Kriging method for estimation of groundwater resources in a basin with scarce monitoring data

CHENGPENG LU[1,2], LONGCANG SHU[1,2], XUNHONG CHEN[3], YUEZAN TAO[4] & YING ZHANG[1,2]

1 *State Key Laboratory of Hydrology-water Resources and Hydraulic Engineering, Hohai University, Nanjing, 210098, China*
thebest@hhu.edu.cn

2 *College of Hydrology and Water Resource, Hohai University, Nanjing 210098, China*

3 *School of Natural Resources, University of Nebraska-Lincoln, Lincoln, Nebraska 68583-0996, USA*

4 *College of Civil Engineering and Hydraulic, Hefei University of Technology, Hefei 230009, China*

Abstract Construction of the water table map is a key step in the assessment of water resources. However, the scarcity of groundwater monitoring data in some basins remains a problem for determination of a reliable variogram model, which is the starting point for kriging interpolation. Researchers have used the secondary variable, the sampling number of which is usually much greater than that of the primary variable, in assisting the spatial interpolation of the primary variable, e.g. by the regression kriging and co-kriging methods. These methods still require a variogram model to characterize the spatial structure of the primary variable. In this study, the authors proposed an approach that derives the variogram model of the groundwater level based on the elevation of the land surface data sets. The measurements of land surface elevation are widely available to researchers, and the density of the data locations is much larger than that of groundwater monitoring records. The land surface elevation was assumed to have a linear relationship with the groundwater level. A relationship between the variogram model for the groundwater level and the variogram model for the land surface elevation were established; the variogram model for the former can be directly inferred from the variogram model of the latter. In the derivation of the groundwater level variogram, the precipitation data can also be taken into account. This approach was implemented for the Nanjing watershed, China. A variogram model of the groundwater level was obtained from the DEM data set of 1000 m × 1000 m grid spacing.

Key words kriging; regression; groundwater resources; geostatistics

INTRODUCTION

Kriging

Kriging is the most important method in geostatistics and is applied widely in many natural science fields. Kriging was originally developed in geostatistics (also known as spatial statistics) by the South African mining engineer called Krige. The mathematics was further developed by Matheron (1963). A classic geostatistics textbook is edited by Cressie (1993). More recent references are given in Martin *et al*. (2004, 2005). Searching for “kriging” via Google.com on 6 February 2009 gave 296 000 hits, illustrating the popularity of this mathematical method.

The core of kriging is the determination of the variogram. According to the general principle of kriging, the steps are: compute the experimental discrete variogram in different points, then fit the discrete experiment points with a mathematical model which is designed as the model variogram. However, it cannot be applied to all conditions. For example, when the data is distributed very sparsely while each point is independent of any other, we cannot use the curve fitting process to obtain the model variogram. In the first condition, supposing the data are correlated to each other in a limited range, generally processing is terminated because the data quantity is much less and the spatial distribution is sparse and discrete. Other methods are developed to solve this problem, and in this paper a new approach based on regression is proposed. For the last condition, the distribution can be considered as a stochastic process, where all the data are random and independent. Therefore, the data do not satisfy the basic assumption of kriging – that the different values of points within a special lag are more or less correlated, and as the lag increases, the correlation reduces. Because of the complete randomness, the deterministic prediction of the variable is unsolvable and impossible.

Copyright © 2009 IAHS Press

Objective

For the estimation of groundwater resources, especially storage, interpolation is key to calculating the quantity of resources. In this paper, a practical case study is presented to estimate the unconfined groundwater storage resources with kriging in the poorly gauged Nanjing basin in China.

The goal of this paper is to introduce kriging generally for Predictions in Ungauged Basins (PUB); the basics of kriging, and the principle of regression kriging are reviewed. Furthermore, a new kriging method based on regression, named Direct Regression Kriging (DRK), is proposed. The rest of this article is the application of DRK in PUB, that is estimation of groundwater resources with kriging, and gives the analysis of different kriging results.

