Geoderma 219–220 (2014) 46–57

Contents lists available at ScienceDirect

Geoderma

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

Covariate selection with iterative principal component analysis for predicting physical soil properties

Matthew R. Levi [*,1], Craig Rasmussen

University of Arizona, Soil, Water, and Environmental Science Department, 1177 E. Fourth St. Shantz Bldg. Room 429, Tucson, AZ 85721-0038, United States

ARTICLE INFO

Article history:
Received 27 December 2012
Received in revised form 6 December 2013
Accepted 10 December 2013
Available online 16 January 2014

Keywords:
Digital soil mapping
Regression kriging
Landsat
Spatial variability
Terrain analysis
Data reduction

ABSTRACT

Local and regional soil data can be improved by coupling new digital soil mapping techniques with high resolution remote sensing products to quantify both spatial and absolute variation of soil properties. The objective of this research was to advance data-driven digital soil mapping techniques for the prediction of soil physical properties at high spatial resolution using auxiliary data in a semiarid ecosystem in southeastern Arizona, USA. An iterative principal component analysis (iPCA) data reduction routine of reflectance and elevation covariate layers was combined with a conditioned Latin Hypercube field sample design to effectively capture the variability of soil properties across the 6250 ha study area. We sampled 52 field sites by genetic horizon to a 30 cm depth and determined particle size distribution, percent coarse fragments, Munsell color, and loss on ignition. Comparison of prediction models of surface soil horizons using ordinary kriging and regression kriging indicated that ordinary kriging had greater predictive power; however, regression kriging using principal components of covariate data more effectively captured the spatial patterns of soil property–landscape relationships. Percent silt and soil redness rating had the smallest normalized mean square error and the largest correlation between observed and predicted values, whereas soil coarse fragments were the most difficult to predict. This research demonstrates the efficacy of coupling data reduction, sample design, and geostatistical techniques for effective spatial prediction of soil physical properties in a semiarid ecosystem. The approach applied here is flexible and data-driven, allows incorporation of wide variety of numerically continuous covariates, and provides accurate quantitative prediction of individual soil properties for improved land management decisions and ecosystem and hydrologic models.

Published by Elsevier B.V.

1. Introduction

Information on the spatial variability of soil properties is required for input to soil erosion models (Chen et al., 2011), hydrology models (Miller and White, 1998; Peschel et al., 2006), site-specific agricultural management (Duffera et al., 2007), and digital soil risk assessments that impact socioeconomic and environmental policy (Carre et al., 2007). Coarse scale soil information masks spatial variability of soil properties important for such landscape modeling at local and regional scales (Lathrop et al., 1995; Singh et al., 2011). The majority of available soils information derives from soil survey efforts that commonly provide little information regarding spatial variability within a soil map unit or accuracy assessments of reported soil properties. This lack of information can present problems for scaling and effectively incorporating soil data into landscape scale models (Wang and Melesse, 2006). Here we develop a robust, data-driven approach for predicting soil physical properties in a continuous raster data format. Specifically, we couple iterative data reduction of covariate layers with model-based sampling design and regression kriging to quantify soil physical properties in a complex semiarid ecosystem.

One of the most important factors for predicting soil properties across the landscape is the distribution of sampling locations. Traditional statistical approaches do not consider spatial correlation of variables or the relative position of sampling locations (Di et al., 1989). These methods can be considered design-based models because they introduce a stochastic element with the determination of sample locations, whereas model-based designs attempt to describe the reality of soil properties that are present as a result of the stochastic soil forming components for a given area (Brus and deGruijter, 1997). While both design- and model-based approaches can be used for predicting soil properties (Brus and deGruijter, 1997), recent efforts have focused on model-based sampling designs for implementing landscape-scale soil prediction models (Minasny and McBratney, 2006). Although many digital soil mapping studies utilize existing soil datasets for developing soil prediction models (Hengl et al., 2007b; Maselli et al., 2008; Ziadat, 2005), estimating soils in an area without existing soil data requires the selection of a sampling design.

