Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

DATA ENHANCEMENT, SMOOTHING, RECONSTRUCTION AND OPTIMIZATION BY KRIGING INTERPOLATION

Hasan Gunes
Hakki Ergun Cekli

Dept. of Mechanical Engineering
Inonu Caddesi, No:87, Istanbul Technical University
34437 Istanbul, TURKEY

Ulrich Rist

Institute of Aerodynamics and Gasdynamics
Pfaffenwaldring 21, University of Stuttgart
70569 Stuttgart, GERMANY

ABSTRACT

The performance of Kriging interpolation for enhancement, smoothing, reconstruction and optimization of a test data set is investigated. Specifically, the ordinary two-dimensional Kriging and 2D line Kriging interpolation are investigated and compared with the well-known digital filters for data smoothing. We used an analytical 2D synthetic test data with several minima and maxima. Thus, we could perform detailed analyses in a well-controlled manner in order to assess the effectiveness of each procedure. We have demonstrated that Kriging method can be used effectively to enhance and smooth a noisy data set and reconstruct large missing regions (black zones) in lost data. It has also been shown that, with the appropriate selection of the correlation function (variogram model) and its correlation parameter, one can control the 'degree' of smoothness in a robust way. Finally, we illustrate that Kriging can be a viable ingredient in constructing effective global optimization algorithms in conjunction with simulated annealing.

1 INTRODUCTION

In this paper, we investigate the ability of Kriging interpolation to smooth and enhance the resolution of experimental data as well as to implement it in global optimization algorithms. Kriging is a statistical tool useful in many disciplines such as geology, thermo-fluid systems, process engineering, environment, meteorology and medicine. It has been named after D.G. Krige, a South African mine engineer who developed the procedure in order to predict mine ore ground water reserves more accurately from multi-point measurements (Krige 1951). Recently, it has been successfully applied to data recovery and reconstruction of randomly generated laminar gappy flow fields of uniform flow past a circular cylinder (Gunes, Sirisup, and Karniadakis 2006). A recent study showed that Kriging interpolation can be used successfully for resolution enhancement and for reconstruction of large spatial gappiness for a mixed convection data (Cekli and Gunes 2006).

Kriging is an unbiased estimation procedure which uses known values and a variogram to determine unknown values. Based on the variogram, optimal weights are assigned to known values in order to calculate the data at unknown points.

The variogram characterizes the spatial continuity or roughness/smoothness of a data set (Davis 2002). The variogram analysis consists of constructing an experimental variogram from the data and fitting a variogram model to the experimental variogram. The experimental variogram is calculated by averaging one half the difference squared of the values over all pairs of observations with the specified separation distance h and possible direction,

$$\gamma(h) = \frac{1}{2L}\sum_{i}^{L}\left(z_i - z_{i+h}\right)^2 \tag{1}$$

where, z_i is the value of the variable at point i, and z_{i+h} is the value of the variable at h separation distance away from point i. γ is called the variance. The variogram model is usually chosen from a set of mathematical functions that describe the spatial relationship. The appropriate model is chosen by matching the shape of the curve of the experimental variogram to the shape of the curve of the mathematical function (i.e., polynomial, exponential, the Gaussian, etc.) (Davis 2002, Isaaks and Srivastava 1989). The selection of a suitable variogram model is a crucial step of Kriging procedure, as it has an important effect on the weights and estimation error.

Kriging gives a linear weight for each known points to estimate a new point, and unlike inverse-distance weighted interpolation, the weights depend on the spatial dependence and the function values of the data set (i.e., on the variogram). We refer (Cekli and Gunes 2006, Cekli 2007, Davis 2002, Isaaks and Srivastava 1989, Cressie 1993) for application of the procedure in detail.

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

In general, Kriging is a computationally expensive procedure and also, depending on the size of the design (known) dataset, a large memory may be required to construct and evaluate the coefficient matrix required for evaluations of weights. In order to eliminate memory problems, increase the speed of the procedure significantly, and to obtain a more robust smoothness, we propose 2D *line* Kriging. This procedure employs a 1D spatial correlation instead of a 2D spatial correlation. 2D line Kriging re-evaluates known points on constant lines of a 2D dataset. In order to have an equivalent 2D correlation, we take constant x and y lines, consecutively, and build a Kriging model to estimate data on each constant line separately. Because this procedure uses correlations in both horizontal and vertical directions (consecutively) it is also effectively a 2D method. It is apparent that 2D line Kriging is suitable only for Cartesian grids. Data smoothing via 2D line Kriging can be comparable to smoothing via digital filters, which we give details in next section.

