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).

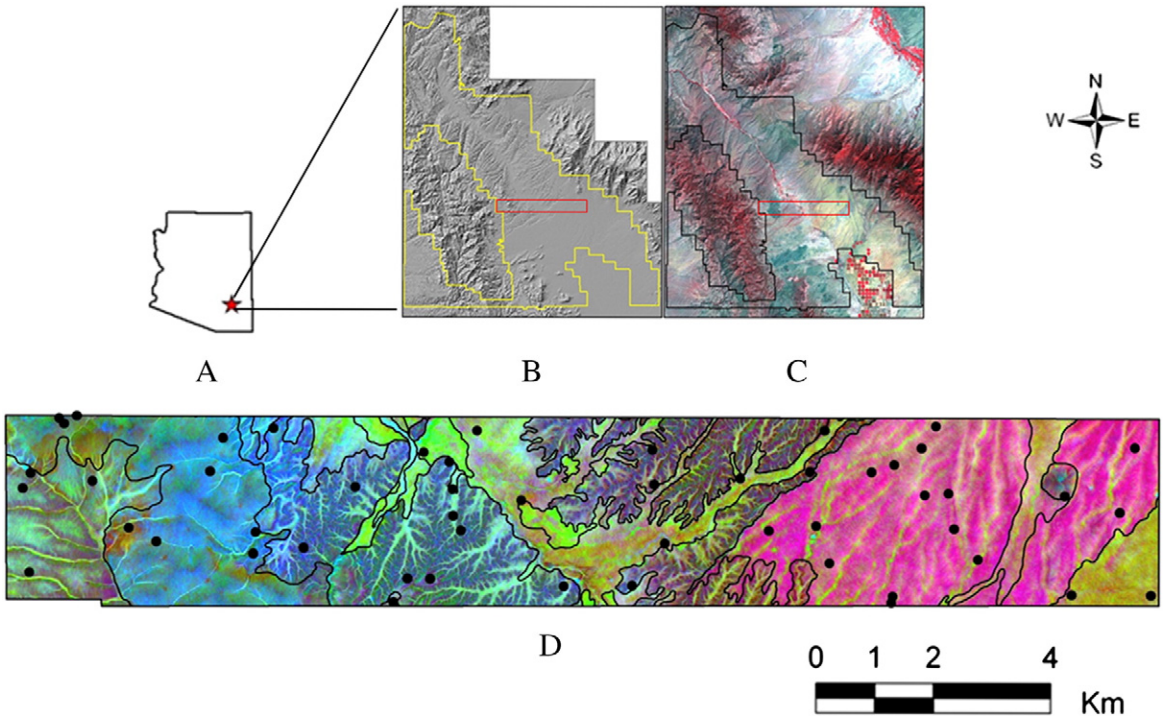

Fig. 1. Location of the soil survey area in southeastern Arizona (A). Processing of auxiliary data was performed across the 160,000 ha survey area (B and C) and used for comparing soil prediction models in the smaller area of interest (6265 ha) (D). Red boxes in (B) and (C) outline the smaller area of interest. Principal components of final covariate layers are shown in the detailed study area and highlight the high relief, as seen in the hillshade of the digital elevation model (B) and the wide range of parent materials which can be distinguished in the Landsat 7 ETM + false color composite image (C). Black lines on the study area represent published soil survey delineations and black points represent the location of 52 sampling locations used for soil prediction (D). Scale bar corresponds to the area of interest (D).

2.2. Project design

This project used a data-driven approach to predict surface soil properties with an underlying soil prediction model similar to the *scorpan* concept proposed by McBratney et al. (2003). Indices representing soil forming factors developed from both surface reflectance data and high-resolution elevation data were combined to provide a robust set of environmental covariates for soil prediction. Integrating surface reflectance and elevation indices provides a powerful set of predictors because it captures both existing surface characteristics and soil forming factors. The first step in this process was to develop an iterative data reduction technique that utilized principal component analysis to distill the information available in remotely-sensed reflectance and elevation covariates (iPCA). A cLHS design was used to establish field sampling locations that represented the variability in the feature space of available covariate layers. Finally, two kriging routines were compared to determine the most effective method of predicting soil properties in these systems.

2.3. Data preprocessing

A DEM derived from interferometric synthetic aperture radar (IFSAR) with 5 m spatial resolution was available for the area surrounding the soil survey polygon and extended beyond approximately 95% of the watershed boundaries at the hydrologic unit code (HUC) 12 level. Watershed extents not covered by the IFSAR data were supplemented with National Elevation Datasets (NED) with a 10 m spatial resolution. NED data were re-sampled to a 5 m spatial extent and combined with the IFSAR data using the Mosaic Wizard in ERDAS Imagine Software version 9.3 (Leica Geosystems, 2008) and clipped to the extent of watershed boundaries. The resulting elevation dataset with watershed extent was prepared for topographic modeling by filling sinks using ArcGIS 9.3 (Environmental Systems Research Institute, 2008). Total curvature was computed with ArcGIS and subsequent analyses of topographic parameters were performed using the SAGA Graphical User Interface — Version 2.0.4 (Conrad, 2006). Terrain analysis was performed with the parallel processing module using a multiple flow direction algorithm (Freeman, 1991) to compute slope and the SAGA wetness index (Boehner et al., 2002). Solar radiation was calculated with the incoming solar radiation module for one year on a 14 day time step using SAGA (Wilson and Gallant, 2000).

Two adjacent Landsat 7 ETM + images collected September 12, 2000 were obtained from the USGS Global Visualization Viewer (path/row 35/37 and 35/38). Data were level 1G products with radiometric and geometric corrections. Each scene was projected to NAD83 UTM Zone 12 North before processing. Scenes were combined using the Mosaic Wizard in ERDAS Imagine Software version 9.3 (Leica Geosystems, 2008) and extracted at the extent of the survey area. Bands 1, 2, 3, 4, 5, and 7 were further processed with panchromatic sharpening using a high pass filter resolution merge of Landsat band 8 to achieve a 14.25 m spatial resolution (Leica Geosystems, 2008) and subsequent re-sampling to 5 m resolution to match the spatial resolution of the elevation dataset. Resulting Landsat bands were atmospherically corrected for simple Rayleigh scattering using the Second Simulation of a Satellite Signal in the Solar System (6S) radiative transfer code web interface (http://modis-sr.ltdri.org/code.html). This included a correction for elevation and did not account for the atmospheric profile or include aerosol information (Levi and Rasmussen, 2011). Reflectance indices representative of soil, vegetation, and geology captured with Landsat band ratios 3/2, 7/3, 3/1, 5/4, 7/5, a calcareous sediment index $(5 - 2)/(5 + 2)$, gypsic index $(5 - 7)/(5 + 7)$, natric index $(5 - 4)/(5 + 4)$, and normalized difference vegetation index (NDVI) $(4 - 3) / (4 + 3)$ (Table 1).

2.4. Data reduction

A data-driven approach was used to interpolate soil variables derived from surface reflectance and topographic parameters (Hengl et al., 2007b) (Table 1). Data reduction involved an iPCA to determine those layers contributing most to observed soil–landscape variance (Nauman, 2009). Prior to iPCA all covariate layers were standardized using a z-score:

$$Z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (1)$$

where Z_{ij} is the z-score of pixel i in layer j, x_{ij} is untransformed value of pixel i of layer j, μ_j is the mean of layer j, and σ_j is the standard deviation of layer j, prior to PCA. The standardized data were grouped into elevation and reflectance indices and each group handled separately for the initial step of the data reduction. The iPCA outputs (eigen matrix and eigenvalues) were used to calculate loading factors (R_{kp}) of each input band using the degree of correlation:

$$R_{kp} = \frac{a_{kp} * \sqrt{\lambda_p}}{\sqrt{Var_k}} \quad (2)$$

where a_{kp} is the eigenvector for band k and component p, λ_p is pth eigenvalue, and Var_k is the variance of band k in the covariance matrix (Jensen, 2005). The absolute value of loading factors for each covariate layer were summed and ranked from greatest to lowest providing a quantitative metric of the total contribution of each covariate layer to the overall variance of the dataset. The number of principal components required to reach 95% cumulative explained variance in the dataset determined the number of covariate layers to retain for subsequent iterations. The covariate layers retained were those with the greatest absolute summed loading factors ensuring that the layers that explain the most variance were retained. This was repeated until all principal components were needed to achieve 95% of cumulative variance. After processing topographic parameters and Landsat reflectance ratios separately, the final layers from each group were merged and this dataset reduced in the same manner. Covariate layers selected with iPCA included Landsat ratio 3/2, the calcareous sediment index, solar radiation, and the SAGA wetness index (Table 2). Reflectance indices captured differences in parent material, and topographic parameters represented relief and aspect controls on microclimate and vegetation patterns; thus, four of the five soil forming factors from Jenny (1941) were represented (Table 2). Final covariate layers from this iPCA were used for field sample design and modeling of soil properties.

