International Journal of Computer Theory and Engineering, Vol. 2, No. 4, August, 2010
1793-8201

A Novel Interpolation Method Using Soft Data and Hard Data

Yi Du[1] and Ting Zhang[2]

Abstract—**Interpolation methods play an important role in many fields such as industrial, geological and military fields for prediction. However, it is quite difficult to predict the unknown information only by some sparse hard data in the process of simulation based on current popular interpolation methods. Accuracy of simulated images can be improved by using soft data and hard data. Multiple-point geostatistics (MPS) originates from geostatistical fields and allows extracting multiple-point structures from training images, after that MPS can copy these structures to the regions to be predicted. To simulate or predict information accurately, an interpolation method using soft data and hard data in MPS is proposed. Dimension reduction is made by filters to reduce the CPU time and memory demand. All similar training patterns fall into a cell in the filter score space, which is created by filters. Finally, a training pattern is randomly drawn from a cell, and then is pasted back onto the unknown region to be predicted. The variogram curves of the simulated images are compared, showing that the structural characteristics of the image simulated by using both soft data and hard data are most similar to those of the training image.**

Index Terms—**interpolation; multiple-point geostatistics; soft data; hard data; filter**

I. Introduction

Interpolation methods for prediction are quite important and significant to the development of many scientific fields, which is also widely used in various fields such as medical, military, geological, meteorologic and mining fields. Although a number of interpolation methods were introduced, the accurate information prediction was still difficult to be realized, especially only with sparse conditional data. When conditional data are not quite available or even there are no conditional data, the ideas of indefinite interpolation can be applied [1]. Interpolation methods are mainly two types: "definite" methods and "indefinite" methods. The "definite" here means that the forms, parameters and results of interpolation functions are mostly definite. "Definite" methods include the inverse distance weighting method, the triangular mesh method, the basis function method, etc. The "indefinite" means that the forms of interpolation functions are indefinite and the selection of parameters in interpolation functions depends on the principles of statistics [1, 2]. The main "indefinite" interpolation methods are kriging and stochastic simulation in geostatistics. Kriging and stochastic simulation, both based on variogram which only describes the relations between two points in space and cannot reconstruct complex patterns such as curvilinear shapes, are called two-point geostatistics [3]. Because of the disadvantages of traditional two-point geostatistics, a new interpolation method called multiple-point geostatistics (MPS) was recently proposed to reproduce complex structures as well as to keep the flexibility of data conditioning. By reproducing high-order statistics, MPS allows capturing structures from a training image, then anchoring them to the specific model data. A training image is a numerical prior model which contains the structures and relationship existing in realistic models [4].

Originally, MPS can only simulate discretized variables, which really has limited its application fields and only is suitable to predict the information with few kinds of states. But in the studies of real information prediction, continuous variables with multiple states are widely existent.

In many fields, there are two types of data: hard data and soft data. It is often considered that hard data are results based on the measurement and observation of objective articles and phenomena, but soft data are statistical data subjectively or vaguely judged by people or equipment. For example, in reservoir characterization, in addition to hard well data, other types of soft data such as seismic data are available. Soft data typically provide an extensive coverage of the field under study although with low resolution. It is necessary to condition the simulated models to all these different types of data to improve the accuracy [3, 4].

Zhang [5, 6] proposed a method based on filters to simulate continuous variables. According to the theories from Zhang, we propose a novel interpolation method based on continuous MPS integrating soft data with hard data. Using soft data and hard data as conditional data, the accuracy of predicted information is improved. Experimental results show that our method is practical and effective.

II. Ideas and Methods

A. Data Templates and Data Events

A training image is scanned by using a data template τ_n that comprises n locations u_α and a central location u. The u_α is defined as: $u_\alpha=u+h_\alpha(\alpha=1,2,\ldots,n)$, where h_α is the vectors describing the data template. For example, in Fig. 1(a), h_α is the 80 vectors in the square 9×9 template. In Fig. 1(b), h_α is the 26 vectors in the cubic 3×3×3 template with a blue center u.

