EG UK Theory and Practice of Computer Graphics (2008)
Ik Soo Lim, Wen Tang (Editors)

Interactive Grid Based Binning for Information Visualization

S.M. Longshaw, M.J. Turner and W.T. Hewitt

Research Computing Services, The University of Manchester, UK

Abstract

Clutter within information visualization (infovis) systems is an area of continuing concern and is receiving increasing research interest. Solutions to the problem vary in their approach, ranging from novel visualizations designed specifically to cope with high data density, through to statistical methodologies such as binning. This paper presents a flexible method that allows interactive placement of a grid based binning system that aims to enhance traditional information visualization techniques. User evaluations employing two specific visualization methods are described using a prototype grid based binning system. The method is shown to be a quick and easy way to visually segment a data domain, while the two visualization techniques presented are shown to provide effective data overview. Due to the abstracted nature of the binning grid, its applicability goes beyond the examples provided in this paper, therefore it could be considered as a generic data reduction and/or overview technique within many systems.

Categories and Subject Descriptors (according to ACM CCS): I.3.6 [Computer Graphics]: Interaction techniques I.4.9 [Image Processing and Computer Vision]: Applications

1. Introduction

Cluttered or dense datasets have always presented a set of challenging issues when developing visualization techniques. Almost all visualizations that directly represent data lose effectiveness after a certain level of complexity has been reached. There can often be a distinct and quantifiable knee-point at which this occurs. Solutions to this issue vary in nature, ranging from data reduction techniques such as Independant/Principle Component Analysis [HKO01, GKWZ07] and Multidimensional Scaling [BG05] through to visualization techniques specifically designed to represent complex datasets [CMR07, ET07, CA04].

One answer is to reduce the size of the dataset by removing items deemed extraneous. However performing this successfully can be a difficult task, especially in a dataset with complex patterns, for example a very dense dataset with similar values all clustered together. Solutions range from manual user driven techniques, through to fully automated algorithms intended to identify and classify data based on initial criteria [HMS01, WF05]. Alternatively the concept of binning can be applied, collecting individual data items into larger classes. Visualizing binned data can lead to insight into the trends and patterns that exist within the dataset, at the expense of detail.

While binning may primarily be viewed as a statistical methodology, recent attempts at integrating binning directly into traditional information visualization (infovis) systems [HLD02] have been made, producing representations that either augment or replace the initial visualization with the visualization of the binned data values.

This paper introduces a system of interactive grid based binning, which has the scope for use in any coordinate based information visualization. It shows how bins can be visually defined in the form of an interactive overlaid grid, on top of classic information visualizations, such as the scatter plot. The cells of the grid are used as defined data boundaries within which simple numerical processing techniques are applied, resulting in a new binned dataset. This data is overlaid on top of the original visualization, ensuring that it is in keeping with the confines of the initial parameters. Currently two numerical methodologies have been explored, however the gridding concept allows for alternative numerical techniques to be applied without alteration to the basic method.

In order to provide proof of concept, a software prototype named Multi View Graphing (MVG) [Lon08], has been developed. MVG is used in this paper to demonstrate the meth-

ods being described through two test cases, using sample data.

The remaining sections present a brief summary of related techniques and then describe and evaluate two clutter reduction visualization methods for this statistical data. The two methods are a Binned Bubble Plot and a Binned Density Map, both of which are designed to characterise the data using an interative grid.

The evaluations and illustrations presented come from real research groups and include results from user-studies with medical electrical lumbar anterior root stimulation data (figures 1-4), an information theory dataset consisting of legal Scrabble words (figure 6) and a life sciences biochemical dataset (figures 5,7-8). This application specific analysis with complete questionnaire user-study results are available on the applications wiki site [Lon08] as well as published in a related longer document [Lon07]. Further user analysis is also available from these references that includes their use within MRI data (supplied by the Wolfson Molecular Imaging Centre) and climate prediction data (supplied by the Hadley Centre for Climate Prediction and Research).

2. Related Work

The basic concept of binning data, by collating points into larger groups, is widely accepted as an informative and useful information visualization tool. Implementations of binning vary widely in not only design but also purpose. When used in image compression for example, the purpose of binning is to reduce the amount of data being stored, while still producing an image which is a good approximation of the original. Binning can also be found in some data analysis tools where its purpose can vary from data density reduction through to methods for in-depth statistical analysis [War94].

The data obtained from applying binning varies based upon how the data is collated; this can range from simply counting how many data points fall within each binning domain through to applying more complex techniques such as Principle Component Analysis [Jol02] of multidimensional data items within each bin. The final result however is usually a single representation for each group.

