Diagnostics/Good Practice

by

David Meek, USDA-ARS NSTL, Ames, IA 50011

I. Spatial diagnostics for analysis residuals or on a stationary series.

In the following passages on recommended practices, I use the following terms:

h is the lag distance,
σ^2 is the variance of the orignal data,
$\gamma(h)$ vs h is the method of moments (MM) empirical semivariogram (ESV),
$\hat{\gamma}(h, \mathbf{c})$ is the modeled semivariogram with **c** vector of the fit parameters,
$\gamma^{\dagger}(h)$ vs h is a robust semivariogram (I use the Cressie-Hawkins estimator),
c_0, c_S, and a_R are the nugget, partial sill (sill=c_0 +c_S), and range parameter from $\hat{\gamma}(h, \theta)$,
$L_{½}$ is the Journal's practical rule domain limit (the maximum h to consider for the $\hat{\gamma}$)
C(h) vs h is the covariogram,
$\rho(h)$ vs h is the correlogram, and
θ is the scale of fluctuation

In the graphical examples I often use use $2\gamma(h)$ and the empirical variogram (EV). I am considering a general spatial model as the superposition of large scale (global or total domain like a trend in time series), one or more intermediate scales, and small (local like autocorrelated residuals in a time series model). In the context of what you sent me, then, these are mainly spatial diagnostics for analysis residuals or stationary series. Developing a sound locally representative spatial model is the goal. The graphs are ones that I already have so I can redo them easily to suit you. I refer to these as plots 1 to 5. Most of the ideas are from a variety of well-known sources.

1. Plot omnidirectional $\gamma(h)$ vs h and directional $\gamma_{\alpha}(h)$ vs h for at least 2 directions if data are sparse ($\alpha = 0$ and $\pi/2$), otherwise consider directional $\gamma_{\alpha}(h)$ for up to 6 angles in the upper half plane (α= 0 to $5\pi/6$ by $\pi/6$).

2. Plot both robust $\gamma^{\dagger}(h)$ vs h and regular $\gamma(h)$ vs h.

3. Plot the $\gamma(h)$ vs h and the pair count vs h with the h = $L_{½}$ line and the $\hat{\gamma} = \sigma^2$ line.

4. Plot $\gamma(h)/h^2$ vs h [I think this is known as a 'Regularity Test'].

5. Plot either one or both $\rho(h)$ vs h and C(h) vs h.

Note I have incorporated these concepts in the following 3 figures.

