ELSEVIER

Journal of Structural Geology 26 (2004) 595–600

JOURNAL OF STRUCTURAL GEOLOGY

www.elsevier.com/locate/jsg

Bootstrapped spatial statistics: a more robust approach to the analysis of finite strain data

Malay Mukul*, Debjani Roy, Shibasis Satpathy, V. Anil Kumar

CSIR Centre for Mathematical Modelling and Computer Simulation, Bangalore 560037, India

Received 3 February 2002; received in revised form 20 August 2003; accepted 20 August 2003

Abstract

Standard spatial statistics involves exploratory data analysis (EDA) and the computation of a semi-variogram prior to kriging from spatial data such as the finite strain data from a thrust sheet. However, the main shortcoming of this method is that uncertainties in the computation of EDA and the semi-variogram parameters cannot be estimated; standard EDA allows the computation of the uncertainties associated with only the sample mean. We use resampling statistics to overcome this and use previously published data from the Sheeprock thrust sheet in the Sevier fold-and-thrust belt in western USA to illustrate Bootstrapped spatial statistics. The use of Bootstrapped spatial statistics in place of standard spatial statistics also simplifies the computation of the best semi-variogram resulting in the computation of a more robust semi-variogram. Kriging estimates obtained using the Bootstrapped semi-variogram indicate that the results are practically insensitive to the uncertainty associated with the estimation of parameters of the semi-variogram used in the Sheeprock example.
© 2003 Elsevier Ltd. All rights reserved.

Keywords: Bootstrapped spatial statistics; Finite strain data; Exploratory data analysis; Variogram; Kriging

1. Introduction

The standard spatial data analysis (e.g. Matheron, 1963; Cressie, 1993; Mukul, 1998) starts with exploratory data analysis (EDA), which typically involves measures of central tendency (such as mean, median, mode), measures of dispersion (such as standard deviation from mean etc.), symmetry (skewness) and peakedness (kurtosis) and their graphical representation to test the behaviour of the data and the underlying structure. Computation of a semi-variogram, which is the key to the spatial statistics method, is carried out next to decide how well a measurement represents another location at a specific distance and direction. The final step in the process is spatial prediction and kriging to estimate the studied spatial variable from unsampled locations in the sampled area using a weighted linear combination of the available samples and the computed semi-variogram in an unbiased manner after smoothing of 'noise' in the data set and declustering it. However, the uncertainties associated with the EDA (except the sample mean) and variogram parameters cannot be ascertained using the standard spatial statistics method. Moreover, the best semi-variogram for a given spatial data set can only be ascertained by comparison of errors between measured and estimated values of the studied spatial variable at different reference points in the population by cross-validation after kriging (Mukul, 1998). This requires that the entire kriging analysis be carried out during each attempt at estimation of the semi-variogram; this involves considerable computational time and labour in the estimation of the best possible semi-variogram and is, therefore, not an efficient approach.

This paper uses resampling statistics (Simon, 1969), later termed the Bootstrap (Efron, 1979), and involves creation of resampled Bootstrap sample sets from a measured base dataset and treating them as multiple data sets (Diaconis and Efron, 1983; Efron and Tibshirani, 1991) to compute uncertainties in the EDA and semi-variogram parameters. It also makes the semi-variogram computation more robust and efficient, providing a distinct improvement over semi-variogram computation using standard spatial statistics.

We illustrate this approach using previously published X/Z strain data collected from 56 locations in a deformed volume of rock (Mukul, 1998) from the Sheeprock thrust sheet (Table 1) in the Provo salient of the Sevier fold-and-thrust

* Corresponding author. Tel.: +91-80-505-1928; fax: +91-80-522-0392.
E-mail address: mlym@cmmacs.ernet.in (M. Mukul).

0191-8141/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsg.2003.08.011

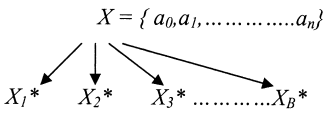
Table 1
Base data values: Sheeprock strain data (Mukul, 1998)

