# Validity Range of the Discrete Gaussian Change-of-Support Model and Its Variant

Jean-Paul Chilès[1]

**Abstract** The discrete Gaussian model is a very popular change-of-support model. It is designed for a stationary random function $Z(x)$ that is regarded as the transform of a stationary random function $Y(x)$ with Gaussian marginal. It is based on an assumption that is not fully true and thus constitutes an approximation to the exact solution. We examine the effect of this approximation on the modeling of the marginal distribution of block values of a lognormal random function. The initial discrete Gaussian model and a variant of that model are considered.

## Introduction

In many situations decision makers are faced with selectivity: extract ore above some cutoff grade, remediate soils whose pollution exceeds some threshold, restrict traffic speed when the ozone concentration is exceeds some limiting value. In each case, data relate to a small support (e.g., a core) that we regard as a point, whereas the decision is taken for a much larger support: selective mining units, decontamination units, average concentration in a city in a given time interval (typically with a low threshold applied to the daily average concentration and a higher threshold for an hourly average concentration). To predict the effect of selectivity it is necessary to take into account that the grade or concentration distribution is all the less dispersed as the support is large. In the framework of the random function theory the result depends on the whole spatial distribution. It can be obtained by Monte Carlo simulations (nonconditional simulations in the global case, conditional simulations in the local case). Even if Monte Carlo simulations are more accessible now than in the past, there are always situations where they require too intensive computations. It is therefore useful to have access to the approximate solutions provided by change-of-support models. Many such models have been developed by Georges Matheron (a synthesis is given in chapter 6 of [1]). The most popular model is the discrete Gaussian model. It should not be used in any situation because it has been developed for random functions such as transforms of stationary Gaussian random functions. The initial model was

---

[1] Center of Geosciences and Geoengineering, MINES ParisTech, 35 rue Saint-Honoré, 77300 Fontainebleau, France
jean-paul.chiles@mines-paristech.fr

proposed in 1976 by Matheron [5]. A simpler variant was proposed in 2007 by Emery [3]. We investigate here the accuracy of this change-of-support model for the modeling of the marginal distribution of block values. This investigation is carried out in the special case of lognormal random functions, which constitutes a large and important class of random functions.

We first recall the assumptions of the discrete Gaussian model and its variant, and explain why they lead to approximations to the true solution. We then describe the principle of the validation method. Finally we explore the validity range of the model and its variant depending on space dimension, variance and covariance function of the logarithmic variable, and support size. In the conclusion we give some indications concerning the modeling of the local block distribution.

## The Discrete Gaussian Model

Let us consider a stationary random function (SRF) $Z(x)$ that can be expressed as the transform of an SRF $Y(x)$ with standard normal marginal distribution. It is therefore of the form $Z(x) = \varphi(Y(x))$ with the transformation function $\varphi = F^{-1} \circ G$, where $F$ is the marginal c.d.f. of $Z(\cdot)$ and $G$ the standard normal c.d.f. Similarly, we can consider that the mean grade $Z(v)$ of the block $v$ is of the form $Z(v) = \varphi_v(Y_v)$ where $Y_v$ is a standard normal random variable and $\varphi_v$ the block transformation function which we want to determine (we assume strict stationarity of $Y$ to ensure that $\varphi_v$ does not depend on the location of the block).

### Model DGM1

Consider a uniform random point $\underline{x}$ within the block $v$. The random variable $Z(\underline{x})$ has for c.d.f. the marginal distribution $F$ of the SRF $Z(\cdot)$ and can be expressed as the transform $\varphi(Y(\underline{x}))$ of the random variable $Y(\underline{x})$. The basic assumption of the discrete Gaussian model proposed by Matheron [5], hereafter referenced as model DGM1, is that the bivariate distribution of the $(Y(\underline{x}), Y_v)$ pair is Gaussian, with a positive correlation coefficient $r$. Matheron deduces from Cartier's relation that the block transformation function $\varphi_v$ is given by

$$\varphi_v(y_v) = \int \varphi(r\, y_v + \sqrt{1-r^2}u)\, g(u)\, du \qquad (1)$$

This defines the distribution of $Z(v)$.

