

Backtransforming Rank Order Kriging Estimates Transformada Reversa das Estimativas de Krigagem Rangueadas

Jorge Kazuo Yamamoto (jkyamamo@usp.br) Departamento de Geologia Sedimentar e Ambiental - Instituto de Geociências - USP R. do Lago 562, CEP 05508-080, São Paulo, SP, BR

Received 04 December 2009; accepted 20 April 2010

ABSTRACT

Kriging of raw data presenting distributions with positive skewness must be avoided because the strong influence of a few high values in the resulting estimates. The solution is to apply data transformation, which changes the shape of original distribution into a symmetric distribution. Kriging of transformed data is performed and then back-transformed to the original scale of measurement. In this paper, we examine the uniform score transform that results in a uniform distribution. Ordinary kriging estimates of uniform score data results in a bell-shaped distribution, since the tails of the distribution are lost in the estimation process because of the smoothing effect. The back-transformation of this bell-shaped distribution result in biased estimates. Therefore, the solution proposed in this paper is to correct the smoothing effect of the rank order kriging estimates before transforming them back to the scale of raw data. Results showed this algorithm is reliable and back-transformed estimates are unbiased in relation to the sample data.

Keywords: Data transformation; Uniform score transformation; Ordinary kriging; Smoothing effect; Back-transformation.

RESUMO

A krigagem de dados apresentando distribuições com assimetria positiva deve ser evitada devido à forte influência dos poucos valores altos nas estimativas resultantes. A solução é a transformada de dados que muda a forma da distribuição original para uma distribuição simétrica. A krigagem dos dados transformados é realizada e então transformada de volta para a escala original de medida. Nesse artigo, nós examinamos a transformada uniforme que resulta em uma distribuição uniforme. A krigagem ordinária de dados uniformes resulta numa distribuição em forma de sino, haja vista as caudas da distribuição terem sido perdidas no processo de estimativa devido ao efeito de suavização. A transformada reversa dos valores apresentando essa distribuição em sino resultará em estimativas enviesadas. Portanto, a solução proposta nesse artigo passa pela correção do efeito de suavização das estimativas ranqueadas antes de transformá-las para a escala dos dados originais. Os resultados obtidos mostraram que esse algoritmo é confiável e as estimativas transformadas para a escala original não são enviesadas em relação aos dados amostrais.

Palavras-chave: Transformação de dados; Transformada uniforme; Krigagem ordinária; Efeito de suavização; Transformada reversa.

INTRODUCTION

Data transformations have been considered as an interesting alternative to kriging, particularly when kriging raw data should be avoided, as, for example, the case of data presenting highly skewed distribution. In geostatistics, the lognormal transformation and normal score transform have been used to normalize skewed distributions. Another data transformation was proposed by Journel and Deutsch (1997, p. 175), which is referred as uniform score transform that is the cumulative frequency for data sorted in ascending order. The uniform score transform produces data uniformly distributed into [0,1]. Therefore, this nonlinear data transformation changes the shape of the original data distribution into a uniform distribution. The great advantage of the uniform score transform is the possibility of integration of data of diverse types, scales, supports and accuracies (Journel and Deutsch, 1997, p. 174). Transformed data can be estimated or simulated at unsampled locations and the resulting values can be back-transformed to the original measurement scale. This method has been used for mapping heavy metal concentrations in contaminated soils (Juang, Lee, Ellsworth, 2001, p. 895). However, kriging estimates show a reduced variance due to the smoothing effect (Journel and Deutsch, 1997, p. 179), which histogram and semivariogram are not reproduced. These authors have improved kriging estimates, transforming them under the constraint of data reproduction (Journel and Deutsch, 1997, p. 179-180), which produces a histogram closer to a uniform distribution, but the semivariogram is not reproduced. According to Journel and Deutsch (1997, p. 181), even using sequential uniform simulation, the uniform distribution was not reproduced because simulated values fell outside the permissible range [0,1]. Once again, when simulated values are transformed under the constraint of data reproduction (Journel and Deutsch, 1997, p. 181), the authors obtained a histogram closer to the uniform distribution. Therefore, the challenge for estimating or simulating rank transformed data is the reproduction of the histogram displaying a uniform distribution. This paper shows how we can reproduce both, the sample histogram and the sample semivariogram of rank transformed data. Actually, the main objective of this paper is to show a new application of the post-processing algorithm for correcting the smoothing effect of ordinary kriging estimates (Yamamoto, 2005, 2007). Moreover, as it is shown in this paper, back-transformed values reproduce the original sample histogram and, consequently, the mean and variance of the original data.

