

The Comparative Analysis of Spatial Structure of Ji Wheat 22 Yield Based on Different Stochastic Samplings

Yujian Yang* and Xueqin Tong

S & T Information Engineering Technology Center of Shandong Academy of Agricultural Science, Information center of agronomy College of Shandong University Jinan 250100, P.R. China
yyjtskhk@gmail.com

Abstract. The spatial variation of crop yield was mainly caused by the complicated factors, such as soil and land factors. Based on sound theoretical reasoning, the paper performed an exploratory investigation of the relationship between the spatial distribution of yield measurement about 849 samplings on Thiessen polygon. Reasonable sample size or sampling density is an assurance of yield estimation surface. To investigate the spatial structure of Ji wheat 22 yield, we developed the comparison analysis of surface interpolation precision by stochastic samplings of 568, 425, 212 and 144 from a total of 849 samplings in zoning area, the other emphasis of the study was explored the influence of the distribution change of yield with the threshold of samplings numbers for the higher estimation precision. We first detected the presence of global spatial autocorrelation characteristics about the distribution of the Thiessen polygon region generated yield measurement samplings, then with the support of IDW interpolation, the higher precision surface was analyzed and obtained by the comparison of 5 samplings mode. High yield Thiessen polygon region of Ji wheat 22 tended to be clustered close to the high yield polygon region, the low yield Thiessen polygon region tended to be clustered close to the low yield polygon region. It is important to estimate the yield distribution for sampling density, samplings method, interpolation methods and sample size, we suggested that a suitable sample size of about 200 samplings should be used for Ji wheat 22 yield developed from the sampling zoning to acquire the higher surface estimation precision, beyond or the samplings number, the estimation precision presented the fluctuation characteristics.

Keywords: Ji wheat 22, Thiessen polygon, Autocorrelation, IDW, Yield.

1 Introduction

Many studies have been conducted in recent years about all kinds of regional variance. The sampling intensity and the position of sample points are critical to obtaining good predictions and for constructing maps from sample data, provide the context for the development of geostatistics, at the same time, Webster also pointed out that it

* Corresponding author, Address: S&T Information Engineering Research Center, Number 202 Gongye North Road, Licheng District of Jinan, 250100, Shandong Province, P. R. China.

is essential to develop the consistence study about the sampling density and the scale of spatial variation for spatial prediction map[1]. For three elements of spatial sampling scale, including sampling extent, sampling spacing and sampling support, different sampling numbers create different influence of estimation precision. About the analogy studies, in Nanjing soil institute of CAS, correlated researcher developed the effects of sample size on spatial characterization of soil fertility properties in an agricultural area the Yangtze River Delta Region, results showed that sample size of 250 is suitable for soils to develop the spatial variability of soil fertilizer indices on country level. In China Agriculture University, Zhang fusuo also explored how the change of sampling numbers have impact on the spatial variability of soil nutrients, the case of Wuhu city, in Anhui province[2,3].

Under soil-crop system framework, crop yield has the more complicated characteristics than the soil characteristics. The spatial variability of yield on the regional scale is the hotspot of soil-crop system study and integration demonstration. Though RS technology became the better data of crop yield estimation, application of a kind of method, theoretically, interpolation methods are also essential for the surface formation of crop yield, correlated study about crop growth information and crop yield of samplings is rarely according to the study literatures using interpolation methods. Crop information is a close relationship to the climatic conditions, soil characteristics and agriculture cultivation status. Soil characteristics has the spatial structure, so crop information including crop yield has the spatial structure. An important objective was developed to explore the spatial structure of wheat yield in the paper, taking the example of Ji wheat 22 cultivar in demonstration base of precision agriculture in Shandong province. Obviously, the spatial variation of Ji wheat 22 yield was caused by complicated factors, the spatial variability of soil and land factors are only a kind of parts, other factors, the spatial variability of weather factors and pests and diseases, the temporal variability, the prediction variation, and so on. A detailed description of crop yield can be found in the relevant literatures(Blackmore,1996)[4,5].