METHODOLOGY

Regression-Kriging (RK) – a hybrid interpolation method

An alternative to kriging is the regression approach, which makes predictions by modelling the relationship between the target and auxiliary environmental variables at sample locations, and applying it to unvisited locations using the known value of the auxiliary variables at those locations (Odeh *et al.*, 1995; Hengl *et al.*, 2004, 2007).

Common auxiliary predictors in geosciences are land surface parameters, remote sensing images, and geological, soil and land-use maps. A common regression approach is linear multiple regression, where the prediction is again a weighted average:

$$Z(u_0) = \sum_{k=0}^{p} \beta_k \cdot q_k(u_0), \quad q_0(u_0) \equiv 1 \tag{1}$$

where $q_k(u_0)$ are the values of the auxiliary variables at the target location, β_k are the estimated regression coefficients and *p* is the number of predictors or auxiliary variables.

Regression Kriging (RK) combines these two approaches: regression is used to fit the explanatory variation and Simple Kriging (SK) with expectation 0 is used to fit the residuals, i.e. unexplained variation:

$$\begin{aligned} Z(u_0) &= m(u_0) + \delta(u_0) \\ &= \sum_{k=1}^{p} \hat{\beta}_k \cdot q_k(u_0) + \sum_{i=1}^{n} \lambda_i \cdot Z(u_i), \end{aligned} \tag{2}$$

where $m(u_0)$ is the fitted drift, $\delta(u_0)$ is the interpolated residual, β_k are estimated drift model coefficients (β_0 is the estimated drift), λ_i are kriging weights determined by the spatial dependence structure of the residual and where $\delta(u_i)$ is the residual at location u_i. The regression coefficients β_k are estimated from the sample by some fitting method, e.g. ordinary least squares.

RK has the advantage that it explicitly separates trend estimation from residual interpolation, allowing the use of arbitrarily complex forms of regression, rather than the simple linear techniques that can be used with KED (Kriging with External Drift). In addition, it allows the separate interpretation of the two interpolated components. In the paper, this regression-kriging is named indirect regression kriging (IRK) to distinguish the following proposed kriging.

Direct regression kriging (DRK)—from regression to kriging

One unique advantage of kriging is that kriging variance is readily available during the computational process, and it offers the confidence level of the interpolation (Zimmerman *et al.*, 1999). However, when the rational variogram cannot be derived, the advantage is invalid in practice. In terms of the experimental variogram, the accuracy of it is due to the quality and quantity of data. When the data is sparse in the ungauged area, a rational variogram is difficult to obtain. Therefore, the core of using kriging in PUB is how to obtain a rational variogram.

The structured variable exists and is known in the special time–space domain. The true value of structured variable can be obtained by de-noising and inversion methods. In practice, we always supposed that the geo-variant conform to the stationarity condition, $Z_c = f(x, y) = 0$, and then the geo-variants are equal to the spatial variation. So if we can determine the spatial variation precisely, the unbiased and optimal estimation can be obtained with kriging. The accuracy of the variogram reflects the density and accuracy of the actual measured data. Ignoring the data accuracy, the density of the data is the key factor whether the variogram is credible or not.

Variograms are categorized as the experimental variogram and the theory variogram. The experimental variogram is the value calculated from the sample, and this variogram is discrete and tendency. However, the theory variogram is the successive analysis expression to fit the sample scatters. The common theoretical models are a spherical model, Gaussian model and exponential model. Under the stationarity condition, the larger the data quantity, obviously the better the tendency is shown, and conversely the experimental variogram scatters distribute dispersed and irregularity, which directly cause the accuracy and reliability of the variogram. So when the scatters cannot embody the tendency, the variogram is incredible and the kriging based on this variogram is unreliable.

For the interpolation of sparse data, it is impossible to obtain more accurate values at ungauged locations using only observed data. However, plenty of research results show that the geoscience elements present good relationships. Using the regression method the unknown variant can be estimated from some related variants. But when the unknown element has strong spatial correlation, the regression method cannot reflect the spatial correlation. In the kriging method, making full use of the spatial correlation is the starting point, and the element which embodies the correlation is the variogram. Considering the strong relation of the variants, the authors derived equations to obtain the variogram of the sparse data by means of the correlated variables. Because the correlated ones are rich and obtainable, the ungauged variable can be estimated indirectly and reasonably.

