Choosing an Adequate Number of Conditioning Data for Kriging

Hong Guo and Clayton V. Deutsch

The number of data to use in kriging should be chosen correctly to ensure there are no artifacts and to maintain reasonable computational cost. A number of numerical studies are presented to support a recommendation for how many data to use. Two criteria (the average kriging weight and standard deviation of the weights) are adopted to determine the selection of conditioning data in different situations. The appropriate number of data to use in kriging depends on the dimensionality of the domain, the variogram. The dimensionality is the most important and an adequate number of conditioning data is finally defined as 30 in 1D, 40 in 2D, and 60 in 3D.

Introduction

Kriging calculates a weighted average of *n* available sample values. Kriging weights depend on the number of conditioning data, how far the estimated location is from sampled location and the configuration of the data. So, a choice of *n* is very important. A limited search can be used to depend less on global stationarity and to improve calculation speed; however, not enough data would lead to conditional bias. Moreover, too few data introduces unwarranted noise in simulated realizations and statistical parameters will not be reproduced. The number of data could be set arbitrarily large to guarantee the precision of results and avoid conditional bias; however, the computational cost can be high since the time to solve the kriging equations increases as the number of data cubed; doubling the number of data leads to an eight fold increase in computer time. The number of samples is usually decided by experience and computational considerations. This paper aims to provide a better understanding of how many data should be used in different situations and choose a representative number as the default value. Four factors are considered: different constraints on the kriging weights, the dimensionality of the domain, the variogram model and the grid spacing.

The data are sorted by variogram distance from the location being estimated. Then, the kriging weights are calculated for many different data configurations. The data in GSLIB have been used for much of the following. Simple and Ordinary kriging were considered in many numerical experiments; however, they behave in a very similar fashion. Following are typical results for 2D data (see Figure 1). The abscissa axis is the number of data used in kriging (n) and the ordinate axis is in the units of the kriging weight. The grid selected here is 1m*1m grid cell in 50m*50m area. Then, for each x-coordinate value (100 in all), there are 2,500 different weights summarized on the plot. The range of kriging weights, including 0.99 quantile of weights, 0.75 quantile of weights, 0.25 quantile of weights, 0.01 quantile of weights, and the average kriging weight (the red line) are shown for the first, second, and so forth.

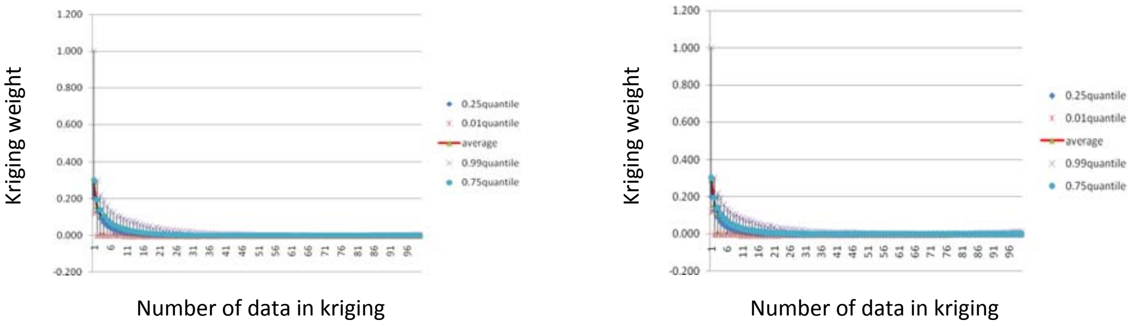


Figure 1 Typical results for 2-D using simple kriging and ordinary kriging (SK on the left and OK on the right). They behave in a very similar fashion.

The goal of this paper is to choose an adequate number of conditioning data for kriging. Then, some criteria are required to define this adequate number. We found that two parameters both decrease when data index increases and show the form of asymptote to 0. One is the average kriging weight. The other is the standard deviation of weights (σ). These two parameters are adopted as criteria to choose an adequate number for kriging. Simple kriging and ordinary kriging behave in a very similar fashion; therefore, the results for simple kriging will be shown in the following. The dimension of the domain is the most important, so the results for 1D, 2D and 3D are successively explored and summarized respectively, and in each case, factors such as variogram model and grid spacing are considered.

Four kinds of grid spacing are selected for simple kriging calculation in 1D, 2D and 3D respectively, see Table1.1. Many numerical experiments show that results derived from different kinds of grid spacing are similar, which indicates that the choice for number of conditioning data have little relation to the grid spacing. Typical results for one kind of grid spacing are shown. The number of data where λ=0 or σ change very little can be considered as the adequate number for kriging. The typical results for 1D can be seen at Table 2 and Figure 2. The adequate number of data to use in 1-D kriging to some degree depends on variogram range. The largest number for the two criteria is 29; therefore, 30 is a safe default for 1D.

