A Short Note on how Response Uncertainty is Affected by Local Data and Parameter Uncertainty

Olena Babak and Clayton V. Deutsch

Centre of Computational Geostatistics
Department of Civil & Environmental Engineering
University of Alberta

An important goal of geostatistical modeling is to assess output uncertainty after processing realizations through a transfer function, for example, the assessment of recoverable reserves in a petroleum or mining context. The decisions of stationarity and a modeling method are critical for reasonable results. Then, response uncertainty is affected by the amount of local data and parameter uncertainty, which may be ignored. Oftentimes, ergodic or statistical fluctuations given a set of parameters and local data are used to measure uncertainty. This short note documents the importance and interactions between local data, parameter uncertainty and ergodic fluctuations.

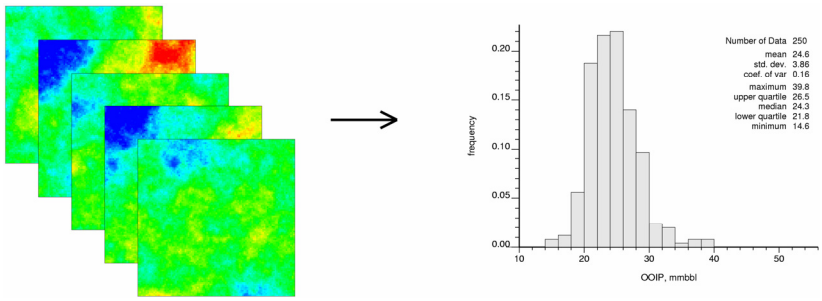
Introduction

Uncertainty assessment is an important goal of modern geostatistics. We are interested in local and global uncertainty. Local uncertainty relates rock properties at specific locations that we could potentially sample in the future. Global uncertainty relates to some calculated statistic that involves many locations simultaneously. We often check local uncertainty by cross validation or new drilling: the proportions of true values falling within specified probability intervals are checked against the width of the intervals. Most often we can tune the geostatistical parameters to achieve fair local uncertainty predictions. Global uncertainty is another matter.

Checking global uncertainty is very difficult. We would need multiple deposits or reservoirs. A simulation exercise could be constructed to generate multiple true distributions; however, the goodness of checking would relate simply to the closeness of the geostatistical algorithm to the simulation algorithm that created the reference truths. Techniques for assessing global uncertainty are not fully developed in geostatistics. A first step in the direction of reliable global uncertainty predictions is the integration of all reasonable data and parameter uncertainty.

Uncertainty is model dependent. An implicit multivariate distribution underlies all of approaches to uncertainty assessment. The uncertainty in simple global statistics may be assessed by analytical calculations, for example, the variance of the average of n-independent data is 1/n that of the data. Earth sciences data are rarely independent, so we either (1) aggregate the scale up so that independence can be reasonably assumed, or (2) consider the spatial correlation between the data. Simple analytical models rarely work because of spatial correlation and the application of a non-linear selection function. We cannot avoid the requirement for a model.

A common approach to handle spatial correlation and a non-linear transfer function is to construct alternative realizations of the spatially distributed variables. These realizations are then passed through the transfer function (to calculate resources or reserves) and uncertainty in the global response is assembled as a histogram of the responses. A schematic is shown below.



The chosen multivariate probability (random function – RF) model is critically important. Often, it is a categorical variable modeling method for rock types followed by a Gaussian-based approach for continuous variables within specific rock types. The key idea of geostatistical simulation is to construct multiple realizations of the rock types and grades and then process them one-at-a-time through a transfer function to calculate response variables such as resource or reserve numbers.

