A SURVEY INTO ESTIMATION OF LOGNORMAL DATA

Jorge Kazuo YAMAMOTO [1] & **Rafael de Aguiar FURUIE** [2]

(**1**) *Instituto de Geociências, Universidade de São Paulo. Rua do Lago, 562. CEP 05508-080.*
São Paulo, SP. Endereço eletrônico: jkyamamo@usp.br
(**2**) *Petróleo Brasileiro S.A. / PETROBRAS. Avenida Elias Agostinho, 665. CEP 27913-350. Macaé, RJ.*
Endereço eletrônico: furuie@petrobras.com.br

ABSTRACT – Lognormal data are very difficult to handle because of its high variability due to the occurrence of a few high values. In geostatistics the solution calls for a data transform, such as the logarithm transform and the indicator transform. Both approaches have been used for estimating lognormal data. Lognormal kriging works on kriging the transformed data and then estimates are back-transformed into the original scale of data. Indicator kriging builds a conditional cumulative distribution function at every unsampled location and estimates are based on the conditional mean or E-type estimate. Usually back-transformed lognormal kriging estimates are mean biased and conditional means from indicator kriging are unbiased. This paper compares both approaches for 27 data sets presenting distributions with increasing positive skewness. Actually 27 exhaustive data sets have been computer generated from which stratified random samples with 90 points were drawn. Estimates were first examined for local accuracy and the associated uncertainties were checked for the proportional effect. Results show that lognormal kriging is still the best approach for lognormal data if we use an algorithm that takes into consideration correcting the smoothing effect before back-transformation.
Keywords: lognormal distribution, lognormal kriging, indicator kriging, proportional effect.

RESUMO – *J.K. Yamamoto & R. de A. Furuie - Um estudo sobre estimativa de dados lognormais*. Dados lognormais são muito difíceis de se trabalhar devido à sua grande variabilidade por causa da ocorrência de uns poucos valores altos. Em geoestatística a solução passa pela transformação dos dados, como a transformada logarítmica e a transformada indicadora. Ambas as aproximações têm sido utilizadas para estimativa de dados lognormais. A krigagem lognormal trabalha sobre os dados transformados e após isto as estimativas são transformadas de volta para a escala original dos dados. A krigagem da variável indicadora constrói uma função de distribuição acumulada condicional em cada ponto não amostrado e as estimativas são baseadas na média condicional ou estimativa do tipo E. Geralmente, estimativas por krigagem lognormal transformadas de volta para a escala original apresentam vieses em relação à média amostral e as médias condicionais derivadas da krigagem da indicadora não são enviesadas. Esse trabalho compara ambas as aproximações para 27 conjuntos de dados apresentando distribuições com assimetria positiva crescente. Na verdade, 27 dados completos foram gerados em computador dos quais amostras aleatórias estratificadas com 90 pontos foram extraídas. As estimativas foram examinadas inicialmente em relação à precisão local e as incertezas foram verificadas para o efeito proporcional. Os resultados mostram que a krigagem lognormal é ainda a melhor aproximação para dados lognormais se usarmos a equação que leva em consideração a correção do efeito de suavização antes da transformada reversa.
Palavras-chave: distribuição lognormal, krigagem lognormal, krigagem da indicadora, efeito proporcional.

INTRODUCTION

Lognormal distributions are very common in mineral deposits of rare metals, diamonds, uranium and other minerals. This distribution is characterized by a positive skewness in such a way that the mean is greater than the median of the distribution. Data displaying lognormal distribution present a great number of low values and a few high values. These high values increase the variance of the data set and make the task of semivariogram calculation and ordinary kriging estimation difficult. Actually, experimental semivariograms are very sensitive to these high values and consequently are useless (Journel, 1983). Journel (1983) proposed two solutions for this problem: trim off high values or transform the original data using functions such as square roots, natural logarithm or normal score transform. Data transformation is a much better solution than trimming off high valued data. The objective of data transform is to obtain a symmetrical distribution. Logarithm transform is a good option used not only in geostatistics but also in other fields. Transformed data are then used for computing and modeling the semivariogram and for ordinary kriging estimation. After that estimates in the transformed domain are back-transformed into the original scale of measurement. For ordinary lognormal kriging it was proved that back-transformation after correcting the smoothing effect of ordinary kriging estimates is the best alternative to get unbiased results (Yamamoto, 2007).

Another approach commonly used for lognormal data was proposed by Journel (1983), based on the indicator transform. According to this approach, instead of estimating at every unsampled location, we build a conditional cumulative distribution function (ccdf). From this conditional cumulative distribution function some statistics can be derived that are the conditional mean or E-type estimate and the conditional variance as well. It is important to note that the conditional variance derived from the indicator approach is much better than the traditional kriging variance, which is considered as just a measure of the spatial configuration of neighboring data (Journel & Rossi, 1989).

The results for both approaches can be compared with each other in terms of unbiasedness, correlation and errors of estimates versus real data. This paper presents the results of a comparison between ordinary lognormal kriging and indicator kriging for estimation of lognormal data.

ORDINARY LOGNORMAL KRIGING

Lognormal kriging was proposed by Journel (1980), who also proposed a back-transform equation based on the kriging variance following the traditional approach for computing the mean of lognormal data. Original data are transformed into logarithms as follows:

$$Y(x) = Log(Z(x)) \qquad (1)$$

By definition if the random variable Z(x) follows a lognormal distribution then *Y*(*x*) will present a normal distribution. Sometimes it is necessary to use another logarithm transform in order to guarantee that 50% of transformed data are less than zero and the other 50% are greater than zero. It can be done by dividing *Z*(*x*) by its median and then taking its logarithm:

$$Y(x) = Log\left(\frac{Z(x)}{Median}\right) \qquad (2)$$

This transform does not change the shape of the resulting frequency distribution but only guarantees the symmetry of transformed data relative to zero.

In geostatistical estimation or simulation the semivariogram model is the point of departure. The experimental semivariogram is computed by using the transformed values. Estimation at unsampled locations can be made using ordinary kriging:

$$Y^*_{OK}(x_o) = \sum_{i=1}^{n} \lambda_i Y(x_i) \qquad (3)$$

Estimates at the unsampled locations are in the logarithmic domain and so they need to be back-transformed into the original scale of measurement. The traditional formula for back-transforming lognormal kriging estimates is based on (Journel, 1980):

$$Z^*_{OLK}(x_o) = \exp\left(Y^*_{OK}(x_o) + \sigma^2_{OK}/2 - \mu\right) * Median \qquad (4)$$

However, this is where the main problem in lognormal kriging appears since back-transformed estimates are usually biased when compared with the original data (Journel & Huijbregts, 1978). Bias of back-transformed estimates is reported in several papers (e.g. Saito & Goovaerts, 2000) because expression (4) is very sensitive to the semivariogram model.

A new approach was proposed by Yamamoto (2007) in which the back-transform is performed after correcting ordinary kriging estimates (equation 3) for the smoothing effect (Yamamoto, 2005). Actually, the ordinary kriging estimator (equation 3) is none other than a weighted average formula and therefore its results will present some smoothing. As a consequence, low values are overestimated and high values underestimated. Comparing the histogram of ordinary kriging estimates with the histogram of transformed data it is possible to realize that the lower and upper tails are lost in the estimation process. Therefore, if we try to back-transform a smoothed histogram we will not get the original data histogram. This is the main idea behind the approach proposed by Yamamoto (2007), details of this approach can be found in the referred paper. Thus, according to Yamamoto (2007), ordinary kriging estimates can be back-transformed by using:

$$Z^{**}_{OLK}(x_o) = \exp\left(Y^*_{OK}(x_o) + Y^*_{NS_o}(x_o)\right) * Median \qquad (5)$$

where $Y^*_{NS_o}(x_o)$ is the smoothing error that is negative when overestimation occurs and positive otherwise.

This way estimates can then be back-transformed into the original scale of measurement. But uncertainties remain in the logarithmic scale and so they cannot be used. A new approach for back-transforming uncertainties was proposed by Yamamoto (2008). According to this proposal, the interpolation standard deviation can be back-transformed as:

$$\begin{aligned}\sigma &= \exp\left(Y_{OK}^{*}(x_o) + S_o\right) * Median - (\\ &- \exp\left(Y_{OK}^{*}(x_o)\right) * Median\end{aligned} \qquad (6)$$

where $Y_{OK}^{*}(x_o)$ is the lognormal kriging estimate at an unsampled location x_o and S_o is the interpolation standard deviation (Yamamoto, 2000) in the logarithmic scale.