UNIFORM SCORE TRANSFORM

The uniform score transform is derived from rank orders $r(x_i)$ associated with a set of n data values $\{z(x_i), i = 1, n\}$; in which $r(x_1) = 1$ and $r(x_n) = n$ (Journel and Deutsch, 1997, p.176). According to these authors, the standardized rank:

$$v(x_i) = r(x_i)/n \tag{1}$$

is the uniform score transform. We can also define the standardized rank as (Saito and Goovaerts, 2000, p. 4230):

$$v(x_i) = r(x_i)/n - 0.5/n$$
 (2)

A uniform distribution on an interval (a,b) has an expected value of:

$$E[X] = \frac{(a+b)}{2}$$

and a variance of:

$$S^2 = \frac{(b-a)^2}{12}$$

After equation (1), *a* is equal to 1/n and *b* is equal to 1. Therefore, the mean is equal to (n + 1)/2n, that is, a little greater than 0.5, and the variance is:

$$\left(\frac{n-1}{n}\right)^2 / 12$$
, i.e. a little less than 1/12.

According to equation (2), a is equal to (1 - 0.5)/n and b is equal to (n - 0.5)/n. Thus the mean is 0.5 and the variance is:

$$\left(\frac{n-1}{n}\right)^2 / 12 ,$$

which is the same value obtained in equation (1) and, therefore, a little less than 1/12. In this paper, equation (2) will be used, since it provides symmetric standardized ranks.

REPRODUCING THE UNIFORM SCORE HISTOGRAM

As mentioned before, the challenge here is the reproduction of the sample histogram that represents a uniform distribution. Clearly, in order to reproduce the sample histogram, it is necessary to obtain the sample variance. The reproduction of the sample semivariogram depends not only on the sample variance but also on the spatial distribution of ranked estimates. We can estimate rank orders at unsampled locations using the well-known ordinary kriging technique:

$$v_{OK}^{*}(x_{o}) = \sum_{i=1}^{m} \lambda_{i} v(x_{i})$$
(3)

where $v_{OK}^*(x_o)$ is the rank order estimate at the unsampled location x_o ; $\{v(x_i) = 1, n\}$ is the set of neighbor ranked data and $\{\lambda_i, i = 1, n\}$ is the set of weights associated with neighbor data.

According to Yamamoto (2000), we can determine the interpolation variance associated with ordinary kriging estimate as:

$$S_{o}^{2} = \sum_{i=1}^{m} \lambda_{i} \left[v(x_{i}) - v_{OK}^{*}(x_{o}) \right]^{2}$$
(4)

where S_o^2 is the interpolation variance. Other variables have the same definition as in equation (3).

It has been proven the heteroscedastic interpolation variance is more reliable than the homoscedastic kriging variance (see Yamamoto, 2000, 2005, 2007). Although the interpolation variance has been used by this author since 1989 (Yamamoto, 1989), it was published only after Journel and Rao (1996) interpreted ordinary kriging weights as conditional probabilities. To make that possible, not only must the ordinary kriging weights be all positive, but also they have to sum up to one. After it has been confirmed that all kriging weights are positive and sum up to one, a conditional cumulative distribution function can be built at any given location x_o , by sorting the n neighboring data into increasing order:

$$z(x_1) \le z(x_2) \le \dots \le z(x_n)$$

The local conditional cumulative distribution function is then modeled as:

$$F(x_o, z_\alpha) = \sum_{i=1}^{\alpha} \lambda_i$$

where λ_i is the weight associated with ith data point and interpreted as the conditional probability for the ith data point.