In practice φ is expressed through its Hermite polynomial expansion

$$\varphi(y) = \sum_{n=0}^{\infty} \varphi_n \, \chi_n(y) \tag{2}$$

where the $\chi_n$'s are the normalized Hermite polynomials (see, e.g., [1], Appendix A.5) and the coefficients $\varphi_n$ are given by

$$\varphi_n = \int \varphi(y) \, \chi_n(y) \, G(dy) \tag{3}$$

Relation (1) then implies that $\varphi_v$ can be expressed in the form

$$\varphi_v(y) = \sum_{n=0}^{\infty} \varphi_n \, r^n \chi_n(y) \tag{4}$$

The variance of $Z(v)$ can be derived from the coefficients $\varphi_n \, r^n$ of the expansion of $\varphi_v$ or from the covariance $C(h)$ of the SRF $Z$. For consistency, $r$ is obtained by equating these two expressions, that is

$$\sum_{n=1}^{\infty} \varphi_n^2 \, r^{2n} = \frac{1}{|v|^2} \int_v \int_v C(x'-x) \, dx \, dx' \tag{5}$$

Considered as an equation in $r$, (5) has a unique solution between 0 and 1. The correlation coefficient $r$ is called the change-of-support coefficient.

Extensions of the model (not considered here) enable the local estimation of a block by disjunctive kriging or in a multivariate Gaussian framework.


## Model DGM2

The variant DGM2 proposed by Emery [3] is simpler but requires the additional assumption that the bivariate distribution of $Y(\underline{x})$ and $Y(\underline{x}')$ for two independent random points within the same block $v$ is Gaussian. In that case, Emery shows that $r^2$ is the variance of the average $Y(v)$ of $Y(\cdot)$ in the block $v$:

$$r^2 = \frac{1}{|v|^2} \int_v \int_v \rho(x'-x) \, dx \, dx' \tag{6}$$

where $\rho(h)$ is the covariance (here a correlogram) of the SRF $Y(\cdot)$. Moreover, $Y_v$ is simply the average $Y(v)$ rescaled to a unit variance by the change-of-support coefficient $r$:

$$Y_v = Y(v) / r \tag{7}$$

This induces large simplifications in the extensions of the model to local estimation, notably in the framework of a multivariate Gaussian assumption.

## *Discussion of the Assumptions*

Does there exist random functions satisfying the above assumptions? It is easy to simulate Gaussian samples with pairwise correlation $r$, thus corresponding to samples of $Y$ with independent random locations in the block $v$ (see [3]). But the author does not know of a random function model for fixed locations leading to such correlations for random locations in $v$. Would such a model exist, it would be specific to that support $v$. The above assumptions should therefore be considered as approximate only.

Let us first consider the situation where $Z$ is a Gaussian SRF, that is, $\varphi$ amounts to an affine transformation (in that case we have no need of the discrete Gaussian model but is interesting to see what it would mean). Models DGM1 and DGM2 give the same value for $r$ because $C(h)$ is proportional to $\rho(h)$, and relation (7) is exact. Since $\underline{x}$ is random in $v$, the bivariate distribution of $Y(\underline{x})$ and $Y_v$ is a mixture of standard bivariate normal distributions with correlations

$$\omega(x) = \frac{1}{r|v|} \int_v \rho(x'-x)\,dx' \tag{8}$$

for $x$ having any possible location in $v$. The average value of $\omega(x)$ when $x$ scans $v$ is $r$, so that the average correlation between $Y(x)$ and $Y_v$ is $r$ but, depending on $r$, it can be an average of very different values. A mixture of such bivariate Gaussian distributions is not a bivariate Gaussian distribution: it is a Hermitian bivariate distribution ([4]; see also [1], pp. 418 and 423). In the one-dimensional case, that is, when $v$ is a segment of length $L$, and for an exponential covariance with unit sill and with scale parameter $a$, the change-of-support coefficient $r$ and the function $\omega(x)$ are respectively given by

$$r^2 = \frac{2a^2}{L^2}\left[\exp\left(-\frac{L}{a}\right) - 1 + \frac{L}{a}\right]$$

$$\omega(x) = \frac{1}{r}\frac{a}{L}\left[2 - \exp\left(-\frac{x}{a}\right) - \exp\left(-\frac{L-x}{a}\right)\right]$$