2 DIGITAL FILTERING

Filtering is used to pass certain frequency components in a signal through the system without any distortion and to block other frequency components. The range of frequencies that is allowed to pass through the filter is called the pass band, and the range of frequencies that is blocked by the filter is called the stop band (Oppenheim and Schafer 1989).

A type of filter which is called low-pass filter passes low-frequency components below a certain specified frequency f_c and blocks all high-frequency components of a signal above f_c.

In Figure 1 the amplitude response of the system is given for different values of cut-off frequency. As shown in Figure 1 the characteristics of the Butterworth filter are that they are maximally flat in the pass band and monotonic overall.

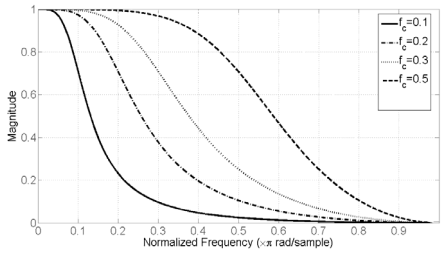


Figure1: Amplitude response of the second order low-pass IIR Butterworth filter for different values of cut-off frequency.

Traditionally, digital filters have been classified into two large families; first those whose transfer function does not have a denominator, and second those whose transfer function has a denominator. Since the filters of the first family admit a realization where the output is a linear combination of a finite number of input samples, they are sometimes called non-recursive filters. For these systems, it is more customary and correct to refer to the impulse response, which has a finite number of non-null samples, thus calling them Finite Impulse Response (FIR) filters. On the other hand, the filters of the second family admit only recursive realizations, thus meaning that the output signal is always computed by using previous samples of itself. The impulse response of these filters is infinitely long, thus justifying their name as Infinite Impulse Response (IIR) filters.

The following equation describes how the output y of a FIR filter is calculated from the input x. This equation simply says that the n^{th} output is a weighted average of the most recent N inputs. The mathematical expression of a FIR filter is,

$$y[n] = \sum_{m=0}^{N} h[m]x[n-m] \tag{2}$$

Since the time extension of the impulse response is $N+1$ samples, we say that the FIR filter has length $N+1$. The transfer function is obtained as the z-transform of the impulse response and it is a polynomial in the powers of z^{-1},

$$H(z) = \sum_{m=0}^{N} h_m z^{-m} = h_0 + h_1 z^{-1} + \ldots + h_N z^{-N} \tag{3}$$

Since such polynomial has order N, we also say that the FIR filter has order N. The IIR filter is the same as the FIR filter, only with an additional summation term which feeds back previous outputs. These filters can produce superior results with much less computational cost, but they are harder to design, and can suffer from stability problems if improperly designed. IIR filter is represented by a difference equation where the output signal at a given instant is obtained as a linear combination of samples of the input and output signals at previous time instants. Moreover, an instantaneous dependency of the output on the input is also usually included in the IIR filter. The difference equation that represents an IIR filter is,

$$y[n] = \sum_{m=1}^{N} a_m y[n-m] + \sum_{m=0}^{M} b_m x[n-m] \tag{4}$$

The array a holds weighting coefficients for feeding back the previous N outputs into the current output value. While the impulse response of FIR filters has a finite time extension, the impulse response of IIR filters has, in general, an infinite extension. The transfer function is obtained by application of the z-transform to the above equation. The result is the rational function $H(z)$ that relates the z-transform of the output to the z-transform of the input,

$$H(z) = \frac{b_0 + b_1 z^{-1} + \ldots + b_M z^{-M}}{1 + a_1 z^{-1} + \ldots + a_N z^{-N}} \tag{5}$$

The filter order is defined as the degree of the polynomial in z^{-1} that is the denominator of the above equation (Mitra 2001).

In this study the digital filters with these to smooth noisy data have been designed by Matlab routines. Routines one can construct a variety of digital filters for certain cut-off frequency and filter order.