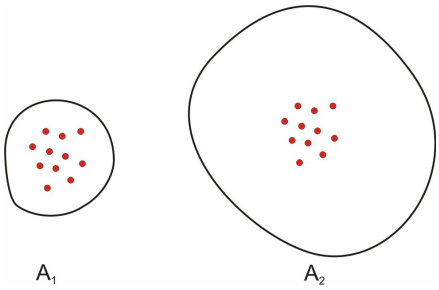
The parameters of the RF model are important. The parameters are the input histogram and the variograms or the training images that contain the spatial features believed to apply to the deposit under consideration. Common practice consists of determining the most representative parameters and constructing each realization with the same input parameters, that is, every realization has the same target rock type proportions (histogram of continuous variables) and target spatial features. The experimental statistics calculated from the realizations will not be the same as the input parameters because of statistical fluctuations. These statistical fluctuations are often referred to as *non-ergodic fluctuations* because they arise due to the finite non-ergodic size of the domain under consideration. There would be no global uncertainty if we were considering a very large (ergodic) domain since high and low areas would average out. The realizations would not be the same – there would, of course, be local uncertainty, that is, differences between the realizations at each location.

> *We conjecture that global uncertainty calculated with fixed input parameters seriously underestimates the true uncertainty.*

The goal of this short note is to document the importance of parameter uncertainty in the assessment of global uncertainty. This is a difficult problem and we make many assumptions including a fixed random function model. Let's see a numerical example.

Importance of Parameter Uncertainty

Global uncertainty will be affected by the amount of local data, the variogram structure – range of correlation, and the size of the domain. The uncertainty in input parameters will be affected mostly by the local data and their spatial correlation. The parameter uncertainty does not depend directly on the size of the area being modeled. Two areas are shown to the right with the exact same configuration of 11 data. The uncertainty in the input parameters would be the same, but the global uncertainty over area A_1 would be different than that over area A_2. The precise difference is not evident. There are two competing considerations. First, the



smaller area of A_1 may have less uncertainty because more of the area is within the range of correlation of the available 11 data. Second, the larger area A_2 may have less uncertainty because there could be areas of higher and lower values away from the data control that *cancel* each other out. A simulation study is one approach to understand which area has the most uncertainty.

There are three factors that must be considered: (1) the number of data, (2) the variogram range of correlation, and (3) the size of the domain. We will consider a fine level of discretization of any domain; therefore, the latter two considerations could be parameterized by the variogram range for a fixed domain size. Increasing the variogram range and decreasing the domain size would have the same effect.

Problem Setup

A spatial context is required. We start with the well known GSLIB data. That data consists of about 100 data that are sampled on a random stratified grid and 40 data that are clustered in high valued areas. We discard the clustered data. This exercise is aimed at global uncertainty and not the challenge of declustering and debiasing. The 2-D area of interest is 50 by 50 distance units. The distribution of data is approximately lognormal with a mean of 2.5 and a standard deviation of 5.0. A transfer function and response variable is required. We chose to block average the results to a scale of 1% the area in each coordinate direction, apply a cutoff of 0.7 and then calculate the total quantity of metal above that cutoff.

The quantity of metal is the global response variable under consideration. Uncertainty in the response variable will be measured by the standard deviation in the quantity of metal calculated from multiple realizations. The standard deviation is used because it is in the units of the data and linearly related to the width of a probability interval (approximately).

We can quantify global uncertainty by generating multiple realizations by sequential Gaussian simulation (SGS), calculating the response variable on each, and then calculating the standard deviation of the outcomes. The conventional approach, without considering parameter uncertainty, would amount to freeze the input histogram and variogram for each realization as that of the input data. The output uncertainty is due entirely to non-ergodic fluctuations; there would be no uncertainty if we had an infinite sized domain. We consider the conventional approach and the approach where uncertainty in the histogram is accounted for.

Uncertainty in the histogram is accounted for by the spatial bootstrap (see last year's CCG report). Given a set of data and the variogram model, we can easily construct multiple realizations of the histogram. We refer to these alternative realizations as parameter uncertainty. SGS could be executed such that a different input histogram is considered for each realization. This is not the conventional approach, but one we would recommend. The results include parameter uncertainty and non-ergodic fluctuations.

