Ersch. in: Information Visualization ; 9 (2010), 3. - S. 194-203

http://dx.doi.org/10.1057/ivs.2010.4

Original Article

Variable binned scatter plots

Ming C. Hao[a,*]
Umeshwar Dayal[a]
Ratnesh K. Sharma[a]
Daniel A. Keim[b] and
Halldór Janetzko[b]

[a]Hewlett Packard Laboratories, 1501, Mill Road, ms 1u-2, Palo Alto, California, USA.
[b]Department of Computer and Information Science, University of Konstanz, Box 78, 78457, Konstanz, Germany.

*Corresponding author.

Abstract The scatter plot is a well-known method of visualizing pairs of two continuous variables. Scatter plots are intuitive and easy-to-use, but often have a high degree of overlap which may occlude a significant portion of the data. To analyze a dense non-uniform data set, a recursive drill-down is required for detailed analysis. In this article, we propose *variable binned scatter plots* to allow the visualization of large amounts of data without overlapping. The basic idea is to use a non-uniform (variable) binning of the x and y dimensions and to plot all data points that are located within each bin into the corresponding squares. In the visualization, each data point is then represented by a small cell (pixel). Users are able to interact with individual data points for record level information. To analyze an interesting area of the scatter plot, the variable binned scatter plots with a refined scale for the subarea can be generated recursively as needed. Furthermore, we map a third attribute to color to obtain a visual clustering. We have applied variable binned scatter plots to solve real-world problems in the areas of credit card fraud and data center energy consumption to visualize their data distributions and cause-effect relationships among multiple attributes. A comparison of our methods with two recent scatter plot variants is included.
Information Visualization (2010) **9,** 194 – 203. doi:10.1057/ivs.2010.4

Keywords: variable binned scatter plots; correlations; patterns; cause–effect; data distribution

Introduction

Motivation

Scatter plots are one of the most powerful tools for data analysis in daily business operations. Analysts face the challenge of understanding underlying data and finding important relationships from which to draw conclusions, such as answering questions on how one variable is affected by another. For example, in credit card fraud analysis, business analysts want to know fraud impact factors (that is, amount, count and region) and distributions. Data center managers want to find the cause-effect of resource consumption to increase energy savings. A scatter plot of power consumptions against temperature can show the impact between these two variables for administrators to improve cooling efficiency.

Scatter plots are widely used, intuitive and easy to understand. However, scatter plots often have a high number of overlapping data points. When there are many data points and significant overlap, scatter plots become less useful. Besides, in analyzing a very dense dataset, the current scatter plots are too coarse and recursion in certain areas is needed. For example, the traditional scatter plot in Figure 1(a) shows 70465 fraud observations, but only about 200 distinct data points are visible in the scatter plot, which may mislead the user in judging the density of the data. In addition, users can not distinguish the details in the bottom left corner. There are several approaches that can be used when this occurs (for details see Section 'Related work').But the difficulties still remain,

This paper has been accepted as one of the best in VDAIO 'Visual analytics of large multidimensional data using variable binned scatter plots' Paper no. 7530-5.

Received: 26 February 2010
Revised: 12 June 2010
Accepted: 14 June 2010

especially when visualizing very large multi-dimensional high-density data sets. Current scatter plots do not provide a complete picture of the data regarding:

- detailed information at record level;
- overlapping data points;
- distribution and clusters in the high-density areas;
- scatter plot only has a fixed scale.

Our contribution

First, our contribution is to allow users to see an entire picture of their data, such as data distribution and correlations. In order to visualize the data distributions and discover cause effect between attributes (variables), we have to solve the overlap problem. Our solution is the *variable binned scatter plot*. First, the data set is binned into proper value ranges. Then, we use density estimation with distortion techniques to place overlapping data points that fall within each bin into corresponding square. The bin size is variable and is computed from the data density. The degree of variation is optimized based on the number of overlapping data points and the available space.

Second, we introduce a third dimension in scatter plots. Analysts are able to use the value of a third attribute as the color of the data points. With the color, data points can be classified into different groups (clusters). This feature is especially useful in placing overlapping data points by certain categories (that is, sales regions). Overlapping data points are sorted by the value of the third attribute and then placed together to form clusters. Variable binned scatter plots can be extended into a variable binned scatter plot matrix to display pairwise relations between multiple attributes.

Each data point is represented by a pixel.[1,2] Because a pixel is the smallest element on the screen, large volumes of data points can be displayed in a single view. Variable binned scatter plots are interactive. Analysts can rubber band an interesting area and zoom into detailed information. Using a recursive drill down, users are able to select one or multiple bins in the variable binned scatter plot to generate a new variable binned scatter plot with refined data value range of *x*-axis and *y*-axis for detailed analysis.

Variable binned scatter plots have been applied with success to real-world credit card fraud analysis and data center thermal management applications. Both applications use variable binned scatter plots to visualize data distributions and impacts among various factors.

This article is structured as follows: Section 'Related work' provides an overview of related work. Section 'Our approach' introduces the variable binned scatter plots basic idea and four basic techniques: pixel cell-based representation, binning and distortion, placement and grouping, and recursive scatter plot generation. In Section 'Applications', we present application examples in which real-world data are used to demonstrate the effectiveness of our techniques. An evaluation of the strengths and weaknesses of our approach versus other variants is presented in Section 'Evaluation'.

Related Work

The scatter plot is a well-known data analysis method to show how much one variable is affected by another. Overlap is always a problem in visualizing high-density data sets using scatter plots. In 1984, Cleveland[3] introduced sunflowers to draw overlapping points and superposition of smoothing methods for enhancing the *x*- and *y*-axes in scatter plots. Cleveland's ideas are great improvements of scatter plots, but they do not solve the overlap problem. In 1999, Lee Wilkinson[4] suggested the usage of semi-transparency to make overlapping data points partially visible.