Consider an attribute S that has K possible states $\{s_k; k=1,2,\ldots,K\}$. A data event d_n of size n, centered at location u, constituted by n vectors u_α in τ_n is defined as [4]:

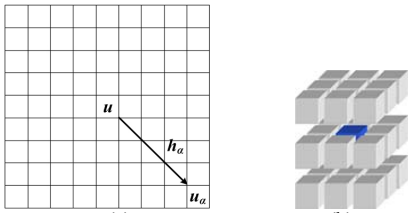
$$d_n=\{S(u_\alpha)=s_{k_\alpha}; \alpha=1,2,\ldots,n\} \quad (1)$$

where $S(u_\alpha)$ is the state at the location of u_α within the

This work is supported by the Innovation Program of Shanghai Municipal Education Commission (09YZ454), and the Knowledge Innovation Project of the Shanghai Education Commission

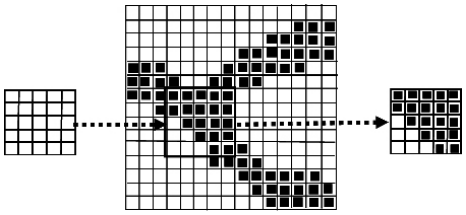
[1]School of Computer and Information, Shanghai Second Polytechnic University , Shanghai, China(email:duyi@mail.ustc.edu.cn).
[2]National Key Laboratory of Science and Technology on C4ISR, Nanjing,China(email: tingzh@mail.ustc.edu.cn).

template. d_n actually means that n values $S(u_1)\ldots S(u_n)$ are jointly in the respective states $s_{k_1} \ldots s_{k_n}$. Fig. 2 illustrates the procedure of capturing a data event with a 5×5 template. Fig. 3 illustrates two data events captured by the data templates displayed in Fig. 1(a) and Fig. 1(b) respectively. The different colors in Fig. 3 mean different states of an attribute.



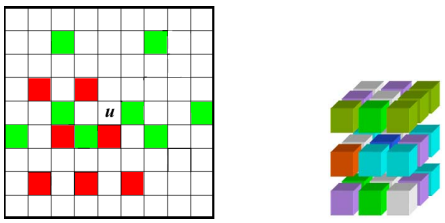
(a) (b)

Figure 1. Data templates.(a)a 2D data template;(b)a 3D data template.



(a) (b) (c)

Figure 2. Procedure of a data event captured by a 5×5 data template. (a) a 5×5 data template;(b) a 15×15 training image;(c)a data event.

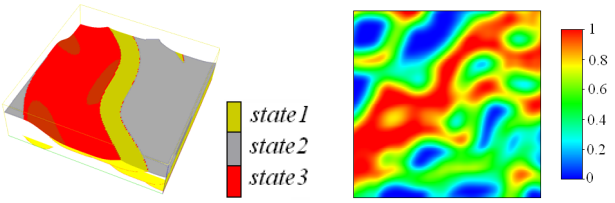


(a) (b)

Figure 3. Data events captured by the data templates displayed in Fig.1. (a) captured by a 2D data template;(b) captured by a 3D data template.

B. Training Images

A training image is considered as a lot of repetitive information with some special characteristics that are supposed to be existent in the fields to be predicted. They are purely conceptual assemblages of patterns, which possibly are not quite accurate, and don't need to honor any conditioning data. A training image can be viewed as prior structural models that show how information should be linked together. Training images can be acquired through many ways such as images of remote sensing, hand-drawn sketches, pictures of geological outcrops, etc [6]. There are two types of training images: discretized images and continuous images. Fig. 4(a) is a discretized 3D training image composed of three kinds of states, which are illustrated by yellow, gray and red. Fig. 4(b) is a continuous 2D training image composed of a continuous variable, whose value varies from 0 to 1, as shown in the state bar in Fig. 4(b).



