A Diagnostic to Assess the Fit of a Variogram Model to Spatial Data

Journal of Statistical Software, Vol. 1, Issue 1, 1996

Ronald Paul Barry
Dept. of Math. Sciences
Univ. of Alaska Fairbanks
Fairbanks, Alaska 99775-6660
ffrpb@aurora.alaska.edu

ABSTRACT

The fit of a variogram model to spatially-distributed data is often difficult to assess. A graphical diagnostic written in S-plus is introduced that allows the user to determine both the general quality of the fit of a variogram model, and to find specific pairs of locations that do not have measurements that are consonant with the fitted variogram. It can help identify nonstationarity, outliers, and poor variogram fit in general. Simulated data sets and a set of soil nitrogen concentration data are examined using this graphical diagnostic.

Key Words: kriging, graphical diagnostic, pocket plots

1. INTRODUCTION

The variogram is both of interest in its own right, and as a component of the kriging equations. With a two dimensional spatial process that may contain outliers or may not even be intrinsically stationary, it is difficult to determine if a fitted variogram really is consonant with the data. Such diagnostics as the pocket plot (Cressie, 1993, p. 42) can help detect limited pockets of nonstationarity. This paper discusses the use of a graphical diagnostic that can be used to investigate the fit of a variogram to data.

2. DIAGNOSTIC

The diagnostic is quite simple. If $\{s_i\}$ are a set of locations and $\{Z(s_i)|i = 1, ..., n\}$ a multivariate normal, intrinsically random process with mean μ and variogram

$$2\gamma(h) := \mathrm{Var}(Z(s) - Z(s+h)),$$

then

$$(Z(s) - Z(s+h))^2/(2\gamma(h))$$

is chi-square with one degree of freedom (Cressie, 1993, p. 96). The diagnostic first plots the n locations, then connects a pair of locations with a line if the locations are either unusually different or unusually similar, compared to the assumed variogram. The user chooses a variogram, and three parameters, the lower threshold $t_l \in [0, 0.25]$, the upper threshold $t_u \in [0.75, 1]$, view.range. Then, a pair of locations s_i and s_j are connected if and only if the two locations are within view.range (the default for view.range is the maximum distance between locations) of each other, and

$$\frac{(Z(s_i) - Z(s_j))^2}{2\gamma(s_i - s_j)} < \chi^2_{t_l}(1) \quad \text{or} \quad \frac{(Z(s_i) - Z(s_j))^2}{2\gamma(s_i - s_j)} > \chi^2_{t_u}(1).$$

If the data really comes from a process with the chosen variogram, the probability that any two locations are connected with a line because the measurements are too similar is t_l and the probability that any two locations are connected with a line because the measurements are too dissimilar is t_u. If we set t_l near zero and t_u near one, an abundance of connected pairs of locations will indicate a poor fit to the variogram model. Further, the pattern of these line segments may give us some insight into what has gone wrong.

Additional arguments can be used to modify the plot. If the argument thick.param is given a positive value, then more extreme lines (where the difference between measurements is much more than expected or much less than expected) will be thicker. If thick.param is set to zero (the default), then all pairs of locations that are connected are connected by segments of the same width. The argument jitter lets us add a small random displacement to each location *in the graph*. If the locations are in a lattice, a little jitter can make it easier to tell which segments connect which locations. If twiddle is set to TRUE (default), a dialog box appears that allows the user to interactively modify the graphical display. On a color screen, I prefer viewing both extremely similar and extremely dissimilar locations connected, on the same graph. With a black and white monitor, I would use the argument sidedness to view, for instance, a graph with only dissimilar locations connected.

In this paper I will describe an S-plus function vario.diag that produces this diagnostic plot. In Section 3 I give several examples of its use. Full documentation of the function is in the Appendix. A copy of the function along with some test data sets is archived with this paper.