Wollenhaupt explored the comparative study of soil characteristics with the support of Kriging and Inverse Distance Weighted method, the study results that the interval of 31.8m with the gridding samplings in the region is suitable, the higher interpolation precision for IDW than Kriging. We developed the spatial estimation comparison study of yield was developed on the precision scale, with the support of Kriging and Inverse Distance Weighted (IDW) method, referred to in document[4], the study indicated that IDW method is more accurate on basis of the sampling sites data than Kriging method, under less than 40m interval between sites for Ji wheat 22 yield, though Kriging method has merits, it not only provides a means of predicting values on a surface given sample data, but also confidence intervals can be attached to each prediction using the prediction standard errors, the other advantage of Kriging is estimating weights for locations based not only on the distance between separated locations but also on the global semivariogram model fitted to the collected data. The results also showed that the estimation precision depended on the threshold of sampling numbers, if the numbers are less than the threshold, the sampling density is in contact with the spatial estimation precision, if the numbers are more than the threshold, the density of the sampling numbers is improved, but the estimation results are not proportional to the density, the detailed contents referred to document [6].

In the study, we discussed the spatial structure of Ji wheat 22 yield based on IDW, firstly, we developed the spatial autocorrelation characteristics based on 849 yield measurement samplings on Thiessen polygon region, then we analyzed the spatial structure of Ji wheat 22 yield in detail, based on different samplings methods, stochastic samplings of 568, 425, 212 and 144 from a total of 849 samplings, which not only the study extension of the spatial structure of Ji wheat 22 yield, but also the typical cases of spatial clustering of the crop yield on the regional scale. Exploring the reasonable samplings numbers on precision scale, and inducing the threshold of sampling density and spatial variability scale, the paper ultimately developed the interpolated map precision of different samplings under the random samplings mode, the optimization samplings numbers for the higher continuous surface.

2 Material and Methodology

2.1 Study Area and Ji Wheat 22 Cultivar

Winter wheat (*Triticum aestivum* L.) is one of the most important crops in China, is planted in about in north-China plain, and it has the highest yield and the largest consumption, its cultivation area is about 22% of the total crop area, and its yield is 20% of the total crop production. Especially, the 65.4% of the total wheat cultivation area, the 75.5% of the crop yield, respectively. Ji wheat 22 was intensively planted in Lingxian country, which is located in the north china, is characterized by a warm temperate monsoon climate, the precipitation and temperature presents the space-time variability, an average annual precipitation of 900mm in the south of sampling area and 480mm in the north of sampling area, the 80% of annual precipitation distributed from June to September and the soil is formed from the sediments of the Yellow River, so the irrigation is the crucial factor for agriculture development. Typically, agriculture in the area is intensified by a double cropping system with a high-yield cultivar and high fertilizer and water inputs. Winter wheat, Ji wheat 22 and summer maize is a typical double cropping system a year in the study area, and the ratio of the total yield of two kinds of crops is about 1/5 in China.

As super high yield wheat cultivar by cross breeding, Ji wheat 22 cultivated by Shandong Academy of Agricultural Science was absolutely planted predominance in the study zoning area, located in Lingxian country, precision demonstration base and the commodity demonstration base of China.

2.2 Experimental Sites and Data Acquisition

The Ji wheat 22 yield obtained from the actual yield of the observation area, with the support of DGPS, the area covered the zoning area about 4 kilometers \times 2 kilometers, 1 square meter area was selected around DGPS location site, the wheat was bounded and weighted using the suitable balance, yield information involved in plant height, the total spike numbers of Ji wheat 22, spike numbers per 666.7 m², grain number per spike, 1000-grain weight, yield per 666.7 m² and biological yield, the above contents were measured and described in the yield measurement process, the corresponding

per 666.7 m² yield of Ji wheat 22 was transformed in yield per hectare, and a total of 849 yield measurement sampling sites of Ji wheat 22 were obtained.

