Geosciences Journal
DOI 10.1007/s12303-014-0023-9

Re-tooling of regression kriging in R for improved digital mapping of soil properties

Christian T. Omuto* *Department of Environmental and Biosystems Engineering, University of Nairobi, P.O. Box 30197-0100, Nairobi, Kenya*
Ronald R. Vargas *FAO, Land and Water Division (NRL), Viale delle Terme di Caracalla, 00100 Rome, Italy*

ABSTRACT: Regression analysis and kriging are popular spatial estimation methods often used in soil science to provide soil information at different spatial resolutions and extent. Attempts have been made to combine them into a method known as regression kriging (RK). With the increasing acceptance of digital soil mapping paradigm, utilization of spatial estimation method such as RK is bound to rise. Although RK is versatile and popular, its current format has deficiencies which can hinder the quality of estimated soil properties. One of the deficiencies of RK is the failure of its regression model to recognize that natural soil occurs in groups with unique response characteristics to soil forming factors. Ideally, these groups should be represented as a family of curves when modelling the landscape. However, the current applications tend to use average models which either block/control the grouping effects or do not statistically recognize them. In this paper, mixed-effects modelling technique is shown for ingenious recognition of soil groupings and consequent improvement of RK accuracy. Mixed-effects modelling allows for simultaneous regression estimation for individual models in a group and for different groups in the landscape. Its implementation in RK has been illustrated using executable scripts in R. It gives better mapping accuracy and reliable maps than the current application in RK. The new RK and its easy implementation in R software are anticipated to provide potential for wide application and eventual contribution to improved soil mapping and application of DSM.

Key words: digital soil mapping, regression kriging, mixed-effects, R

1. INTRODUCTION

Digital soil mapping (DSM) is increasingly gaining worldwide acceptance as a means for fulfilling the demand for accurate soil information at different spatial resolutions and extent. Although DSM has many components, spatial estimation of soil properties is perhaps the most actively researched and applied (Lagacherie et al., 2007; Hattermink et al., 2008). The spatial estimation endeavours to produce soil maps in which variability within soil mapping units is sufficiently accounted for. In the traditional soil maps, this variability was depicted as homogeneous or was generally described but not spatially represented. Recent studies show that opportunities still exist for improving the spatial estimation of soil properties by accounting for more variability within- and between-soil mapping units (Brus et al., 2008). This paper identified this opportunity in a popularly used DSM mapping method and attempted to improve it.

Many methods exist in the literature for soil spatial estimation such as fuzzy membership, multivariate statistical methods, geostatistics, decision tree analysis, among others (Scull et al., 2005; Hengl et al., 2007). The geostatistics method seems to be popular with many researchers perhaps because it is often easily implemented in numerous available GIS software (Hengl et al., 2007). Examples of geostatistical applications include regression kriging, simple kriging, Bayesian kriging, etc. (Cui et al., 1995). In regression kriging (RK), spatial estimation involves statistical modelling of the deterministic and stochastic components of the soil variables in the landscape. The deterministic component represents the large-scale trends while the stochastic component represents the small-scale autocorrelation. The large-scale trends are usually modelled using regression analysis while the autocorrelation trends are modelled with kriging analysis (Odeh et al., 1995). The co-occurrence of regression and kriging analyses gives the name of regression kriging.

Although there are some arguments against RK approach (Lark et al., 2006), it still remains versatile, easily implementable and compatible with many modelling software (Carré and Girard, 2002; Herbst et al., 2006; Hengl et al., 2007). This aspect of RK is especially important for implementing the DSM paradigm. This paper sought to improve the accuracy of RK by targeting its regression part using mixed-effects modelling. Mixed-effects modelling is a form of regression analysis that can simultaneously model nested relationships. It is especially suitable in situations where unique relationships exist for certain individuals within a group and for different groups in a population. It has a potential in soil mapping because soil properties have unique relationships with soil forming factors in different catena in the landscape. Although these unique relationships have been recognized by soil scientists, they have not been adequately represented in the modelling process for producing soil maps (Zhu et al., 2001; McBratney et al., 2003). Mixed-effects modelling approach presents the opportunity for recognizing such relationships and eventually contributing to accurate soil mapping.

*Corresponding author: thineomuto@yahoo.com

The approach has been used in other studies with nested relationships, which is a promise for successful application in soil mapping. Some researchers have used it to improve the modelling accuracy and efficiency (Pinheiro and Bates, 2000; Faraway, 2006) while others have used it as a tool for incorporating environmental covariates in modelling soil properties (Omuto et al., 2006; Omuto and Gumbe, 2009). These applications encourage the need to test it in regression kriging of soil properties. The objective of this study was to show how it can improve the performance of regression kriging in digital soil mapping.

2. IMPROVING REGRESSION PART OF RK

2.1. Use of Mixed-Effects Modelling

Regression kriging comprises of regression and kriging analyses. In regression analysis, a regression model is used to model the relationship between the target soil property and its predictors. This relationship is believed to take care of the large-scale trends present in soil properties at the landscape scale. A parametric generic form of this relationship is given in Equation (1).

$$\mathbf{y} = f(\mathbf{x}, \beta) + \mathbf{e} \tag{1}$$

where **y** is a vector of the target soil property, **x** is a vector of its predictors, f is a function linking **y** and **x**, β is a vector of the regression model parameters, and **e** is a vector of the regression residuals.