3. SIMULATIONS

In all of the simulated data sets I used the locations used by Kay Gross in her study of nitrogen availability in three successional plant communities (Gross et al., 1995). To understand the use of the graphical variogram diagnostic, I generated several known random processes at these locations, and then examined the diagnostic plots. In all cases I used a view.range of 5 meters (to isolate shorter-range patterns in variability), a jitter of 0.1 to make it clearer which locations were connected by lines, a thick.param of 0.1, to emphasize unusually similar or dissimilar measurements. Unless otherwise stated, no variogram function was specified. Then the default variogram is used, which corresponds to independent random variables. The processes I examined had the following variograms: constant (independent random variables), exponential. I also examined an exponential process with a spatial outlier, and a lognormal process.

3.1. Independent Random I generated an independent random process with mean zero and variance one at all locations, and examined the the diagnostic plots. Figure 1 shows unusually similar locations connected, and figure 2 shows unusually dissimilar locations connected. These two figures will be the standard with which we will compare other plots.

Note that in figure 1 the orientation and position of the line segments is quite random. In figure 2 note that the lines are not thick, indicating that these differences are not extremely large. Note also that the typical pattern is a star-like pattern of segments radiating from a location that is a local "outlier". These should be checked as outliers in general, but some of these high or low differences between neighbors are inevitable in a set of independent random variables.

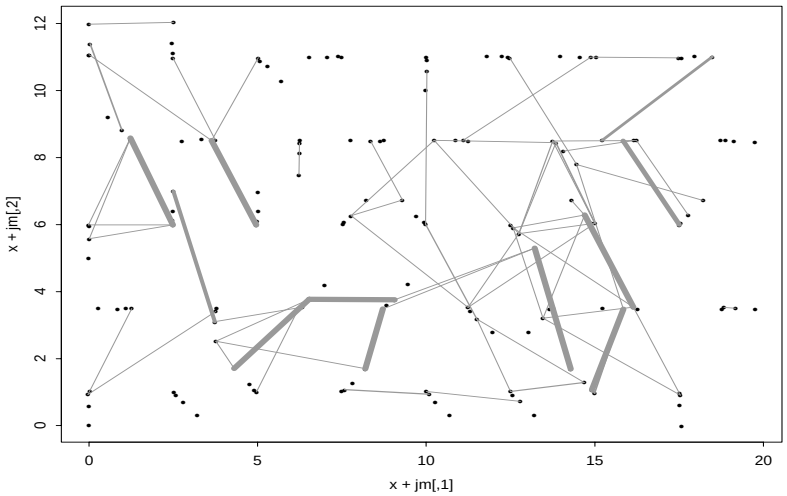

Figure 1. Identical and independent random variables at each location. The assumed variogram is constant and similar measurements are connected.

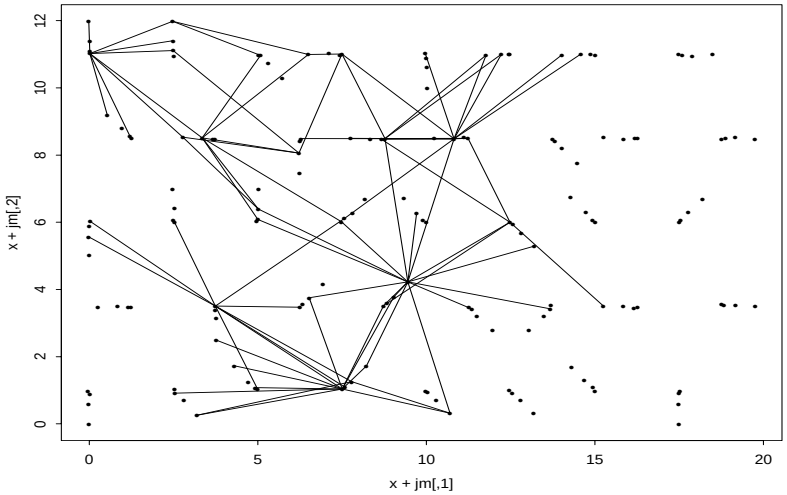

Figure 2. Identical and independent random variables at each location. The assumed variogram is constant and dissimilar measurements are connected.

