

Visual Analytics of Large Multi-Dimensional Data Using Variable Binned Scatter Plots

Ming C. Hao, Umeshwar Dayal, Ratnesh K. Sharma
Hewlett Packard Laboratories, Palo Alto, CA

Daniel A. Keim, Halldór Janetzko
University of Konstanz, Germany

ABSTRACT

The scatter plot is a well-known method of visualizing pairs of two-dimensional continuous variables. Multi-dimensional data can be depicted in a scatter plot matrix. They are intuitive and easy-to-use, but often have a high degree of overlap which may occlude a significant portion of data. In this paper, 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 plots all the data points that fall within each bin into corresponding squares. Further, we map a third attribute to color for visualizing data distribution and clustering. Analysts are able to interact with individual data points for record level information. We have applied these techniques to solve real-world problems on credit card fraud and data center energy consumption to visualize their data distribution and cause-effect among multiple attributes. A comparison of our methods with two recent well-known variants of scatter plots is included.

Keywords: Variable Binned, Scatter plots, Correlations, Clusters, Cause-Effect, Data Distribution

1. INTRODUCTION

1.1 Motivation

The scatter plot is 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 (i.e., amount, count, and region) and distribution. 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. For example, the traditional scatter plot in Figure 1A has 70,465 fraud observations, but only about 200 distinct data points are visible in the scatter plot, which may mislead the user about the distribution and density of the data. There are several approaches that can be used when this occurs (see section 2). But the difficulties still remain, 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.
- Non-overlapping data points.
- Patterns and distributions in the high density areas.

1.2 Our Contribution

In order to visualize the data distributions and discover cause-effect between attributes (variables), we have to solve the overlapping issue. Our solution is the *variable binned scatter plot*. First, the dataset 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.

Further, variable binned scatter plot uses 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) to show distributions. This feature is especially useful in placing overlapping data points by certain categories (i.e., sales regions). Overlapping data points are sorted by the value of the

third attribute and then are placed together to form clusters. Binned scatter plots can be extended into a variable binned scatter plot matrix to display pair-wise relations between attributes.

Each data point is represented by a pixel [11]. 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. 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 distribution and impacts among various factors.

This paper is structured as follows: Section 2 provides an overview of related work. Section 3 introduces the variable binned scatter plots basic idea and three basic techniques: pixel cell-based representation, binning, data point placement and grouping. In section 4, we present application examples in which real-world data are used to demonstrate the effectiveness of our technique. An evaluation of the strengths and weaknesses of our approach versus other variants is presented in section 5.

2. 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 [1] introduced sunflowers to draw overlapping points and superposition of smoothing methods for enhancing the x-y axes in scatter plots. Cleveland's ideas are great enhancements of scatter plots, but they do not solve the overlap problem. In 1999, Lee Wilkinson [3] suggested the usage of semi-transparency to make overlapping data points partially visible. Later, Jittering [4] used different markers to visually identify overlapping data points.

In the book by Antony Unwin et al. [2], 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 axes by certain factors. JMP 8 Software [4] 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% of the data. Carr [5] 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 paper. Bowman's smooth contour scatter plot [9, 10] applies smoothing techniques for data analysis. Bachthaler' continuous scatter plots [6] 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 paper, we introduce the variable binned scatter plot technique. We combine the best features of the above methods (e.g., 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 distributions, patterns, and clusters.

3. OUR APPROACH

3.1 Basic Idea of Variable Binned Scatter Plots

Binning is an efficient approach [5] to reduce the complexity of large volumes of multi-dimensional data by dividing 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, c_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 c to show a third variable. Variable binned scatter plots employ color (c_i) to cluster related data points.

Figures 1A to 1C illustrate the progression from traditional scatter plots to variable binned scatter plots. The scatter plot in Figure 1A has many overlapping points. There is no indication as to which data points are overlapping, potentially resulting in a misleading data representation. In Figure 1B, 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

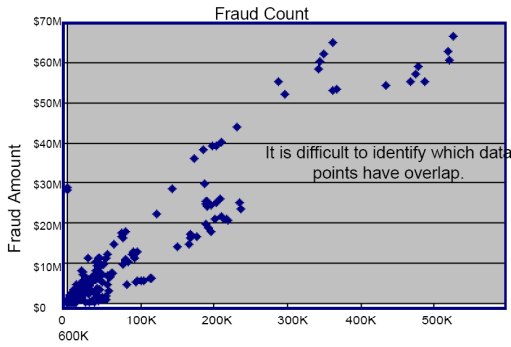


Figure 1A: Traditional Scatter Plots (70,465 data points)
Most data points overlap; only ~200 data points are visible.