Perhaps the most common form of binning is that of simple collation, whereby numerical boundaries are defined and all data that falls within those boundaries becomes a member of that binning class. Classically this type of binned data will usually then be represented by an appropriate visualization such as a Histogram. As the binned values are conceptually seperate from the original data, a level of abstraction exists between the two, that when bridged, may provide a more insightful binning system.

One area that has received little attention is the definition of the bins themselves. Systems that offer the ability to apply binning either do so through complex control panels or as automated systems with simple input parameters [The02, WR07]. The advantages offered through visual interaction and feedback are, on the whole, unexplored. Perhaps due to its intuitive nature, it is possible to find examples of a simple grid being used to define binning areas [GBLM05]; however the grid is rarely referred to in any great detail. Indeed the grid often appears to be located after the binning itself has taken place, in order to highlight how the bins were initially defined, rather than the other way round. This methodology appears to be especially prevalent within the discipline of Astrophysics [GBLM05, BGL*04].

3. The Binning Grid

Information visualization techniques such as the scatter plot are often designed around a Cartesian coordinate scheme. Simple segmentation of such a space naturally leads to a uniform grid structure.

To show the binning grid concept, an application has been produced that allows multivariate datasets to be visualized using some of the more common information visualization techniques. For this paper we have highlighted scatter plots in both 2 and 3 dimensions as well as Parallel Coordinates. The application also allows for interactive data brushing within each of these visualizations, that can be used to augment the grid based binning.

Initially the grid is generated according to the width and height of the dataset being viewed, with the number of segmentations in the X and Y directions defined by the user. The size and placement of each bin is therefore dependant on the granularity of the binning desired, an example of this can be seen in figure 1. When applying this to the Parallel Coordinates visualization, the lack of a derived X component means that only a single line of bins is generated along each axis.

In providing the grid as an overlay to the original data, rather than an abstracted concept, it becomes easier to see what level of granularity is suitable for the current data being viewed, as well as providing an obvious visual cue representing the bins themselves. In a more traditional system of bin definition, this can often be difficult to gauge, resulting in trial and error.

One issue raised by considering the granularity different grid sizes offer, was if the dataset itself can actually define the minimum size each bin should be set too. Presuming the grid has not been rotated or translated then the answer for an integer dataset is almost certainly yes. For example if the smallest value for a variable within an integer dataset is 1 and the largest 10 and we map the variable to the X axis of a 2D scatter plot, there is nothing to be gained by defining a grid with more than 10 bins in the X direction. If rotation or translation is applied to the grid then this is no longer the case as the user is able to affect at which value of X the first

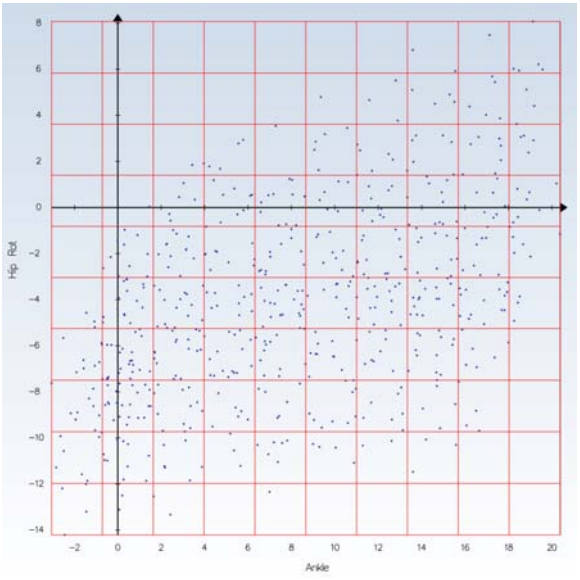

Figure 1: *2D scatter plot with a 10 x 10 binning grid overlay, defined in red.*

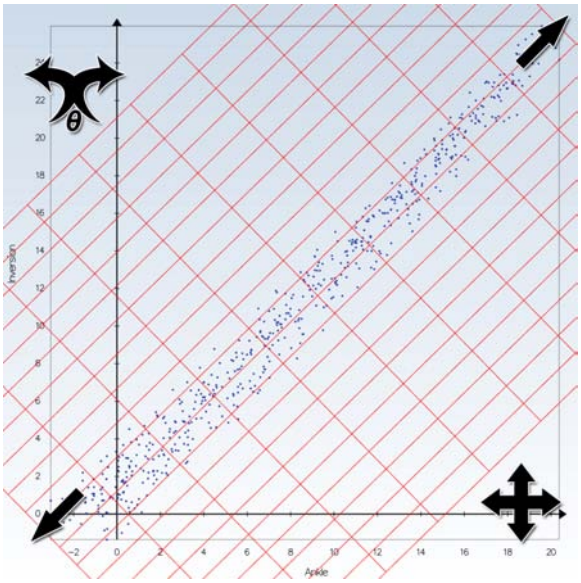

Figure 2: *2D scatter plot with a 10 x 21 binning grid overlay, defined in red. The grid has been rotated by roughly 45 degrees to the left (-X), translated down and to the right and finally its scale has been increased in the X direction.*