3.2 Exponential Variogram I generated a random process with an exponential variogram

$$2\gamma(h) = 2 - 2\exp(-h/4),$$

and plotted the two diagnostic plots in figures 3 and 4. Figure 3 displays unusually similar

locations. Here the some of the lines are thicker than in the independent cases, and more of the locations are connected, indicating a tendency for similar locations to have similar measurements. It is figure 4 where the non-independence becomes apparent. Only two pairs of locations are displayed as being abnormally dissimilar. A sparse plot of dissimilar locations is a clear sign that there is autocorrelation at distances as large as view.range (here 5 meters).

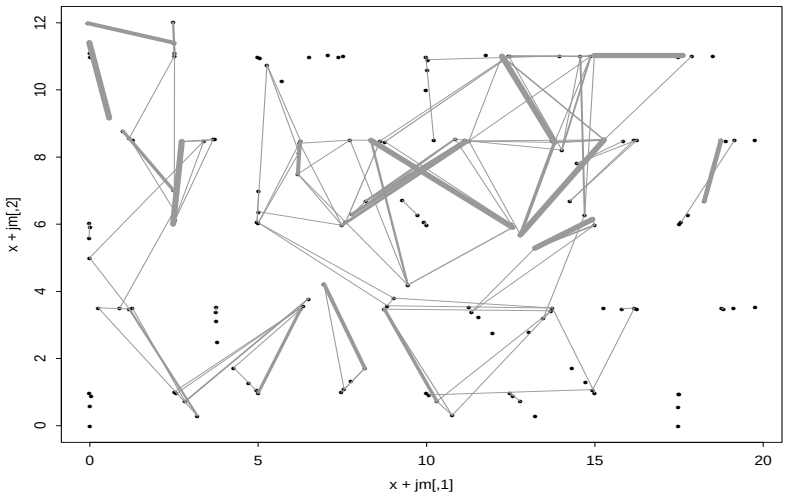

Figure 3. Simulated measurements have an exponential variogram. The assumed variogram is constant and similar measurements are connected.

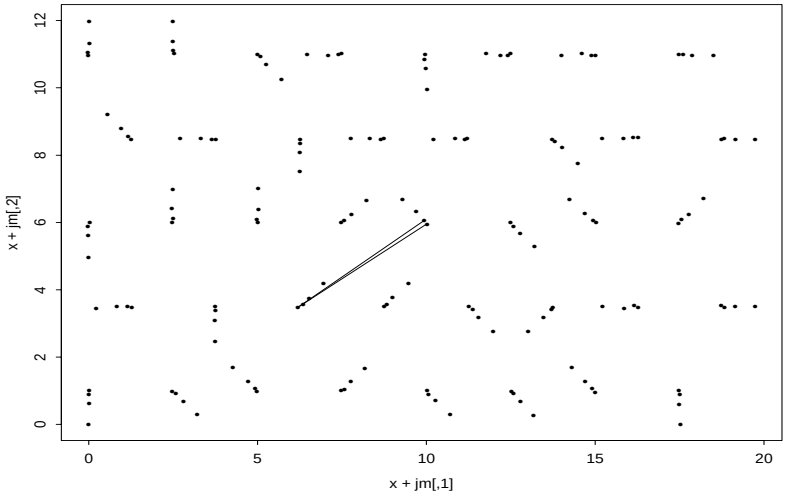

Figure 4. Simulated measurements have an exponential variogram. The assumed variogram is constant and dissimilar measurements are connected.

