Subject: Re: Frustrated by 2 Data Plotting problems
Posted by penteado on Sat, 28 May 2011 19:29:06 GMT
View Forum Message <> Reply to Message

On May 28, 12:21 pm, David Fanning <n...@idlcoyote.com> wrote:
> But this sort of proves my point. If I run your program
> with 1 percent of the points, the "visualization" doesn't
> change in any material way, but the time is reduced by
> a factor of 1000.

It does not change in that case, but it can easily not be the case. I have one particular application where I can have millions of points to plot, and the visualization would change substantially if I took a random subsample.

All it takes is for the distribution of points to be very non-uniform along it. Then the random subsample might (in some cases probably would) miss those few points that have very different characteristics (because, say, nearly all points fall in the same region, with a lot of overlap, but only one in a 1000 will fall in a distinct region in the plot). A common situation, for instance, when one works with the spatial distribution of observations, where some regions, due to geometry / instrument constraints, are only observed rarely.

The plot may have a lot of overlapping points, but still be interesting. As long as the overlapping points do not cover everything, there is room to have the different (frequently the most interesting) points falling in other regions. And this may not show well in 2D histograms, which may not resolve well those few odd points. That is the reason why in some visualizations I used both a scatterplot and a 2D histogram: the histogram shows the distribution well where there is a lot of overlap, while the scatterplot shows well the uncommon points.