The conditional cumulative distribution function derived from ordinary kriging weights has a conditional expectation that is equal to the ordinary kriging estimate (3) and a conditional variance that is none other than the interpolation variance (4). Because of the smoothing effect, the variance of ordinary kriging estimates is less than the sample variance and is expressed as:

$$Var\left[V_{OK}^{*}(x)\right] < \left(\frac{n-1}{n}\right)^{2} / 12$$

If we want to reproduce the sample histogram, estimates have to be corrected by adding a certain amount, in order that the variance of corrected estimates be equal to the sample variance. The correcting amount can be obtained using simulation methods or a post-processing algorithm proposed by Yamamoto (2005, 2007), as follows:

$$v_{OK}^{**}(x_{o}) = v_{OK}^{*}(x_{o}) + v_{NS_{o}}^{*}(x_{o}) factor$$
(5)

where $v_{NS_o}^*(x_o)$ is the correcting amount added to ordinary kriging estimates and *factor* is a number that makes the variance of corrected estimates equal to the sample variance. For details on the algorithm for correcting the smoothing effect of ordinary kriging estimates, please see Yamamoto (2005, 2007). The optimum value that makes the variance of corrected estimates equal to the sample variance is determined by the equation below:

$$factor = \frac{-2Cov\left(V_{OK}^{*}(x), V_{NS_{o}}^{*}(x)\right) + \sqrt{\Delta}}{2Var\left[V_{NS_{o}}^{*}(x)\right]}$$

where
$$\Delta = 4Cov \left(V_{OK}^*(x), V_{NS_o}^*(x) \right)^2$$

- $4Var \left[V_{NS_o}^*(x) \right] \left(Var \left[V_{OK}^*(x) \right] - Var \left[V(x) \right] \right)$

is the discriminant of the quadratic function; and

$$Var\left[V(x)\right] = \left(\frac{n-1}{n+1}\right)^2 / 12$$
 is the sample variance

This method reproduces the sample histogram. If the histogram of rank orders is recovered, the histogram of back-transformed values will be similar to the histogram of original values.

The only way of restoring the sample variance is by adding a random variable in order that the variance of corrected estimates be equal to the sample variance. This is possible if and only if this random variable has a zero mean in order to guarantee unbiasedness of ordinary kriging estimates and variance and covariance that restores the sample variance. For more details on this subject, please see Yamamoto (2008).

MATERIALS AND METHODS

For testing the proposed method, three exhaustive data sets have been considered in this study. These data sets were derived from the well-known true.dat (Deutsch and Journel, 1992, p. 35), modifying the secondary variable by mathematical transformation, and resulting in a new variable. The first variable represents a perfect normal distribution after a normal score transform of the secondary variable:

$$Z_{GAUSS} = G^{-1}(Z)$$

The second variable was derived from an exponential function:



 $Z_{LOG} = 0.004115 \exp(1.098612 * Z_{GAUSS})$

Therefore, this variable represents a lognormal distribution. And the third variable is the sum of both variables:

$$Z_{SUM} = Z_{GAUSS} + Z_{LOG}$$

All exhaustive data sets generated according to previous equations are shown in Figure 1. Table 1 presents summary statistics for variables of the exhaustive data sets. The first variable represents a normal distribution and the second a lognormal distribution, whereas the third variable represents a positively skewed distribution, but cannot be considered a lognormal distribution.

From these exhaustive data sets, samples of 121 data points were drawn based on stratified random sampling. These samples have the same location of data points for



Figure 1. Image maps for exhaustive data sets: **A.** normal distribution, **B.** lognormal distribution, **C.** positive skewness distribution.

the three data sets. Statistics for the samples drawn from exhaustive data sets are shown in Table 2. Histograms are shown in Figure 2. These samples were used for inferring their exhaustive models.

s were used for inferring for all samples. All calculations were carried

All original variables were converted into standardized ranks according to equation (2). Since data locations of all

three samples are the same, standardized rank histogram (Figure 3A) and semivariogram (Figure 3B) were the same for all samples.

All calculations were carried out based on this semivariogram model. Ordinary kriging and corrected ordinary kriging estimates were calculated according to

Table 1. Summary statistics for variables of exhaustivedata sets.

