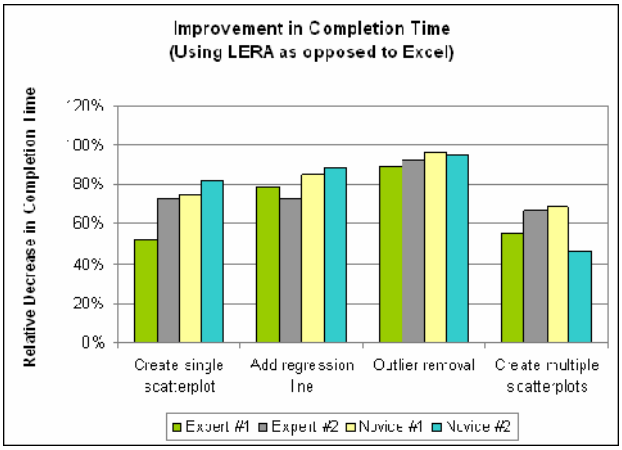


Figure 8. Improvement in task completion time by using LERA as opposed to Excel for four different tasks.

Users were also asked to indicate in a post-questionnaire which of LERA's features they liked. In particular, the following set of features was found to be beneficial by the users in the study:

- Volume of information provided in a mouseover textbox for a particular point is a vast improvement over simply provided the x- and y-coordinates of a data point.
- The ability to specify colour for a variable was well received, particularly by the expert participants who had found this feature lacking in other applications they had used for conducting regression analysis.
- Linked mouseover and highlighting was found to be very useful, particularly by the expert users who indicated that this feature would greatly aid the comparison of the relationships between different variables.

However, there were also some features that the users either did not like, or found confusing:

- It was too easy to remove outlier points (i.e., through a simple mouse click).
- The inability to undo the removal of outlier points was found to be disconcerting to users who asked how they could add the removed points back to the scatterplot.
- The colour scheme used for the scatterplot display were found to be overwhelming and one of the expert users suggested that using LERA for a long period of time would be straining on one's eyes.

Finally, the users were asked to suggest what additions to LERA would be useful:

- Zooming and panning capability to be used in conjunction with linked highlighting to help the user investigate clusters at a finer level of detail.
- A way to undo the removal of points (either outliers or lassoed points).
- Ability to control colour coding of a data series (i.e., being able to represent different regions with different colours and/or allow the user to determine groupings)
- Application of statistical methods to automatically identify outliers for the user.

7 Discussion

7.1 Strengths and Weaknesses

Evaluation of the LERA system revealed both strengths and weaknesses. The most significant strength of the system is that it allows the user to quickly construct a display that conveys both large-scale relationships and specific pieces of information from a data set that is too large to permit unaided searching. By selecting X and Y mappings, users were quickly able to isolate one relationship in the data. By hovering the mouse over a point, users obtained specific information about a data point. Multiple scatterplot views and linked highlighting and labels add to this strength. Small multiples permit comparison between different fields of many different agencies. Linked labeling allows the user to quickly obtain a set of relevant field values associated with a single agency of interest.

The most important weaknesses of the system are a result of limitations in the current implementation. The most significant problems in the current version of LERA have to do with the fact that mappings are generated before filtering takes place. As a result, a scatterplot can contain mostly empty space and colour encodings can be sub-optimal, based on the total number of categories rather than the number of filtered categories. A more complete system would make an effort to customize colour sets and axis ranges based on filtered data, while trying to minimize changes as a result of user input.

Filtering in general could also be improved to permit many different queries on the data set beyond just states, although state filtering is arguably the most useful for our data set. The interface for automatic outlier selection is also somewhat lacking. Full details of potential improvements are given in Section 7.

8 Future Work

Many features that would be useful to include in the LERA system remain unimplemented due to time constraints. These features include:

- More robust filtering and more complete aggregation. It is possible to support range queries on any data field. Similarly, aggregation can be done over different fields using different statistical measures.
- Zooming and panning as suggested in the user study and which can also help alleviate the problem of visual resolution in display when there are a large number of scatterplots displayed [19].
- Additional scagnostics such as number of detected outliers or least-squares error [13].
- Cues to help the user keep track of scatterplot reordering. Currently, the reordering function redraws all scatterplots in a new position once the new order has been calculated. Scatterplots often look similar and it can be difficult to determine which scatterplot moved where.
- Manual "drag and drop" style reordering of scatterplots and possibly multiple scatterplot scales. For example, it would be possible to enable the user to construct a display with one large scatterplot and four small scatterplots.
- Multi-scale banking [6]. Axis scaling can be changed to improve judgments of trends shown in the scatterplots.
- Additional methods of statistical analysis such as quadratic or exponential regression curves.
- Data cleaning functions and better handling of invalid entries.
- More user control over colour encoding. Colour palettes could be configurable. Users evaluating LERA also thought it would be useful to support a wider range of colour encodings. For example, they wanted to be able to choose one colour for one state and a second colour for all others.
- LERA could also be generalized to handle other tabular data sets. In order to do this it would have to support configuration files that define the types of data fields and the structure of menus presented.