In contrast, when the true model (exponential variogram with sill two and range four) is assumed, figures 5 and 6 indicate a good fit. In particular, there are now relatively dissimilar pairs of measurements taken within five meters of each other. Figures 5 and 6 resemble figures 1 and 2, where the true model was a constant variogram and the test variogram was also constant. Generally, what we look for in these plots as signs of a poor fit is 1) very few connected pairs of locations 2) many pairs of locations connected by thick segments 3) many pairs of locations connected, only in one part of the plot. Also, if a location is connected to many of its neighbors, it should be investigated as a possible outlier.

Figure 7 is a plot of distance $|s_i - s_j|$ against the empirical $(z(s_i) - z(s_j))^2$ for all pairs of locations, the variogram cloud. The variability of the data at large distances is very high (for example, for distances larger than 5 meters, the mean empirical variogram is 3.4, with a variance of 15.5), and most of the pairs of locations are farther than 5 meters apart (9804 pairs out of 12720). Since the points in a variogram cloud are not associated with locations, it is impossible to notice nonstationarity or anisotropy with these plots.

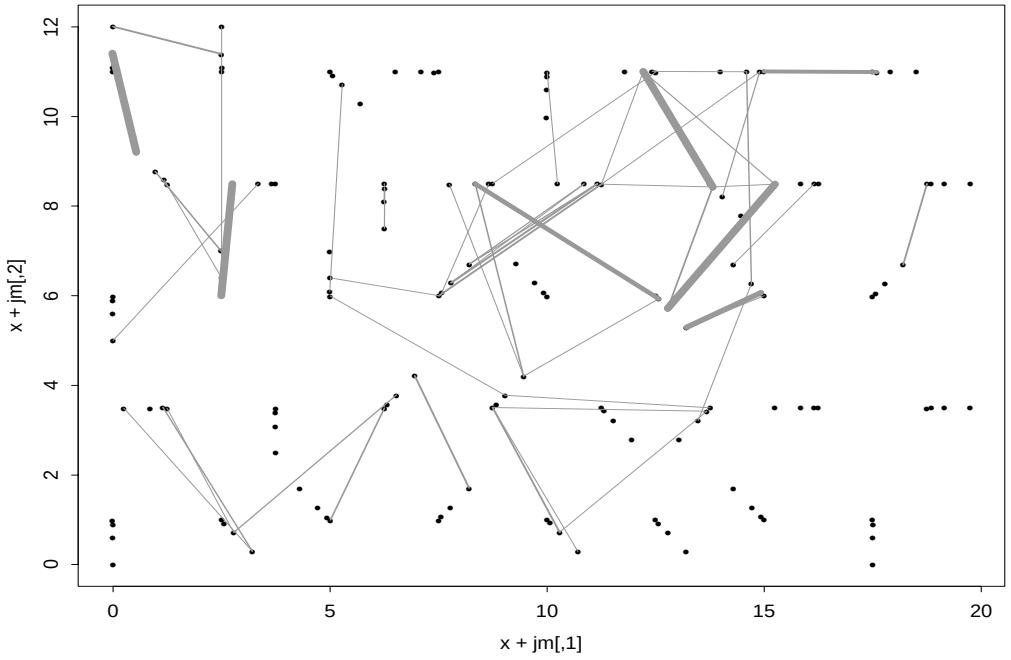

Figure 5. Simulated measurements have an exponential variogram. The assumed variogram is exponential and similar measurements are connected.