We vary the number of data and the variogram range. Five cases for the number of data are considered – 5, 10, 20, 40, and 100. The uncertainty should be the largest for the fewest data and uncertainty should decrease as the number of data increases. Five cases for the range of correlation are considered – 1, 5, 10, 20, and 50. The dimensionless ranges would be 0.02, 0.1, 0.2, 0.4, and 1.0. The uncertainty should be the least with a small range of correlation because of the fixed domain size. The change of uncertainty as the range of correlation increases is not so predictable. On one hand we might think that the uncertainty should increase with an increase in range because the non-ergodic fluctuations will not cancel each other out. On the other hand, we think the uncertainty should decrease because the data are more correlated to all locations in the

domain of interest. The usefulness of stochastic simulation is apparent – we will directly see the affect of these two variables.

The dataset of 100 is randomly decimated to construct the data sets of 5, 10, 20 and 40 data. The results would be sensitive to the specific set of data chosen; therefore, we randomize the chosen data and consider different datasets. We did not fully automate all of the simulation, so we consider 5 different sets of data for each n. The results will be affected by the multivariate features of the GSLIB data and the SGS simulation algorithm.

In summary, the impact of parameter uncertainty on the global uncertainty (measured using ergodic fluctuations and local data) after a non-linear transfer function can be assessed using the following simple procedure:

1. Choose s random subsets of size $n_1, n_2, \ldots, n_s$ from original data set of size n, where $n \geq n_i, \quad i = 1, \ldots, s.$
2. For each of the s chosen subsets
    a) Simulate
        i) L (100) geostatistical realizations with fixed input parameters (parameter uncertainty is not accounted for);
        ii) L (100) geostatistical realizations with different input parameters that are sampled from their distribution of uncertainty (parameter uncertainty is accounted for);
    b) Apply the non-linear transfer function to each realization of the two scenarios in 2.a);
    c) Calculate uncertainty for the transformed set of L realizations for each scenario in 2.a).
3. Repeat steps 1-2 several times, say, $k = 5$.
4. Average results for the global uncertainty over k for each of the s subsets and for each of the two scenarios in 2.a).

The normal scores variogram used in the simulation exercise was constrained to have a nugget effect of 30% and a variable range. The real data have a range of about 10. There is a slight inconsistency in the results when the range is set larger than 10 – the data are not that correlated. This may lead to an artifact in the results. A fully simulation based exercise may be setup for future analysis.

The realizations were constructed at a resolution of 500 by 500 grid nodes. The block size for the quantity of metal calculation was held constant relative to the variogram range so that the range parameter is a domain size parameter.

Some Results

Figure 1 shows some results for the very first case – 5 data and a variogram range of 1. The spatial bootstrap and the SGS programs are standard programs no particular problems were encountered. 5 cases times 5 different n cases times 5 different a cases times 100 realizations (12500 total realizations) were generated and processed.

Figures 2, 3, and 4 show the uncertainty results in a series of plots. Note that the scales on these plots are quite different. In general the results are what we expect. The results on Figure 3 are interesting because there are many different effects confounding the results including the domain size, variogram range, and different data sets.

Conclusions and Future Work

Decisions of stationarity and a modeling methodology are arguably the most important factors in determining output uncertainty in any practical modeling study. The effects of local data, ergodic fluctuations and parameter uncertainty were investigated with a small example. Some conclusions:

- The uncertainty drops quickly as the number of data increases. In general, the decrease is proportional to 1/n as we would expect with independent data. The spatial correlation between the data was accounted for with the spatial bootstrap.
- The uncertainty due to the size of the domain / range of correlation is not as predictable. In general, as the range of correlation increases, the data are more correlated with each other (the uncertainty due to parameter uncertainty goes up) and the data are more correlated to the locations being predicted (the uncertainty goes down).
- Global uncertainty when parameter uncertainty is accounted for is, on average, 4 to 10 times higher than when we rely simply on nonergodic fluctuations. This confirms our initial conjecture that parameter uncertainty is a very important aspect of global uncertainty characterization.