It is important to note that we cannot simply obtain the interpolation standard deviation in the original scale of measurement by applying $\exp(S_o) * Median$. Actually we have to add to it the term which brings the uncertainty into the range of logarithmic values.

INDICATOR KRIGING

The indicator approach is based on the indicator transform of the original data as follows (Journel, 1983):

$$I(x; z_c) = \begin{cases} 1 \text{ if } Z(x) < z_c \\ 0 \text{ if } Z(x) \geq z_c \end{cases} \qquad (7)$$

where z_c is the cutoff grade or a reference value.

The mean of an indicator variable is the probability that the random variable is less than the cutoff grade:

$$m = E[I(x; z_c)] = P(Z(x) < z_c) \qquad (8)$$

The variance of an indicator variable can be written as:

$$\begin{aligned}Var[I(x; z_c)] &= E\left[I^2(x; z_c)\right] - \left(E[I(x; z_c)]\right)^2 \\ &= m - m^2 = m(1 - m)\end{aligned} \qquad (9)$$

Noting that $E\left[I^2(x; z_c)\right] = E[I(x; z_c)]$ the variance can also be expressed in terms of probabilities:

$$\begin{aligned}Var[I(x; z_c)] &= P(Z(x) < z_c)(1 - P(Z(x) < z_c)) = \\ &= P(Z(x) < z_c)P(Z(x) \geq z_c)\end{aligned} \qquad (10)$$

With this new variable the indicator semivariogram is computed and modeled for the indicator kriging approach. The indicator kriging estimator is (Journel, 1983):

$$I_{OK}^{*}(x_o; z_c) = \sum_{i=1}^{n} \lambda_i I(x_i; z_c) = P(Z(x_o) < z_c) \qquad (11)$$

This means we are estimating the probability that the random variable at an unsampled location x_o is less than the cutoff grade z_c. The uncertainty associated with the indicator kriging estimate after (11) is as follows:

$$S_o^2 = \sum_{i=1}^{n} \lambda_i \left[I(x_i; zc) - I_{OK}^{*}(x_o; zc)\right]^2 \qquad (12)$$

Actually this is the interpolation variance according to Yamamoto (2000). Developing this expression we get:

$$S_o^2 = \sum_{i=1}^{n} \lambda_i I^2(x_i; zc) - \left(I_{OK}^{*}(x_o; zc)\right)^2$$

Note that this is similar to expression (9). Thus, the interpolation variance can be interpreted as a product of probabilities as shown in (10):

$$S_o^2 = P(Z(x_o) < zc)P(Z(x_o) \geq zc) \qquad (13)$$

The same interpretation cannot be done with the kriging variance because it depends on the semivariogram model.

It is important to mention we are estimating the probability for just a cutoff grade. However, if we are interested in building a conditional cumulative distribution function we need to estimate the probability for several cutoff grades. Therefore, we have to split the original data distribution into a number of cutoff grades in such a way that we can build a conditional cumulative distribution function. Just for illustration purposes Table 1 shows some the first and the last percentiles for a number of cutoff grades.

TABLE 1. Sampled intervals of the distribution after splitting into a number of cutoff grades.

Nb. cutoff	First percentile	Last percentile
9	10	90
19	5	95
39	2.5	97.5
49	2	98
79	1.25	98.75
99	1	99

As we can see, even when dividing the original distribution into 100 intervals (or 99 cutoff grades), only 98% of the data are considered for building the conditional cumulative distribution function.

If we choose 19 cutoff grades, it means we have to compute and model 19 indicator semivariograms. Besides, indicator semivariograms computed for cutoff grades representing the tails of the distribution will present great statistical fluctuations. For example, if the first percentile is 5% we will have only 5% of data equal to one and 95% equal to zero. Therefore, only 5% of the data will form pairs (the squared difference must be greater than zero) that can be considered in the semivariogram computation. The same happens in the upper tail, in which 95% of data will be equal to one and 5% equal to zero. Once again, only 5% of the data will form pairs for semivariogram calculation. Other than that, often we have problems for building the conditional cumulative distribution function mainly when order relation occurs (Hohn, 1999).

Thus, a practical solution for this problem was proposed by Deutsch & Journel (1992) which is based on the median indicator semivariogram. This is the best semivariogram because 50% of data are equal to one and the other 50% equal to zero, meaning all data will form pairs for semivariogram computation. The median indicator semivariogram is used for all other cutoff grades and order relation will never occur. This approach will be considered in this paper. For illustration purposes let us consider a conditional cumulative distribution function presented in Figure 1.

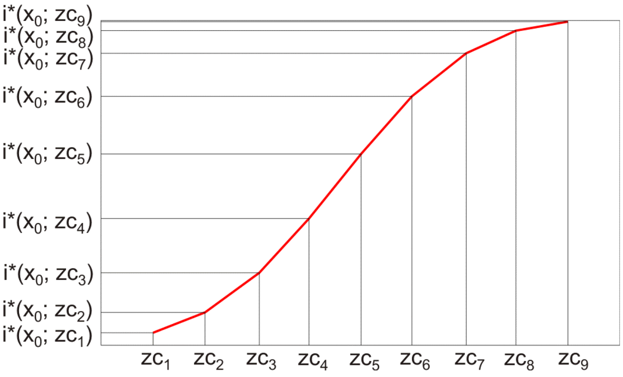

FIGURE 1. Illustrating a conditional cumulative distribution function built from 9 deciles.

From the conditional cumulative distribution function we can derive two statistics: the conditional mean or E-type estimate (Deutsch & Journel, 1992) and the conditional variance:

$$Z_E^*(x_o) = \sum_{i=2}^{K} \left(\frac{zc_i - zc_{i-1}}{2} \right) \left(I_{OK}^*(x_o; zc_i) - I_{OK}^*(x_o; zc_{i-1}) \right) \quad (14)$$

$$\sigma^2(x_o) = \sum_{i=2}^{K} \left(I_{OK}^*(x_o; zc_i) - I_{OK}^*(x_o; zc_{i-1}) \right) \left(zc_i - Z_E^2(x_o) \right)^2 \quad (15)$$

The great advantage of this method is that the conditional mean and the conditional variance derived from the conditional cumulative distribution function are in the original scale of measurement.

MATERIALS AND METHODS

In this section we want to show how synthetic data can be computer generated. What we need is the spatial distribution of a random variable. For instance we can start from the well known public domain data set named *true.dat* (Deutsch & Journel, 1992). This data set presents two variables named: primary and secondary. Since the primary variable is a simulated variable, it was chosen to work with the secondary variable. Then this secondary variable from *true.dat* was transformed into a normal distribution N(0,1) using the procedure described in Deutsch & Journel (1992). Figure 2 shows the original secondary variable and the normal score transformed new variable. Since the original data represent a spatial phenomenon, we can consider them as an exhaustive set of data or a known population.

We can check parameters for both populations as given in Table 2.

Population parameters (Table 2) confirm a Gaussian distribution after the normal score transform of the secondary variable from *true.dat* (Deutsch & Journel, 1992).

From this normal score transformed variable we can derive a lognormal distribution by raising *e* (2.71828) to a power equal to the normal score:

$$Z_{Log} = \exp(Z_{Gauss}) \quad (16)$$

By definition we have an exact lognormal distribution because if we take the logarithm of Z_{Log} we have Z_{Gauss}, which presents a normal distribution. Figure 3 illustrates a typical lognormal distribution.

Population parameters for the new random variable Z_{Log}, which presents a typical lognormal distribution, are presented in Table 3.

In Table 3 we can observe that a coefficient of variation equal to 1.254 means a typical lognormal distribution.

If we multiply the random variable Z_{Gauss} by a constant K ($K > 1$) in equation (16) we will obtain other lognormal distributions, but if the constant K is less than one, other positively skewed distributions with coefficients of variation less than 1.254 are generated.

$$Z_{Log_K} = \exp\left(Z_{Gauss} * K * 0.1\right) \qquad (17)$$

In equation (17) we multiply K times 0.1 in such a way we can use K as an integer constant. Starting from K equal to 1 to K equal to 27 we will have 27 synthetic exhaustive data sets. Just for illustration purposes we show only population parameters (Table 4) and image maps for K=1 and for K=27 (Figure 4).