Suppose that there are two variables *Z* and *X*, the variable *Z* are underestimated and the amount of data *Z* is *n*. The variable *X* is a correlated variable of *Z*, and the amount of data *X* is *m*. Comparing *m* and *n*, *m* is much larger than *n*. So there are *n* pairs of (*Z*, *X*), and *m–n* individual data *X*. Because of the scarcity and lower spatial uniformity of *Z*, it is unable to depict the spatial variability of *Z*. However, the data *X* are distributed all over, and can easily be obtained to construct the variogram exactly.

We calculated the variograms of *Z* and *X* separately, and plotted scarcity points. If the trends of two series are similar and the correlation coefficient of the two variants are large enough, we can derive the variogram of *Z* from that of *X* as follows:

By means of the mapping relationship of *Z* and *X*, the equation of variance function can be derived by the known variable:

$$Z = f(X) \Rightarrow C(Z) = f(C(X)) \tag{3}$$

where *C*()is the variance function. This function is substituted by variogram and covariance. In practice, the theory variogram of *Z* can be obtained from equation (3).

If *Z* and *X* conform to a linear relation: $Z = aX + b$, then:

$$\begin{aligned} \gamma_Z(h) &= \frac{1}{2}E\left[(Z(u+h) - Z(u))^2\right] \\ &= \frac{1}{2}E\left[(aX(u+h) - aX(u))^2\right] \\ &= \frac{1}{2}a^2E\left[(X(u+h) - X(u))^2\right] \\ &= a^2\gamma_X(h) \end{aligned} \tag{4}$$

And if *Z* and *X*, *Y* accord with multivariate linear correlation, $Z = aX + bY + c$, where *Y* is another correlated variable, then:

$$
\begin{aligned}
\gamma_Z(h) &= \frac{1}{2}E\left[(Z(u+h)-Z(u))^2\right] \\
&= \frac{1}{2}E\left[(aX(u+h)+bY(u+h)-aX(u)-bY(u))^2\right] \\
&= \frac{1}{2}E\left[(a(X(u+h)-X(u))+b(Y(u+h)-Y(u))^2\right]
\end{aligned}
\tag{5}
$$

From the definition of covariance, the equation can be written as: $C(h)=E\left[(Z(u+h)-m)(Z(u)-m)\right]$. When the regional variable conforms to second order stationarity condition, the expectation exists and is equal to a constant: $E[Z(x)]=m$ (constant) $\forall x$.

In the study area, the spatial covariances of regional variable exist and are stationary:
$C(h)=E\left[(Z(u+h)Z(u)\right]-m^2 \quad \forall x,\forall h$ when $h=0$, then $C(0)=Var[Z(u)] \quad \forall x$

The stationarity of covariance means that variance and variogram are stationary too. The followed relation is established:

$$C(h)=C(0)-\gamma(h) \tag{6}$$

In addition, the following definitions are given:

$$
\begin{aligned}
C_X(h) &= E\left[(X(u+h)-m)(X(u)-m)\right] \\
C_Y(h) &= E\left[(Y(u+h)-m)(Y(u)-m)\right] \\
E_{XY}(h) &= E\left[X(u+h)Y(u)\right] \\
E_{YX}(h) &= E\left[Y(u+h)X(u)\right]
\end{aligned}
\tag{7}
$$

The binary correlation can be converted to:

$$
\begin{aligned}
\gamma_Z(h) &= a^2(C_X(0)-C_X(h))+b^2(C_Y(0)-C_Y(h))-ab(E_{XY}(h)+E_{YX}(h))+2abE_{XY}(0) \\
&= a^2\gamma_X(h)+b^2\gamma_Y(h)-ab(E_{XY}(h)+E_{YX}(h))+2abE_{XY}(0)
\end{aligned}
\tag{8}
$$

Based on the known data, the experimental values of covariance can be calculated, and the theory covariance fitted and substituted into equation (8) to obtain the variogram of the object variable.