Sampling plans has a large effect on the continuous surface formation, stochastic sampling does not ensure even coverage of the area to be sampled, so gridding sampling combined with stochastic samplings in the study, a total of 849 sampling sites of Ji wheat 22 yield measurement, stochastic samplings of 568, 425, 212 and 144, according to the ratio of 2/3,1/2,1/4,1/6 from a total of 849 samplings, as illustrated in Fig. 1.

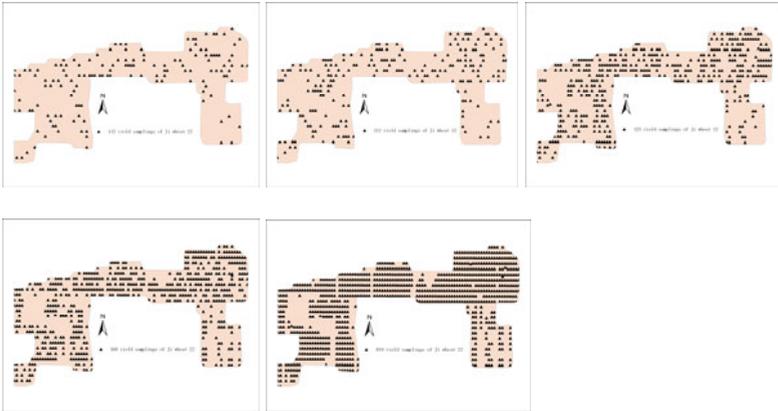


Fig. 1. Distribution of yield measurement sites on precision scale on 5 kinds of sample modes

The statistical characteristics was analyzed, results showed the mean value of wheat yield is 6716.3kg/hectare, the maximum value is 10364.1kg/hectare, the minimum value is 1828.0kg/hectare, the range of yield was 1828.0kg/hectare-10364.1kg/hectare. The above description explained the characteristics of wheat yield rather than the regional characteristics of the whole sampling area. Coefficient of variation of was 0.2 which suggested wheat yield belonged to the medium variation. Consequently, it is necessary to apply spatial interpolation methods together with GIS to solve this problem and reveal the spatial structural characteristics.

2.3 Inverse Distance Weighting Method

Inverse Distance Weighting (IDW) are interpolation techniques in which interpolated estimates are made based on values at nearby locations weighted only by distance from the interpolation location. IDW make assumptions about spatial relationships except the basic assumption that nearby points ought to be more closely related than distant points to the value at the interpolate location. The formula used for Inverse Distance Weighting is:

$$Z_j = \frac{\sum \left[\frac{z_i}{(h_{ij} + s)^p} \right]}{\sum \left[1/(h_{ij} + s)^p \right]} \tag{1}$$

Where Z_j : estimated value for location j ; z_i : measured sample value at point i ; h_{ij} : distance between Z_j and z_i ; s : smoothing factor; p : weighting power. The weighting power p defines the rate at which weights fall off with h_{ij} , the distance between the interpolated and sample locations. A value of 1-5 is typical, the smoothing factor s reduces the likelihood that any one sample value will overly influence an estimated value for a given interpolation location. IDW is an exact interpolator, so where an interpolation location j coincides with a sample location i , $Z_j = z_i$ and a sharp "peak" or "valley" may result; setting $s > 0$ reduces this peaking effect when it occurs[7].

IDW emphasized that attribute values close in distance terms tends to be similar but that the similarity weakens as distance separation increases, which is consistent for the Tobler Law by estimating unknown measurements as weighted averages over the known measurements, "Everything is related to everything else, but near things are more related than distant things"[8]. IDW give differential weights to observations based on their proximity to the missing value, it is the nearest sites that should be given most weight in any imputation, so the method estimators depend only on geometric relationships between the sites and the estimates they yield are dependent on neighboring size and how the weights are specified and the method do not provide an estimate of the possible error associated with the imputation.