Equation (1) assumes an average response of all individuals in a population. In a graphic illustration of the relationship between **y** and **x** such as shown in Figure 1a, it's represented as a single line around which the individuals coalesce. The parameters (or coefficients) of this regression model are obtained using the parameter estimation methods such as least-squares and optimization (Kottegoda and Russo, 1998). In this study, Equation (1) was referred to as the *single* model.

This study proposes to use mixed-effects modelling to replace Equation (1) which is the current application in RK. Mixed-effects modelling is a unique regression analysis that can simultaneously model nested hierarchical relationships. Its parameters for the high hierarchy (e.g., at the landscape-scale) relationship are known as *fixed-effects* while the parameters associated with individual groups within the landscape are known as *random-effects* (Pinheiro and Bates, 2000). Fixed- and random-effects together form the mixed-effects modelling whose statistical formulation for Equation (1) is as follows,

$$\begin{aligned} \mathbf{y}_i &= f_i(\mathbf{x}, \phi) + \mathbf{e}_i \\ \phi_i &= \mathbf{D}*\beta + \mathbf{B}*\mathbf{b}_i \quad i = 1,2,..., m \\ \mathbf{b}_i &\sim N(0, \psi), \quad e_i \sim N(0, \sigma^2) \end{aligned} \tag{2}$$

where m is the number of groups/classes in the population, ϕ is the vector of regression model parameters consisting of fixed-effects β and random-effects $\mathbf{b}_i$, **D** and **B** are design matrices for solving Equation (2), and ψ is the variance-covariance matrix for the random-effects (Laird and Ware, 1982). The random-effects ($\mathbf{b}_i$) in Equation (2) are associated with i groups/classes because they represent random variations of the groups/classes around the population average estimates (or the fixed-effects) (Fig. 1b). The fixed-effects are the population average estimates and do not have the grouping effects. They, therefore, behave like the estimates in the *single* model in Equation (1).

The random-effects in Equation (2) provide the opportunity for separating groups of individuals with similar response characteristics. This separation consequently accounts for

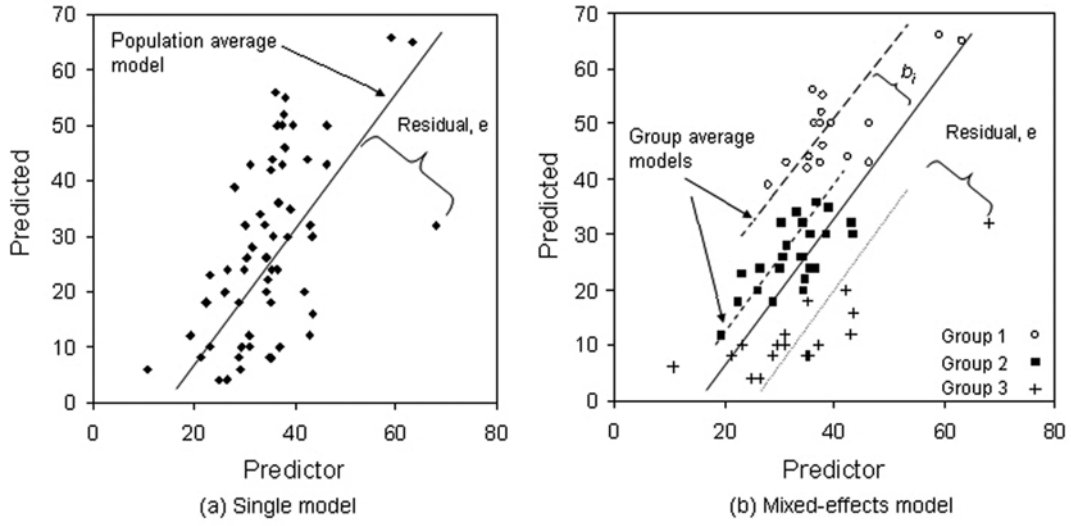

Fig. 1. Conceptual differences between single and mixed-effects models.

the variability between-groups in the landscape, which substantially shrinks the resultant final residual variance and eventually improves the estimation accuracy.