9 Conclusion

LERA is an interactive information visualization system designed to support the exploration of the effects of various programs and policies on crime rates by crime analysts. It has a number of features designed to facilitate this task: regression lines, manual outlier removal and automatic detection and removal of outliers, small multiples, focus and context and, filtering. Through user testing, we discovered that LERA can be useful for supporting the tasks faced by a crime analyst in determining which particular programs a law enforcement agency's limited funds should be applied to. And, that LERA was easier to use than Excel, the current tool used by most crime analysts for regression analysis. Finally, at the top of our list for future enhancements to LERA is to adapt our system to be more flexible in the choice of colour sets and the range of axes scales.

References

[1] Brewer, C. Color Use Guidelines for Data Representation. Proceedings of the Section on Statistical Graphics, American Statistical Association, 1999, pp. 55-60.

[2] Choosing Colors for Data Visualization. Maureen Stone. Retrieved December 10, 2007 from www.perceptualedge.com/articles/b-eye/choosing_colors.pdf.

[3] Cleveland, William S. The Elements of Graphing Data. New Jersey: Hobart Press, 1994.

[4] Crime Analysis in America. Prepared for US Department of Justice Office of Community Oriented Policing Services. Center for Public Policy, University of South Alabama (March 2002).

[5] Crime Trends. US Federal Bureau of Investigation (FBI) Uniform Crime Reports. Retrieved October 20, 2007 from http://bjsdata.ojp.usdoj.gov/dataonline/Search/Crime/Crime.cfm.

[6] Heer, J. and Agrawala, M. Multi-Scale Banking to 45°. IEEE Transactions on Visualization and Computer Graphics 2006: 701-708.
[7] Heer, J., Card, S.K., and Landa, J.A. prefuse: a toolkit for interactive information visualization. In CHI Human Factors in Computing Systems, 2005.
[8] Heer, J. and Robertson, G. Animated Transitions in Statistical Data Graphics. Proc. IEEE Information Visualization, 2007.
[9] Law Enforcement Management and Administration Statistics (LEMAS). US Department of Justice, Bureau of Justice Statistics. Retrieved October 20, 2007 from http://bjsdata.ojp.usdoj.gov/dataonline/Search/Law/Law.cfm.
[10] MacEachren, A., Dai, X., Hardisty, F., Guo, D. and Lengerich, G. Exploring High-D Spaces with Multiform Matrices and Small Multiples. Proc Info Vis 2003: 31-38.
[11] Rao, R. and Card, S. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. ACM SIGCHI 1994: 318-322.
[12] Rogowitz, B. and Treinish, L. How Not to Lie with Visualization. Computers In Physics 1996: 268-273.
[13] Seo, J. and Shneiderman, B. A Rank-by-Feature Framework for Unsupervised Multidimensional Data Exploration Using Low Dimensional Projections. In Proceedings of the IEEE Symposium on information Visualization 2004.
[14] Slocum, T., McMaster, R., Kessler, F. and Howard, H. Thematic Cartography and Geographic Visualization. Upper Saddle River: Pearson Prentice Hall, 2005.
[15] Tufte, Edward R. Envisioning Information. Cheshire: Graphics Press, 1990.
[16] Vischeck. Retrieved December 11, 2007 from http://www.vischeck.com/vischeck/vischeckImage.php.
[17] Ware, Colin. Information Visualization Perception for Design. San Diego: Academic Press, 2000.
[18] Wegman, E.J. Hyperdimensional Data Analysis Using Parallel Coordinates. Journal of the American Statistical Association 1990 (85): 664-675.
[19] Wilkinson, L., Anand, A., and Grossman, R. Graph-Theoretic Scagnostics. Proc Info Vis, 2005.

APPENDIX A.1

Figure 9. Selection of Y-value (i.e., Robbery rate)