$\omega(x)$ is minimal at the extremes ($x = 0$ or $L$) and maximum for $x = L/2$. The contrast between the largest correlation and the smallest one is

$$\frac{\omega(L/2)}{\omega(0)} = \frac{2}{1 + \exp\left(-\frac{1}{2}L/a\right)}$$

This ratio is equal to 1 for $L/a = 0$, remains close to 1 when $L/a$ remains moderate (with a value of about 1.245 for $L/a = 1$) and increases to 2 when $L/a$ increases to infinity. Figure 1 shows the graph of $\omega(x)$ for some values of $L/a$. The approximation of a constant correlation $\omega(x)$ thus seems to be acceptable when $L/a$ is not too large.
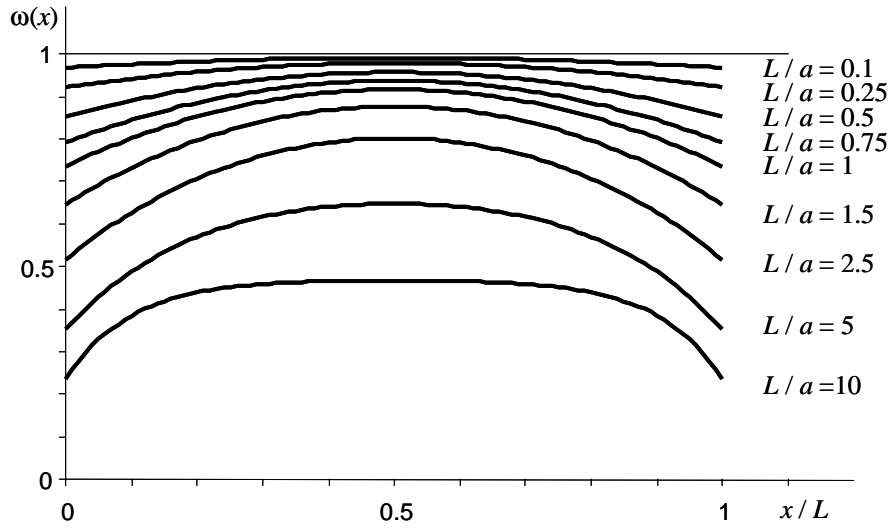


**Figure 1** Graph of the correlation $\omega(x)$ of $Y(x)$ and $Y_v = Y(v)/r$ for several values of $L/a$ ($Y$: Gaussian SRF; $v$: segment with length $L$; $a$: exponential-covariance scale parameter).

The situation is more critical for the bivariate distribution of $Y(\underline{x})$ and $Y(\underline{x}')$ because it is a mixture of standard bivariate Gaussian distributions with correlations $\rho(x' - x)$, thus ranging from 1 when $x = x'$ to $\rho(L)$ when $x$ and $x'$ are at opposite corners of the block $v$ and $L$ denotes their separation vector (we assume here isotropy). In the above one-dimensional case, $L$ is the length of the segment and this minimum correlation is equal to $\exp(-L/a)$: it can be very small if $L$ is large with respect to $a$.