Summary statistics	Z _{GAUSS}	Z _{LOG}	Z _{SUM}
N° of data	2500	2500	2500
Mean	5.000	1.815	6.815
Standard deviation	0.997	2.617	3.428
Coefficient of variation	0.199	1.441	0.503
Maximum	8.353	39.782	48.140
Upper quartile	5.674	2.096	7.769
Median	4.999	0.999	5.999
Low er quartile	4.325	0.476	4.802
Minimum	1.647	0.025	1.672



Table 2. Summary statistics for samples drawn fromexhaustive data sets.

Summary statistics	Z _{GAUSS}	Z_{LOG}	Z _{SUM}
N° of data	121	121	121
Mean	4.998	1.751	6.749
Standard deviation	0.950	2.368	3.177
Coefficient of variation	0.190	1.352	0.471
Maximum	7.537	16.227	23.763
Upper quartile	5.656	2.057	7.713
Median	4.959	0.956	5.915
Low er quartile	4.309	0.468	4.778
Minimum	2.626	0.074	2.700



Figure 2. Sample histograms for: A. variable with a normal distribution, B. variable with a lognormal distribution, C. variable with a positive skewness distribution.



Figure 3. Uniform histogram of standardized rank data (A) and standardized rank semivariogram model (B).

equations (3) and (5), respectively. The ordinary kriging estimation was obtained using the twelve nearest neighbor points searched by the quadrant procedure. Firstly, we want to show that the method for correcting the smoothing effect of ordinary kriging estimates reproduces the sample histogram and the sample semivariogram. Furthermore, we also want to show that the back-transformed values are reliable and unbiased.

RESULTS AND DISCUSSION

First, let us analyze the standardized rank statistics (Table 3). When examining this table, it is observed that ordinary kriging estimates have reduced variance. On the contrary, corrected ordinary kriging estimates reproduce the mean and variance, but the median has a small bias. However, it has to be checked if this small difference in the upper tail of the distribution affects back-transformation to original values.

When original data is transformed into a new scale, for instance, rank data, the main concern is the backtransformation procedure. As we know, ordinary kriging estimates have a serious drawback, which is known as the smoothing effect. Therefore, if smoothed kriging estimates (calculated after equation - 3) are backtransformed, the resulting estimates will also be smoothed and, consequently, biased (loss of lower and upper tails and reduced variance). Thus, the solution is to correct the smoothing effect (equation 5) and back-transform the corrected estimates to the original scale of measurement. Please, see Table 4 - back-transformed estimates from both conventional ordinary kriging estimates and Table 5 corrected ordinary kriging estimates.

Table 3. Standardized rank statistics.

Summary statistics	V(x)	$V_{OK}^{*}(x)$	$V_{OK}^{**}(x)$
N° of data	121	2500	2500
Mean	0.500	0.504	0.500
Standard deviation	0.286	0.202	0.286
Coefficient of variation	0.573	0.402	0.572
Maximum	0.992	0.967	0.984
Upper quartile	0.774	0.665	0.734
Median	0.496	0.502	0.511
Low er quartile	0.248	0.354	0.248
Minimum	0.008	0.028	0.008

Table 4. Summary statistics for back-transformed estimates

 obtained from conventional rank order ordinary kriging

 estimates (equation 3).

Summary statistics	Z _{GAUSS}	Z_{LOG}	Z _{SUM}
N° of data	2500	2500	2500
Mean	4.987	1.196	6.183
Standard deviation	0.567	0.801	1.341
Coefficient of variation	0.114	0.670	0.217
Maximum	6.802	7.242	14.044
Upper quartile	5.362	1.488	6.849
Median	4.970	0.967	5.937
Low er quartile	4.635	0.670	5.305
Minimum	3.291	0.153	3.444

After comparing summary statistics in Tables 4 and 5, we conclude that the back-transformation of corrected rank order kriging estimates produces the best results when compared with sample data (Table 2). Figures 4, 5 and 6 illustrate both procedures for back-transforming rank order kriging estimates.

In Figures 4, 5 and 6, we can see how the postprocessing algorithm for correcting the smoothing effect works properly. As observed in these figures, an almost bell shaped histogram resulting from smoothed ordinary kriging estimates is corrected to a uniform distribution. Therefore, we can reproduce as close as possible the sample histogram, when corrected estimates with uniform distribution are back-transformed.