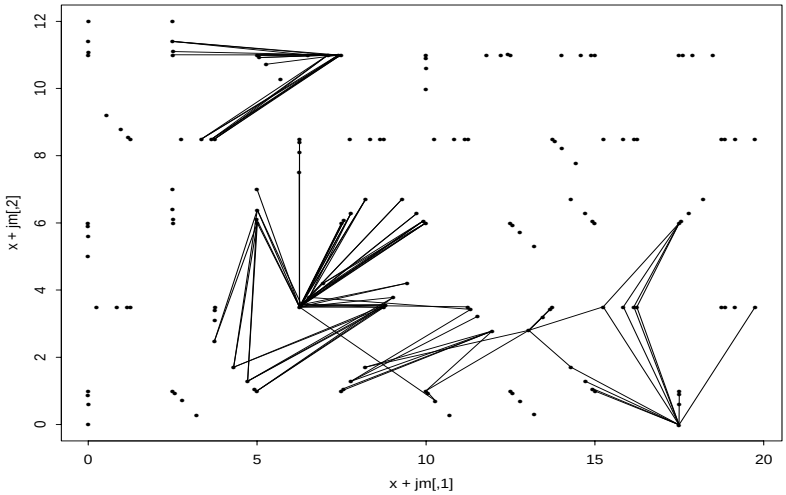

Figure 6. Simulated measurements have an exponential variogram. The assumed variogram is exponential and dissimilar measurements are connected.

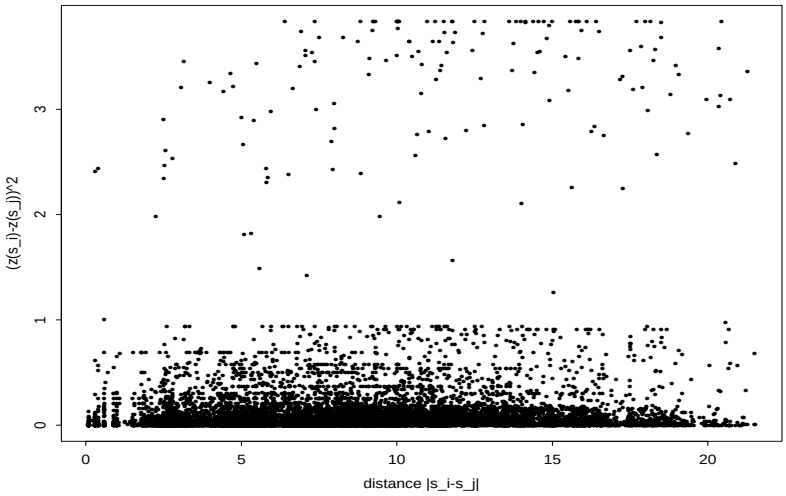

Figure 7. Variogram cloud from data simulated with an exponential variogram

3.3 Exponential Variogram with Spatial Outlier Next I generated a random process with the same exponential variogram $2\gamma(h) = 2 - 2\exp(-h/4)$, but I then replaced the smallest $z(s_i)$ with the largest $z(s_i)$. This creates an outlier in the sense that it is dissimilar from its neighbors, but that is not, in the absence of neighbor information, an unusually large value.

In this example I used the true exponential variogram, instead of the default assumption of independence. The plot flagging similar measurements was uninteresting (it looked like the corresponding plot for the random process), but in figure 8, the plot that emphasizes dissimilar pairs, the outlier is identified: it is at the location (13,6), with many segments radiating from it.

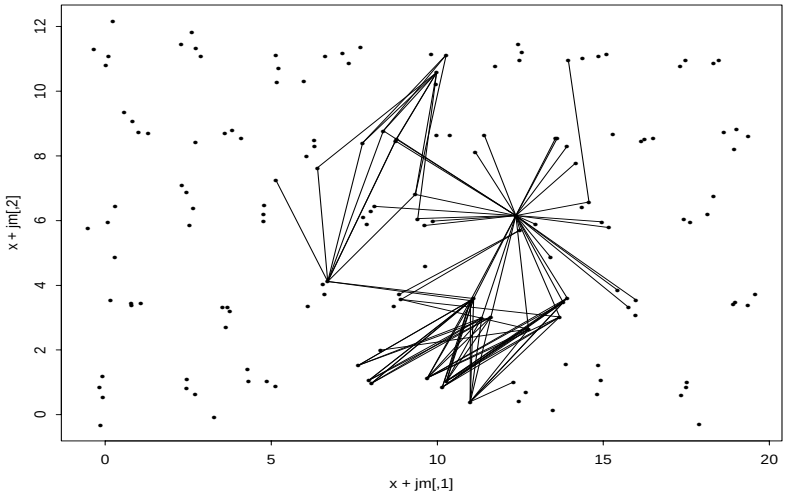

Figure 8. Simulated measurements have an exponential variogram with an outlier at (13,6). The assumed variogram is constant and dissimilar measurements are connected.