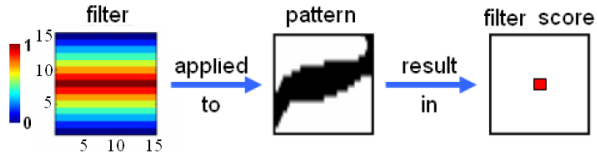
(a) (b)

Figure 4. Training images.(a) a discretized training image;(b) a continuous training image.

C. Filters and Filter Scores

Continuous MPS simulation uses a few filters to classify different patterns in a training image to realize dimension reduction. A filter is defined like a data template centered at u but with a set of weights attached to each node in the locations of the filter. When scanning the training image using a filter, we can put a filter over a local pattern, and then apply this filter to the pattern to obtain a value of combining the filter weights and the state values of the training pattern below the filter. The value is called "filter score", and each score is considered to be a summary of a pattern, resulting in a tremendous dimension reduction [6, 7].

Fig. 5 illustrates the process of obtaining a filter score with a specific 2D filter [6]. Fig. 5(a) is a 2D filter with 15×15 nodes and different weights. The state values in the filter are shown in different colors. This filter is used to scan the pattern shown in Fig. 5(b) and a filter score is obtained finally, as shown in Fig. 5(c). It is seen that the 15×15 nodes are represented by a value (the red node in Fig. 5(c)), resulting in a dimension reduction from 15×15 to 1×1.



(a) (b) (c)

Figure 5. Illustration of obtaining a filter score of a training pattern. (a) a filter; (b) a training pattern; (c) a filter score.

In the 2D condition, the filter score is defined as:

$$S_k(i,j) = \sum_{y=-m}^{m} \sum_{x=-m}^{m} f_k(x,y) T(i+x, j+y), \quad k = 1,\ldots,6 \qquad (2)$$

where $S_k(i,j)$ is the filter score; (i,j) is the coordinate of the central node u in the data template; $f_k(x,y)$ is k-th filter; x and y vary from $-m$ to m; $T(i+x,j+y)$ is the value of a local pattern located at $(i+x,j+y)$ in the training image; $2m+1$ is the number of nodes in the X and Y directions respectively. In the 3D condition, there are totally 9 filters. Each filter is defined to characterize different aspects of the local training pattern. The corresponding filter score is defined as:

$$S_r(i,j,k) = \sum_{z=-m_2}^{m_2} \sum_{y=-m_1}^{m_1} \sum_{x=-m_1}^{m_1} f_r(x,y,z) T(i+x, j+y, k+z), \quad r = 1,\ldots,9 \qquad (3)$$

where $2m_2+1$ is the number of nodes in the Z direction; $f_r(x,y,z)$ is the r-th filter defined over the 3D template withthe size of $n=(2m_1+1)^2(2m_2+1)$ locations.

The following are nine filters defined in the 3D condition [6].