The proposed procedure for back-transforming rank order kriging estimates seems to reproduce sample histograms quite well. Since it is very difficult to compare histograms, it is much easier to reach a conclusion about the effectiveness of the proposed method using the comparison of cumulative frequency distribution (Figure 7).

Figure 7 shows the reproduction of the sample histogram of back-transformed corrected rank order kriging estimates. P-P plots also confirm this observation.

Figures 8, 9 and 10 present back-transformed histograms for both smoothed and corrected rank order kriging estimates. Comparing these figures, it is clear that the post-processing algorithm adds some information that was missing due to the smoothing effect.

Since the exhaustive data sets (Figure 1) are available, we can compare actual values vs. estimates in scattergrams (Figures 11, 12 and 13).

Table 5. Summary statistics for back-transformed estimatesobtained from corrected rank order ordinary krigingestimates (equation 5).

Summary statistics	Z _{GAUSS}	Z _{LOG}	Z _{SUM}
N° of data	2500	2500	2500
Mean	4.990	1.770	6.760
Standard deviation	0.974	2.381	3.215
Coefficient of variation	0.195	1.345	0.476
Maximum	7.325	12.884	20.209
Upper quartile	5.608	1.952	7.560
Median	4.987	0.986	5.974
Low er quartile	4.310	0.469	4.779
Minimum	2.626	0.074	2.700

Once again, scattergrams of back-transformed estimates after correcting the smoothing effect show better correlation and are not biased in relation to the 1:1 line. Moreover, the post-processing algorithm improves local accuracy in terms of the correlation between actual values and backtransformed estimates.

Geolo

Let us also analyze the semivariogram reproduction of rank order kriging estimates. Figure 14 displays the semivariogram model with experimental semivariograms computed from smoothed (empty circle) and corrected (full circle) rank order kriging estimates. Since the post-processing algorithm (equation 4) recovers almost the full sample variance, the experimental semivariogram of corrected estimates has a sill closer to the sill of the semivariogram model.

Now we can see how the post-processing algorithm works. Equation (5) is composed of two terms: the ordinary kriging estimate and the smoothing error. Thus, summing up these two terms, we obtain the corrected estimates. Figure 15 displays histograms of these two terms and the resulting histogram of corrected estimates. Smoothing errors in Figure 15B have a bell shaped distribution with a mean of zero and standard deviation of 0.1088.

It is important to check the reproduction of variance after equation (5). Applying the variance operator to equation (5), we obtain the variance of corrected estimates as:

$$Var\left[V_{OK}^{**}(x)\right] = Var\left[V_{OK}^{*}(x)\right] + Var\left[V_{NS_{o}}^{*}(x)^{*} factor\right] + 2Cov\left(V_{OK}^{*}(x), V_{NS_{o}}^{*}(x)^{*} factor\right)$$

where $Var\left[V_{OK}^{**}(x)\right]$ is the variance of the corrected estimates;

Var $\left[V_{OK}^{*}(x)\right]$ is the variance of the ordinary kriging estimates

and $Var\left[V_{NS_o}^*(x)^* factor\right]$ is the variance of the correcting

amounts multiplied by a constant factor.

Table 6 shows the components of the variance of the corrected estimates, which sum up to 0.819. This result is very close to 0.8197, which is the sample variance. Note that the sample variance tends to 1/12 (0.08333) when the sample size tends to infinity. In this case study, where the sample size is equal to 121, the sample variance is only 0.08197.

It is also important to understand how smoothing errors are related to rank order kriging estimates. Figure 16 shows that there is a linear relationship between smoothing errors and rank order kriging estimates. Therefore, these errors are not independent of estimates, as stated by Yamamoto (2008, p. 589).



Figure 4. Procedures for back-transforming rank order kriging estimates for variable Z_{GAUSS} .





Figure 5. Procedures for back-transforming rank order kriging estimates for variable Z_{LOG}.



Figure 6. Procedures for back-transforming rank order kriging estimates for variable Z_{SUM}.





Figure 8. Back-transformed rank order kriging estimates for normal variables: A. smoothed rank estimates, B. corrected rank estimates.



Figure 9. Back-transformed rank order kriging estimates for lognormal variables: A. smoothed rank estimates, B. corrected rank estimates.