In Figure 4 (B) we cannot see anything but two spots showing higher values. The color scale is divided into arithmetic scale in such a way that practically all of the area is painted red.

Now we have 27 exhaustive data sets representing 27 different spatial phenomena. From these exhaustive data sets we have drawn a sample based on the stratified random sampling technique. Moreover, all samples have the same locations as shown in Figure 5.

Summary statistics for all 27 samples are presented in Table 5.

Regarding semivariogram models we have to compute experimental semivariograms for all logarithm transformed data and only one semivariogram for the indicator variable. Semivariogram models for logarithm transformed data look like the semivariogram model shown in Figure 6, with a range equal to 12 and sills scaled according to the constant K. Table 6 presents sills for all semivariogram models for logarithm transform data.

The semivariogram model for the indicator variable (Figure 7) is the same for all samples because all samples have the same location and data points were calculated using equation (17).

This paper intends to compare both approaches in terms of local precision and associated uncertainties. Both lognormal kriging and indicator kriging were run, by which we wanted to estimate a regular grid of 50 by 50 nodes that is exactly equal to the exhaustive grids. Instead of 2500 nodes, we estimate 2290 nodes located within the convex hull (Figure 8).

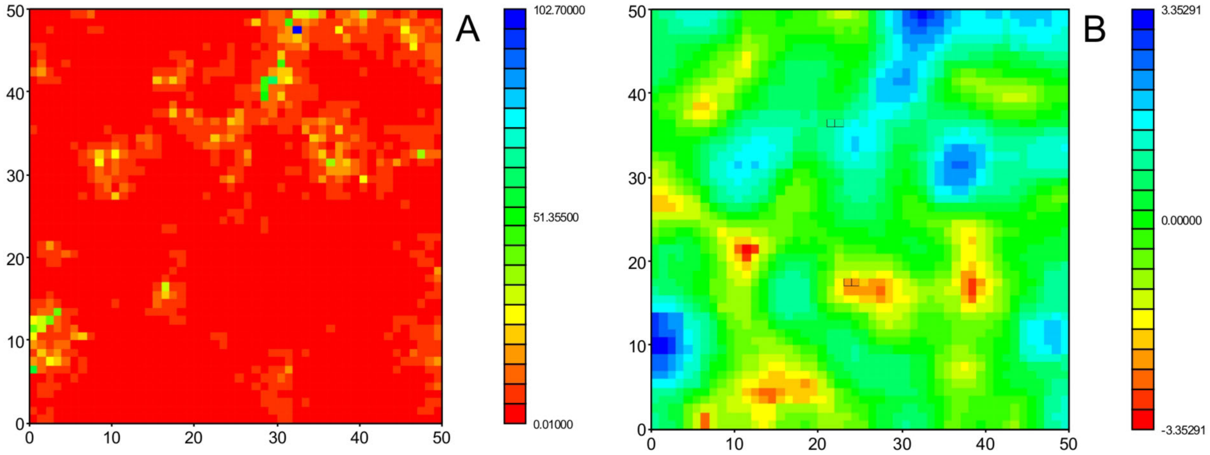

FIGURE 2. Image map of the secondary variable (A) and of the normal score transformed variable (B).

TABLE 2. Population parameters for the secondary variable and after normal score transform.

Parameters	Secondary variable	Normal score
N	2500	2500
Mean	2.580	0.000
Standard deviation	5.151	0.997
Coef. variation	1.996	Undefined
Maximum	102.70	3.353
Upper quartile	2.555	0.674
Median	0.959	-0.001
Lower quartile	0.333	0.675
Minimum	0.010	-3.353

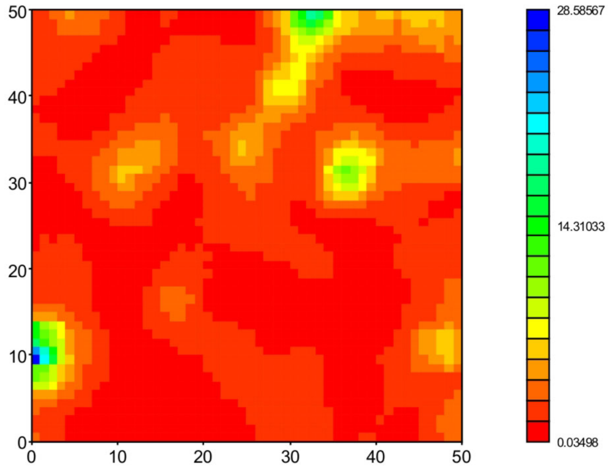

FIGURE 3. Image of a typical lognormal distribution.

TABLE 3. Population parameters for the new random variable presenting lognormal distribution.

Parameters	Z_{Log}
N	2500
Mean	1.640
Standard deviation	2.057
Coef. variation	1.254
Maximum	28.586
Upper quartile	1.961
Median	0.999
Lower quartile	0.509
Minimum	0.035

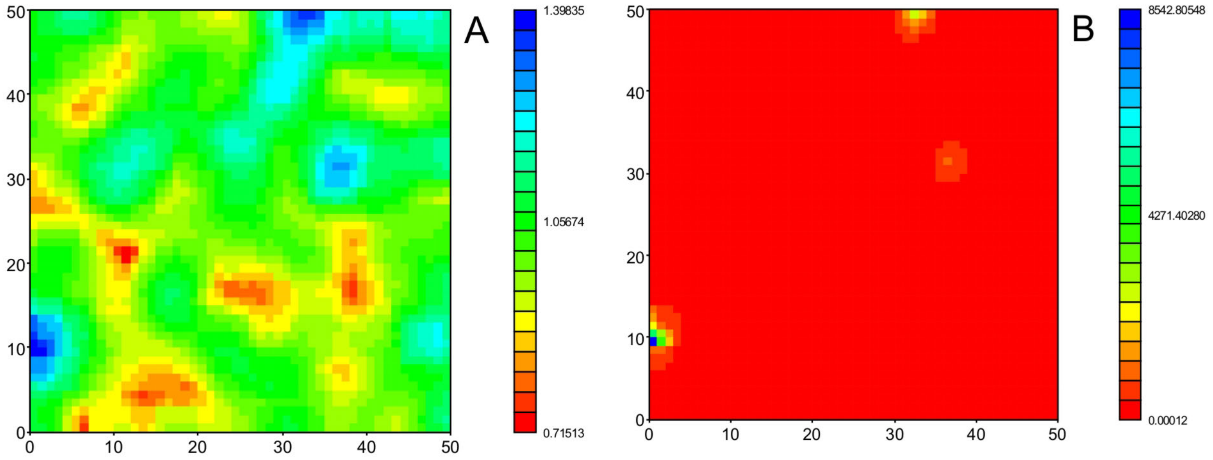

FIGURE 4. Image maps for exhaustive data sets generated after expression (16) with K=1 (A) and with K=27 (B).

TABLE 4. Population parameters for K = 1 and for K = 27.

Parameters	Z_{Log1}	Z_{Log27}
N	2500	2500
Mean	1.005	30.368
Standard deviation	0.100	246.236
Coef. variation	0.100	8.108
Maximum	1.398	8542.805
Upper quartile	1.070	6.163
Median	1.000	0.999
Lower quartile	0.935	0.162
Minimum	0.715	0.000

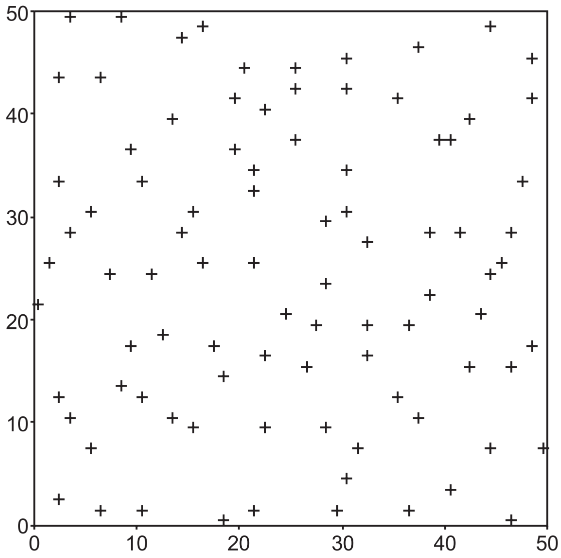

FIGURE 5. Location map for samples drawn from exhaustive data sets (sample size = 90).