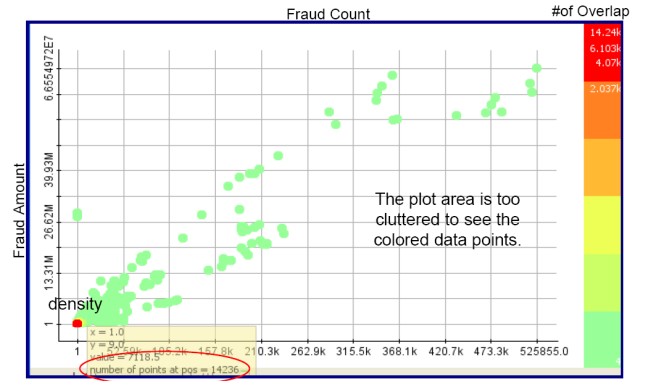


Figure 1B: Color the overlapping points by the number of data points which have the same (xi,yi) position

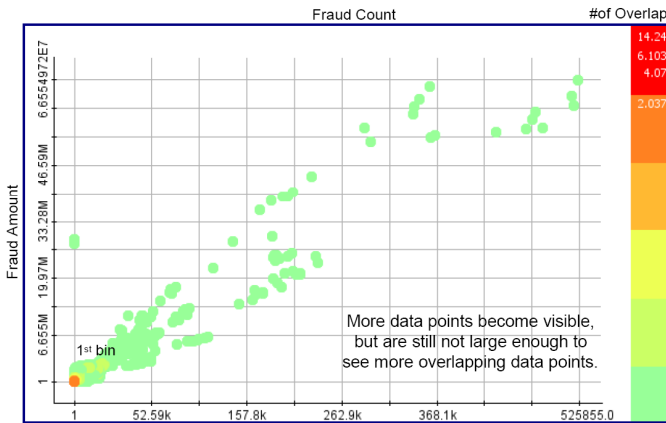


Figure 1C: Slightly enlarge the cluttered area with less overlap.

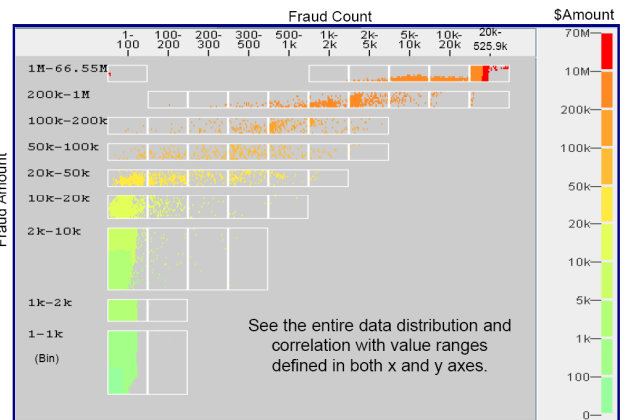


Figure 1D: Variable Binned Scatter Plot without overlapping show credit card fraud amount distribution and correlation (low: green; medium: yellow; high: red)

Figure 1: From Traditional Scatters Plot to Variable Binned Scatters Plot Without Overlapping (x-axis: fraud count, y-axis: fraud amount, color: #of overlapping data points/fraud \$Amount)

points. In the scatter plot in Figure 1C, 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 1D 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. Analysts 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 datasets.

3.2 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 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 [7] are also provided for analysts to select a focused area in a scatter plot and then apply automated analysis