2.5. Sampling design

The goal of the sampling design for this study was to determine the minimum number of sampling locations that could effectively represent the variability of feature space for each covariate layer while also distributing the locations across geographic space to represent all soil features in the study area. Following Minasny and McBratney (2006), a cLHS routine was used to identify sampling locations in the field using publically available MATLAB code (http://www.iamg.org/CGEditor/index.htm). A wide range of sample numbers (n = 25, 50, 100, 200, 500) were identified using the cLHS design to facilitate the most efficient use of sampling locations due to cost and time constraints. Box-and-Whisker plots of extracted covariate data from each sample size were compared to the full covariate layers and the number of sampling sites was determined by the lowest number of samples that still captured the greatest variation in the original covariate layers (mean, skewness, range, etc.). We found that 50 samples provided the smallest set of sample locations that still accurately represented the distributions of each of the original covariate layers. Due to some inaccessible locations we substituted 2 locations of the original sample with locations derived from an additional iteration of the cLHS design and added 2 locations in underrepresented areas for a total of 52 sampled locations in the study area (Fig. 1). Additional samples were not taken due to restrictions on the timing and the number of soil samples to be sampled and analyzed.

The points sampled here had a sample density of 120 ha per point, which was similar to or higher than several recent digital soil mapping efforts (Gessler et al., 1995; Li, 2010; McKenzie and Ryan, 1999; Neild et al., 2007). Webster and Oliver (1992) recommended at least 50–100 points for satisfactory variogram estimates and Hengl et al. (2007a) strongly recommended the use of regression kriging if there are more than 50 total observations and at least 10 observations per predictor used in regression to prevent over-fitting of the model. With 52 sampled locations and 4 predictor variables used in the regression, we fit the recommended constraints of variogram estimates and regression kriging.

2.6. Field sampling and laboratory analysis

Soils were sampled by genetic soil horizon from 0 to 30 cm. Field descriptions followed National Cooperative Soil Survey standards and included horizon identification, texture, diagnostic horizons, surface coarse fragments by volume determined by ocular methods, coarse fragments of each horizon, parent material, dominant vegetation cover, and landform (Schoeneberger et al., 2002). Coarse fragments were estimated in three categories where gravels (GR) were 2–75 mm in diameter, cobbles (CB) were 75–250 mm, and stones (ST) were 250–600 mm (Soil Survey Division Staff, 1993).

Sieved samples were prepared for particle size analysis with pretreatments of sodium acetate (NaAOc — pH 5) to remove soluble salts and sodium hypochlorite (NaOCl — pH 9.5) to remove organic matter (Jackson, 2005). Samples were air dried and homogenized by gently grinding with a metal spatula and a mortar and pestle. Depending on the particle size, between 0.2 and 0.1 g of homogenized

Table 1
Initial data layers used for iterative PCA data reduction in the study area in southeastern Arizona.

Index	Source	Software	Feature	Reference
3/2	Landsat	ERDAS Imagine v. 9.2	Carbonate radicals	Boettinger et al. (2008)
7/3	Landsat	ERDAS Imagine v. 9.2	Ferrous Fe	Boettinger et al. (2008)
3/1	Landsat	ERDAS Imagine v. 9.2	Fe oxide	Leica Geosystems (2008)
5/4	Landsat	ERDAS Imagine v. 9.2	Ferrous	Leica Geosystems (2008)
7/5	Landsat	ERDAS Imagine v. 9.2	Clay; hydroxides	Boettinger et al. (2008) and Leica Geosystems (2008)
Calcareous sediment index	Landsat	ERDAS Imagine v. 9.2	Calcareous sediment; igneous rocks	Boettinger et al. (2008)
Gypsic index	Landsat	ERDAS Imagine v. 9.2	Gypsiferous soils	Neild et al. (2007)
Natric index	Landsat	ERDAS Imagine v. 9.2	Natric soils	Neild et al. (2007)
NDVI	Landsat	ERDAS Imagine v. 9.2	Vegetation	Huete et al. (1985)
Curvature	IFSAR	ArcGIS v. 9.3	Water and sediment flux	Moore et al. (1991)
SAGA wetness index	IFSAR	SAGA GIS v. 2.0.4	Water table depth; evapotranspiration	Boehner et al. (2002) and Freeman (1991)
Solar radiation	IFSAR	SAGA GIS v. 2.0.4	Energy input; available moisture	Wilson and Gallant (2000)
Slope percentage	IFSAR	SAGA GIS v. 2.0.4	Runoff and soil loss; soil thickness	Freeman (1991)

Table 2
Final data layers resulting from iterative PCA data reduction in the study area in southeastern Arizona.

Index	Landscape feature or process	Soil forming factor represented
Landsat 3/2	Carbonate radicals, red alluvial fans	Parent material
Calcareous sediment index	Mafic vs felsic parent material	Parent material
Solar radiation	Aspect, available moisture, vegetation	Climate, organisms, relief
SAGA wetness index	Landform, water/sediment flux	Climate, organisms, relief

soil was weighed into 15 ml auto-sampler tubes and dispersed first with deionized water using an automatic rotator for 24 h and second with 5 ml of 5% sodium hexametaphosphate ($(NaPO_3)_6$) and rotated for an additional 24 h to ensure dispersion of soil particles. After dispersion, the samples were processed using a Beckman Coulter LS 13 320 Laser Diffraction Particle Size Analyzer and USDA equivalent sand, silt, and clay fractions obtained from the results. Loss on ignition (LOI) was performed as a proxy for soil organic matter by heating samples to 360 °C in a muffle furnace for a 2 h combustion (Konen et al., 2002). Munsell soil color was determined on sieved soil using a Minolta CR-200 handheld digital chromameter (Minolta Camera Co., Ltd., Osaka, Japan). Soil redness rating (RR) was determined as:

$$RR = \frac{(10-\text{Hue})^{*}\text{Chroma}}{\text{Value}} \quad (3)$$

where Hue, Chroma, and Value are derived from Munsell soil color (Torrent et al., 1983).

2.7. Soil prediction models

Soil prediction models of surface soil properties were developed from the 52 sampled locations. Prediction model development was performed with both ArcGIS 9.3 and the statistical programming language R version 2.14.0 (R Development Core Team, 2011). A logit transformation was performed using the 'boot' package in R (Canty and Ripley, 2011) to approximate a normal distribution for the non-normally distributed soil property data where (Hengl et al., 2004):

$$z_{++} = 1n\left(\frac{z^{+}}{1-z^{+}}\right); 0<z^{+}<1 \quad (4)$$

and z_{++} is the logit transformed variable, z^{+} is the target variable standardized to the 0 to 1 range:

$$z_{+} = \frac{z-z_{\min}}{z_{\max}-z_{\min}}; z_{\min}<z<z_{\max} \quad (5)$$

and $z_{\min}$ and $z_{\max}$ are the physical minimum and maximum values of z. The physical values of each variable were within and not equal to the minimum and maximum values to prevent ln(0) situations in the logit transformation. Percent sand, silt, clay, LOI, and coarse fragments were reported with values between 0 and 1. A value of 0.1 was added to all fractions of coarse fragments prior to transformation to prevent ln(0) situations of the sites with no coarse fragments, and a range of 0–15 was used for RR, as this reflects the possible range of values.

Shapefiles of point data attributed with measured soil variables were imported to R using the 'sp' package (Bivand et al., 2008) in preparation for regression and kriging. Ordinary kriging of both the logit-transformed variables and the residuals resulting from stepwise linear regression was performed using the 'gstat' package of R (Pebesma and Wesseling, 1998). The 'gstat' package cannot automatically estimate anisotropy parameters when modeling the variogram. Therefore, we determined variogram anisotropy using ArcGIS and applied the information to variogram modeling in R. Variogram models that minimized the root mean square error (RMSE) and had a standardized RMSE closest to 1.0 from cross validation in ArcGIS were selected to provide inputs to variogram modeling in R.

Modeling of soil properties was performed with regression kriging using the selected covariate layers. Regression kriging results were both visually and quantitatively compared to ordinary kriging, as it is one of the most common geostatistical approaches used in environmental landscape studies (Li and Heap, 2011). Regression kriging models were developed using principal components of the final covariate layers used in the sampling design (RK). Regression kriging was used instead of cokriging to avoid the complexity of making predictions beyond bivariate predictions. Further, the use of principal components reduces potential error introduced through multi-collinearity of predictor variables (Hengl et al., 2003).