This is a preliminary study. A more complete setup will have to be considered with dimensionless groups and nonlinear transfer functions that are carefully thought out in the first place. There is no question that uncertainty in the input histogram must be considered for realistic global uncertainty characterization. The spatial bootstrap should be used extensively. The affect of parameter uncertainty on local uncertainty should be assessed.

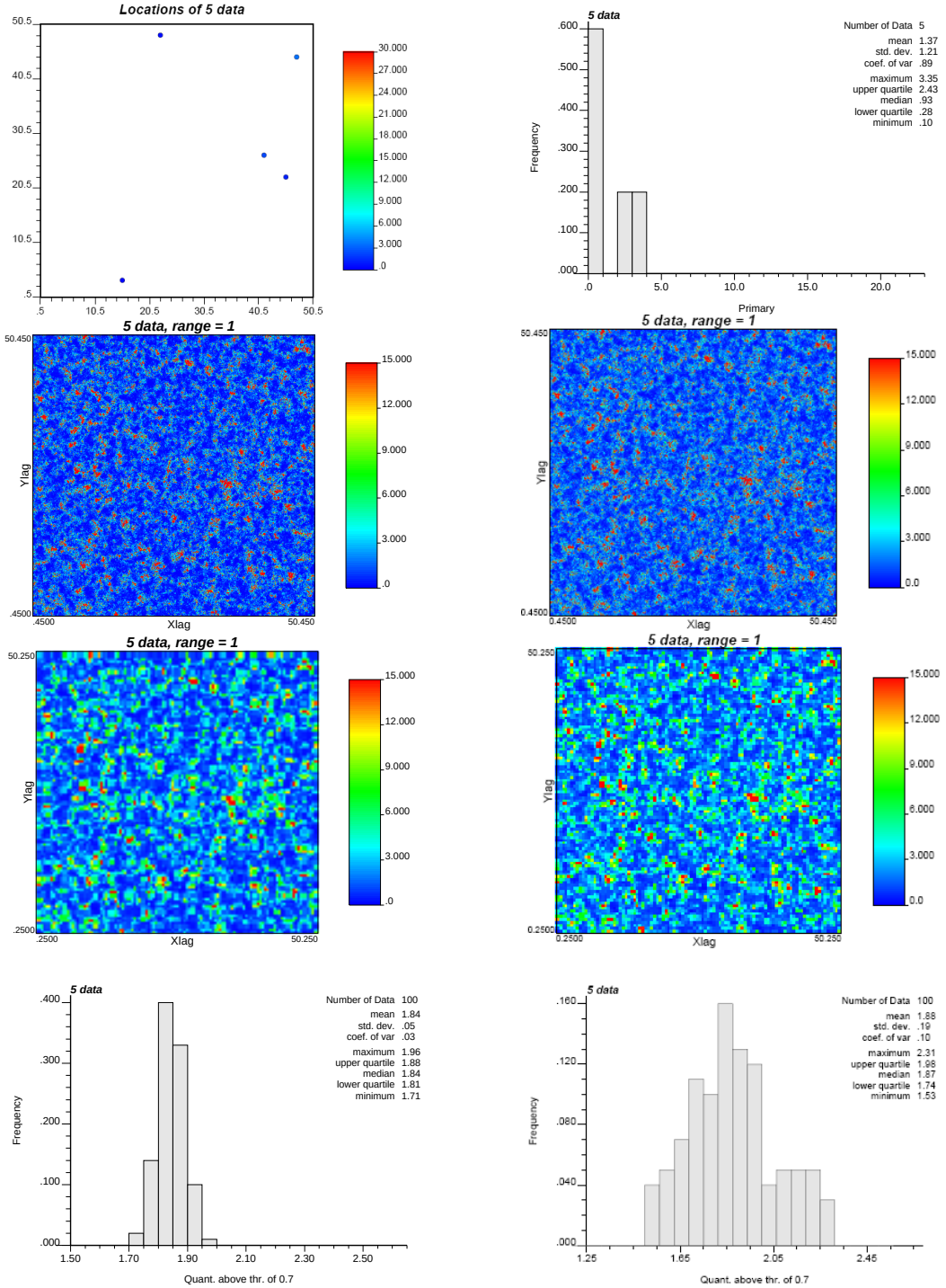


Figure 1: Setting of the problem for the first (out of 5) randomly chosen data set of size 5. The location map and histogram of the data are shown on the top. Example realizations of SGS on the grid of 500 by 500 blocks and their respective scaled to 0.5 by 0.5 block results with and without parameter uncertainty taken into account are shown in the middle right and left, respectively. Resulting distributions of block 'quantity' above threshold of 0.7 are shown on the bottom.