For the multivariable issue, the variogram of the estimated variable is combined with several correlated variables in a complex manner.

The method (DRK) comprises several main steps.

(a) Selection of correlated variable. The number of correlated variables and the significance of the variables should be determined in this step. Using the F-test and the correlation coefficient, we judge whether the auxiliary variables and the regionalized object variable are strongly correlated.

(b) Observation of the variograms. Draw the variograms of the different variables separately, and compare the variogram trend to construct a fitted variogram. The DRK is applicable to those cases where the less sparse points cannot present the trend properly.

(c) Comparison with the spatial correlation. Plot the standardized experiment variograms in one drawing and observe whether the trends of different scatters are similar with each other. If the standardized variograms show obvious differences, this kriging is unsuitable.

(d) Derivation the variogram by regression. Based on the large amount of data, the variogram, variance and other information of the auxiliary variable should be obtained first, and then in terms of equation, the unrecognized variogram of sparse data is obtained.

(e) Interpolation and extrapolation of the unknown variable with kriging.

APPLICATION OF REGRESSION KRIGING

To demonstrate the usefulness of direct regression kriging (DRK) method and the difference between DRK and indirect regression kriging (IRK), we applied it in a practical case. In the case,

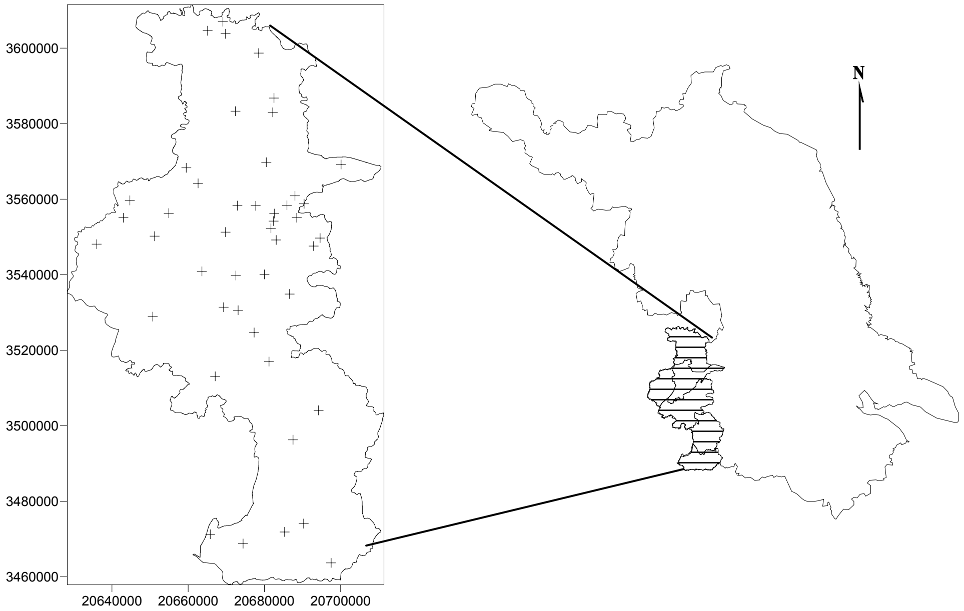

Fig. 1 Study area and the position of observed wells (+ indicates observed well).

we will use selected data points from a digital elevation model (DEM) for Nanjing city to estimate an elevation surface, and then take the raingauge data into account to obtain the water table.

Shallow groundwater plays an important role in river valleys and deltas around the world. It is a source of drinking water, supplements soil moisture necessary for agricultural production during the dry season and is the life blood of terrestrial and wetland ecosystems. In hilly terrain with permeable soils and in most humid climate zones, groundwater constitutes an important part of runoff, while groundwater depth is a key indicator of the antecedent moisture condition of a catchment. Thus, knowledge about the spatio–temporal variation of groundwater elevation and groundwater depth is important for many hydrological applications. In fact, there are limited gauges in the hilly terrain and from this sparse data it is difficult to obtain the correct spatial information. In kriging, sparse data cannot be used to form a suitable variogram. In this section, a case study about regression kriging is described.