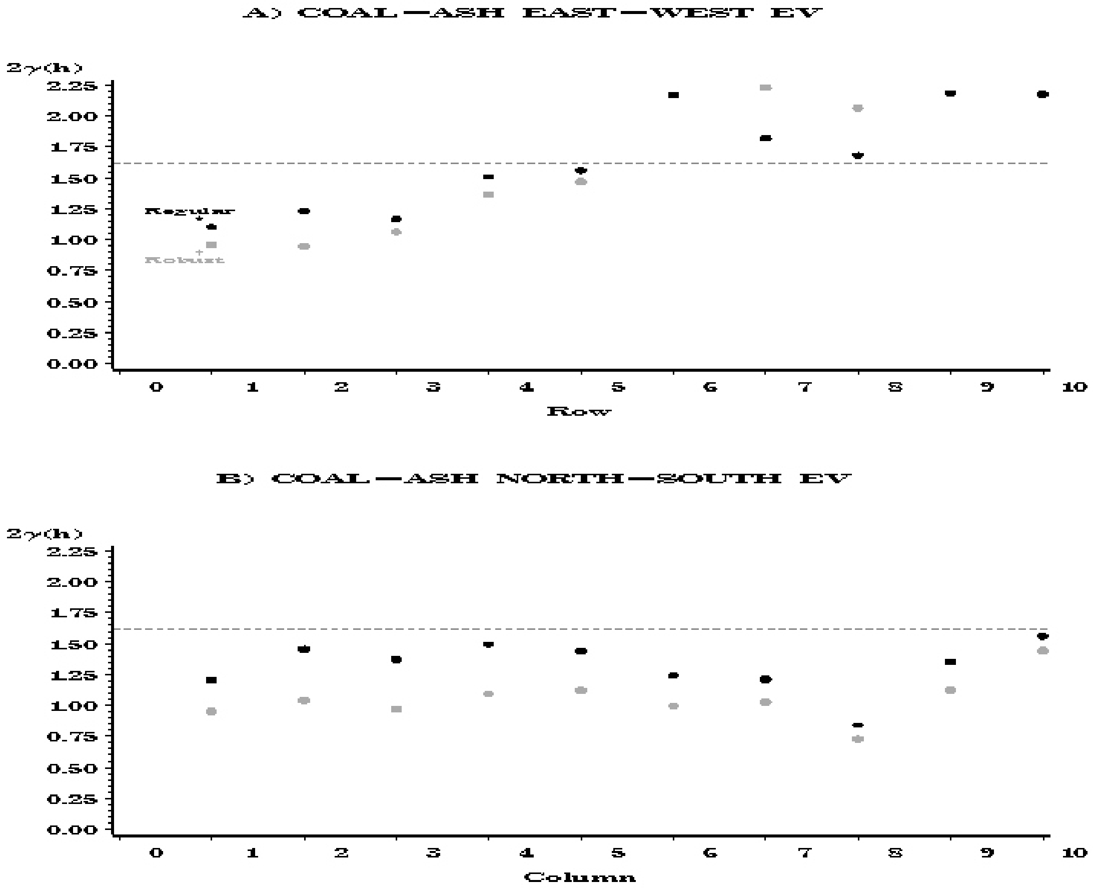

Figure 1. Directional variograms with regular and robust estimates.

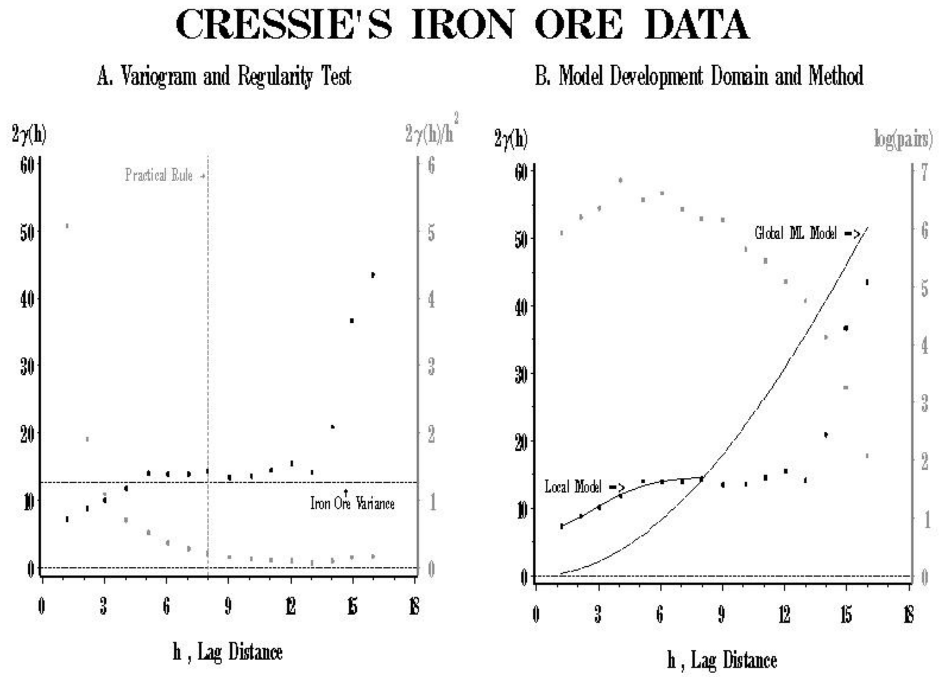

Figure 2. Variogram diagnostic plots.