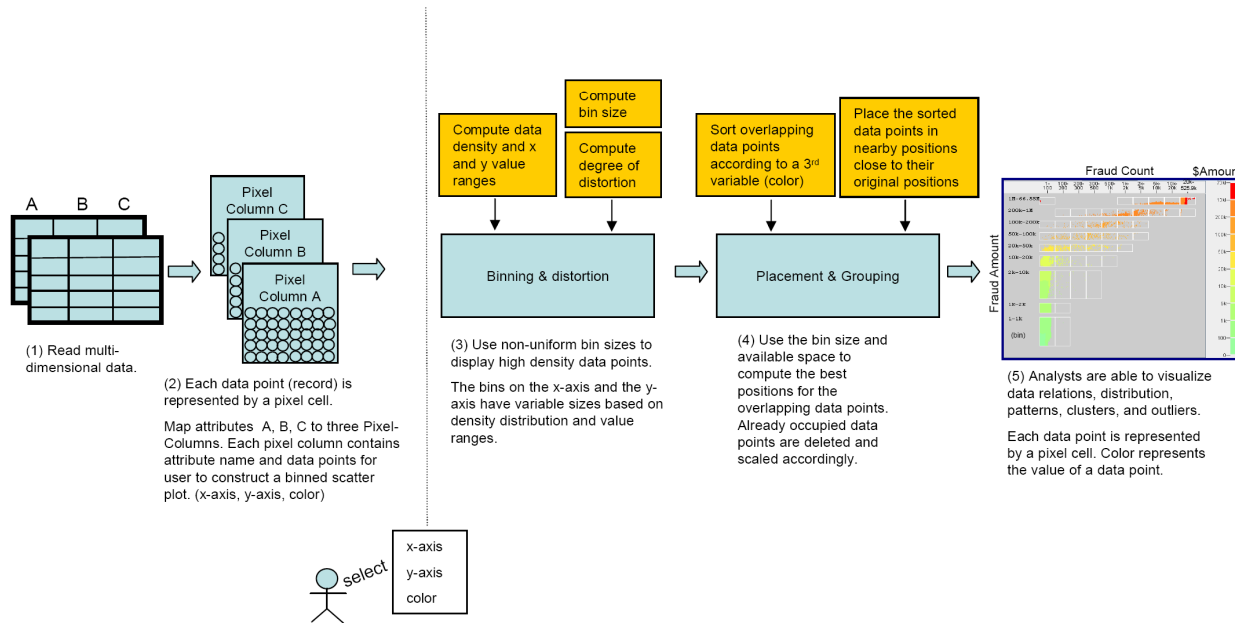


Figure 2: A Variable Binned Scatter Plots Construction Pipe Line

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

2. Binning & Distortion

Based on Carr et al.'s definition [5], 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-axes is determined according to the data value ranges which are computed from the incoming data and their density distribution. The following illustrates the overall binning algorithm using a non-uniform graphical density display [12]:

- 1) Determine the density distribution and value ranges in x and y directions.
- 2) Assign the number of bins in x and y directions and compute the bin size based on data density distributions and value ranges.
- 3) Determine bin width according to the total window width divided by the number of bins on the x-axis.
- 4) 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. For example, in a fraud dataset, a data point P with (x, y) = (172K, 35M) is positioned within the bin (20K-525.9K, 1M-66.55M) where x=20K-525.9K and y=1M-66.55M in Figures 3 and 4.

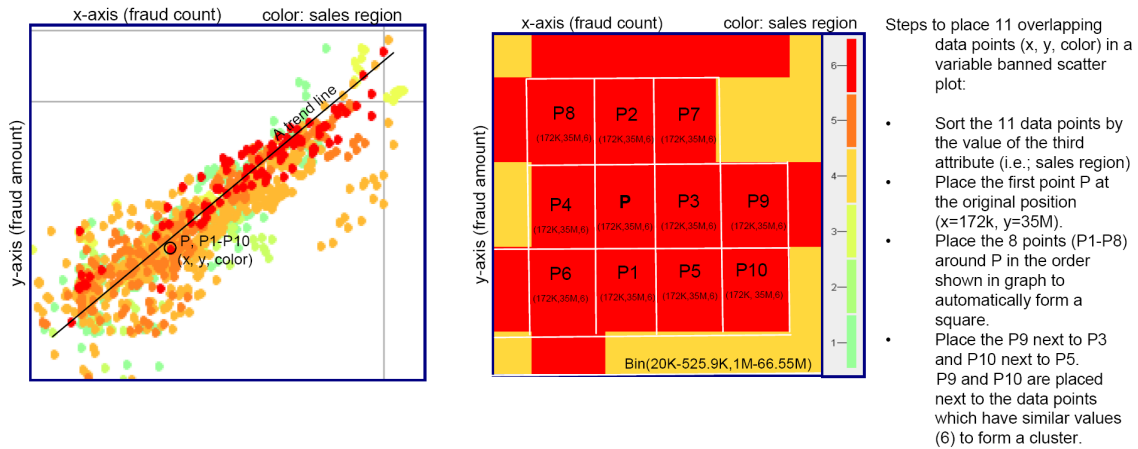


Figure 3A: Traditional Scatter Plot with overlapping. Data point P is overlapped by p1 through P10.

Figure 3B: Variable Binned Scatter Plot without overlapping. Data point P1 to P10 is ordered and placed around P to form a red square cluster.

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 3A 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: (1) the number of overlaid data points is unknown and (2) the value of overlaid data points is not visible. Figure 3B 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 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.

4. APPLICATIONS

4.1 Credit Card Fraud Analysis

Fraud is one of the major problems faced by many companies in the banking, insurance, and telephony 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 distribution, 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. Which sales regions, fraud amounts, and payment types have the most fraud?
- Q3. Are there any outliers and what are their causes?

