Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., author comment AC1
https://doi.org/10.5194/gmd-2021-174-AC1, 2021

Reply on RC1

Mirko Mälicke

Author comment on "SciKit-GStat 1.0: a SciPy-flavored geostatistical variogram estimation toolbox written in Python" by Mirko Mälicke, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-174-AC1, 2021

I thank Anonymous Referee #1 for the positive and constructive review of my work. Please find the *Referee comments (in italic)* below, followed by my answer.

Specific comments:

- *The pancake dataset was used as example data for the demonstration of the functions in SciKit-GStat and the author described the advantage of it in appendix A. But I still think a complicated real-world data example such as precipitation should be given to support the powerful of the package. Also, such case can give more realistic variogram usage clues to the users.*
  Answer: As mentioned, the reasoning for this unconventional dataset is described in the manuscript. A complicated geoscientific example is used in figure 6 and 7 and one can question if the shown variogram is representing a valid spatial model for that sample. With figure 2, it is more obvious to me how SciKit-GStat can be used to estimate a variogram. Nevertheless, a more geoscientific example, that geoscientists might relate easier to, can be added.
  I suggest to add a detailed tutorial based on more geoscientific data to the source of SciKit-GStat, as that is the right place for a detailed demonstration. For this to happen, a dataset is needed, which actually can be distributed under a MIT license. If a tutorial like this can be compiled, I suggest to add an excerpt of the tutorial to the manuscript, either by an additional sub-figure in figure 2 showcasing the default variogram plot, or by an appendix about more sophisticated examples, or preferably both.
- *The author declared that the SciKit-GStat version is 1.0, but only version 0.6 (or 0.6.6) of it could be found in github and online document website.*
  Answer: The referee is right, I should have made this clearer somehow. Version 0.6.X actually is the release candidate for version 1.0. There are no missing versions or something. I did not call it a 1.0.0-rc1 though, as there was active development in the package from other contributors at the time. Additionally, there might be additional code changes, during the discussion phase, that I want to add to the 1.0 version as well. With that in mind, and with respect to the time span of an open discussion of the manuscript, I decided to keep the 0.6.X as the current versioning to decrease the number of needed release candidate versions. With the revisions of the manuscript, the 1.0 version will be released and be in line with the manuscript. The current version is 0.6.9 and there are at least two additional iterations due to the valuable comments in this review.
- *Why non-Gaussian geostatistics are not covered in SciKit-GStat?*

Answer: non-Gaussian geostatistics are not covered, because I never used them, nor did any of the users ever request non-Gaussian methods. Additionally, I do not oversee the existing literature enough to only summarize, which methods actually all belong to 'non-Gaussian geostatistics'. I basically read the two publications about Copulas (Bárdossy 2006, Bárdossy and Li, 2008) and the generalized sub-Gaussian Model (Guadagnini et al., 2018). In both cases, the workflow is largely incompatible with SciKit-GStat. At the example of Guadagnini et al. (2018): Consider their figure 5 showing the workflow: I am not even sure at which step the non-Gaussian variant of SciKit-GStat would be involved, as some are clearly pre-processing and post-processing (such as process simulations for ie. flow or transport).
To wrap this up: Implementing only what I have heard of non-Gaussian geostatistics so far is already way out of scope for the software as well as the presented manuscript and beyond that, I doubt that the mentioned publications summarize the field of non-Gaussian geostatistics in its entirety.

