Analysis of variograms with various sample sizes from a multispectral image

Huihui Zhang[1], Yubin Lan[2], Ronald E. Lacey[1], Yanbo Huang[3], W. Clint Hoffmann[2], D. Martin[2], G. C. Bora[4]

(1. *Department of Biological and Agricultural Engineering, Texas A&M University, College Station, Texas, USA*; 2. *USDA-ARS*, 2771 *F&B Road, College Station, Texas, USA*; 3. *USDA-ARS*, 141 *Experiment Station Road, Stoneville, Mississippi, USA*; 4. *Department of Agriculture and Biosystems Engineering North Dakota State University, Fargo, North Dakota, USA*)

Abstract: Variogram plays a crucial role in remote sensing application and geostatistics. It is very important to estimate variogram reliably from sufficient data. In this study, the analysis of variograms computed on various sample sizes of remotely sensed data was conducted. A 100×100 - pixel subset was chosen randomly from an aerial multispectral image which contains three wavebands, Green, Red and near-infrared (NIR). Green, Red, NIR and Normalized Difference Vegetation Index (NDVI) datasets were imported into R software for spatial analysis. Variograms of these four full image datasets and sub-samples with simple random sampling method were investigated. In this case, half size of the subset image data was enough to reliably estimate the variograms for NIR and Red wavebands. To map the variation on NDVI within the weed field, ground sampling interval should be smaller than 12 m. The information will be particularly important for Kriging and also give a good guide of field sampling on the weed field in the future study.

Keywords: variogram, multispectral image, geostatistics

DOI: 10.3965/j.issn.1934-6344.2009.04.062-069

Citation: Zhang Huihui, Lan Yubin, Lacey Ronald E, Huang Yanbo, Hoffmann W Clint, Martin D, Bora G C. Analysis of variograms with various sample sizes from a multispectral image. Int J Agric & Biol Eng, 2009; 2(4): 62－69.

1 Introduction

The techniques of spatial statistics were first developed and formalized in the 1950s. Recently, with the development of GIS, spatial statistical techniques have drawn considerable attention and have been widely applied in spatial data modeling and analysis for natural sciences such as geophysics, biology and agriculture. There are numerous studies demonstrating the benefits of the spatial analysis techniques to agricultural management[1-3]. Geostatistics is a part of the spatial statistics. Geostatistical analysis consists of computing

Received date: 2009-02-18 **Accepted date:** 2009-11-18

Biographies: Huihui Zhang, Ph. D. Candidate, Department of Biological and Agricultural Engineering, Texas A&M University College Station, TX 77843, Tel: 979-260-3750. Email: zhh815@ neo.tamu.edu; **Yubin Lan**, Ph.D., Agricultural Engineer, USDA-ARS-SPARC-APMRU, Email: yubin.lan@ars.usda.gov; **Ronald E. Lacey**, Ph.D., Professor, Department of Biological and Agricultural Engineering, Texas A&M University, College Station, TX 77843, Tel: 979-845-3967. Email: ron-lacey@ tamu.edu; **Yanbo Huang**, Ph.D., Agricultural Engineer, USDA-ARS- Aptru, 141 Experiment Station Rd., Stoneville, MS, 38776, Tel: (662)820-7715. Email: Yanbo.Huang@ars.usda.gov; **W. Clint Hoffmann**, Ph.D., Agricultural Engineer, USDA-ARS-SPARC-APMRU, 2771 F&B Road College Station, TX 77845, Tel: 979-260-9521. Email: clint.hoffmann@ars.usda.gov; **D. Martin,** Research Engineer, USDA- ARS-SPARC-APMRU, 2771 F&B Road College Station, TX 77845, Tel: 979-260-9290; **G. C. Bora,** Ph.D., Assistant Professor, Department of Agriculture and Biosystems Engineering, North Dakota State University, Fargo, ND 58108, Tel: 701-231-7271. Email: ganesh.bora@ndsu.edu.

Corresponding author: Yubin Lan, Ph.D, Agricultural Engineer, USDA-ARS-SPARC-APMRU, 2771 F&B Road, College Station, TX 77845, Tel: (979)260-3759. Email: yubin.lan@ars.usda.gov.

some function such as variogram (also referred to as semivariogram) to characterize the spatial variation in a region of interest.