TABLE 5. Summary statistics for samples drawn from exhaustive data sets (all samples are composed of 90 data points).

Var	Mean	Sdv	CV	Max	UQ	Med	LQ	Min
ZLog1	1.001	0.093	0.093	1.246	1.052	0.998	0.933	0.772
ZLog2	1.010	0.189	0.187	1.552	1.108	0.996	0.871	0.596
ZLog3	1.028	0.291	0.283	1.933	1.166	0.995	0.813	0.460
ZLog4	1.056	0.403	0.382	2.408	1.227	0.993	0.759	0.355
ZLog5	1.093	0.530	0.485	3.000	1.291	0.991	0.708	0.274
ZLog6	1.142	0.676	0.592	3.738	1.359	0.990	0.661	0.211
ZLog7	1.202	0.848	0.705	4.656	1.431	0.988	0.617	0.163
ZLog8	1.277	1.053	0.825	5.801	1.506	0.986	0.576	0.126
ZLog9	1.367	1.299	0.950	7.226	1.586	0.984	0.537	0.097
ZLog10	1.476	1.597	1.082	9.002	1.669	0.983	0.502	0.075
ZLog11	1.606	1.959	1.220	11.214	1.757	0.981	0.468	0.058
ZLog12	1.761	2.401	1.364	13.970	1.850	0.979	0.437	0.045
ZLog13	1.945	2.941	1.512	17.404	1.948	0.977	0.408	0.034
ZLog14	2.165	3.604	1.664	21.681	2.051	0.976	0.381	0.027
ZLog15	2.427	4.416	1.820	27.009	2.160	0.974	0.355	0.021
ZLog16	2.739	5.415	1.977	33.647	2.274	0.972	0.332	0.016
ZLog17	3.112	6.643	2.135	41.915	2.395	0.971	0.309	0.012
ZLog18	3.557	8.153	2.292	52.216	2.522	0.969	0.289	0.009
ZLog19	4.090	10.013	2.448	65.049	2.656	0.967	0.270	0.007
ZLog20	4.729	12.305	2.602	81.035	2.797	0.966	0.252	0.006
ZLog21	5.495	15.128	2.753	100.951	2.946	0.964	0.235	0.004
ZLog22	6.417	18.617	2.900	125.760	3.103	0.962	0.219	0.003
ZLog23	7.526	22.904	3.043	156.667	3.268	0.960	0.205	0.003
ZLog24	8.865	28.202	3.181	195.169	3.443	0.959	0.191	0.002
ZLog25	10.481	34.741	3.315	243.133	3.626	0.957	0.178	0.002
ZLog26	12.436	42.814	3.443	302.885	3.820	0.955	0.166	0.001
ZLog27	14.804	52.784	3.566	377.321	4.024	0.954	0.155	0.001

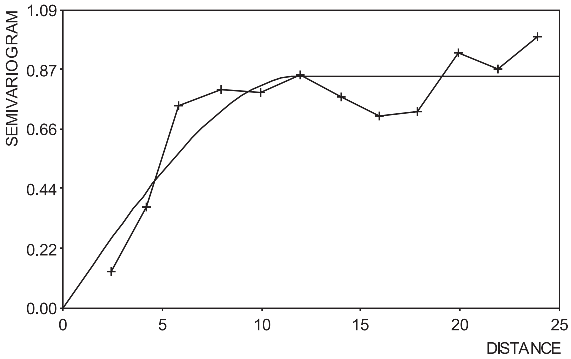

FIGURE 6. Semivariogram model computed for K=10 (lognormal data) after logarithm transform.

TABLE 6. Sill values according to the constant K.

K	Sill	K	Sill	K	Sill
1	0.00831	10	0.85092	19	3.08402
2	0.03424	11	1.02732	20	3.39226
3	0.07724	12	1.22928	21	3.76525
4	0.13571	13	1.44705	22	4.10953
5	0.21656	14	1.65888	23	4.49576
6	0.30501	15	1.91092	24	4.88800
7	0.41587	16	2.16976	25	5.29681
8	0.54544	17	2.44217	26	5.73418
9	0.68831	18	2.75343	27	6.18572

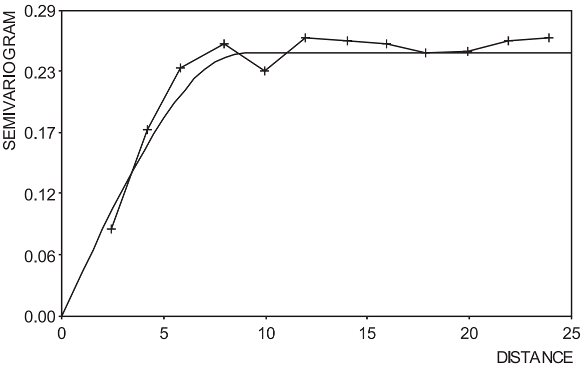

FIGURE 7. Semivariogram model for the median indicator variable.

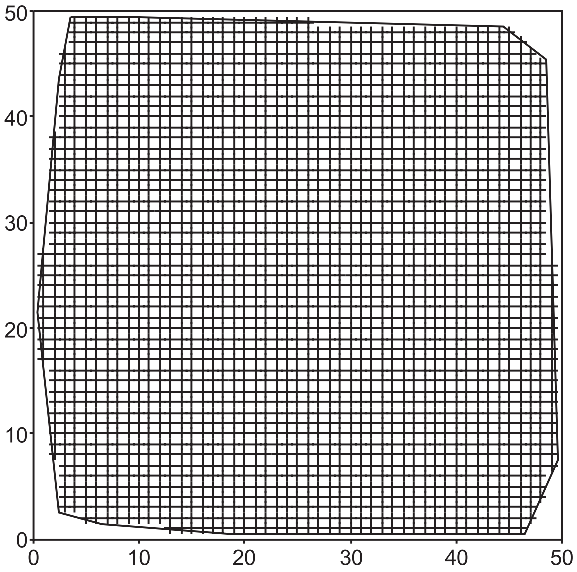

FIGURE 8. Regular grid within the convex hull, calculated after Yamamoto (1997).

RESULTS AND DISCUSSION

First of all the results for lognormal kriging estimates are shown. Actually, back-transformed estimates after expressions (4) and (5) are examined. Tables 7 and 8 present summary statistics for back-transformed lognormal kriging estimates.

Next, the results for indicator kriging that is the conditional mean (E-type estimate) calculated as equation (14), are shown. Summary statistics for E-type estimates are shown in Table 9.

Thus we want to know how different methods work when compared to the samples. Actually, the samples are taken as a representation of the population that is the object of study and therefore the closer the estimates are to the sample data the best inference we can do about the population. Figures 9, 10 and 11 show box plots for back-transformed lognormal kriging estimates after equation (4), after equation (5) and E-type estimates from conditional distributions built from indicator kriging approach, respectively.

Comparing equations (4) and (5) it is possible to verify that Journel's approach (Journel, 1980) produces estimates that are mean biased as reported in literature such as (Journel & Huijbregts, 1978). However, the median of back-transformed estimates are not biased. Moreover, these estimates do not reproduce the full variability of data sets. The approach after Yamamoto (2007) presents the best results, reproducing all basic statistics as close as possible to the sample data.

Examining Figure 11 it is possible to assert that E-type estimates present means very close to the sample means. However, medians are strongly biased because of the loss of information on the lower tail (see minimum values). The upper tails of distributions are reasonably well reproduced by the indicator approach.

We can compare the different approaches by comparing their cumulative frequency distributions. Actually, just the back-transformed lognormal kriging estimates after equation (5) and E-type estimates derived from the indicator kriging approach are compared to sample data because of limitations in the computer program used for three distributions. Instead of showing all 27 samples we present six samples that illustrate the performance of different approaches for estimating lognormal data (Figure 12).

Figure 12 just reconfirms what was seen in previous figures, which is that the best approach is provided by the back-transformed estimates after equation (5). Although E-type estimates are not mean biased, distributions get further from the sample distribution as the coefficient of variation increases. Therefore, lognormal kriging seems to be the best approach for lognormal data. However, it is clear that this approach presents best results after correcting the smoothing effect of ordinary kriging estimates by the use of equation (5).