3 CREATING AND DISTURBING TEST DATA

We create a synthetic (analytical) data set and disturb it by adding random noise as follows.

$$u(x,y) = \left(x^5 - y^3 + y^2\right) e^{-\left(x^2+y^2\right)} + rand(x,y) \quad (6)$$

Then, the root-mean-square error (*rms*) of the disturbed data can be evaluated by taking the difference of actual data and the disturbed data. The following equation defines the *rms* error,

$$rms = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[z_A(x,y) - z_p(x,y)\right]^2} \quad (7)$$

where, z_A is the actual value and z_P is the disturbed or estimated value. The performance of the different methods can be evaluated by comparing the *rms* values.

4 KRIGING INTERPOLATION FOR OPTIMIZATION PROBLEMS

A possibility of using Kriging interpolation in optimization algorithms is investigated. For example, in the field of combinatorial optimization the aim is to develop efficient techniques to find global minimum or maximum values of a function of many independent variables. With this context, we construct a combinatorial optimization algorithm using Kriging interpolation in combination with Simulated Annealing (SA) algorithm. The difficulty in many optimization algorithms is that they effectively find a local minima, but they cannot get away from there to the global minima.

4.1 Simulated Annealing

Simulated Annealing (SA) is a generic probabilistic meta-algorithm for the global optimization problems, as introduced in (Kirkpatrick, Gelatt, and Vecchi, 1983). Its major advantage over other methods is an ability to avoid becoming trapped at local minima. SA is based on an analogy involving heating and controlled cooling of a material to increase the size of its crystals and thus reduce the defects (the annealing process in metallurgy). In SA algorithm, the current solution is replaced by a random "nearby" solution, chosen with a probability that depends on the difference between the corresponding function values and on a global parameter *T* (called usually the temperature), that is gradually decreased during the process. The dependency is such that the current solution changes almost randomly when T is large, but increasingly "downhill" as T goes to zero. The allowance for "uphill" moves saves the method from becoming stuck at local minima. The algorithm is based upon that of (Metropolis et al. 1953), which was originally proposed as a means of finding the equilibrium configuration of a collection of atoms at a given temperature.

In this paper, a Kriging model is built and validated from available data. In general, the data may come from experiments or computer simulations such as CFD analyzes and the task is to determine the optimum variables for this data set. For the experiments or computations, it is important to determine the design sites effectively. We use Latin Hypercube Sampling (LHS) algorithm to determine design sites. LHS is based on random numbers and ensures that all portions of the vector space are represented (McKay, Conover, and Beckman, 1979). Having determined the design sites by LHS through the data field, we build Kriging model based on design sites for the optimization algorithm. Once Kriging model is build, we start with initial values of variables and predict a value for this point by Kriging interpolation and set the predicted solution as the best solution of the problem. Then we change the values of the variables nearby the current values and estimate a solution for this configuration. If the current configuration is better than the best configuration, we set it as the best configuration, if not, it is treated probabilistically according to the probability function given as follows

$$P = \exp\left[\left(e - e_n\right)/T\right] \quad (8)$$

where, e is the best value and e_n is the last estimated value, T is the controlling parameter (temperature) of the problem and it is decreased during the optimization process. The parameter T can be calculated using as,

$$T = T_0\left(1 - k/K\right)^c, \quad (9)$$

where T_0 is the initial temperature, c is a parameter controlling the temperature decreasing (cooling) rate (if $c = 1$, the cooling is linear), k is the number of evaluated steps (i.e., current step) and K is the maximum step number.

Uniformly distributed random numbers are generated between interval (0,1) and compared with probability function value and when the probability function value is greater than random value, the current configuration is taken as the best configuration. These steps are repeated for a sufficient time and finally a best solution is obtained.

5 RESULTS

5.1 Smoothing by Kriging Interpolation

Different variogram models such as exponential, spherical, spline and the Gaussian model have been used to smooth the noisy dataset (i.e., modifying the design points by Kriging). It is noted that all the reported variogram models except for the Gaussian model, reproduce the original

noisy (disturbed) data (within the numerical accuracy). This means that only the Gaussian model has a capability to smooth data. In other words, only the Gaussian model changes the values of the given points from experiment while the other models merely re-produce the given data values.