When $Z$ is not Gaussian, the transformation function $\varphi$ is not linear. The variogram of $Z$ is therefore not proportional to the variogram of $Y$. Notice that the change-of-support coefficient of model DGM2 depends only on the variogram of $Y$ and does not depend on the transformation $\varphi$. This is not the case for model DGM1, because the covariance $C$ of $Z$ depends also on the transformation $\varphi$:

$$C(h) = \sum_{n=1}^{\infty} \varphi_n^2 \, \rho(h)^n$$

The correlograms $\rho(h)^n$ are less and less structured when $n$ increases. For example, if $\rho(h)$ is an exponential covariance with scale parameter $a$, $\rho(h)^n$ remains exponential, but with scale parameter $a / n$, and therefore tends to a pure nugget effect when $n$ tends to infinity. As a consequence, the change-of-support coefficient $r$ of model DGM1 is larger than that of model DGM2, with a difference that increases as the block $v$ is large and the terms $\varphi_n$ with large $n$ dominate in the development of the transformation function $\varphi$. The efficiency of model DGM2 will be similar to that of model DGM1 when the block size is small and the $\varphi_n$'s decrease rapidly with $n$.

Note that the change-of-support coefficient $r$ cannot simultaneously satisfy relations (5) and (6) (except if $Z$ is bivariate Gaussian). Unlike model DGM1, model DGM2 therefore does not preserve the variance of $Z(v)$ expressed in the right-hand side of relation (5), which is an important parameter of the block distribution. This may be critical if the ratio $r_1/r_2$ of the coefficients $r$ provided by models DGM1 and DGM2 is significantly larger than 1.

## Validation of the Discrete Gaussian Model

We will check the validity of the assumptions of the discrete Gaussian models for modeling the marginal block distribution. This corresponds to the so-called global change-of-support, by contrast with the local change-of-support which consists in predicting the distribution of block values conditional on data available in the block and its neighborhood.

The result depends on the spatial distribution of the random function $Z$ and we consider the ideal situation where this spatial distribution is known. In practice, we will consider the case where $Z$ is a lognormal SRF (that is, its logarithm is a Gaussian SRF). We can assume without lack of generality that it has a unit mean. It is then characterized by its logarithmic standard deviation $\sigma$ and by the correlogram $\rho(h)$ of its logarithm. The random function $Z(x)$ is thus of the form

$$Z(x) = \exp\left( \sigma Y(x) - \frac{\sigma^2}{2} \right)$$

It can also be expressed as

$$Z(x) = \sum_{n=0}^{\infty} (-1)^n \frac{\sigma^n}{\sqrt{n!}} \chi_n \left( Y(x) \right) ,$$

which corresponds to a development of the transformation function $\varphi$ with

$$\varphi_n = \frac{\sigma^n}{\sqrt{n!}}$$

The $\varphi_n$'s decrease rapidly when $n$ increases if $\sigma$ is small, but not so much if $\sigma > 1$. For $\sigma = 3$, for example, $\varphi_n$ increases up to $n = 8$ and then decreases slowly. The logarithmic assumption thus includes very contrasted behaviors. This can also be seen on the variance of $Z$: it is given by

$$\Sigma^2 = e^{\sigma^2} - 1$$

Since $Z(x)$ has unit mean, its coefficient of variation is close to $\sigma$ when $\sigma$ is small (0.53 for $\sigma = 0.5$) but takes large values when $\sigma$ is large: 1.31 for $\sigma = 1$, 7.32 for $\sigma = 2$, 90.01 for $\sigma = 3$. Such large a logarithmic standard deviation as 3 has been sometimes reported for permeability in hydrogeology [7] but the main variable in that case is log-permeability rather than permeability. When dealing with ore grade or pollution concentration, we are of course interested in $Z(x)$ itself rather than its logarithm, and such a logarithmic standard deviation would be extreme. Would it correspond to an actual situation, we would be in a very uncomfortable situation because a simple parameter such as the arithmetic mean of $Z$ would require a large number of observations to be estimated safely. Of course it is possible to deduce it from the mean and the variance of log $Z$ but then we rely heavily on the assumption of a lognormal distribution, which cannot be taken for granted.

A specificity of lognormal SRFs is that DGM models lead to a lognormal distribution for $Z(v)$, with logarithmic variance $r^2 \sigma^2$: This is the well-known permanence of lognormality. It can be shown easily by application of (1) or (4). Another specificity is that the correspondence between the covariance $C(h)$ of $Z$ and the correlogram $\rho(h)$ of $Y$ takes the simple form

$$C(h) = \exp\left(\sigma^2 \rho(h)\right) - 1 ,$$

which facilitates the computation of the right-hand side of equation (5).