In the book by Antony Unwin *et al.*,[5] a number of interesting visualization techniques were introduced regarding scatter plots, such as drawing overlap points with slightly bigger sizes and reducing the *x*- and *y*-axis by certain factors. JMP 8 Software[6] generates scatter plots with nonparametric density contours and marginal distributions to show where the data is most dense. Each contour line in the curved shape encloses 5 per cent of the data. Carr[7] uses a hexagonal-shaped symbol whose size increases monotonically as the number of observations in the associated bin increases, and HexBin scatter plots[8] determine the brightness value of each HexBin cell depending on the number of data points in the cell. All three techniques, Unwin's distortion, Carr's binning and the HexBin visualization techniques, are close to the method presented in this article. Bowman's smooth contour scatter plot[9,10] applies smoothing techniques for data analysis. Bachthaler's continuous scatter plots[11] are different from the above scatter plots. Continuous scatter plots are used for visualizing spatially continuous input data instead of discrete data values.

The above approaches provide excellent methods for data correlation analysis. However, analysts are not able to see and access all data points, especially if the third variable mapped to color is of high importance. In this article, we introduce the variable binned scatter plot technique with recursive visual analytics. We combine the best features of the above methods (for example, binning and zooming) with distortion to find the best placement for the overlapping data points to enable analysts to quickly discover distribution and clusters. Also, we use color to visualize the third attribute, while the previous approaches use color to represent density. This feature helps the users to quickly identify patterns and clusters.

Our Approach

Basic idea of variable binned scatter plots

Binning is an efficient approach[3] to reduce the complexity of large volumes of multi-dimensional data by dividing

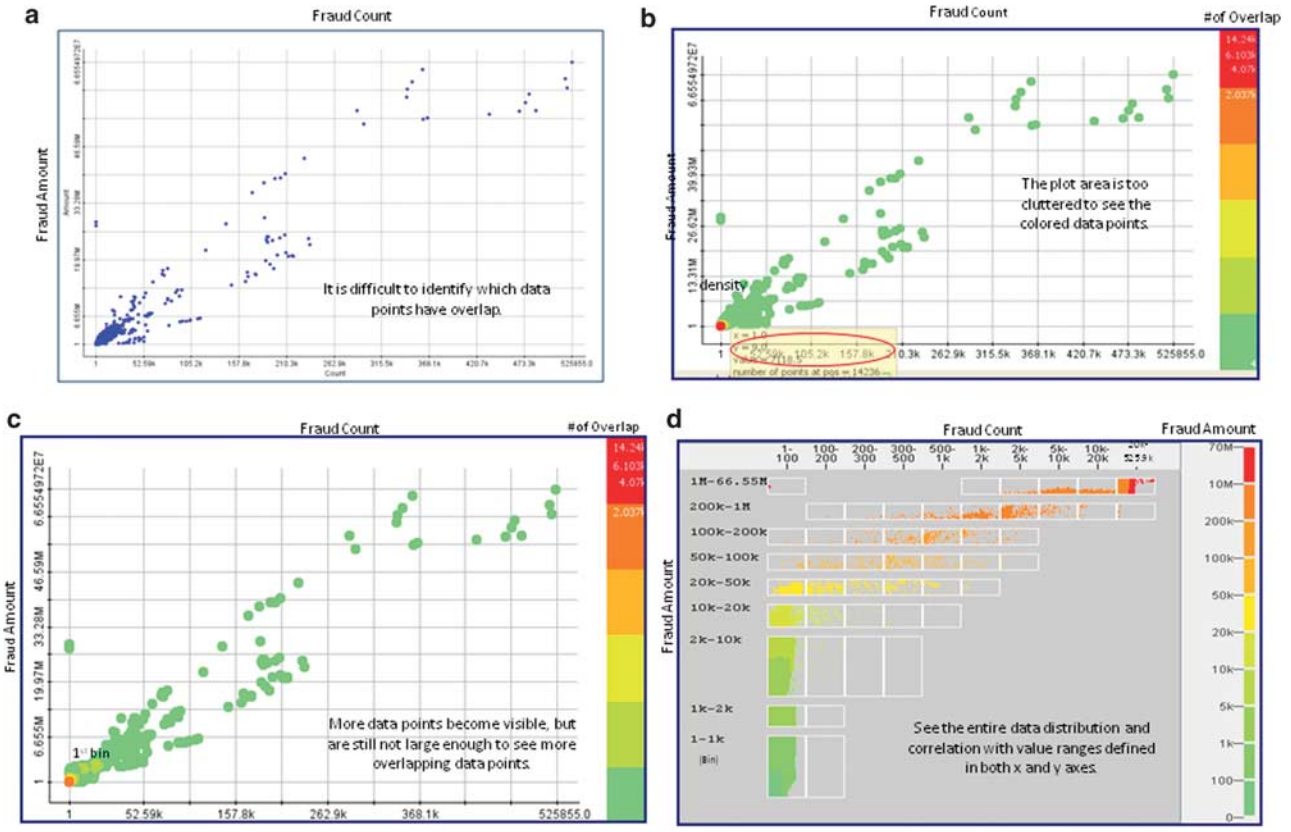

Figure 1: (**a**) Traditional scatter plots (70 465 data points). Most data points overlap: only 200 data points are visible; (**b**) Color that overlapping points by the number of data points which have the same (x_i, y_i) position; (**c**) Slightly enlarge the cluttered area heaving less overlap; (**d**) Variable binned scatter plot without overlapping show credit card fraud among distribution and correlation (low: green; medium: yellow; red).

the plot area into a number of value ranges. We introduce the concept of the variable binned scatter plot to manage large volumes of data that overlap. Variable binned scatter plots are derived from traditional scatter plots to address the issue of overlapping. Variable binned scatter plots group data in a two-dimensional space based on the densities of pairs of variables. Each data point defined as (x_i, y_i, z_i) for i from 1 to n consists of a pair of two variables, x and y. A scatter plot of x_i against y_i shows the relationship between x and y; and color z to show a third variable. Variable binned scatter plots employ color (z_i) to cluster related data points.