X (m)	*Y* (m)	*X*/*Z* ratio	*X*-orientation	Sequence #
810.000	3470.000	1.245000	300.0000	1
975.000	3640.000	1.180000	341.9900	2
1150.000	3830.000	1.325000	92.7500	3
1880.000	4600.000	1.302000	90.2000	4
2185.000	4850.000	1.233000	306.7300	5
125.000	2340.000	1.305000	322.5100	6
290.000	2530.000	1.301000	005.7000	7
495.000	2715.000	1.339000	147.3400	8
665.000	2920.000	1.375000	282.6900	9
1000.000	3295.000	1.225000	233.4000	10
1160.000	3495.000	1.222000	311.6200	11
1325.000	3655.000	1.163000	224.1400	12
1495.000	3855.000	1.067000	17.0600	13
2200.000	4560.000	1.285000	64.7000	14
2395.000	4760.000	1.127000	288.2000	15
335.000	2170.000	1.319000	180.4600	16
525.000	2370.000	1.220000	242.9000	17
695.000	2545.000	1.276000	244.9700	18
865.000	2745.000	1.260000	352.8000	19
1355.000	3300.000	1.226000	316.2300	20
1505.000	3495.000	1.351000	48.6100	21
1715.000	3670.000	1.127000	237.0600	22
1865.000	3855.000	1.230000	61.7200	23
2030.000	4055.000	1.270000	43.0700	24
2205.000	4210.000	1.400000	345.0600	25
2370.000	4375.000	1.318000	208.0200	26
550.000	2005.000	1.250000	173.8900	27
865.000	2395.000	1.196000	200.9200	28
1040.000	2600.000	1.357000	197.4500	29
1700.000	3325.000	1.243000	271.9700	30
2040.000	3695.000	1.149000	98.6000	31
2225.000	3870.000	1.303000	194.0400	32
2405.000	4055.000	1.219000	16.0300	33
2560.000	4165.000	1.143000	319.5100	34
725.000	1840.000	1.383000	152.9000	35
885.000	2040.000	1.301000	340.5500	36
1220.000	2400.000	1.356000	255.6400	37
2035.000	3340.000	1.160000	58.9900	38
925.000	1665.000	1.308000	123.2100	39
2065.000	2970.000	1.284000	318.3700	40
2420.000	3345.000	1.372000	154.1200	41
1115.000	1490.000	1.482000	70.1300	42
1265.000	1700.000	1.269000	131.9600	43
1440.000	1890.000	1.308000	199.9100	44
1775.000	2275.000	1.352000	71.8200	45
2410.000	2980.000	1.207000	181.4100	46
1465.000	1525.000	1.301000	237.7400	47
1625.000	1740.000	1.299000	188.6300	48
2785.000	3010.000	1.316000	324.2500	49
2965.000	3170.000	1.315000	001.5200	50
1525.000	1175.000	1.284000	289.6200	51
1655.000	1365.000	1.266000	317.4500	52
1985.000	1755.000	1.344000	341.0200	53
2645.000	2460.000	1.604000	257.3100	54
2815.000	2645.000	1.533000	43.0300	55
2980.000	2820.000	1.266000	141.5300	56

belt (Fig. 1) in western United States. The strain ratios used in this study are an example of spatial variables, which require quantification of both location and magnitude. This approach can work equally well with tensor-averaged strain data (Brandon, 1995) where the Bootstrapped variogram analysis could be applied to each of the six independent components of the observed tensor.

2. Bootstrapped spatial data analysis

We estimate the semi-variogram prior to kriging by introducing a couple of extra steps into the standard procedure for the analysis of spatial data as carried out in Mukul (1998) and Mukul and Mitra (1998). First, the base data set (Table 1) is resampled using the Bootstrap resampling technique (Diaconis and Efron, 1983; Efron and Tibshirani, 1991). The name 'Bootstrap' refers to use of the original data set X to generate new data sets X^* (Efron and Tibshirani, 1991). This involves creating artificial data subsets by randomly selecting elements from the original data set. The selection is carried out with replacement, i.e. the selected elements are replaced back into the base data set and are available for the next selection. Each subset, therefore, has the same number of elements as the original set but is different because an individual element can be chosen once, more than once, or not at all. Therefore, if the original data set X contains $X = \{a_0, a_1, a_2,\ldots,a_n\} = \{a_i, i = (0,n)\}$, the Bootstrap sample made from X denoted by $X^* = \{a_0^*, a_1^*, a_2^*,\ldots, a_n^*\} = \{a_i^*, i = (0,n)\}$, could be a possible selection sequence in creating X^* from X. Each a_i^* is one of the original a_i values, randomly selected (with replacement). Thus, the selection sequence could read $a_0^* = a_6$, $a_1^* = a_0$, $a_2^* = a_1$, etc. The Bootstrap random sampling algorithm described above allows creation of Bootstrap data files $X_1^*, X_2^*, X_3^*,\ldots,X_B^*$. Thus:



A pseudo random number generator (PRNG) was used to randomly sample from the base data set (Table 1). In a given bootstep (B), the selection process was iterated 56 times to get a Bootstrap data set containing the same number of data elements as the original base data file (Table 1). A maximum of 500 bootsteps were used to generate 500 Bootstrap data files.