Figure 10. Mouseover information (and crosshair) provided by LERA

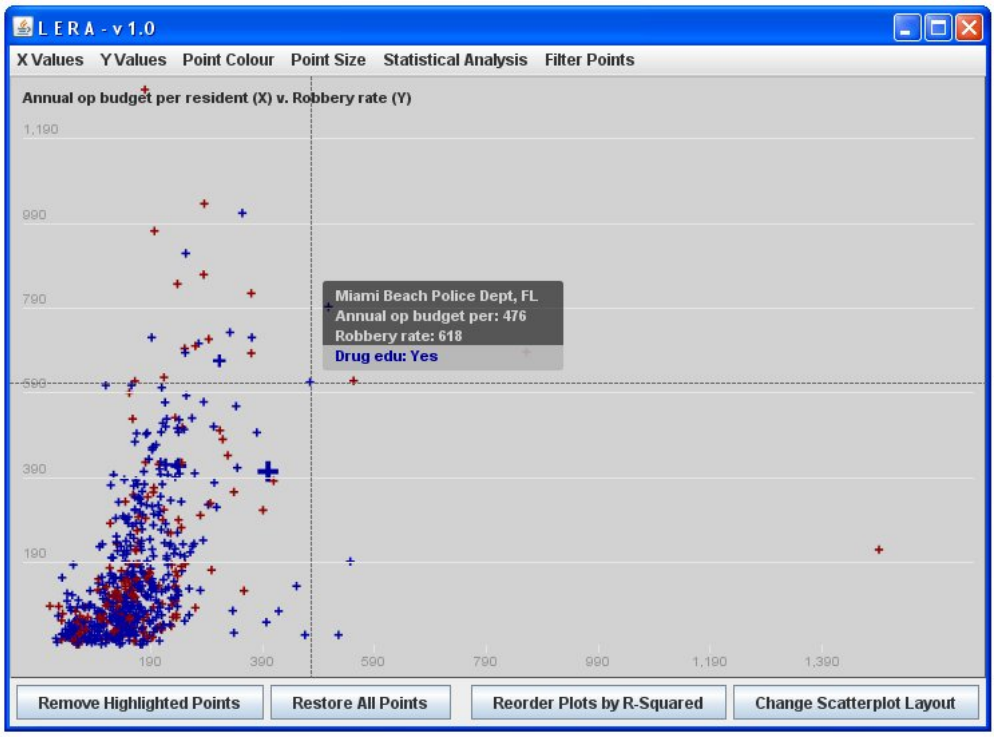


Figure 11. User specification of threshold for automatic outlier highlighting

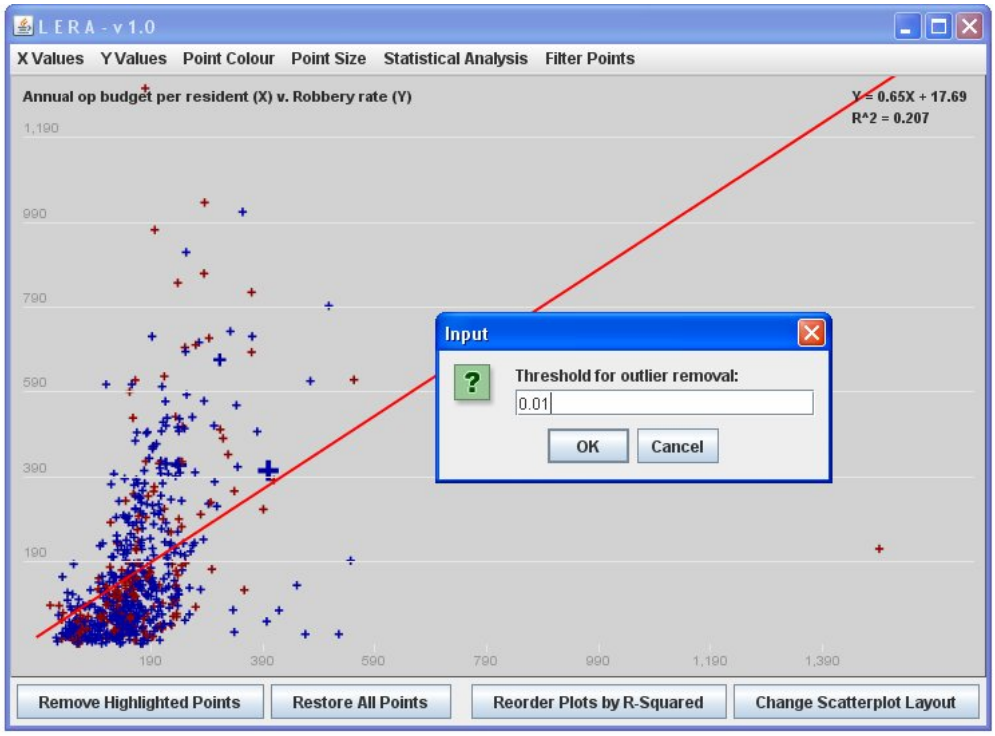


Figure 12. Single scatterplot with a regression line and largest 1% of outliers highlighted (i.e., 1% of data points farthest from the regression line)

