International Soil and Water Conservation Research 4 (2016) 151–158

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

International Soil and Water Conservation Research

journal homepage: www.elsevier.com/locate/iswcr

Original Research Article

Independent principal component analysis for simulation of soil water content and bulk density in a Canadian Watershed

Alaba Boluwade*, Chandra A. Madramootoo

Department of Bioresource Engineering McDonald Campus of McGill University, 21111 Lakeshore Road, Ste-Anne-De-Bellevue, Quebec H9X3V9, Canada

ARTICLE INFO

Article history:
Received 22 January 2016
Received in revised form
9 August 2016
Accepted 10 September 2016
Available online 19 September 2016

Keywords:
Spatial variability
Soil properties
Agricultural decision-making
Independent principal component analysis

ABSTRACT

Accurate characterization of soil properties such as soil water content (SWC) and bulk density (BD) is vital for hydrologic processes and thus, it is importance to estimate θ (water content) and ρ (soil bulk density) among other soil surface parameters involved in water retention and infiltration, runoff generation and water erosion, etc. The spatial estimation of these soil properties are important in guiding agricultural management decisions. These soil properties vary both in space and time and are correlated. Therefore, it is important to find an efficient and robust technique to simulate spatially correlated variables. Methods such as principal component analysis (PCA) and independent component analysis (ICA) can be used for the joint simulations of spatially correlated variables, but they are not without their flaws. This study applied a variant of PCA called independent principal component analysis (IPCA) that combines the strengths of both PCA and ICA for spatial simulation of SWC and BD using the soil data set from an 11 km^2 Castor watershed in southern Quebec, Canada. Diagnostic checks using the histograms and cumulative distribution function (cdf) both raw and back transformed simulations show good agreement. Therefore, the results from this study has potential in characterization of water content variability and bulk density variation for precision agriculture.

1. Introduction

The accurate characterization of soil water content (SWC) and other soil properties such as bulk density (BD) are paramount in guiding agricultural management decisions and for increasing the potential of the soil for crop productivity. Grote, Anger, Kelly, Hubbard, and Rubin (2010) reported that the characterization of SWC could be difficult due to its spatial and temporal variability. Also, obtaining the adequate soil sample size to characterize the heterogeneities of SWC can be expensive (Grote et al., 2010). Their accurate estimate is important for maximizing crop yield, sustainable irrigation practice and reduction in the negative impact on the environment.

Soil properties such as BD equally influence the productivity of the soil for crop productivity. BD is an important factor for soil nutrients retention. Also management and several factors such as land use, geomorphology and soil heterogeneities affect BD (Geypens, Vanongeval, Vogels, & Meykens, 1999). SWC and BD parameters vary in space and time Zhang et al. (2014). There is spatial correlation among these variables (Bivand, Pebesma, & Gómez- Rubio, 2008) due to various management decisions such as irrigation application, soil tillage and so forth. Farmers hope to be productive, enhance plant growth and maximize yield (Fulton, Wells, Shearer, & Barnhisel, 1996). There is reduction of 10–20% in yield due to soil compaction (Kisekka, Migliaccio, Muñoz-Carpena, Schaffer & Khare, 2014). Kisekka et al. (2014), Zang et al. (2014), Li and Xinmei (2014), Ngailo and Vieira (2012), Delbari, Afrasiab, and Loiskandl (2009), and Bourennanea et al. (2007) have all emphasized the importance of spatial correlation of soil properties such as BD and SWC.

The spatial correlations between bivariate or multivariate correlated variables can be characterized by variogram and cross variogram (Bivand et al., 2008). With an appropriate linear model of co-regionalization (LMC), the variables can be estimated or simulated. These simulations require processing of massive nodes, which make computation very difficult. Contributors to this complexity include the tedious inference and modeling of the cross-variograms, and computational inefficiency, substantially increased with the number of variables being simulated.

Principal component analysis (PCA) approach has been attempted in the past (Dimitrakopoulos & Makie, 2008; Goovaerts, 1997; Wackernagel, 1995) to solve this problem. A PCA method's major disadvantage involves the inability to remove cross-

* Corresponding author.
E-mail address: alaba.boluwade@mail.mcgill.ca (A. Boluwade).
Peer review under responsibility of International Research and Training Center on Erosion and Sedimentation and China Water and Power Press.

http://dx.doi.org/10.1016/j.iswcr.2016.09.001

correlations at distances other than zero (Goovaerts, 1997).

In this paper, we are proposing another technique, which combines the strengths of PCA and independent component analysis (ICA) (ICA is method that does not involve variable transformation and the random vector of spatially correlated variables can be decomposed into independent components (IC). The combination of both PCA and ICA is called independent principal component analysis (IPCA). IPCA has been used in the medical field by Yao, Coquery, and Cao (2012). The current study is the first time it has been applied in domain of geostatistics. IPCA technique overcomes the shortcomings of the techniques above because of its ability to reduce noise and better reflect the internal structure of the data set. When compared with PCA, IPCA removes correlation at all lag distance. Also, initial normal score transformation is not required unlike minimum and maximum autocorrelation factors (MAF). In case of ICA, IPCA do not suffer from high dimensionality (Yao et al., 2012).