cell begins. In the case of non integer data the answer is less obvious, the ability to define the lowest size a bin should be, would require the smallest distance between any two points within the dataset to be found in each applicable dimension. Again user rotation or translation of the grid may render this value incorrect.

The amount of data that is contained within each bin is dependent on the visualization style it is being applied to. If it is being applied to a 2D scatter plot, each data point will be represented by two values, while in a 3D scatter plot each will be representative of three values. In an environment where complex Glyphs are used to visually represent many data dimensions, each binned item will provide many values. How these values are collated is an area for exploration, this paper presents two possible solutions but other methodologies exist.

3.1. Grid Interaction

Once the basic binning grid and level of granularity have been interactively defined, it may then be necessary to alter the orientation, placement and scale of the grid. For example, a grid based layout clamped to the original coordinates of a scatter plot is unlikely to be optimal if the dataset produces a compact slope of points or a region of interest that is not ideally aligned to the initial grid.

In order to introduce flexibility to the binning grid system, it is possible to interactively rotate and translate the grid in real-time on top of the original visualization. The system uses simple mouse interaction to allow the user to place the grid at the orientation they feel will produce the best binning areas for their dataset. The downside to this method is that the placement is not as accurate as it may be if it were performed numerically, however it is proposed that the ease of the system outweighs any minor accuracy issues this may introduce. Figure 2 demonstrates the range of movement applicable to the binning grid within a 2D scatter plot. This range of movement could be extended and applied to a grid within any other coordinate based visualization, including those in 3D.

It is worth noting at this point that datasets which are dispersed evenly across their whole data range may in fact be best suited to a binning grid orientated and located according to the initial coordinate system. The ability to define and quantify the effect that the placement of the grid will have before it is used to bin, is perhaps an area worthy of further study.

4. Bubble Plot and Density Map Visualizations

While previous sections of this paper have been devoted to the specifics of actually defining the binning grid, the question of how to use the geometrical separation that the grid defines is more open ended. Two methods are presented here, both based around widely used concepts and both suited specifically to the geometry generated by a grid.

The first visual style presents each bin of data as a single representative disc or sphere. Each object is placed within

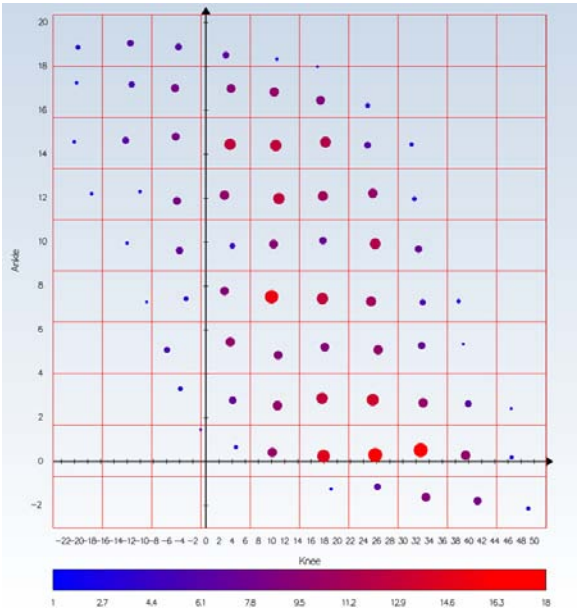

Figure 3: *A Binned Bubble Plot within the confines of a 2D scatter plot. The 10 x 10 binning grid utilised has been left in place for illustrative purposes.*

the layout of the original visualization according to the arithmetic mean of the coordinates of the binned data and then scaled and coloured according to the number of data items within that bin.

The second method counts how many values exist in each bin and assigns a greyscale value accordingly. Each bin within the grid is then filled with this generated colour.