Figure 1(a)–(d) illustrate the progression from traditional scatter plots to variable binned scatter plots. The scatter plot in Figure 1(a) has many overlapping points. There is no indication as to which data points are overlapping, potentially resulting in a misleading data representation. In Figure 1(b), the overlapping data points use the color to denote the number of data points which have the same (x_i, y_i) position, but the plot area is too dense to see all the colored data points. In the scatter plot in Figure 1(c), the first bin is slightly enlarged resulting in less overlap. More data points become visible, but it is still not possible to see all data points with their distributions and patterns.

Variable binned scatter plot in Figure 1(d) uses a non-uniform (variable) binning of the x and y dimensions and plots all the data points that fall within each bin into the corresponding square area. These square areas are scaled to allow each data point to be shown without any overlap. The relative position of a data point within a bin is retained as accurately as possible. Users are now able to visualize all the data points without losing information. Users are able to visualize the impact between two variables accurately and quickly, and without misrepresentations of the data. Variable binned scatter plots enhance the traditional scatter plots in analyzing very large and dense data sets.

Construction of variable binned scatter plots

Figure 2 illustrates a pipeline on how to construct a variable binned scatter plot using the following techniques:

(1) *Use of pixel cells to represent data points in a binned scatter plot*: Variable binned scatter plots use the smallest

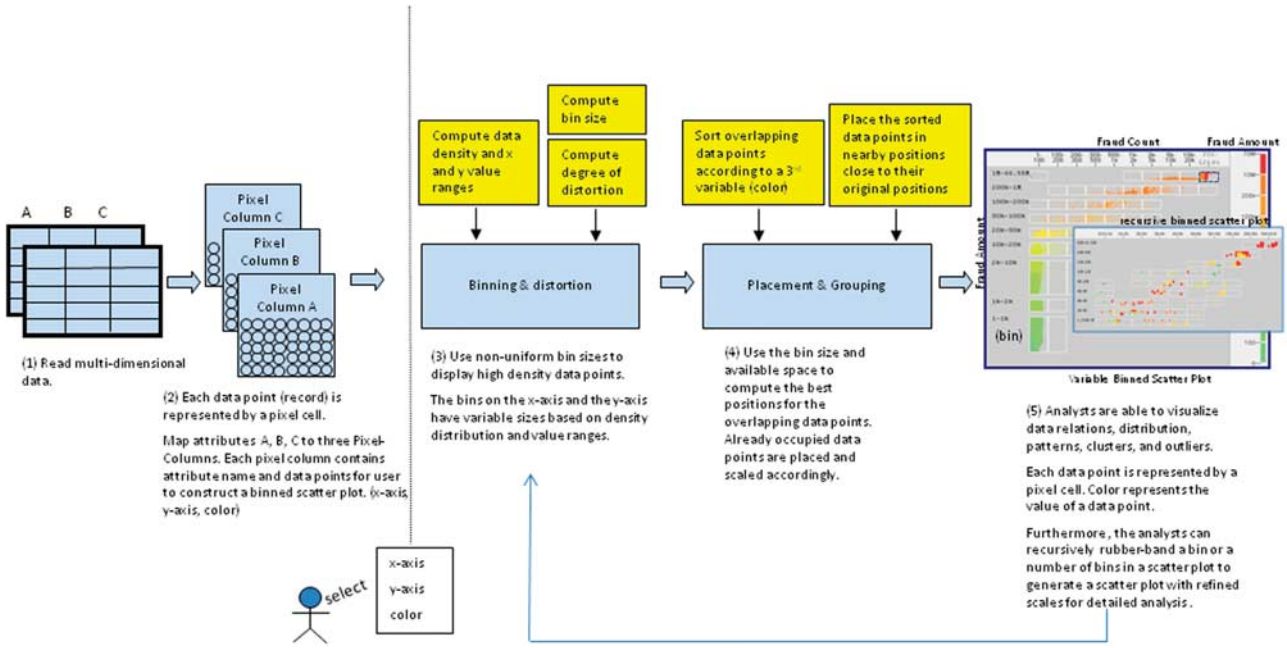

Figure 2: A Variable Binned Scatter Plots Construction Pipe Line.

element on the screen, such as a pixel, to represent a data point. Analysts are able to view large volumes of data points in a single display. Data points of binned scatter plots are interactive. Analysts can zoom-in on a data point to view specific data attributes. Intelligent visual queries[12] are also provided for analysts to select a focused area in a scatter plot.

Then apply automated analysis methods to identify characteristics of the selected data as well as their relationships to other attributes and data points.

(2) *Binning and distortion*: Based on Carr *et al.*'s definition,[3] binning is an approximation for density of the joint distribution of two variables. Our variable binned scatter plots use non-uniform bin sizes to display high-density areas. A bin contains the data points which have their (x, y)coordinates within defined x and y value ranges. The binning of the x- and y-axis is determined according to the data value ranges which are computed from the incoming data and their density distribution. When the data size grows, the bin size will be enlarged accordingly. Therefore, there are no overlapping data points in the variablebinned scatter plots. The following illustrates the overall binning algorithm using a non-uniform graphical density display:[13]

- Determine the density distribution and value ranges in x and y directions.
- Assign the number of bins in x and y directions and compute the bin size based on data density distributions and value ranges.
- Determine bin width according to the total window width divided by the number of bins on the x-axis.
- Determine bin height for each row according to the maximum number of data points of all bins in the corresponding row.

In our current application, the bins on the x-axis use equal widths based on the window size. The bins on the y-axis have different heights according to the maximum number of data points within the bins in the row.