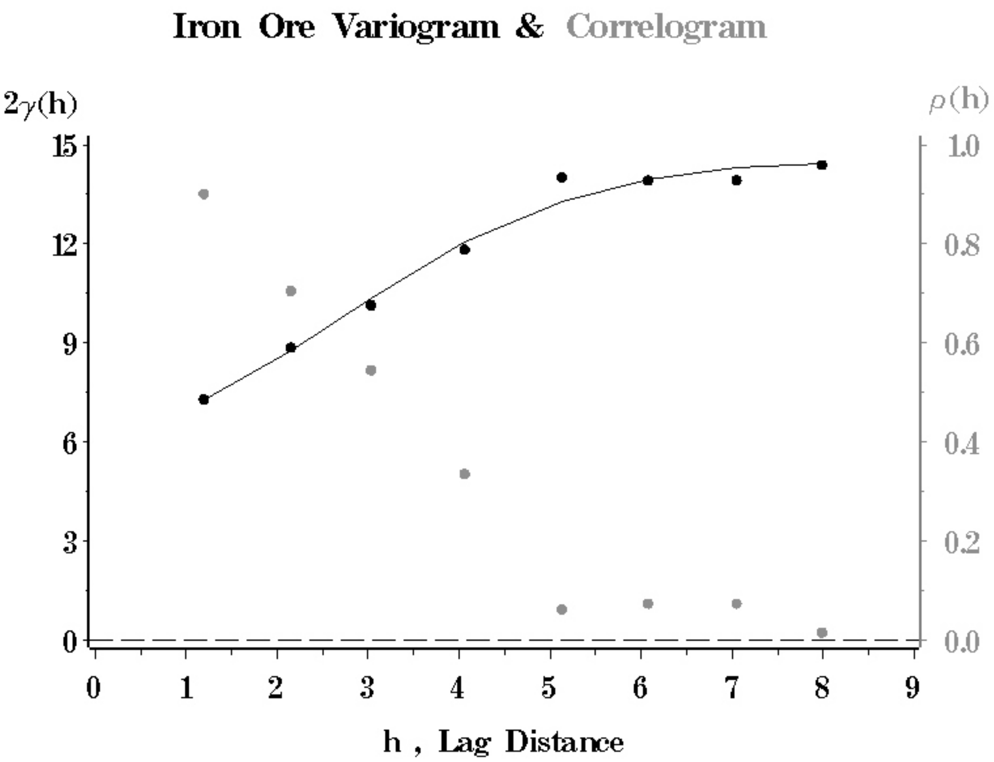

Figure 3. The correlogram as a diagnostic.

II. Assessments.

1. Estimator adequacy:

A. Can isotropy be assumed? Construct plot 1 (compare top to bottom panels in Fig. 1). Are the EVs in each considered direction reasonably similar? Note these EVs are from Cressie's raw coal-ash data, not median polished data. Final recommendations were a two part model - a linear trend in the E-W direction (large scale) and directional spherical semivariogram model in the N-S direction. The domain was irregular with 23 rows and 17 columns. For Fig. 1, only 10 lags were shown because $L_{1/2} \approx 10$ was estimated using the geometric mean of both directions because that is what one could use for the omnidirectional plot (I have not included this one but I can if you want). Of course to model the N-S behavior, one should only consider $\gamma_N(h) \leq 17^{1/2}$.

B. Are the $\gamma(h)$'s OK? Construct plot 2 (observe the different shaded dots in both panels in Fig. 1). Many programs give just a mean value of $\gamma(h)$ for each h but do not show or assess the distribution. Even if at each h the distribution is unimodal it may be nonnormal and heavy tailed. Some programs give $\gamma^{\dagger}(h)$ estimates in addition to $\gamma(h)$. In this case, one can at least set an ad hoc rule on $\gamma(h) - \gamma^{\dagger}(h)$. Moreover, ask: "Are all of the original observations acceptable?" A whole other but related issue is binning for irregularly spaced data. I have not addressed this issue at all here.

2. Stationarity:

Can some form of stationarity be assumed? Mention of some EDA methods that can be done are outlined in the KSU kriging paper so I can focus on the small scale. For the analysis on the EV or ESV, construct the plots in 3, 4, or 5 above.

Ask, does $\gamma(h)/h^2 \to 0$ as $h \to L_{1/2}$ [see gray-scale points and use right axis in Fig. 2A]? Recall that the exponent in a power model is constrained to the interval ,$0 \leq p < 2$. A power model can be a trend indicator. A plot with $\gamma(h)/h^2 > 0$ near or after $L_{1/2}$ shows a need for $p \geq 2$.

Or does $\rho(h) \to 0$ as $h \to L_{1/2}$ [see gray-scale points and use right axis in Fig. 3]? Recall that in time-series, an exponential decay pattern in ρ that continues past a quarter of the length the original series is considered to be a trend indicator. In spatial analysis $L_{1/2}$ is more commonly used. For the series shown in Fig. 2 the stationarity assumption was considered reasonable.

FYI: I have an artificial data set example that compares a pure trend model to an EV under the stationarity assumption. If you want it, let me know.

3. Model/Series Properties:

A. How well determined is $\gamma(h)$ at each h? Construct the second graph in plot 3 [see gray-scale points and use right axis in Fig. 2B]. Here $L_{1/2} = 8$ was used although one could also use 9. The local series is well represented for $h \leq L_{1/2}$ but, as one would expect, the pair count drops considerably afterward. So, in turn, ask: What does the series represent? The data set is one I use to show what you get by using MM estimators to develop the EV and model the local behavior. The local model is the Gaussian in which the parameters were estimated with nonlinear least squares using Cressie's practical compromise weight. In contrast, the global Gaussian model is shown. It was estimated with an unweighted and unrestricted maximum likelihood procedure applied to original coal-ash series. This method, in effect, uses all $\gamma(h)$ values over h domain and weights them equally (this result is why one fixes p(h) in SAS proc mixed with the noiter option).