## *Validation Method*

Several authors already checked the validity of model DGM1 or DGM2 in the lognormal case [1, 2, 3, 6]. The principle is to compare the block transformation function given by the DGM model with the "exact" block transformation function obtained with a large number of simulations.

Matheron [6] conducted a first exercise in 1981. With the computation capabilities available at that time, this check was limited to the one-dimensional case with an exponential covariance (the random function is then a Markov random process), and used a limited number of simulations. We extend it to 2D and 3D, with a finer discretization of the block, and a much larger number of simulations, enabling relatively large values for $\sigma$ to be considered. The principle of the checking is as follows:

- Consider a block $v$ of the $d$-dimensional space, defined as the union of $M$ discrete points forming a regular grid.
- Build $N$ unconditional simulations $Y_k(x)$: $k = 1, \ldots, N$, of a point-support stationary Gaussian random function $Y(x)$ with zero mean, unit variance, and covariance $\rho(h)$.
- Build the corresponding simulations $Z_k(x) = \exp\left(\sigma Y_k(x) - \sigma^2 / 2\right)$.
- Calculate the simulated block values $Z_k(v)$, namely the average values of $Z_k(x)$ among all $x$ defining the block $v$.
- Sort the $Z_k(v)$ by increasing values; let $W_k(v)$ denote the sorted series.
- Use $(W_k(v) + W_{k+1}(v)) / 2$ as the $k / N$ th quantile of the distribution of $Z(v)$, that is as the value of $\varphi_v(y)$ for $y = G^{-1}(k / N)$.
- Compare to the values predicted by models DGM1 and DGM2, that is to $\exp(r \sigma y - r^2 \sigma^2 / 2)$ with the corresponding change-of-support coefficient $r$.

Note that we do not consider the average value at $N$ grid nodes as an approximation to the average in a block $v$ of $\mathbb{R}^d$: We substitute the problem in the discrete space to the problem in the continuous case, so that there is no approximation in the approach. The simulations are generated with the discrete spectral method, which produces perfectly Gaussian simulations (up to the quality of the pseudo-random number generator). Would it not be possible to exactly reproduce the desired covariance for $Y$, we would replace it by a covariance as close as possible to it and check this model. For consistency, the integrals in (5) and (6) are replaced by discrete sums. We use a large number of simulations (up to 100,000).

Note that this approach, fully similar to that of Matheron [6], is slightly different from that used by Emery [3] to check model DGM2. Indeed Emery simulated standard Gaussian values with pairwise correlations all equal to $r$, thus corresponding to the values of $Y$ at independent random points in $v$.

## *Results*

We focus on a spherical correlogram for $Y$, and on three contrasted block sizes $L$ with respect to the range $a$ of the correlogram: $L / a = 0.1$, 1, and 10.

When $L / a = 0.1$, the change-of-support coefficient $r_2$ of model DGM2 is close to 1 (0.980 in 1D, 0.951 in 2D, 0.933 in 3D) and the coefficient $r_1$ of model DGM1 remains very close to $r_2$ even for a large $\sigma$ value (for $\sigma = 3$ we obtain 0.981 in 1D, 0.955 in 2D, 0.938 in 3D). Both models lead to similar block distributions, very close to the true one (see [1] p. 454 for the 2D case with $\sigma = 1.5$).