3.4. Lognormal Process Finally, I generated a lognormal random process at the same set of locations. Here, $Z(s_i) = \exp(Y(s_i))$, where $Y(s_i)$ is a intrinsically stationary random process with variogram $2\gamma(h) = 2 - 2\exp(-h/4)$. Figures 9 and 10 show the similar and dissimilar pairs, respectively. In figure 9 we see one of the signs of nonstationarity: there are patches in which almost all of the measurements are unusually similar. One such patch is near the top center of the plot. A lognormal random process tends to make the neighbors of a small value much more similar than expected, and tends to make the neighbors of a large value much more dissimilar than expected. In figure 10 this is indicated by a patch of mutually very dissimilar measurements.

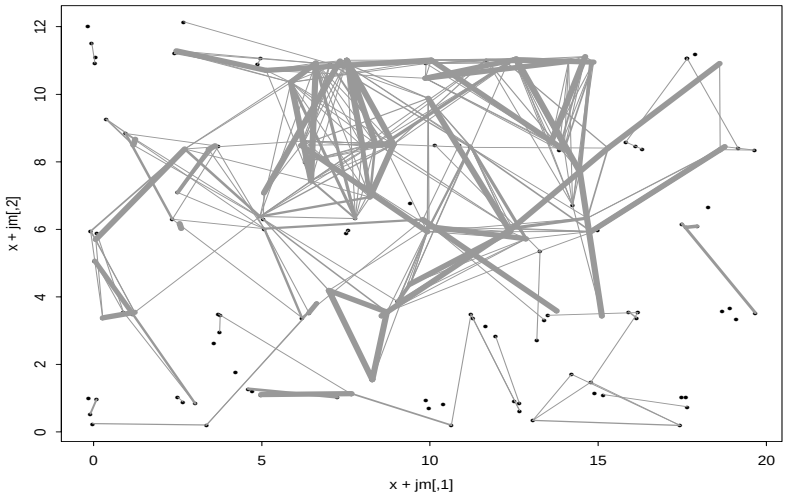

Figure 9. Simulated measurements from a lognormal process. The assumed variogram is constant and similar measurements are connected.

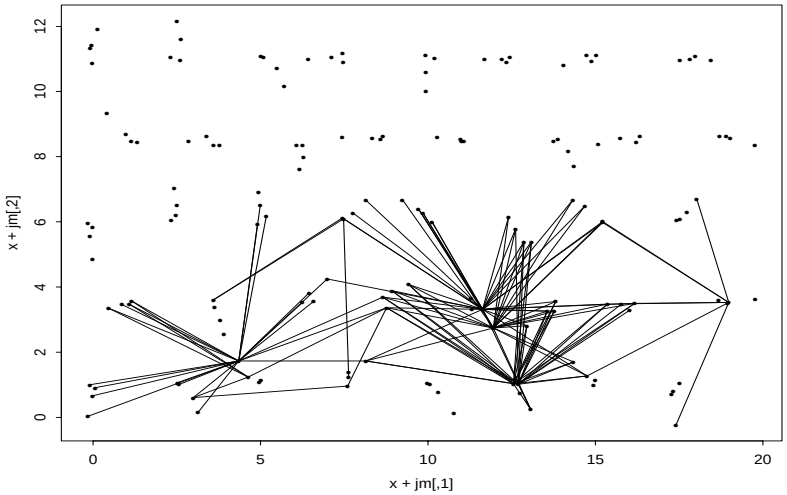

Figure 10. Simulated measurements from a lognormal process. The assumed variogram is constant and similar measurements are connected.