3 Results and Analysis

3.1 Spatial Autocorrelation Characteristics

The presence of spatial dependence means that values for the same attribute measured at locations that are near to one another tend to be similar, and tend to be more similar than values separated by larger distances. We construct the Thiessen polygons to develop spatial autocorrelation characteristics of 849 yield measurement points. Thiessen polygons can be used to apportion a point spatial file into regions known as Thiessen or Voronoi polygons, and each region contains only one input spatial file point, all points are triangulated into a triangulated irregular network (TIN) that meets the Delaunay criterion. From the principle, the perpendicular bisectors for each triangle edge are generated, forming the edges of the Thiessen polygons, the locations at which the bisectors intersect determine the locations of the Thiessen polygon vertices. On the basis of Thiessen polygons construction process, each region has the unique property that any location within a region is closer to the region's point than to the point of any other region. The Thiessen polygon of 849 yield measurement sites generated was displayed in Fig. 2.

The spatial autocorrelation function, examines the correlation of a random process with itself in space, not only showing the correlation between a series of points or a map and itself for different shifts in space or time, but visualizing the spatial variability of the phenomena under study. In general, large numbers of pairs of points that are close to

each other on average have a lower variance than pairs of points at larger separation, the autocorrelation quantifies this relationship and allows gaining insight into the spatial behavior of the phenomenon under study clarify the spatial interactions of model, so we need to understand the spatial connectivity between each country unit.

The spatial weight matrix is the fundamental tool used to represent the spatial connectivity between regions units, based on constructing a K -nearest neighbouring weights file. For the spatial weight matrix $W(k)$ which we use in this study is based on the K -nearest neighbouring function rules, K -function is a second-moment measure as it is closely related to the second-order intensity of a stationary isotropic point process, it captures the spatial dependence between different country regions of the point process, in a short word, the general form of the k -nearest neighbouring weight matrix $W(k)$ is defined as the formular (2):

$$\begin{aligned}
 w_{ij}(k) &= 0 \text{ if } i = j \\
 w_{ij}(k) &= 1 \text{ if } d_{ij} \leq d_i(k) \text{ and } w_{ij}(k) = w_{ij}(k) / \sum_j w_{ij}(k) \quad (2) \\
 w_{ij}(k) &= 0 \text{ if } d_{ij} > d_i(k)
 \end{aligned}$$

Where $d_i(k)$ is a critical cut-off distance defined for each region i . More precisely, $d_i(k)$ is the k^{th} order smallest distance between regions i and j such that each region i has exactly k neighbors. We use the K -nearest neighbor weight matrix and we check for robustness in the study[9].

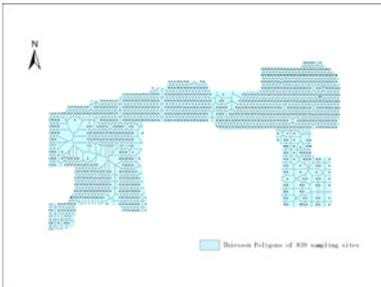


Fig. 2. Thiessen polygon of 849 yield measurement sites (left)

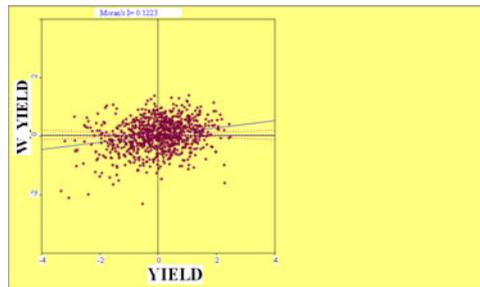


Fig. 3. Moran scatterplot of the yield measurement in 2010 (right)

The Moran's measurements which indicated spatial association of the Ji wheat 22 yield was positively spatially autocorrelated (Moran's I value is equal to 0.1223, mean value in theory is equal to 0.0004, the experience mean value is equal to -0.0012, the experience standard deviation value is equal to 0.0239) with 499 times Monte-Carlo tests, and the statistics are significant with $p=0.002$ for 849 sampling sites. Moran's I coefficient referred to Fig. 3. Results indicated that yield of Ji wheat 22 cultivar presented the positive autocorrelation, high yield Thiessen polygon region

of the cultivar tended to be clustered close to the high yield polygon region, the low yield Thiessen polygon region tended to be clustered close to the low yield polygon region. According to the Thiessen polygon of four quadrants of Moran scatter plots, from univariate local spatial autocorrelation results of local Moran's I computerization, 98 Thiessen polygon regions which presented the autocorrelation is significant characteristics in confidence interval of 95%.