$$f_1(x,y,z)=1-\frac{|y|}{m_1}\in[0,1], y=-m_1,...,+m_1 \quad (4)$$

$$f_2(x,y,z)=1-\frac{|x|}{m_1}\in[0,1], x=-m_1,...,+m_1 \quad (5)$$

$$f_3(x,y,z)=1-\frac{|z|}{m_2}\in[0,1], z=-m_2,...,+m_2 \quad (6)$$

$$f_4(x,y,z)=y/m_1\in[-1,1], y=-m_1,...,+m_1 \quad (7)$$

$$f_5(x,y,z)=x/m_1\in[-1,1], x=-m_1,...,+m_1 \quad (8)$$

$$f_6(x,y,z)=z/m_2\in[-1,1], z=-m_2,...,+m_2 \quad (9)$$

$$f_7(x,y,z)=\frac{2|y|}{m_1}-1\in[-1,1], y=-m_1,...,+m_1 \quad (10)$$

$$f_8(x,y,z)=\frac{2|x|}{m_1}-1\in[-1,1], x=-m_1,...,+m_1 \quad (11)$$

$$f_9(x,y,z)=\frac{2|z|}{m_2}-1\in[-1,1], z=-m_2,...,+m_2 \quad (12)$$

The f_1~f_3 filters are respectively used to characterize the average structures in the North-South, East-West and Top-Bottom directions. The f_4~f_6 filters are used for gradient in the North-South, East-West and Top-Bottom directions. The last three filters f_7~f_9 characterize the curvature also in the three directions.

D. Process of Simulation Using Soft Data and Hard Data

The 3D nine filters will be applied one by one to scan a 3D training image. A local training pattern captured by the nine filters will be characterized by nine filter scores. Finally, a nine-dimensional filter score space is created and each score in the filter score space corresponds to a local training pattern.

Similar patterns in the training image are stored in the same groups. These groups are called "cells", which include all the patterns with close filter scores. The patterns in the same cell are averaged to generate a value, which is called a "prototype" and can represent the patterns in the cell. Normally, a filter score space will be partitioned for two times to create cells and sub-cells. Fig. 6 illustrates a two-step partition of a two-dimensional filter score space. The first step is to partition the space with solid lines to create 9 cells. S_1 and S_2 are respectively the maximal filter scores in the two dimensions. The blue dashed lines mean the second partition of the score space, which partition the current cells to create sub-cells. Each black point in a cell corresponds to a training pattern. During the simulation, we can define a random visiting path of all unsampled nodes. For each node to be simulated, we can acquire the filter score of the current data events by filters. After comparing the filter score from the region to be simulated and that from the training pattern, a training pattern closest to the current conditioning data is drawn and pasted back onto the region to be simulated. Loop until all the nodes in the visiting path are simulated.

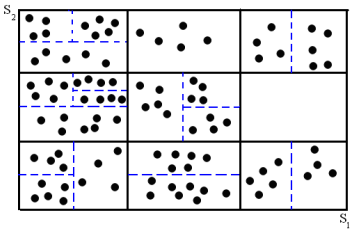


Figure 6. Illustration of partition for a two-dimensional filter score space.

Only hard data will be used in the above process. However, to improve the accuracy of simulation, soft data can be involved in the simulation. The following notations are used:

- *A* denotes a set of hard local data within a neighborhood defined by a template τn centered at location u. The conditioning data set A includes both the original hard data and simulated values at previously visited nodes.
- *B* denotes the soft data.

The following procedure is used when simulating each unknown node in the visiting path [5, 6]:

Step 1. Acquire the data event A within the template τn. If there are no conditional data in A, search the prototype that is closest to the soft data B and then paste it to the region to be simulated.

Step 2. If A is not empty, the soft data B is used to fill in the unknown nodes or pixels in the data template. Then search for the closest prototype to the full data event and paste it to the region to be simulated.

Step 3. Loop step 1 and step 2 until all the nodes in the visiting path are simulated. Then one stochastic image has been generated.

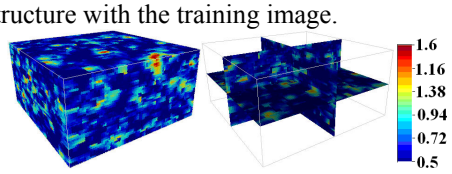
The above procedure involves the soft data and hard data simultaneously, which will improve the simulation accuracy.

