Histogram modelling and simulations in the case of skewed distributions with a 0-effect: issues and new developments

Jacques DERAISME Geovariances,
Jacques RIVOIRARD Mines Paris Tech Centre of Geosciences.

Geovariances
Where no one has gone before

MINES ParisTech

Outline

- Introduction
- Methodology
  - Histogram modelling with dispersion of data
  - Normal score transform of data with zero-effect.
- Case study
- Conclusions

Introduction

- Histogram modeling (equivalently c.d.f. F(z) modeling) is required in non linear geostatistical techniques.
- It is equivalent to modeling the (non decreasing) gaussian anamorphosis function $z = \Phi(y)$.
- The difficulty is that the data gives a partial description of the whole distribution. In particular the tail is not well represented while the impact on the recoverable resources estimates or confidence intervals is crucial.
- A second issue is related to normal score transform of data in case of zero-effect, which is necessary when using conditional simulations in the multi-gaussian case.

Histogram Modeling Methodology

- Gaussian anamorphosis can be developed into Hermite polynomials Hn. Historically this was required to perform bi-gaussian Disjunctive Kriging of recoverable resources at mining cutoffs.
- Because of truncation at a given order, the polynomial development is not everywhere a non-decreasing function, particularly at large values (oscillations).
- It seems appropriate to model the distribution first, then to develop it into Hermite polynomials only if required by the application.

Histogram Modeling Methodology

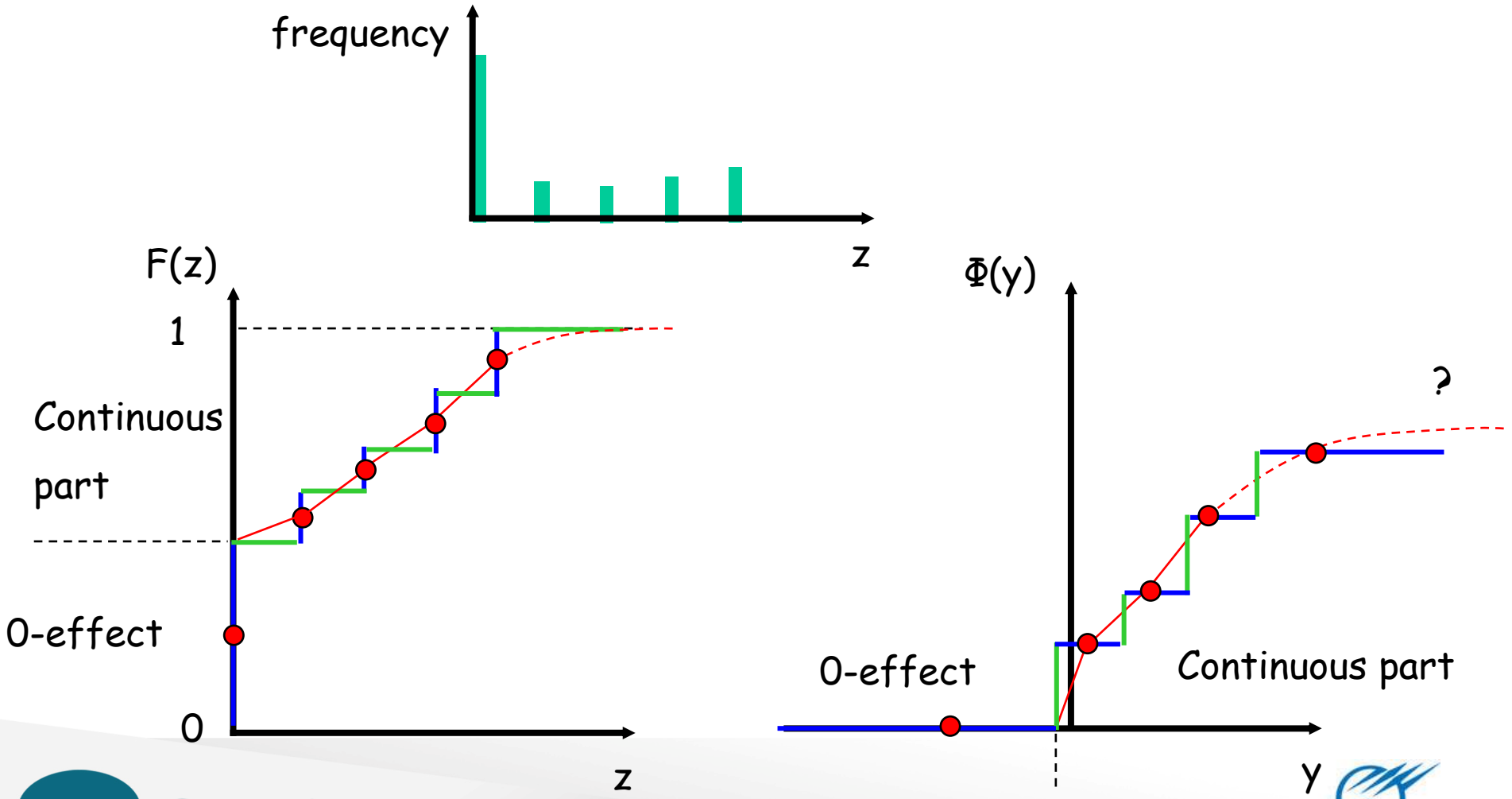
- The empirical distribution of data from a continuous variable is discrete and finite, even with a large number of data.
- We aim at getting a model which is a continuous distribution with possibly additional atoms (equal values for many data, e.g. 0-effect).