The PRNG was chosen, in SGI IRISX6.5 Operating System, such that its period was much larger than the total number of random numbers in the sequence and both low and high order bits of the random number expressed high randomness. We used a non-linear additive feedback random number generator that generated pseudo-random numbers in the range from 0 to $2^{31} - 1$. The period or cycle length of the generator is 34359738352 (16 times of $2^{31} - 1$) and is much higher than the total random numbers required to generate the random sequence (56×500 for 500 random file each of size 56 lines). Further, the non-linear

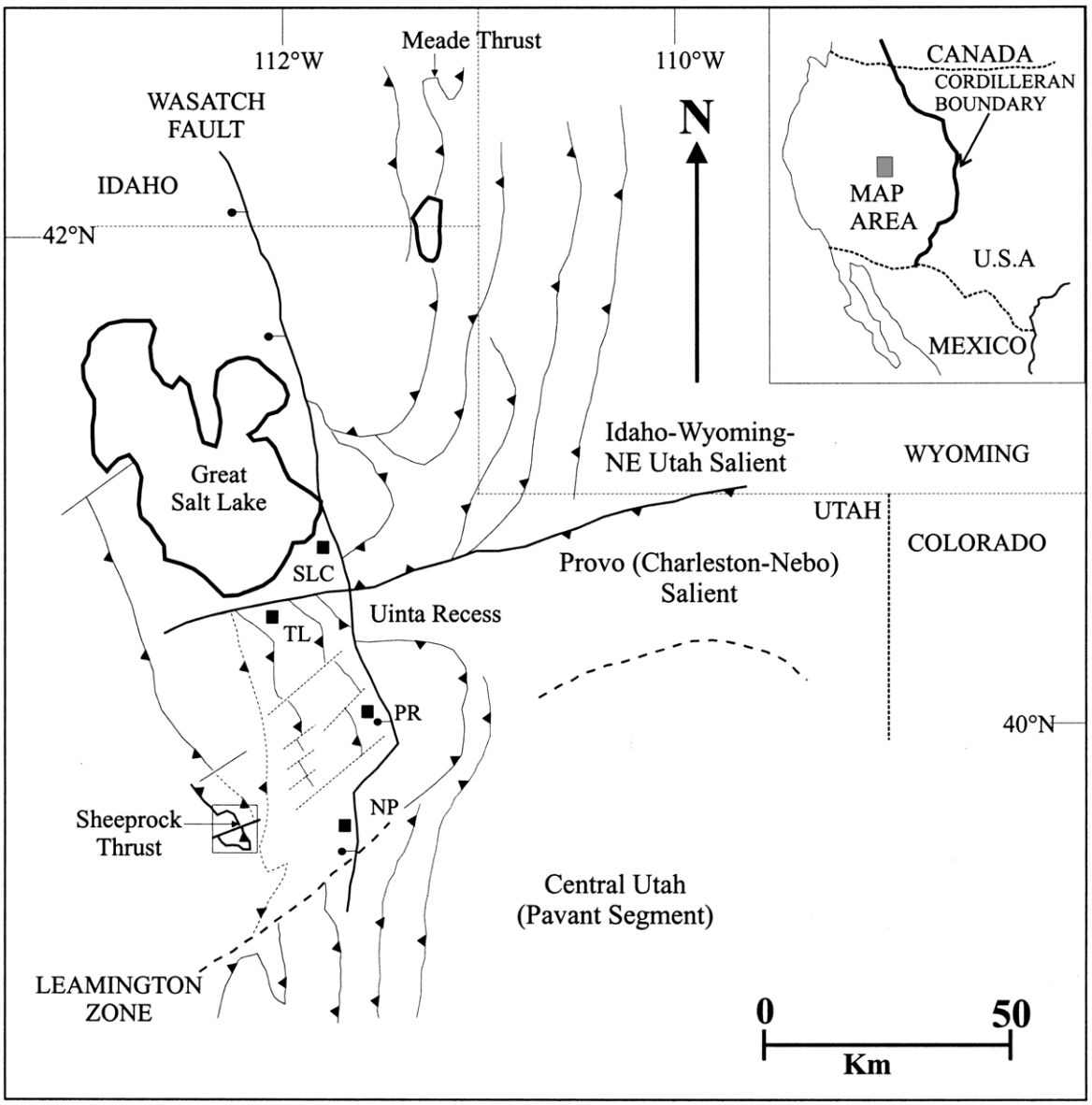


Fig. 1. Simplified map of the Sevier fold-and-thrust belt in Idaho–Wyoming and northern Utah. The Sheeprock thrust sheet, from which the data used in this study was collected, has been boxed. Symbols: SLC, Salt Lake City; TL, Tooele; PR, Provo; NP, Nephi.