Since we departed from exhaustive data sets we know the real value at every estimated location. Thus, we can compare estimates in terms of local precision by computing correlation coefficients, RMS errors, mean errors and mean absolute errors (Figure 13). In terms of correlation coefficients the indicator approach shows the lower values and between the two approaches for back-transforming estimates equation (5) provides better correlations for data sets 1 to 12

TABLE 7. Summary statistics for back-transformed lognormal kriging estimates after equation (4).

Var	Mean	Sdv	CV	Max	UQ	Med	LQ	Min
ZLog1	0.996	0.057	0.058	1.226	1.034	0.997	0.956	0.790
ZLog2	0.996	0.115	0.116	1.502	1.069	0.993	0.913	0.623
ZLog3	0.999	0.175	0.175	1.842	1.106	0.990	0.873	0.492
ZLog4	1.005	0.237	0.235	2.257	1.143	0.986	0.834	0.389
ZLog5	1.015	0.302	0.297	2.767	1.182	0.983	0.797	0.307
ZLog6	1.029	0.372	0.362	3.391	1.222	0.980	0.762	0.242
ZLog7	1.046	0.448	0.429	4.157	1.264	0.976	0.728	0.191
ZLog8	1.067	0.532	0.499	5.095	1.307	0.973	0.695	0.151
ZLog9	1.092	0.625	0.573	6.246	1.351	0.970	0.605	0.119
ZLog10	1.122	0.730	0.651	7.656	1.397	0.966	0.635	0.094
ZLog11	1.156	0.849	0.735	9.384	1.445	0.963	0.607	0.074
ZLog12	1.196	0.986	0.824	11.502	1.494	0.960	0.580	0.059
ZLog13	1.242	1.142	0.920	14.099	1.545	0.956	0.554	0.046
ZLog14	1.294	1.324	1.023	17.281	1.597	0.953	0.530	0.037
ZLog15	1.354	1.536	1.135	21.182	1.651	0.950	0.506	0.029
ZLog16	1.421	1.783	1.255	25.964	1.708	0.947	0.484	0.023
ZLog17	1.497	2.074	1.385	31.825	1.766	0.943	0.462	0.018
ZLog18	1.583	2.416	1.526	39.009	1.826	0.940	0.442	0.014
ZLog19	1.681	2.820	1.678	47.815	1.888	0.937	0.422	0.011
ZLog20	1.792	3.299	1.841	58.609	1.952	0.934	0.403	0.009
ZLog21	1.917	3.867	2.018	71.839	2.018	0.931	0.385	0.007
ZLog22	2.059	4.543	2.207	88.056	2.087	0.927	0.368	0.006
ZLog23	2.220	5.348	2.410	107.934	2.158	0.924	0.352	0.004
ZLog24	2.402	6.309	2.626	132.299	2.231	0.921	0.336	0.003
ZLog25	2.611	7.458	2.857	162.165	2.307	0.918	0.321	0.005
ZLog26	2.848	8.832	3.101	198.772	2.386	0.915	0.307	0.002
ZLog27	3.119	10.497	3.360	243.642	2.467	0.912	0.293	0.002

TABLE 8. Summary statistics for back-transformed lognormal kriging estimates after equation (5).

Var	Mean	Sdv	CV	Max	UQ	Med	LQ	Min
ZLog1	1.000	0.092	0.092	1.245	1.060	0.998	0.938	0.772
ZLog2	1.009	0.186	0.185	1.549	1.123	0.997	0.880	0.596
ZLog3	1.027	0.286	0.278	1.928	1.190	0.995	0.825	0.460
ZLog4	1.054	0.394	0.374	2.400	1.261	0.993	0.774	0.355
ZLog5	1.090	0.515	0.473	2.987	1.337	0.992	0.726	0.274
ZLog6	1.136	0.654	0.576	3.718	1.417	0.990	0.681	0.211
ZLog7	1.195	0.816	0.683	4.628	1.502	0.989	0.639	0.163
ZLog8	1.266	1.007	0.795	5.760	1.591	0.987	0.600	0.126
ZLog9	1.351	1.234	0.913	7.170	1.687	0.985	0.562	0.097
ZLog10	1.453	1.506	1.036	8.924	1.787	0.984	0.528	0.075
ZLog11	1.575	1.835	1.165	11.107	1.894	0.982	0.495	0.058
ZLog12	1.720	2.234	1.299	13.825	2.007	0.980	0.464	0.045
ZLog13	1.891	2.719	1.438	17.208	2.128	0.979	0.435	0.034
ZLog14	2.093	3.312	1.582	21.418	2.255	0.977	0.408	0.027
ZLog15	2.332	4.036	1.731	26.658	2.390	0.976	0.383	0.021
ZLog16	2.615	4.922	1.883	33.181	2.532	0.974	0.359	0.016
ZLog17	2.950	6.010	2.037	41.300	2.684	0.972	0.337	0.012
ZLog18	3.347	7.346	2.195	51.404	2.844	0.971	0.316	0.009
ZLog19	3.819	8.988	2.353	63.982	3.014	0.969	0.297	0.007
ZLog20	4.381	11.009	2.513	79.636	3.195	0.968	0.278	0.006
ZLog21	5.050	13.498	2.673	99.121	3.386	0.966	0.261	0.004
ZLog22	5.849	16.566	2.832	123.374	3.588	0.964	0.245	0.003
ZLog23	6.804	20.349	2.991	153.560	3.803	0.963	0.230	0.003
ZLog24	7.948	25.017	3.147	191.132	4.030	0.961	0.215	0.002
ZLog25	9.321	30.780	3.302	237.898	4.271	0.960	0.202	0.002
ZLog26	10.970	46.697	3.605	368.554	4.797	0.956	0.178	0.001
ZLog27	12.954	46.697	3.605	368.554	4.797	0.956	0.178	0.001

TABLE 9. Summary statistics for E-type estimates from ordinary kriging approach after equation (5).

Var	Mean	Sdv	CV	Max	UQ	Med	LQ	Min
ZLog1	0.999	0.054	0.055	1.217	1.034	0.999	0.959	0.809
ZLog2	1.006	0.110	0.109	1.486	1.074	1.004	0.927	0.655
ZLog3	1.022	0.168	0.165	1.819	1.119	1.011	0.903	0.534
ZLog4	1.046	0.231	0.221	2.232	1.174	1.025	0.884	0.438
ZLog5	1.080	0.302	0.279	2.742	1.236	1.042	0.869	0.362
ZLog6	1.123	0.382	0.340	3.373	1.306	1.066	0.863	0.303
ZLog7	1.178	0.474	0.402	4.154	1.386	1.096	0.859	0.256
ZLog8	1.246	0.583	0.468	5.120	1.479	1.129	0.862	0.220
ZLog9	1.328	0.711	0.536	6.314	1.579	1.169	0.870	0.192
ZLog10	1.426	0.865	0.607	7.790	1.698	1.213	0.885	0.170
ZLog11	1.542	1.050	0.681	9.615	1.828	1.264	0.905	0.149
ZLog12	1.681	1.274	0.758	11.872	1.985	1.322	0.925	0.132
ZLog13	1.846	1.545	0.837	14.663	2.171	1.392	0.946	0.119
ZLog14	2.091	1.874	0.918	18.114	2.409	1.461	0.965	0.108
ZLog15	2.272	2.274	1.001	22.381	2.653	1.536	0.994	0.100
ZLog16	2.546	2.762	1.085	27.660	2.945	1.624	1.034	0.093
ZLog17	2.871	3.356	1.169	34.188	3.279	1.730	1.069	0.088
ZLog18	3.257	4.082	1.253	42.262	3.708	1.837	1.108	0.085
ZLog19	3.716	4.968	1.337	52.250	4.137	1.977	1.155	0.083
ZLog20	4.263	6.050	1.419	64.606	4.624	2.128	1.198	0.082
ZLog21	4.915	7.374	1.500	79.891	5.181	2.277	1.253	0.082
ZLog22	5.693	8.993	1.580	98.803	5.954	2.440	1.321	0.077
ZLog23	6.625	10.975	1.657	122.202	6.830	2.623	1.388	0.070
ZLog24	7.741	13.402	1.731	151.155	7.839	2.825	1.453	0.063
ZLog25	9.079	16.374	1.803	186.983	8.977	3.069	1.524	0.056
ZLog26	10.688	20.017	1.873	231.321	10.269	3.295	1.603	0.051
ZLog27	12.623	24.482	1.940	286.194	11.921	3.564	1.687	0.046