Initial evaluation of soil prediction models using regression kriging of principal components of the final four covariate layers indicated that strong aspect differences were introduced by the solar radiation information and estimated differences in soil properties that were highly unlikely. For example, predicted values of clay and sand percentage on north- and south-facing slopes on hills of the same parent material were 10–20% different for soils that were less than 100 m away. Based on these spurious predictions, solar radiation was removed from the set of covariate layers and a PCA of the remaining three covariates were used as predictors, i.e., Landsat band 3/2, calcareous sediment index, SAGA wetness index. Relationships between soil properties and the three remaining covariates were extracted using backward stepwise linear regression as the first step in regression kriging using the 'MASS' package in R (Venables and Ripley, 2002). Model selection was determined by minimizing the Akaike Information Criterion (AIC) (Akaike, 1974). Principal components of covariate layers were the linearly uncorrelated variables used to predict soil properties. Prior to applying the regression equations to the raster data, areas representing cattle ponds were masked out using ArcGIS to remove pixels representing surface water.

Regression model residuals were interpolated using ordinary kriging. Residual variogram development was performed as above, using a combination of ArcGIS to determine variogram anisotropy and the 'gstat' package in R to perform ordinary kriging of the residuals. Kriged residuals were added to regression model results for final prediction maps.

2.8. Model validation

Model validation was performed with leave-one-out cross validation and comparison of the predicted values at interpolation points (Pebesma and Wesseling, 1998). Normalizing measures of model performance is also useful for comparing relative prediction error for transformed variables for which the variance cannot simply be back-transformed (Hengl et al., 2004); thus, logit transformed variables were used for the cross validation to determine normalized mean square error (NMSE)

$$\text{NMSE} = \frac{\frac{1}{n}\sum_{i=1}^{n}(p_i-o_i)^2}{s^2} \quad (6)$$

where n is the number of observations, p_i is the predicted value at location i, o_i is the observed value at location i, and s^2 is the variance of the observed samples (Li and Heap, 2011). A Pearson rank correlation

Table 3
Pearson correlation coefficients of measured soil properties of surface soils at 52 locations and candidate auxiliary data layers applied to iterative PCA data reduction. Shaded rows correspond to auxiliary data layers selected with the iterative PCA data reduction. Values in bold are significant at the $\alpha = 0.05$ level.

	Sand[a]	Silt	Clay	GR	CB	CF_total	RR	LOI	Mean_abs
LS 3/1[b]	**0.55**	**−0.33**	**−0.53**	−0.14	**−0.38**	**−0.43**	**0.81**	**−0.60**	0.47
LS 3/2	**0.46**	**−0.29**	**−0.44**	−0.15	−0.27	**−0.34**	**0.79**	**−0.51**	0.41
LS 5/4	**0.46**	−0.17	**−0.51**	−0.17	−0.21	**−0.30**	**0.41**	**−0.38**	0.32
LS 7/3	**0.56**	−0.20	**−0.62**	−0.24	**−0.41**	**−0.52**	**0.44**	**−0.53**	0.44
LS 7/5	**0.53**	**−0.30**	**−0.52**	−0.12	**−0.30**	**−0.33**	**0.45**	**−0.53**	0.39
Calc_sed	**0.50**	−0.15	**−0.56**	**−0.28**	**−0.41**	**−0.56**	**0.65**	**−0.48**	0.45
Gypsic	**−0.54**	**0.32**	**0.52**	0.10	**0.28**	**0.31**	**−0.44**	**0.53**	0.38
Natric	**0.48**	−0.20	**−0.51**	−0.15	−0.20	**−0.28**	**0.41**	**−0.37**	0.33
NDVI	−0.26	0.23	0.22	−0.04	−0.04	−0.06	−0.26	0.22	0.17
Slope	−0.21	−0.09	**0.32**	**0.29**	**0.34**	**0.51**	**−0.37**	**0.37**	0.31
WI	0.27	0.03	**−0.37**	**−0.35**	**−0.39**	**−0.60**	**0.41**	**−0.38**	0.35
Curv	−0.02	−0.17	0.12	0.03	0.24	0.23	0.00	0.10	0.11
S_rad	−0.07	0.15	0.01	−0.22	−0.16	**−0.30**	0.20	−0.16	0.16

[a]Sand, Silt, Clay, GR, CB, and CF_total represent percent sand, silt, clay, gravel, cobble, and total coarse fragments; RR is redness rating derived from Munsell soil color; LOI is loss on ignition; Mean_abs is the mean of absolute values of correlations for each candidate auxiliary data layer.
[b]LS 3/1, LS 3/2, LS 5/4, LS 7/3, and LS 7/5 represent Landsat band ratios; Calc_sed is the calcareous sediment index; Gypsic is the gypsic index; Natric is the natric index; NDVI is the normalized difference vegetation index; Slope is percent slope; WI is the SAGA wetness index; Curv is total curvature; and S_rad is solar radiation.

coefficient was used to compare the observed and predicted values of each variable at all 52 locations using the leave-one-out cross validation.

Patterns of predicted soil properties were compared to published soil survey data as an additional validation of the soil prediction models. Although the goal of this project was to provide detailed raster predictions of soil properties across the landscape, we also wanted to ensure that we captured the general patterns of soil variability represented in the vector soil survey data. Mean values of sand, silt, and clay were extracted from OK and RK predictions for each map unit polygon and compared to the mean representative value for each polygon reported in SSURGO (surface horizons) using a paired *t*-test. The same representative values of surface soil properties from SSURGO were presented for spatial comparisons of predictions.

3. Results

3.1. Relationships between soil properties and covariate data

Candidate covariate layers applied to the iPCA data reduction showed varying degrees of correlation with the eight measured soil properties (Table 3). Of the 13 candidate covariates, 11 were significantly correlated with CF_total, 10 with clay, CF_total, and RR, and 8 with sand. In general, Landsat indices had stronger correlations with measured properties than topographic parameters. Summing across all measured variables, the strongest correlations were found for Landsat 3/1 and the calcareous sediment index and weakest correlations were found for curvature and solar radiation. Covariate layers selected with the iPCA demonstrated significant correlation with measured physical soil properties (Table 4). Both Landsat ratio 3/2 and the calcareous sediment index were significantly correlated with seven of the eight measured soil properties. The SAGA wetness index was significantly correlated with six of the eight properties while solar radiation only showed significant correlation with one soil property. Weak correlation of solar radiation to soil properties coupled with the unrealistic predictions of soil properties resulting from the strong aspect differences resulted in eliminating solar radiation as a predictor variable. The strongest correlation between covariate layers and soil properties was between Landsat ratio 3/2 and RR ($r = 0.79$). This was followed by RR and the calcareous sediment index (0.65) and CF_total and SAGA wetness index (0.60).

Soil properties showed significant correlations with one another, as CB, clay, silt, RR, and LOI each shared significant correlations with six of the other seven soil properties (Table 4). GR shared the fewest number of significant correlations with other properties. Sand and clay shared the strongest correlation of all measured properties ($r = -0.92$) followed by a strong negative correlation between sand and LOI (-0.73).

Covariate layers demonstrated moderate correlation, in particular, between Landsat ratio 3/2 and the calcareous sediment index ($r = 0.71$). A final PCA was performed on the three covariate layers, Landsat ratio 3/2, calcareous sediment index, and SAGA wetness index, to address problems associated with multi-collinearity of predictor variables in regression model development, with resulting low correlation coefficients between the principal components (Table 4).

3.2. Performance of coupled iPCA–cLHS design

The spatial patterns of landscape variability captured by the iPCA data reduction demonstrated strong visual correspondence with the published soil survey (Fig. 1 and Table 5) indicating the iPCA captured soil–landscape variation as described in the mapping process. Furthermore, the cLHS design based on iPCA output produced a spatial sampling scheme that well represented the spatial variability of soil survey map units (Fig. 1). The cLHS design stratified the sampling locations randomly in feature space and the resulting spatial structure is geographically dispersed, as determined by a nearest neighbor ratio (observed mean distance/expected mean distance) of 1.19 for $n = 52$ points (p-value = 0.0072). Strong correspondence of the spatial patterns of soil map units to cLHS selected sample locations indicated the combination of iPCA and cLHS may serve as effective tools for soil sample design for both soil survey and digital soil mapping. The sampling design also captured a wide range of soil types, as reflected in the variability in measured soil properties (Table 6). Sand and clay had the widest range of values with ranges of >60%. Silt had the lowest variability, as indicated by a low coefficient of variation, whereas coarse fragments had high variability.

Table 4
Pearson correlation coefficients of measured soil properties of surface soils at 52 locations and auxiliary data layers. Values in bold are significant at the $\alpha = 0.05$ level. Shaded rows and columns correspond to principal components of Landsat ratio 3/2, calcareous sediment index, and SAGA wetness index.