Study area

The study region is complex, with inland highlands running northeast to southwest and wide and flat plains along the Yangzi and Chu rivers and Shijiu Lake. The area of this city is 6597 km^2, and there are 46 wells within the administrative boundary (Fig. 1). The average well density is larger than 140 km^2 per well. Nanjing city is subdivided by five catchments, which are Yangzi River basin, Chu River basin, Qinhuai River, Yiyang River basin and Tai Lake basin. According to the aquifer characteristics, the underground water can be divided into pore water, fissure water and karst water. Because the unconfined water level has a good linear relationship with elevation, and the water table is similar to the land surface, DEM derived elevation is used as a correlative variable first. Then the precipitation is taken into account to obtain the water level. Considering the bottom of the aquifer at –30 m, –20 m, –10 m, the groundwater resource can be calculated by the aquifer volume timing specific yield (*Sy*). The purpose of this study is to compare the different kriging results, and *Sy* is steady in specific aquifer, therefore, the aquifer volume is substituted for the groundwater resource in this paper.

Elevation, precipitation and water level data

Elevation data is obtained from the DEM data set of 1000 m × 1000 m grid spacing. The water level data and precipitation data employed are the average observed data in 2001. There are 36 raingauges and 46 observed wells.

Steps of DRK

First, we examine the spatial correlation structure of the data from the 46 wells. A variogram was constructed (Fig. 2, left). This variogram showed an obvious nugget effect, which means different points do not have correlation. This is clearly against the continuity and stationarity of groundwater. It is true that the sample size of 46 may be too small to develop a reasonably stable empirical variogram (Burrough & McDonnell, 1998). Second, the variogram of elevation was constructed (Fig. 2, right). The theoretical variogram selected was a spherical distribution with goodness of fit.

In addition, the standardized variogram scatters are plotted in Fig. 3. The water level points are sparsely distributed and surround the DEM points (line), so we believe the two variograms of these variables are similar and conform to the conditions of DRK.

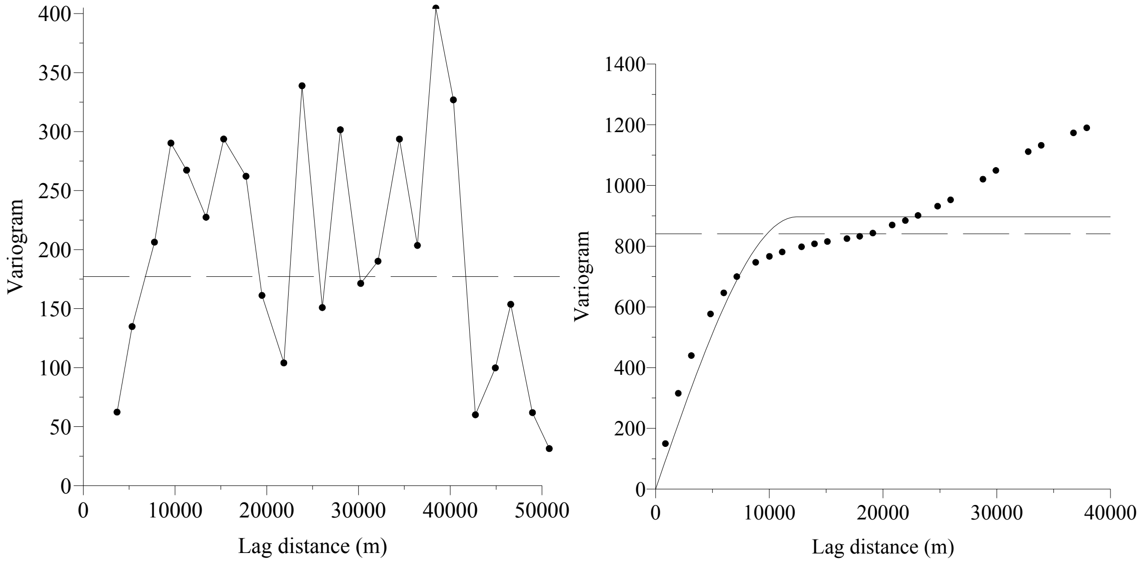

Fig. 2 The variograms of water level (left) and surface elevation (right).