Histogram Modeling Methodology

- We propose to represent the histogram by intervals (y_i, y_{i+1}) or $(F(z_i), F(z_{i+1}))$.
- This includes
  - theoretical as well as empirical or discretized distributions
  - possibly the case of only 1 piece, eg lognormal.
- Then
  - for discrete distribution or atoms, $\Phi(y)$ is constant on piece(s)
  - else, behaviour of $\Phi(y)$ within each (y_i, y_{i+1}) piece, or of $F(z)$ within each (z_i, z_{i+1}) interval, to be given.

Histogram Modeling Methodology

From empirical to continuous distribution



Here, red points represent medians within intervals

Histogram Modeling Methodology

- Essential aim, here: to go from a staircase empirical distribution to a continuous distribution.
- The method is:
  - to represent each stair by one point (eg median or mean over each stair, related respectively to « frequency » or « empirical » inversion)
  - to **interpolate** between these points (linear, power…)
    - linear interpolation of T(z) is different from linear interpolation of Φ(y)
    - interpolation can possibly be used to rediscretize finely the distribution

Histogram Modeling Methodology

Extrapolation of large values:

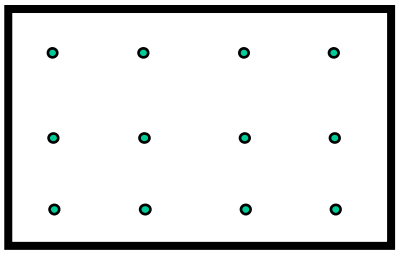
- We will focus on the possible tail of large values of a positively skewed distribution (the reverse to be obtained by symmetry)
- Empirical distribution gives often a poor description of the tail
  - « control points » can be added to refine the empirical distribution, without changing its staircase aspect
- Major issues:
  - Choosing a bound or not?
  - Behaviour towards this bound, or unbounded behaviour?

Histogram Modeling Methodology

- The previous interpolation and, above all, extrapolation, result in an increase of variance from the empirical variance of sample data
- In fact an increase of variance is natural, but could be better controlled

- Typical case:

  systematic grid with random origin



Additivity relationship where V* represents all samples within domain V:

$D^2(O|V) = D^2(O|V^*) + D^2(V^*|V)$

ie $\text{var}(Z(x)-Z(V)) - \text{var}(Z(x)-Z(V)^*) = \text{var}(Z(V)^*-Z(V))$

increase of variance = global estimation variance !!