(3) *Placement and grouping*: variable binned scatter plots place the non-overlapping data points according to their x and y coordinates within the corresponding bins. The overlapping data points are sorted according to the value of the third attribute to form groups in two-dimensional space. The placement algorithm uses the available space around the already occupied data points to compute the best location for the data points that would otherwise be overlapping in a traditional scatter plot. Data points with the same x and y coordinates are sorted and placed in nearby neighborhood according to the similarity of the third attribute (color).

Figure 3(a) illustrates 11 data points with the same (x, y) coordinates. The data point P is overlapped by the data points P1 through P10. Overlap causes two problems in visualizing data distributions and patterns: (i) the number of overlaid data points is unknown and (ii) the value of overlaid data points is not visible. Figure 3(b) shows how to place the overlapping data points to form a square group around the data point P ordered by the third variable values. If the neighborhood position is already occupied, then the bin axes will be proportionally enlarged and will push the already occupied data points away along

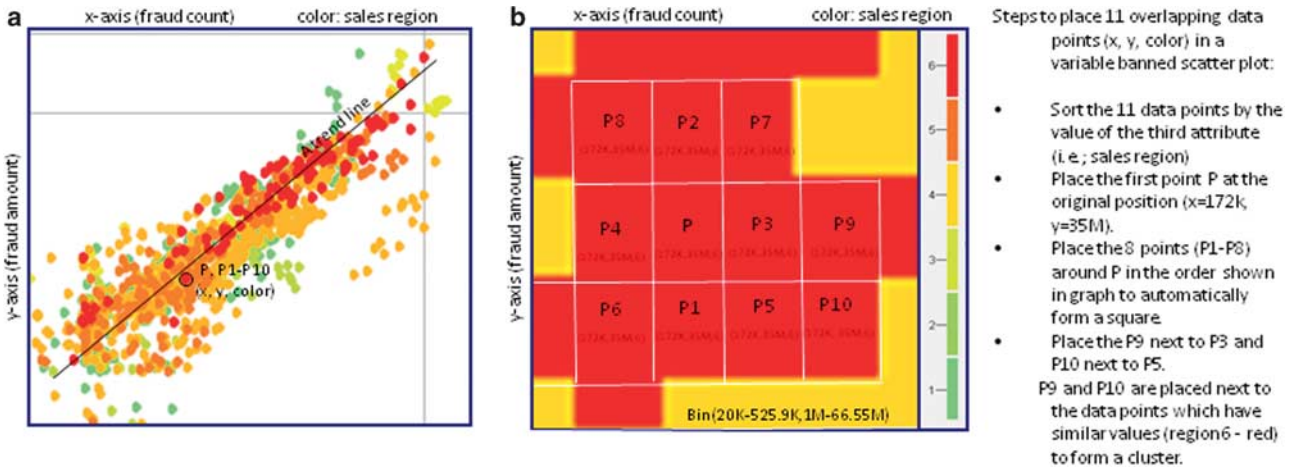

Figure 3: (**a**) Traditional scatter plot with overlapping data point P is overlapped by P–P10; (**b**) Variable binned scatter plot without overlapping data point P1–P10 are order and placed around P to from a red square cluster.

the x (toward right) and y (toward top) directions. As the result of this placement process, a red square for sales region 6 is constructed from the 11 overlapped data points. Our algorithm recursively searches for an x and y location. The placement of the next data point i is at a $\min[(x_i - x_p) + (y_i - y_p)]$, where (x_i, y_i) is an unoccupied location. If necessary, we need to increase the bin size either along the x-axis or y-axis. For example, the 12th data point will be placed at the bottom of 'P9' in Figure 3(b). This location is unoccupied and has a smallest distance from p.

(4) *Recursive binned scatter plots generation*: users can select adjacent bins or a region to construct a subset of the variable binned scatter plot. The algorithm uses the data points from the selected bins to generate a new variable binned scatter plot for the selected data points with a new binning adapted to the data distribution of the selected data. The (x, y) coordinates are defined by the value range of the data points from the selected bins. The color remains the same. The scale is computed with a refined scale based on the new data ranges. Figure 4 shows a sequence of recursive variable binned scatter plots (plots #1, #2 and #3) in a credit card fraud analysis application (Section 'Credit card fraud analysis'), generated by the user rubber-banding a high-density area (x = 20 K–525.9 K, y = 1 M–66.5 M, color = region). In the variable binned scatter plot #2, users are able to detect the correlation between fraud amount and fraud count using the refined scale from fraud amounts in the ranges 1.254 M, 3 M..., to 66.55 M. The analyst can further select another group of bins in plot #2 to compare fraud distribution in the refined data ranges, such as fraud amount 100 K–525.9 Kand fraud count 12 M–66.55 M. The result is shown in the plot #3.

Applications

Credit card fraud analysis

Fraud is one of the major problems faced by many companies in the banking, insurance and telephone industries. Large volumes of dollars in fraudulent transactions are processed yearly on credit card payments. Transforming raw transaction data into valuable business intelligence to support fraud analysis will save companies millions of dollars. Fraud analysis specialists require visual analytics tools that help them to better understand fraud behavior, geographical locations and correlated factors as well as identify exceptions.

Typical questions in fraud analysis are:

Q1. What is the fraud distribution and which are the most correlated attributes?
Q2. Are there any outliers and what are their causes?
Q3. Which sales regions and purchase amount have the most fraud?

