Graph Builder Contour Plots in JMP® 15

2019-US-30MP-247

Daniel R. Schikore

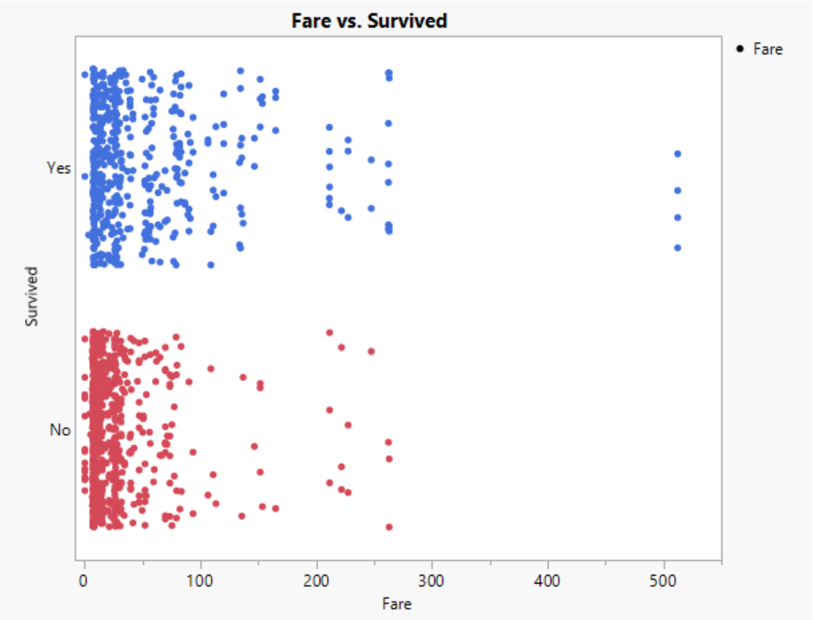
Principal Software Developer, JMP

Introduction

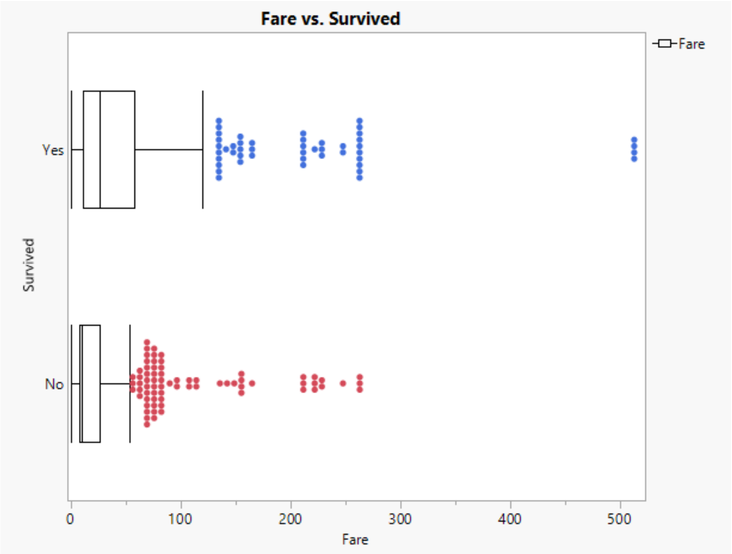
Contour plots are a common visualization technique for summarizing the shape of a dense collection of data. Contours in 2D are curves along which a continuous function has a constant value. Domain-specific terminology for contours is often based on the underlying function – isobars, isotherms, isopleths, and many more. In JMP, contours are used in several platforms, including Bivariate, Contour Plot, Contour Profiler, Graph Builder, and others. This paper will concentrate on the six different forms of contours that are available in Graph Builder, including two types of 1D density contours, three types of 2D density contours, and triangulations for 2D value contours. The strengths and weaknesses of each contour visualization techniques will be discussed, as well as the various options that are supported for the different types, including smoothing parameters, outliers, and alpha-shapes for non-convex domains.