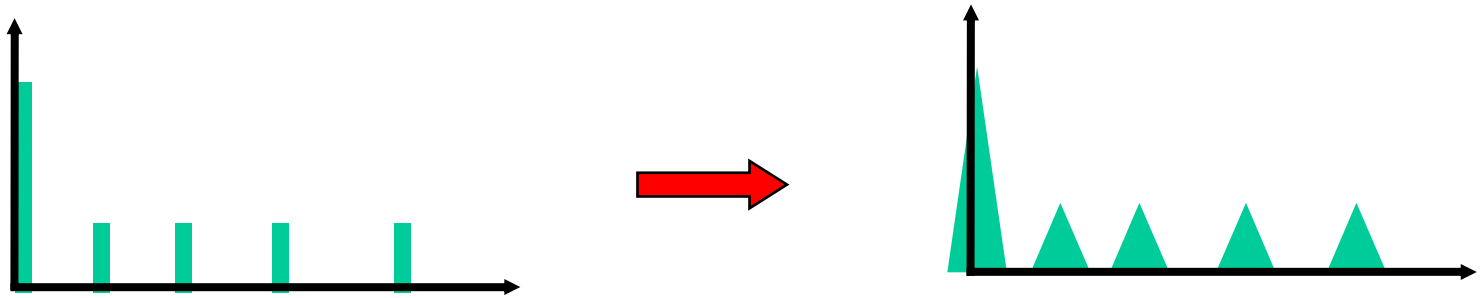
Histogram Modeling Methodology

- An idea is to consider the distribution model as a *dispersed* version of the empirical distribution, with a slightly higher variance equal to empirical sample variance + ~ global estimation variance
(this being expected to lie between nugget/n and sill/n for a variogram with sill and a regular sampling).

Histogram Modeling Methodology

- A way to obtain a distribution model which is a *dispersed* version of the empirical distribution consists in replacing each data value z_α by a distribution with mean z_α and appropriate variance

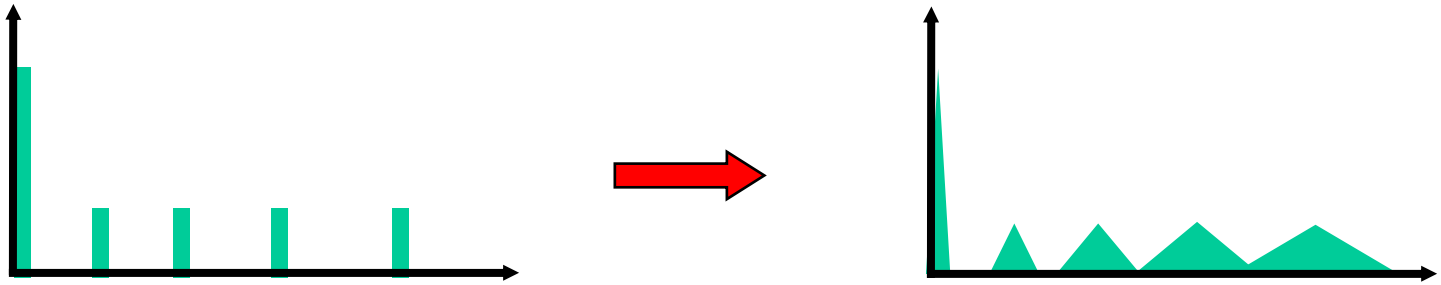
  - Replacing each data value z_α by a gaussian (or other) distribution with mean z_α and constant variance s^2 results in an increase in variance equal to s^2



  - This is appropriate if the distribution is unbounded: e.g. dispersing a zero-value would give negative values which may not be acceptable

Histogram Modeling Methodology

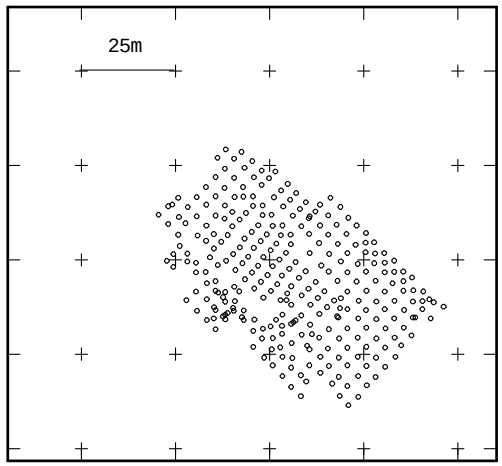
- Otherwise the variance can be modulated. For instance, in the case of a positive (non-negative) distribution with positive skewness, each data value z_α could be replaced by a gaussian (or lognormal, etc) distribution with mean z_α and std dev s z_α so that the increase in variance is s^2 sum z_α^2 /n (for weights 1/n)