The determination of the spatial variability of field parameters is usually based on the concept that sampled values at nearby locations are more similar than those from further apart. Measurements from the field are usually gathered as point data, such as an individual plant. Geostatitical analysis methods can be used to interpolate the measurements to create a continuous surface map or to describe its spatial pattern[4]. As a powerful tool in geostatistics, variogram describes the spatial dependence of data and gives the range of spatial correlation, within which the values are correlated with each other and beyond which they become independent. The parameters of the best fitted model for a variogram can be used for Kriging[5,6]. Kriging has been recommended as the best method to interpolate point data since it minimizes the error variance using a weighted linear combination of the data[7]. Therefore, it is very important to estimate variogram reliably from sufficient data and modeled properly[8]. The effect of sampling on the accuracy of sample variogram was studied from independently generated random fields[9] and from experimental data[10]. Brus and De Grunijtar[11] concluded that design-based sampling strategies based on classical sampling theory offered unprecedented potentials for estimation of theoretical variograms. A sampling configuration, simple random sampling design, was tested for estimating the variograms of three soil properties in that study. Gascuel-Odoux and Boivin[12] investigated the consistency of the sample variograms and spatial estimates by a sub-sampling procedure. They took five series of 20 sub-samples with different sample size of data points from the initial sample and found that the consistency of both experimental and fitted variograms increased with sample size.

Remotely sensed imagery constitutes a record of distinct spatial properties of the Earth's surface. Images can be treated as "field" data depicted by varied digital numbers (DN). These spectral values of pixels are spatially auto-correlated and their spatially dependent structures can be represented by variogram. Variogram has been estimated and investigated in a wide range of remote sensing applications[13-20]. Woodcock et al.[21-22] had studied the sensitivity of variogram by varying parameters of scene models in calculating explicit variograms, simulating images and real digital images. They found that the height of the variogram was affected by the density of the coverage of the objects in the scene and the range of the variogram changed with the size of the object. They also found that when the variance in the distribution of the sizes of objects increased, the shape of the variogram curve becomes more round. Atkinson and Emery[23] explored the relationship between wavelength and spatial structure which was summarized by the variogram. Several studies have applied the geostatistical analysis on various agricultural applications, such as soil properties[24-27], crop yield monitor data[28-30], and crop qualities[31-33].

With advanced multispectral imaging systems, aerial images are now collected in several bands. Few studies investigated how the variogram changes with various sample sizes of data by given a certain remotely sensed imagery. The objective of the study was to explore the effect of sample size of image data on sample variogram estimation and find out how much image data could be used to estimate variogram reliably. Design-based simple random sampling was used for the sub-sampling procedure.

2 Material and methods

2.1 The image

The imaging system used to acquire multispectral image is TerraHawk® Aerial Imaging System. An MS4100 multi-spectral camera (Geospatial Systems, Inc., Rochester, NY) is the central component of the airborne multi-spectral imaging system. The image sensors are charge coupled device (CCD) array sensors with spectral sensitivity from 400-1000 nm and support three standard models for RGB, Color Infrared (CIR) and RGB/CIR with blue band between 437 and 483 nm, green band between 520 and 560 nm, red band between 640 and 680 nm, and near-infrared (NIR) band between 767 and 833 nm. They approximated Landsat satellite (NASA, Washington DC and USGS, Reston, VA) bands. NIR,

Red, and Green bands can be combined to make CIR image, which is often called "false-color" image. This band combination makes vegetation appear as shades of red. Brighter reds indicate more vigorously growing vegetation. Soils with no or sparse vegetation will range from white to greens or browns depending on moisture and organic matter content[34].

A CIR aerial image was obtained over the Texas AgriLife Research Farm (30.524588°N, 96.407181°W), College Station, Texas in Feb 2009. The field had been left fallow for the previous eight months and thus, was inundated with both broadleaf and grass weeds. The acquired raw image was calibrated and processed into reflectance image. A 100 × 100-pixel subset was randomly chosen from the reflectance image and highlighted in Figure 1.