When $L / a = 1$, the coefficient $r_2$ is equal to 0.79 in 1D, 0.59 in 2D, and only 0.46 in 3D. Moreover, $r_1$ increases significantly with $\sigma$ when $\sigma$ exceeds 1, as can be seen in Figure 2.
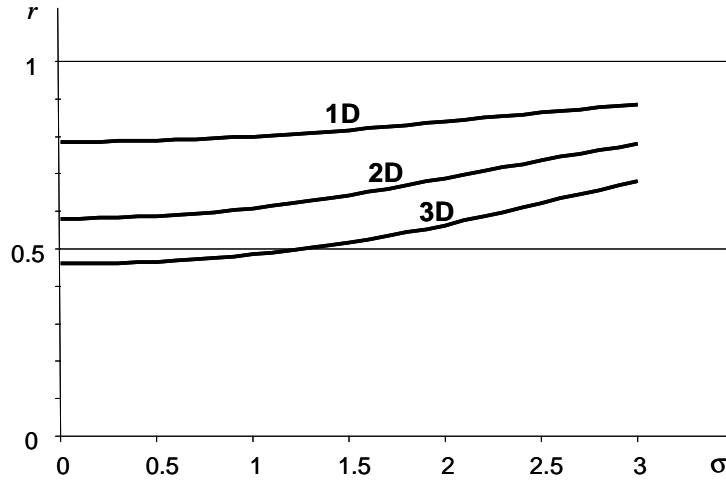


**Figure 2** Graphs of the change-of-support coefficient of model DGM1 for $L / a = 1$ as a function of the logarithmic standard deviation $\sigma$, in the 1D, 2D and 3D cases. The value at the origin coincides with the change-of-support coefficient of model DGM2, which does not vary with $\sigma$.

Figures 3 and 4 show the results obtained with 100,000 simulations for $\sigma = 1$ and 2 respectively. Model DGM1 quite perfectly reproduces the true transformation function—and thus the block distribution. Model DGM2 gives good results as far as large values are not considered but comes with a slight bias for $y > 2$ when $\sigma = 1$, and a significant bias when $\sigma = 2$.
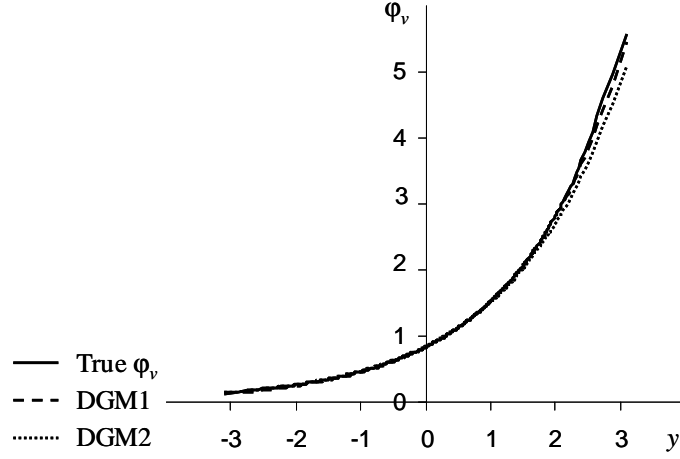
**Figure 3** Validity of the DGM approach for a 2D block with size *L* equal to range *a* and a logarithmic standard deviation $\sigma = 1$: "True" block transformation function $\varphi_v$ (determined from 100,000 simulations) and approximations provided by models DGM1 and DGM2.
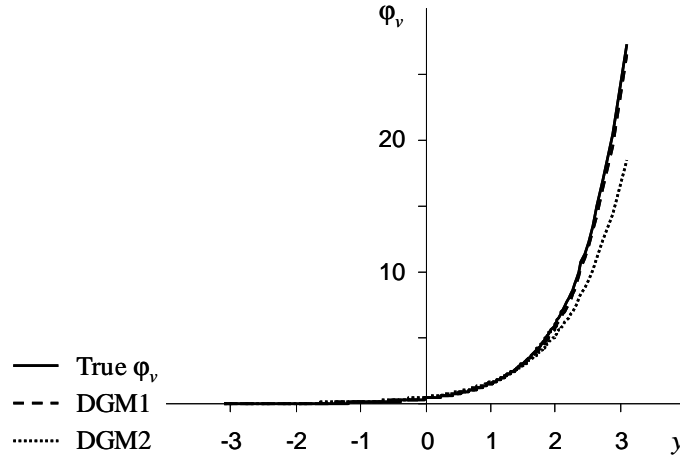
**Figure 4** Validity of the DGM approach for a 2D block with size *L* equal to range *a* and a logarithmic standard deviation $\sigma = 2$: "True" block transformation function $\varphi_v$ (determined from 100,000 simulations) and approximations provided by models DGM1 and DGM2.