Plot#1 in Figure 4 shows a binned scatter plot with 70 465 fraud records. Analysts use it to analyze fraud distributions and correlations among different attributes (that is, amount, count and region) to answer the first question. In a variable binned scatter plot, each fraud data point is represented by a pixel. Because there are no overlapping data points in a variable binned scatter plot, analysts are able to visualize fraud distribution at each data point along the x and y directions. The binning of the x and y directions is determined according to the fraud amount and fraud count. The color of a data point represents the sales regions 1–6 where the fraud occurred. Plot #1 shows that the fraud amount is almost increases linearly with fraud count. The fraud amount is highly impacted by the

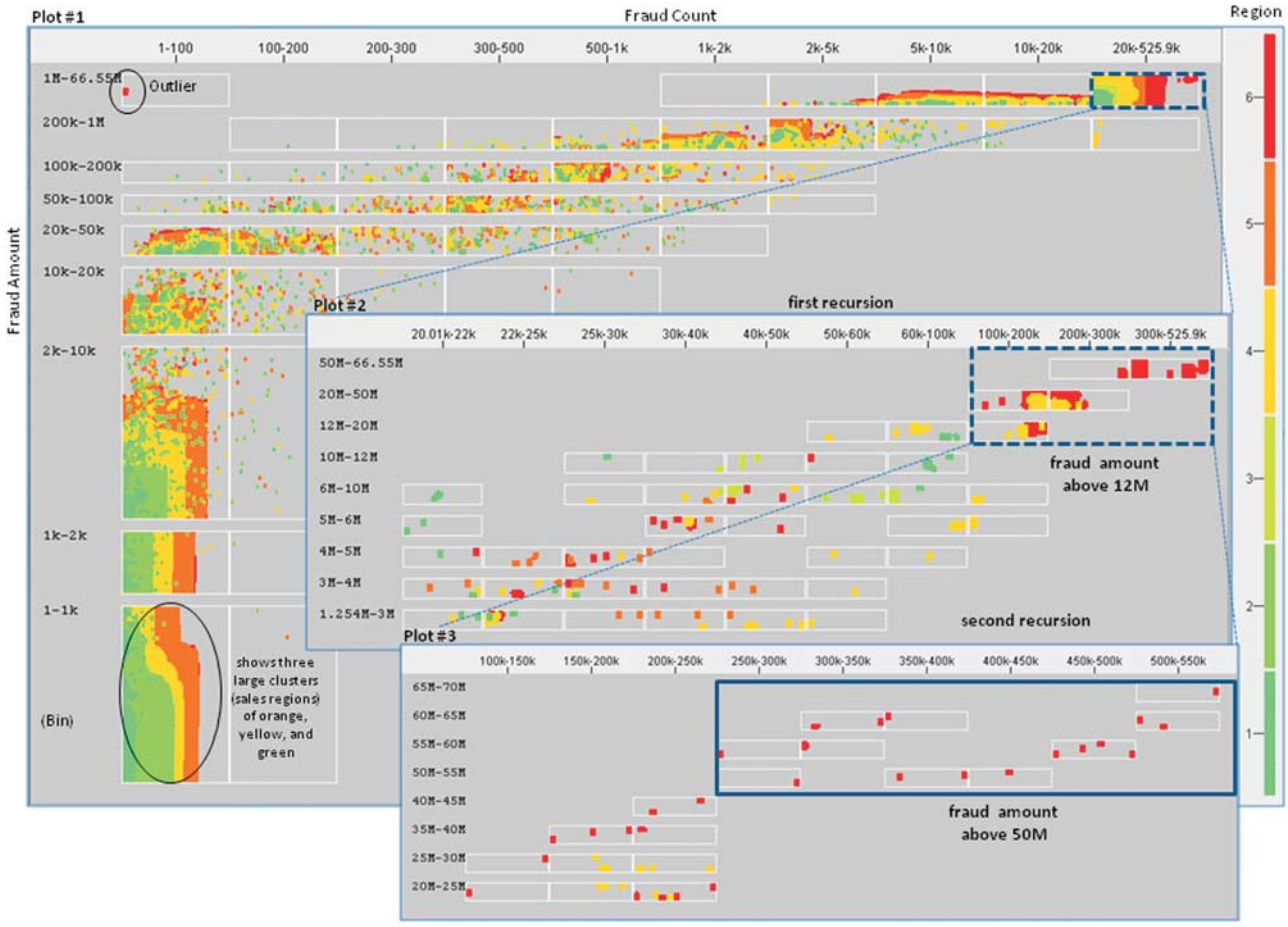

Figure 4: A Credit Card Fraud Analysis Variable Binned Scatter Plot (*x*-axis: Fraud Count, *y*-axis: Fraud Amount, Color: Regions 1 to 6).

fraud count. In addition, there is an outlier at the top left bin ($x = 1$–100, $y = 1$ M–66.55 M) with a low fraud count of 5 but a very high fraud amount of $28.107 million that occurred in region 6 (red).

In order to answer the third question on finding which sales region in which fraud amount range (bin) has the most fraud, we optimize data point placement, so that data points from the same sales regions (colors) are placed together (the technique is described in Section 'Our approach') for analysts to see the fraud regional distribution. There are three large clusters (orange, yellow and green) in bin ($x = 1$–100, $y = 1$–1 K). Sales region 1 (green) has the lowest fraud amount and count. The smallest cluster is sales region 6 (red) but with the highest fraud amount and count. To find which fraud amount and region have the most fraud, analysts can first select a bin, such as bin ($x = 20$ K–525.9 K, $y = 1$ M–66.55 M) and then recursively drilldown to generate the binned scatter plots #2 and #3 with refined scales. From plot #2, analysts can learn that the most frauds came from region 6 (red) and the purchase amount is above $1.254 million. From plot #3, analysts can also learn that the highest fraud amount is in the refined data ranges ($x = 250$ K–525.9 K, $y = 50$ M–70 M). Using the above information, the company is able to place strict control on certain sales regions and purchase amount, such as sales region 6 and a purchaseamount above $50 million.