Figure 1 Aerial multispectral image of Texas AgriLife Research Farm field, College Station, Texas obtained in Feb, 2009

2.2 Variogram and fitting model

The sample variogram was computed by Matheron's method of moments (MoM) estimator[27]. The spatial variance between the digital numbers of any two distinct pixels would depend on their separation distance, lag *h*. The semivariance, $\gamma(h)$, between any two pixels at a lag *h* can be expressed as:

$$g(h)=\frac{1}{2}E[z(x)-z(x+h)]^2 \quad (1)$$

Where: $\gamma(h)$ is the semivariance at lag distance *h*; $z(x)$ is the digital number of the pixel at location *x*. In the region of interest, suppose there will be $m(h)$ pairs of pixels separated by a particular lag *h*. Their semivariance is given by equation:

$$\hat{g}(h)=\frac{1}{2m(h)}\sum_{i=1}^{m(h)}[z(x_i)-z(x_i+h)]^2 \quad (2)$$

where $\hat{g}(h)$ is an unbiased estimate of the variance of these $m(h)$ pairs of pixels; $m(h)$ is the number of pairs of pixels separated by a lag h for i=1, 2, ⋯, $m(h)$; $z(x_i)$ and $z(x_i+h)$ are the digital numbers of $z(x)$ at locations x and $x+h$, respectively. $\hat{g}(h)$ is a useful measure of dissimilarity between spatially distributed regionalized variables. The larger $\hat{g}(h)$ is, the less similar the pixels. The similarity between two pixels increases with decrease in the value of $\hat{g}(h)$.

When a variogram is plotted using discrete experimental data points, it is called an experimental or sample variogram. A theoretical model can be fitted through the experimental data points to quantify spatial patterns. The shape and description of a "classic" variogram[13, 15] is shown in Figure 2.

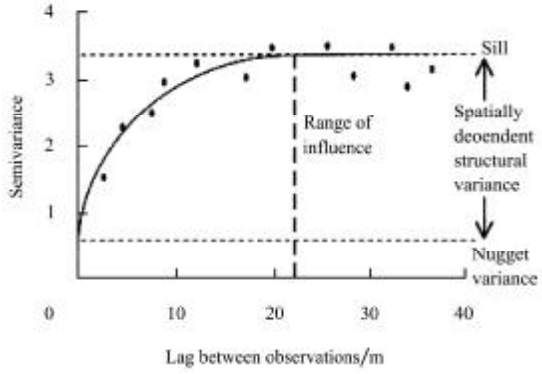


Figure 2 Shape and description of a "classic" variogram[13,15]

There are three key terms in each model, the sill, the range, and nugget variance. The sill corresponds to the overall variance in the dataset and the range is the maximum distance of spatial autocorrelation[35]. The nugget variance is the positive intercept of the variogram and can be caused by measurement errors or spatial sources of variation at distances smaller than the sampling interval or both.

The spherical model is the most commonly used model for experimental data[36] and expressed as:

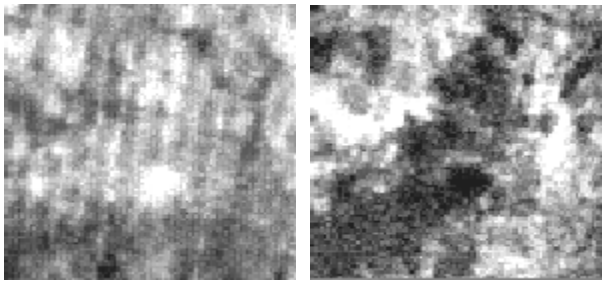
$$g(h)=\begin{cases} c_0 & \text{when } h=e \text{ (a very small lag)} \\ c_0+\mathrm{c}\left(\frac{3h}{2a}-\frac{1}{2}\left(\frac{h}{a}\right)^3\right) & \text{when } 0<h\le a \\ c_0+\mathrm{c} & \text{when } h>a \end{cases} \quad (3)$$

where c_0 is the nugget variance, $c+c_0$ is sill, h is the lag

and a is the range. All variograms computed in this study are all fitted with spherical model.