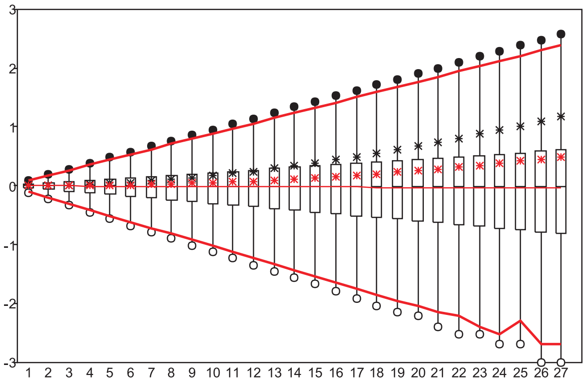

FIGURE 9. Box plots for all samples compared with back-transformed lognormal kriging estimates after equation (4). Legend: box = lower quartile, median and upper quartile of sample statistics; star = mean; open circle = minimum; full circle = maximum; black = sample and red = estimates. All values are represented in logarithmic scale.

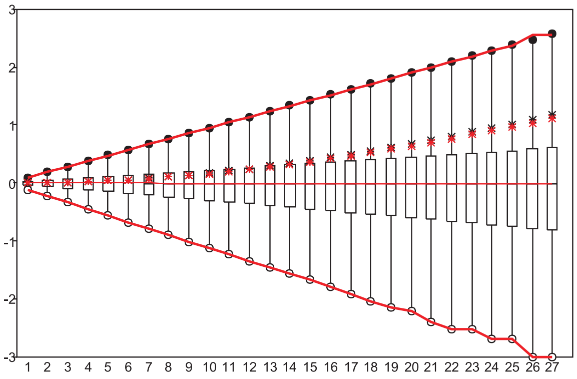

FIGURE 10. Box plots for all samples compared with back-transformed lognormal kriging estimates after equation (5). Legend: box = lower quartile, median and upper quartile of sample statistics; star = mean; open circle = minimum; full circle = maximum; black = sample and red = estimates. All values are represented in logarithmic scale.

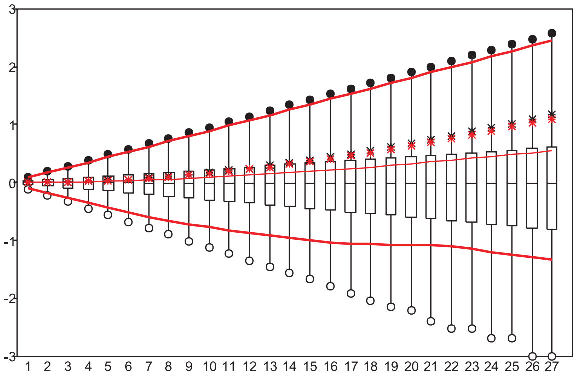

FIGURE 11. Box plots for all samples compared with E-type estimates derived from indicator kriging conditional distributions. Legend: box = lower quartile, median and upper quartile of sample statistics; star = mean; open circle = minimum; full circle = maximum; black = sample and red = estimates. All values are represented in logarithmic scale.

whereas equation (4) gives better correlations for data sets from 13 to 27. RMS errors are very close to each other and all methods give approximately the same values. Mean errors are expected to be close to zero. In our case study, back-transformed estimates after equation (4) gave the poorest results. Mean absolute errors both show back-transforming approaches giving errors close to each other, but E-type estimates from the indicator kriging approach results in the largest errors. In general, the indicator approach seems to present poorer results when compared to lognormal kriging.

Figure 13 shows both methods for back transforming lognormal kriging estimates producing very similar results because we are examining the mean values that resulted from averaging over 2290 data. Then, what is the difference between equations (4) and (5) for back-transforming lognormal kriging estimates? Comparing scattergrams presented in Figure 14 we verify that both correlations to the exhaustive data are similar to each other, but the slopes of the regression lines are different, being that Figure 14A shows a slope greater than one and in Figure 14B the slope is less than one. Figure 15 illustrates a scattergram of actual values versus E-type estimates from indicator kriging in which the regression line has a slope greater than one.

The slope of the regression line is calculated as (Yamamoto, 2005):

$$Slope = \frac{\rho_{X,Y} * S_Y}{S_X}$$

where $\rho_{X,Y}$ is the correlation coefficient; S_X is the standard deviation for variable X and S_Y is the standard deviation for variable Y.

The slopes of the regression lines on scattergrams were calculated for all data and illustrated in Figure 16. Looking at this figure we can verify certain differences. As we can see just back-transformed estimates after equation (5) present slopes closer to one, while for the other approaches slopes are always greater than one, showing that these two last approaches present some smoothing effect. The only method which removes this effect is the back-transformation after equation (5). It is important to observe that the smoothing removal does not mean loss of local accuracy and that the corrected estimates reproduce the sample histogram (Figure 12).

This way it is possible to examine the uncertainties associated with both lognormal kriging and indicator kriging. As we know, lognormal data present the proportional effect, which means that the variance increases when data values increase. Finney (1941) realized that a number of biological and other populations show the standard error of an individual observation approximately proportional to the magnitude of the observations. According to Manchuk et al. (2009) the proportional effect is becoming important as long as geostatistical procedures involve complex data sets and geometrically complicated models presenting unstructured grids and map elements of variable size.

On the other hand, Rocha & Yamamoto (2000) showed that for distributions presenting negative skewness the variance decreases when data values increase.

In this case study we are handling lognormal data, thus it is interesting to examine the relationship between estimates and uncertainties. For lognormal kriging we have considered the relationship between back-transformed interpolation standard deviations (equation 6) and back-transformed estimates after equation (4). For the indicator kriging approach we used the conditional standard deviations versus E-Type estimates. For these pairs of variables correlation coefficients were computed as displayed in Figure 17. As we can see in this figure, correlation coefficients increase as the coefficient of variation increases. The lognormal approach presents correlation coefficients a little bit greater than those shown by the indicator kriging approach. Data sets presenting coefficients of variation greater than 1.254 can be considered lognormal distributions. Examining Table 5 it is possible to verify that variable ZLog11 presents a coefficient of variation equal to 1.220, which is very close to. 1.254. In Figure 17 we can see that for coefficients of variation greater than 1.220 correlation coefficients do not increase as much, reaching a sill after data set number 18 approximately.