Contours in 1D

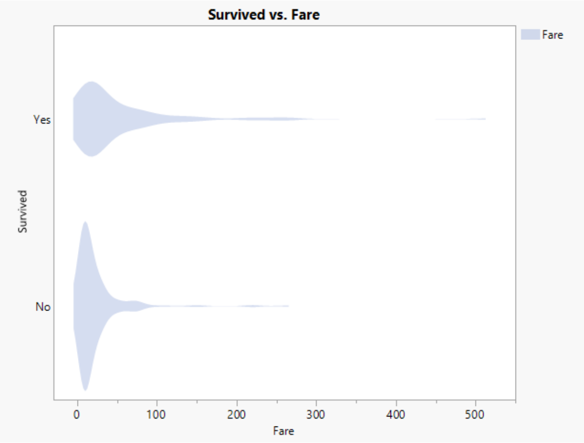
The scatterplot is a direct way to visualize continuous data and is often the default way to get a visual representation of continuous data. When the points are very dense, it becomes difficult to characterize the distribution of the points, and especially to make comparisons between multiple groups. In this example using the Titanic sample data in JMP, the fares paid by the survivors of the tragedy are compared to the fares paid by those that perished. Some observations can be made from the scatterplots, such as that the outliers in the survivor group that paid roughly twice what the others (in either group) paid.

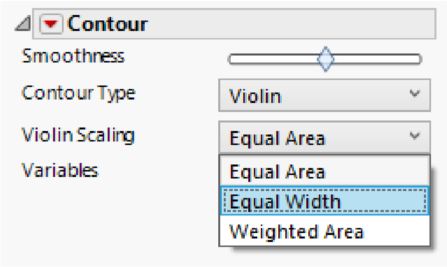


To compare these two groups further, some additional analysis is necessary to characterize the groups. One widely used technique for summarizing 1D data is the boxplot, shown here. The boxplot will be useful for comparison with the new contour techniques. There are several variations on the visual representation for a boxplot, but they are all based on quartiles. The box extends from the first quartile to the third quartile, so it contains 50% of the expected observations. A line is drawn at the median, giving additional detail. There are several variations on how to draw the whiskers – in this example they extend to the outermost point that lies within 1.5x the interquartile range from the box.



The Graph Builder Contour element uses a probability density function for a given kernel bandwidth to illustrate the shape of a 1D point distribution. The default view for 1D contours is the Violin plot, which plots the density function with symmetry about the 1D axis. The magnitude of the curves is determined by a scaling option, with options for Equal Area, Equal Width, or Weighted Area.




Contour

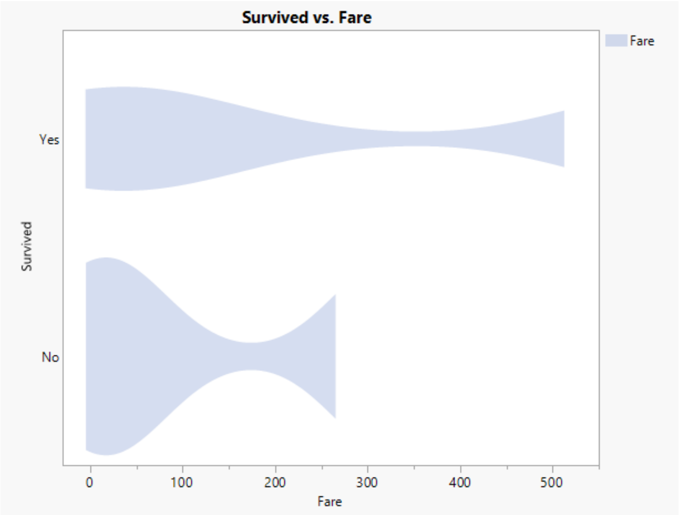
Smoothness

Contour Type: Violin

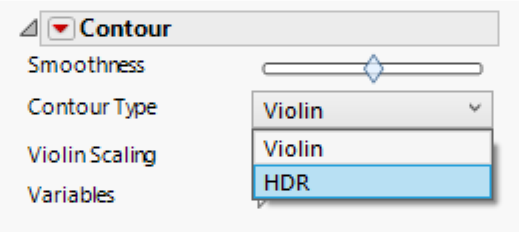
Violin Scaling: Equal Area

Variables

New in JMP 15 is the ability to control the kernel bandwidth via a smoothness parameter. The parameter is normalized, with the default value of 0 leading to a default computed bandwidth based on the data characteristics. Smoothness values in [-1,0) lead to less smoothing, and smoothing values in (0,1] result in more smoothing, within computed limits.