Table 1 Grid spacing selected for 1D, 2D and 3D

dimensionality of the domain	domain size	grid number	grid cell size
1D	50m*1m	25*1	2
		50*1	1
		100*1	0.5
		125*1	0.4
2D	50m*50m	25*25	2*2
		50*50	1*1
		100*100	0.5*0.5
		125*125	0.4*0.4
3D	50m*50m*50m	25*25*25	2*2*2
		50*50*50	1*1*1
		100*100*100	0.5*0.5*0.5
		125*125*125	0.4*0.4*0.4

Table 2 The typical results for 1D simple kriging.

grid	grid cell size	number of locations	range	data index where λ=0	data index where σ rarely change	The largest number of two criteria
100*1	0.5	100	5	3	4	4
			20	10	26	26
			35	21	29	29
			50	20	26	26
			65	21	26	26

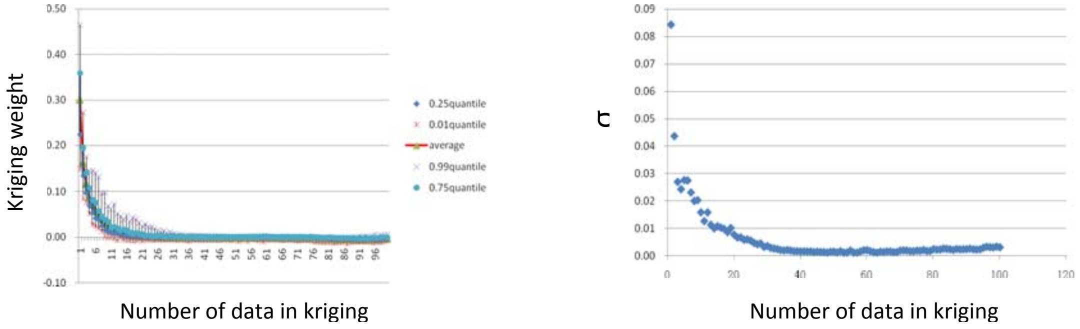


Figure 2 Change of kriging weights and standard deviation of the weights in 1-D. The grid is 100 with 50 variogram range. λ = 0 when number of data is 20. σ rarely changes when thenumber of data is 26.

The typical results for 2D can be seen at Table 1.3 and Fig 1.3. The adequate number of data to use in 2D kriging also depends on variogram range. The largest number of two criteria for different variogram range is 39, so 40 can be defined as the default data for 2D.

Table 3 The typical results for 2D simple kriging.

grid	grid cell size	number of locations	range	data index where λ =0	data index where σ rarely change	The largest number of two criteria
50*50	1*1	2500	5	14	20	20
			20	19	31	31
			35	24	35	35
			50	29	37	37
			65	32	39	39

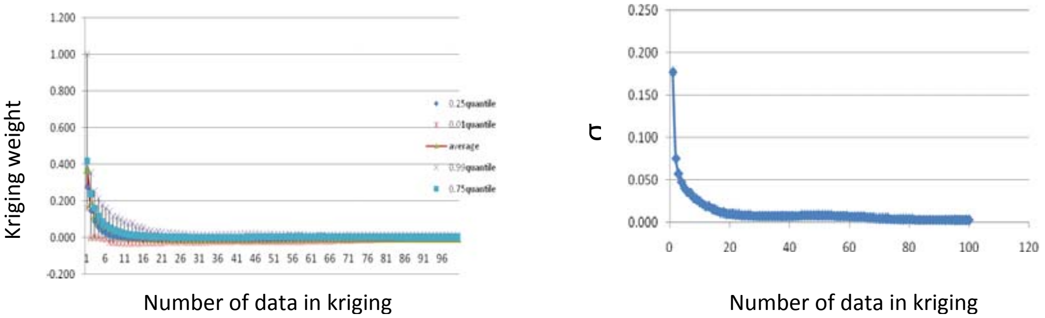


Figure 3 Change of kriging weights and standard deviation of the weights in 2D. The grid is 50*50 with 20 variogram range. λ = 0 when number of data is 19. σ rarely changes when the number of data is 31.

The typical results for 3D can be seen at Table 1.4 and Fig 1.4. The adequate number of data to use in 3D kriging depends on variogram range like 1D. The largest number of two criteria for different variogram range is 59, so 60 can be defined as the default data for 3D.