B. Is there a sill and doe it ≈ $\frac{1}{2}\sigma^2$? Construct plot 3 and inspect it [see Fig. 2A]. Sometimes the sill is not $\frac{1}{2}\sigma^2$. This finding may be real and needs to be considered. Often though it is so because there is one or more other factors present like trend, anisotropy, outliers, etc. Try to eliminate these possibilities first. Inspecting the EV may lead to other interesting properties that need to be addressed like nested structures (like complete Clarke's Adit-Silver series [I have a file and plot if you want them]).

C. What are reasonable models to consider? Inspect the EV, regularity plot, and correlogram. Asymptotic and bounded models assume a sill. The model fit, even if poor, insures positive definiteness. A conditions like $\gamma(h)/h^2$ or $\rho(h) \rightarrow 0$ as h↑, especially if $\rho(h) \approx 0$ after some h value, is needed to make the assumption of a sill reasonable. Alternatively, if the assumption is bad, then consider unbounded forms but be wary of possible trend, intermediate or larger scale spatial variability.

4. Spatial correlation and further analysis:

How serious (of what consequence) is the spatial correlation? Again construct or refer plots 3, 4, and 5. The major effect on an analysis that considers the spatial correlation is that the effective degrees of freedom (df) for any test are less and so the associated probability (P-value) for the test is less than it would be if all the original data were independent. Most software outputs the adjusted df and P-value. There are some ad hoc calculations that can suggest when the effect of spatial relationship can be considered strong. There two separate concepts but both can occur together. First, consider $a_R/L_{\frac{1}{2}}$ or preferably $\theta/L_{\frac{1}{2}}$ for bounded models. If either one approaches or exceeds 1 then considerable spatial correlation is indicated. Of course, an unbounded model unless restricted is defined over the entire range of lags. Second, consider the ratio of $c_0/(c_0 + c_S)$ or $\gamma(1)/(\sigma^2$ or other ad hoc sill estimate). The smaller the ratio, e.g., < ½, the more consequential the effect. Note that in some software packages the choice of spatial correlation forms is restricted. Studies have shown that it is better to try to account for spatial dependence in some way than not at all.

III. Some suggestions on computer/numerical methods.

1. Model starting values:

How do you estimate initial values for the model? With a trained eye, you can guess them from the plot or list. Some ad hoc rules are suggested in the scale of fluctuation paper. First some suggestions for a long series. For c_0 do a linear regression, preferably weighted, and use the intercept of the model developed with the first few lags. If considering a highly curved form, do the same but use $\log(\gamma(h))$ and transform the estimate. For $c_0 + c_S$ estimate the mean $\gamma(h)$, preferably weighted, using a few lags around $L_{\frac{1}{2}}$. Guess a_R from the EV for a bounded line or spherical model by considering the h where $\Delta\gamma(h) \approx 0$ and convert it to θ. Otherwise, if θ were known by another way, use the known θ relations to a_R to estimate a_R for other suitable models. This method may not always work but if the models are reasonable candidate, I have found θ

estimates to nearly identical. For example, consider $\gamma(h)$ series for STP from Sauer and Meek (2003):

Model	Distance Parm.	Range/Practical Range	θ_E	$2\theta_E$
Spherical	58 m	58 m	44 m	88 m
Exponential	22 m	65 m	44 m	88 m

This is why I think using θ often provides more insight and can be better than just considering a_R. Note that θ is generally reported in random field literature. For short series, try $c_0 = \gamma(h)/2$ or $\gamma(h)/4$ for steeper curves and $c_0 + c_S = \frac{1}{2}\sigma^2$ [or alternative suitable variance estimate] or $\gamma(h)$ at the h where $\Delta\gamma(h) \approx 0$.

2. Model routine:

Are you sure of your $\hat{\gamma}(h, \mathbf{c})$ model? What if it does not converge or converges to unacceptable values? Note I use the pragmatic compromise weight to start. Assuming the convergence criterion is acceptable, here are some things try: If $c_0 < 0$ then drop it and retry. Appropriately numerically integrate the $\gamma(h)$ series then use an integral semivariogram form (ISV) of the model. Next, of course, one can always resort to the use of generalized least squares (GLS) methodology.