GR[a]														
-0.24	CB													
0.57	0.66	CF_total												
-0.06	0.37	0.25	Clay											
-0.3	0.41	0.11	0.35	Silt										
0.18	-0.5	-0.24	-0.9	-0.7	Sand									
-0.05	-0.4	-0.36	-0.4	-0.4	0.52	RR								
-0.09	0.62	0.45	0.63	0.58	-0.7	-0.6	LOI							
-0.3	-0.4	-0.57	-0.5	-0.18	0.49	0.74	-0.54	PC1						
-0.15	-0.04	-0.15	0.16	0.32	-0.26	-0.4	0.19	-0.2	PC2					
0.13	0.17	0.24	0.22	-0.11	-0.12	0.08	0.0	-0.1	-0.1	PC3				
-0.15	-0.3	-0.34	-0.4	-0.3	0.46	0.79	-0.51	0.87	-0.6	0.26	LS 3/2			
-0.3	-0.4	-0.56	-0.6	-0.15	0.5	0.65	-0.48	0.92	-0.3	-0.5	0.71	Calc_sed		
-0.22	-0.16	-0.3	0.01	0.15	-0.07	0.2	-0.16	0.37	0.03	-0.05	0.28	0.33	S_rad	
-0.4	-0.4	-0.6	-0.4	0.03	0.27	0.41	-0.38	0.77	0.43	-0.03	0.46	0.61	0.36	WI

[a]GR, CB, and CF_total, Clay, Silt, and Sand represent percent gravel, cobble, total coarse fragments, clay, silt, and sand; RR is redness rating derived from Munsell soil color; LOI is loss on ignition; PC1, PC2, and PC3 are principal components of LS 3/2, Calc_sed, and WI; LS3/2 is Landsat ratio of band 3 over band 2; Calc_sed is the calcareous sediment index; S_rad is solar radiation; WI is the SAGA wetness index.

3.3. Ordinary kriging

Modeled soil properties exhibited varying degrees of spatial dependence (Table 7). Ordinary kriging of the logit transformed data produced generalized predictions of soil properties with smooth transition of high to low values. For example, ordinary kriging of clay produced generalized predictions that captured major trends in the study area, namely the high clay content of the basalt hills region in the western portion of the study area (Fig. 2). Predictions of sand with ordinary kriging were highest in the granitic alluvial soils in the eastern portion of the study area (Fig. 3). The ordinary kriging predictions of silt looked very different from clay and sand (Fig. 4), as the anisotropy present in the spatial structure of the logit-transformed silt variable produced a complex pattern of high and low estimates of silt content that likely did not reflect actual distributions of silt in the study area. This high level of anisotropy was likely a result of poor spatial dependency of silt in the study area and the lack of sufficient landscape-scale patterns of silt content.

Spatial autocorrelation of clay modeled with ordinary kriging had a range of 9.4 km compared to 11.7 km for sand and 5.9 km for silt. Sand had a higher nugget than both clay and silt suggesting more measurement error or micro-scale variation at spatial scales too fine to detect in this study (Table 7). The distance of spatial dependence for GR was 12.6 km and 9.8 km for CB, while the range of total coarse fragments (CF_total) was an average of the two components at 10.4 km. The range of the spatial dependence for RR was 10.5 km for ordinary kriging and 5.9 km for LOI. Relatively small nugget effects were similar for the semivariograms of LOI suggesting most of the fine scale spatial structure was captured with the distribution of sample locations.

Expressing nugget semivariance as a fraction of the total semivariance provides a way to classify the spatial dependence of soil variables by identifying the proportion of short range variability that cannot be described by the statistical model based on the variogram. Cambardella et al. (1994) suggested the following classes of spatial dependence: strong spatial dependence if the ratio was ≤0.25, moderate for ratios between 0.25 and 0.75, and weak for ratios >0.75. Using this classification, the logit-transformed variables of clay and sand showed moderate spatial dependence and silt had strong spatial dependence (Table 7). GR, CB, CF_total, and LOI had moderate spatial dependence and RR showed a strong spatial dependence.

Evaluation of the range of predicted values suggests that the ordinary kriging approach did not capture the range of measured values, as the measured values of sand ranged from 1 to 75% and the prediction map only estimated values up to 64%. Measured values of clay ranged from 9 to 61% and the maximum predicted value from ordinary kriging was 48%. Predicted values for silt were very similar to the measured values. Though predictions from ordinary kriging provided adequate information for generalized soil property maps, they lacked the detail necessary for high-resolution hydrology and landscape modeling.

3.4. Regression

Multiple linear regression models indicated significant relationships between measured soil properties and covariate layers although adjusted R^2 values did not show particularly strong predictability (Table 7). The regression between the RK model and RR showed the best relationship relative to all other regression models. GR and silt exhibited the weakest relationships with covariate layers, whereas RR, LOI and total coarse fragment content exhibited higher R^2.

3.5. Regression kriging

Predictions of soil properties using RK represented landscape features present on the landscape and captured detailed variation in soil properties that aligned well with the soil survey map units (Figs. 2, 3, and 4). The prediction map of clay from the RK model provided detailed patterns of clay estimates that corresponded to landscape features. High estimates of clay corresponded to the basalt hills in the western portion of the study area, whereas low estimates were closely related to soil map units representing granitic alluvial fans and drainage networks in the eastern and central portions of the study area. The RK model of sand also produced detailed estimates of sand content that correspond with realistic patterns in the study area (Fig. 3). High

Table 5
Comparison of sample distribution from the conditioned Latin Hypercube design to stratification of published soil map units.

MU[a]	Area (ha)	# pts	% area	% pts
145	315	2	0.05	0.04
175	23	0	0.00	0.00
235	1656	15	0.26	0.29
255	492	5	0.08	0.10
260	160	1	0.03	0.02
275	30	1	0.00	0.02
280	2	0	0.00	0.00
300	942	9	0.15	0.17
320	1845	14	0.29	0.27
325	799	5	0.13	0.10
Sum	6265	52	1	1

[a] MU is soil map unit; # pts represents the number of sampled locations within each map unit; % pts represents the percentage of sampled locations in each map unit.

Table 6
Summary statistics for measured soil properties of surface soils at 52 locations.

	Clay[a]	Silt	Sand	GR	CB	CF_total	LOI	RR
	%							
Min	8.90	13.00	1.00	0.00	0.00	0.00	0.01	0.20
Max	70.10	46.80	75.40	55.00	50.00	55.00	0.06	4.70
Median	16.20	26.10	58.10	5.00	0.00	12.50	0.02	1.40
Range	61.20	33.90	74.50	55.00	50.00	55.00	0.05	4.40
Std	16.40	9.00	21.30	12.70	13.60	16.40	0.01	1.10
Skewness	1.50	0.50	−0.90	1.70	1.80	0.50	1.00	1.00
Kurtosis	4.00	2.20	2.50	5.50	4.90	1.80	3.30	3.30
CV	0.71	0.33	0.43	1.19	1.79	0.89	0.61	0.64

[a] Clay, Sand, Silt, GR, CB, and CF_total represent percent sand, silt, clay, gravel, cobble, and total coarse fragments, LOI is loss on ignition, and RR is redness rating derived from Munsell soil color.

Table 7
Semivariogram model parameters of two kriging methods for logit-transformed surface soil properties using 52 sample points.

Property[a]	Method[b]	Model	Nugget	psill	range	Nug:sill	Adj. R^{2c}	p-Value[d]
Clay	OK	Sph	0.28	0.38	9368	0.42	–	–
Sand	OK	Sph	0.74	0.87	11,671	0.46	–	–
Silt	OK	Sph	0.04	0.16	5854	0.18	–	–
GR	OK	Sph	2.26	2.06	12,592	0.52	–	–
CB	OK	Sph	3.00	3.87	9768	0.44	–	–
CF_total	OK	Sph	1.20	1.87	10,420	0.39	–	–
RR	OK	Sph	0.08	0.49	10,474	0.14	–	–
LOI	OK	Sph	0.20	0.11	5926	0.65	–	–
Clay	RK	Sph	0.32	0.20	18,415	0.62	0.34	<0.001
Sand	RK	Sph	0.75	0.46	19,482	0.62	0.21	0.001
Silt	RK	Sph	0.09	0.13	9556	0.42	0.09	0.018
GR	RK	Sph	2.18	1.98	14,098	0.52	0.02	0.157
CB	RK	Sph	1.74	2.84	3852	0.38	0.24	<0.001
CF_total	RK	Sph	1.19	0.40	19,750	0.75	0.39	<0.001
RR	RK	Sph	0.09	0.17	11,445	0.36	0.51	<0.001
LOI	RK	Sph	0.20	0.13	16,332	0.61	0.37	<0.001

[a] Clay, Sand, Silt, GR, CB, and CF_total represent percent sand, silt, clay, gravel, cobble, and total coarse fragments, LOI is loss on ignition, and RR is redness rating derived from Munsell soil color.
[b] OK is ordinary kriging of logit-transformed variables and RK is for the residuals of regression kriging with PCs as predictors.
[c] Adj. R^2 is adjusted R^2 values resulting from backward step-wise multiple linear regression of surface soil properties modeled with RK.
[d] p-Value is from backward step-wise multiple linear regressions.

estimates of sand were in the granitic alluvial fans on the eastern portion of the study area and in the drainage networks throughout the center of the study area. These predictions independently captured differences in sand content that corresponded well with the soil map units.