additive feedback algorithm generated numbers with high randomness in both low and high order bits. This eliminated the risk of sequence repetition within the entire sequence and consequent period sampling of same data sequence. System time was used as the initial seed for the generator.

2.1. Bootstrapped exploratory data analysis (BEDA)

BEDA aims to improve EDA computations by estimating the uncertainties associated with each EDA statistic as opposed to only the mean. Given a data set $X = \{a_0, a_1,\ldots,a_n\}$, and $\bar{a}$ to be its mean, the standard error of mean $\bar{a}$ or its root-mean-square error during EDA is computed by:

$$\mathrm{Se}(\bar{a}) = \left\{\sum_{i=1}^{n} \frac{(a_i - \bar{a})^2}{n(n-1)}\right\}^{1/2} \quad (1)$$

BEDA statistics can now be computed for each of the files X_1^*, $X_2^*,\ldots,X_B^*$. For any BEDA statistic $t(X_b^*)$; $b = 1, 2,\ldots,B$), we can compute the standard error of the corresponding BEDA statistic $t(X_b^*)$ by:

$$\left\{\sum_{b=1}^{B} \left[t(X_b^*) - \bar{t}\right]^2 / B(B-1)\right\}^{1/2} \quad (2)$$

where $\bar{t}$ is the average of the B Bootstrap replications $t(X_b^*)$ given as:

$$\frac{1}{B}\sum_{b=1}^{B} x_b^* \tag{3}$$

Here $t(X_b^*)$ could be mean, median, mode, skewness or peakedness in the context of BEDA. It has also been established (Efron and Tibshirani, 1991) that a stable estimate of Bootstrap standard error requires $B \sim 200$ and, therefore, at least 200 Bootstrap files need to be analysed during BEDA.

BEDA results for the Sheeprock strain data (Table 1) are given in Table 2. The errors fluctuated for initial bootsteps (B) and stabilized beyond $B \geq 400$. Results for $B = 500$ have been tabulated in Table 2.

2.2. Bootstrapped semi-variogram computation

One of the most frequently cited problems with kriging is the necessity of using an estimated model for the semi-variogram function (Hardy, 1977; Armstrong, 1984; Philip and Watson, 1986). Brooker (1986) pointed out that the kriging variance is "robust to most errors likely to be made in semi-variogram selections" except that it may be sensitive to incorrect choice of the nugget value. This was also confirmed by Armstrong (1984). However, there still was a need to quantify the uncertainty associated with the semi-variogram estimation. Chung (1984), Dubrule (1983) and Campbell (1988) explored the use of resampling techniques in geostatistics concluding that the distribution of resulting Bootstrap estimates of the semi-variogram parameters should approximate the corresponding true distribution and thus provide a non-parametric way to estimate uncertainties associated with the original estimator.

To estimate the Bootstrapped semi-variogram for the Sheeprock strain data (Table 1), we need to determine the uncertainties or standard errors associated with the parameters a and c of the exponential semi-variogram function in Eq. (6). Given $X = \{a_0, a_1,\ldots,a_n\}$ and Bootstrap samples X_1^*, $X_2^*,\ldots,X_B^*$ from X, we estimated the semi-variogram function (Eq. (6)) from each X_b^*, $b = 1, 2,\ldots,B$, where $B \leq 500$, thereby computing B semi-variogram functions, each with their own range (a) and sill (c) parameters. For a given B, we computed the average of the a^* and c^* values following the appropriate form of Eq. (3):

$$\bar{a}^* = \frac{1}{B}\sum_{b=1}^{B} a_b^* \text{ or } \bar{c}^* = \frac{1}{B}\sum_{b=1}^{B} c_b^* \tag{4}$$

The standard error of a^* and c^* for a given B was then computed from an appropriate form of Eq. (1):

$$\left\{\sum_{b=1}^{B}(a_b^* - \bar{a}^*)^2/B(B-1)\right\}^{1/2} \text{ or } \left\{\sum_{b=1}^{B}(c_b^* - \bar{c}^*)^2/B(B-1)\right\}^{1/2} \tag{5}$$