Also new in JMP 15 is a the HDR (Highest Density Region) plot, a new option for 1D contours. The HDR plot uses the same density function as the violin but applies concepts similar to the box plot to reduce the display to a few quantitative measures. A black line is drawn at the highest point in the density function, and dark rectangles are drawn to represent the 50% highest cumulative probability. Lighter rectangles are drawn to represent the 99% cumulative probability, and any points outside of the 99% probability are considered outliers. The same smoothness parameter is available for the HDR plot and Violin plot.



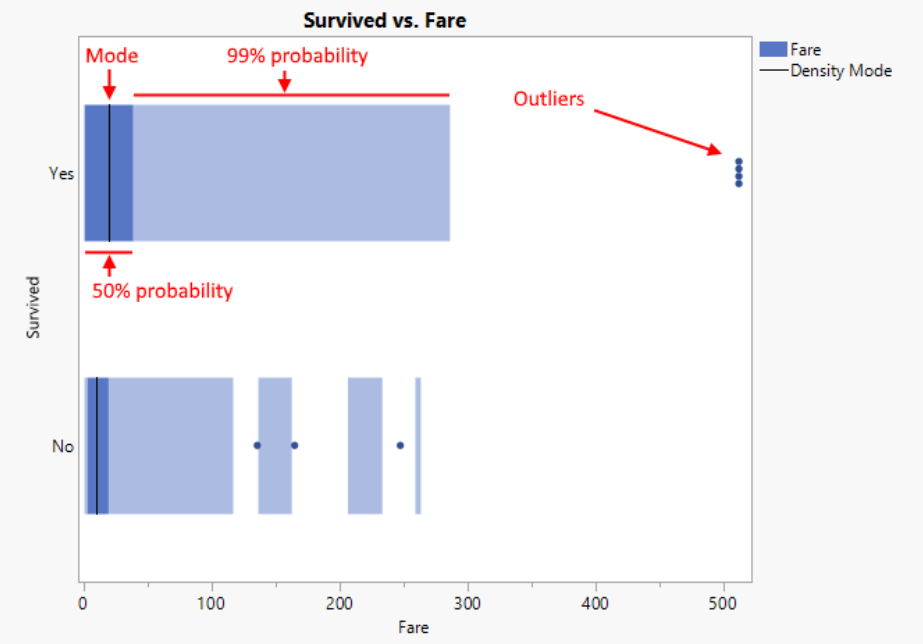
Contour

Smoothness

Contour Type: Violin

Violin Scaling

Variables



Survived vs. Fare