2.2. Regression Analysis in R

2.2.1. Modelling large-scale trends

Equations (1) and (2) for modelling the large-scale trends during soil spatial estimation can be efficiently implemented in R using downloadable packages obtained from http://cran.r-project.org/web/packages/available_packages_by_name.html. An important step in implementing these regression models is to supply the vector of training points **x** and **y**. These training points contain the soil property to be spatially estimated (or **y** vector) and its predictors (or **x** vector). The following example illustrates how executable R scripts can be used to implement the regression models in R. In this example, soil clay content (CLAY) is used as the **y** variable and the following GIS layers as the **x** predictors: elevation (DEM), topography (landform), remote sensing index of vegetation greenness (NDVI), rainfall, latitude, and longitude. It's important to note that the **x** and **y** variables must be in the same working directory in R and are georeferenced to a similar spatial projection.

```
>#Load the necessary libraries for regression analysis
>
> library(foreign)
> library(nlme)
> library(sp)
> library(rgdal)
>
>#import the vector of predictors and georeferenced soil property
>
>samplingpoints=read.table("clay.txt", header=T)# vector of clay content
> predictors=readGDAL("xdist.asc") # for longitude
> predictors$ydist=readGDAL("ydist.asc")$band1# for latitude
> predictors$dem=readGDAL("dem.asc")$band1 # for DEM
> predictors$landform=readGDAL("landform.asc")$band1 # for Landform
> predictors$ndvi=readGDAL("ndvi.asc")$band1 # for NDVI
> predictors$rain=readGDAL("rain.asc")$band1 # for rainfall
> predictors$xdist=predictors$band1 #for renaming the first band
> predictors$band1=NULL #for removing the unnecessary space of renamed band1
>
># Include the geographic projection of the input data
>
> coordinates(samplingpoints)=~X+Y
> proj4string(samplingpoints)=CRS("+proj=utm +ellps=WGS84") # using UTM for this example
> proj4string(predictors)=CRS("+proj=utm+ellps=WGS84")
>
># Align the soil properties (vector y) with the predictors (vector x) in one training file
>
> predictors.ov=overlay(predictors, samplingpoints)
> samplingpoints$xdist=predictors.ov$xdist
> samplingpoints $ydist=predictors.ov$ydist
> samplingpoints $dem=predictors.ov$dem
> samplingpoints $ndvi=predictors.ov$ndvi
> samplingpoints $rain=predictors.ov$rain
> samplingpoints $landform=predictors.ov$landform
>
># Now the regression analysis can be carried out for the data in the training file
>
> por.lm=lm(sqrt(CLAY)~(xdist+ydist+dem+landform+ndvi+rain), samplingpoints)# Equation (1)
> por.lme=lme(sqrt(CLAY)~(xdist+ydist+dem+ndvi+rain), random=~1|landform, samplingpoints)# Equation (2)
>
># Compare the coefficient of determination for the two regression models
>
> cor(fitted(por.lm),samplingpoints$CLAY)^2 # for Equation (1)
 [1] 0.4980632
> cor(fitted(por.lme),samplingpoints$CLAY)^2 # for Equation (2)
 [1] 0.6412688
```

The potential of mixed-effects modelling in improving regression analysis is seen in better indicators of goodness of fit than in the single model. For example, the coefficient of determination (r^2), as an indicator of goodness of fit, was higher in mixed-effects model (e.g., in the above example, $r^2 = 0.64$) than in the *single* model (0.50). Similarly, the residual standard error of the mixed-effects model (in the above example, = 0.033) was lower than the *single* model (0.038). When comparing models, high coefficient of determination (r^2) and low residual standard error are agreed as statistical indicators of better predictive performance (Kottegoda and Russo, 1998; Pinheiro and Bates, 2000). Further assessment of the mixed-effects model showed that its random variations were split into two: the random-effects component with a standard deviation of 0.02 and the residuals component with a residual standard error of 0.033. The *single* model had the residuals as the only random variation with a residual standard error of 0.038. Its residual standard error was inflated because it incorporated the random variations due to the soil group characteristics. This is because the model does not recognize the natural soil grouping. This deficiency costs the model its accuracy in soil spatial estimation.

2.2.2. Modelling small-scale autocorrelation using kriging

Small-scale autocorrelation are modelled using the residuals from the regression stage of RK as the input variable. It is important to note that these residuals not only contain the small-scale autocorrelation but also the measurement and modelling errors. The modelling errors are due to regression deficiencies of the regression part of RK while the measurement errors come from sampling/measurement of input variables.

In kriging, the guiding mathematical hypothesis is the intrinsic stationarity that requires the mean and semivariance to depend strictly on the separation distance between samples and not on the coordinate position of the data (Journel and Huigbregts, 1978). Two conditions must be satisfied in this hypothesis:

1. The mean residual exists and doesn't depend on the geographic locations

$$\hat{e}(s) = m \tag{3}$$

2. The variance of the residual increment $[e_i(s+h) - e_i(s)]$ exists and doesn't depend on the geographic locations but on the difference vector h

$$\text{var}[e_i(s+h) - e_i(s)] = E[e_i(s+h) - e_i(s)]^2 = 2\gamma(h) \tag{4}$$

where $e(s)$ is the residual from Equation (2), s is the geographic coordinates with latitude and longitude components, $\gamma(h)$ is the semivariance, and E is mathematical notation for expectation. The semi-variance is usually estimated by Equation (5).

$$\gamma(h) = \frac{1}{2W(h)} \sum_{i=1}^{W(h)} [e_i(s+h) - e_i(s)]^2 \tag{5}$$

where $W(h)$ is the number of pairs of e which are located h distance apart. A graph of semivariance with h is called experimental semivariogram (Nielsen and Wendroth, 2003). The experimental semivariogram often depict scatter just like many experimental data. There are mathematical models which can model the experimental semivariogram so that they can be used to extend to all distances (Gotway, 1991). The modelled semivariogram is then used to characterize the small-scale autocorrelation and for kriging estimation. It contains at least three parameters: nugget, sill, and range which explain the spatial structure (Isaaks and Srivastava, 1989). These parameters, the experimental data, and modelled semivariogram are shown in Figure 2. The description of these parameters has been given in various publications (see for example Nielsen and Wendroth, 2003).