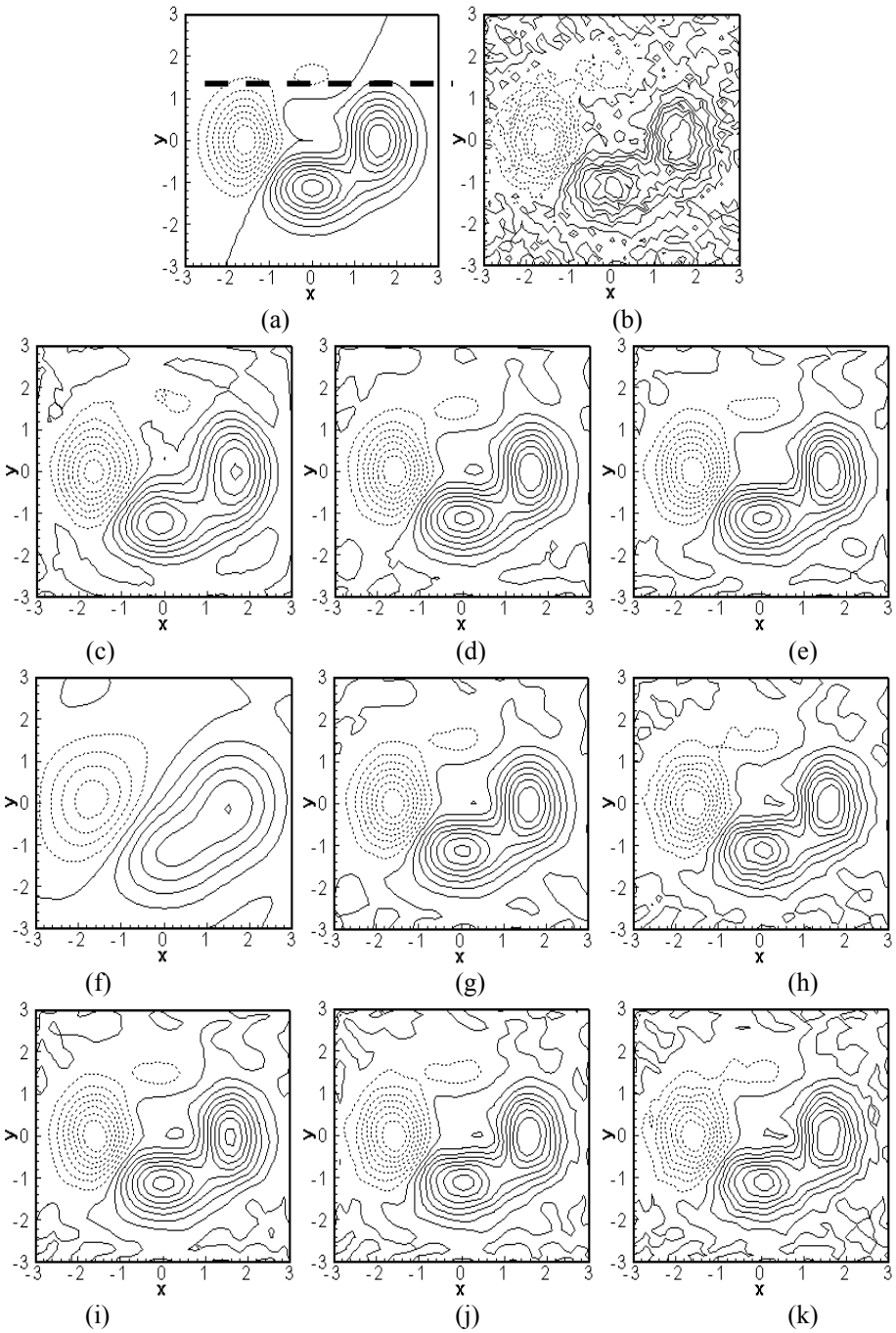


Figure 2: (a) Actual data, (b) disturbed data (rms = 0.05), (c) smoothed data by Kriging (θ = 0.1) (rms = 0.035), (d) smoothed data by Kriging(θ = 0.5) (rms = 0.016), (e) smoothed data by Kriging (θ = 1) (rms = 0.02), (f) smoothed data by IIR Butterworh filter, f_c = 0.1 (rms = 0.085), (g) smoothed data by IIR Butterworh filter, f_c = 0.3 (rms = 0.0154), (h) smoothed data by IIR Butterworh filter, f_c = 0.5 (rms = 0.0227), (i) smoothed data by 2D line Kriging (θ = 0.5) (rms = 0.019), (j) smoothed data by 2D line Kriging (θ = 1) (rms = 0.023), (k) smoothed data by 2D line Kriging (θ = 2) (rms = 0.029). Kriging interpolations are based on Gaussian model.