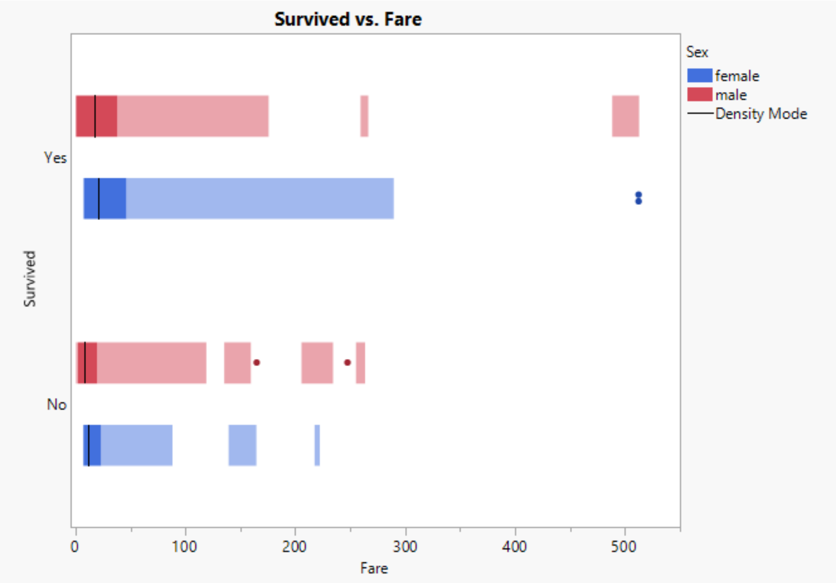
Mode

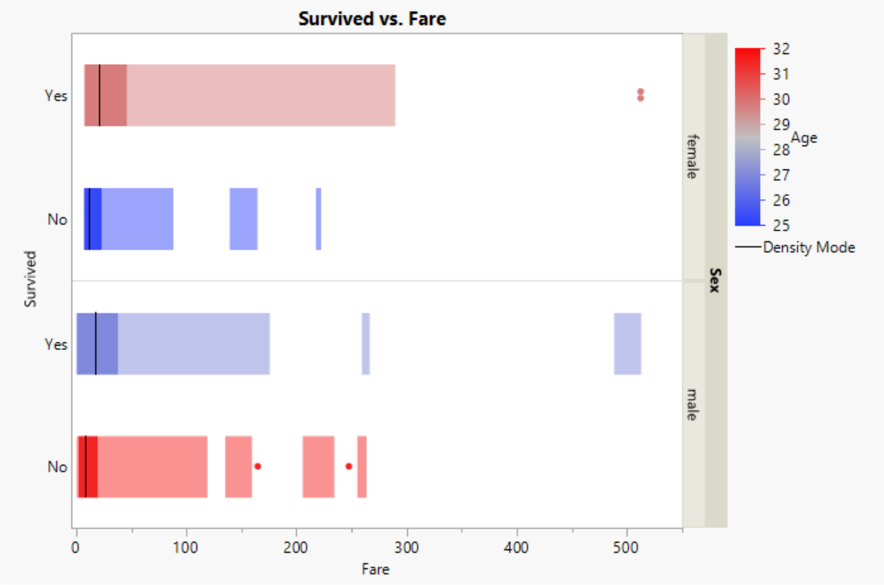
99% probability

Outliers

50% probability

The fill color or line color can be change from the legend, and colors are also used when adding columns to the Overlay or Color roles. The X Grouping and Y Grouping roles provide additional layout possibilities.

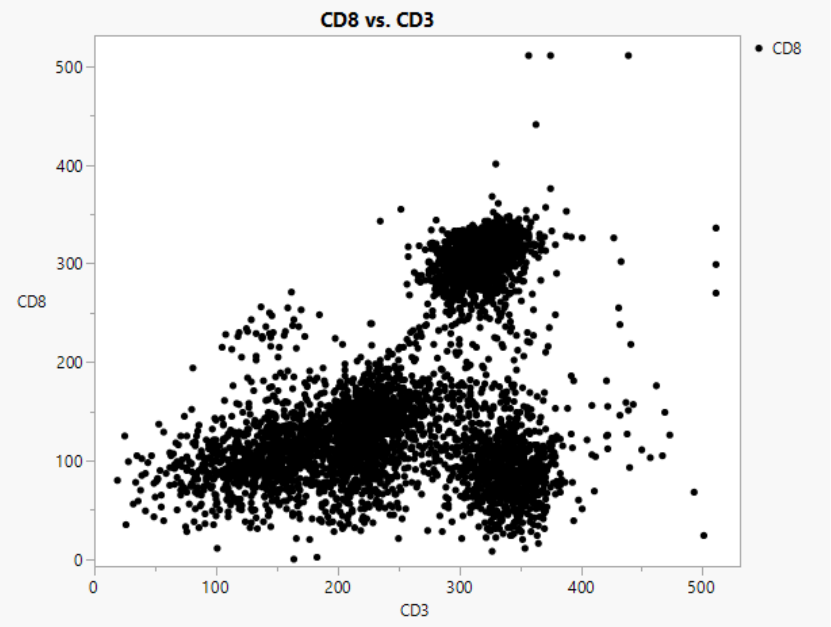




One distinction between box plots and HDR plots is that the 50% and 99% probability regions in an HDR plot are not necessarily continuous. When the underlying data is not unimodal, this can give a better picture of the underlying distribution of the data. Compared to the Violin representation, the mode lines and outliers give a more quantitative presentation of the underlying density function.