Variograms of sand and clay residuals from RK had a lower sill than the original variables (Table 7) indicating a smaller variance in the residuals, relative to the original variables. There was limited gain in predictive power of regression kriging models over ordinary kriging for GR indicated by the weak regression models and the similar semivariograms for each model (Table 7). Though the spatial dependence was moderate, the nugget effect was high (>2) suggesting much of the spatial variability of GR was at scales too fine to detect.

The nugget:sill ratio indicated moderate spatial dependence for regression residuals of clay, sand, and silt for RK. With the exception of percent GR and CB, the nugget:sill ratio increased for variograms of regression residuals, suggesting the regressions removed a considerable portion of the spatial dependence from the original variables. Residuals from the RK regression showed only a moderate spatial dependence (Table 7).

Predicted values for RK effectively represented the ranges of measured values. Measured values of sand ranged from 1 to 75% and predictions from RK ranged from 9 to 81%. Measured values of clay ranged from 9 to 61% and predicted values from RK ranged from 7 to 72%. Predicted values for silt were also similar to the measured values for RK methods.

3.6. Goodness-of-fit

Goodness-of-fit from leave-one-out cross validation indicated that OK had the highest correlation between predicted and observed values for seven of the eight measured properties (Fig. 5). Only GR was predicted better with the RK method. The mean R^2 for the cross

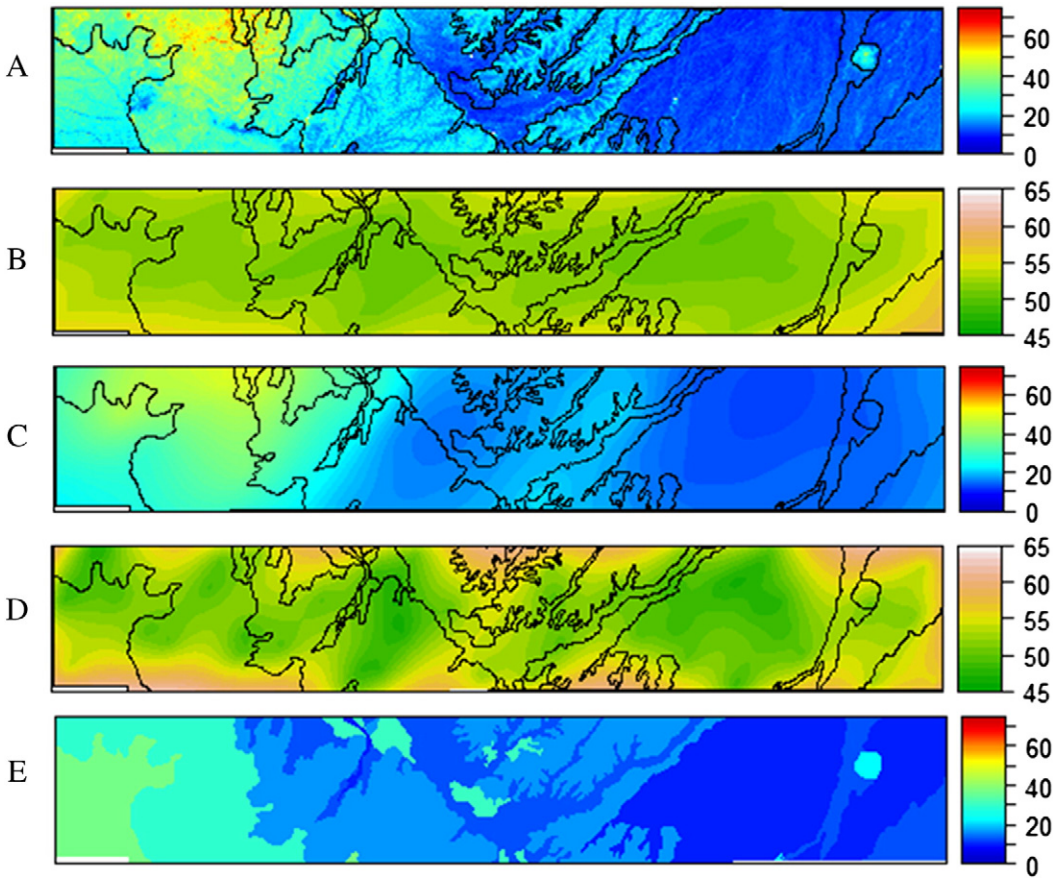

Fig. 2. Prediction maps and relative prediction error of clay produced with regression kriging using principal components of covariate layers (A, B) and ordinary kriging (C, D). Black lines represent soil map unit boundaries. Panel E represents the weighted average of clay content for surface soil horizons in all map unit components derived from the USDA SSURGO data product.

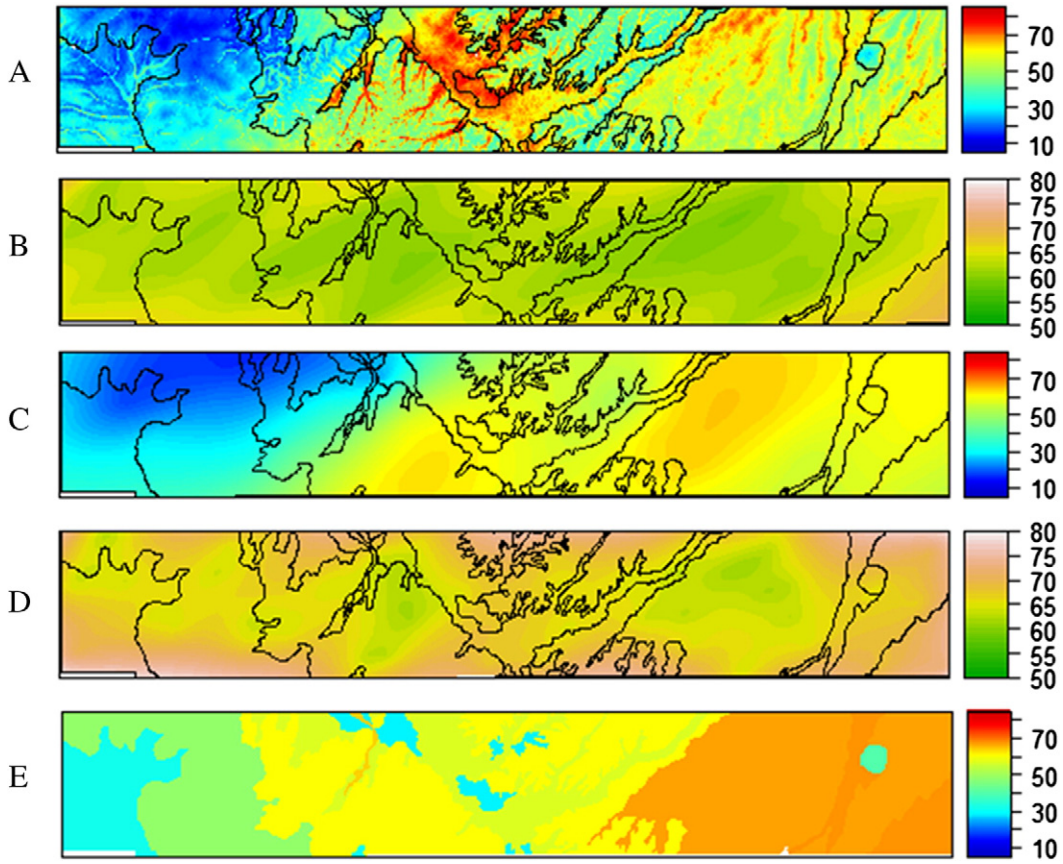

Fig. 3. Prediction maps and relative prediction error of sand produced with regression kriging using principal components of covariate layers (A, B) and ordinary kriging (C, D). Black lines represent soil map unit boundaries. Panel E represents the weighted average of sand content for surface soil horizons in all map unit components derived from the USDA SSURGO data product.

validation was 0.50. The NMSE provided a metric to compare the bias of model prediction for all methods. The lowest NMSE was achieved with RK for seven of the eight properties (Fig. 5); higher correlation generally corresponded to a lower NMSE.