3.2 Spatial Structure Analysis of Wheat Yield

Spatial autocorrelation of yield measurement mapped the spatial clustering, to detect spatial structures of distributional yield samplings, the continuous surface results should select interpolation method to solve. Kriging is estimating weights for locations based not only on the distance between separated locations but also on the global semivariogram model fitted to the data. But Estimation precision of IDW is higher estimation precision than Kriging method in author's other paper study[6]. IDW are affected by clusters in the data points, the autocorrelation characteristics of yield polygon analyzed the outlier tendency, satisfied the outlier stationary condition for IDW interpolated map. We obtained and developed the comparison analysis of surface interpolation precision by stochastic samplings of 568, 425, 212 and 144 from a total of 849 samplings, referred to Fig.4. On the whole, with the samplings increasing, from 144, 212 to 425,568,849, the interpolation surface presented the corresponding changes, the interpolation surface of wheat yield showed that the lower yield located in the east of the middle part, the higher yield located in the south of the middle part, the total distribution of wheat yield, the groovy regions, wheat yield is higher in the west of the region than the east of the region. IDW estimation of 144 samplings of yield measurement presented the fairly thick distribution. However, the IDW estimation of 212 yield measurement has the relatively thinner distribution and the better optimization pattern, the different block shape, which explained the crucial points of estimation precision caused by the samplings number change. From the interpolation surface of 144 samplings, of the estimation distribution, which reflected the high yield or low yield in different areas, indicated that the change of samplings numbers has the key impact on the estimation precision. Maps generally provide some interesting hotspot, the obvious high value area of block shape located in the north of the west part in the zoning area, centered in the high yield area, the continued low value distributed extending the south part. For the interpolation surface of 212 samplings, the high value area still is the same area of surface of 144 samplings, with a difference, the block scope is smaller than the surface of 144 samplings, the secondary high value area is clear. The new content was described as the unobvious crossing area of "high yield and low yield" for 212 samplings estimation, or no sudden changes, but the two high yield area of 144 samplings surface, and the gradient changes from the high yield area, the low yield area is located in the south-east part of estimation area, the relatively moderate estimation distribution was presented, which showed the non-sensitive estimation of IDW method for the high and low observed points, for continuous surface based on different random samplings(144,212,425,568), there existed in relatively moderate distribution mode, no sudden changes situation[10].

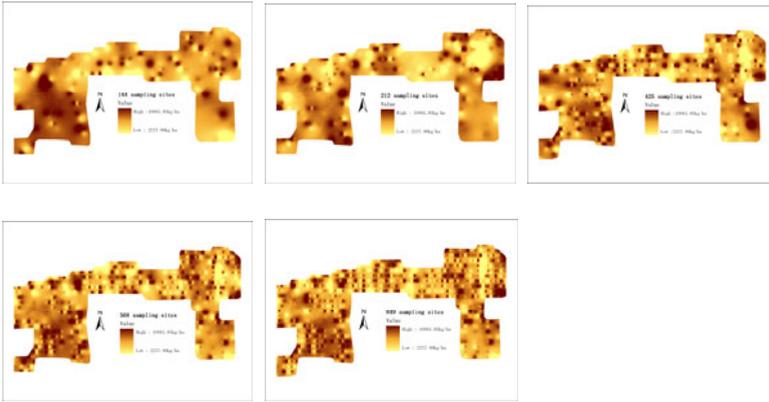


Fig. 4. Spatial distribution map of yield measurement sites based on 5 kind of modes