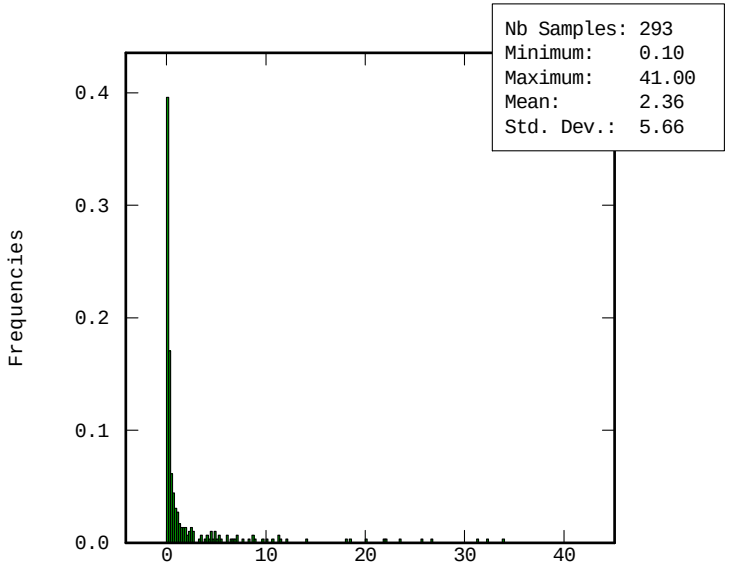


- Thus, in this method, the variance can be directly used as an input control parameter
- A further fine discretization appears as a convenient way to store the modeled distribution

Case study

Example of data with a skewed distribution

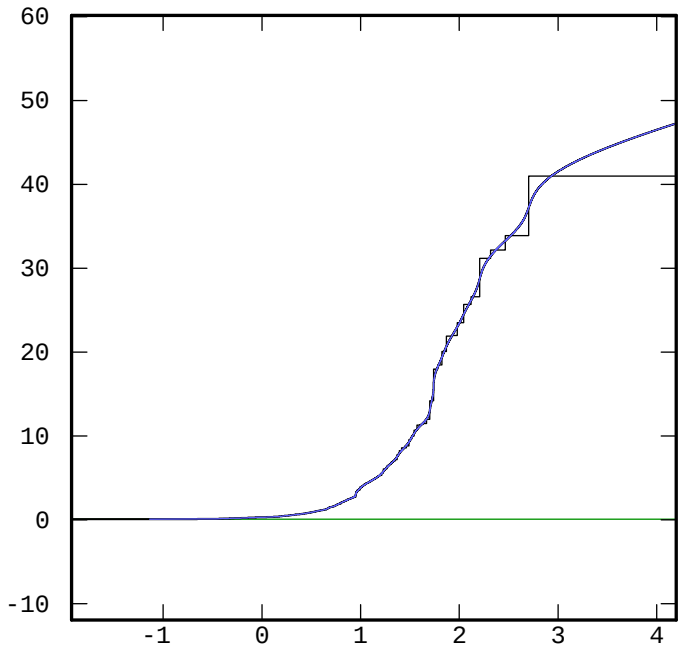




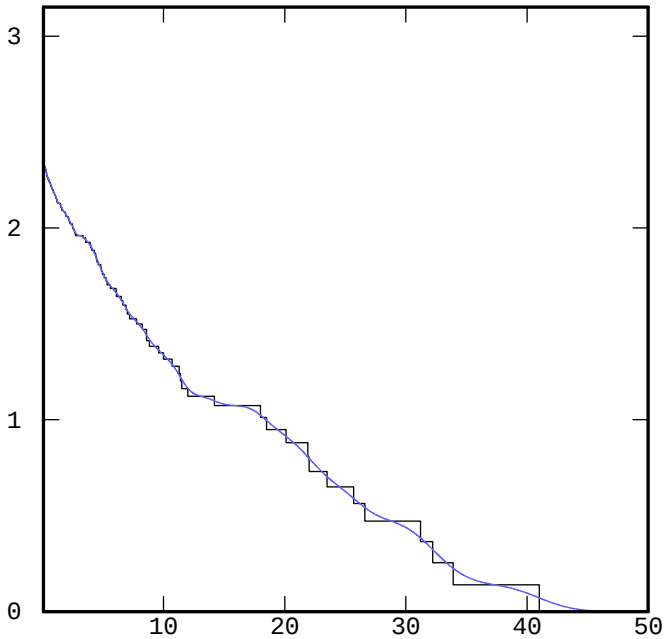
Case study

Gaussian anamorphosis with dispersion of the data allows extending beyond the maximum of the data.

The metal quantity versus cutoff is changed accordingly.



distribution



Gaussian anamorphosis with lognormal dispersion of the data.