4. DATA ANALYSIS

Nitrate concentration, in NO_3 in $\mu g/g$, was measured at 160 locations at Turner's Field in 1991, at the W. K. Kellogg Biological Station (Gross et al., 1995). Figures 11 and 12 show the unusual similarities and unusual dissimilarities, respectively.

These figures indicate nonstationarity (perhaps a lognormal random process). In figure 11, there seems to be several outliers, and a patch of dissimilar measurements in the upper left corner. Transformation of the data might be appropriate. In this case, the measurements are left-censored, as there are 36 measurements with the value 0. With nontransformed data this presents no problem, since a pair of measurements that are truncated to zero would have been unusually similar even if not truncated. However, if I transform the data, perhaps by taking $z'(s_i) = \log(z(s_i) + c)$, where c is a small, positive constant, the left-truncated measurements will remain left-truncated, thus unusually similar, even though the true (non-truncated) values might not have been unusually similar after the transformation. This could result in a severe overestimation of the short-range correlations. The argument censoring should be set to "mark" (to give the corresponding segments a third color) or to "remove" (to delete the corresponding segments) to avoid misinterpretation.

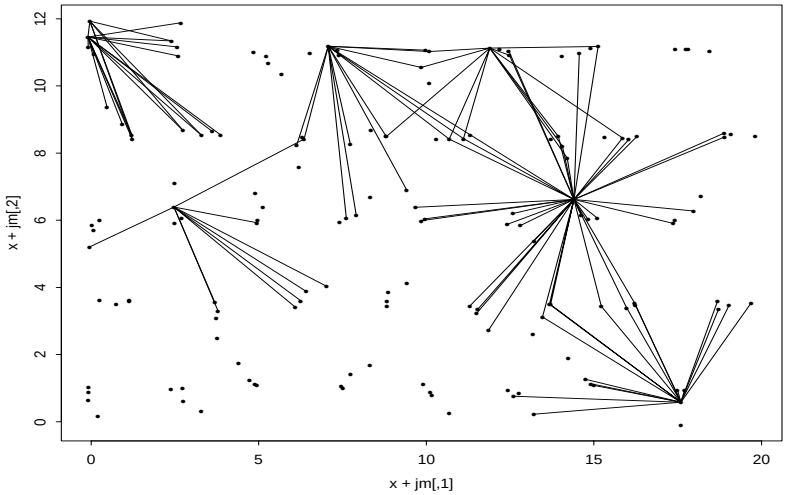

Figure 11. Nitrate concentration data, fitted with a constant variogram. Similar measurements are connected.

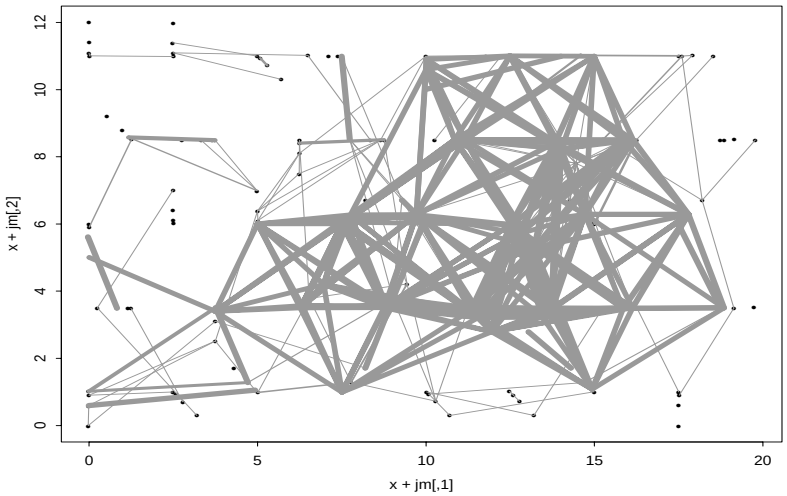

Figure 12. Nitrate concentration data, fitted with a constant variogram. Dissimilar measurements are connected.