Kriging estimation is a linear function of neighbouring sampling points and is given by

$$e_o(s_o) = \sum_{i=1}^{n} \lambda_j e_j(s_i) \tag{6}$$

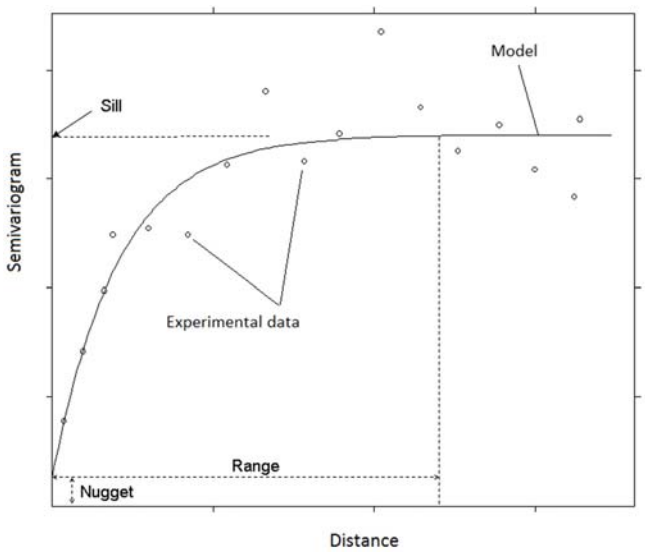

Fig. 2. Example of semivariogram.

on condition that the sum of weights $\sum\lambda_j = 1$ and the variance is minimum between the true and estimated values. The minimum variance is obtained when

$$\sum_{j=1}^{n} \lambda_j * \lambda(h_i - h_j) + \mu - \gamma(h_i - h_0) = 0 \tag{7}$$

where $\gamma(h)$ is the semivariance obtained from the semivariogram model, $e_o(s_o)$ is the krigged residual at location s_o, and μ is a Lagrangian multiplier.

2.3. Kriging Analysis in R

Many kriging methods exist in the literature such as ordinary kriging, simple kriging, Bayesian kriging, etc. The R packages for implementing these methods are also available and include Gstat, geoR, automap, etc. The following example shows how kriging can be implemented in R to model small-scale autocorrelation. The input for kriging is the vector of residuals from the regression part of RK.

```
>#Continue from the regression part
>#Load the packages for kriging
>
> library(spacetime)
> library(gstat)
> library(automap)
>
>#Recall the residuals from the regression model and use them to fit the variogram
> variogram = autofitVariogram(residuals(por.lme)~1, samplingpoints)
> plot(variogram)
>
>#Implement ordinary kriging and show the results
```

```
> krig = autoKrige(residuals(por.lme)~1, samplingpoints, predictors)
> plot(krig)
```

If the above kriging example is repeated for the residuals from the *single* model, it will be seen that the *nugget* variance is higher than that obtained from the mixed-effects model. *Nugget* is the local variance describing the non-spatial variability, such as the measurement error, the temporal variability, errors carried from the regression part, and the spatial variability at the scale too small to be captured by the applied sampling scheme. If all other factors are assumed constant for the two regression models (that is, measurement errors, temporal variability, and spatial variability), then the difference in *nugget* between the *single* and mixed-effects models could be attributable to differences in modelling errors from the regression part of the RK method. Hence, the high *nugget* variance in the *single* model could imply that the model had more modelling errors than the mixed-effects model. This further confirms that the *single* model has inaccuracies which can negatively influence the spatial predictive performance of RK.

The final output of spatially estimated soil property is obtained by combining the outputs of the two stages: regression and kriging parts. The following example shows how this can be implemented in R.

```
>#Predict and back-transform the regression part for the whole study area
> predictors$PredCLAY=(predict(por.lme, predictors))^2
> predictors$Predclay=predictors$PredCLAY+krig$krige_output$var1.pred
>
>#Display the outputs
>
> Pred.plt=spplot(predictors["Predclay"], col.regions=bpy.colors(), scales=list(draw=TRUE,cex=1), sp.layout=list("sp.points", pch="+",col="black",fill=T, samplingpoints))
> var.plt=spplot(krig$krige_output["var1.var"], col.regions=bpy.colors(),scales=list(draw=TRUE,cex=1), sp.layout=list("sp.points",pch="+",col="black", fill=T, samplingpoints))
> print(Pred.plt, split=c(1,1,2,1), more=TRUE)# Estimated map
> print(var.plt, split=c(2,1,2,1), more=FALSE)# Kriging variance map>
>#The outputs can also be exported to other software
> write.asciigrid(predictors["Predclay"], "predclay.asc")
```

3. APPLICATION IN SPATIAL ESTIMATION OF PERCENT CLAY CONTENT IN KENYA

3.1. Input Data

This example is shown for spatial estimation of clay content in Kenya. The training data used consisted of 374 georeferenced samples of clay content (Fig. 3). 350 of these samples were obtained from www.isric.org on 4th October 2010, 19 from Tana River Development Authority (TARDA) on 19th September 2010 (Tana River Development Authority, 1987) and the remaining five samples from the Ministry of Livestock Development (MOLD) on 11th October 2010 (Touber, 1991). The clay content data was for topsoil (i.e., from 0 cm to 20 cm from the soil surface) (Table 1). Description of the laboratory methods for its estimation has been given in Batjes (2008), Touber (1991) and Tana River Development Authority (1987).