4.1. Binned Bubble Plot

In order to utilise all of the data encoded into each point on either a 2D or 3D scatter plot, a simple Glyph based visualization was required. Each bin is used to generate a set of coordinates and a value for the number of points contained within its confines. This value is then mapped to the radius of a disc or sphere, which is placed at the binned coordinates. This would mean that a bin within a 2D scatter plot, containing 2 data items with coordinates at (1, 2) and (3, 6), would produce a disc at (2, 4) with a radius and colour generated from a count value of 2. An example of a Binned Bubble Plot within a 2D scatter plot can be seen in figure 3.

This form of visualization has proven to be especially successful when applying the binning grid to cluttered integer datasets. This is due to the fact that while we are able to replace areas of high data density with a bubble representation, we are also able to retain the initial integer positioning of each value. A more in depth study is presented in section 5.1.

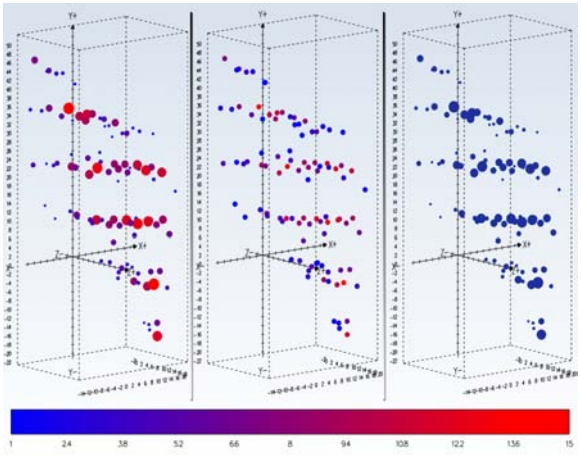

Figure 4: *A Binned Bubble Plot within the confines of a 3D scatter plot. The left image shows the initial visualization, the middle shows the same but with the bubbles radii homogenised and the right hand image shows the bubbles colour map homogenised.*

The positional nature of the Binned Bubble Plot also transfers well into three dimensions. When applied to a cluttered 3D scatter plot, the resultant visualization quickly and clearly defines a reasonably accurate overview of the initial data. However it becomes clear that if either the colour map is homogenised across all of the spheres, or if all of the spheres are rendered at the same radii, it is more difficult to extract information during user evaluation. This effect can be seen in figure 4, where the same Binned Bubble Plot is shown in its initial state and then with the two distinguishing effects disabled. We must therefore conclude that the overall success of the Binned Bubble Plot is due to the combination of both the colour mapping and variable sizing of each bubble.

4.2. Binned Density Map

To provide an alternative and simpler use of the binning grid than found in the Binned Bubble Plot, a mapping between the amount of points in each cell and a grayscale value between pure white and black is generated. The geometry provided by the binning grid is then used as a bounding area to be filled. It is not a requirement of this method that the mapping of each grayscale value is to the number of points within the bin. Indeed there are examples of visualizations similar to this being used in a more complex manner [GBLM05, BGL*04], however when used with this easily comprehended value the Binned Density Map was found to be a powerful visualization.

As only the amount of points in each cell is being utilised in the method, the density map can also be applied to visualizations with only one derived dimension such as Parallel Coordinates.