Data center thermal monitoring

Cooling is the major operational cost in a data center. The chiller consumes over 600 KW of power in order to keep a normal temperature for the daily IT load in a data center with 500 racks and 11 air conditioning units. Chillers consume power to extract heat from the warm water and provide cold water to the air conditioning units to keep the data center temperature cool. Visual monitoring of the utilization of chillers and power consumption and their impacts on temperatures can greatly reduce operating expenses and equipment downtime.

Questions of a data center service manager's frequent concern are:

Q1. What is our daily power consumption? How do we optimize the cooling system performance?

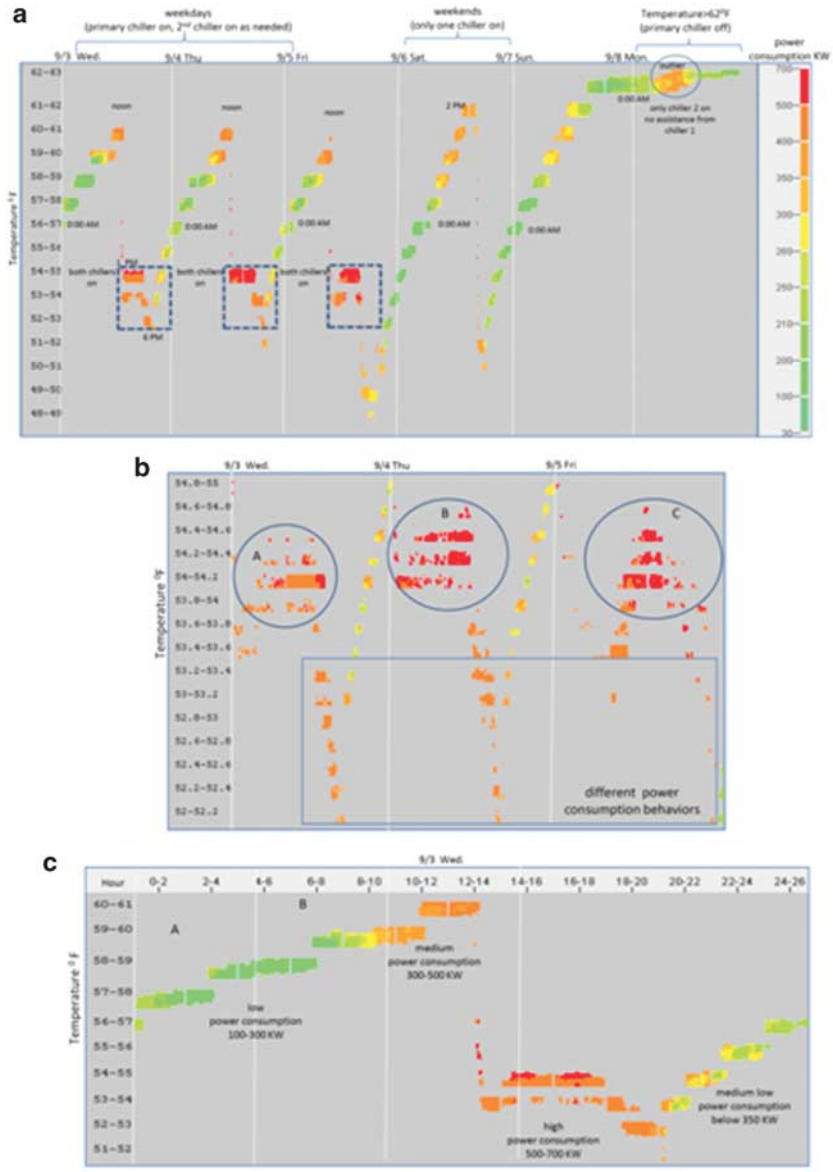

Figure 5: (**a**) Data Center Temperature Time Series Visual Analysis Using Variable Binned Scatter Plots (x-axis: days 9/3 to 9/8, y-axis: Temperature °F, color: Power Consumption KW). (**b**) Power Consumption Visual Analytics Using Recursive Binned Scatter Plots Generated from the three rubber-band areas in Figure 5 (x-axis: days 9/3 to 9/5, y-axis: Temperature °F, color: Power Consumption KW); (**c**) Visual Analysis of Hourly Temperature and Power Cost-Effect Using Recursive Binned Scatte Plots Generated from the 9/3 time series in Figure 5A (x-axis: 9/3, y-axis: Temperature °F, color: Power Consumption (KW).

Q2. How is the chiller operating? Are there any problems?
Q3. What are the cause-effects of the power consumption on temperature?

To answer the above three questions, we have used variable binned scatter plots to enable administrators to visualize the relationships and impacts among these three thermal factors: temperature, power consumption and chiller utilization. Figure 5(a) shows a time series scatter plot. Each data point is represented by a pixel and defined with three attributes (x, y, color) where the x-axis represents the time line from 9/3 to 9/8. The y-axis represents temperature. The color of the data point represents the power consumption which is needed by the chiller, from low (green) to medium (yellow, orange) to high (red). With variable binned scatter plots, the overlapping data points are sorted and placed as close as possible to their original locations as described in Section 'Our approach'.

Figure 5(a) shows, in weekdays (9/3, 9/4 and 9/5), the power consumption is higher during the day and early evening (more orange and red), than during the early morning and late evening (mostly green). This result

helps the administrators to optimize cooling system performance. During weekends (9/6 and 9/7), less power is used (most green and yellow). But there is an exception is found on 9/8 (Monday). The temperature remains high (above 62°F) and even the power consumption is above 400 KW (orange), between 10:00 and 20.00.

Recursive drilldown in variable binned scatter plots

To answer the second question on how the chiller is operating, the analysts can rubber-band the three high power consumption areas (on day 9/3, 9/4 and 9/5, hours 14:00–18:00, temperature 52°F–55°F) to generate a recursive plot in Figure 5(b).