The predictors of clay content tested were annual average rainfall amounts, Digital Elevation Model (DEM), land use, Normalized Difference Vegetation Index (NDVI), landform, and geology. These predictors represented the soil forming factors (i.e., climate, organism, relief, parent material, and time) that are known to influence the spatial distribution of soil properties (McBratney et al., 2003). DEM and landform represented relief, NDVI and land use represented organism, rainfall represented climate, and geology represented parent material. DEM was downloaded from http://srtm.usgs.gov on 15th September 2009. It was a map of surface elevation in metres above sea level at a spatial resolution of 90 m. Land use was extracted from land use map developed by AFRICOVER (www.africover.org accessed on 16th September 2009). The map contained 73 land use/cover types mapped at the scale of 1:200,000. NDVI data was the remote sensing image of average vegetation index of 16-day composite MODIS images between January 2000 and December 2009. It had a spatial resolution of 250 m and was obtained from http://pekko.geog.umd.edu/usda/apps on 14th August 2010. The landform was obtained from the landform map of Kenya using the technique developed by Iwahashi and Pike (2007). It had 18 landform types mapped at 1,000 m spatial resolution. The geology data was originated from the geology map. The map was obtained from www.rcmrd.org on 10th October 2010 and had 42 classes of lithology mapped at the scale of 1: 1,000,000. The rainfall data was the mean annual rainfall amount (between January 1982 and December 2009) for 152 weather stations in Kenya. It was obtained from the Kenya Meteorology Department (www.kmd.go.ke) on 22nd August 2010.

All vector input data were converted into spatial raster data format to aid pixel-based analysis. They included rainfall, land use, and geology maps. The rainfall data was converted to rainfall image by use of the simple kriging method (Nielsen and Wendroth, 2003). The land use and geology maps were converted to images using vector-to-raster algorithms in ERDAS Imagine® (ERDAS, 2003; Teng et al., 2008). After the conversion, the input raster datasets were re-sampled to 1 km spatial resolution to ensure consistency with all input data. This resolution was chosen for convenience to harmonize data ahead of RK.

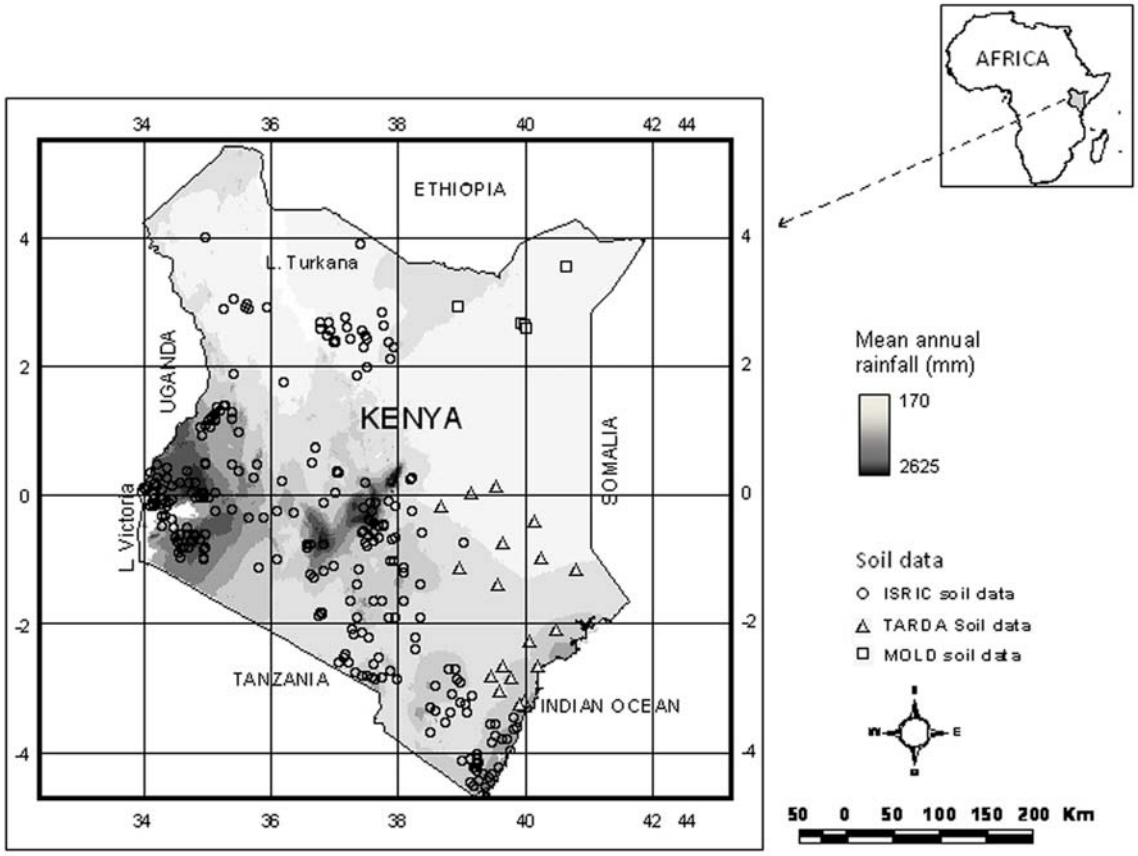

Fig. 3. Location of training points for predicting clay content in Kenya.