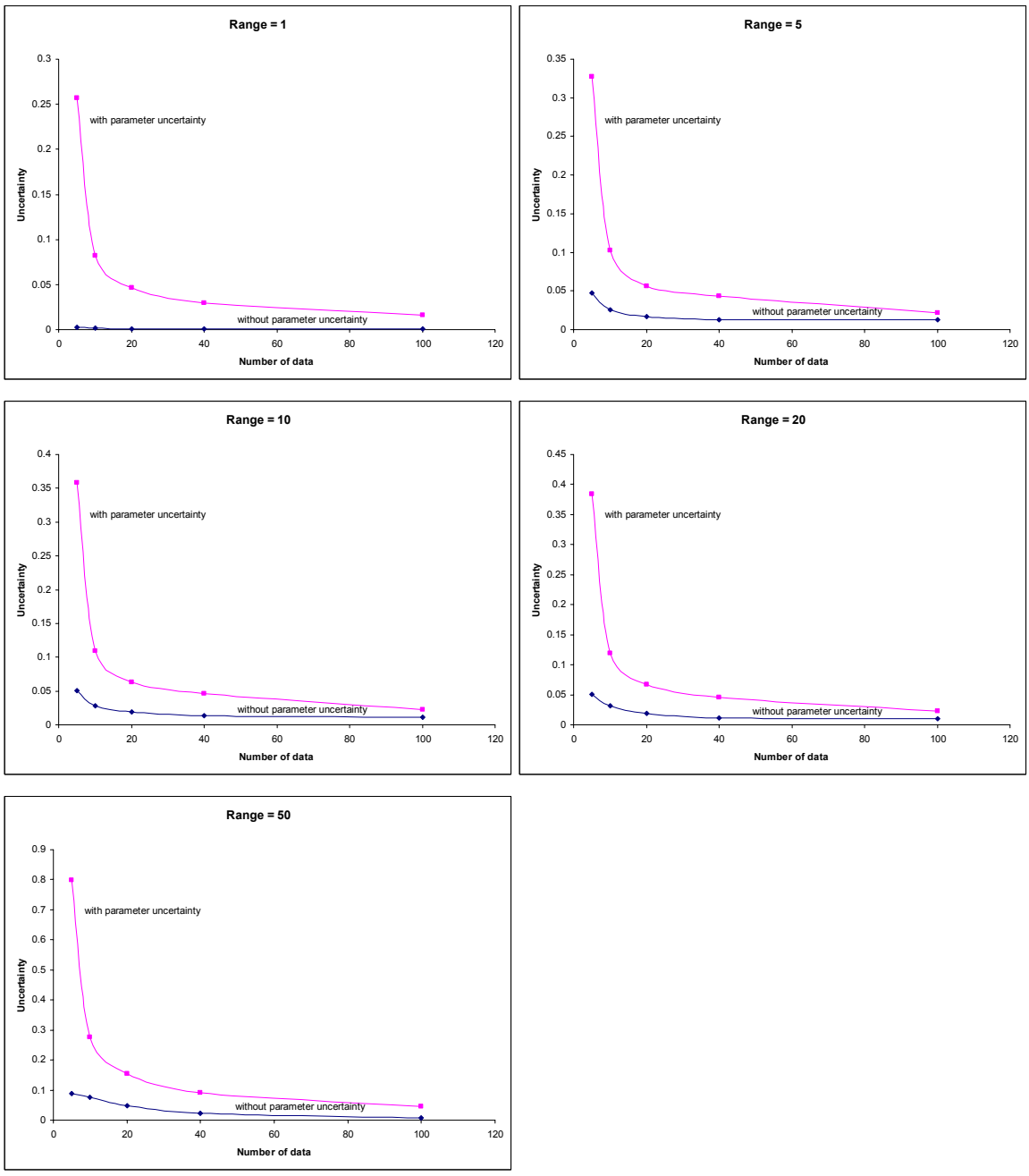


Figure 2: The uncertainty versus the number of data for different ranges. One curve shows the uncertainty with fixed parameters, the other with the histogram considered as uncertain.