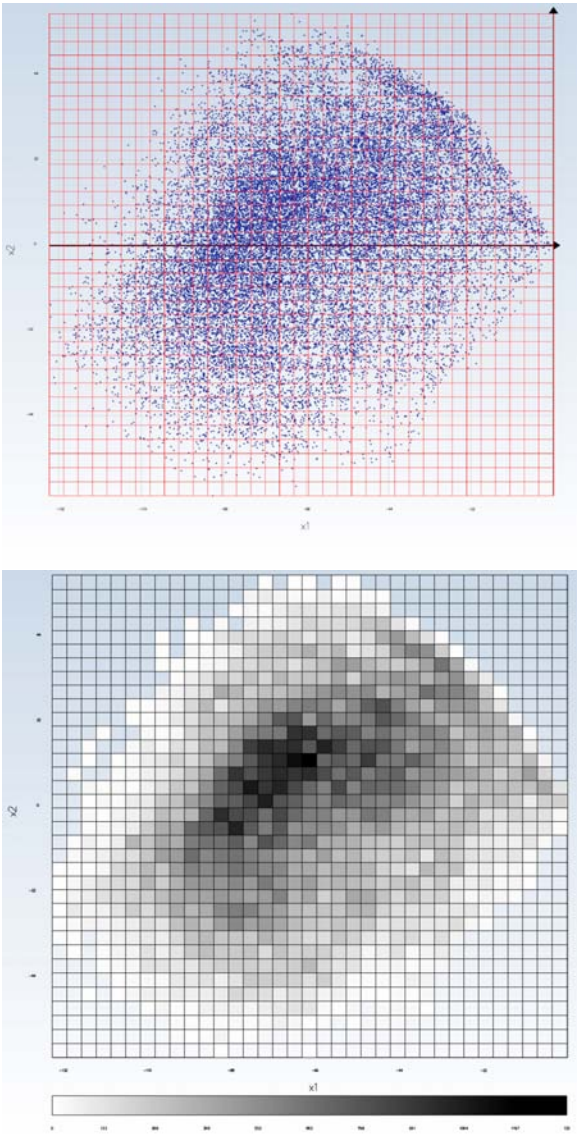

Figure 5: *The top image shows an initial dense and cluttered 2D scatter plot with a 35 x 35 binning grid overlay, the bottom image shows the Binned Density Map Visualization that resulted.*

The binned density map has proven to be most useful when used with dense, non integer datasets, where a traditional visualization problem of a physical lack of available pixels can become most evident. This is usually caused due to a dataset containing values that are close to each other, thus visually each point lies within close proximity of at least one other; introducing ambiguity as to which points are being represented by which pixels. Indeed it is not uncommon for coordinate based information visualizations such as a scatter plot or Parallel Coordinates to end up looking like a solid geometric shape rather than a collection of discrete objects.

Due to the fact the geometry provided by the binning grid is directly utilised in the Binned Density Map visualization, the ability to rotate, translate and resize the grid before binning also results in the visualization following these changes. In datasets that do not optimally follow a traditionally orientated grid, this can result in each bin being better aligned with the data, effectively reducing the aliasing effect seen when the visualization is presented using the initial Cartesian coordinates of the scatter plot.

5. Application and Evaluation of the Method

In order to present the methods described in this paper in a practical context and to show their merits, two different datasets are presented in this section of the paper. In section 6.1 a purely integer dataset is examined using the Binned Bubble Plot methodology, followed by a dense and highly cluttered Life Sciences dataset in section 6.2.

In depth real world studies, beyond those presented in this paper, have also been carried out using the MVG prototype [Lon07]. While these studies examine techniques other than just the binning grid concept, they show that its inclusion into a rounded infovis system is beneficial. Primarily the studies that benefited most from the binning grid were cluttered or dense datasets. Studies also show that by allowing advanced brushing to dictate which data the binning is performed on, an extra level of flexibility is introduced, which participating users found useful.

5.1. Integer Datasets

The following study utilises a dataset which has been generated by mapping each letter from an amalgamation of the words from the official 2006 US and Canada (TWL) and 2007 International (SOWPODS) 8 letter Scrabble(©Mattel Inc.) lexicons [Che07] to an appropriate integer value. Thus ‘a’ becomes ‘0’, ‘b’ becomes ‘1’ through to ‘z’ becoming ‘25’. The result is an 8 dimensional dataset containing 29766 rows.

If we wish to perform a simple analysis of this dataset in order to find the most common pair of letters that start each word, it can be achieved with relative ease by using a 2D scatter plot combined with a Binned Bubble Plot. As is often the case with integer datasets, we are initially presented with a neat but uninformative scatter plot, as can be seen in the top portion of figure 6. While we can see where letters lie in the dataset, we are unable to see how often this occurs at each location. The solution is to define a 26 x 26 binning grid (thus ensuring 1:1 granularity with the original dataset) and generating a Binned Bubble Plot, as can be seen in the bottom portion of figure 6.