The relative prediction error for each soil property–prediction method combination reflects the proportion of variance in each respective dataset. For clay estimates, the maximum relative prediction error was smaller for RK than OK (Fig. 2). The lowest error for OK was in areas close to sampled locations, whereas error associated with the RK prediction was more evenly distributed across the entire study area. Prediction error showed a similar trend for sand (Fig. 3) with lower error in the RK prediction relative to OK. The relative prediction error for silt was more widely distributed from the OK model than for the RK prediction (Fig. 4).

Predicted values of sand, silt, and clay were comparable to surface texture data reported in SSURGO map unit polygons (Table 8). Results from OK illustrated no significant differences between modeled values of sand, silt, or clay and representative values reported in SSURGO. RK predictions of sand and clay were not significantly different from SSURGO data; however, mean predicted values of silt were significantly different from values reported in SSURGO map units.

4. Discussion

4.1. Selected covariate data

Comparison of the candidate covariates for iPCA to measured soil property values provided insight that was helpful for selecting covariates. The four covariates selected with iPCA represented a subset of the candidate layers that captured a range of soil–landscape features in our study area. Although solar radiation was selected with the iPCA data reduction, it showed very weak correlations with all measured properties and we elected to remove solar radiation due to unrealistic predictions of soil properties. A simple correlation of measured soil properties with covariates could provide a means to sort the predictability of properties of interest with regression; however, selection of covariates would still have to be based on expert opinion or some arbitrary threshold. One benefit of the iPCA approach is that it provides a clear method of determining the number of covariates to retain for prediction models. One interesting result of comparing all covariates to the soil properties of interest is that RR and LOI shared very similar correlations with the covariates which illustrates the well-established relationships between soil color and soil organic matter content.

All measured soil properties had significant correlations with at least one of the covariate layers selected with the iPCA. Though the strongest relationships were found between RR and auxiliary data, moderate correlations between LOI, sand, silt, clay, and CF_total suggested the iPCA technique performed well to identify important covariate layers for digital soil mapping techniques. Similar to our findings, Csillag et al. (1993) found that a stepwise PCA was useful for identifying covariates for classifying the salinity status of soils from California and Hungary. These data reduction methods may be more useful for digital soil mapping applications than band selection methods for image visualization such as the optimum index factor (Chavez et al., 1982) or the Scheffield index (Sheffield, 1985) because the data reduction methods can easily be applied to select more than the three bands selected for red, green, and blue visualization.

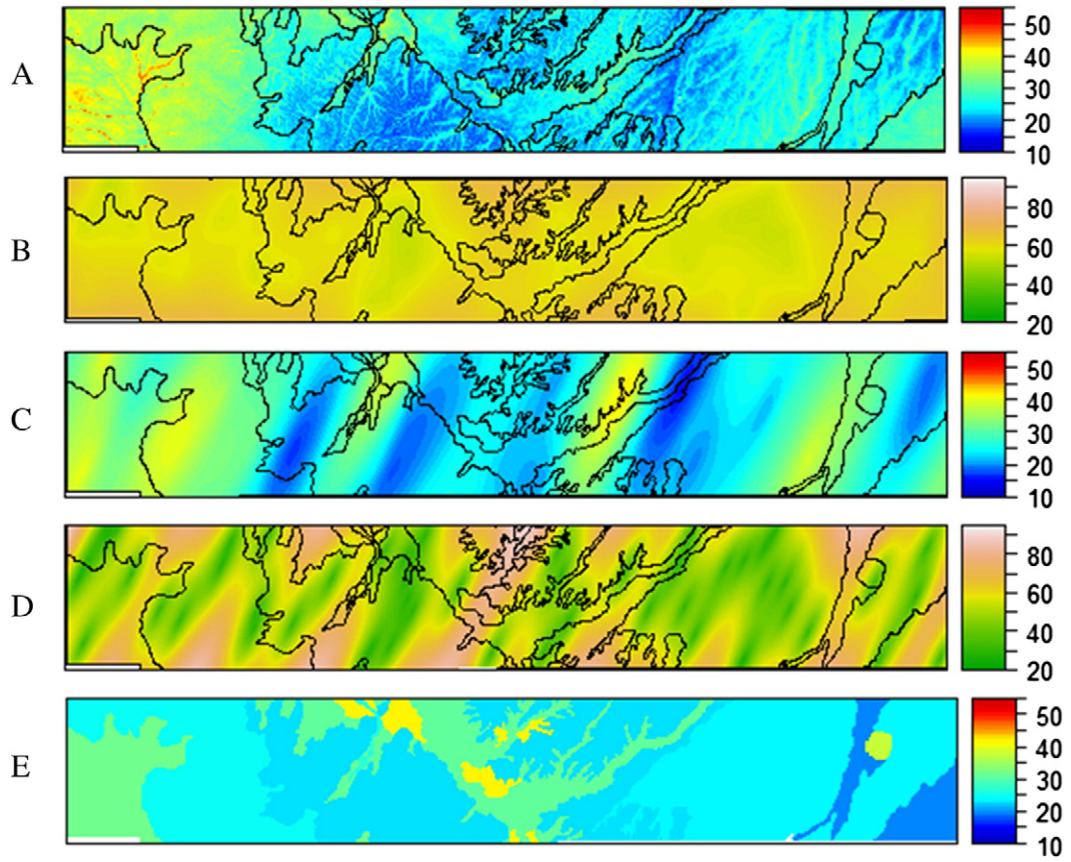

Fig. 4. Prediction maps and relative prediction error of silt produced with regression kriging using principal components of covariate layers (A, B) and ordinary kriging (C, D). Black lines represent soil map unit boundaries. Panel E represents the weighted average of silt content for surface soil horizons in all map unit components derived from the USDA SSURGO data product.

More complex soil prediction models have been shown to increase predictive power (Motaghian and Mohammadi, 2011); however, obtaining adequate high-quality data to aid prediction is likely more important than employing more complex prediction techniques (Minasny and McBratney, 2007). The iPCA–cLHS approach optimized sampling locations and maximized the likelihood of developing successful soil prediction models with a minimal number of observations.

4.2. cLHS design

The cLHS sampling scheme effectively captured the spatial variability of soils in the study area and provided the foundation for prediction of soil properties with both ordinary kriging and regression kriging approaches. Sampling designs can optimize locations for different facets of the geostatisical process including variogram estimation (Bogaert and Russo, 1999) or kriging (van Groenigen, 2000) where kriging requires evenly dispersed sample locations and variogram estimation requires a range of short and long distances between points (Marchant and Lark, 2007). This is because samples that are close in feature space tend to be close geographically (Hengl et al., 2003). Although statistically dispersed, the geographic distribution of sample locations includes a wide range of distances between points with a random distribution on the landscape. Furthermore, the sample locations represented the equivalent of a stratified random design with respect to the area of published soil map units. The distribution of points in feature space, geographic space, and proportionally across the soil map units indicate cLHS was an effective sample design for prediction of soil attributes across the study area.

4.3. Regression kriging vs. ordinary kriging and regression

RK produced estimates of soil properties that corresponded to the landscape features and soil map units present in the study area and had the lowest NMSE for seven out of eight modeled properties and moderate correlations of observed and predicted values. Ordinary kriging had higher NMSE; however, the adjusted R^2 of observed and predicted values was higher for seven of eight properties. Some studies have found that regression kriging outperforms both non-spatial and pure geostatistical methods (Odeh et al., 1994, 1995) while others have found minimal improvement using a regression kriging approach (Eldeiry and Garcia, 2010; Li, 2010). Prediction of soil properties using non-spatial models other than regression have also been improved by kriging residuals (Motaghian and Mohammadi, 2011; Scull et al., 2005). Landscape patterns delineated with RK were similar to those delineated by the published soil survey verifying that general soil patterns were captured. Furthermore, RK provided detailed spatial information of within map unit variability not currently captured in available soil survey data.