Figure 4A shows a binned scatter plot with 70,465 fraud records. Analysts use it to analyze fraud distributions and correlations among different attributes (i.e., amount, count, and sales region) to answer the first question. In a variable binned scatter plot, each fraud data point is represented by a pixel. Because there is no overlapping data points in



Figure 4A: A Credit Card Fraud Analysis Variable Bin Scatter Plot (x-axis: Fraud Count, y-axis: Fraud Amount, Color: Region 1-6)

Shows:

- Fraud region distribution in all value ranges including high density areas.
- High correlation between fraud amount and fraud count. Both variables have a strong linear relationship. They are correlated.
- Each bin has various sizes of clusters from different regions.
- Bin (x=1-100, y=1-1K) has three large clusters from regions 2,3,and 5.
- The clusters with the highest fraud amount and count is in bin (x=20K-525.9K, y=1M-66.55M) from region 6 (red).
- One outlier discovered in bin (x=1-100, y=1M-66.55M) with a high fraud amount (\$28.107M) but a low fraud count (5) from region 6.

Drilldown from bin (x=20K-525.9K, y=1M-66.55M) to find which payment type has the most fraud.

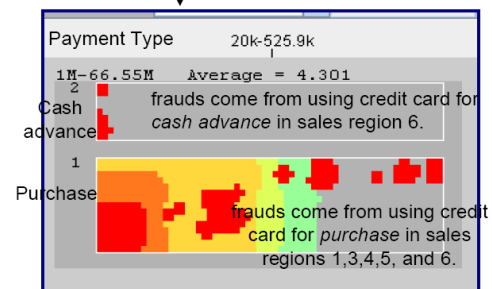


Figure 4B: shows that using credit card for purchase has more frauds than using credit card for cash advance (more fraud regions)

Figure 4: A Variable Binned Scatter Plots on Fraud Analysis

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 direction is determined according to the fraud amount and fraud count. The color of a data point represents the sales region (1-6) where the fraud occurred. Figure 4A shows that the fraud amount is almost increases linearly with fraud count. The fraud amount is highly impacted by the fraud count. However, there is an outlier at the top left bin (x=1-100, y=1M-66.55M) with a low fraud count but a very high fraud amount of \$ 28.107M. This exceptional credit payment might be a potential problem or error.

Data points that would be overlapped in traditional scatter plots are now represented as clusters distributed inside a bin. In order to answer the second 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 which from the same sales regions (colors) are placed together (the technique is described in Section 4) for analysts to see the fraud regional distribution. From Figure 4A, analysts observed that fraud comes from all the sales regions in each bin, except sales region 6. Sales region 6 only appears in the bins over \$1M. There are three large clusters (orange, yellow, and green) in bin (x=1-100, y=1-1K).

To find which fraud amount and payment types have the most fraud, analysts can first select a bin, such as bin (x=20K-525.9K, y=1M-66.55) and then query on the payment type as shown in Figure 4B. Analysts can learn that the most frauds came from using credit cards to *purchase* rather than to get *cash advances*. Purchase has larger clusters from different five sales regions (green to red) while cash advance has only two small size clusters from sales region 6 (red). Overall, analysts are able to observe that the largest cluster is formed by sales region 1 (green) with the lowest fraud amount and count. The smallest cluster is sales region 6 (red) but with the highest fraud amount and count. Using the above information, the company is able to place strict control on certain sales regions and payment types, such as sales region 6 and a purchase amount above \$1M.

4.2 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.

Examples of a data center service manager’s frequent concerns are:

- Q1. What is my data center daily temperature distribution? How do I optimize the cooling system performance?
- Q2. How is the chiller operating? Are there any problems?