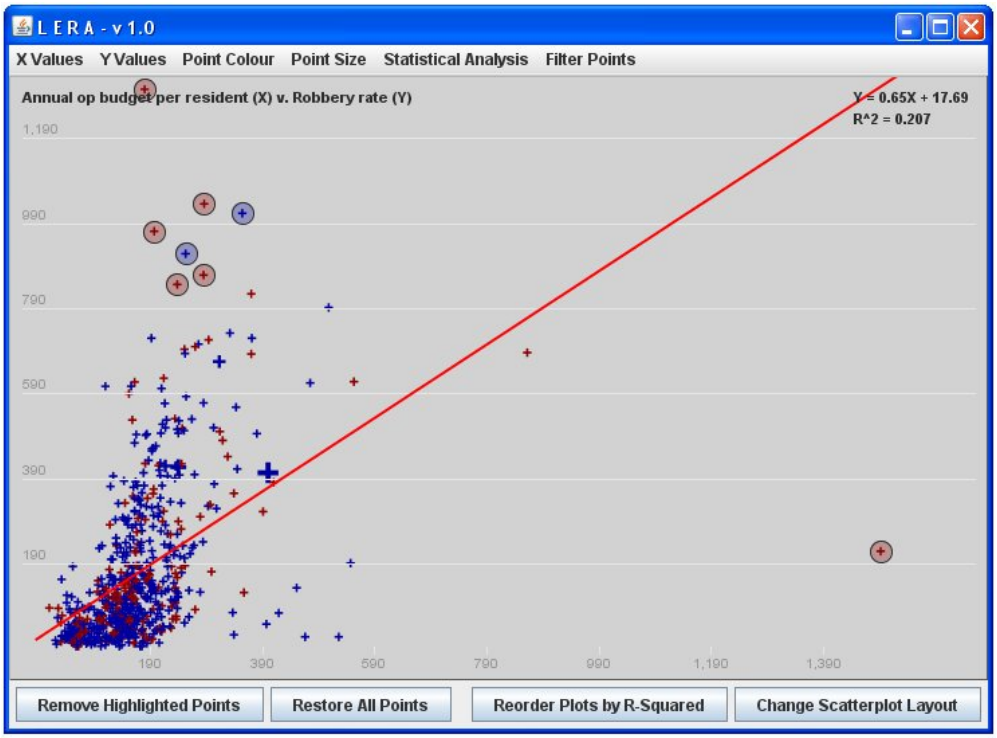


Figure 13. Context (aggregate) for California, focus for all other states

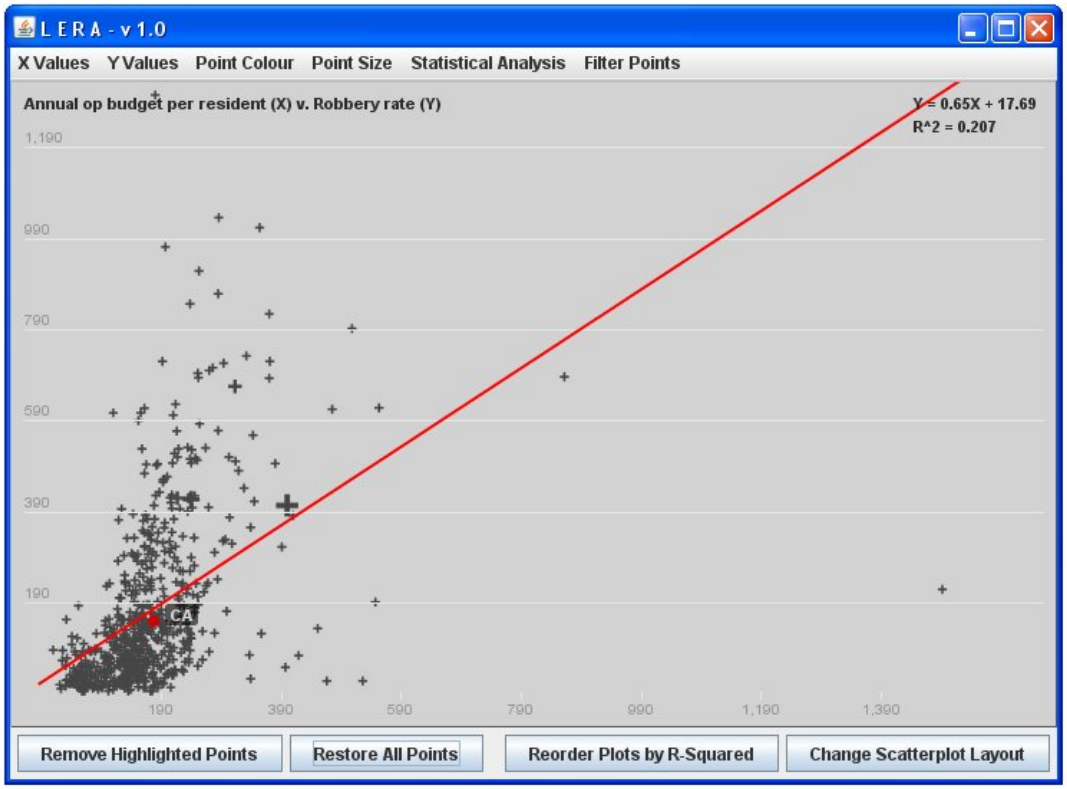


Figure 14. Window for specifying multiple scatterplots to be displayed in a small multiple grid.