2.3 Data collection and analysis

The image was processed in the Environment for Visualizing Images (ENVI) software package (Version 4.5, ITT Visual Information Solution, www.ittvis.com). A 100×100-pixel subset was randomly selected from the image. The subset image and its NDVI were shown in Figure 3a and b, respectively. Each subset comprised of total of 10 000 pixels. Since the spatial resolution of the image was 0.51 m, it covered 0.51×0.51 m^2 area of the field.



a. subset image b. NDVI

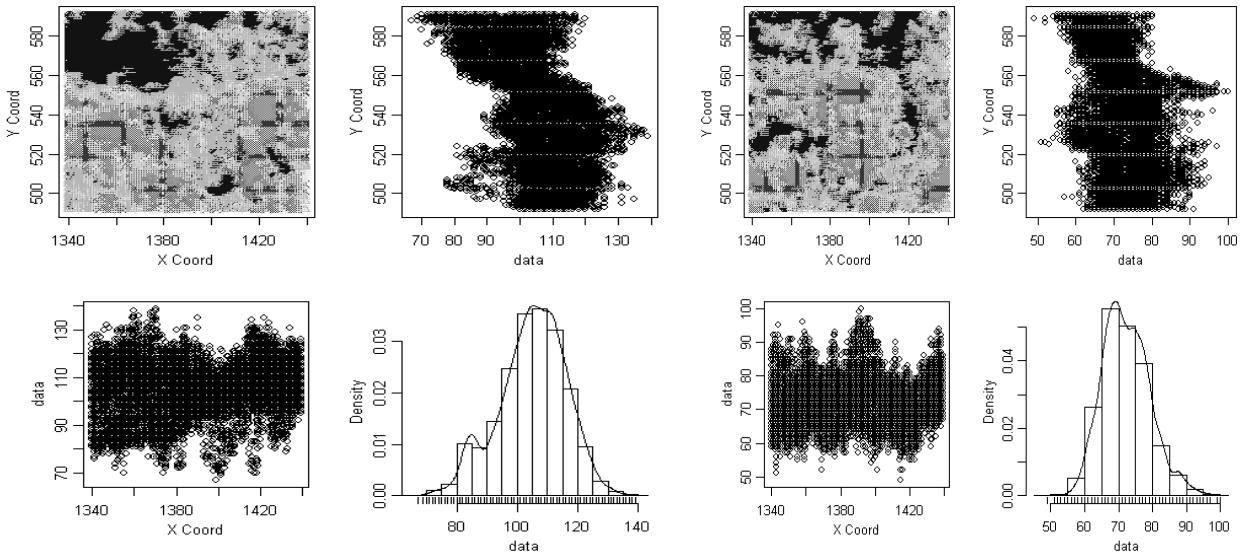
Figure 3 Comparison between the subset image and NDVI

These two data sets were exported as ASCII text files. The data file from original image contained the coordinates of the pixels and the pixel values of three wavebands, NIR, Red and Green. The data file of NDVI image contained the coordinates and the NDVI values of the pixels. These two data files were then imported into R statistical software(R 2.8.1, www.r-project.org) and converted into four geostatistical data sets, NIR, Red, Green wavebands and NDVI, with as.geodata function. Variograms were computed for each of the NIR, Red, Green wavebands and NDVI. The spherical model was fitted to those variograms and the sill, nugget and range were identified. To investigate how variograms and those parameters change with sample size, these four geodata sets were randomly sub-sampled in R. The sub-samples were taken independently from each other. For each sub-sample, locations were selected randomly and independently.