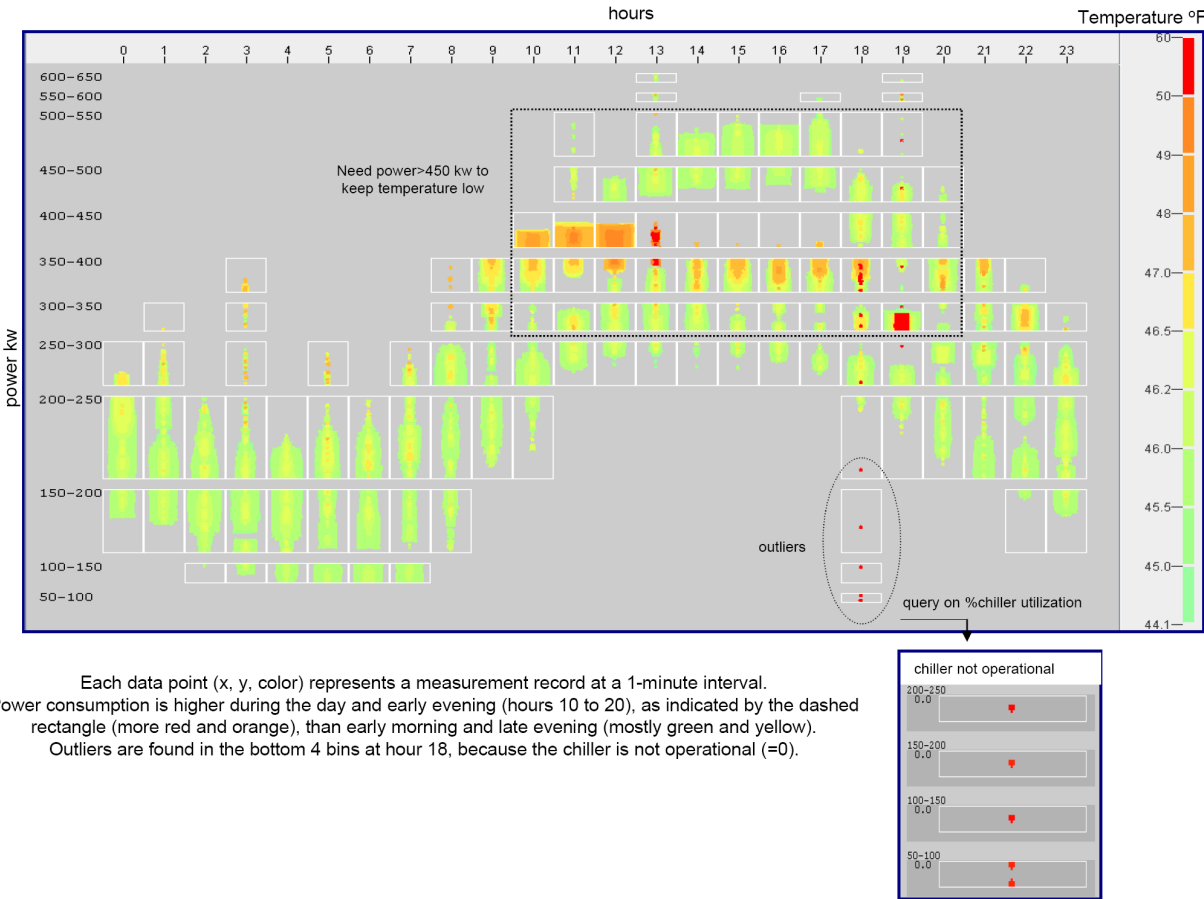


Figure 5: Hourly Temperature Visual Analysis Using Variable Binned Scatter Plot (x-axis: daily hours (0-23), y-axis: Power (KW), color: temperature)

Q3. What are the cause-effects of the chiller utilization and the power consumption on temperature?

To answer the above three questions, we have used variable binned scatter plots to enable data center administrators to visualize the temperature distribution, patterns, and outliers. To visualize %chiller utilization and power consumption impacts on temperatures, we map temperature as the third attribute in the variable binned scatter plot matrix in Figure 5. Each data point is represented by a pixel and defined with three attributes (x, y, color) where the x-axis represents the %chiller utilization and the y-axis represents the power in KW which runs the chiller. The color of the data point represents the temperature, from low (green) to medium (yellow, orange) to high (red). With variable binned scatter plots, the overlapping data points are sorted and placed close to their original locations as described in Section 4.

Figure 5 shows that the temperature is higher during the day and early evening (more orange and red), as indicated by the dashed square, than during the early morning and late evening (mostly green and yellow). This result helps the administrators to optimize cooling system performance. From this observation, administrators are able to use less power (under 300 KW) in the early morning and late evening and then gradually increase the power (above 300 KW) between 10 am to 8 pm. Especially during the peak hours 11 am to 1 pm, power could be increased greater than 450 KW for the chiller to cool down the temperature less than 46.5°F.

Five outliers appear in the bottom three bins with power under 250 KW at hour 18. Administrators can issue a visual intelligent query [7] to find the root-cause of the problem. The result of the query is shown in the bottom right window: The high temperatures are caused by the chiller not being operational at that time (i.e., %chiller utilization is 0.0%).