OK likely had better predictions than RK because the spatial autocorrelation of logit-transformed variables was greater than the correlation between the variables and the covariates (Eldeiry and Garcia, 2010). Performance of individual techniques is largely determined by the local or regional relationships that exist between covariate layers and soil properties, sample locations, and the choice of prediction method. In our study, the regression of measured soil properties and PCs visually separated landscape features; however, the regression models had a relatively low R^2. The combination of regression with kriging of

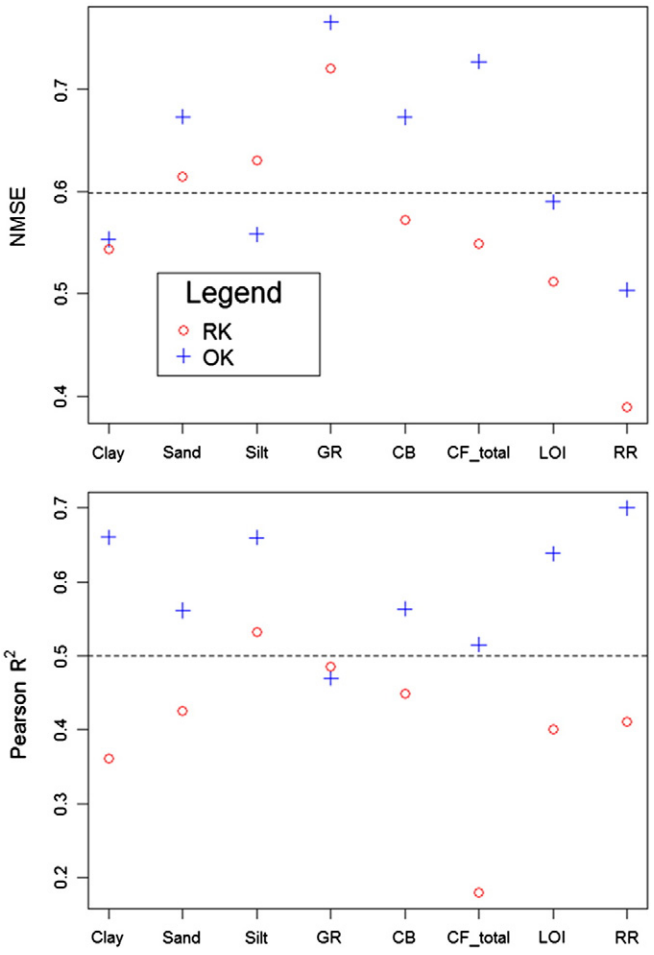

Fig. 5. Comparison of goodness-of-fit of surface soil properties modeled with regression kriging with PCs as predictors (RK) and ordinary kriging (OK) using Pearson correlation and the NMSE. Clay, Sand, Silt, GR, CB, and CF_total represent percent sand, silt, clay, gravel, cobble, and total coarse fragments, LOI is loss on ignition, and RR is redness rating derived from Munsell soil color. The dashed lines indicate the mean values of NMSE and R^2 for respective plots.

residuals improved predictions relative to the regression alone by accounting for spatial variability of regression model error.

5. Conclusion

Data reduction using an iPCA combined with a cLHS design produced a sampling design that effectively captured the variability of soil types as a function of the relative area of the published soil map. This minimal dataset of 52 sample locations represented the variability of both feature space and geographic space, and effectively predicted a range of soil physical properties in this 6265 ha study area, demonstrating the efficacy of the coupled iPCA–cLHS–RK approach. The detailed variation in soil properties captured with RK aligned well with soil survey map units, both spatially and in magnitude, and provided a means to characterize the spatial variability of important soil properties within map units. Improvements in the prediction model could be made with additional field sampling to better define the spatial structure of the data; however, the method presented here can optimize the distribution of sample locations in similar circumstances when time and financial resources are limited. The combination of iterative data reduction with a structured sampling design and a robust soil prediction model can incorporate a wide variety of numerically continuous covariates to improve soil sampling efforts. This approach can reduce the time and money needed to provide detailed soil information and associated errors to landscape models relevant to hydrology, agriculture, geosciences, and atmospheric sciences.

Table 8
Comparison of modeled surface sand, silt, and clay to representative values from published SSURGO data. Values are p-values from paired t-tests between SSURGO data and predictions from ordinary kriging (OK) and regression kriging (RK) by map unit. n = 27 for sand and silt and n = 28 for clay because one SSURGO map unit polygon did not report values for sand or silt. Values in bold are significant at the $\alpha = 0.05$ level.

	OK	RK
Sand	0.38	0.33
Silt	0.12	**0.02**
Clay	0.40	0.96

Acknowledgments

This research was supported by the USDA-Natural Resources Conservation Service of Arizona, Cooperative Agreement #68-9457-8-466, NSF EAR/IF #0929850, and the Arizona Agricultural Experiment Station ARZT-1367190-H21-155. The authors would like to thank the handling editor and two anonymous reviewers for their helpful comments that greatly improved the manuscript.

References

Akaike, H., 1974. New look at statistical-model identification. IEEE Trans. Autom. Control 19 (6), 716–723.
Beaudemin, M., Fung, K.B., 2001. On statistical band selection for image visualization. Photogramm. Eng. Remote Sens. 67 (5), 571–574.
Bishop, T.F.A., McBratney, A.B., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103 (1–2), 149–160.
Bishop, T.F.A., Minasny, B., McBratney, A.B., 2006. Uncertainty analysis for soil-terrain models. Int. J. Geogr. Inf. Sci. 20 (2), 117–134.
Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V., SpringerLink (Online service), 2008. Applied spatial data analysis with R, Use R! Springer, New York; London xiv (374 p).
Boehner, J., Koethe, R., Conrad, O., Gross, J., Ringeler, A., Selige, T., 2002. Soil regionalisation by means of terrain analysis and process parameterisation. In: Micheli, E., Nachtergaele, F., Montanarella, L. (Eds.), Soil Classification 2001. European Soil Bureau, Research Report No. 7, EUR 20398 EN, Luxembourg, pp. 213–222.
Boettinger, J.L., Ramsey, R.D., Bodily, J.M., Cole, N.J., Kienast-Brown, S., Nield, S.J., Saunders, A.M., Stum, A.K., 2008. Landsat spectral data for digital soil mapping. In: A.E.H.e.a (Ed.), Digital Soil Mapping with Limited Data. Springer, pp. 193–202.
Bogaert, P., Russo, D., 1999. Optimal spatial sampling design for the estimation of the variogram based on a least squares approach. Water Resour. Res. 35 (4), 1275–1289.
Brown, D.E., 1994. Biotic Communities: Southwestern United States and Northwestern Mexico. University of Utah Press, Salt Lake City.
Brown, D.E., Lowe, C.H., 1994. Biotic Communities of the Southwest. University of Utah Press, Salt Lake City.
Brungard, C.W., Boettinger, J.L., 2010. Conditioned Latin hypercube sampling: optimal sample size for digital soil mapping of arid rangelands in Utah, USA. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienest-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Production, and Environmental Application. Progress in Soil Science 2. Springer, Dordrecht; London, pp. 67–75.
Brus, D.J., deGruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80 (1–2), 1–44.
Cambardella, C.A., Moorman, T.B., Novak, J.M., Parkin, T.B., Karlen, D.L., Turco, R.F., Konopka, A.E., 1994. Field-scale variability of soil properties in central Iowa soils. Soil Sci. Soc. Am. J. 58 (5), 1501–1511.
Canty, A., Ripley, B., 2011. Boot: Bootstrap R (S-Plus) Functions.
Carre, F., Girard, M.C., 2002. Quantitative mapping of soil types based on regression kriging of taxonomic distances with landform and land cover attributes. Geoderma 110 (3–4), 241–263.
Carre, F., McBratney, A.B.., Mayr, T., Montanarella, L., 2007. Digital soil assessments: beyond DSM. Geoderma 142 (1–2), 69–79.
Chavez, P.S.J., Berlin, G.L., Sowers, L.B., 1982. Statistical method for selecting Landsat MSS ratios. J. Appl. Photogr. Eng. 8 (1), 23–30.
Chen, T., Niu, R.Q., Li, P.X., Zhang, L.P., Du, B., 2011. Regional soil erosion risk mapping using RUSLE, GIS, and remote sensing: a case study in Miyun Watershed, North China. Environ. Earth Sci. 63 (3), 533–541.
PRISM Climate Group. 2008. Oregon State University, http://www.prism.oregonstate.edu/, created 31 Oct 2008.
Conrad, O., 2006. SAGA — program structure and current state of implementation. In: Böhner, J., McCloy, K.R., Strobl, J. (Eds.), SAGA—Analysis and Modeling Applications. Verlag Erich Goltze GmbH, pp. 39–52.
Csillag, F., Pasztor, L., Biehl, L.L., 1993. Spectral band selection for the characterization of salinity status of soils. Remote Sens. Environ. 43 (3).
Di, H.J., Trangmar, B.B., Kemp, R.A., 1989. Use of geostatistics in designing sampling strategies for soil survey. Soil Sci. Soc. Am. J. 53 (4), 1163–1167.