Results of the above computation for $B = 500$ are given in Table 3. The sill value c becomes stable for $B \geq 300$. The range value, a, was more variable and was computed to be 759.32 ± 33.99 for $B = 500$. Given that the best semi-variogram for the Sheeprock data, based on comparison of errors between measured and estimated X/Z ratios at different reference points in the population by cross-validation after kriging (Mukul, 1998), was found to be an exponential semi-variogram ($\gamma(|h|)$):

$$\gamma(|h|) = c\left[1 - e^{\frac{-3|h|}{a}}\right] \tag{6}$$

where $c = 0.008$ and $a = 750$, and h the separation vector between sample locations, the Bootstrapped semi-variogram was, therefore, re-computed to be:

$$\gamma(|h|) = (0.0078 \pm 0.0004)\left[1 - e^{\frac{-3|h|}{(759.32\pm33.99)}}\right] \tag{7}$$

This result is more robust than the Mukul (1998) result given in Eq. (6) because uncertainties in the arbitrary constants a and c have now been determined.

3. Kriging and cross-validation using Bootstrapped variogram

We next use the Bootstrapped semi-variogram of Eq. (7) to carry out kriging using the GEO-EAS 1.2.1 program (Englund and Sparks, 1991) as in Mukul (1998). The validity of the kriged results was also tested by cross-validation as in Mukul (1998). First, the kriging and cross-validation parameters in GEO-EAS 1.2.1 were fine-tuned to improve the Mukul (1998) estimations further between the 25th and 75th percentiles of the distribution (Table 3, Column 2). Next, the sensitivity of the kriged results to the uncertainty in the computation of the sill value ($c = 0.0078 \pm 0.0004$) in the Bootstrapped semi-variogram (Eq. (7)) was tested. This was done by kriging and cross-validation using $a = 759.32$

Table 2
DA and BEDA results for Sheeprock X/Z axial ratios (Table 1)

Statistical parameters	Mean	Median	Mode	Skewness	Peakedness
EDA results (Mukul, 1998)	1.2832 ± 0.0127	1.2845. Not possible	1.2870. Not possible	0.4726. Not possible	4.8335. Not possible
BEDA result (for $B = 500$)	1.2839 ± 0.0005	1.2866 ± 0.0006	1.2921 ± 0.0013	0.5685 ± 0.0230	4.5355 ± 0.0477

Table 3
Cross-validation results (with corresponding errors in brackets) of Kriging using Bootstrapped semi-variogram

	Measured value	Modified Mukul (1998): $c = 0.008$; $a = 750$	$c = 0.0082$; $a = 759.32$	$c = 0.0074$; $a = 759.32$	$c = 0.0078$; $a = 725.33$	$c = 0.0078$; $a = 793.31$
Minimum	1.067	1.165 (9.18%)	1.165 (9.18%)	1.165 (9.18%)	1.166 (9.28%)	1.163 (9.00%)
25th percentile	1.225	1.240 (1.22%)	1.243 (1.47%)	1.243 (1.47%)	1.243 (1.47%)	1.242 (1.39%)
Mean	1.283	1.280 (0.23%)	1.280 (0.23%)	1.280 (0.23%)	1.280 (0.23%)	1.280 (0.23%)
Median	1.284	1.281 (0.23%)	1.281 (0.23%)	1.281 (0.23%)	1.281 (0.23%)	1.281 (0.23%)
75th percentile	1.319	1.307 (0.91%)	1.307 (0.91%)	1.307 (0.91%)	1.308 (0.83%)	1.308 (0.83%)
Maximum	1.604	1.41 (12.09%)	1.408 (12.22%)	1.408 (12.22%)	1.407 (12.28%)	1.409 (12.16%)

and $c = 0.0082$ and 0.0074 values in the semi-variogram (Eq. (7)). The results (Table 3) reveal that kriged estimates are not sensitive to the uncertainty in the estimation of the sill value of the semi-variogram. Finally, the sensitivity of the kriged results to the uncertainty in the computation of the range ($a = 759.32 \pm 33.99$) in the Bootstrapped semi-variogram (Eq. (7)) was tested; kriging and cross-validation with $c = 0.0078$ and $a = 725.33$ and 793.31 was carried out. Results (Table 3) indicate that kriged estimates are nearly insensitive to the uncertainty in the estimation of the range of the semi-variogram as well.