III. Experimental Results and Analyses

A. Comparison with Predicted Results Using Hard Data only and Unconditional Simulation

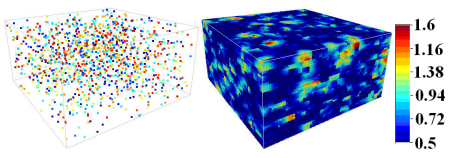
Fig. 7 shows the training image (80×80×40 voxels), whose value varies from 0.5 to 1.6. Fig. 7(a) and (b) are respectively the exterior and cross-sections (X=40, Y=40, Z=20) of the training image. The average of the training image is 0.6121, and the variance is 0.041. The training image provides reference data for us to evaluate the simulated results. As shown in Fig. 8(a), 0.5% sample points used as hard data for continuous MPS simulation are randomly extracted from the training image, whose average is 0.6335. The soft data with the size of 80×80×40 voxels are shown in Fig. 8(b).

Then our proposed MPS method is tested. The simulated results (80×80×40 voxels) using both soft data and hard data are shown in Fig. 9. It is seen that the simulated image has similar structure with the training image.



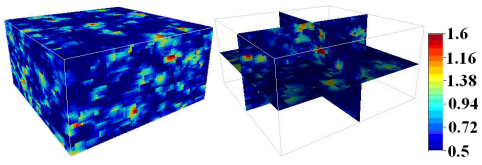
(a) (b)

Figure 7. The training image. (a) exterior; (b) cross-sections (X=40, Y=40, Z=20).



(a) (b)

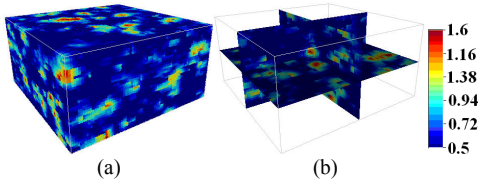
Figure 8. Hard data and soft data. (a) hard data extracted from the training image; (b) soft data.



(a) (b)

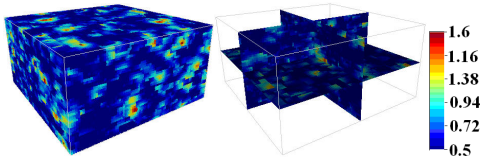
Figure 9. Simulated results using soft data and hard data. (a) exterior; (b) cross-sections (X=40, Y=40, Z=20).

Using MPS, one realization (80×80×40 voxels) of unconditional simulation and one realization (80×80×40 voxels) of simulation using hard data only were generated. The simulated results are shown respectively in Fig. 10 and Fig. 11, in which the structures existing in the training image are well reproduced.



(a) (b)

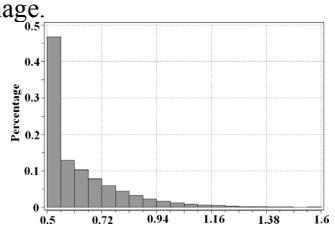
Figure 10. Simulated results of unconditional simulation. (a) exterior; (b) cross-sections (X=40, Y=40, Z=20).



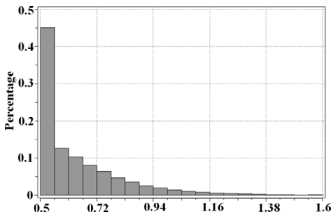
(a) (b)

Figure 11. Simulated results using hard data only. (a) exterior; (b) cross-sections (X=40, Y=40, Z=20).

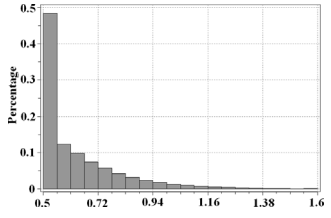
The histograms of the training image, the image using soft data and hard data, the image using hard data only and the image of unconditional simulation are shown in Fig. 12, respectively. It is seen that the distribution of simulated values is quite similar between all the simulated images and the training image.



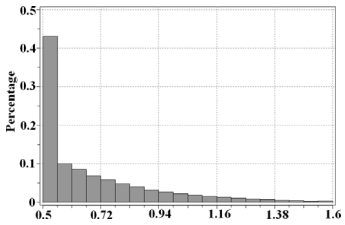
(a) training image



(b) using soft data and hard data



(c) using hard data only



(d) unconditional simulation

Figure 12. The histograms of the training image, the image using soft data and hard data, the image using hard data only and the image of unconditional simulation.