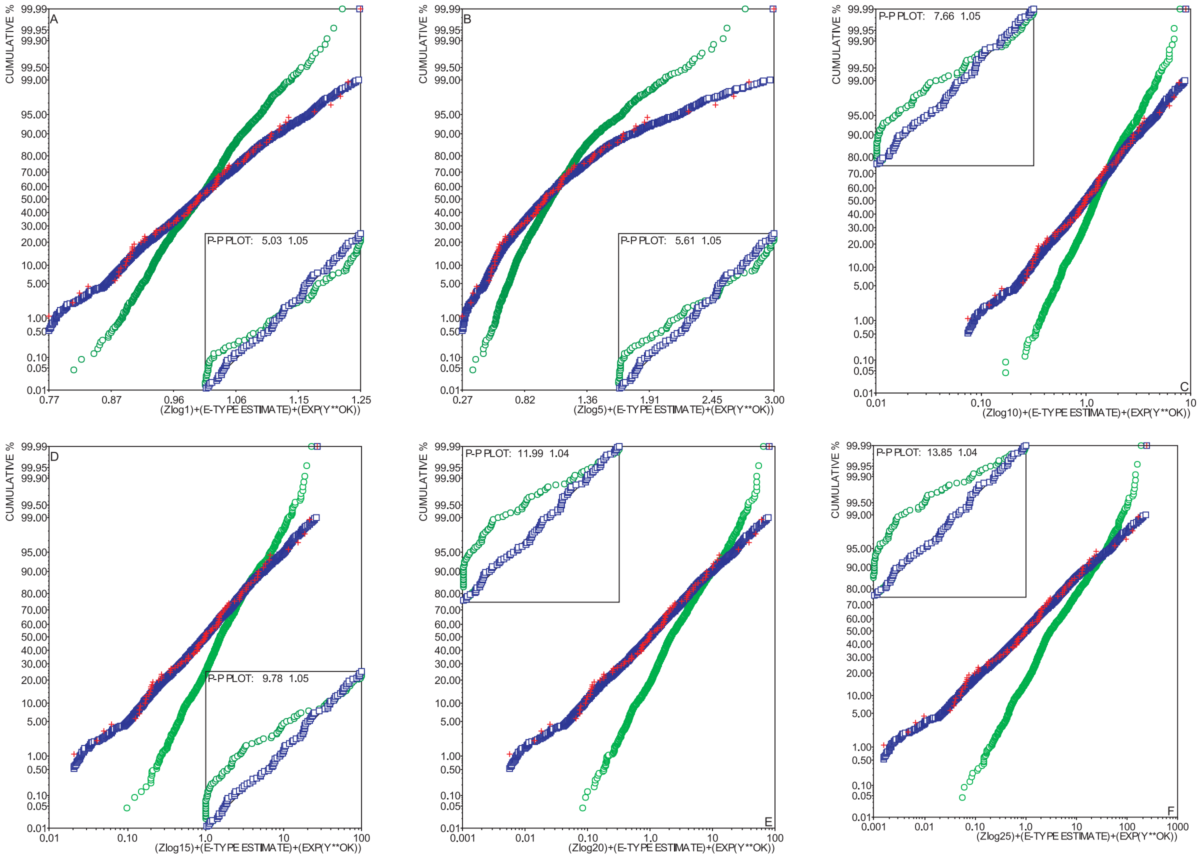

FIGURE 12. Comparing estimated cumulative frequency distributions with sample distributions. Legend: red cross = sample; green circle = E-type estimates from indicator kriging; blue square = back-transformed lognormal kriging estimates after equation (5).

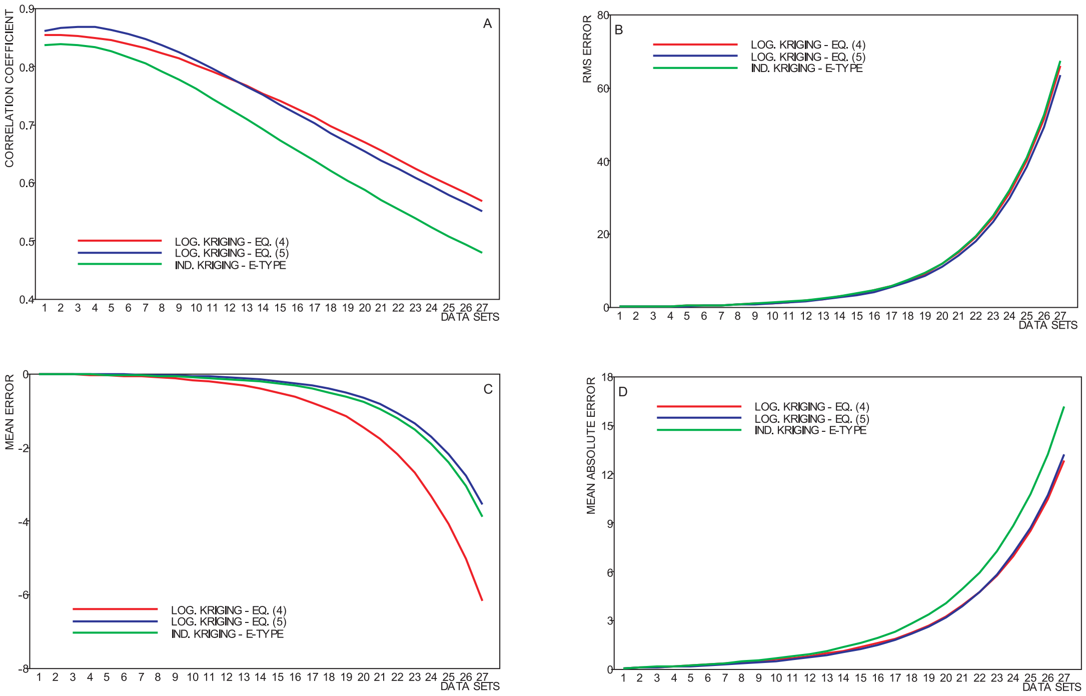

FIGURE 13. Statistics comparing real and estimated values: correlation coefficient (A); RMS error (B); Mean error (C); Mean absolute error (D).

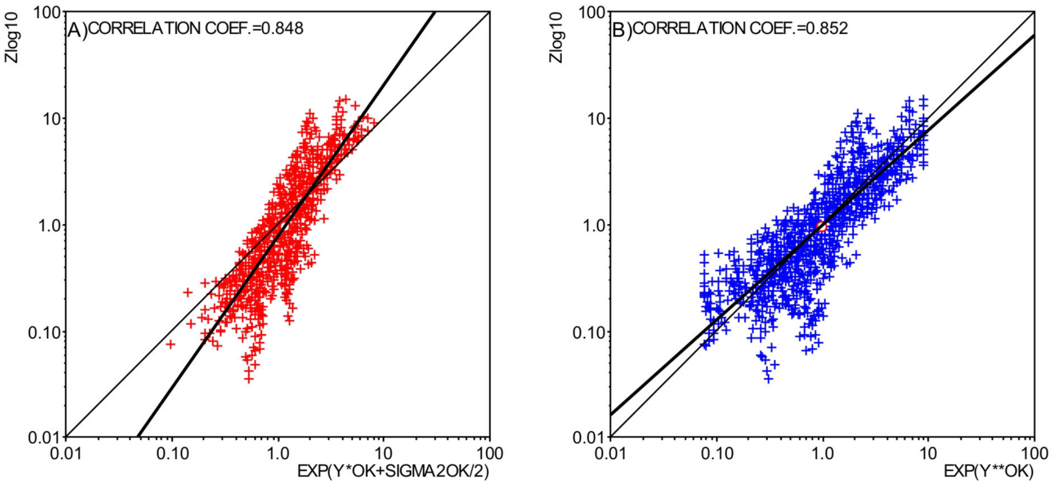

FIGURE 14. Scattergrams of actual values versus lognormal kriging estimates: (A) back-transformed estimates after equation (4); (B) back-transformed estimates after equation (5).

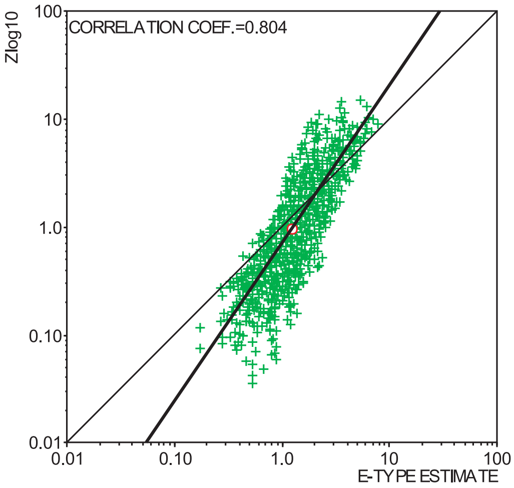

FIGURE 15. Scattergrams of actual values versus E-type estimates from indicator kriging.

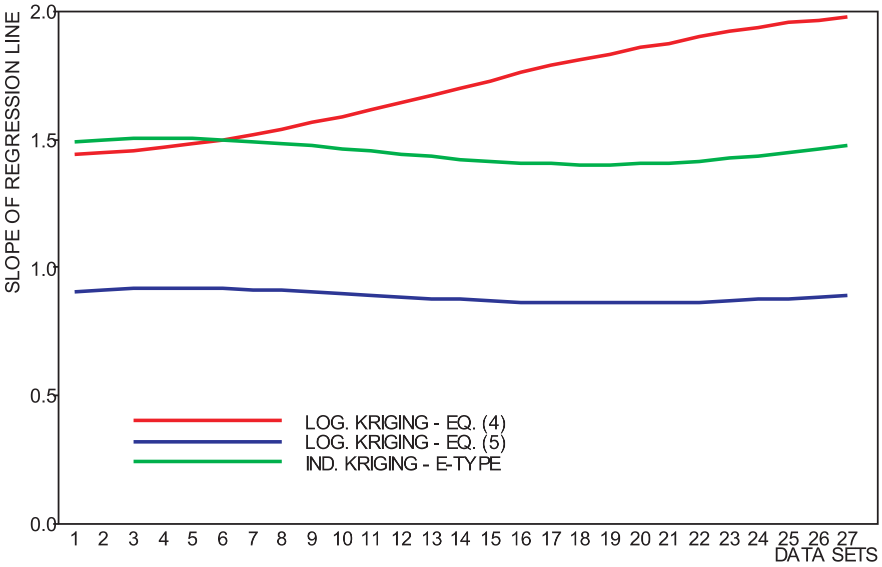

FIGURE 16. Slopes of regression lines calculated on scattergrams.