4. Conclusions

Any experimental measurement of a physical quantity and computations of mathematical parameters based on them is subject to a certain amount of intrinsic random error that needs to be characterized and quantified precisely in the system being studied. The standard spatial statistics method does not permit computation of uncertainties associated with the EDA and semi-variogram parameters. Bootstrapped spatial statistics allows computation of the uncertainties associated with the EDA and semi-variogram parameters and also makes the computation of the best semi-variogram more efficient. Thus, the range (a) = 750 m and sill (c) = 0.008 values associated with the best exponential semi-variogram computed by Mukul (1998) for the Sheeprock *X/Z* strain data was recomputed, using Bootstrapped spatial statistics, to be 759.32 ± 33.99 and 0.0078 ± 0.0004, respectively. However, kriged estimates computed using the Bootstrapped semi-variogram indicate that the results are not sensitive to the uncertainty associated with the estimation of parameters of the semi-variogram used in the Sheeprock example. However, this might not be the case always and kriged estimates need to be tested for sensitivity to uncertainty associated with the semi-variogram used. Therefore, the Bootstrap method of computing EDA parameters and the best semi-variogram should be used in the analysis of spatial data.

Acknowledgements

This work was funded by the Department of Science and Technology (DST), India, Grant Nos. ESS/CA/A9-49/97 and ESS/23/VES/134/2001 to MM. An excellent critical review and comments by Mark Brandon, an anonymous referee and JSG editor Richard Norris helped greatly in improving the quality of this paper. MM also thanks John Tarduno for pointing out that the spatial statistics parameters needed to have uncertainties associated with them and Mark McNaught for e-mail discussions on Bootstrap. We also benefited by discussions with M.K. Sharada and N.K. Indira.

References

Armstrong, M., 1984. Improving the estimation and modelling of the variogram. In: Verly, G., David, M., Journel, A.G., Marechal, A. (Eds.), Geostatistics for Natural Resources Characterization, D. Reidel Publishing Company, Dordrecht, pp. 1–19.

Brandon, M., 1995. Analysis of geologic strain data in strain-magnitude space. Journal of Structural Geology 17, 1375–1385.

Brooker, P.I., 1986. A parametric study of robustness of kriging variance as a function of range and relative nugget effect for a spherical semivariogram. Mathematical Geology 18, 477–488.

Campbell, K., 1988. Bootstrapped models for intrinsic random functions. Mathematical Geology 20, 699–715.

Chung, C.F., 1984. Use of the Jackknife method to estimate autocorrelation functions (or variograms). In: Verly, G., David, M., Journel, A.G., Marechal, A. (Eds.), Geostatistics for Natural Resources Characterization, D. Reidel Publishing Company, Dordrecht, pp. 55–69.

Cressie, N., 1993. Statistics for Spatial Data, John Wiley and Sons, New York.

Diaconis, P., Efron, B., 1983. Computer-intensive methods in statistics. Scientific American 248, 116–126.

Dubrule, O., 1983. Cross validation of kriging in a unique neighborhood. Mathematical Geology 18, 477–488.

Efron, B., 1979. Bootstrap methods; another look at the Jackknife. The Annals of Statistics 7, 1–26.

Efron, B., Tibshirani, R., 1991. Statistical data analysis in the computer age. Science 253, 390–395.

Englund, E., Sparks, A., 1991. GEO-EAS 1.2.1 Geostatistical Environmental Assessment Software. Environmental Monitoring Systems Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Las Vegas.

Hardy, R.L., 1977. Least squares prediction. Photogrammetric Engineering and Remote Sensing 18, 445–448.

Matheron, G., 1963. Principles of geostatistics. Economic Geology 58, 1246–1266.

Mukul, M., 1998. A spatial statistics approach to the quantification of finite strain variation in penetratively deformed thrust sheets: an example from the Sheeprock thrust sheet, Sevier fold-and-thrust belt, Utah. Journal of Structural Geology 20 (4), 371–384.

Mukul, M., Mitra, G., 1998. Finite strain and strain variation analysis in the Sheeprock thrust sheet: an internal thrust sheet in the Provo salient of the Sevier fold-and-thrust belt, Central Utah. Journal of Structural Geology 20 (4), 385–405.

Philip, G.M., Watson, D.F., 1986. Matheronian geostatistics—quo vadis? Mathematical Geology 18, 93–117.

Simon, J.L., 1969. Basic Research Methods in Social Science, Random House, New York.