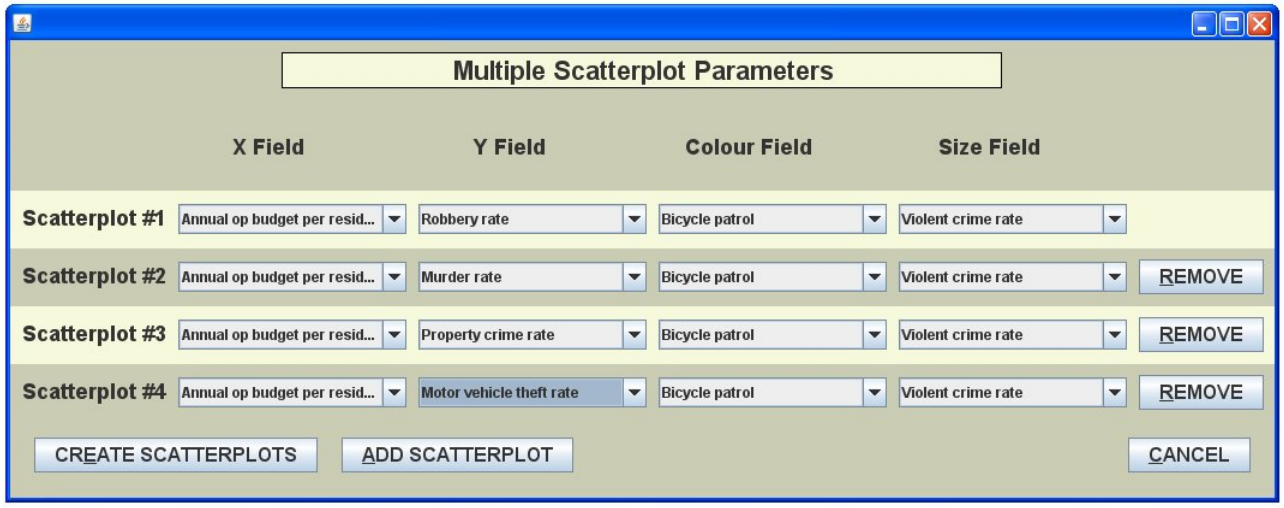


Figure 15. Example of the colour scheme for a categorical variable in a small multiple of four scatterplots. Note that each scatterplot in a small multiple has its own menu for visually encoding its variables (X Values, Y Values, etc) and for creating a regression line, detecting outliers, filtering points and creating an aggregate point.

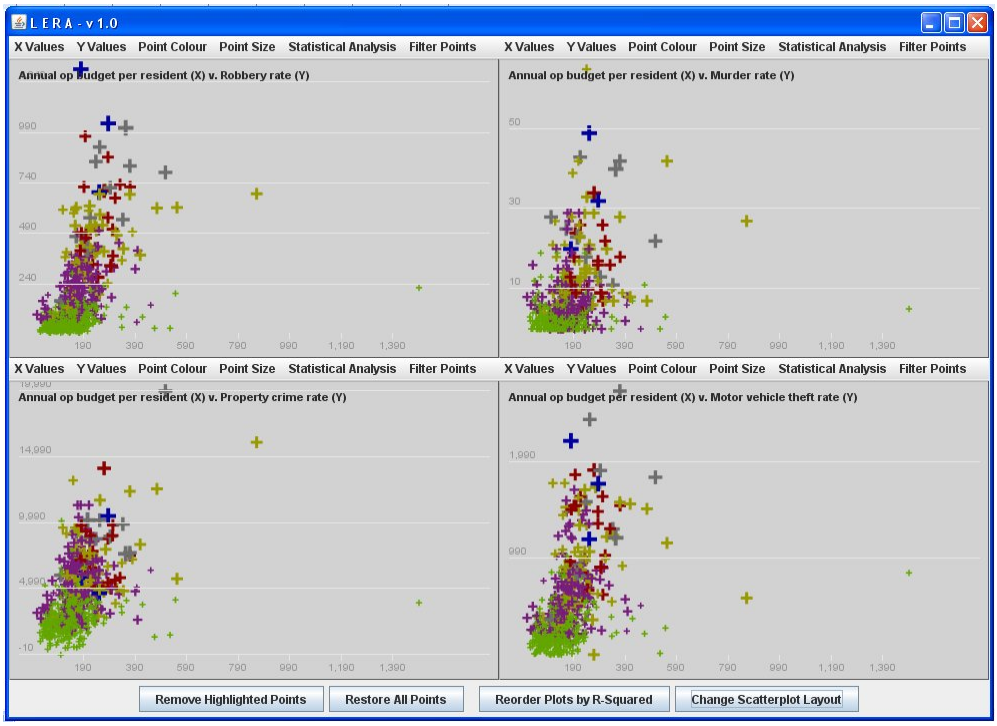


Figure 16. Outlier points selected (outlined circular blur patches) in upper left scatterplot. Linked highlighting indicates the position of these points in the other scatterplots.

