Subject: Re: "Correct" Data Philosophy
Posted by Kenneth P. Bowman on Thu, 17 Dec 2009 21:23:16 GMT
View Forum Message <> Reply to Message

In article <MPG.25940db1221ff3269896aa@news.giganews.com>,
David Fanning <news@dfanning.com> wrote:

> Folks,
>
> Every couple of weeks I get an e-mail from someone whose
> data is "missing" and they want to replace it with the
> "correct" value. These e-mails bug me because if the
> data is "missing" how the hell would I know what the
> "correct" value is suppose to be?
>
> But, generally speaking, they want some method to
> guess at the "correct" values by looking around the
> neighborhood, shuffling their feet, etc. I guess we
> have all been tempted to fudge data, if only for
> aesthetic reasons, so maybe it is a legitimate request.
>
> What would you tell them to do?
>
> If I get some good suggestions I'll write an article
> so I can get rid of these requests in the future. :-)
>
> Cheers,
>
> David

The problem of estimating values where you have no data is
very common and often very difficult. The best approach depends
on the character of the data, the size of the gaps, the methods used,
and the purpose of the analysis.

It is very important to not mislead yourself or your readers.
My first recommendation is *not* to fill gaps whenever possible --
instead, adapt your analysis and display methods to the data.
If you are displaying an image or contour, for example, show
the viewer where the data is missing with a special color
and don't display contours where there is no data.

If I am plotting global maps of 5 deg x 5 deg data, it should
look chunky (pixelated), not smooth. That reminds the viewer
what the actual resolution of the data is.

If you need to do a Fourier transform, consider using
least-squares estimation rather than interpolating

and using an FFT.

If the data is smooth and the gaps are small, interpolation will probably work well. If the data is noisy and the gaps are large, it is possible that nothing will work well.

If you do fill gaps, always test the impact on your results. Does it matter whether you use linear or cubic interpolation, for example?

In the end, you need to be confident that your results do not depend significantly on how you chose to estimate the missing data.

Cheers, Ken