Therefore, the aim of this paper is to apply IPCA for spatial joint simulation of SWC and BD. This is relevant in the hydrologic processes such as infiltration, erosion and flooding studies (Delbari et al., 2009). IPCA would by-pass the complex matrix inversion in the direct and cross variogram analysis. The goal of the technique is to obtain independent components (IC) that provide accurate structure of the data sets and achieve stochastic simulations better than MAF, PCA and ICA. The joint simulation of IC will be performed using sequential Gaussian simulation (SGS) technique.

2. Methodology

2.1. Stochastic IPCA modeling procedure

As discussed under "Introduction section", IPCA combines both PCA and ICA (Yao et al., 2012). PCA is first applied to the $X[nx\,p]$, where n is the sample size and p represents the number of attributes. X can also be a centered data matrix using the singular value decomposition (SVD) to extract the loading vectors (Yao et al., 2012):

$$X(u)=U(u)DV^T \tag{1}$$

Where

- $X(u)$ is the centered data matrix and u represents a geographical location
- U is an $n\,X\,p$ matrix whose columns are uncorrelated
- D is a $p\,X\,p$ diagonal matrix with diagonal elements d_j
- V^T is the transpose of the orthogonal matrix with $V^TV=I_p$

From Eq. (1), the columns of V contain the loading vectors and they are whitened (Yao et al., 2012).

The next step involves applying ICA to the whitened vectors using the fastICA algorithm (, 1998, 1999; Langlois, Chartier, & Gosselin, 2010; Yao et al., 2012). Detailed review about fastICA algorithm can be seen at (Hyvärinen, 1999; Langlois et al., 2010). Therefore we will not discuss its procedure.

The IPCA can therefore be summarized as (Yao et al., 2012):

- *Application of PCA using SVD*: This is applied on the centered data matrix$X$. This generates the loading vectors V. The number of components$S$is chosen to reduce dimensionality.
- *Application of ICA:* The fastICA algorithm is applied on vectors V to obtain the independent loading vectors, m^T
- *Projection of the centered data matrix X:* This is applied on the S independent loading vectors to obtain the independent components.
- *Ordering the IPCS:* The kurtosis value is used to order the IPCS.

2.2. Geostatistical modeling procedure

The second stage of this methodology involves using the independent components obtained from above for stochastic joint simulation. We selected Sequential Gaussian Simulation technique for this purpose. The major advantage of the SGS technique in stochastic simulation is that it provides an estimate of both the mean and standard deviation of the variable at each grid node (Boluwade & Madramootoo, 2013; Lin, 2008). SGS chooses a random deviate (through a Monte Carlo technique) from the Gaussian distribution, selected according to a uniform random number representing the probability level (Boluwade & Madramootoo, 2013; Lin, 2008). SGS also generates set of realizations that have statistics similar to that of the conditioning data set and equally quantifies the spatial uncertainty (Boluwade & Madramootoo, 2013; Lin, 2008).

In this study, we summarized the second stage as (Lin, 2008; Boluwade & Madramootoo, 2013):

- Normal score transformation for the independent components
- Create a random path for the grid D of locations (u) to be simulated with a data X (u) of n such that, $X(u)=\left\{x_{1(u)..,}x_2(u).........x_n(n)\right\}$ with initialize i=1
- Let the algorithm visit ith node of the grid G and estimate the mean and variance using Simple Kriging technique conditioned on the values of the data in the neighbor.
- Draw a random value from the Gaussian distribution of the considering node. This value is considered an SGS estimate.
- This estimate is added back and treated as an observation for next visit. This is done until all the nodes are visited. When this is finished (i.e., when $i = D$), we obtained one realization, Li
- Next, the realizations are back-transform of the generated realizations first from normal score and later into data space using coefficients (the mixing matrix)
- Rescaling of the means (μ) for each corresponding components.

2.3. Study area

This study was conducted at Castor watershed (Fig. 1) which is located in southern Quebec, Canada. This watershed has a total area of 11 km^2 under reduced tillage system. It drains into Pike River, which is a tributary to the Mississquoi Bay that is located on the northeastern part of Lake Champlain. This location was chosen because the Castor watershed is on the downstream part of the Pike River watershed where there are intensive agricultural activities. The land use pattern has been estimated to be 44% corn, 28% grass and 20% cereal (Beaudin, Michaud, & Desjardins, 2005; Boluwade, & Madramootoo, 2013). The data sets were collected during the 2011 summer season.. A random stratified sampling (according to the soil classes) was done to cover the spatial extent of the study area (Fig. 1). Each of the points was Geo-referenced and imported into a global positioning system (GPS) which has a positional accuracy of ± 2 m. The sampling depth used was 0–0.30 m. The soil samples were collected using standard core size. Gravimetric technique was used to determine both BD and SWC. The total soil sample size is 144.

3. Results and discussion

3.1. Descriptive statistics and spatial autocorrelation of soil properties

Table 1 and Fig. 2 show the descriptive statistics and histogram