Table 1. Statistics summary of clay content and its predictors

Statistic	Clay (%)	DEM (m)	Rainfall (mm)	NDVI (–)	Y-coord (m)	X-coord (m)
Minimum	10	0	2.7	0.100	–4.68	33.90
Maximum	89	4695	1096	0.982	5.41	41.90
Average	40.5	471	341	0.251	0.36	37.90
Skewness	0.32	0.48	1.22	0.44	1.72	1.11

3.2. Spatial Estimation of Clay Content

The predictors to use in the regression analysis were chosen depending on the strength of the correlation analysis between them and clay content, and between themselves (Table 2). Since clay content and these predictors were positively skewed, they were normalized with Box-Cox transformation (Eagleson and Muller, 1997) before the correlation analysis. DEM, NDVI, and X-coordinates were chosen as the best predictors of clay content because of their favourable correlation indices (Table 2).

The general mixed-effects model in Equation (2) was used for the regression modelling. The random-effects were fitted for each land use, landform, and geology classes. A sequential analysis as given in Faraway (2006) was done to select the best grouping variable from land use, landform, and geology classes. Landform and geology were identified as the best grouping variables for modelling clay content in Kenya. It's possible the selection was adequate since topography (landform) and parent material (geology) have been shown in many publications to influence clay formation (see for example, Barshad (1957), Hay (1960), and Price et al. (2005)).

Table 2. Summary of correlation (r) between clay content and its predictors

	DEM	Rainfall	NDVI	Y-coord	X-coord
DEM	1				
Rainfall	0.34	1			
NDVI	0.53	0.5	1		
Y-coord	0.37	–0.25	–0.04	1	
X-coord	–0.61	–0.35	–0.27	–0.25	1
Clay	0.54	0.44	0.61	0.33	0.70

In order to check the estimation accuracy of the model, the model was calibrated on randomly selected three-quarters of the samples and validated on the remaining one-quarter. After the accuracy assessment, all the samples were included in the spatial estimation to produce the final map of clay content in the study area.

3.3. Performance Comparison for Spatial Estimation Models

In order to compare output map by the mixed-effects and *single* models, the models were first used in RK to produce clay content map of Kenya. Figure 4 shows the final maps of estimated soil clay content using the mixed-effects and *single* models.

Both maps had fuzzy boundaries between different levels of clay content. They also showed that the eastern and south-eastern parts of the country had low clay content while the western and central parts had high clay content (Fig. 4). This pattern was similar to the pattern of clay content map of Kenya produced by Batjes (2008) using the traditional method of mapping with polygons.

Compared to the measured clay content in Table 1, the map produced by the mixed-effects model had nearly similar range (i.e., the difference between the maximum and minimum values) while that produced by the *single* model had a narrow range (Fig. 4). In terms of holdout validation, the mixed-effects model had 58% estimation success while the *single* model had 14% estimation success. Further comparison was made from a 2D plot of the measured clay content (on x-axis) verses estimated values (on y-axis) in order to establish which model produced uniform scatter around the 1:1 line. The comparison showed skewed distribution for the *single model*. It had several plotted points below the 1:1 line for values of clay content above 50%. It also had more plotted points above the 1:1 line than those below the line for values of clay content below 25%. These distribution characteristics showed that the model did not have a balanced estimation throughout the range of measured values. The mixed-effects model showed a fairly balanced estimation throughout the range of measured values. It had nearly similar scatter of points for the low and high values of measured clay content. However, the plotted points were not quite close to the 1:1 line, which implied that the model was also not very accurate. No concrete explanation was available for this observation. However, it was believed that the low accuracy was attributable to inaccuracies in the input data used instead of the models used. The data used were secondary data from different sources.

Although the above validation statistics were not very high, they seemed to favour the mixed-effects model more than the *single* model using the same dataset. They suggested that the map produced by the mixed-effects model seemed closer to the actual values than that produced by the current