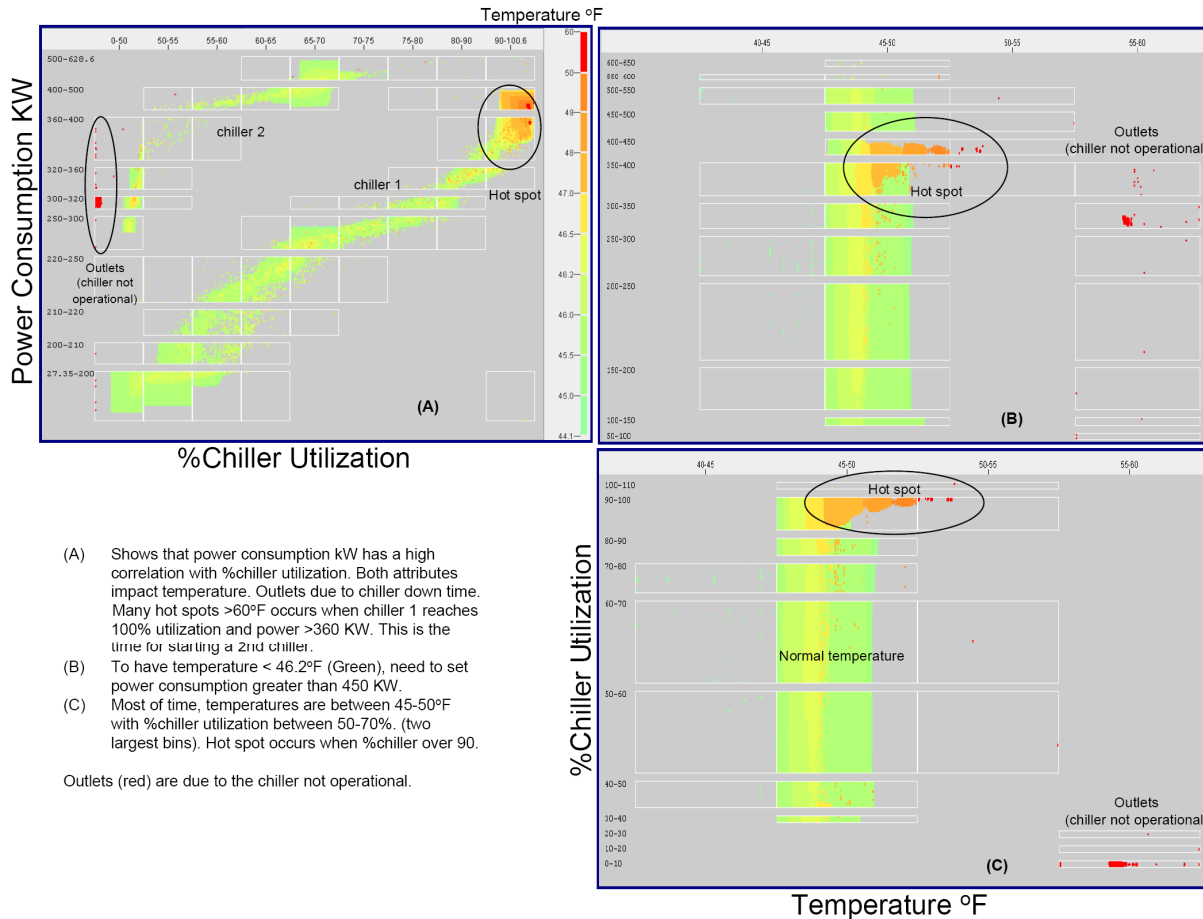


Figure 6: A Variable Binned Scatter Plot Matrix for Three Thermal Attributes. (Color represents the temperature of a data point.)

To answer the third question, we employ the variable binned scatter plot matrix to visualize the relationships and impact among three attributes: power, chiller, and temperature. Figure 6(A) shows that power consumption has a high correlation with %chiller utilization. Interestingly, administrators also notice that there are two chillers being used together when one of the chiller's utilization is over 100% as seen in the top right two bins (most yellow and orange). With the assistance of the second chiller and over 450 KW of power consumption, the temperatures are brought back to normal (green). From these facts, we can conclude that temperature is highly impacted by both the %chiller utilization and power consumption KW. From Figure 6(B) and 6(C), administrators can quickly learn how to optimize the power and chiller resources to reduce the data center cooling costs.

5. 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 7A-7D 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 7E and 7F. 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 clearly show fraud and thermal distribution in the high density areas, marked by the dashed rectangle (Figures 7E and 7F) than both hexbin scatter plot (Figures 7A and 7B) and the smooth contour scatter plot (Figure 7C and 7D). Smoothed contour scatter plots show linearly increasing overlaps with different shades which are better than hexbin scatter plots.