- *Why the procedures that can fit a model directly based on unbinned data are not implemented in SciKit-GStat?*
  Answer: I also have to admit, that I am not familiar with these procedures in a sense, that I never used them. If I am not mistaken, one uses the unbinned data to find variogram model parameters in a maximum likelihood approach. I think it is worth mentioning, that the unbinned data can be plotted by the variogram instance and the data underlying the plot is also accessible as instance properties (For those unfamiliar with class properties in Python, these can be thought of dynamic instance attributes, that are calculated at access time). That means, finding the optimal parameters outside of SciKit-GStat and then pass them back into the class is already possible. The SciPy package includes various solvers to minimize a function. The formulation of this function would be the only remaining task to the user, which is a very common workflow in SciPy. It has to return the negative log-likelihood from a given set of variogram parameters. This is already possible and a frequent approach while working with SciPy.
  In terms of examples, Lark (2000) or Marchant and Lark (2007) are two publications I am aware of. I am ad hoc not sure if i.e. equation 14 of Lark (2000) can be applied just like this to the data as it is managed in SciKit-GStat in a generalized and automatic way. Equations 9 and 10 need to be adapted for all models in SciKit-GStat (involving other parameters) and SciKit-GStat handles the variogram model parameters different, which would need some adaptions.
  Beyond an example, there are substantial hurdles for implementations into the core of SciKit-GStat. One of the most basic design decisions underlying the Variogram class in SciKit-GStat, was to bind each instance to the used data, as briefly described in section 4.1.1 (p.18). This also includes the one dimensional array of distances (which is the row-wise upper triangle of the distance matrix) , the array of residuals in same order and the array of lag class indices in same order for all point pairs. Thus, the binning is the most fundamental processing step in SciKit-GStat, which technically cannot be suppressed. In an unbinned approach, the Variogram would always present bins, derived from default settings, along with the parameters or on the plots, which I find most counter-intuitive.
  In addition, the maximum likelihood estimation involves the inversion of the autocorrelation matrix for each iteration (eq. 9 in Lark (2000)). From Github issues, I am aware, that Variograms are frequently estimated for samples with n > 10,000 using SciKit-GStat. The autocorrealtion matrix has the shape n x n and this step might have serious performance implications. This might harm the interactivity aspect of SciKit-GStat seriously. Lark (2000) also mentions, that the maximum likelihood approach is not preferable for sample n > 150.
  Thus, changing this behavior is not desired and would involve the refactoring of almost 3000 lines of code. The right approach would be to introduce a new Variogram class tailored for these methods only. Although some parts of the class might be reusable, this is essentially a new major development and therefore out of scope.

Nevertheless, I will try to take the most out of the comment. For a preliminary examination, I used the last two days to script an application of Lark (2000), as I am not aware of any Python implementations. The fitting takes substantially longer and, right now, fails to find parameters too often (not sure why yet). The whole procedure uses only the distance matrix of the Variogram class, which illustrates my point of implementation effort.
I'll do my best in resolving issues and, if feasible, add the script as a tutorial-showcase to the documentation anyway. It could be helpful as a comparison of methods.

Minor comments:

- *Color bars should be plotted in figure 1, 4, and 8.*
  Answer: Figure 1 is a RGB image, for which a color bar is not really useful, I think. But I get the point that a reference is missing and suggest to add a subplot of the red channel, which in fact was used in the manuscript, and add a color bar to that one.
  A color bar will be added to figure 4 and to figure 10 (figure 8 does not hold any continuous information, but 10 does. The referee might have confused these two figures)
- *Line 99, change "SciKit-Gstat" to "SciKit-GStat".*
  Answer: Thank you for pointing out.
- *Line 212, remove one "all".*
  Answer: Thank you for pointing out.
- F*igure 3 was not explained clearly.*
  Answer: Agree. The figure caption and the referencing section in the manuscript will be checked and improved.
- *Figure 6, it's better to set transparent color to the surface part so the distribution of the scatter data could be clearer.*
  Answer: I agree with the referee. This figure shows the default plot of SciKit-GStat using the plotly backend. At the time of development, there was an issue with surface opacity in that library. But that seems to be resolved and I will happily update the figure in the manuscript. Additionally, the source code will be changed, as I actually spotted an error in the code, while formulating the answer to this comment, that prevents the user from setting the opacity correctly.
- *The data source of figure 6 was not described.*
  Answer: Thank you for pointing out. A section about the datasource will be added to section 2.1.

References

Bárdossy, András. "Copula□based geostatistical models for groundwater quality parameters." Water Resources Research 42.11 (2006).

Bárdossy, András, and Jing Li. "Geostatistical interpolation using copulas." Water Resources Research 44.7 (2008).

Guadagnini, Alberto, Monica Riva, and Shlomo P. Neuman. "Recent advances in scalable non-Gaussian geostatistics: The generalized sub-Gaussian model." Journal of hydrology 562 (2018): 685-691.

Lark, R. M. "Estimating variograms of soil properties by the method□of□moments and maximum likelihood." European Journal of Soil Science 51.4 (2000): 717-728.

Marchant, B. P., and R. M. Lark. "Robust estimation of the variogram by residual maximum likelihood." Geoderma 140.1-2 (2007): 62-72.