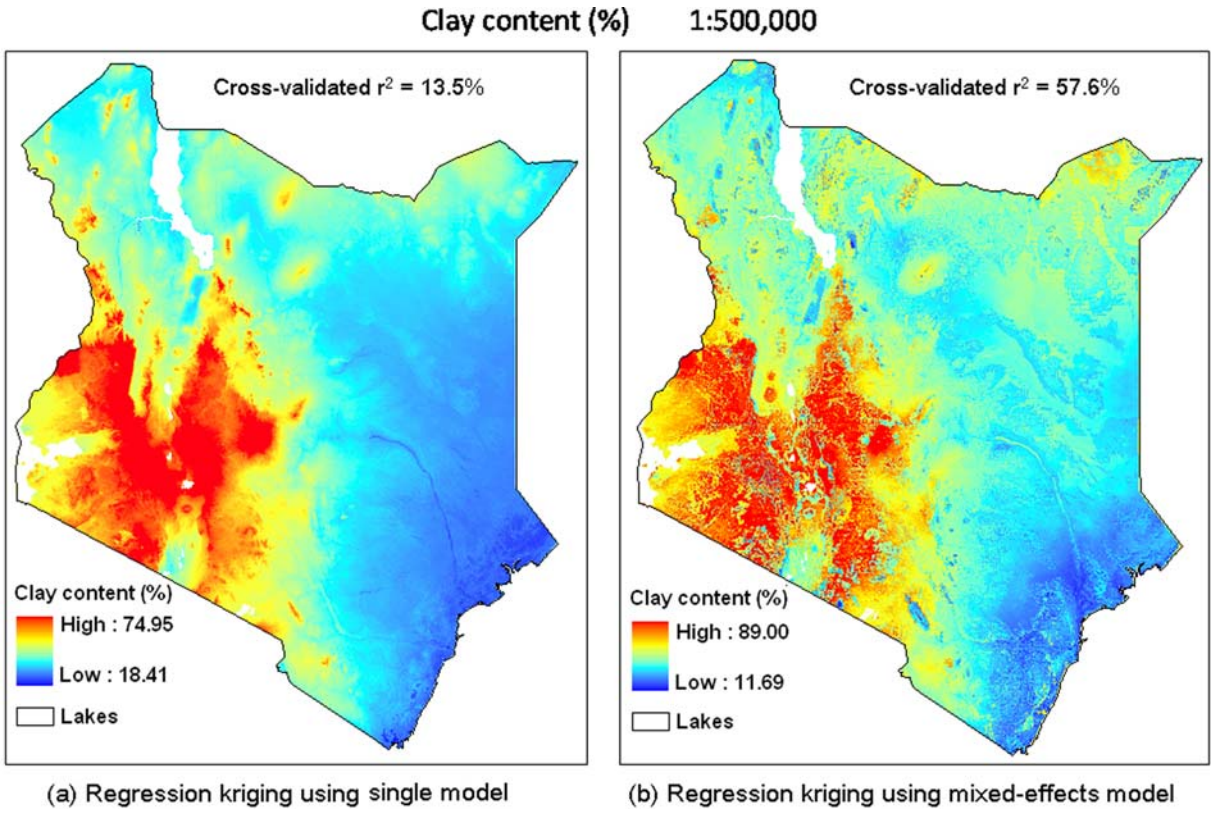

Fig. 4. Estimated topsoil clay content in Kenya.

application in RK. Hence, the spatial pattern depicted in Figure 4b should be more realistic than Figure 4a.

4. CONCLUSIONS

This paper used mixed-effects modelling approach to improve the performance of regression kriging (RK). RK, which is a spatial estimation method, contains the regression and kriging parts. Any deficiency in these parts can negatively affect the overall performance of RK. In this paper, it was shown that the current regression application has a deficiency that can limit better predictive performance of RK. The deficiency is the inability of the current regression models to recognize the natural soil groupings in the landscape. Natural soil occurs in groups (catena) with unique response characteristics to soil forming factors. During soil spatial estimation, these groups ought to be recognized and incorporated in the modelling process. The current regression application in RK do not incorporate these groups; hence, resulting into inflated regression residuals during modelling. These inflated residuals are passed as unexplained variations in the regression analysis. In this study, mixed-effects approach is introduced with capabilities of further explaining the regression residuals and improving the regression accuracy. Mixed-effects approach contains random-effects which can be associated with groups in a population. These random-effects were used in this paper to model the natural soil groupings in the landscape. Consequently, they helped to further explain soil group variation which was otherwise lumped into unexplained residual variation in the current RK application. Since the endeavour of soil spatial prediction is to explain modelling variations as far as possible, the opportunity introduced by random-effects in the mixed-effects modelling should be welcome. In this paper, it was shown how this opportunity contributed to improved accuracy. Evidence of the improvement was shown with high coefficient of determination for the regression part of RK, regression residuals with low errors passed to the kriging part of RK, and high validation statistics for the final output map.

In addition to improved performance of mixed-effects in RK, this study also showed how mixed-effects in RK can be easily implemented using freely downloadable and versatile R software. Executable scripts were shown using available R packages to produce spatially estimated soil property map. All together, the new RK and its easy implementation in R software are anticipated to provide the potential for wide application and eventual contribution to improved soil mapping and application of DSM.

ACKNOWLEDGMENTS: A special acknowledgement is given to different authors who produced the R packages used in this study. The data used in this study was obtained from ISRIC (www.isric.org), Kenya Meteorological Department (www.kmd.ac.ke), Geography Department of the University of Maryland (http://pekko.geog.umd.edu), Regional Centre for Mapping of Resources for Development (www.rcmrd.org), Tana River Development Authority (TARDA), and the Government of Kenya. The study was financed by International Foundation for Science (www.ifs.se) through the project number C/3953-2.

REFERENCES

Barshad, I., 1957, Factors affecting clay formation. Clays and Clay Minerals, 6, 110–132.