The cross-validation method which mainly described the sampling estimation required from the other samplings interpolation results was used to the comparative study of the prediction value and the actual value, the computerization of MSE was created by the way in the study, the smaller MSE value, the better prediction of the samplings estimation, with a difference, The correlation of prediction value and observed value is an important index, the larger, the better[3]. Based on the IDW interpolation method, the MSE of different samplings (144,212,425,568) interpolation results has the small difference. Moreover, the analysis of the Pearson's correlation coefficient between predicted value and observed value of different samplings was computerized by the R software tool in the study, the results indicated that the value is equal to -0.051, 0.211, 0.01 and 0.039 for 144 measurement samplings, 212 measurement samplings, 425 measurement samplings and 568 measurement samplings of Ji wheat 22 yield in confidence interval of 99%, respectively. According to this analysis, 0.211 was regarded as the better correlation characteristics in the study, which is regarded as the better prediction precision of 212 samplings, or the number of samplings is probably the threshold of the low prediction value and the high prediction value, means that the better significance characteristics over the threshold. But the value of Pearson's correlation coefficient is equal to 0.113 of the prediction value and observed value on basis of 849 samplings measurement of Ji wheat 22 in the study, showed the fluctuation was embodied by the estimation precision with the measurement yield samplings increasing. The results of correlation coefficient of prediction value and observed value about the different measurement numbers illustrated that the non-linear relationship is between the estimation precision and samplings numbers, and the threshold of sampling numbers is also the best estimation precision for the prediction areas, the more samplings cannot indicate the higher estimation for the special sampling regions, the suitable, optimization sampling plans and numbers is the only law to the best precision of the estimation precision.

Theoretically, the more samplings number, the better estimation precision, the samplings cost constrained the sampling numbers, so the optimization sampling numbers attracted the attention of researchers. Different random samplings

(144,212,425,568,849) predicted the yield surface distribution, with the increasing sampling numbers, the larger of the correlation of prediction value and observation value, but the threshold existence between sampling numbers and estimation precision, beyond threshold of sampling numbers, the increasing scope decreased, below threshold of sampling numbers, the correlation change scope increased. As the above mentioned, IDW interpolation results indicated that if the sampling numbers are less than the threshold, the sampling density is in contact with the spatial estimation precision. If the numbers are more than the threshold, the density of the sampling numbers is improved, but the estimation results are not proportional to the density. Considering the estimation precision and cost at the fields, the suitable yield measurement sampling numbers should be 200, or is regarded as the threshold, which has the higher estimation precision of yield measurement samplings in zoning area[1,10].

4 Conclusions and Discussion

The study revealed the presence of legitimate structures in the yield data in zoning area, which involved in the discrete valued how extended the continuous surface. The principal focus of this work was the comparative analysis and threshold of samplings data interpolated continuous surface for yield measurement, and the paper did not deal with otherwise important issues. During the interpolation, the method which is “best” depends on the yardstick we choose, using all the nearby sites and accounting for the possibility of outlier is the better selection, IDW estimators depend only on geometric relationships between the sites and the estimates yield dependent on neighboring size and how the weights are specified and the method, which improved the outliers[10].

The sample data between the sample data points and the locations ignored the spatial correlation or dependence meaning the large prediction errors[11]. Spatial autocorrelation of Ji wheat 22 yield solved the problem, results showed yield of Ji wheat 22 cultivar presented the positive autocorrelation, high yield thiessen polygon region of the cultivar tended to be clustered close to the high yield polygon region, the low yield thiessen polygon region tended to be clustered close to the low yield polygon region. 98 thiessen polygon regions which presented the autocorrelation are significant aggregation characteristics in confidence interval of 95%.

From the theory and method, the interpolation surface was discussed from different samplings(144,212,425,568,849) in the study, considering the estimation precision and cost at the fields, the higher precision of IDW interpolation comparison results of Ji wheat 22 yield, the suitable samplings number of 200 in the zoning area about 4 kilometers \times 2 kilometers, or the threshold of sampling numbers constrained the estimation precision, over 200, the higher estimation precision is not presented in the interpolation surface, the fluctuation precision changes, which probably is a close relation of requirements any method for estimating or predicting values on a spatially continuous surface, such as data structure, sampling density, experimental area shape.