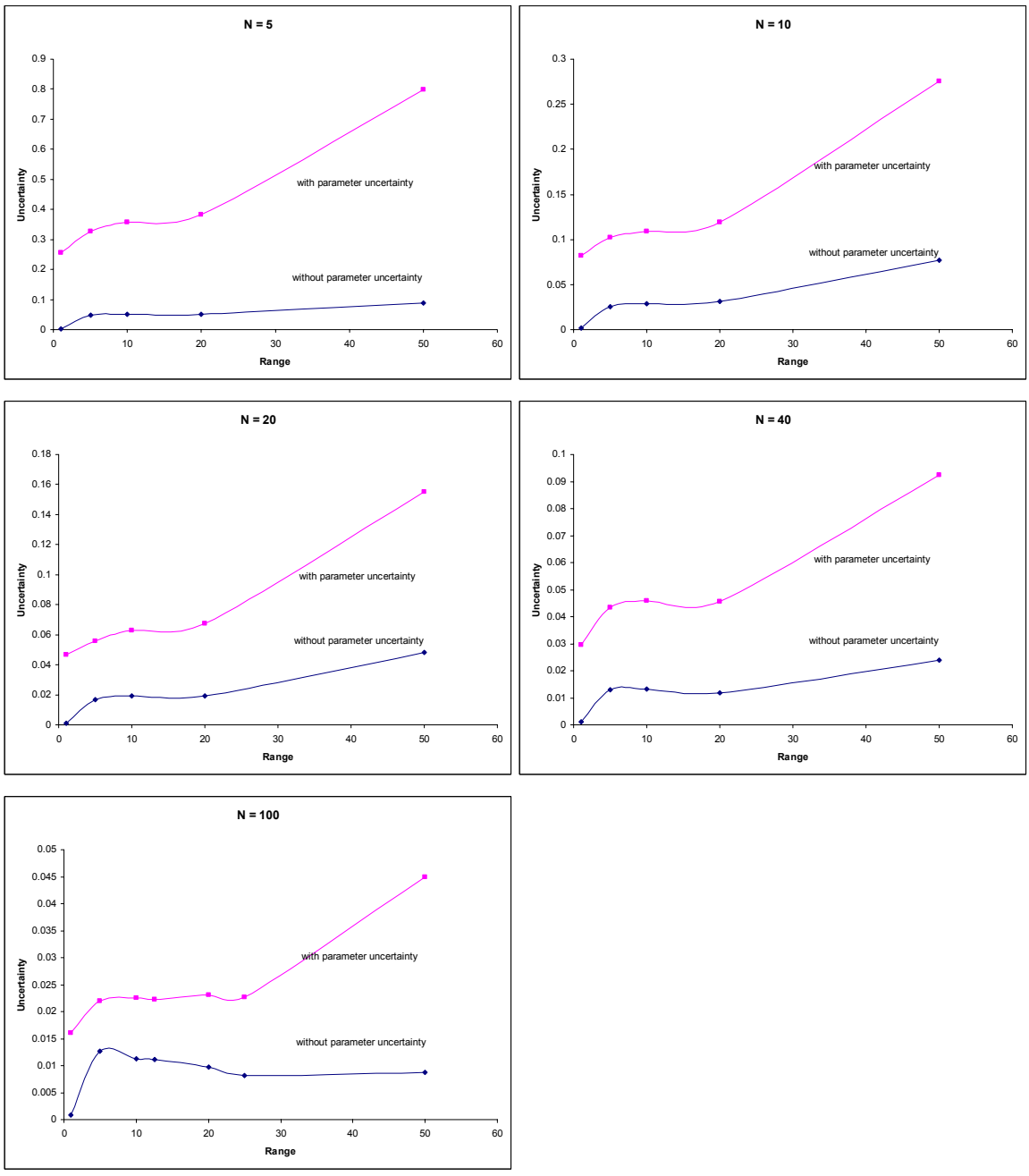


Figure 3: The uncertainty versus the range for different numbers of data. One curve shows the uncertainty with fixed parameters, the other with the