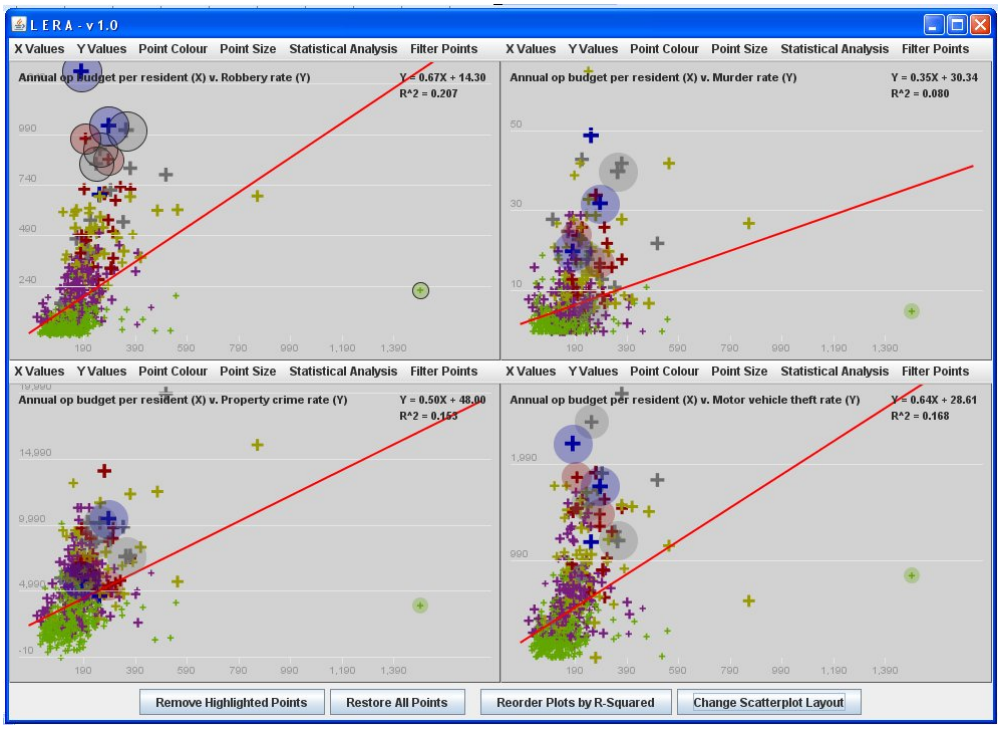


Figure 17. Example of lassoing shown in upper left scatterplot by the dotted black line.

