Atmos. Chem. Phys. Discuss., 7, S7442–S7446, 2007
www.atmos-chem-phys-discuss.net/7/S7442/2007/

Atmospheric
Chemistry
and Physics
Discussions

Interactive comment on "Method for evaluating trends in greenhouse gases from ground-based remote FTIR measurements over Europe" *by* T. Gardiner et al.

Anonymous Referee #3

Received and published: 4 December 2007

This paper presents some useful interpretation of data from six NDACC sites within European longitudes, extending in latitude from 28 to 79 deg N. The main thrust of the paper is its presentation of methods for statistical analysis of the data, but in this respect it falls short in several respects.

Bootstrap analysis is well established as one of several non-parametric statistical tools used to avoid some of the asssumptions of classical statistics where those assumptions are dubious. Gardiner et al. correctly observe that the familiar linear regression technique for trend analysis assumes independently identically distributed (usually Normal) residuals that can be treated additively. Bootstrap analysis avoids the assumption of normality, or indeed any parameterised distribution, but in its simplest form it does not avoid the other assumptions.

Indeed, identical distribution of residuals (homoscedasticity) is still assumed, and anomalous results can arise if it is untrue. Trace gas concentrations that are very variable but strictly non-negative can be a problem in this way, as for example in the height retrieval by optimal estimation used in FTIR work. The examples in Gardiner et al.'s Figure 1 suggest that this issue (constancy of the distribution of residuals) should at least be checked for ethane, for example.

Heteroscedasticity can be readily corrected by weighting, as in classical statistics, and transform or link functions (as in generalised linear models) can adjust for non-normality also, as does bootstrap analysis. The real problem is the first assumption, independence, as it is all too obviously untrue in geophysical data series. Although Gardiner et al. acknowledge this issue, they dismiss it too easily. Multilinear regression to remove seasonal ('intra-annual') variation is very straightforward, and other geophysical correlates (QBO, SOI, solar cycle, etc.) are often used at longer time scales.

The problem is that even after all known or suspected factors are included in the regression model, the residuals are still not independent. This is evident in Gardiner et al.'s Figure 1, where the residuals (blue cross values minus red triangle values) are clearly not white noise.

One approach to this problem is to treat the residuals as an autoregressive (usually AR(1)) time series, and then back out the purely additive residual from that (e.g., Tiao et al., JGR, 95, 20507-20517, 1990). Bootstrap analysis, randomly resampling the residuals so that they genuinely are temporally independent, does something different again. I haven't quite worked out what that is, and Gardiner et al. have not enlightened me. If, as I suspect, they are not certain either, then a somewhat different approach should perhaps be taken.

The paper notes in section 5.2 that "95% confidence intervals ... were also calculated under the (unsupported) assumption that Gaussian statistics apply". Assuming this means that non-normality was shown to be not supported, rather than just not shown to be supported (i.e., untested), I would prefer to actually see the extent to which it fails (e.g., by quantile-quantile plot). "In most cases [classical results] were comparable to the bootstrap resampled results", so the misfit is perhaps minor. Combining site trends, weighted by inverse variance, is also readily possible by traditional methods. By my calculation it gives 0.258 +- 0.053 %/yr for the combined N2O trend, compared with the 0.245 +- 0.044 %/yr for the bootstrap, and 0.25 %/yr from AGAGE and NOAA/CMDL; all equal, effectively.

Thus I am not persuaded that the bootstrap technique has produced much better, or even particularly different, results, so there may be no strong argument for it in this instance. Equally, there is no evidence against it here, and bootstrap analysis is well established in its own right for this sort of application. In particular, it is ’robust’ against the effect of outliers, which can invalidate classical statistics. Gardiner et al. could comment on whether outliers are a feature of their datasets. I wonder too why confidence limits are all given as symmetric (+- 0.XX) in their Table 3, when they don’t arise that way from the 2.5 and 97.5 percentiles.

Even in a paper for which the focus is on illustrating applicability of a statistical method, or perhaps especially in such a paper, more care is needed with interpreting the few results shown. "The results show broadly similar trends in [the] troposphere and stratosphere for most species, except for ozone, and to a lesser extent carbon monoxide,...". It is true that for these species the signs of the tropospheric and stratospheric trends are opposite, but the confidence limits for CO overlap, and for O3 they don’t miss by nearly as much as do those for CH4, N2O, and even HCFC-22, in proportionate (i.e., probability) terms. The whole point of confidence limits is to determine when different measurements are close enough to be considered to come from the same distribution, and therefore eligible to be combined into a single value.

The same question as for trop-strat comparison arises with the site comparison of measurement and model for different species, but again the authors seem more concerned with sign than confidence intervals.

The bootstrap method, as applied here, must give the same trends as simple regression with the same predictors. Resampling the residuals and adding them back to the linear (in ordinate, not abscissa) model to fit again would look dubious if it biased the trend, as Gardiner et al. acknowledge and check. The purpose of different analyses, then, is to refine not the trend but its confidence limits.

I was initially alarmed to see the suggestion that "In addition to the statistical validation of the model, the results of the trend analysis of the data from the UFTIR sites can be compared to ... the output from atmospheric models, to give external validation of the results". Models are validated against measurements - not the other way around. On a more careful interpretation, it is reasonable to validate a technique for determining confidence limits by using it for comparisons where we think we already know the answer. If we are confident that things really are from the same distribution, about one in 20 measures should stand out as in disagreement at the 95% level. More than that suggests that confidence limits are too narrow; much less and they may be too wide.

In summary, the paper needs: - some demonstration at the outset of why normality is questionable; - a discussion of the assumption of identical distribution of residuals; - acknowledgement of the problem that an (inevitably) imperfect model leaves some structure (dependence) in the residuals and, if possible, some comment on how this affects bootstrap results; - a focus, in the comparisons and discussion, on how well the confidence limits from their analysis match expectations of what they should convey.

With those additions, the paper would be well worthy of publication in ACP.

Minor points:

’Data’ starts as a plural noun, and becomes singular later (e.g., p 15783, line 26). Stick with the former.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

p 15783, line 18 ’which cover the latitude range from’ would be better as ’which range in latitude from’. Discrete values cannot really cover a range.

p 15784, line 17 ’varied between’ would read better as ’ranged from’ here. Tropopause height varies with time, but the sense here is latitudinal dependence.

p 15785, line 13 ’slope estimate that would be observed if the data was’ should read ’slope that would be computed if the data were’

p 15785, line 16 ’gradient parameter is also associated wiht a normal (Gaussian) distribution’ is true, but, as usual when its standard error is derived from the data rather than given a priori, the statistic $<$slope$>$/$<$s.e. of slope$>$ has a t-distribution, with n-m d.o.f. for n data points and m parameters in the model.

p 15788, line 5 ’experimental data analysis’ is ambiguous. Does it mean exploratory data analysis, where different techniques are tried, or just analysis of experimental data?

p 15792, lines 24 and 28 Figure 3 is Figure 2

p 15794, line 14 ’data sets’

Interactive comment on Atmos. Chem. Phys. Discuss., 7, 15781, 2007.