Batjes, N.H., 2008, ISRIC-WISE Harmonized Global Soil Profile Dataset. Report No. 2008/002. ISRIC-World Soil Information, Wageningen, 59 p.

Brus, D.J., Bogaert, P., and Heuvelink, G.B.M., 2008, Bayesian maximum entropy estimation of soil categories using a traditional soil map as soft information. European Journal of Soil Science, 59, 166–177.

Carré, F. and Girard, M.C., 2002, Quantitative mapping of soil types based on regression kriging of taxonomic distances with landform and land cover attributes. Geoderma, 110, 241–263

Cui, H., Stein, A., and Myers, M.E., 1995, Extension of spatial information, Bayesian kriging and updating of prior variogram parameters. Environmetrics, 6, 373–384.

Eagleson, G.K. and Muller, H.G., 1997, Transformations for smooth regression models with multiplicative errors. Journal of Royal Statistical Society, 59, 173–189.

ERDAS, 2003, ERDAS Imagine 8.7 Field Guide™. Leica Geosystems GIS and Mapping LLC, Georgia, 698 p.

Faraway, J.J., 2006, Extending the Linear Model with R: Generalized Linear Models, Mixed Effects and Nonparametric Regression Models. Taylor and Francis, Boca Raton, 331 p.

Gotway, C.A., 1991, Fitting semivariogram models by weighted least squares. Computers & Geoscience, 17, 171–172.

Hay, R.L., 1960, Rate of clay formation and mineral alteration in a 4000-year-old volcanic ash soil on Saint Vincent, B.W.I. American Journal of Science, 258, 354–368.

Hengl, T., Heuvelink, G.B.M., and Rossiter, D.G., 2007, About regression kriging from equations to case studies. Computers & Geosciences, 33, 1301–1315.

Herbst, M., Diekkrüger, B., and Vereecken, H., 2006, Geostatistical co-regionalization of soil hydraulic properties in a micro-scale catchment using terrain attributes. Geoderma, 132, 206–221.

Isaaks, E.H. and Srivastava, R.M., 1989, An Introduction to Applied Geostatistics. Oxford University Press, New York, 592 p.

Iwahashi, J. and Pike, R.J., 2007, Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. Geomorphology, 86, 409–440.

Journel, A.G. and Huijbregts, C.J., 1978, Mining Geostatistics. Academic Press, London, 604 p.

Kottegoda, T.N. and Russo, R., 1998, Statistics, Probability and Reliability for Civil and Environmental Engineers. McGraw-Hill, New York, 735 p.

Lagacherie, P., McBratney A.B., and Voltz, M., 2007, Digital Soil Mapping. An Introductory Perspective. Elsevier, Amsterdam, 658 p.

Laird, N.M. and Ware, J.H., 1982, Random–effects models for longitudinal data. Biometrics, 38, 963–974.

Lark, R.M., Cullis, B.R., and Welham, S.J., 2006, On spatial estimation of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. European Journal of Soil Science, 57, 787–799.

McBratney, A.B, Santus M.L.M., and Minasny, B., 2003, On digital soil mapping. Geoderma, 117, 3–52.

Nielsen, D. and Wendroth, O., 2003, Spatial and Temporal Statistics: Sampling Field Soils and their Vegetation. Catena-Verlag, Reiskirchen, 404 p.

Odeh, I.O.A, McBratney, A.B., and Chittleborough, D.J., 1995, Further results on estimation of soil properties from terrain attributes: Heterotopic co-kriging and regression kriging. Geoderma, 67, 215–226.

Omuto, C.T. and Gumbe, L.O., 2009, Estimating water infiltration and retention characteristics using a computer program in R. Computers & Geosciences, 35, 579–585.

Omuto, C.T., Minasny, B., McBratney, A.B., and Biamah, E.K., 2006, Nonlinear mixed-effects modelling for improved estimation of infiltration and water retention parameters. Journal of Hydrology, 330, 748–758.

Pinheiro, J.C. and Bates, D.M., 2000, Mixed-Effects Model in S and S-plus. Springer-Verlag, New York, 529 p.

Price, J.R., Velbel, M.A., and Patino, L.C., 2005, Rates and time scales of clay-mineral formation by weathering in Saprolitic Regoliths of the Southern Appalachians from geochemical mass balance. Geological Society of America Bulletin, 117, 783–794.

Scull, P., Franklin, J., and Chadwick, A.O., 2005, The application of classification tree analysis to soil type estimation in a desert landscape. Ecological Modelling, 181, 1–15.

Tana River Development Authority (TARDA), 1987, Development Plan. Report Vol. II. TARDA, Nairobi, 219 p.

Teng, J., Wang, F., and Y. Liu, 2008, An efficient algorithm for raster-to-vector data conversion. Annals of GIS, 14, 54–62.

Touber, L., 1991, Wet Season Restrictions in Accessibility of Livestock in Mandera and Wajir Districts. Ministry of Livestock and Development, Government of Kenya, Nairobi, Kenya, 57 p.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., and Simonson, D., 2001, Soil mapping using GIS, expert knowledge and fuzzy logic. Soil Science Society of America Journal, 65, 1463–1472.

Manuscript received September 30, 2013
Manuscript accepted April 30, 2014