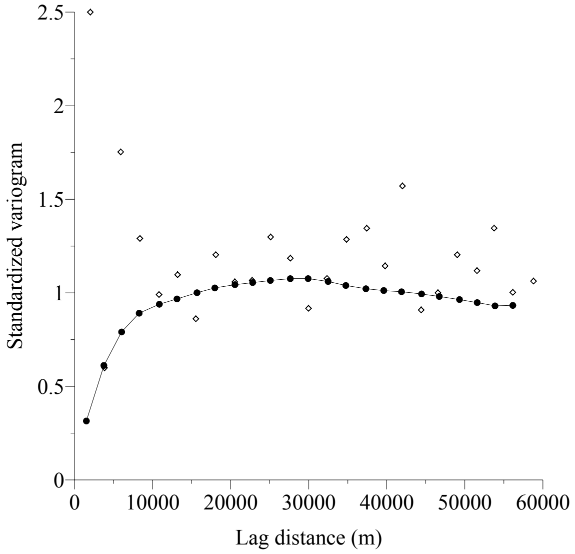

Fig. 3 The integrated variogram scatter (◊ the water level variogram, and ● surface elevation).

The linear relationship between water level and elevation was constructed as equation (9):

$$H = 0.5244Z - 1.8551 \tag{9}$$

$$F = 7.89 > 7.23 = F_{0.01}(1,45)$$

The correlation coefficient R = 0.7924, and the F variance test is satisfied. The correlation of variances is significant. Using equation (4), $\gamma_H(h) = 0.275\gamma_Z(h)$.

Based on the experiment variogram, the theoretical variogram of elevation selected was a spherical model with the following specification:

$$\gamma_Z(h) = \begin{cases} C(\frac{3}{2}\frac{h}{a} - \frac{1}{2}\frac{h^3}{a^3}) & 0 < h \le a \\ C & h > a \end{cases}$$

where the sill C is 896.7, the range a is 1000, the main anisotropic direction Az is 39.08°, and the ratio of anisotropic T is 1.857. The variogram of water level is obtained, and the spatial interpolation can be obtained with kriging.

Besides this single relationship, precipitation and terrain were taken into consideration. The binary regression equation was constructed:

$$H = 0.0176P + 0.5665Z - 1.6344 \tag{10}$$

$$F = 33.29 > 5.14 = F_{0.01}(2,43)$$

The correlation coefficient R = 0.7795 and the F variance test is satisfied. Calculation of the relevant variograms and other essential variables is shown as Table 1. To compare the results of DRK, we implement the IRK with single and binary scenarios.

Table 1 The parameters of binary linear regression kriging.

Function / variant	a	b	$\gamma_P(h)$	$\gamma_Z(h)$
Expression / value	0.0176	0.5665	$6930(1-\exp(-\frac{3h^2}{59720^2}))$	$888.1(1-\exp(-\frac{3h^2}{8928^2}))$
Function / variant	$E_{PZ}(h)$	$E_{ZP}(h)$	$E_{PZ}(0)/E_{ZP}(0)$	
Expression / value	$24686-0.1953h$	$24409-0.0033h$	24547.5	

RESULTS

For checking out the validation of this method, cross validation is used in this paper. The results of cross-validation are showed in Table 2. And we draw the contours of water level which use the single DRK and binary DRK in Fig. 4.

Comparing the results, the errors of DRK are not larger than OK and IRK. We can conclude that DRK is an efficient method for spatial interpolation.

In different conditions, we calculated the groundwater resources that are listed in Table 3. In terms of IRK, the trends obtained by different variants regression can lead to deviation of groundwater storage from 9.5% to 17.3%, for the single linear regression and binary linear regression methods, when the bottom elevation of the unconfined aquifer ranges from –30 m to –10 m.

Table 2 Cross-validation of interpolation errors, using different kriging methods.