Table1.4 The typical results for 3D simple kriging.

grid	grid cell size	number of locations	range	data index where λ =0	data index where σ rarely change	The largest number of two criteria
25*25*25	2*2*2	15625	5	15	51	51
			20	29	59	59
			35	33	57	57
			50	37	57	57
			65	39	57	57

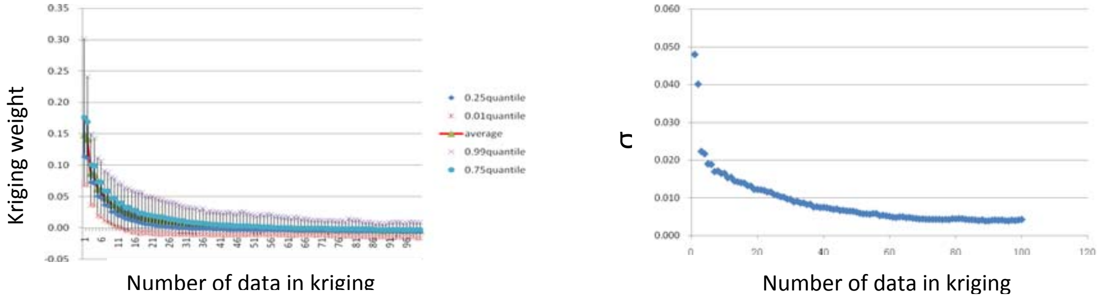


Figure 4 Change of kriging weights and standard deviation in 3D. The grid is 25*25*25 with 35 variogram range. λ = 0 when number of data is 33. σ rarely changes when the number of data is 57.

Although the variogram range has some effect to the results for 1D, 2D and 3D, its influence becomes less with the increase of variogram range. For example in 3D, the adequate number of data in kriging becomes a constant after the range becoming 70% of the domain size, see Figure 5.

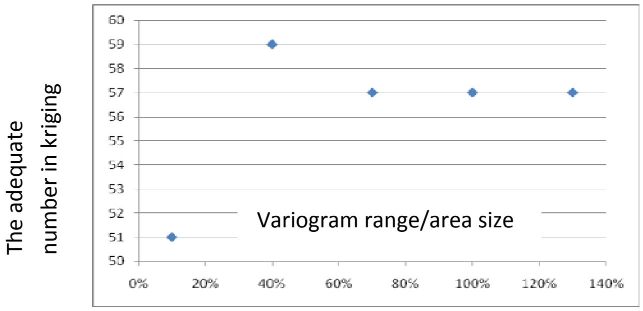
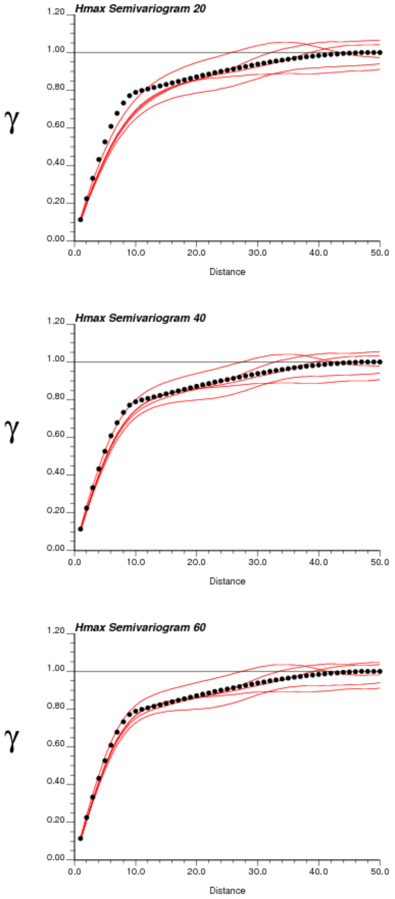


Figure 5 Change of the adequate number in kriging with dimensionless variogram range in 3-D, where the adequate number is the largest number of two criteria in Table 1.4.

Another consideration is decent variogram reproduction. The figure to the right gives an example for unconditional variogram reproduction using different data number in maximum horizontal direction. Data numbers are 20, 40, and 60 from top to bottom. The black point line is the theoretical variogram model, and the red solid lines represent results from simulation realizations. The more is data number used, the better is the result of variogram reproduction (the red solid lines close to the black point line). 60 data appears to achieve decent variogram reproduction in 3-D, which accord with the choice made before.



Conclusion

In order to choose an adequate number of conditioning data to maintain decent results for kriging but not cost too much computational time, a number of numerical studies in different situations are made to get the adequate data number for kriging. It can be found that the dimensionality is the most important. The variogram has a little effect to the choice, and the grid spacing has nothing to do with the results. Using the largest data number where λ =0 and σ rarely changes as the adequate number for kriging, the default number for krging in 1D,2D and 3D are ultimately choosed, and they are 30, 40 and 60 respectively.