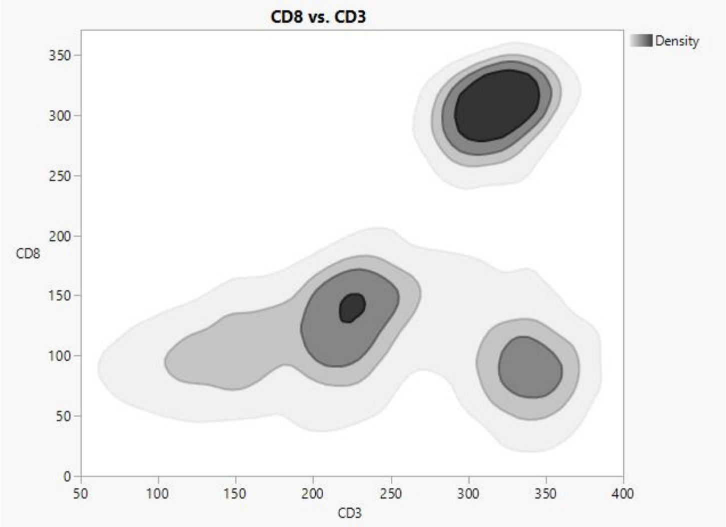
Contours in 2D

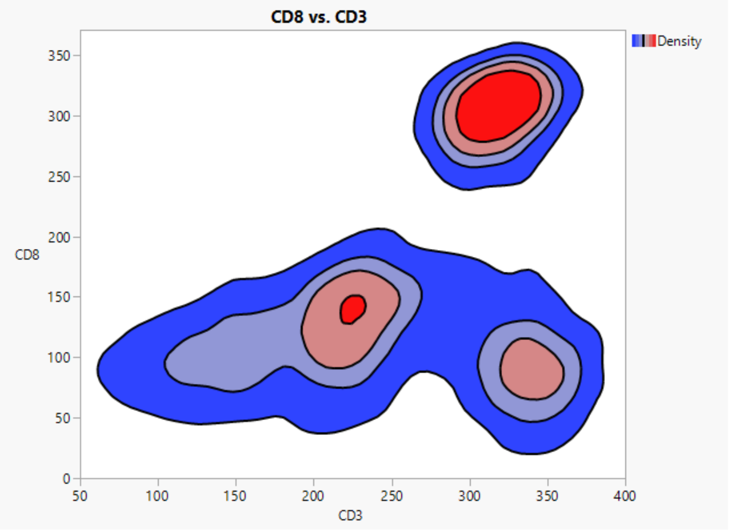
Scatterplots for two continuous variables have some of the same disadvantages as scatterplots in 1D. When the data is dense, it can be difficult to understand the distribution of the points. One common practice is to use transparency for the points in a 2D scatterplot. The regions with the most overstriking will appear darker in the resulting plot.



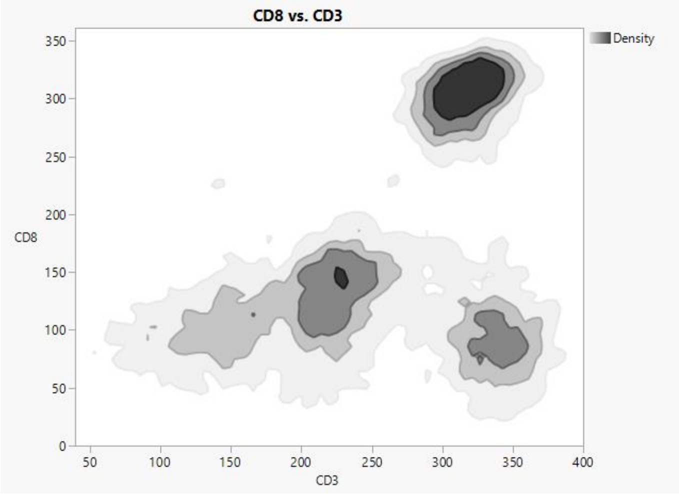
2D Density contours

The Graph Builder contour element uses a bivariate nonparametric density computation to extract density contour to directly display regions of various densities. By default, 4 levels are displayed. JMP 15 adds new visualization properties for 2D density contours, including line/fill options (from the property panel), and density gradient and transparency gradient options (from the legend).