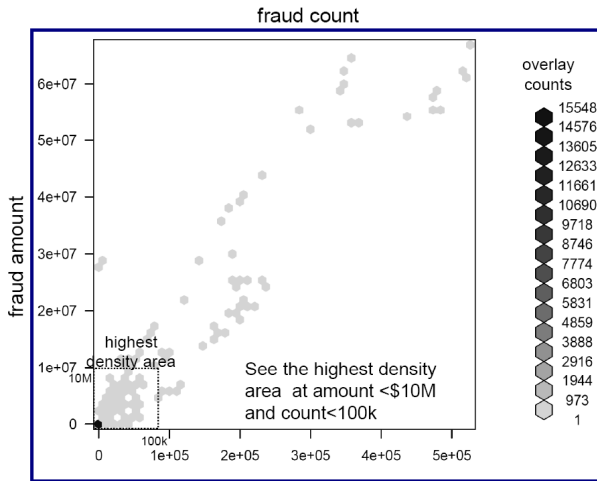
In most applications, the majority of data points occur in the high density areas. In order to analyze the data distribution, patterns, and outliers in those areas, both variants require an extra step of zooming into the high density areas (i.e., dashed rectangles). The variable binned scatter plots 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 4A (Section 4.1). 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 be visible 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.

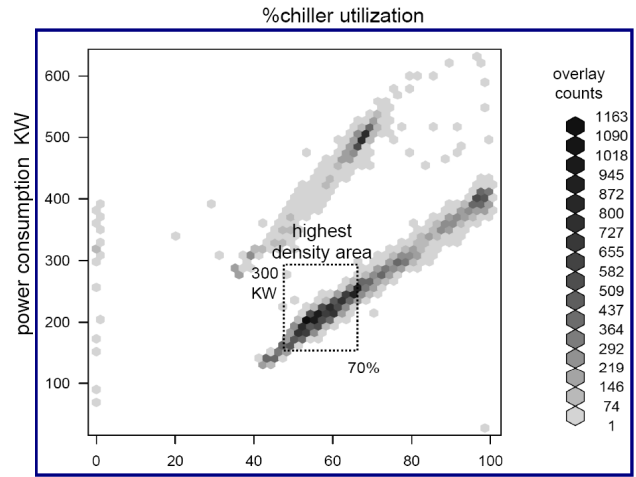
The weaknesses of the variable binned scatter plot include:

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.

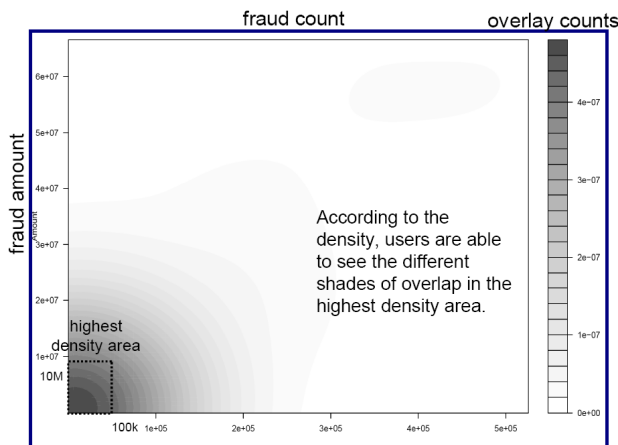
In summary, both hexbin and smooth contour scatter plots are able to provide a quick overview of data density and correlations. Variable binned scatter plots visualize the entire data distribution but also allow access to each individual data point for users to retrieve information at the record level. These three variants of scatter plots complement one another.



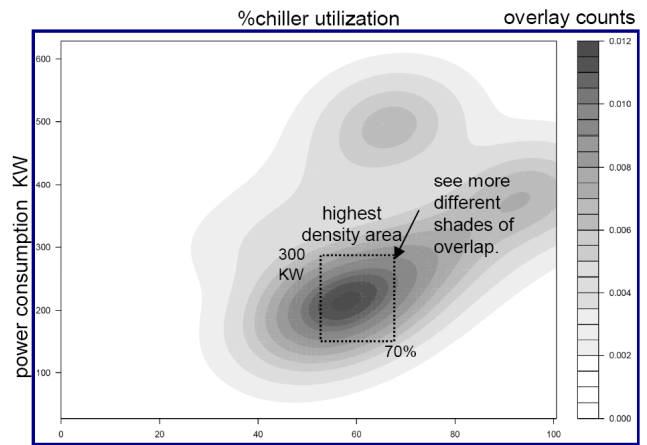
7A: Hexbin Scatter Plot: only see one data point with high overlaps in a dashed rectangle.