The frontpiece is a full-color graph of this data, with all pairs of locations included (by setting **view.range**=max($|s_i - s_j|$), all lines of equal thickness (**thick.param**=0). Yellow connects unusually dissimilar locations and blue connects unusually similar locations. The patch of low and unusually similar measurements is the apparent blue patch in the middle of the plot. High-variability patches and outliers are at the center of yellow "stars".

5. CONCLUSION

The diagnostic program vario.diag allows us to evaluate the goodness of fit of a fitted variogram model. By comparing the empirical variogram $(z(s_i) - z(s_j))^2$ with the fitted

variogram $2\hat{\gamma}(s_i - s_j)$ at every pair of locations, it allows us to detect undetected short-range correlation, outliers and nonstationarity.

6. ACKNOWLEDGMENTS

I thank Kay Gross of the W. K. Kellogg Biological Station for the use of the nitrate concentration data. I also thank Andrea Corbett for conveying the nitrate concentration data to me. The associate editor and a reviewer gave many helpful comments. The Journal of Statistical Software is available at
http://www.stat.ucla.edu/journals/jss

APPENDIX

A graphical diagnostic for two-dimensional variograms.

USAGE

vario.diag((**x**, **z**, **vario**, **indep** = F, **l.threshold** = 0.05, **u.threshold** = 0.95, **thick.param** = 0, **view.range** = NA, **jitter** = 0, **twiddle** = T, **sidedness** = "both", censoring = "ignore")

REQUIRED ARGUMENTS

x	A $2 \times n$ matrix contining the x and y coordinates of n locations
z	A vector of n measurements at the locations in x

OPTIONAL ARGUMENTS

vario	A function yielding a variogram, it must be a function of two vector arguments, x1 and x2. If vario is missing, the constant variogram sill=2*var(z) is assumed (corresponding to independence).
l.threshold	Connects locations s_1 and s_2 if $(z(s_1) - z(s_2)^2/(2\gamma(s_1 - s_2)) \leq \chi^2(l.threshold)$.
u.threshold	Connects locations s_1 and s_2 if $(z(s_1) - z(s_2)^2/(2\gamma(s_1 - s_2)) \geq \chi^2(u.threshold)$.
thick.param	If thick.param>0, then line thickness will indicate how unlikely the observed values of $(z(s_1) - z(s_2))^2$ are
view.range	Locations farther apart than view.range are never connected. By default, view.range = $\max \lvert s_i - s_j \rvert$.
jitter	An IID N(0,jitter) perturbation is added to all locations when the locations are graphed. Jitter does not effect the variogram

	calculations.
twiddle	If twiddle=T, a dialog box is displayed which allows interactive modification of the arguments of the diagnostic graph
sidedness	If sidedness="both", both low and high values of $(z(s_1) - z(s_2))^2$ are indicated. Other options are sidedness="low" (low values only displayed) and sidedness="high"
censoring	This argument controls what happens when $z(s_i) = z(s_j)$. If **censoring**="ignore", these are displayed as though not censored measurements, with the same color as low values. If **censoring**="remove", these locations are not connected. If **censoring**="mark", they are connected, but displayed in a different color.

SIDE EFFECTS

Displays a plot of the graphical diagnostic. If twiddle=T, it also displays a dialog box to allow modification of the plot.

The line segments in the graph are given one of three colors, color one goes to unusually low values, color two to unusually high values, and color three to segments that are censored, when censoring="mark". For display in black and white, sidedness="low" or "high" should be selected, to differentiate between abnormally similar and abnormally dissimilar pairs.

REFERENCES

Cressie, Noel A.C. (1993) Statistics for Spatial Data, revised edition. John Wiley and Sons, Inc. New York.

Gross, K. L., Pregitzer, K.S., and Burton, A.J. (1995). "Spatial variation in nitrogen availability in three successional plant communities", *Journal of Ecology*, 83, 357-367