The memory demand, CPU time, average and variance of the simulated results are shown in TABLE I (we used a computer with a 2G Athlon CPU, 2G DDR memory and a Windows Server 2003 OS). It is seen that the memory demand and CPU time in the condition of using soft data and hard data are obviously less than those of two other conditions. Besides, the average and variance of the image simulated by using hard data and soft data are closest to those of the training image (Recall that the average and variance of the training image are 0.6121 and 0.041).

TABLE I. THE MEMORY DEMAND, CPU TIME, AVERAGE AND VARIANCE OF IMAGES SIMULATED IN THREE CONDITIONS

	using soft data and hard data	**using hard data only**	**unconditional simulation**
maximum memory(M)	425	522	643
CPU time (second)	1321	1676	1981
average	0.6330	0.6564	0.6887
variance	0.045	0.055	0.067

B. Comparison with two-point geostatistics

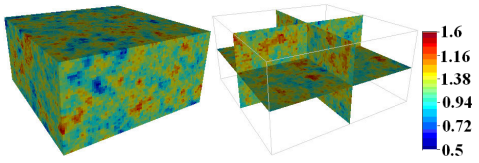
For comparison, two two-point geostatistical methods, which are one of the stochastic simulation methods called SGSIM (sequential Gaussian simulation) and SK (simple kriging), are used to simulate the images. Sample points shown in Fig. 8(a) are used as conditional data for two-point geostatistical simulation. The exterior and cross-sections (X=40, Y=40, Z=20) of simulated results using SGSIM and SK are shown in Fig. 13 and Fig. 14. The histograms of the SGSIM-simulated image and the SK-simulated image are shown in Fig. 15, respectively. It is seen that the distribution of simulated values is quite different from that of the training image. The average and variance of the simulated results using SGSIM and SK are respectively shown in TABLE II, obviously differing from those of the training image.

C. Comparison of Variogram

Variogram can reflect the relativity and variability of a spatial variable in certain directions, which is used as the evaluation method for MPS simulation. If an attribute of two images has the similar variogram curves in the same direction, then the structures of this attribute are similar in this direction; otherwise, the structures of this attribute in the two images are different [8].

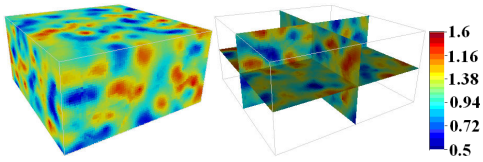
Suppose that the distances between two neighboring nodes are all one, and then the distances between two

computing nodes in the directions of X, Y and Z are respectively 79, 79 and 39. Figure 16 shows the variogram curves of the training image, the image using soft data and hard data, the image using hard data only, the image of unconditional simulation, the SGSIM-simulated image and the SK-simulated image in the directions of X, Y and Z. In Figure 16, the variogram of the training image is most similar to that of the simulated image using soft data and hard data, demonstrating that the simulated structures using soft data and hard data are closest to those of the training image.



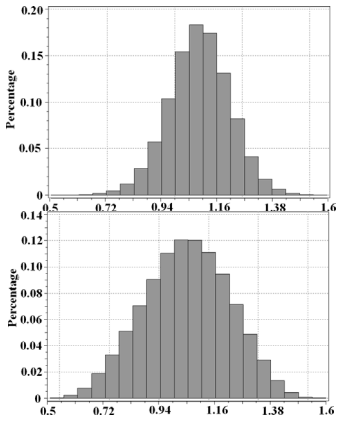
(a) (b)

Figure 13. Simulated results using SGSIM. (a) exterior; (b) cross-sections (X=40, Y=40, Z=20).



(a) (b)

Figure 14. Simulated results using SK. (a) exterior; (b) cross-sections (X=40, Y=40, Z=20).