Figure 5(b) shows the generated scatter plot which has a different distance scale. The plot contains three different power consumption patterns (A, B and C). In order to keep their high temperature under 54.6°F, the administrator has to turn on both chillers with power consumption above 500 KW. Administrators can quickly find that pattern B consumes more power than patterns A and C (more red data points). In addition, administrators notice that pattern C has different power consumption pattern form patterns A and B. Pattern C has less data points under 53°F. Administrators can use this information to identify impact of energy efficiency measures within the data center.

To answer the third question on the cause–effects of the power consumption on temperature and time of a day, the administrator can select the entire bin on day 9/3 as shown in the Figure 5(a) and generate a recursive variable binned scatter plot. The resulting plot is shown in Figure 5(c).

Figure 5(c) shows different power consumption patterns from low (green), medium (yellow), high (orange and red). This result helps the administrators to optimize cooling system performance. From this observation, administrators are able to use less power (green, under 250 KW) in the early morning and late evening and then gradually increase the power (yellow and orange, over 300 KW) between 10:00 and noon. Especially during the peak hours 14:00–18:00, power could be increased greater than 500 KW (red) for the chiller to cool down the temperature to less than 55°F.

Evaluation

There are many well-known variations of the traditional scatter plot that try to solve the overlap problem of scatter plots. The *HexBin*[8] and *smoothed contour* scatter plots[9,10] are two recent variants which are also available in the R statistics software. We will address the question 'Can the HexBin and smooth contour scatter plots achieve the same results as our variable binned scatter plot?'

Figures 6(a)–(d) shows the HexBin and smoothed contour scatter plots with the same number of fraud records (70 465) and the same data center resource consumption data (43 204) as the variable binned scatter plot shown in Figures 6(e) and (f). An evaluation of the strengths and weaknesses of the three approaches follows.

The strengths of the variable binned scatter plot include:

(1) Variable binned scatter plots show fraud and thermal distribution in the high-density areas, marked by the dashed rectangle (Figures 6(e) and (f)) more clearly than either HexBin scatter plots Figures 6(a) and (b) or the smooth contour scatter plots (Figures 6(c) and (d)). Smoothed contour scatter plots show linearly increasing overlaps with different shades which are better than HexBin scatter plots.
In most applications, the majority of data points occur in the high-density areas. Both variants require an extra step of zooming into these areas (that is, dashed rectangles). The variable binned scatter plots provide a big picture of the entire distribution without additional drilldown. Furthermore, the variable binned scatter plot can quickly identify clusters as well as reveal hidden structures in the dense areas.
(2) Variable binned scatter plots map the value of a third attribute to color in order to visualize the extra dimension by clustering data points as shown in Figure 6(f) (that is, power KW). In the HexBin scatter plot, it is not possible to use color to represent a different attribute at same time. Variable binned scatter plots have one more dimension to use than HexBin and smooth contour scatter plots allowing the third attribute to bevisible in the same scatter plot.
(3) Since the data is aggregated in HexBin and smooth contour scatter plots, it is not possible for users to interact with a data point for detailed information. All data points in variable binned scatter plots are accessible and readily viewable.
(4) Variable binned scatter plots provide a recursive drilldown capability which allows analysts to view detailed information by using refined scales.

The weaknesses of the variable binned scatter plot include:

(1) The HexBin and smooth contour scatter plots show a better trend line than the variable binned scatter plot. Trend line is visible with the variable binned scatter plots, but requires users to follow the bins (value ranges).
(2) The HexBin and smooth contour scatter plots use different shading to visualize data density while the variable binned scatter plots introduce some distortion to visualize all data points.

In summary, both HexBin and smooth contour scatter plots are able to provide a quick overview of data density and correlations. Variable binned scatter plots visualize