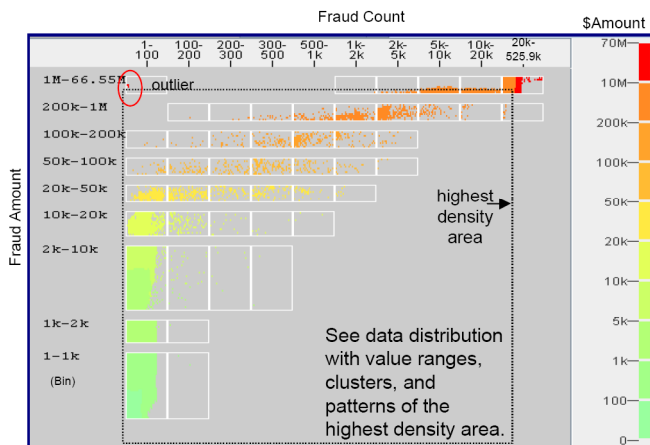
7B: Hexbin scatter plot shows high overlapping area inside a dashed rectangle.



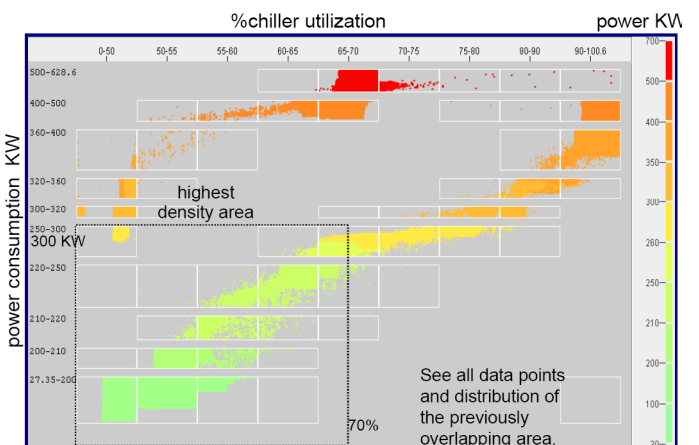
7C: Smooth contour scatter plot shows different degree of overlapping area in the high density area.



7D: Smooth contour scatter plot shows different density areas more clearly.



7E: Variable Binned Scatter Plot: shows all data points and their distributions with value ranges in the highest density area.



7F: Variable Binned Scatter Plot shows all data points and their distribution, correlations, and clusters in the highest density area.

Figure 7: Evaluation of Hexbin (7A, 7B) and Smoothed Contour (7C, 7D) Scatter Plots with Variable Binned Scatter Plots (7E, 7F)

6. CONCLUSION

In this paper, we introduce new variable binned scatter plots used for visual analysis of data distributions and cause-effect of large volumes of multi-dimensional data besides correlation between pairs of variables. Variable binned scatter plots are simple and easy to use. They allow users to visually analyze large transaction data sets at the record level. The number of bins in x and y is computed from the x and y density distribution and value range. A minimal distortion is introduced to provide space for the overlapping data points. Also, we are able to use color for a third attribute of the data to help analysts quickly identify patterns and clusters. An evaluation of the recent HexBin scatter plots demonstrates the benefits of using variable binned scatter plots to reveal distribution in a high density area. This technique reveals interesting distributions and patterns which otherwise would have been lost. Our future work will be in the area of recursive scatter plots which allow analysts to further analyze an important area in a scatter plot.

REFERENCES

- [1] Cleveland W. S. , The Many Faces of a Scatterplot. Robert McGill Journal of the American Statistical Association, Vol. 79, No. 388 (Dec. 1984), pp 807-822.
- [2] Unwin A., Martin T., Heike H, Graphics of Large Datasets, Springer, NY, 2006, pp 39-193.
- [3] Wilkinson, L. The grammar of Graphics, New York, Springer, 1999. Singh, Mala, MrExcel, Ohio, USA.
- [4] JMP 8 Software. www.jmp.com/software, new 64-bit computers and visual analytics tools.
- [5] Carr, D. B., Littlefield, R. J., Nicholson W. L., and Kuttkefuekdm J. S. Scatterplot Matrix Techniques for Large N, "Journal of American Statistics Association" 82, 424-436. 1987.
- [6] Bachthaler S. and Weiskopf D. Continuous Scatterplots, IEEE Transactions on Visualization and Computer Graphics, Vol. 14, No. 6, November/December 2008.
- [7] Hao, M., Dayal, U., Keim, D. A., and Morent, D., Intelligent Visual Analytics Queries. IEEE Symposium on Visual Analytics Science and Technology, pp. 91-98, 2007.
- [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 University Press, Oxford.
- [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] Keim, D. A., Kriegel, H. P., and Ankerst, M. Recursive pattern: A Technique for Visualizing Very Large Amounts of Data. Proc. IEEE Visualization, pp. 279-286, 1995.
- [12] Hao, M., Dayal, U., Method for visualizing graphical data sets having a non-uniform graphical density for display. US patent number 7,046,247, issued in May, 2006.