Normal score transform of data with 0-effect

- 0-effect means that a significant proportion of values are equal, say $Z\alpha=0$.
- P(Z=0) = P(Y<yc), hence a 0-effect corresponds to a threshold on the gaussian variable.
- But the normal score transforms for the 0-effect cannot be obtained easily, as the normal score transform for all data should:
  - be normally distributed
  - share a consistent structure.

Normal score transform of data with 0-effect

- The method consists in generating Gaussian values for 0-effect using a Gibbs sampler:
  - start from gaussian random values
    equal to the normal score transform of Z>0 data,
    and to gaussian values <yc for the data with Z=0.
  - then generate a gaussian value for each data Z=0, conditionally on values at all other data points.
  - repeat the procedure iteratively in order to get the desired gaussian distribution and variogram.

Normal score transform of data with 0-effect

- For applying that method we need the variogram model of the gaussian variable.
- This variogram cannot be inferred directly since the gaussian values for the zero-effect are ignored.
- But in a gaussian model the covariance of a transform of the gaussian variable Y can be expressed from the covariance of the gaussian variable Y.

$$Cov\{f[Y(x+h)]\, f[Y(x)]\} = \int f(t)\, f(u)\, g_{\rho(h)}(t,u)\, dt\, du - [E\{f[Y]\}]^2$$

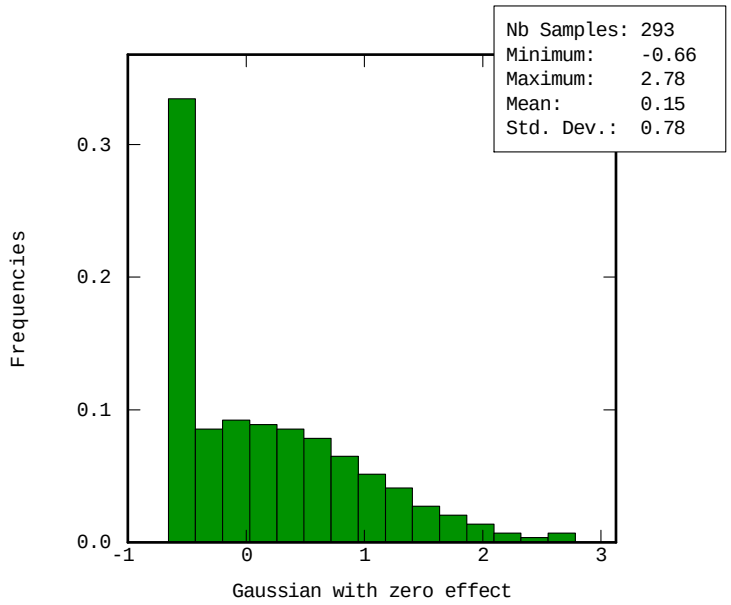
with the bi-gaussian standard p.d.f. $$g_\rho(t,u) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{t^2 - 2\rho tu + u^2}{2(1-\rho^2)}\right)$$

Therefore we can determine indirectly the covariance of Y from this of f(Y), taken here as Y truncated at the threshold yc.

Case study

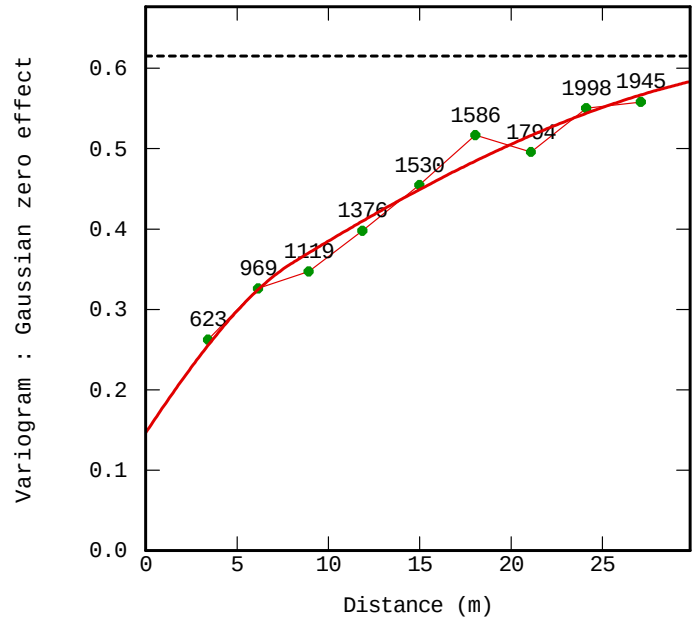
In then previous example about 40% of the data were equal to the detection limit.