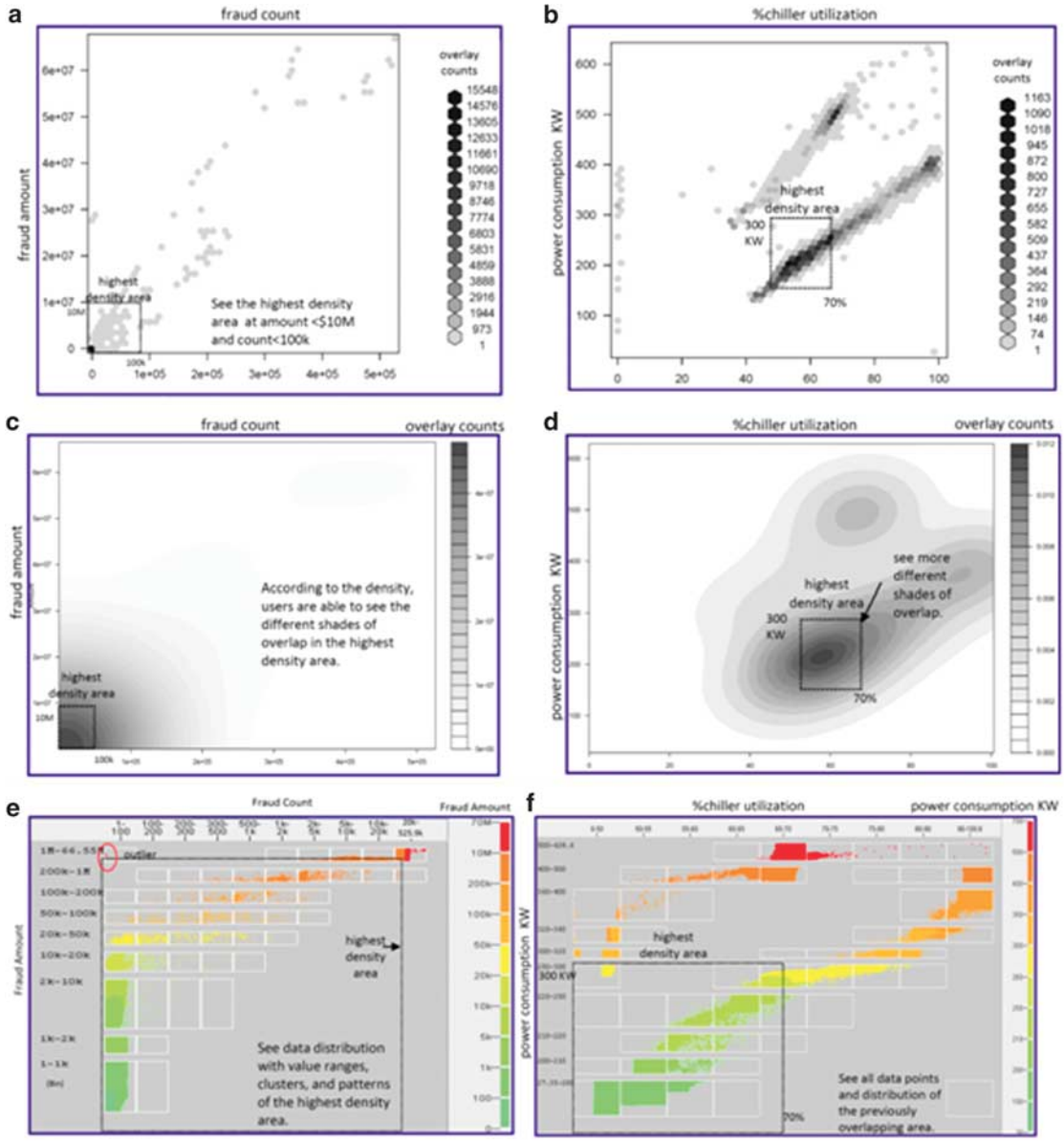

Figure 6: (**a**) Hexbin Scotter Plot: Only see one data point with high overlaps in a dashed rectangle. (**b**) Hexbin Scotter plot show: high overlapping area inside a dashed rectangle. (**c**) Smooth contour scotter plot shows different degree of overlapping area in the high-density area. (**d**) Smooth contour scotter plot shows different density area more clearly. (**e**) Variable Binned Scatter Plot: shows all data points and their distributions with value ranges in the highest density area. (**f**) Variable Binned Scatter Plot shows all data points and their distribution, correlations, and clusters in the highest density area.

the entire data distribution but also allow access to each individual data point for users to retrieve information at the record level. These three variants of scatter plots complement one another.

Conclusion

In this article, we introduce variable binned scatter plots with recursive drill down for a visual analysis of data distributions and cause–effect of multi-dimensional data in addition to correlation between pairs of variables. Variable binned scatter plots resolve the overlapping issue and allow users to visually analyze large data sets at the record level. A minimal distortion is introduced to provide space for the overlapping data points. We are able to use color for a third attribute of the data to help analysts quickly identify patterns and clusters. The recursive drill down capability of the variable binned scatter plots provides detailed information with refined scales for users to further analyze

important areas in a plot. An evaluation of the recent HexBin and smooth contour scatter plots demonstrates the benefits of using variable binned scatter plots to reveal clusters and distribution in a high-density area.

We have applied variable binned scatter to the real-world applications, such as credit card fraud, Most fraud analysts and data center administrators (novice) can quickly identify the dense areas which used to be over-lapped data points and . Our future work will be in the area of visual prediction using scatter plots to study the trends. In addition, we would like to read in live data stream to construct variable binned scatter plots and detect anomalies in real time.

Acknowledgements

The authors thank Meichun Hsu for her encouragements, and thank Alex Zhang, Manish Marwah and Cullen E. Bash for providing comments and suggestions.

References

1 Keim, D.A. (2000) Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 6(1): 59–78.

2 Keim, D.A., Kriegel, H.P. and Ankerst, M. (1995) Recursive pattern: A Technique for visualizing very large amounts of data. Proceedings of the IEEE Visualization, pp. 279–286.

3 Cleveland, W.S. (1984) The many faces of a scatterplot. *Robert McGill Journal of the American Statistical Association* 79(388): 807–822.

4 Wilkinson, L. (1999) *The Grammar of Graphics*. New York, Singh Mela, Mr Excel, Ohio, USA: Springer.

5 Unwin, A., Martin, T. and Heike, H. (2006) *Graphics of Large Datasets*. New York: Springer, 39–193.

6 JMP 8 Software. New 64-bit computers and visual analytics tools, http://www.jmp.com/software.

7 Carr, D.B., Littlefield, R.J., Nicholson, W.L. and Kuttkefuekdm, J.S. (1987) Scatterplot matrix techniques for large N. *Journal of American Statistics Association* 82: 424–436.

8 HexBin scatter plot released by R System in January 2009, https://stat.ethz.ch/pipermail/r-help/2009: documented http://rss.acs.unt.edu/Rdoc/library/hexbin/doc/hexagon_binning.pdf.

9 Bowman, A.W. and Azzalini, A. (1997) *Applied smoothing techniques for data analysis: The Kernel Approach with S-Plus Illustrations*. Oxford: Oxford University Press.

10 Bowman, A.W. and Azzalini, A. (2003) Computational aspects of nonparametric smoothing with illustrations from the SM library. *Computational Statistics and Data Analysis* 42: 545–560.

11 Bachthaler, S. and Weiskopf, D. (2008) Continuous scatterplots. *IEEE Transactions on Visualization and Computer Graphics* 14(6).

12 Hao, M., Dayal, U., Keim, D.A., and Morent, D. (2007) Intelligent visual analytics queries. *IEEE Symposium on Visual Analytics Science and Technology* 91–98.

13 Hao, M., and Dayal, U. (2006) Method for visualizing graphical data sets having a non-uniform graphical density for display. US patent number 7,046,247 issued in May, 2006.