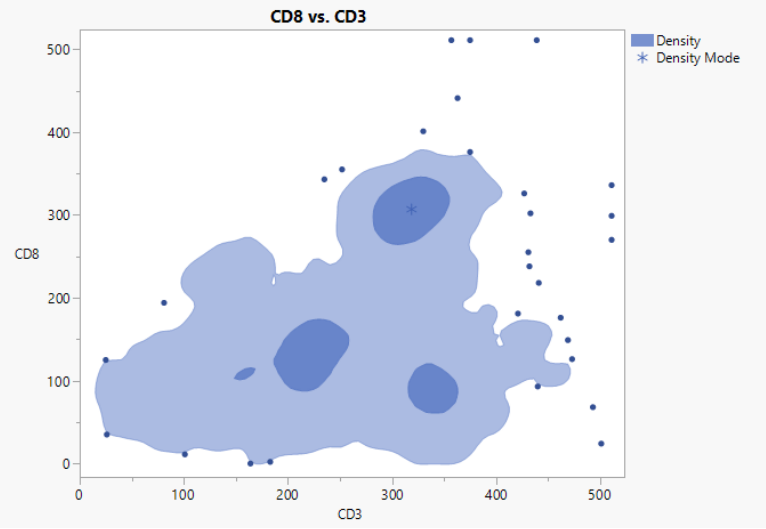


Also new in JMP 15 is the ability to change the smoothness for 2D density contours – as in the 1D case, there is a single normalized smoothness parameter ranging from [-1,1]. The default values of 0 represents a default kernel bandwidth in each dimension, negative values decrease the smoothing, and positive values increase the smoothing.



2D HDR contours

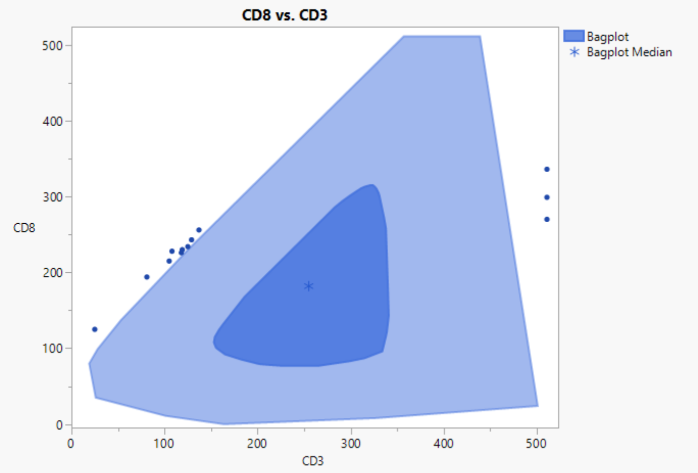
The concepts of HDR (Highest Density Region) plots for 2D density functions are directly analogous to the 1D case. Using the same bivariate density function, a marker (*) is drawn at the point of highest probability, a dark contour is drawn for the area of 50% cumulative probability, a lighter contour is drawn at 99% cumulative probability, and anything outside of the 99% region is (optionally) shown as an outlier.



2D Bagplot

Another new representation in JMP 15 is the bagplot. The bagplot was designed as an analog to the box plot for 2D data. In 2D the "order" of the points is not given. This is resolved by first computing the Tukey depth of each point. Consider all lines going through any point. The Tukey depth is the minimum number of points lying to one side of any of these points. Points that lie near the edge will have low Tukey depth, and points that lie near the center of a dense collection of points will have the highest depth. Further, for each depth value *i > 1* you can take the convex hull of all the points with *depth > i*. These hulls define a natural nesting of the points from outside to inside.