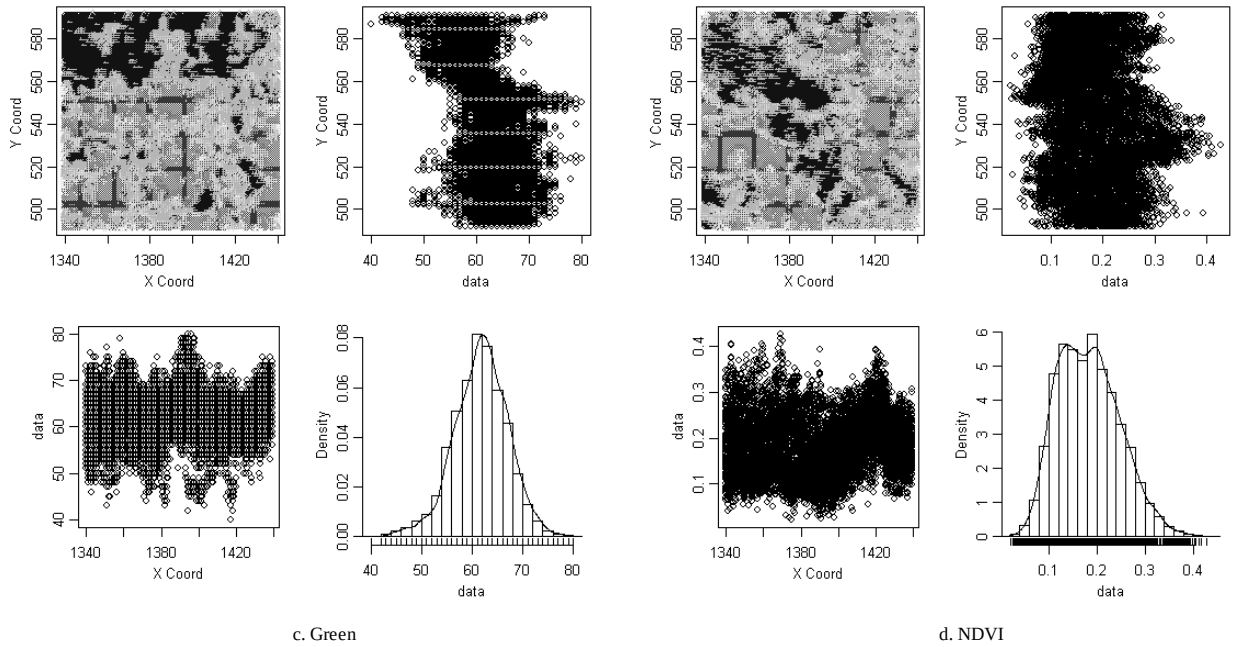
3 Results and discussion

3.1 Full data sets analysis

Geostatistical methods are optimal when data are normally distributed and stationary (mean and variance do not vary significantly in space). Significantly deviation from normality and stationarity can cause problems. The scatter plots of four geodata sets were shown in Figure 4. Each scatter plot consists of four subplots, which were x versus y coordinate plot, data versus y coordinate plot, x coordinate versus data plot, and the histogram plot. By looking at the histograms in the subplots of three wavebands and NDVI, severe deviation from normality was not observed. It can be noticed that there were some hint of NW- SE trends from the subplots of data versus y coordinate for NIR and Green wavebands.