Choosing models and fitting them to data remain among the most controversial topics in creating the continuous surface, Bayesian model should be probably better solved the issues, hierarchical bayesian model and bayesian maximum entropy

(BME). Bayesian method makes a decision according to the posterior probability distributions combination with the probability density function and prior probability distributions, the method has merits which taken into account the error of sample estimation and prior probability and knowledge. The estimation of Bayesian maximum entropy (BME) has a set of criteria and the relative theoretical merits, spatial-temporal covariance models available was taken into account, such as covariance variogram and semivariogram, ordinary and generalized, separable and non-separable, at the same time, the spatiotemporal structure of the yield distribution combination with the temporal variation was adequately represented[13,14].

Acknowledgements. This work was supported by Innovation Fund Projects of Shandong Academy of Agricultural Science(SAAS) (2006YCX035), we thank the relevant researcher in SAAS for their constructive comments, such as professor Kong Ling`an in crop research institute of SAAS, We also wish to thank the relative researchers for their support of S&T Information Engineering Research Center of SAAS (P. R. China). Especially, we also wish to thank Dr. Shuyun, Liu; Fengyun, Wang; Wenjie, Feng for the data acquisition.

References

1. Webster, R., Oliver, M.A.: *Geostatistics for Environmental Scientists*. John Wiley, Chichester (2001)
2. Wang, Z.-G., Zhao, Y.-C., Biao, H., Jeremy, D., Sun, W.-X.: Effects of Sample Size on Spatial Characterization of Soil Fertility Properties in an Agricultural Area of the Yangtze River Delta Region, China, *Soils*. 42(3), 421–428 (2010)
3. Sun, Y., Wu, C., Zhu, K., Cui, Z., Chen, X., Zhang, F.: Influence of interpolation method and sampling number on spatial prediction precision of soil P. *Chinese Journal of Applied Ecology* 20(3), 673–678 (2009)
4. Blackmore, S.: Precision farming, an overview. *Agriculture Engineering*, 85–87 (Autumn, 1997)
5. Blackmore, S., Godwin, R.J., Taylor, J.C., et al.: Understanding variability in Four Fields in the United Kingdom. In: *Processing of the 4th International Conference on Precision Agriculture*, Paul, July 19-22, pp. 3–18 (1998)
6. Yang, Y., Liu, S., Feng, W., Shang, M.: Comparison Study on the Spatial Estimation of Ji Wheat 22 Yield on the Precision Scale. In: *2011 3rd International Conference on E-business and Information System Security Proceedings*, IEEE Catalog Number:CFP1129F-PRT, Special Session Forestry&Agriculture Engineering and Computer Science, Produced, pp. 576–579. IEEE eXpress Conference Publishing (2011)
7. Schabenberger, Oliver, Gotway, Carol, A.: *Statistical Methods for Spatial Data Analysis* Texts in Statistical Science. CRC Press (2005)
8. Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46(2), 234–244 (1970)
9. Russo, G., Reggiani, A., Nijkamp, P.: Spatial activity and labour market patterns: A connectivity analysis of commuting flows in Germany. *Ann. Reg. Sci.* 41, 789–811 (2007)

10. Haining, R.: *Spatial data analysis: theory and practice*. Cambridge University Press (2004)
11. Yang, Y., Yang, J., Li, S., et al.: Comparison of spatial interpolation methods for maize growth period. *Transactions of the CSAE* 25(9), 163–167 (2009) (in Chinese)
12. Yang, Y., Zhu, J., Zhao, C., Liu, S., Tong, X.: The spatial continuity study of NDVI based on Kriging and BPNN algorithm. *Mathematical and Computer Modelling* 54, 1138–1144 (2011)
13. Yu, H.-L., Kolovos, A., Christakos, G.: Steve Warmerdam and Boris Dev. Interactive spatiotemporal modelling of health systems: the SEKS–GUI framework. *Stoch Environ Res Risk Assess* (2007)
14. Wang, et al.: Hand, foot and mouth disease: spatiotemporal transmission and climate. *International Journal of Health Geographics* 10, 25 (2011)