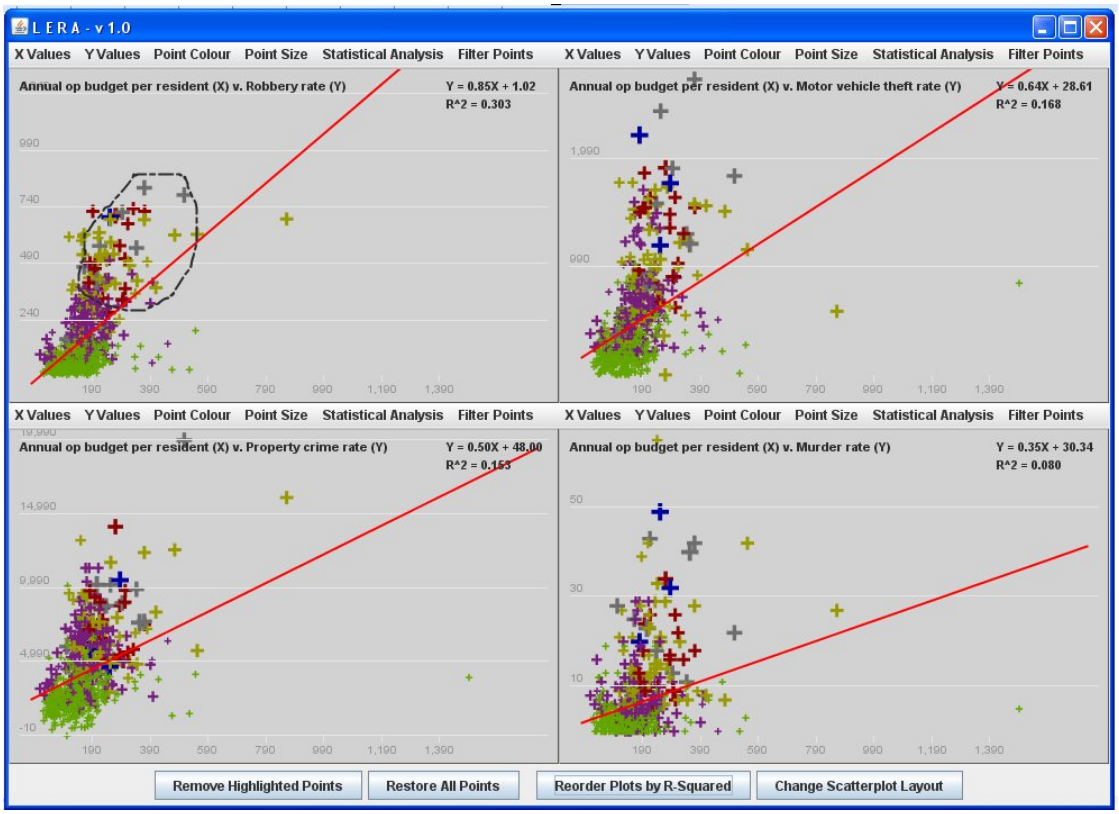


Figure 18. Screenshot after lassoing performed in Figure 16.

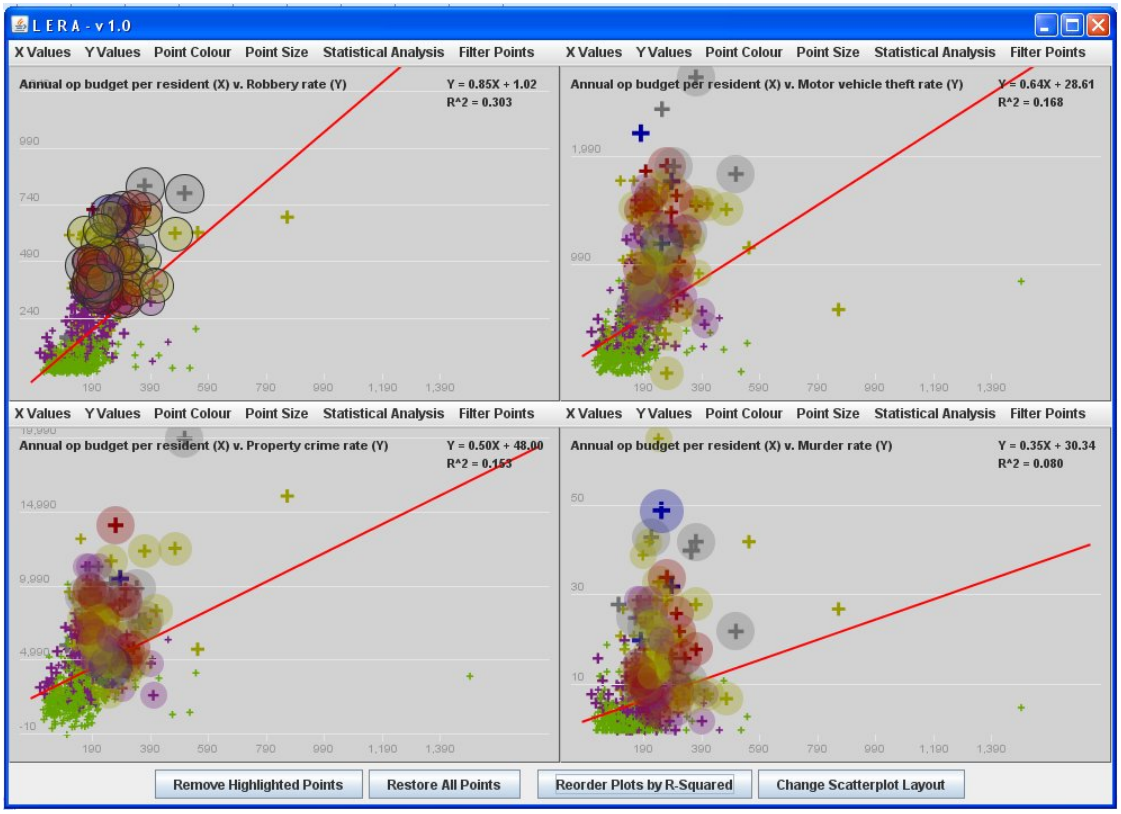
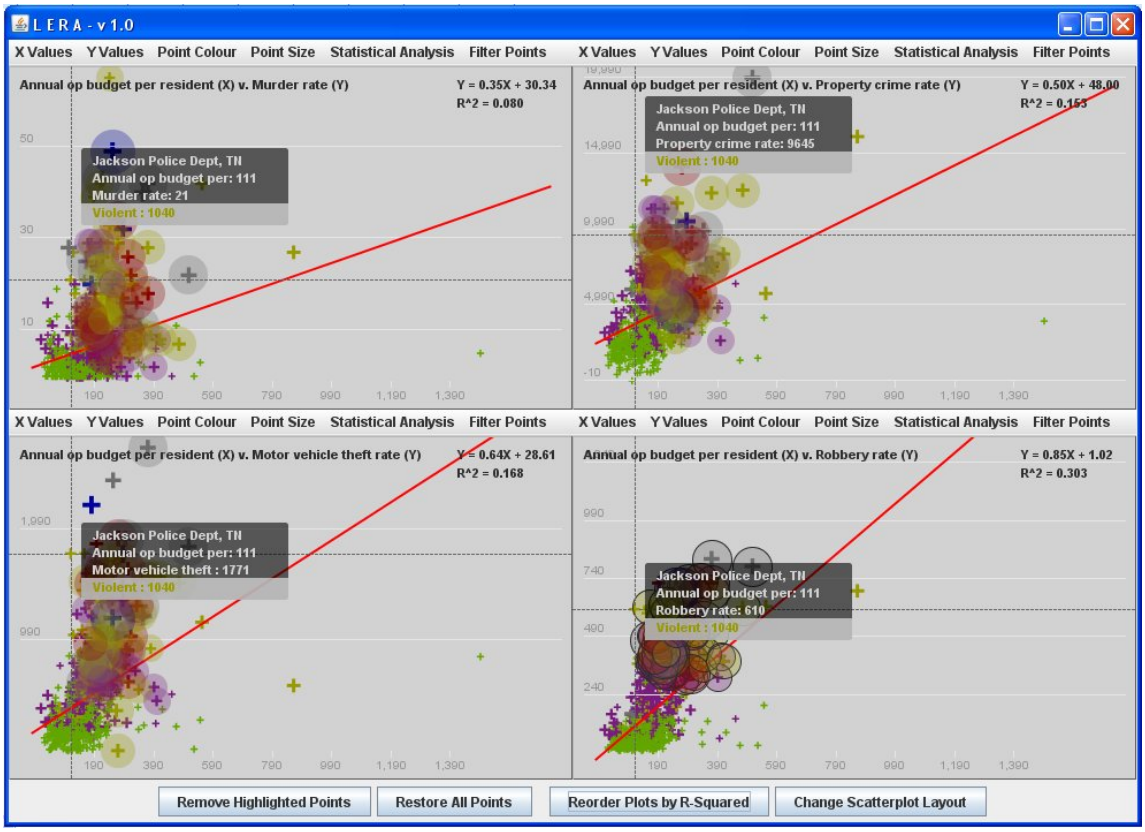