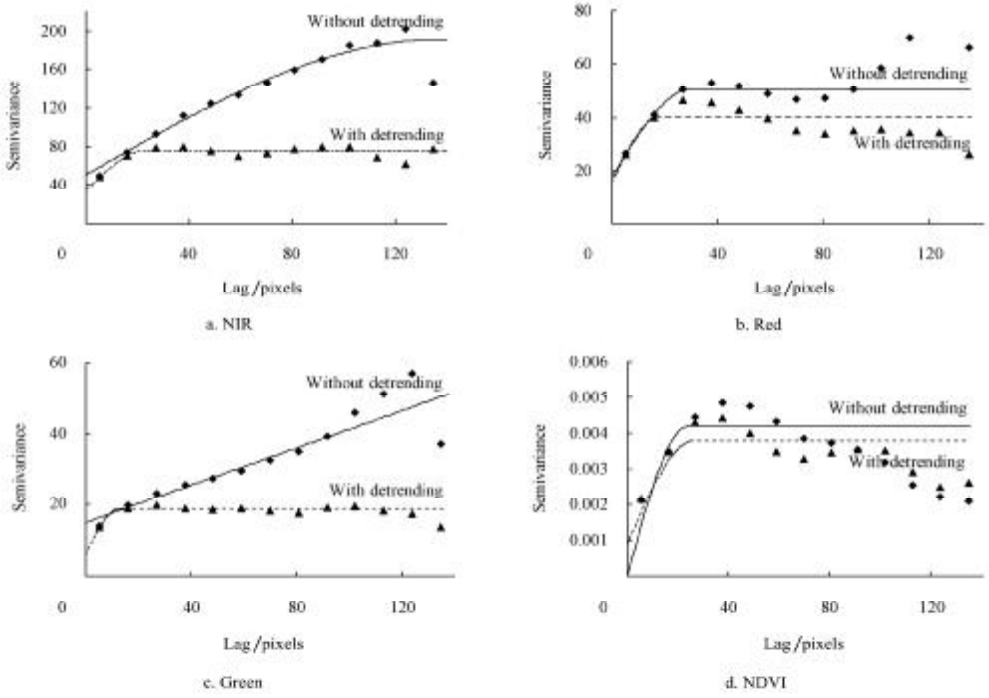
a. NIR b. Red



c. Green d. NDVI

Figure 4 Plots of geodata sets: NIR, Red, Green, and NDVI

Variograms were computed on them and fitted with spherical models (Figure 5). If the semivariance increases steadily over the separation distance, it is often indicative of a significant spatial trend in the variable. The variograms for Green waveband data indicated a significant trend. A spatial trend usually results in a negative correlation between variables separated by large lags. Trend surface fit is always needed. After trend surface fit for NIR, the range of variogram dropped from 137 to 23 pixels; the sill decreased from 141 to 41; and the nugget variance reduced from 50 to 34. For Red, detrending only reduced the range by 7 pixels but the partial sill and nugget variance increased slightly. For Green, both range and partial sill of variogram decreased dramatically with detrending. There is no noticeable difference between two variograms for NDVI.



a. NIR b. Red

c. Green d. NDVI

Figure 5 Variograms computed on NIR, Red, Green wavebands and NDVI with and without detrending

3.2 Sub-sampling on the NIR band

The sample variograms for NIR were computed with samples sizes 50, 100, 300, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, and 9000 pixels (Figure 6).

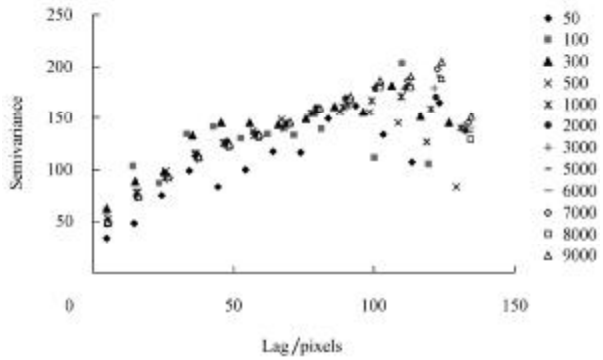


Figure 6 Sample variograms with sample sizes from 50 to 9000 pixels for NIR

All the parameters of variograms fitted with spherical models for NIR are summarized in Table 1. The range and partial sill of the sample size 50 were similar to those of the sample size 10000. The range was 50 pixels for the sample size 100. The range was increasing until the sample size reaches 5000, which is half of the total sample size. After that, the range became almost stable. The nugget, sill, and nugget to sill ratio became stable after the sample size1000.

Table 1 Parameters of variograms fitted with spherical models with various sample sizes for NIR waveband

Sample size(pixel)	Range(pixel)	Nugget	Sill	Nugget/Sill/%
50	138	35	161	22
100	50	50	138	36
300	60	46	151	30
500	83	48	154	31
1000	107	55	165	33
2000	116	49	177	27
3000	128	49	183	27
4000	118	49	176	28
5000	136	50	187	27
6000	131	48	184	26
7000	136	52	188	27
8000	130	49	182	27
9000	139	51	193	26
10000	137	51	192	26

3.3 Sub-sampling on the Red band

The sample variograms for Red were computed on samples size 50, 100, 300, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, and 9000 pixels (Figure 7).

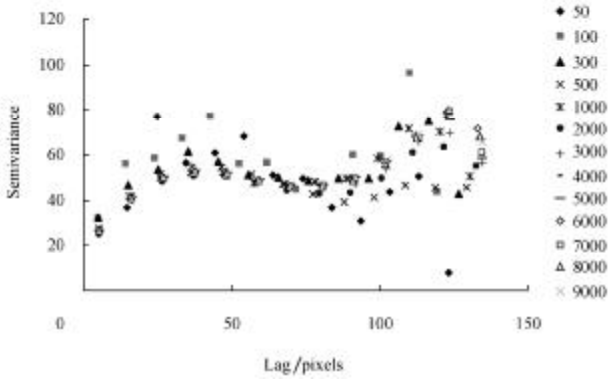


Figure 7 Sample variograms with sample sizes from 50 to 9000 pixels for Red.