We can now determine that for eight letter words the most

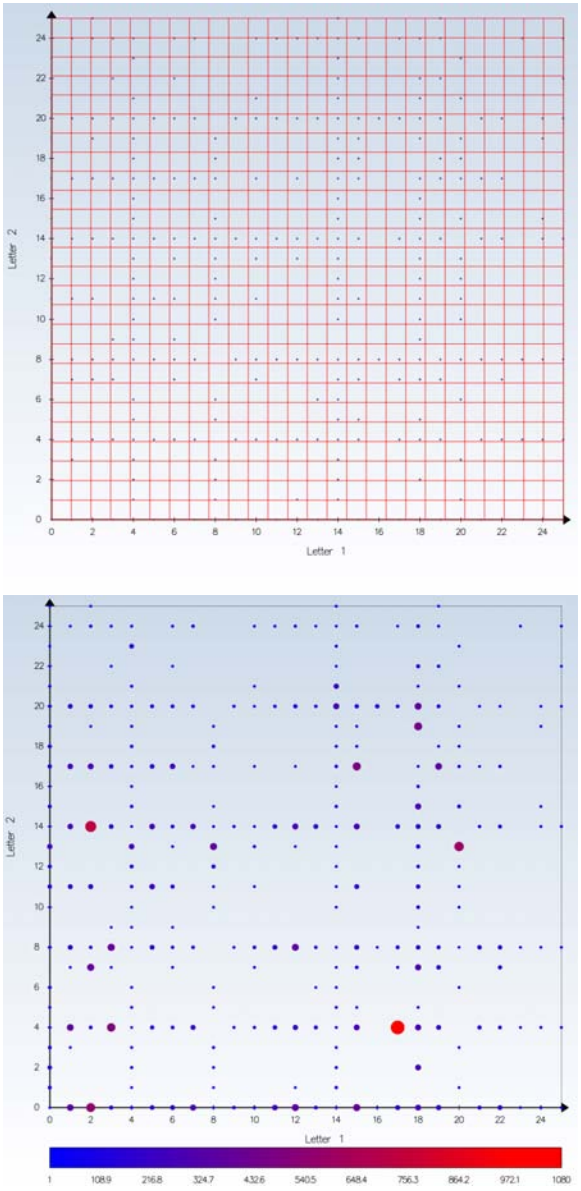

Figure 6: *The top image shows the initial state of an integer 2D scatter plot showing the first two characters of the Scrabble dataset. Although we can distinguish the individual values that exist, it is not clear how many of each pair there are. The bottom image is a Binned Bubble Plot with a bin:data granularity of 1:1, achieved using a 26 x 26 binning grid.*

common first two letters are 'R E' (17, 4). If we then wish to move into three dimensions and see which three letters are most likely to start an eight letter word in the scrabble dictionaries, the same technique can be utilised to great success. This time we would utilise a 26 x 26 x 26 binning grid

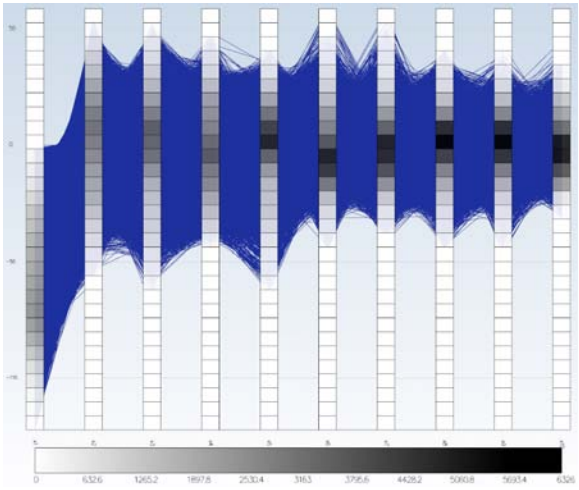

Figure 7: *A non normalised Parallel Coordinates visualization, showing the first 10 variables from the sample processed Life Sciences dataset. The initial presentation of the blue lines is too cluttered and dense to be of real use. The Density Map overlaid on each axis provides a quick and clear insight into the number of points that lie in each area.*

within a 3D scatter plot. Then as with most 3D visualizations, through user interaction such as rotation of the visualization, we could determine the locations of bubbles of interest. Using these tools along with brushing, information theorists have been able to visually discover unusual probability distributions.

5.2. High Density Datasets

To demonstrate the success of the Binned Density Map when used with highly cluttered non integer datasets, this case study utilises a dataset [Kom08] generated by running Principle Component Analysis on a large life sciences dataset and then saving the ten topmost results. It contains 10 dimensions and 26733 data rows. These values have all been scaled by a factor of 10 for visualization purposes and its final dimension ignored due to the fact it only contains binary values.

If we were to generate a non normalised Parallel Coordinates visualization of the whole dataset, we are presented by the mass seen in figure 7. Although a very general overview of the trends within each variable can be seen, it is difficult to extract any real detail into how the data lies. However, also in figure 7, a binning grid of 30 bins per variable has been applied creating an overlaid Binned Density Map. Through this we are able to start to determine where the densest region of points lie. The more bins we introduce, the finer the detail we will be presented with. In figure 7 we can see that the densest area lies in the eighth variable, roughly between 0 and 6 and that 6326 points lie within this bin.