In Figure 2, the actual solution of the test data, its disturbed case and the smoothing result by various procedures (e.g, Kriging interpolation with the Gaussian model, digital filtering for different cut-off frequencies, and 2D line Kriging) are shown. The rms-error values are given to compare the procedures. When the results of 2D line Kriging are compared with those of ordinary Kriging, it is seen that 2D line Kriging leads to smoother data.

The rms-error value for noisy data is 0.05 and if it is smoothed by the Gaussian model with correlation parameter $\theta = 0.1$ it becomes $rms = 0.035$, and for $\theta = 0.5$ and $\theta = 1$, we have $rms = 0.016$ and $rms = 0.02$, respectively. It can be clearly seen that the correlation parameter in the Gaussian model can be used to control the level of smoothing. A high value of correlation parameter means a low smoothness and vice versa.

2D line Kriging works reasonably for data smoothing and correlation parameter has an importance on the result. In addition, the smoothing results by a 2D IIR Butterworth type low–pass, double-sided (zero-phase) filter with different cut-off frequencies are shown in Figure 2. For a small cut-off frequency, a very smooth result can be obtained but it is significantly different from the actual data, i.e., it has a higher value of rms-error than the noisy data! On the other hand, a large cut-off frequency gives a relatively smooth but still noisy data and has a large value of rms-error. When the cut-off frequency is selected appropriately, the digital filtering gives a result that is quite smooth and close to the actual data with a low value of rms-error. As can be seen in Figure 2 by dashed line, we also take a constant y line and plot actual, disturbed and smoothed data for a detailed comparison in Figure 3.

5.2 Enhancement by Kriging Interpolation

We obtained the test data with a fine mesh (41x41 mesh) and reduced resolution to some other coarse meshes (such as 7x7, 9x9, etc.) then tried to estimate the original values again by Kriging interpolation, and compared estimation performance for each different resolutions. In Figure 4, the contour plots of the original data and its "low-resolved" versions are given. These low resolved data are enhanced to the same grid points of the original data and the contour plots of enhanced data are given on the right column in Figure 4. As we know the actual values of interpolated data we can calculate *rms* error value for the data enhancement by (7).

5.3 Reconstruction of Black Zone by Kriging Interpolation

In order to illustrate that Kriging interpolation can be used to recover large missing regions in a data set, we create black zones (continuous large missing regions) with different sizes by discarding all the existing data in a rectangular zone in the data set which are shown in Figure 5. Using Kriging, the black zones are reconstructed. A Gaussian model is employed using the available data outside of the black zone. In Figure 5, the contour plot of the original data is shown and the large missing region (black zone) is outlined by a dashed rectangular line and shaded. In Figure 5, the contour plot of the reconstructed data is given in the right column. When we compare the two contour plots, we see that the black-zone region is recovered reasonably well by Kriging interpolation. A detailed comparison is shown in Figure 6, where the original data and the reconstructed data on a constant line on x = -1,05 are shown for reconstruction of two sizes of black zone in the test data.