All the parameters of variograms fitted with spherical models for Red are presented in Table 2. The sample sizes of 50 and 100 appeared to be pure nugget models, which mean there was no spatial dependence in the data. From the sample size 300, the range gradually increased until the sample size 5000 with exception of sample size 1000, which might be caused by the randomly sampling process by the computer. Beyond the sample size 5000, all parameters became stable.

Table 2 Parameters of variograms fitted with spherical models with various sample sizes for Red waveband

Sample size(pixel)	Range (pixel)	Nugget	Sill	Nugget/Sill/%
50	361	54	54	100
100	656	58	58	100
300	14	10	53	19
500	14	10	53	19
1000	505	44	80	55
2000	28	16	48	33
3000	29	17	50	34
4000	31	19	51	37
5000	30	17	50	34
6000	30	17	50	34
7000	30	17	49	35
8000	29	17	50	34
9000	30	17	50	34
10000	30	17	50	34

3.4 Sub-sampling on the Green band and NDVI

Similar procedures had been undertaken for Green and NDVI. For Green, all variograms computed on all subsample sizes were similar to the variogram without detrending shown in Figure 5. For NDVI, all the variograms and model parameters were remarkably

consistent. The scale of spatial dependence of the NDVI was 24 pixels, which was about 12 m on the ground.

4 Conclusions

In this study, the analysis of variograms computed on various sample sizes of remotely sensed data was conducted. A 100 × 100-pixel subset was chosen randomly from an aerial multispectral image which contains Green, Red and NIR wavebands over our weed study field. Green, Red, NIR and NDVI datasets were imported into R software for spatial analysis. By fitting with spherical models, behaviors of the major parameters of those variograms were investigated. In this case, it turned out that half size of the subset image data was enough to reliably estimate the variograms for NIR and Red wavebands. To map the variation on NDVI within the weed field, sampling interval should be smaller than 12 m. The information will be particularly important for kriging and also give a good guide of fieldwork in the future study.

[References]

[1] Stein A, Brouwer J, Bouma J. Methods for comparing spatial variability patterns of millet yield and soil data. Soil Science Society of American Journal, 1997; 61: 861－870.

[2] Stewart C M, McBratney A B, Skerrit J H. Site-specific durum wheat quality and its relationship to soil properties in a single field in Northern New South Wales. Precision Agriculture, 2002; 3(2): 155－168

[3] Zhang H H, Lan Y B, Lacey R E, et al. Analysis of vegetation indices derived from aerial multispectral and ground hyperspectral data. International Journal of Agricultural and Biological Engineering, 2009; 2(3): 33－40.

[4] Cressie N A C. Statistics for Spatial Data. New York, N.Y.: John Wiley & Sons, 1993.

[5] Matheron G. Principles of geostatistics. Economic Geology, 1963; 58(8): 1246－1266.

[6] Stein A, Corsten L C A. Universal kriging and cokriging as a regression procedure. Biometrics, 1991; 47(2): 575－588.

[7] Panagopoulos T, Jesus J, Antunes M, Beltrao J. Analysis of spatial interpolation for optimizing management of a salinized field cultivated with lettuce. European Journal of Agronomy, 2006; 24: 1－10.

[8] Oliver M A, Webster R. How geostatistics can help you. Soil use and Management, 1991; 7(14): 206－217.

[9] Webster R, Oliver M A. Sample adequately to estimate variogram of soil properties. Journal of Soil Science, 1992; 43: 177－192.

[10] Van Meirvenne M, Hofman G. Sampling strategy for quantitative soil mapping. Pedology, 1991; 41: 263－275.

[11] Brus D J, De Gruijter J. Estimation on non-ergodic variograms and their sampling variance by design-based sampling strategies. Mathematical Geology, 1994; 26(4): 437－454.

[12] Gascuel-Odoux C, Boivin P. Variability of variograms and spatial estimates due to soil sampling: a case study. Geoderma, 1994; 62: 165－182.

[13] Curran P J. The semivariogram in remote sensing: An introduction. Remote Sensing of Environment, 1988; 24(3): 493－507.