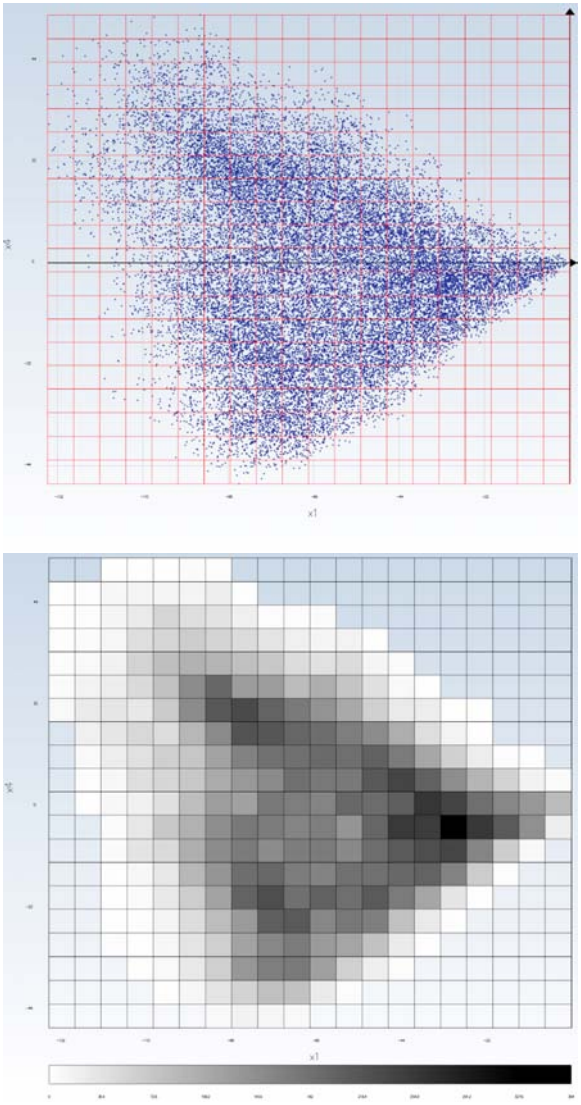

Figure 8: *The top image shows the initial state of a cluttered and complex 2D scatter plot. The bottom image shows the Density Map that has resulted from applying a 20 x 20 binning grid.*

As is common practice following an examination using Parallel Coordinates we then examine two specific data dimensions within a 2D scatter plot, in this case the first and fourth variables from the life sciences dataset. As is depicted in the top portion of figure 8, we are initially presented with a dense cloud of pixels, while we can roughly see areas of higher point density, it is difficult to determine specific detail. If we then apply a 20 x 20 binning grid to the same visualization, we are presented with the image seen in the bottom portion of figure 8. We are now able to determine that the densest area of points roughly lies between [(-34, -7), (-27, -2)] and that 364 points fall within this catchment area. Further analysis is then possible to discover new correlations or trends. This can be achieved by either extracting the area of interest and generating new normalised scatter plots, or by increasing the granularity of the binning grid and generating new Binned Density Maps.

6. Conclusions

This paper presents a method to interactively define a grid for binning data, within traditional information visualization techniques. Through visual representation and user interaction, the concept of overlaying a grid on top of coordinate based visualizations has proven to be a useful and powerful method of clutter reduction and data overview. The application of the method to three separate styles of information visualization has been explored and shown to be largely successful.

The ability to physically rotate and move the grid from its initial position is also demonstrated in this paper. This ability is important when the binning technique and resultant visualization being used makes direct use of the binning grid itself. Indeed, by altering the orientation of the grid it was possible to decrease the effect of aliasing seen when the majority of the data passed through the cells at an angle of around 45 degrees and the Binned Density Map visualization was utilised.

Two possible uses of the geometry provided by the binning grid were also presented by this paper. Both utilise fairly simplistic statistical methodologies to collate the data contained in each bin. This is then utilised to generate appropriate visualizations. The Binned Density Map has been shown to be a useful and appropriate way to visualize the binning of dense, non integer datasets. While the Binned Bubble Plot is well suited to integer datasets, and those with many overlapping values.

7. Future Work

Many different avenues of exploration exist beyond this initial presentation of the binning grid method. The most pressing is the exploration of a non uniform grid. This could mean a uniform grid that is then deformed by the user or a grid defined interactively cell by cell. The latter method would provide the ability to introduce discontinuities into the grid structure.

Geometric shapes other than the rectangle and cuboid can also provide a useful binning area definition and should be explored. Shapes that pack exactly such as triangles or hexagons could still provide a continuous grid while packing more bins into the same space.

The ability to translate, rotate and scale the binning grid has mainly been considered in two dimensions within this paper; movement within three dimensional visualizations

could provide some interesting binning opportunities and should be explored. Similarly, other methods of deformation such as skewing the grid or applying a fish-eye lens distortion may prove useful. Applying a lens effect to the grid may provide a magnifying effect to that specific area. It should also be possible to automatically define the best placement and size for a grid dependant on the current visualization and underlying data.