Abbreviations: cLHS, Conditioned Latin Hypercube sampling design; iPCA, Iterative principal component analysis; NED, National elevation dataset; RK, Regression kriging.

* Corresponding author. Tel.: +1 575 646 3557.
E-mail address: mrlevi21@email.arizona.edu (M.R. Levi).
[1] Present address: USDA-ARS Jornada Experimental Range, MSC 3JER, Box 30003, New Mexico State University, Las Cruces, NM 88003, USA.

0016-7061/$ – see front matter. Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.geoderma.2013.12.013

Developing a sampling design provides the opportunity to address particular questions of interest and allows the incorporation of special considerations that can maximize the potential for accurately predicting soil properties. In addition to the selection of sample locations in geographic space (i.e., X and Y coordinates), a considerable amount of attention has been focused on spreading sampling locations in the feature space of available auxiliary data (Brungard and Boettinger, 2010; Hengl et al., 2003; Minasny and McBratney, 2006). An optimal sampling design for an area where functional relationships between soil properties and auxiliary information are not known should aim to simultaneously represent geographical space and feature space of available data (Hengl et al., 2003). One method of achieving this is with a conditioned Latin Hypercube sampling design (cLHS) to create sample locations that represent the variability of available covariate data (Minasny and McBratney, 2006). Stratification of sample locations in both feature space and geographic space can optimize deterministic and stochastic prediction models by providing the necessary sampling structure for each technique (Hengl et al., 2003; McBratney et al., 2000).

Interpolation methods such as ordinary kriging provide coarse estimates of soil variability with limited gain in information relative to vector based soil maps. Ordinary kriging is one of the most common geostatistical approaches used in digital soil mapping and is often used for comparison purposes against other spatial modeling methods (Bishop and McBratney, 2001; Li and Heap, 2011; Scull et al., 2005). Auxiliary information is often available for a given area and presents the opportunity of using hybrid prediction models that combine non-spatial prediction methods like regression with spatial methods such as kriging (Hengl et al., 2004, 2007a; McBratney et al., 2000). The term regression kriging was first coined by Odeh et al. (1994) and refers to using regression to extract information from sampled locations using covariate layers and then modeling the residuals with ordinary kriging. Kriging of residuals can minimize problems associated with uncertainty in the secondary information (Bishop et al., 2006).

There are multiple approaches to digital soil mapping that use a wide variety of covariate data. For example, surface reflectance data such as Landsat (Eldeiry and Garcia, 2010; Neild et al., 2007), SPOT (Carre and Girard, 2002), IKONOS (Eldeiry and Garcia, 2008), and MODIS (Hengl et al., 2007a) have all been used for soil prediction models. Digital elevation models are also common data sources for soil prediction and come in a variety of spatial resolutions (Hengl et al., 2007b; McKenzie and Ryan, 1999; Ziadat, 2005). If global soil mapping efforts are to be successful for projects like the GlobalSoilMap project (Sanchez et al., 2009), a method of identifying important auxiliary variables from the numerous available data sets is needed to determine the best data for input to soil prediction models. Tesfa et al. (2009) used correlation filtering in association with an importance measure from random forests to determine explanatory variables important for modeling soil depth. Another example is the optimum index factor, which is based on the variance and correlation of different reflectance band ratios (Chavez et al., 1982). In some cases, selection is based on expert knowledge and the availability of data for a given area. Though numerous methods have been employed to select important layers of information from the plethora of available data, band selection methods often produce different results (Beaudemin and Fung, 2001). A standard approach to selecting input data to soil prediction models has yet to be developed. Here we used an iterative principal component analysis (PCA) data reduction process similar to Hengl et al. (2007b) as a data-driven approach to determine important covariate layers.

