The Effects of Underlying Asymmetry and Outliers in data on the Residual Maximum Likelihood Variogram: A Comparison with the Method of Moments Variogram

R. Kerry[1], M. A. Oliver[2]

[1]Department of Geography, Brigham Young University, Provo, Utah, USA
Telephone: +1 (801) 422 4311
Fax: +1 (801) 422 0266
Email: Ruth_Kerry@byu.edu

[2]Department of Soil Science, University of Reading, Whiteknights, Reading, Berkshire, RG6 6DW, England
Telephone: +44 (0)118 378 6557
Fax: +44 (0) 118 378 6660
Email: m.a.oliver@reading.ac.uk

1. Introduction

Accurate maps of soil properties are required to make informed management decisions for dealing with agricultural land and contaminated sites in a site-specific way. These can be obtained by geostatistical analysis provided the data are adequate. The method of moments variogram (MoM) is usually computed, modelled and its parameters used for kriging. Webster and Oliver (1992) showed that 100-150 data are required to compute a reliable MoM variogram and these should be spaced at an appropriate interval. Kerry and Oliver (2003) showed that ancillary data could be used to determine an appropriate soil sampling interval. However, if the sampling interval indicated is large relative to the size of the study area, only a small sample size would be needed to resolve the spatial variation. This would mean that there would be too few data to compute the MoM variogram.

A possible solution to this problem is to use the Residual Maximum Likelihood (REML) variogram. It has been suggested that fewer samples are required to estimate this reliably than the MoM variogram (Pardo-Igúzquiza, 1998; Lark, 2000 and Kerry and Oliver, 2007a). However, the REML variogram assumes the process is second order stationary and it is computationally intensive. The variogram is generally sensitive to asymmetry in the distribution because it is based on variances. Kerry and Oliver (2007b) examined the effects of this on the MoM variogram. Here we describe a preliminary investigation of whether the effects of underlying asymmetry and outliers on the REML variogram are similar to those on the MoM variogram and whether the implications for the accuracy of interpolation are similar.

2. Methods

2.1. Simulation of Data

Four hypothetical fields of data, 200 m by 200 m with points spaced at 20-m intervals (100 points) and a standard normal distribution, $N(0,1)$, were produced by simulated annealing (Deutsch and Journel, 1992) from spherical variograms with a range of 75m, a sill of 1 and nugget:sill ratios of 0, 0.25, 0.50 and 0.75.

Data with underlying asymmetric distributions were generated from the above fields following the procedure of Lark (2000) to give positive skewness coefficients of 0.75, 1.5 and 5.0. A constant of 4 was added to the simulated normal scores to make all values positive and each value was raised to the power α (α = 2.05, 3.35 and 37.0) to create the desired degree of asymmetry in the data. These data were then standardized to zero mean and unit variance.

The normally distributed data are realizations of a primary Gaussian process. To contaminate these data with outliers at a proportion (0.05) of the sites, contaminants were drawn from random, normally distributed populations with different means, $N_C(1,1)$, $N_C(1.25,1)$, $N_C(1.5,1)$…. $N_C(10,1)$, and added to the original values of the primary process to give skewness coefficients of 0.5, 1.0, 1.5, 2.0 and 3.0.

2.2. Geostatistical Methods

The REML variograms were computed on the simulated data by Pardo-Igúzquiza's (1997) MLREML program. The spherical and exponential functions were computed and the model with the smallest negative log-likelihood function (NLLF) was taken as the most appropriate. The simplex method was used to minimize the NLLF.

Cross-validation was used to assess quantitatively the performance of each variogram. The method involved removing each datum in turn and then kriging at the point with the model parameters and neighbouring data points. The mean squared error (MSE) and median squared deviation ratio (MeSDR) from cross-validation were calculated. Lark (2000), recommended the MeSDR to determine the best model for kriging with skewed data because it is not affected by asymmetry. It was determined by dividing the squared errors by the kriging variances for each data point, and then ordering the values; the middle value was taken as the MeSDR. When the correct model is used for kriging, the MeSDR should be close to 0.455.

3. Results

3.1. Underlying Asymmetry

Figure 1 shows that underlying asymmetry does not have a consistent affect on the sill height of REML variograms and this was also the case for MoM variograms (Kerry and Oliver, 2007b). However, the variation in the form of the variogram and its departure from the generating function are different for the REML and MoM variograms. There are marked differences in the variograms for all nugget:sill ratios when the skewness is 5. For coefficients of skewness of 0.75 and 1.5 departures from the generating functions are moderate for nugget:sill ratios of 0 and 0.25, but are negliable for ratios of 0.5 and 0.75.