[14] Jupp D L B, Strahler A H, Woodcock C E, et al. Autocorrelation and regularization in digital images. I. Basic theory. Geoscience and Remote Sensing, IEEE Transactions, 1988a ; 26(4): 463－473.

[15] Jupp D L B, Strahler A H, Woodcock C E, et al. Autocorrelation and regularization in digital images. II. Simple image models. Geoscience and Remote Sensing, IEEE Transactions, 1988b; 27: 247－258.

[16] Curran P J, Atkinson P M. Geostatistics and remote sensing. Progress in Physical Geography, 1998; 22(1): 61－78.

[17] Cohen W B, Spies T A, Bradshaw G A. Semivariograms of digital imagery for analysis of conifer canopy structure. Remote Sensing of Environment, 1990; 34(3): 167－178.

[18] Atkinson P M, Curran P J. Defining an optimal size of support for remote sensing investigations. Geoscience and Remote Sensing, IEEE Transactions, 1995; 33(3): 768－776.

[19] Oliver M A, Webster R, Slocum K, et al. Filtering SPOT imagery by kriging analysis. International Journal of Remote Sensing, 2000; 21(4): 735－752.

[20] Oliver M A, Shine J A, Slocum K R, et al. Using the variogram to explore imagery of two different spatial resolutions. International Journal of Remote Sensing, 2005; 26(15): 3225－3240.

[21] Woodcock C E, Strahler A H, Jupp D L B, et al. The use of variograms in remote sensing: I. Scene models and simulated images. Remote Sensing of Environment, 1988; 25(3): 323－348.

[22] Woodcock C E, Strahler A H, Jupp D L B, et al. The use of variograms in remote sensing: II. Real digital images. Remote Sensing of Environment, 1988; 25(3): 349－379.

[23] Atkinson P M, Emery D R. Exploring the relation between spatial structure and wavelength: Implications for sampling reflectance in the field. International Journal of Remote Sensing, 1999; 20(13): 2663－2678.

[24] Cambardella C A, Karlen D L. Spatial analysis of soil fertility parameters. Precision Agriculture, 1999; 1: 5－14.

[25] Hengl T G, Heuvelink B M, Stein A. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma, 2004; 120: 75－93.

[26] Iqbal J, Thomasson J A, Jenkins J N, Owens P O, Whisler F D. Spatial variability of soil physical properties of alluvial soils. Soil Science Society of America, 2005; 69: 1338－1350.

[27] Ge Y, Thomasson J A, Morgan C L, Searcy S W. VNIR diffuse reflectance spectroscopy for agricultural soil determination based on regression-kriging. Transaction of the ASABE, 2007; 50: 1081－1092.

[28] Yang C, Everitt J H. Relationships between yield monitor data and airborne multidate multispectral digital imagery for grain sorghum. Precision Agriculture, 2002; 3: 373－388

[29] Dobermann A, Ping J L. Geostatistical integration of yield monitor data and remote sensing improves yield maps. Agronomy Journal, 2004; 96: 285－297.

[30] Miao Y, Mulla D J, Robert P C. Spatial variability of soil properties, corn quality and yield in two Illinois, USA fields: implications for precision corn management. Precision Agriculture, 2006; 7: 5－20.

[31] Kravchenko A N, Bullock D G. Spatial variability of soybean quality data as a function of field topography: I. Spatial data analysis. Crop Science, 2002; 42: 804－815.

[32] Johnson R M, Richard E P, Jr. Sugarcane yield, sugarcane quality, and soil variability in Louisiana. Agronomy Journal, 2005; 97: 760－771.

[33] Ge Y, Thomasson J A, Sui R, Morgan C L, Searcy S W, Parnell C B. Spatial variation of fiber quality and associated loan rate in a dryland cotton field. Precision Agriculture, 2008; 9: 181－194.

[34] Stein A, Meer Van Der F D, Gorte B, et al. Saptial statistics for remote sensing. Kluwer Academic Publishers, 1999.

[35] Matheron G. Les variables re´gionalise´es et leur estimation: une application de la the´orie defonctions ale´atoires aux sciences de la nature. Paris: Masson et Cie, 1965.

[36] Webster R, Oliver M A. Geostatistics for environmental scientists (2nd Ed.). John Wiley & Sons Ltd., Chichester, 2007.