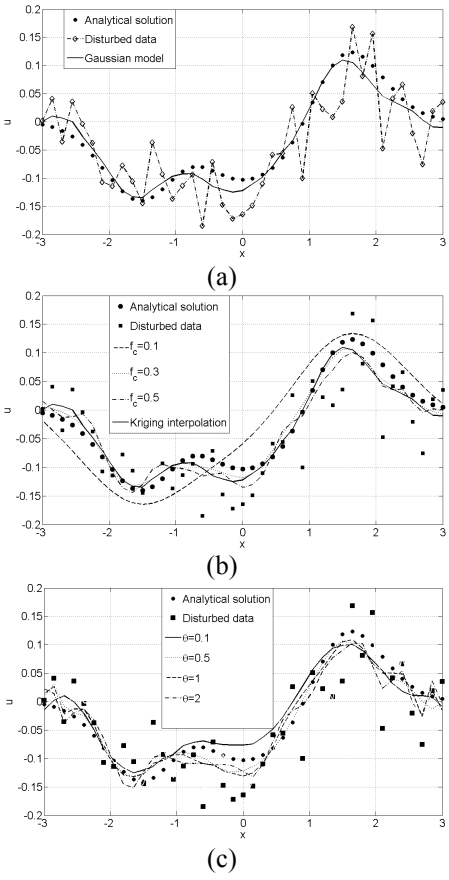


Figure 3: Data distribution on y = 1,35, (a) Ordinary Kriging, (b) IIR Butterworth Filter, (c) 2D line Kriging

5.4 Optimization by Kriging Interpolation

In Figure 7, three examples for global optimization process are given to account the effect of the probability function values. We again use the same test data given in (6) to investigate the optimization problem. The contour of the actual test data are plotted in Figure 7. Next, we obtain the

Kriging model of the test data employing only for 50 LHS points shown as black circles in Figure 7. Kriging model is validated by re-calculation and comparison the existing data. By using this Kriging model in our optimization process, our task is to determine the highest value of the variable in the data set. The maximum value of the test data is at (*x, y*) = (1.8, 0). In Figure 7, the black dots show trial sites and the trajectory of the best points during the optimization process is shown by dashed line. When a new best point is found, it is denoted by a triangular symbol. When we take a close look at Figure 7, it can be seen that for the same initial temperature (T_0), for larger values of *c,* the probability function *P* values reach to zero more rapidly and it effects the location of trial sites.

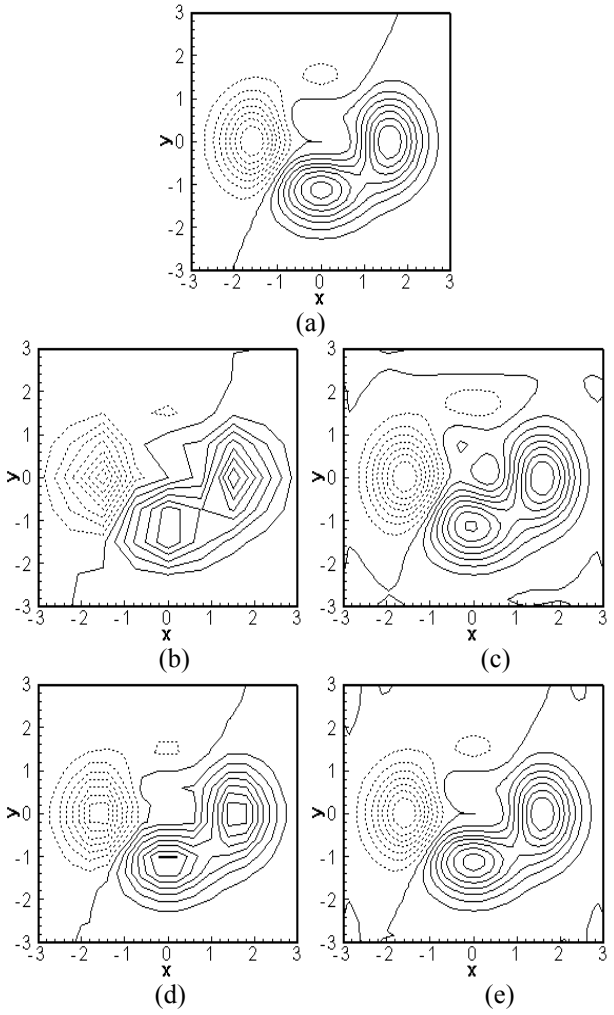


Figure 4: Enhancement of coarse data, (a) actual data, (b) low-resolved data (9x9 mesh), (c) enhancement of data to 41x41 mesh from 9x9, (rms = 0.0194) (d) low-resolved data (16x16 mesh), (e) enhancement of data to 41x41 mesh from 16x16 (rms =$3.398*10^{-4}$).