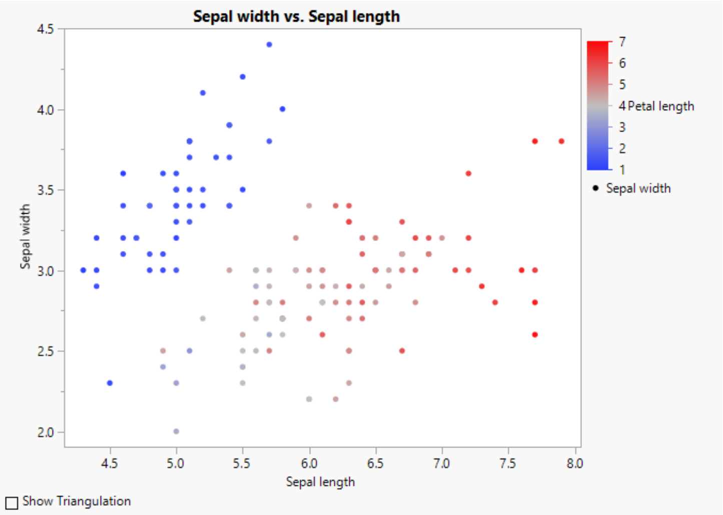
Given these concepts, the bagplot consists of four elements. A marker (*) is shown at the median of the highest depth points. The two hulls that separate the points in half are interpolated to determine a 50% probability region, which is shown in the dark color. This is referred to as the *bag*. The bag is inflated by 3x relative to the median to determine the *loop*, which is not shown. The convex hull of all points inside the loop is shown in a lighter color and is called the *fence*. All points outside the loop are considered outliers.

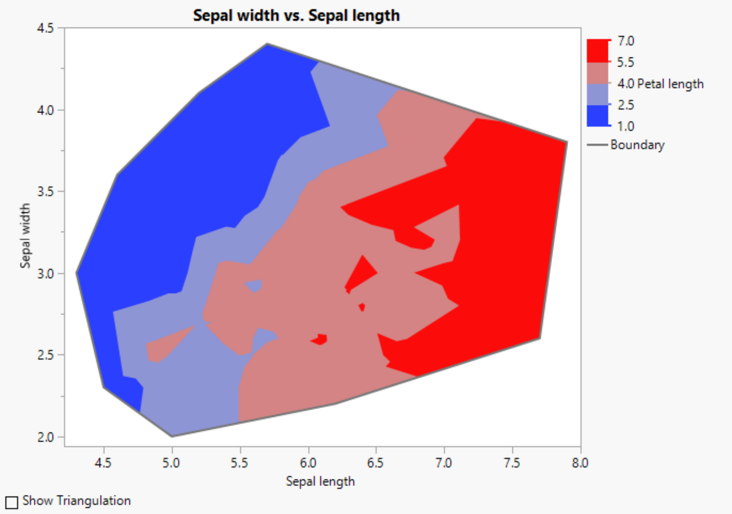


The bag plot is a very close 2D analog of the 1D box plot, and similar distinctions can be made when comparing the bag plot to the 2D HDR plot. The bagplot is better suited to summarizing the distribution of unimodal data, while the 2D HDR plot has an advantage when giving a summary view of multimodal data.

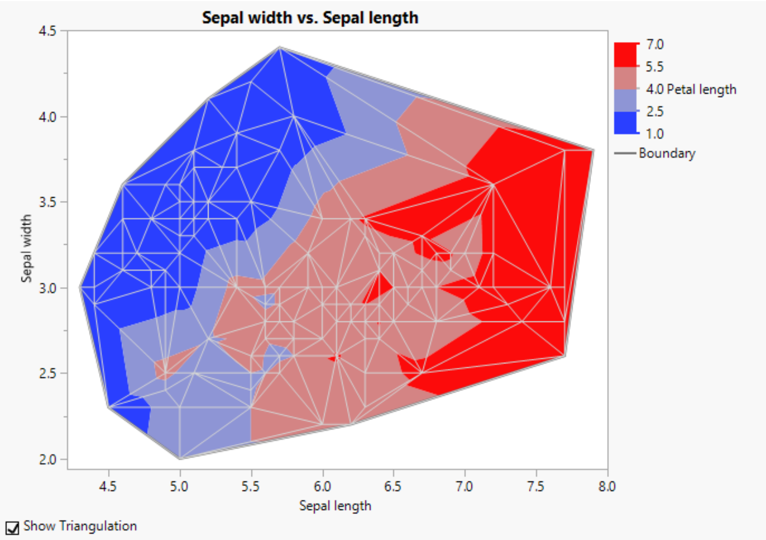
Value Contours

When an additional color role is specified in addition to two continuous columns for X and Y, the Graph Builder contour plot shows value contours, like the Contour Plot platform.