Figure 19. An example of linked mouseover. The point being hovered over in one scatterplot will be indicated in all scatterplots by a crosshair and by a textbox giving the values of the variables encoded by that point in each scatterplot.



APPENDIX A.2

USER STUDY TASKS

Task #1: Create a scatterplot with Annual operating budget per resident as the x-value (independent variable) and Robbery rate as the y-value (dependent variable). Describe the relationship between Annual operating budget per resident and Robbery rate (e.g., clustered in one area, spread out, etc).

Task #2: Add a regression line (also called a trendline) to the scatterplot you created in Task #1. Describe the slope of the trendline (e.g., increasing from bottom left to top right).

Task #3: There is one "outlier" highlighted (circled in red) in the provided scatterplot. Please remove it. Does the slope of the regression line change after removing the outlier?

Task #4: Create a scatterplot that plots an average for California (i.e., an average across its law enforcement agencies) and single points for all other states (i.e., one point per law enforcement agency as is the default for scatterplots). The average data point for California is represented by the average Annual operating budget per resident across all law enforcement agencies in that state for the x-value and the average Robbery rate across all law enforcement agencies in that state for the y-value. What are the x- and y-values for California's average?

Task #5: Create four scatterplots with Annual operating budget per resident as the x-value in each scatterplot. The y-value for each of the four scatterplots are: (1) Scatterplot #1: Robbery rate (2) Scatterplot #2: Property crime rate (3) Scatterplot #3: Aggravated assault rate and (4) Scatterplot #4: Motor vehicle theft rate. Add a regression line to each. Which scatterplot has the steepest regression line?

Task #6: Are the points in Scatterplot #1 in Task #5 clustered in a similar shape across the other three scatterplots (Scatterplots #2, 3 and 4)? In other words, for the points (law enforcement agencies) that make up the main shape of the distribution of points in Scatterplot #1, do these law enforcement agencies show a similar shape across the other three scatterplots (Scatterplots #2, 3 and 4)?