Geosci. Instrum. Method. Data Syst. Discuss.,
https://doi.org/10.5194/gi-2019-10-RC1, 2019

Geoscientific Instrumentation Methods and Data Systems Discussions

Interactive comment on "Soil CO_2 efflux errors are log normally distributed – Implications and guidance" *by* Thomas Wutzler et al.

Anonymous Referee #1

Received and published: 16 September 2019

The application of the log-normal approach can be made even more convincing and consistent

Contents:

Random errors are usually modelled with a normal distribution and a common error standard deviation. The paper shows that both assumptions are inadequate, at least for single measurements from soil CO2 efflux devices. These findings may well generalize to many other environmental measurements. The alternative model of a log-normal, multiplicative random error appears much more suitable and plausible and leads to more efficient and appropriate analyses.

Theory:

The theory on the lognormal distribution, the distribution of sums of such variables and the variance of sums of correlated variables is nicely summarized.

There is, however, a misnomer: The notion of "confidence interval" should be reserved for the interval that covers the true PARAMETER value with the given confidence level, but it is used to name an interval in which approximately 95% of the observations should be contained. This may be called a scatter interval. It is related, but not identical to a prediction interval or a tolerance interval.

Application:

Fig. 1: The comparison of the two main methods - using the normal or the lognormal assumption - is first examined by a respective qq plot of residuals. The residuals are obtained as the differences between 4 individual observations and their average either on the original or the log scale. However, these 4 observations stem from 4 measurement devices in 4 fixed places, and an inspection of Fig. 3 shows that they are clearly subject to systematic differences. Thus, the 4 residuals do not represent 4 independent random errors. In addition, the 4 residuals are collected from many half hours and then shown in a single qq plot. As the authors show in Fig.2 (and also emphasize in the text), these groups of 4 do not have the same variances in case of the "normal method". Therefore, they should not be shown in a common qq plot.

Fig. 2 (left panel) and the related comments show quite convincingly that the assumption of constant standard deviations does not hold - they rather increase with expected values. This disappears when the lognormal distribution is used (right panel), since for this model, the standard deviation is proportional to the expectation. However, an alternative would be to model the random error as normal with standard deviation proportional to the mean. (Note that I would NOT prefer this approach over the log-normal model!)

Fig. 4 shows that the scatter interval ("confidence interval") for values aggretated over a day are most often considerably narrower for the lognormal method. In the text, the

authors call them "the same" and discuss the few exceptions instead. Note, however, that one should first make sure that the intervals produced by the two methods indeed show approximately the same percentage of covering the observations.

In summary, while the theory has a good potential to improve the methodology, the way it is applied is not convincing.

A suggestion:

Fig. 3 suggests that the measurements follow a model

$$Y_{tk} = h(t) \cdot \gamma_k \cdot \epsilon_{tk}$$

or, on the log scale,

$$\log(Y_{tk}) = g(t) + \beta_k + \log(\epsilon_{tk})$$

where t is the time, k, the measurement device (chamber), $g() = \log(h())$, smooth functions of time, and Y_{tk} and ϵ_{tk} are the observations and the (lognormal) random error. Thus, it would be adequate to fit this model (on the log scale) and then show its adequacy using diagnostic plots (residuals against fitted values and time, qq plot).

(The smooth function g may - at least for other target variables than CO2 efflux - advantageously be related to explanatory variables such as a daily and/or a seasonal cycle and environmental variables, still allowing for a smooth additional term.)

This model can be fitted to half hourly measurements or daily averages. (The averages may be the usual arithmetic means or robust version of them.)

If the above model fits well, without showing heteroscadasticity, an alternative method for aggregating measurements to daily averages may be used: the fitted values can be aggregated, and the correction factor for getting an expectation from the first parameter of a log-normal, $\exp(\sigma^2/2)$, can be applied to the result. This is similar in spirit as using the estimated parameters of the lognormal distribution obtained from the 4 measurement devices, as done in the paper, but that method wrongly includes the systematic

effect γ_k in the random variability of the error term.

Discussion:

The model just described has consequences for

- spatial aggregation: The dominant term in such means are the mean of the γ_k. They can be interpreted as random effects if the locations of the measurement devices are randomly chosen from the plot (or region). This is why aggregating over time first ("space last") produced a more plausible interval (Fig. 6). The authors correctly explain this fact in other words - that this occurs "because it wrongly assumed true replicates when in reality there are only pseudo-spatial replicates."
- temporal aggregation: If there is a daily or seasonal cycle, the time points should not be treated as a stationary time series. If the smooth function g looks like a stationary function without further patterns, then it can be integrated into the correlated error term.

The authors say: "We argued that several spatially distributed chambers correspond to samples of a lognormal distribution."
It would be difficult to judge the distributional form from 4 independent values γ_k. Thus, this amounts to an over-interpretation.

Process error:

The authors give many reasons why the magnitude of the process error may depend on the magnitude of the CO2 efflux itself. In fact, in biology and other fields, all kinds of variation typically depend on the magnitude of the variable under consideration, and it would be more necessary to find explanations if for some process this is not the case.

This is even true for measurement errors. I do not see an intrinsic reason why the measurement error should be better described by a normal than a log-normal distribution.

Hopefully, it is small enough that the two models cannot be distinguished. The authors might adapt their comments on measurement errors accordingly.

This contradicts their recommendation to “use the normal assumption for high- frequency, e.g. hourly measurements,” for which I do not find a convincing reason.

Recommendations:

These might be recast along the following line:
The lognormal model and methods are more appropriate for this kind of data. The feature of standard deviations being proportional to expectations is the dominant tigger of the improvements. When measurements are aggregated to means or sums, the central limit theorem grants that the fit of the normal distribution gets better, but it will rarely be better than the fit of the log-normal for environmental and similar measurements, even after aggregation. However, the use of the log-normal is most needed when considering single measurements, for example, when they are compared with model values from any models (as the authors mention it).

Since this discussion suggests major changes of the analyses and interpretations, I do not go into details of the text at this stage.