Figure 10. Back-transformed rank order kriging estimates for added variables: A. smoothed rank estimates, B. corrected rank estimates.



Figure 11. Scattergrams of actual values vs. estimates for normal variables: A. back-transformed smoothed rank estimates, B. back-transformed corrected rank estimates.



Figure 12. Scattergrams of actual values vs. estimates for lognormal variables: A. back-transformed smoothed rank estimates, B. back-transformed corrected rank estimates.



Figure 13. Scattergrams of actual values vs. estimates for added variables: A. back-transformed smoothed rank estimates, B. back-transformed corrected rank estimates.



Figure 14. Semivariograms of ordinary kriging estimates (empty circle), corrected ordinary kriging estimates (full circle) and sample data (asterisk).



Figure 15. Illustrating the post-processing algorithm: A. rank order kriging estimates, B. smoothing errors, C. corrected rank order kriging estimates.



$Var\left[V(x)\right]$	$Var\left[V_{OK}^{*}(x)\right]$	$Var\left[V_{NS_{o}}^{*}(x)^{*} factor\right]$	$2Cov\left(V_{OK}^{*}(x),V_{NS_{o}}^{*}(x)\right)$ * factor)	Sum
0.08197	0.04097	0.01184	0.02909	0.08190



Figure 16. Scattergram of smoothing errors vs. rank order kriging estimates with a positive correlation coefficient.

CONCLUSIONS

This paper shows a new application of the postprocessing algorithm. In this application, the challenge was the reproduction of the sample histogram with uniform distribution. The post-processing algorithm reproduced the uniform score histogram very well and, consequently, the back-transformed estimates had cumulative frequency distributions that closely matched the sample distributions. Therefore this procedure guarantees unbiased backtransformed estimates, making it possible to use uniform score transform as proposed in the literature.

ACKNOWLEGEMENTS

The author wishes to thank the comments and suggestions of an anonymous reviewer that improved this manuscript. The author also wants to thank the National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq) that funded the research (under grant nº 303505/2007-9).

REFERENCES

DEUTSCH, C. V.; JOURNEL, A. G. *GSLib*: geostatistical software library and user's guide. NewYork: Oxford University Press, 1992. 340 p.

JOURNEL, A G.; DEUTSCH, C. V. Rank order geostatistics: a proposal for a unique coding and common processing of diverse data. In: BAAFI, E. Y; SCHOFIELD (Ed.). *Geostatistics Wollongong '96*, Netherlands: Kluwer Academic Publishers, 1997. v. 1, p. 174-187.

JOURNEL, A. G.; RAO, S. E. *Deriving conditional distributions from ordinary kriging*, Palo Alto: Stanford Center for Reservoir Forecasting, 1996. 25 p.

JUANG, K. W.; LEE, D. Y.; ELLSWORTH, T. R. Using rank-order geostatistics for spatial interpolation of highly skewed data in a heavy-metal contaminated site. *Journal Environmental Quality*, Madison, v. 30, p. 894-903, 2001.

SAITO, H.; GOOVAERTS, P. Geostatistical interpolation of positively skewed and censored data in dioxincontaminated site. *Environmental Science Technology*, Easton, v. 34, p. 4228-4235, 2000.

YAMAMOTO, J. K. A new method for ore reserve estimation and modeling. *Brasil Mineral*, São Paulo, v. 68, p. 52-56, 1989.

YAMAMOTO, J. K. An alternative measure of the reliability of ordinary kriging estimates. *Mathematical Geology*, New York, v. 32, p. 489-509, 2000.

YAMAMOTO, J. K. Correcting the smoothing effect of ordinary kriging estimates. *Mathematical Geology*, New York, v. 37, p. 69-94, 2005.

YAMAMOTO, J. K. On unbiased back-transform of lognormal kriging estimates. *Computacional Geosciences*, Netherlands, v. 11, p. 219-234, 2007.

YAMAMOTO, J. K. Estimation or simulation? That is the question. *Computacional Geoscience*, Netherlands, v. 12, p. 573-591, 2008.

Table 6. Components for the variance of	f corrected or	rdinary rank or	der kriging estimates.
---	----------------	-----------------	------------------------