Table 1 shows that there is a general increase in the difference between the parameters of the fitted variograms and the generating function as skewness increases for both REML and MoM variograms. The nugget:sill ratios and sills for the MoM variograms tend to be closer to those of the generating function, whereas for the REML variograms it is the range that is closer. Figure 1 and Table 1 show that when the skewness coefficient is large the form of the experimental variogram becomes erratic and is difficult to model, whereas there are still parameters for the REML variograms. Table 2 shows that for both REML and MoM variograms the MSEs become larger as asymmetry increases and the MeSDRs depart more from 0.455. The MSE and MeSDR values are similar for both types of variogram, but for all skewness coefficients >0 the MeSDR is closer to 0.455 for MoM and the MSEs are smaller for REML for skewness coefficients of 1.5 and 5.0.

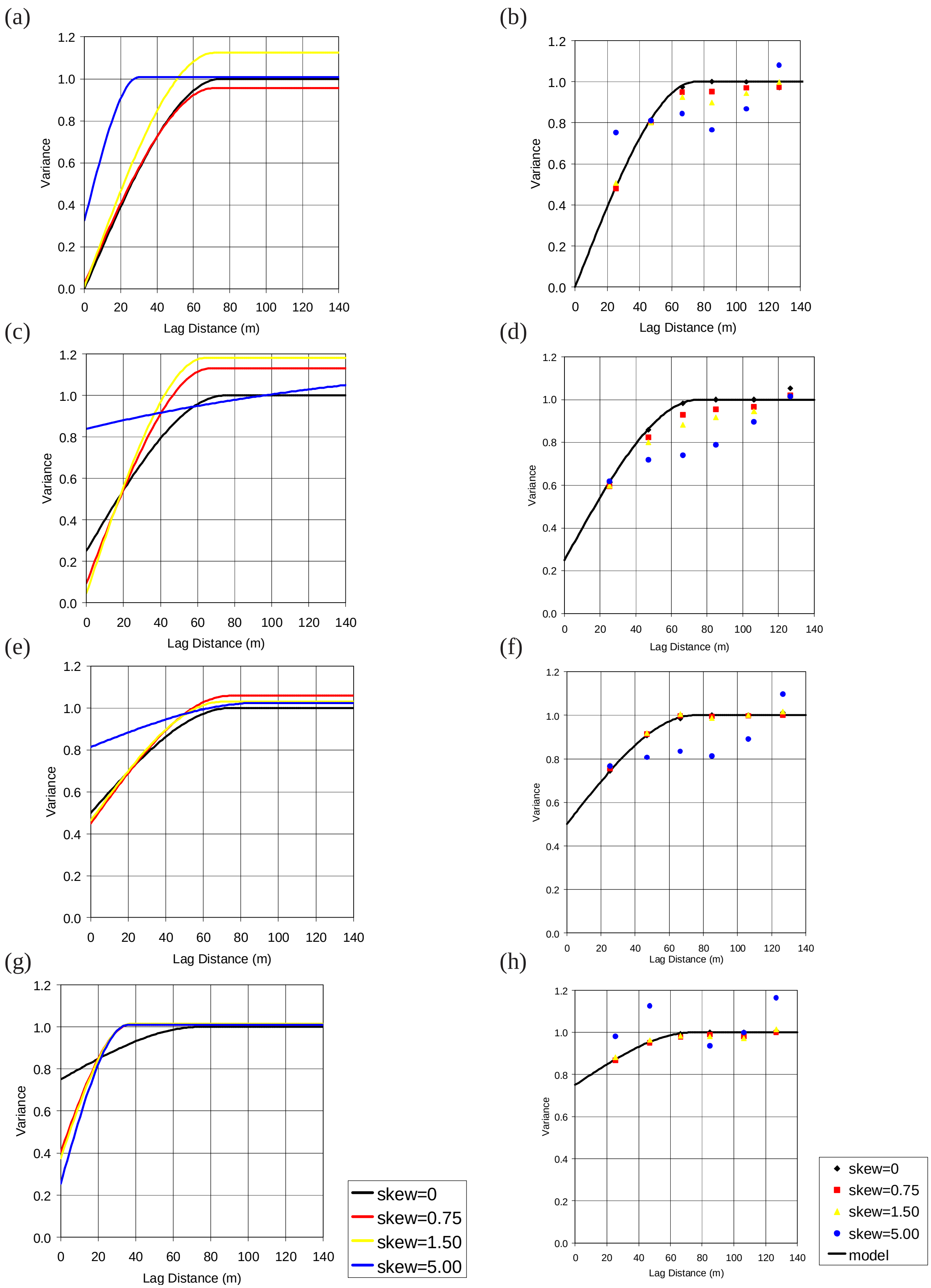

Figure 1. Residual Maximum Likelihood (REML – a, c, e and g) and Method of Moments (MoM – b, d, f and h) variograms computed on data with different levels of underlying asymmetry and nugget:sill ratios of 0 (a and b), 0.25 (c and d), 0.50 (e and f) and 0.75 (g and h).

Table 1. Differences between the parameters of the target variogram and those calculated by residual maximum likelihood and the method of moments using data with underlying asymmetry

Skew due to underlying	Nugget of target	Difference between target and actual variogram parameters					