.	OK1	OK2	Single DRK	Binary DRK	Single IRK	Binary IRK
Mean	–0.1212	–2.993E-15	1.023E-15	–0.1234	–0.3378	–0.1806
Root Mean Square	14.48	13.08	13.08	14.36	7.272	8.366

OK1: in this method the variogram is derived by linear fit; OK2: in this method the variogram is directly

equal to the one of elevation.

Table 3 Calculated groundwater resources of OK, DRK and IRK in different scenarios.

Bottom level	OK1	OK2	Single DRK	Binary DRK	Single IRK	Binary IRK	Average value	E_A	E_M	E_B
-30	2.75	2.92	2.92	2.76	2.66	2.94	2.82	9.87%	10.46%	5.79%
-20	2.09	2.26	2.26	2.10	2.00	2.28	2.16	12.88%	13.98%	7.60%
-10	1.43	1.60	1.60	1.44	1.34	1.61	1.50	18.53%	20.90%	11.08%

Unit of the groundwater resources: 10^{11} m^3.

E_A: maximum error compared with average value; E_M: maximum error between each data series; E_B: error between single DRK and binary DRK.

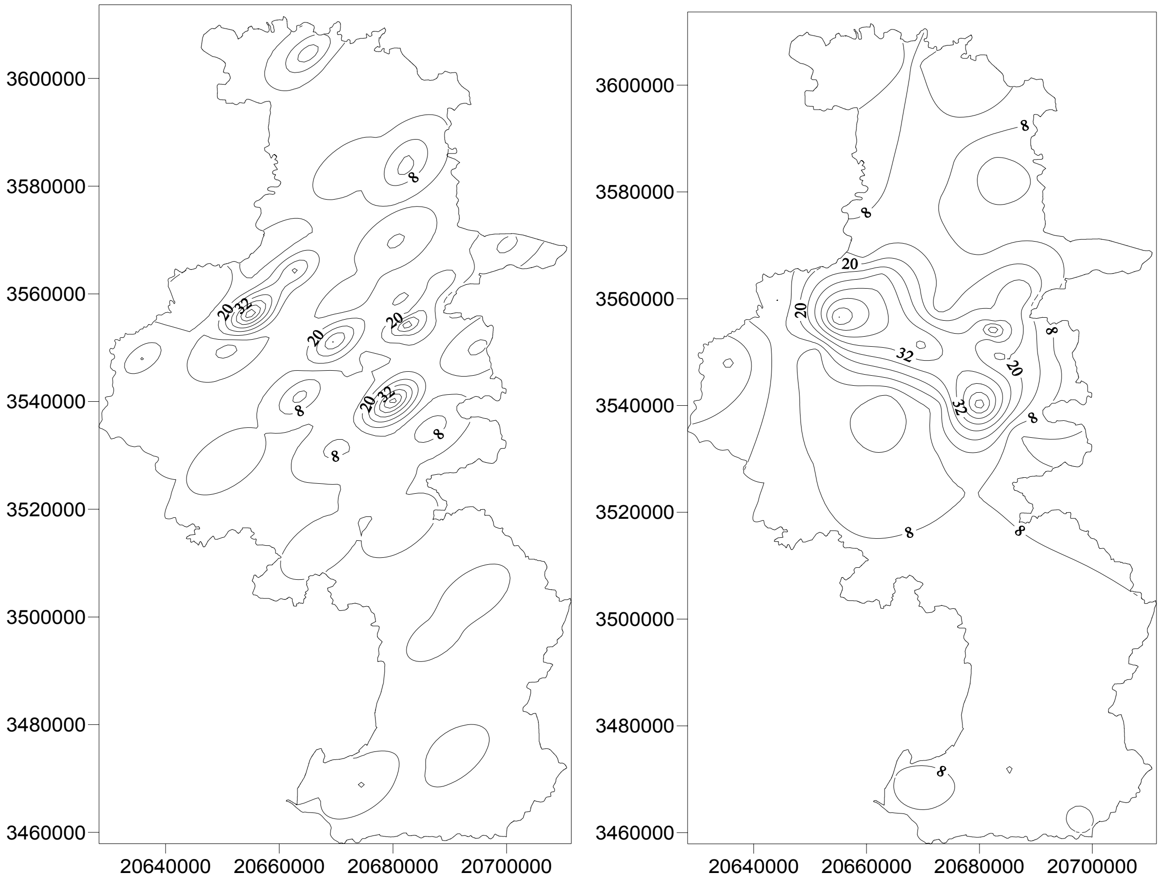

Fig. 4 The contours of water level using DRK methods (the single DRK is on the left, the binary DRK is on the right).