(a) SGSIM (b) SK

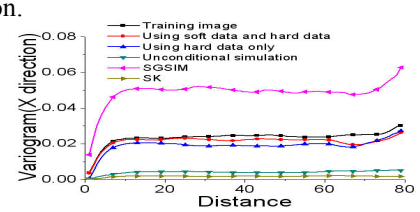
Figure 15. The histograms of the SGSIM-simulated image and the SK-simulated image.

TABLE II. THE AVERAGE AND VARIANCE OF THE SGSIM-SIMULATED IMAGE AND THE SK-SIMULATED IMAGE

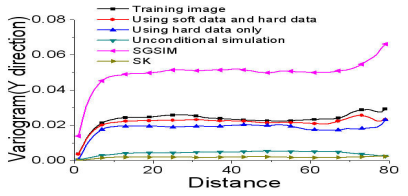
	SGSIM	SK
average	1.0632	1.0856
variance	0.056	0.064

IV. CONCLUSIONS

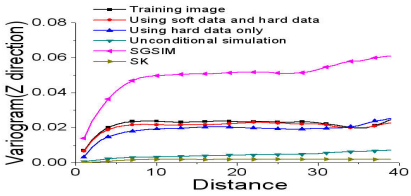
A novel interpolation method based on continuous MPS using soft data and hard data is proposed to realize the continuous simulation of unknown information. Nine filters are used to characterize the average, gradient and curvature of a pattern respectively, by which the dimensions can also be largely reduced. Soft data and hard data are both used as conditional data during simulation to improve the accuracy of information prediction. Experimental results show that the structures simulated by using soft data and hard data are most similar to those of the training image. Although the soft data and hard data are integrated, memory demand and CPU time will be less than those using hard data only and those of unconditional simulation. The experimental results also prove that the performance of our method is better than that of two-point geostatistics in predicting the unknown information.



(a) X direction



(b) Y direction



(c) Z direction

Figure 16. Variogram curves of the training image, the image using soft data and hard data, the image using hard data only, the image of unconditional simulation, the SGSIM-simulated image and the SK-simulated image.

REFERENCES

[1] T. Zhang, D. T. Lu, and D. L. Li, "Porous media reconstruction using a cross-section image and multiple-point geostatistics," Proceedings of ICACC 2009, Singapore, pp. 24-29, Jan. 2009.

[2] T. Zhang, D. T. Lu, and D. L. Li, "A statistical information reconstruction method of images based on multiple-point geostatistics integrating soft data with hard data," Proceedings of ISCSCT 2008. Shanghai, China, vol.1, pp. 573-578, Dec. 2008.

[3] D. T. Lu, T. Zhang, J. Q. Yang, D. L. Li, and X. Y. Kong, "A reconstruction method of porous media integrating soft data with hard data," Chinese Science Bulletin, vol.54, No.11, 2009, pp. 1876-1885.

[4] S. Strebelle, "Conditional simulation of complex geological structures using multiple-point statistics," Mathematical Geology, vol.34, No.1, 2002, pp. 1-21.

[5] T. F. Zhang, "Filter-based training pattern classification for spatial pattern simulation," Ph.D. dissertation, USA: Stanford University, 2006, pp. 11-26.

[6] T. F. Zhang, S. Bombarde, S. Strebelle, and E. Oatney, "3d porosity modeling of a carbonate reservoir using continuous multiple-point statistics simulation," 2005, SPE paper # 96308, pp. 1-6.

[7] J. B. Wu, "4D Seismic and Multiple-Point Pattern Data Integration Using Geostatistics," Ph.D. dissertation, USA: Stanford University, 2007, pp. 163-173.

[8] N. Remy, A. Boucher, and J. B. Wu, "Applied Geostatistics with SGeMS: A Users' Guide," Cambridge University Press, 2009, pp. 108-134