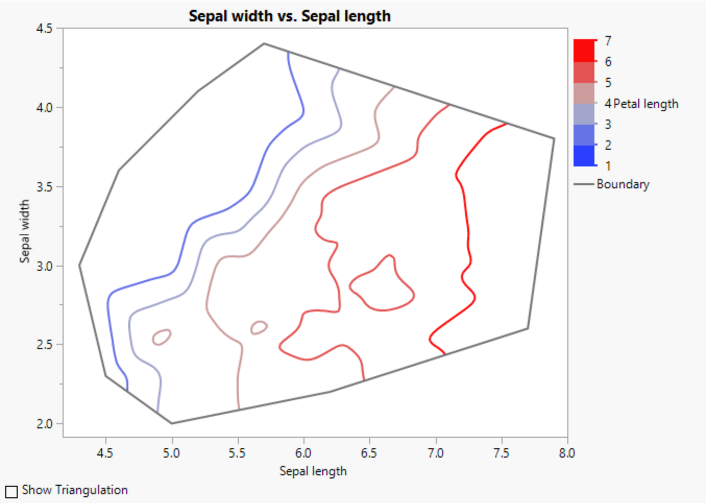




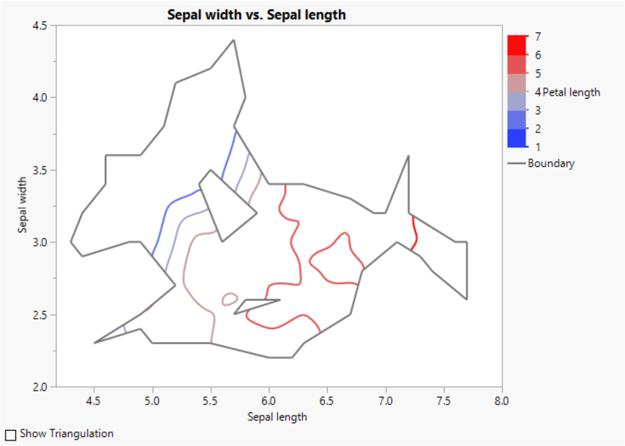
Contours in this case are computed using a piecewise linear interpolation of the input data based on a Delaunay triangulation. The triangulation is not shown within the Graph Builder platform – in the example below the triangulation was added via JSL.



JMP 15 adds new options for showing the boundary, filled contours, and line contours. For value contours, smoothing is done by interpolating the triangular data to a fixed grid and applying gaussian smoothing to the grid. The smoothed contours are then clipped to the boundary of the triangulation.



A Delaunay triangulation always results in a convex region. If the points are not uniformly distributed, this can lead to interpolation between points that are spread over large areas. The Alpha Shape is a subset of the Delaunay triangulation that filters triangles that are very large or very long and skinny. The filter looks at the radius of the circumscribing circle for each triangle, removing those that are larger than some limit. The resulting subset could have holes or multiple components.



Summary

The new HDR methods provide a quantitative view of density-based contours for both 1 and 2 continuous variables. All Graph Builder density contours now support a smoothing parameter, and the bagplot extends boxplot concepts to 2D. Alpha shapes control the shape of the domain for triangulation-based contours. All of the contour types have new controls for drawing lines or filled contours, and 2D density contours also have additional control of the color palette and transparency.

References

H. Edelsbrunner, D. Kirkpatrick, R. Seidel, On the shape of a set of points in the plane, IEEE *Transactions on Information Theory*, Vol. 29, Issue 4 (July 1983), 551-559.

Rob J. Hyndman, Computing and Graphing Highest Density Regions, *The American Statistician*, Vol. 50, No. 2 (May 1996), 120-126.

Peter J. Rousseeux, Ida Ruts, John W. Tukey, The Bagplot: A Bivariate Boxplot, *The American Statistician*, Vol. 53 (1999), 382-387.

J.W. Tukey, Mathematics and the Picturing of Data, *Proceedings of the International Congress of Mathematicians*, Vancouver, 1975, 523-531.