		REML			MoM		
asymmetry	variogram	Sill	Nugget:sill	Range	Sill	Nugget:sill	Range
0.75	0 %	-0.043	0.027	-3.381	-0.039	0.032	-9.680
1.50	0 %	0.124	0.008	-3.484	-0.053	0.025	10.600
5.00	0 %	0.009	0.323	-44.934	*	*	*
0.75	25 %	0.131	-0.166	-7.914	-0.017	0.035	25.500
1.50	25 %	0.182	-0.211	-10.622	0.020	-0.044	39.030
5.00	25 %	0.239	0.427	484.353			
0.75	50 %	0.058	-0.075	-0.894	-0.003	0.038	-8.533
1.50	50 %	0.029	-0.050	-5.137	0.002	0.065	-7.680
5.00	50 %	0.024	0.295	14.920	*	*	*
0.75	75 %	0.009	-0.349	-38.141	-0.007	-0.204	-15.690
1.50	75 %	0.014	-0.379	-38.044	-0.008	-0.235	-23.430
5.00	75 %	0.009	-0.498	-39.183	*	*	*

* linear or power functions fitted or experimental variogram was pure nugget

Table 2. Mean squared errors and median squared deviation ratios, from cross-validation using residual maximum likelihood (REML) and method of moments (MOM) variogram model parameters from data (0% nugget generating variogram) with underlying asymmetry and asymmetry caused by randomly located outliers

Skewness of		REML		MoM	
Data		MSE*	MeSDR*	MSE*	MeSDR*
Underlying asymmetry	0	0.369	0.394	0.369	0.394
	0.75	0.355	0.301	0.353	0.425
	1.5	0.435	0.206	0.438	0.239
	5.0	1.037	0.0007	1.070	0.002
Outliers	0	0.369	0.394	0.369	0.394
	0.5	0.992	0.215	0.982	0.199
	1.0	1.395	0.199	1.373	0.188
	1.5	1.900	0.170	1.854	0.185
	2.0	2.603	0.144	2.605	0.148
	3.0	5.626	0.112	5.528	0.082

*A constant of 4 was added to values in each original data set so that all values were just positive for the logarithmic transformation.