In Figure 7a, there is a better distribution of trial sites than in Figure 7b. It can also be seen that for a larger initial temperature, we obtain relatively larger P values and it reaches zero only at the end of optimization process. This behaviour results in a well distributed trial sites through the optimization domain as it can be seen in Figure 7c.

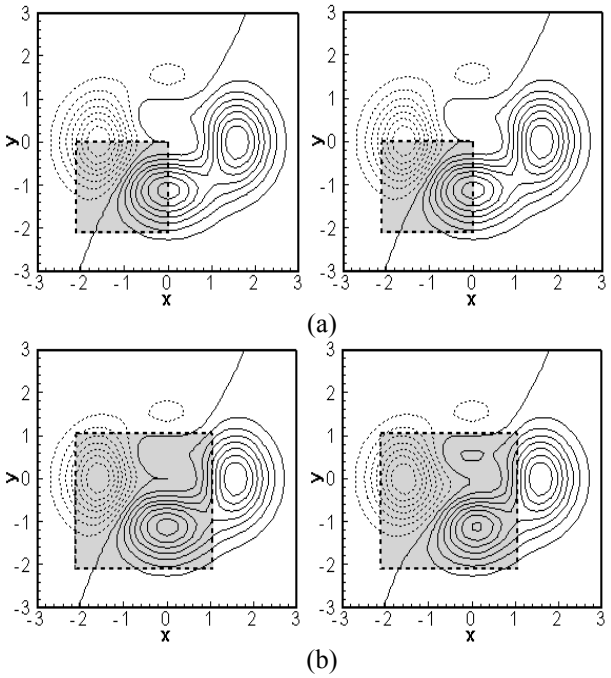


Figure 5: Reconstruction of black zones, (a) 12.25% missing, (b) 26.7% missing. (Left: actual data, right: reconstructed data).

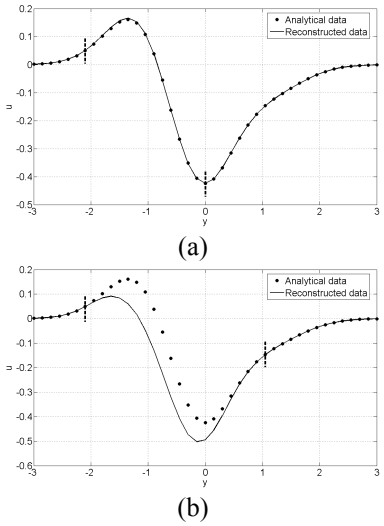
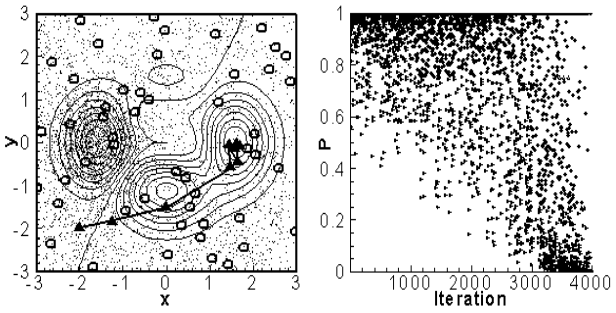


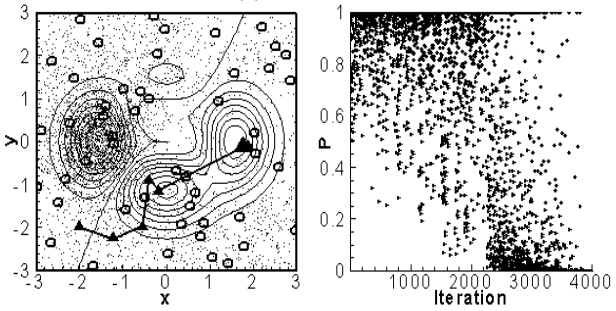
Figure 6: Data distribution on x = -1,05, (a) 12.25%, (b) 26.7%.