The objectives of this study were to develop a data-driven soil prediction model for estimating physical soil properties of surface horizons in a semiarid ecosystem using a combination of surface reflectance and digital elevation model (DEM) covariates. We integrated iPCA for selecting covariate layers, a conditioned Latin Hypercube to design the sampling plan, and a hybrid geostatistical approach for soil property prediction. With this approach in mind, our hypotheses were 1) that covariate layers selected with the iterative data reduction technique would have a strong correlation with physical soil properties, 2) the cLHS design would produce a statistically robust sampling scheme to capture the spatial variability of soils in the study area, and 3) integrating covariate layers with spatial statistics using regression kriging would improve the prediction of soil properties on the landscape relative to either regression or ordinary kriging alone.

2. Materials and methods

2.1. Study area

The study area represents a sub-region of a recently mapped soil survey area (Graham County, AZ, Southwestern Part) of approximately 160,000 ha located 30 km north of the town of Wilcox in southeastern Arizona (Fig. 1). This soil survey represents a Soil Survey Geographic (SSURGO) data product that was mapped as a third order soil map with a mapping scale ranging from 1:20,000 to 1:63,360. The larger survey area includes a wide elevation gradient ranging from 910 to 1970 m asl with adjacent mountain ranges to the east and west that have maximum elevations of 3267 and 2336 m, respectively, that strongly influence local soil–landscape relationships. The current study was focused on a smaller area of interest of approximately 6265 ha with an elevation gradient of 1273 to 1655 m asl (Fig. 1). This area was selected because it represents the variability of landscape positions, geology, surface reflectance, and soils found in the surrounding areas. Soils in the study area were mapped as Argiustolls in the western third, Paleargids and Haplocambids in the eastern third, Haplogypsids and Gypsitorrerts in the central third, and Torrifluvents, Torriorthents, and riverwash in the drainages with areas of rock outcrop distributed throughout portions of the upland landscape positions (Soil Survey Staff, 2011).

Sedimentary basin fill deposits, including dissected and inset alluvial fans and fan terraces, cover the study area and range in age from Holocene to early Miocene-aged (20 Ma) materials (Richard et al., 2000; Wilson and Moore, 1958). Areas to the east consist of large, gently sloping alluvial fans formed from material eroded from Middle Proterozoic granitic rocks (1400–1450 Ma) and Early Proterozoic rocks (1600–1800 Ma) that include granite schist, gneiss, sandstone, andesite, and rhyolite, whereas basin fill deposits in the western portion of the study area consist of material eroded from Middle Miocene to Oligocene age volcanic rocks (20–30 Ma) that include andesite, rhyolite, and basalt, and are expressed on the landscape as a large alluvial fan composed predominantly of rhyolitic materials and an area of hills formed on residual basalt. Pliocene to Middle Pleistocene age lacustrine deposits that contain abundant carbonate and gypsum deposits occupy the center of the survey area (Fig. 1) (Melton, 1965). The major drainage network drains to the N–NW and stream channels are actively cutting back into the lacustrine sediments.

The wide variation in elevation, landform, and soils supports a diverse range of vegetation types across the study area. This area occupies the transition zone between Sonoran and Chihuahuan Deserts, which differ in their annual precipitation regimes and dominant vegetation communities (Brown, 1994; Neilson, 1987). Semi-desert grassland makes up the majority of the study area (Brown and Lowe, 1994) and includes a variety of grasses, forbs, shrubs, leaf succulents, and cacti (Brown, 1994).

The climate is semiarid with mean annual precipitation that ranges from 403 to 472 mm and has a bi-modal distribution with maximum rainfall during the summer monsoon and winter months (PRISM Climate Group, 2008). Mean annual air temperature ranges from 16 to 17 °C with average minimum temperature ranges from 9 to 10 °C and the average maximum temperature ranges from 23 to 25 °C. The soil temperature regime is thermic (15–22 °C), and soil moisture regimes include aridic and ustic, with the transition between the two occurring in the foothills of the neighboring mountain ranges (Soil Survey Staff, 2011, 2012).