Comparing the results when the variograms are determined by the elevation directly, or indirectly, the interpolation values are the same, and the difference lies in the estimated variance. With the variance of regression coefficient of elevation and water level, the estimated variance is changed. But when the precipitation is taken into account, the estimated resource is less than the former, and according to the increased bottom elevation of the aquifer, the smaller proportion that the reduced volume occupied ranged from 5.79% to 11.08%.

In the results of all the kriging method we implemented in this study, the resources estimated by the binary IRK are largest, and the results of single IRK are the least. We calculated the average value of al the results for different bottom levels, and the differences between single IRK and binary IRK account for 9.87%, 12.88% and 18.53% of the average values. When the denominators, perhaps the true values, become the results of single IRK, the ratios are larger and can reach 10.46%, 13.98% and 20.90%, respectively.

CONCLUSIONS

From the results of the analysis above, we find that the different kriging method can produce distinct groundwater resources; the maximum error surpass 20% when the bottom elevation is –10 m. At different bottom elevation levels, the estimation of OK is less then the single DRK and binary DRK, and it lies between the single IRK and binary IRK. It shows that the estimation from the method proposed in the paper can generate optimistic results. So for positive researchers, this method can be applied deliberately.

Based on the regression equation of primary variant and auxiliary variant, the relationship between primary variant and auxiliary variant are constructed. Because of the data scarcity of the primary data set, this indirect method of obtaining the variogram is more rational than the original method.

The IRK method is a classic tool which was implemented to estimate the unknown variant by the auxiliary variant. In this study, we found that the chosen correlated variant can make great uncertainty in the results. Taking the average value as the criterion, the relative errors of different IRKs come to 9.9% and 18.5%. If the least one is taken as the criterion, the pessimistic errors can even reach 10.5% and 20.1%. Compared with the IRK method, the results of DRK lie between the IRKs, are closer to expectations and have fewer fluctuations. The pessimistic errors are only 5.79% and 11.08%, approximately half of the IRKs.

Acknowledgements This research was supported from the 111 Project under Grant B08048, Ministry of Education and State Administration of Foreign Experts Affairs, China, and was partially funded by National Basic Research Program of China entitled "The fundamental theoretical problems concerning the processes of rocky desertification in the mountainous areas of Southwest China and adaptively ecological rehabilitation (2006CB403200)".

REFERENCES

Burrough, P. A. & McDonnell, R. A. (1998) *Principles of Geographical Information Systems*. Oxford University Press, Oxford, UK.

Cressie, N. A. C. (1993) *Statistics for Spatial Data* (revised edn). Wiley, New York, USA.

Hengl, T., Heuvelink, G. & Stein, A. (2004) A generic framework for spatial prediction of soil variables based on regression kriging. *Geoderma* **122**(1–2), 75–93.

Hengl, T., Gerard B. M. H. & David G. R. (2007) About regression-kriging: from equations to case studies. *Computers & Geosciences* **33**, 1301–1315.

Martin, J. D. & Simpson, T. W. (2004) On using Kriging models as probabilistic models in design. *SAE Transactions J. Materials & Manufacturing* **5**, 129–139.

Martin, J. D. & Simpson, T. W. (2005) Use of Kriging models to approximate deterministic computer models. *AIAA J.* **43**(4), 853–863.

Matheron, G. (1963) Principles of geostatistics. *Economic Geology* **58**(8), 1246–1266.

Odeh, I., McBratney, A. & Chittleborough, D. (1995) Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma* **67**(3–4), 215–226.