Finally, as this paper has presented a method by which to define bins themselves, there are many possible ways in which the bin definitions can be utilised to apply current or new binning techniques. This paper has touched on simple collation in the form of the Binned Density Map and also a more complex method in the form of the Binned Bubble Plot, however the possibilities for applying current methodologies such as Principle Component Analysis or even new methods designed specifically around the gridding concept surely exist. Different methods of binning the data should also lead to new and interesting visualizations. One of the methods presented within this paper (Binned Density Map) utilises the binning grid itself to visualize the results of the binned data, the usage however is simplistic in nature and further study may reveal more complex methods of integrating the grid into the resultant visualization.

8. Acknowledgements

The authors would like to thank the Hadley Centre for Climate Prediction and Research, Peter Hoornaert, Louise Lever, Samantha Mills and all those who offered their time, constructive feedback and data.

References

[BG05] BORG I., GROENEN P.: *Modern Multidimensional Scaling: Theory and Applications (2nd Edition)*. Springer-Verlag New York Inc., 2005.

[BGL*04] BASTIAN N., GIELES M., LAMERS H., GRIJS R., SCHEEPMAKER R.: The star cluster population of m51: Ii. age distribution and relations among the derived parameters, 2004.

[CA04] CARPENDALE S., AGARAWALA A.: Phyllotrees: Harnessing nature's phyllotactic patterns for tree layout. In *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization* (Washington, DC, USA, 2004), IEEE Computer Society, p. 215.3.

[Che07] CHEW J.: Scrabble lexicon (common to both twl and sowpods). http://www.math.toronto.edu/~jjchew/scrabble/lists, 2007.

[CMR07] CAAT M., MAURITS N., ROERDINK J.: Functional Unit Maps for Data-Driven Visualization of High-Density EEG Coherence. In Museth et al. [MMY07], pp. 259–266.

[ET07] ELMQVIST N., TSIGAS P.: TrustNeighborhoods: Visualizing Trust in Distributed File Sharing Systems. In Museth et al. [MMY07], pp. 107–114.

[GBLM05] GIELES M., BASTIAN N., LAMERS H., MOUT J.: The star cluster population of m51: Iii. cluster disruption and formation history, 2005.

[GKWZ07] GORBAN A., KEGL B., WUNSCH D., ZINOVYEV A.: *Principal Manifolds for Data Visualization and Dimension Reduction*. Springer-Verlag Berlin and Heidelberg GmbH and Co. K, 2007.

[HKO01] HYVARINEN A., KARHUNEN J., OJA E.: *Independent Component Analysis*. John Wiley and Sons Inc, 2001.

[HLD02] HAUSER H., LEDERMANN F., DOLEISCH H.: Angular brushing of extended parallel coordinates. In *INFOVIS '02: Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)* (Washington, DC, USA, 2002), IEEE Computer Society, p. 127.

[HMS01] HAND D., MANNILA H., SMYTH P.: *Principles of Data Mining (Adaptive Computation and Machine Learning)*. MIT Press, 2001.

[Jol02] JOLLIFFE I.: *Principal Component Analysis (2nd Edition)*. Springer-Verlag New York Inc., 2002.

[Kom08] KOMAREK P.: Dataset: 'ds1.10'. http://komarix.org/ac/ds, 2008.

[Lon07] LONGSHAW S.: *Multi View Graphing: Synchronous Linked Multi Visualization utilising Brushing, Binning and Clustering*. Master's thesis, The University of Manchester, UK, September 2007.

[Lon08] LONGSHAW S.: Multi view graphing. http://kato.mvc.mcc.ac.uk/rss-wiki/Multi_View_Graphing_%28MVG%29, 2008.

[MMY07] MUSETH K., MÖLLER T., YNNERMAN A. (Eds.):. *Eurographics/ IEEE-VGTC Symposium on Visualization* (Norrköping, Sweden, 2007), Eurographics Association.

[The02] THEUS M.: Interactive data visualization using Mondrian. *Journal of Statistical Software 7*, 11 (11 2002).

[War94] WARD M.: XmdvTool: integrating multiple methods for visualizing multivariate data. In *VIS '94: Proceedings of the conference on Visualization '94* (Los Alamitos, CA, USA, 1994), IEEE Computer Society Press, pp. 326–333.

[WF05] WITTEN H., FRANK E.: *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann, 2005.

[WR07] WOLFRAM RESEARCH I.: Mathematica edition: Version 6.0, 2007.