MSE is the mean squared error

MeSDR is the median squared deviation ratio

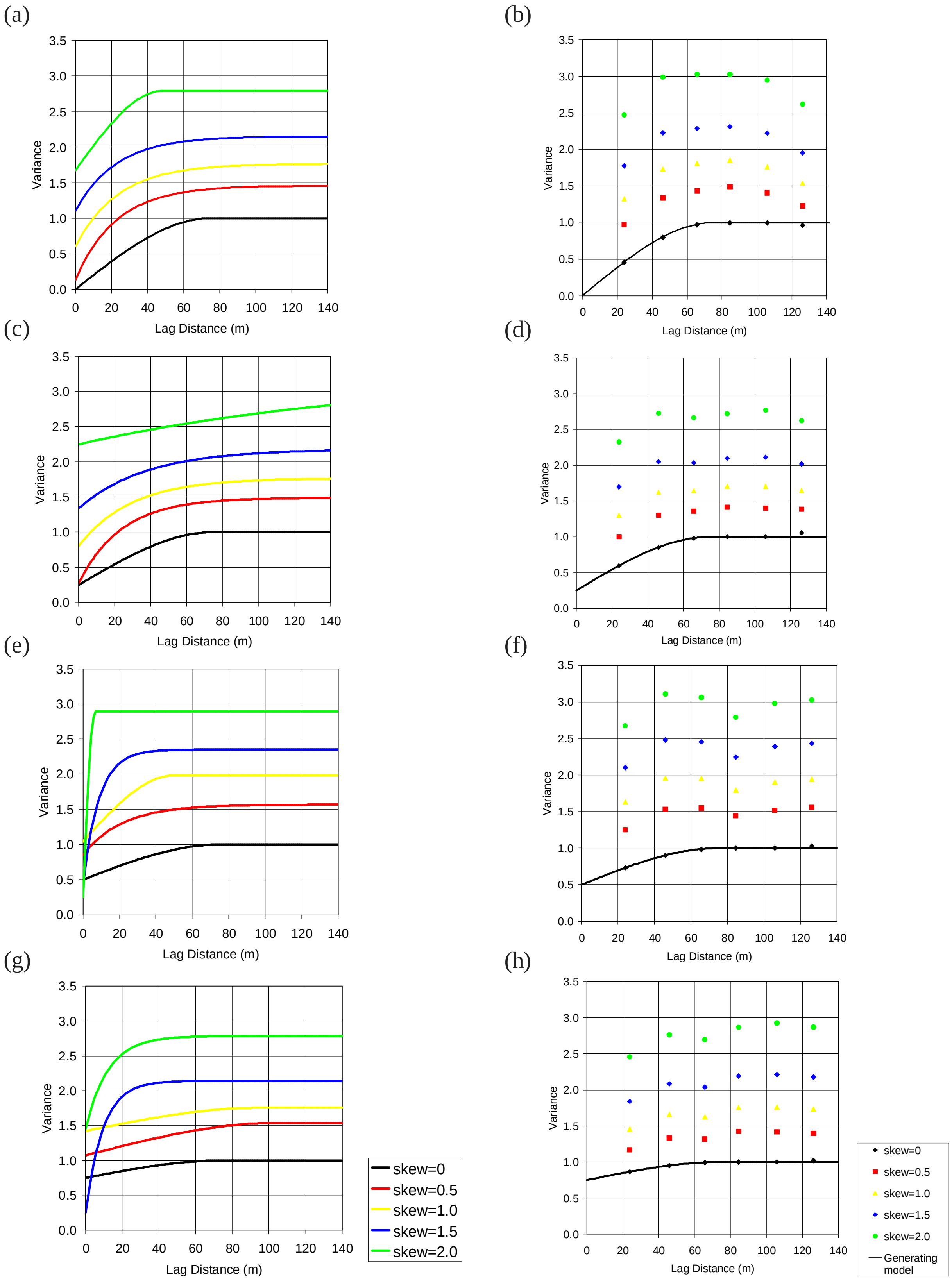

Figure 2. Residual Maximum Likelihood (REML – a, c, e and g) and Method of Moments (MoM – b, d, f and h) variograms computed on data with different levels of skewness caused by outliers and nugget:sill ratios of 0 (a and b), 0.25 (c and d), 0.50 (e and f) and 0.75 (g and h).

3.2. Outliers

Figure 2 shows that asymmetry caused by outliers has a more serious effect on the form of the variogram than underlying asymmetry in that the departures from the generating function are greater (compare scale of y axis in Figs. 1 and 2). The sills of the MoM and REML variograms move up the y axis as skewness increases. This is usually associated with a progressive increase in the nugget variance (Kerry and Oliver, 2007c), but this is not always so for the REML variograms, especially when the nugget:sill ratio of the generating function and skewness are large. This suggests that the REML variogram might be more appropriate for interpolation in such situations. Table 3 shows, however, that the departures of the parameters of the fitted models from those of the generating function are generally greater for the REML than MoM variograms. A comparison of Tables 1 and 3 shows that departure of the REML and MoM parameters from those of the generating model are an order or two of magnitude larger when the skewness is caused by outliers. Table 2 shows that the MSE and MeSDR values are similar for both types of variogram, but the MSEs for the REML variograms are a little larger. The MSEs are larger and the MeSDR values less appropriate when skewness is caused by outliers compared with underlying asymmetry (compare results for skewness 1.5 for underlying asymmetry and outliers in Table 2).