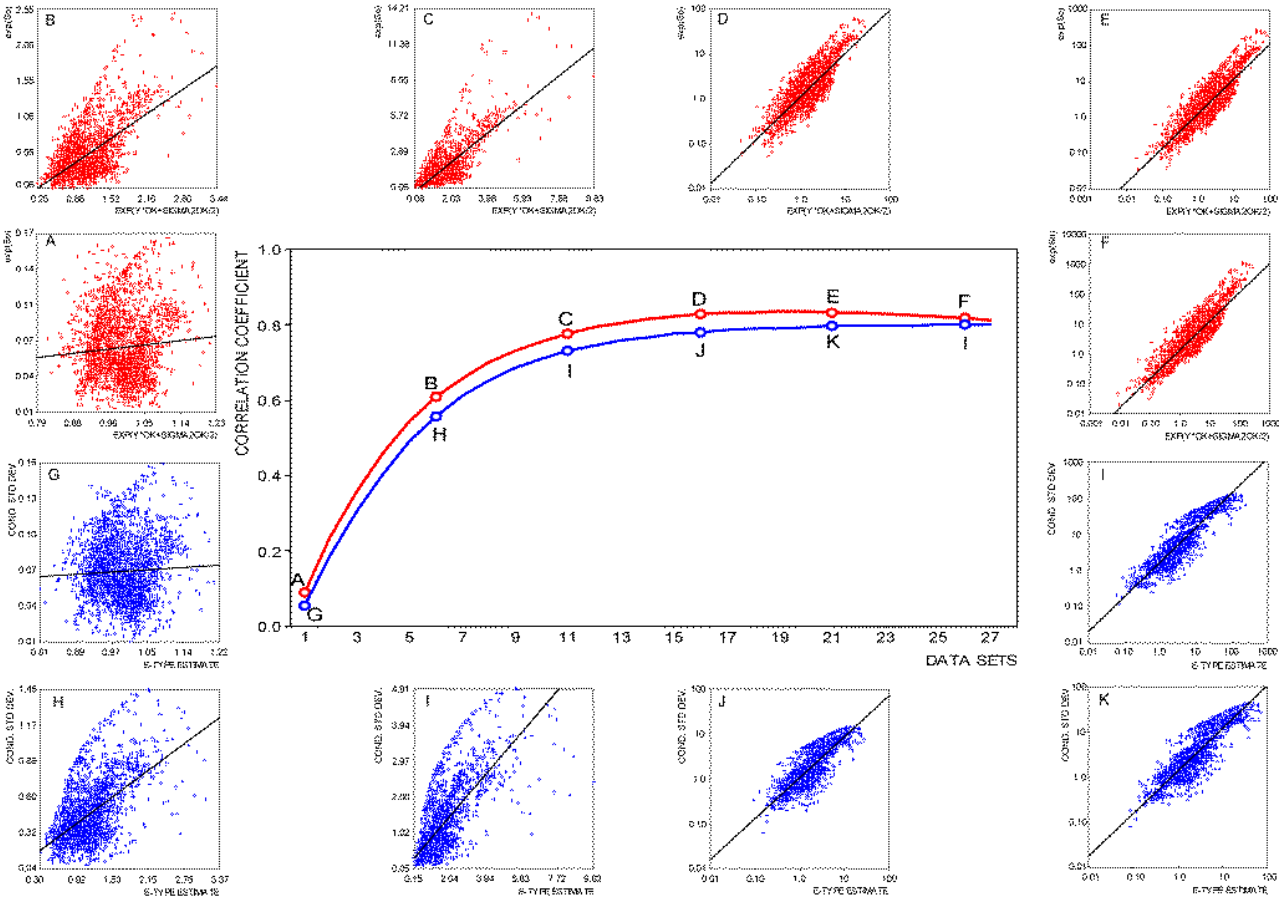

FIGURE 17. Correlation coefficients showing the proportional effect of lognormal data. Legend: lognormal kriging approach (red); indicator kriging approach (blue).

CONCLUSIONS

In this paper two approaches for estimating lognormal data were examined. A systematic and intensive study was carried out to verify the performance of the mentioned methods. The lognormal kriging approach is still the best approach to lognormal data. Equation (5) provides back-transformed lognormal kriging estimates that are closer to sample data. Actually, the closer estimates are to the sample data, the better is the inference about the population. Since we do not know anything about the population that the sample comes from, the best solution is to retain estimates of the spatial phenomenon as close as possible to sample statistics. In this sense, equation (5) provided estimated distributions that are not mean or median biased. Although indicator kriging resulted in estimates with unbiased means, the other basic statistics are very poor when compared with sample statistics. Both approaches represent very well the proportional effect for lognormal distributions.

BIBLIOGRAPHIC REFERENCES

1. DEUTSCH, C.V. & JOURNEL, A.G. **GSLib: geostatistical software library and user's guide**. New York: Oxford University Press, 340 p., 1992.
2. FINNEY, D.J. On the distribution of a variate whose logarithm is normally distributed. **Journal Royal Statistical Society**, Supp. 7, p. 155-161, 1941.
3. HOHN, M.E. **Geostatistics and petroleum geology**. Dordrecht: Kluwer Academic Publishers, 235 p., 1999.
4. JOURNEL, A.G. The lognormal approach to predicting local distribution of selective mining unit grades. **Mathematical Geology**, v. 12, p. 285-303, 1980.
5. JOURNEL, A.G. Nonparametric estimation of spatial distributions. **Mathematical Geology**, v. 15, p. 445-468, 1983.
6. JOURNEL, A.G. & HUIJBREGTS, C.J. **Mining geostatistics**. London: Academic Press, 600 p., 1978.

7. JOURNEL, A.G. & ROSSI, M.E. When do we need a trend model in kriging?. **Mathematical Geology**, v. 21, p. 715-739, 1989.
8. MANCHUK, J.G.; LEUANGTHONG, O.; DEUTSCH, C.V. The proportional effect. **Mathematical Geoscience**, v. 41, p. 799-816, 2009.
9. ROCHA, M.M. & YAMAMOTO, J.K. Comparison between kriging variance and interpolation variance as uncertainty measurements in the Capanema Iron Mine, State of Minas Gerais – Brazil. **Natural Resources Research**, v. 9, p. 223-235, 2000.
10. SAITO, H. & GOOVAERTS, P. Geostatistical interpolation of positively skewed and censored data in a Dioxin-contaminated site. **Environmental Science Technology**, v. 34, p. 4228-4235, 2000.
11. YAMAMOTO, J.K. CONVEX_HULL: a Pascal program for determining the convex hull for planar sets. **Computers & Geosciences**, v. 23, p. 725-738, 1997.
12. YAMAMOTO, J.K. An alternative measure of the reliability of ordinary kriging estimates. **Mathematical Geology**, v. 32, p. 489-509, 2000.
13. YAMAMOTO, J.K. Correcting the smoothing effect of ordinary kriging estimates. **Mathematical Geology**, v. 37, p. 69-94, 2005.
14. YAMAMOTO, J.K. On unbiased backtransform of lognormal kriging estimates. **Computers & Geosciences**, v. 11, p. 219-234, 2007.
15. YAMAMOTO, J.K. Assessing uncertainties for lognormal kriging. In: ZHANG, J. & GOODCHILD, M.F. (Eds.), **Spatial uncertainty**. Proceedings of the 8th International Symposium on Spatial Accuracy in Natural Resources and Environmental Sciences, v. 1, p. 62-69, 2008.

Manuscrito Recebido em: 23 de dezembro de 2009
Revisado e Aceito em: 15 de março de 2010