Duffera, M., White, J.G., Weisz, R., 2007. Spatial variability of Southeastern US Coastal Plain soil physical properties: implications for site-specific management. Geoderma 137 (3–4), 327–339.
Eldeiry, A.A., Garcia, L.A., 2008. Detecting soil salinity in alfalfa fields using spatial modeling and remote sensing. Soil Sci. Soc. Am. J. 72 (1), 201–211.
Eldeiry, A.A., Garcia, L.A., 2010. Comparison of ordinary kriging, regression kriging, and cokriging techniques to estimate soil salinity using LANDSAT images. J. Irrig. Drain. Div. Am. Soc. Civ. Eng 136 (6), 355–364.
Environmental Systems Research Institute, 2008. ArcGIS version 9.3. ESRI, Redlands, CA.
Freeman, T.G., 1991. Calculating catchment-area with divergent flow based on a regular grid. Comput. Geosci. 17 (3), 413–422.
Leica, Geosystems, 2008. ERDAS Imagine version 9.2. Leica Geosystems Geospatial Imaging, LLC, Atlanta, GA.
Gessler, P.E., Moore, I.D., McKenzie, N.J., Ryan, P.J., 1995. Soil–landscape modeling and spatial prediction of soil attributes. Int. J. Geogr. Inf. Syst. 9 (4), 421–432.
Hengl, T., Rossiter, D.G., Stein, A., 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Aust. J. Soil Res. 41 (8), 1403–1422.
Hengl, T., Heuvelink, G.B.M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120 (1–2), 75–93.
Hengl, T., Heuvelink, G.B.M., Rossiter, D.G., 2007a. About regression-kriging: from equations to case studies. Compt. Rendus Geosci. 33 (10), 1301–1315.
Hengl, T., Toomanian, N., Reuter, H.I., Malakouti, M.J., 2007b. Methods to interpolate soil categorical variables from profile observations: lessons from Iran. Geoderma 140 (4), 417–427.
Huete, A.R., Jackson, R.D., Post, D.F., 1985. Spectral response of a plant canopy with different soil backgrounds. Remote Sens. Environ. 17 (1), 37–53.
Jackson, M.L., 2005. Soil Chemical Analysis: Advanced Course, 2nd ed. UW-Madison Libraries Parallel Press, Madison, WI, USA.
Jenny, H., 1941. Factors of Soil Formation; a System of Quantitative Pedology. McGraw-Hill Publications in the Agricultural Sciences, 1st ed. McGraw-Hill book company, inc., New York, London.
Jensen, J.R., 2005. Introductory Digital Image Processing: A Remote Sensing Perspective, 3rd. ed. Prentice Hall, Upper Saddle River, NJ.
Konen, M.E., Jacobs, P.M., Burras, C.L., Talaga, B.J., Mason, J.A., 2002. Equations for predicting soil organic carbon using loss-on-ignition for north central US soils. Soil Sci. Soc. Am. J. 66 (6), 1878–1881.
Lathrop, R.G., Aber, J.D., Bognar, J.A., 1995. Spatial variability of digital soil maps and its impact on regional ecosystem modeling. Ecol. Model. 82 (1), 1–10.
Levi, M.R., Rasmussen, C., 2011. Considerations for atmospheric correction of surface reflectance for soil survey applications. Soil Surv. Horiz. 52 (2), 48–55. http://dx.doi.org/10.2136/ssh2011-52-2-5.
Li, Y., 2010. Can the spatial prediction of soil organic matter contents at various sampling scales be improved by using regression kriging with auxiliary information? Geoderma 159 (1–2), 63–75.
Li, J., Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. Ecol. Inf. 6 (3–4), 228–241.
Marchant, B.P., Lark, R.M., 2007. Optimized sample schemes for geostatistical surveys. Math Geol. 39 (1), 113–134.
Maselli, F., Gardin, L., Bottai, L., 2008. Automatic mapping of soil texture through the integration of ground, satellite and ancillary data. Int. J. Remote Sens. 29 (19), 5555–5569.
McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97 (3–4), 293–327.
McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.
McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89 (1–2), 67–94.
Melton, M.A., 1965. The geomorphic and paleoclimatic significance of alluvial deposits in southern Arizona. J. Geol. 73 (1), 1–38.
Miller, D.A., White, R.A., 1998. A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling. Earth Interact. 2 (2), 1–26.
Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32 (9), 1378–1388.
Minasny, B., McBratney, A.B., 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. Geoderma 140 (4), 324–336.
Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital terrain modeling — a review of hydrological, geomorphological, and biological applications. Hydrol. Process. 5 (1), 3–30.
Motaghian, H.R., Mohammadi, J., 2011. Spatial estimation of saturated hydraulic conductivity from terrain attributes using regression, kriging, and artificial neural networks. Pedosphere 21 (2), 170–177.
Nauman, T., 2009. Digital Soil–landscape Classification for Soil Survey using ASTER Satellite and Digital Elevation Data in Organ Pipe Cactus National Monument, Arizona. (M.S. Thesis) Univ. of Arizona (169 pp.).
Neild, S.J., Boettinger, J.L., Ramsey, R.D., 2007. Digitally mapping gypsic and natric soil areas using Landsat ETM data. Soil Sci. Soc. Am. J. 71 (1), 245–252.
Neilson, R.P., 1987. Biotic regionalization and climatic controls in western North America. Vegetatio 70 (3), 135–147.
Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63 (3–4), 197–214.
Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1995. Further results on prediction of soil properties from terrain attributes — heterotopic cokriging and regression-kriging. Geoderma 67 (3–4), 215–226.
Pebesma, E.J., Wesseling, C.G., 1998. Gstat: a program for geostatistical modelling, prediction and simulation. Comput. Geosci. 24 (1), 17–31.
Peschel, J.M., Haan, P.K., Lacey, R.E., 2006. Influences of soil dataset resolution on hydrologic modeling. J. Am. Water Resour. Assoc. 42 (5), 1371–1389.
R. Development Core Team, 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (ISBN 3-900051-07-0, URL http://www.R-project.org/.).
Richard, S.M., Reynolds, S.J., Spencer, J.E., Pearthree, 2000. Geologic map of Arizona. Arizona Geological Survey, Map 35, scale 1:1,000,000.
Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.d.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vagen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.-L., 2009. Digital soil map of the world. Science 325 (5941), 680–681.
Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., Broderson, W.D.e, 2002. Field book for Describing and Sampling Soils, Version 2.0. Natural Resources Conservation Service, National Soil Survey Center, Lincoln, NE.
Scull, P., Okin, G., Chadwick, O.A., Franklin, J., 2005. A comparison of methods to predict soil surface texture in an alluvial basin. Prof. Geogr. 57 (3), 423–437.
Sheffield, C., 1985. Selecting band combinations from multispectral data. Photogramm. Eng. Remote Sens. 51 (6).
Singh, H.V., Kalin, L., Srivastava, P., 2011. Effect of soil data resolution on identification of critical source areas of sediment. J. Hydrol. Eng. 16 (3), 253–262.
Soil Survey Division Staff, 1993. Soil survey manual. Soil Conservation Service. U.S. Dept. of Agriculture Handbook No. 18, Washington, D.C.
Soil Survey Staff, 2011. Soil Survey Geographic (SSURGO) Database for Graham County, Arizona, Southwestern Part. Natural Resources Conservation Service, United States Department of Agriculture (Accessed 30 August 2011, Available online at http://soildatamart.nrcs.usda.gov).
Soil Survey Staff, 2012. Available Soil Survey Data [Map]. United States Department of Agriculture Natural Resources Conservation Service.
Tesfa, T.K., Tarboton, D.G., Chandler, D.G., McNamara, J.P., 2009. Modeling soil depth from topographic and land cover attributes. Water Resour. Res. 45.
Torrent, J., Schwertmann, U., Fechter, H., Alferez, F., 1983. Quantitative relationships between soil color and hematite content. Soil Sci. 136 (6), 354–358.
van Groenigen, J.W., 2000. The influence of variogram parameters on optimal sampling schemes for mapping by kriging. Geoderma 97 (3–4), 223–236.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, 4th ed. Springer, New York.
Wang, X., Melesse, A.M., 2006. Effects of STATSGO and SSURGO as inputs on SWAT Model's Snowmelt Simulation. J. Am. Water Resour. Assoc. 42 (5), 1217–1236.
Webster, R., Oliver, M.A., 1992. Sample adequately to estimate variograms of soil properties. J. Soil Sci. 43 (1), 177–192.
Wilson, J.P., Gallant, J.C., 2000. Terrain analysis: principles and applications. John Wiley and Sons, New York.
Wilson, E.D., Moore, T., 1958. Geologic map of Graham and Greenlee Counties, Arizona. Arizona Bureau of Mines and The University of Arizona, scale 1:375,000.
Ziadat, F.M., 2005. Analyzing digital terrain attributes to predict soil attributes for a relatively large area. Soil Sci. Soc. Am. J. 69 (5), 1590–1599.