Table 3. Differences between the parameters of the target variogram and those calculated by residual maximum likelihood and the method of moments using data with asymmetry caused by outliers

	Nugget	Difference between target and actual variogram parameters					
Skew due	of target	**REML**			**MoM**		
to outliers	variogram	Sill	Nugget:sill	Range	Sill	Nugget:sill	Range
0.5	0 %	0.458	0.091	-6.885	0.394	0.261	-18.000
1.0	0 %	0.763	0.341	-3.801	0.740	0.299	-24.500
1.5	0 %	1.149	0.512	-7.473	1.270	0.432	-20.660
2.0	0 %	1.786	0.600	-26.934	1.996	0.459	-26.400
3.0	0 %	4.399	0.066	-53.057	4.895	0.623	-34.210
0.5	25 %	0.488	-0.064	-3.438	0.393	0.051	3.300
1.0	25 %	0.766	0.205	12.561	0.678	0.115	-16.690
1.5	25 %	1.180	0.364	37.926	1.069	0.150	-22.800
2.0	25 %	2.245	0.443	439.868	1.724	0.243	-29.720
3.0	25 %	4.758	0.565	483.270	4.387	0.304	-36.230
0.5	50 %	0.569	0.042	-10.083	0.519	-0.125	-35.300
1.0	50 %	0.983	0.027	-25.393	-1.000	-0.500	-75.000
1.5	50 %	1.351	-0.326	-48.797	1.178	-0.041	-37.300
2.0	50 %	1.894	-0.414	-68.260	2.460	0.078	-39.300
3.0	50 %	5.049	-0.469	-70.426	-1.000	-0.500	-75.000
0.5	75 %	0.539	-0.053	29.016	0.410	-0.009	38.600
1.0	75 %	0.755	0.061	19.468	0.752	0.005	43.400
1.5	75 %	1.141	-0.632	-46.828	1.196	0.021	47.200
2.0	75 %	1.786	-0.226	-38.208	1.898	0.054	35.000
3.0	75 %	4.868	-0.615	-53.854	5.225	0.099	41.600

4. Conclusions

It seems possible to compute a reliable variogram with fewer samples by REML than by MoM and this could potentially reduce the cost of soil mapping. The results suggest that the REML variogram is affected similarly, but not identically, to the MoM variogram by asymmetric data. Both types of variogram are more stable with underlying asymmetry and this translates into more reliable estimates than where data are contaminated by outliers. The REML variograms had slightly larger MSEs than the MoM ones when data were contaminated by outliers, but the MeSDRs were sometimes more appropriate for the former. The similarity in behaviour of the two variograms suggests that REML might be a viable alternative to the MoM variogram even when data are skewed. Further investigation with many more realizations is necessary to corroborate these findings. Also, the influences of the spatial configuration of outliers, the size of data set and various standard data transformations on how asymmetry affects the REML compared to the MoM variogram need to be investigated. The development of robust estimators of the REML variogram should also be examined.

5. Acknowledgments

All data for MoM variograms has been transcribed from Kerry and Oliver (2007b) and Kerry and Oliver (2007c).

6. References

Kerry, R. and Oliver, M.A. 2003. Variograms of ancillary data to aid sampling for soil surveys. *Precision Agriculture*, 4:253-270.

Kerry, R and Oliver, M. A. 2007a. Sampling requirements for variograms of soil properties computed by the method of moments and residual maximum likelihood. *Geoderma*. Accepted.

Kerry, R. and Oliver, M.A. 2007b. Determining the Effect of Skewed Data on the Variogram. I. Underlying Asymmetry. *Computers & Geosciences*. Accepted.

Kerry, R. and Oliver, M.A. 2007c. Determining the Effect of Skewed Data on the Variogram. II. Outliers. *Computers & Geosciences*. Accepted.

Lark, R. M. 2000. Estimating variograms of soil properties by the method-of-moments and maximum likelihood. *European Journal of Soil Science,* 51: 717-728.

Pardo-Igúzquiza, E., 1997. MLREML: a computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood. *Computers & Geosciences*, 23: 153–162.

Pardo-Igúzquiza, E. 1998. Maximum likelihood estimation of spatial covariance parameters. *Mathematical Geology*, 30: 95-107.

Webster, R., Oliver, M.A., 1992. Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, 43: 177–192.