After pseudo-normal score transform we get a truncated gaussian distribution.



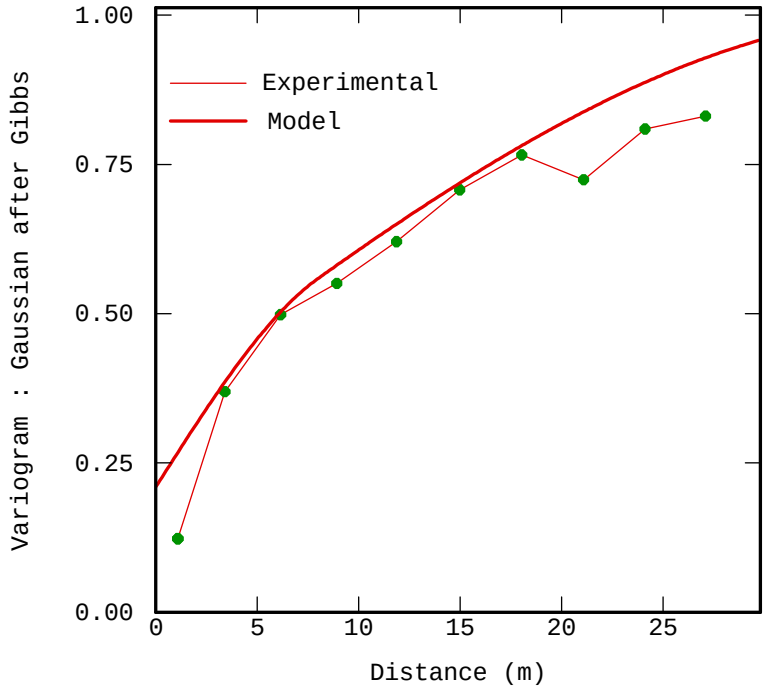
Case study

The variogram model of the underlying gaussian variable can be chosen so that the variogram of the truncated gaussian variable is fitted.



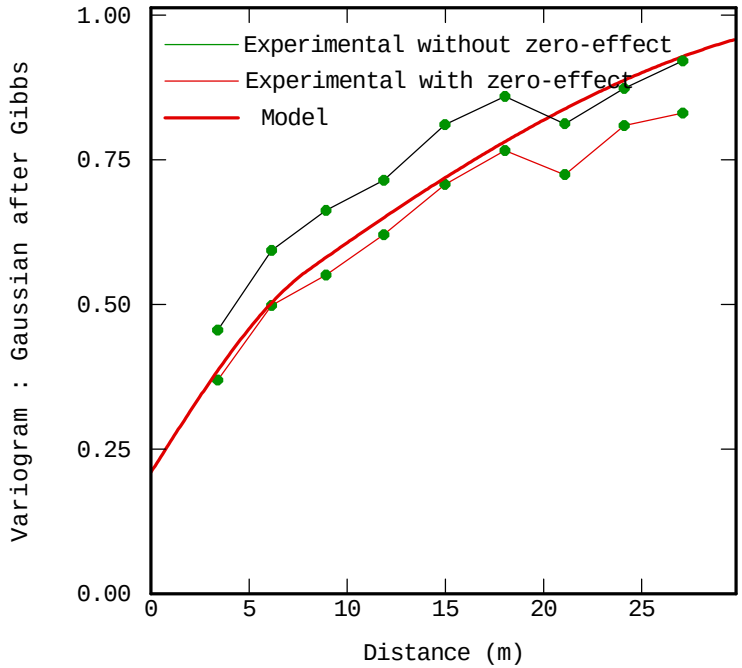
Case study

After the Gibbs sampler we can compare the experimental variogram calculated on the gaussian variable, with the input variogram model.



Case study

With a « classical » approach (normal score transform by frequency inversion) we would get a higher variogram because of the arbitrariness of the gaussian values assignment.



Conclusions

- Histogram modeling by dispersion: a flexible way to model the empirical histogram of data while controlling variance and tail
- Normal score transform of data with zero-effect: allows generating at datapoints values that are normally distributed and have a consistent structure.
  It is a preliminary step before gaussian simulation, either monovariate or in relation with other transformed gaussian fields e.g. representing continuous variables or indicators.