6 CONCLUSION

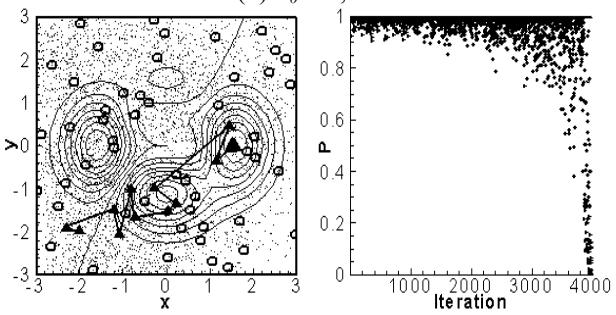
Kriging interpolation and digital filters are investigated in order to smooth and enhance the resolution of a test data. The investigations have shown that selection of model variogram and its correlation parameter are important for predictions. This selection step needs experience and well understanding on physical aspects of the problem. The level of smoothing can be controlled by the correlation parameter. In order to construct a fast and effective procedure, we employ a new type of Kriging (i.e., line Kriging) that uses one-dimensional correlation instead of two- or three-dimensional correlations. Digital filters such as Butterworth type filter were also applied for data smoothing purposes. It has been shown that Kriging can be used in conjunction with simulated annealing for global optimization problems.



(a) $T_0 = 1$, $c = 1$



(b) $T_0 = 1$, $c = 2$



(c) $T_0 = 10$, $c = 1$

Figure 7: Simulated annealing with Kriging interpolation, (left: process of finding maxima, right: probability function values)

ACKNOWLEDGMENTS

The authors would like to acknowledge support from their respective universities to perform this research as well as the financial support for exchange visits provided by Tübitak and internationales Büro of the German ministry of education and research (BMBF) under the project TUR 05/003.

REFERENCES

Cekli, H.E., and H. Gunes. 2006. Spatial resolution Enhancement and reconstruction of mixed convection data using Kriging method. In *ASME International Mechanical Engineering Congress and Exposition*. November 5-10, Chicago, Illinois, USA.

Cekli, H.E. 2007. *Enhancement and smoothing methods for experimental data: application to PIV measurements of a laminar separation bubble*. M.Sc. thesis, Department of Mechanical Engineering, Istanbul Technical University, Istanbul, Turkey.

Cressie, N.A.C. 1993. *Statistics for Spatial Data*. New York: Wiley.

Davis, J.C. 2002. *Statistics and data analysis in Geology*. New York: J. Wiley.

Gunes, H., S. Sirisup, and G.E. Karniadakis. 2006. Gappy data: To Krig or not to Krig. *Journal of Computational Physics*. 212: 358-382.

Isaaks, E.H., and R.M. Srivastava. 1989. *Introduction to Applied Geostatics*. New York: Oxford University Press.

Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi. 1983. Optimization by Simulated Annealing. *Science.* 220: 671-680.

Krige, D.G. 1951. A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. of the Chem., Metal. and Mining Soc. of South Africa.* 52: 119-139.

McKay, M.D., W.J. Conover, and R.J. Beckman. 1979. A comparision of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*. 21.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *J. of Chemical Physics*. 21: 1087-1092.

Mitra, S.K. 2001. *Digital Signal Processing*. 2nd ed. McGraw-Hill.

Oppenheim, A. V., and R.W. Schafer. 1989. *Discrete-Time Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall.

AUTHOR BIOGRAPHIES

HASAN GUNES is an Associate Professor in the Department of Mechanical Engineering at the Technical University of Istanbul, Turkey. He received a Ph.D. degree in mechanical engineering from Lehigh University, USA. His current research and teaching interests are in computational fluid mechanics, numerical methods, application of various data reconstruction, smoothing, enhancement and optimization techniques such as kriging and proper orthogonal decomposition for thermo-fluid systems. His e-mail address is <guneshasa@itu.edu.tr>

ULRICH RIST studied aerospace engineering at the University of Stuttgart and works as a professor at the institute of aerodynamics and gasdynamics. His main interests are in direct numerical simulations of instability and transition in boundary layers, flow visualizations and flow control. His e-mail address is <rist@iag.uni-stuttgart.de>

HAKKI ERGUN CEKLI is currently a Ph.D. student at the Fluid Dynamics Laboratory of Eindhoven University of Technology. He received a master's degree in Mechanical Engineering at Istanbul Technical University. The subject of his current work is to modulate wind tunnel turbulence using an active grid and study the response on modulation in space and time. His e-mail address is <h.e.cekli@tue.nl>