When $L / a = 10$, the coefficient $r_2$ is equal to 0.271 in 1D, 0.077 in 2D, and only 0.022 in 3D, and like in the preceding case $r_1$ can have much larger values when $\sigma$ exceeds 1. Model DGM1 presents a slight bias at the extremes of the distribution, whereas DGM2 is biased everywhere but close to the median (see [1] p. 454 in the 2D case with $\sigma = 1.5$).

## Conclusion

This validation exercise shows that the original DGM1 model of Matheron [5] gives a very good approximation to the true block distribution, except for extreme values when the logarithmic standard deviation is very large. The variant DGM2 of Emery [3] is also a very good approximation to the true distribution provided that we are in one of the following situations: (i) the block size is small with respect to the range or (ii) the logarithmic standard deviation is not too large or (iii) we are not interested in the distribution of high grades. In such a case, this variant can be applied safely.

Further checks are in progress. The approximation of the DGM models has to be quantified also in 1D and in 3D (the approximation is all the less valid as the space dimension increases). It is also interesting to examine other covariance models for $\rho(h)$: When $\sigma$ is fixed, the solution provided by DGM models depends only on $r$ but two covariance models that would give the same $r$ value do not necessarily give the same true block distribution. Finally, the presence of a nugget effect extends the validity of DGM models but this has to be quantified.

We only addressed the global change of support problem. The local change of support (prediction of the block distribution conditional on neighboring data) has been examined in [2]: Cressie considers unbiased lognormal estimators that are exponentials of the simple or ordinary kriging estimators of $Y(v)$, which amounts to the assumptions of model DGM2. The experiment compares this ordinary lognormal kriging with the optimal solution provided by conditional expectation, obtained by a Monte Carlo method. The results indicate that ordinary lognormal kriging performs well in situations where the block size is small with respect to the range, the lognormal standard deviation is not too large, and the neighborhood is sparse. The first two conditions are required for the global model to be efficient. The third one expresses that conditional expectation makes better use of numerous data than an estimator whose form is limited to the exponential of a linear combination of the logarithms of the data.

## Bibliography

1   J.-P. Chilès and P. Delfiner: *Geostatistics: Modeling Spatial Uncertainty*, 2nd edition. John Wiley & Sons, New York, USA, 2012.

2   N. Cressie: Block kriging for lognormal spatial processes. *Mathematical Geology*, 38(4):413–443, 2006.

3   X. Emery: On some consistency conditions for geostatistical change-of-support models. *Mathematical Geology*, 39(2):205–223, 2007.

4    G. Matheron: A simple substitute for conditional expectation: The disjunctive kriging. In: *Advanced Geostatistics in the Mining Industry,* M. Guarascio, M. David, and C. Huijbregts, eds. Reidel, Dordrecht, Holland, 221–236, 1976.

5    G. Matheron: Forecasting block grade distributions: The transfer functions. In: *Advanced Geostatistics in the Mining Industry,* M. Guarascio, M. David, and C. Huijbregts, eds. Reidel, Dordrecht, Holland, 237–251, 1976.

6    G. Matheron: Remarques sur le changement de support. Technical Report N-690, Centre de Géostatistique, Fontainebleau, France, 1981.

7    D.A. Zimmerman, G. de Marsily, C.A. Gotway, M.G. Marietta, C.L. Axness, R. Beauheim, R. Bras, J. Carrera, G. Dagan, P.B. Davies, D.P. Gallegos, A. Galli, J. Gómez-Hernández, P. Grindrod, A.L. Gutjahr, P.K. Kitanidis, A.M. Lavenue, D. McLaughlin, S.P. Neuman, B.S. Ramarao, C. Ravenne, and Y. Rubin: A comparison of seven geostatistically-based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow. *Water Resources Research*, 34(6):1373–1413, 1998.