histogram considered as uncertain.

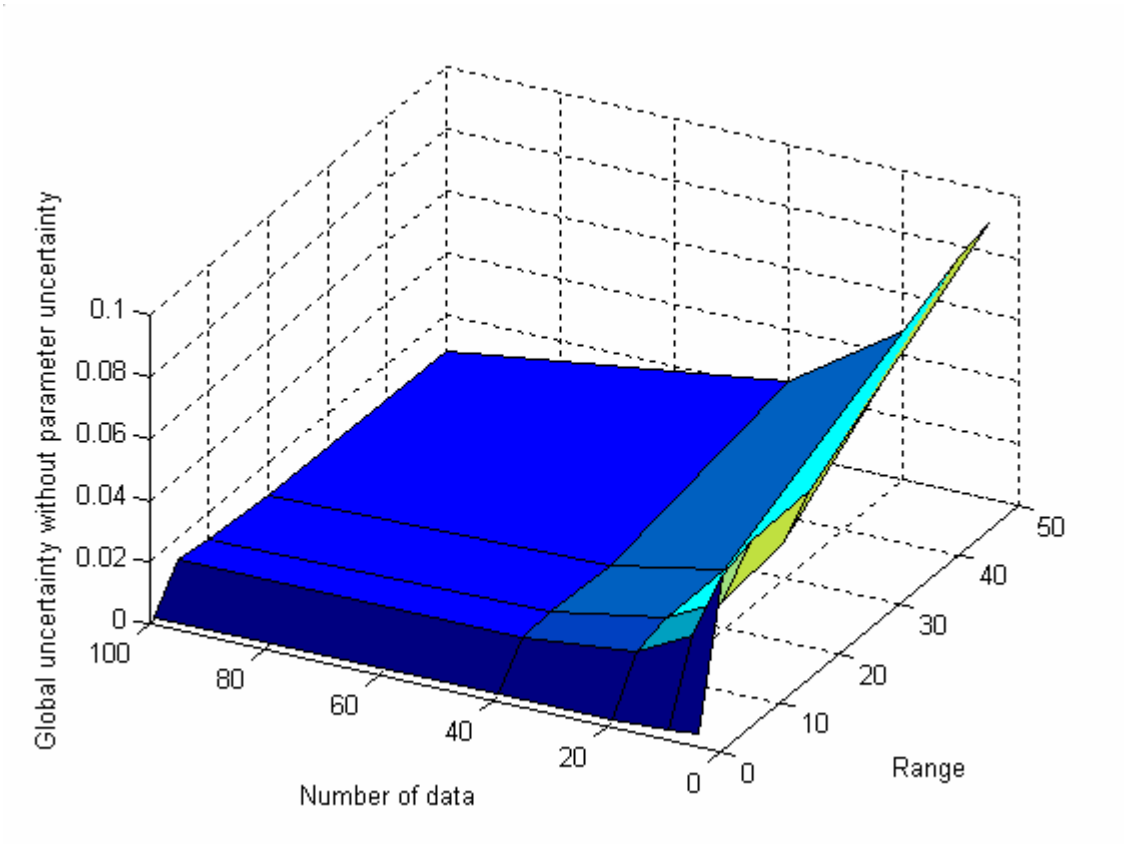


Figure 4: A plot of the global uncertainty (without parameter uncertainty) versus the number of data and the range..