Geoderma 326 (2018) 22–41

Contents lists available at ScienceDirect

Geoderma

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

Regression kriging as a workhorse in the digital soil mapper's toolbox☆

H. Keskin[a,b], S. Grunwald[a,*]

[a] Soil and Water Sciences Department, 2181 McCarty Hall, PO Box 110290, University of Florida, Gainesville 32611, FL, USA
[b] Republic of Turkey Ministry of Forestry and Water Affairs, General Directorate of Combating Desertification and Erosion Söğütözü Cad. No: 14/E, Ankara, Turkey

ARTICLE INFO

Handling Editor: A.B. McBratney

Keywords:
Regression kriging
Digital soil mapping
Soil spatial predictions models
Pedometrics
Accuracy

ABSTRACT

Appropriate scale, justifiably reliable, categorical and continuous spatial soil information is urgently needed to address environmental problems and ensure sustainability of ecosystem services at local, regional and global scales. Regression Kriging (RK) is one of the most popular, practical and robust hybrid spatial interpolation techniques in the digital soil mapper's toolbox that enables the modeling of soil distribution patterns at multiple scales in space and time. Several theoretical and applied aspects of RK have been discussed; however, there are no review studies, which quantify the essential factors affecting the performance of RK. Materials for this review were gathered from high-quality international soil science journals: *Catena*, *Geoderma*, and *Soil Science Society of America* from 2004 to 2014. A total of 142 different models from 40 different articles were examined. The following criteria were considered to evaluate their impacts on the prediction efficiency of RK: i) soil geographic region, ii) area of extent, iii) spatial resolution, iv) target soil properties and/or classes v) sampling design, vi) sampling size and density, vii) sample depth viii) soil-environmental factors as predictors, ix) methods of transformation, x) factor analysis, xi) regression type, xii) model used for variogram, xiii) nugget to total sill ratio, xiv) spatial autocorrelation range, xv) coefficient of variation of observed dataset, xvi) evaluation method (note that in previous publications the term 'validation' has been used extensively in publications in pedometrics) and xvii) coefficient of determination. The historical development of RK, limitations and strengths of current RK studies, research gaps, and future trends in RK are discussed. A major finding is the inverse relationship between the accuracy of RK models and the variation of soil properties in the original datasets. Novel modified RK methods are proposed for further investigation to predict soil properties and classes.

1. İntroduction

At the beginning of the 21st century, advances in computational power, geographic information systems, remote sensing and statistical methods have collectively enabled pedologists to produce state-of-the-art, reliable, categorical and continuous spatial soil information at multiple scales in space and time, which empower environmental scientists to model and policy makers to deal with wicked environmental problems, such as land degradation, climate change, food and water security, biodiversity and ecosystem functions protection (Bouma and McBratney, 2013; Hartemink and McBratney, 2008; Grunwald et al., 2016). Consequently, providing high-quality, justifiably reliable, reproducible spatiotemporal soil information with quantified uncertainty has been the major focus in digital soil mapping (DSM) which has now shifted from the research phase into an operational phase (Minasny and McBratney, 2015). Modeling and decreasing inaccuracies in DSM is an essential requirement in the quest to comprehend variability in soil properties and/or classes at multiple scales. A better understanding of soil variability will pave the way for a better understanding of geopatterns on the Earth's surface (Bockheim and Gennadiyev, 2010).

Inherently soil variation poses a significant problem to achieve accurate digital soil models (Burrough et al., 1994). Two general, yet distinct approaches have been offered to account for the soil variation: discrete modeling of soil variation (polygon-based) and continuous modeling of soil variation (pixel-based) (Heuvelink and Webster, 2001). While the first approach partitions the soil into more and less homogeneous classes, the latter elucidates the soil-landscape as a continuum. Traditional soil classification, which uses a polygon-based soil map unit model, suffers from numerous drawbacks. As Hartemink et al. (2010) articulated the maps produced by traditional soil classification methodology are static, inflexible, and require further steps to integrate with grid-based digital soil sources. Polygon-based models are often devoid of specifying the uncertainty (Grunwald, 2006), while more recently uncertainty assessment to quantify map unit composition has gained more attention. Altogether, these handicaps largely contributed to the decrease in funding to pedological research in the late 1990's (Basher,

☆ In memoriam to I.O.A. Odeh who was instrumental to develop Regression Kriging. This paper is dedicated to honor his profound contributions to pedometrics.

* Corresponding author.
E-mail addresses: hkeskin@ormansu.gov.tr (H. Keskin), sabgru@ufl.edu (S. Grunwald).

https://doi.org/10.1016/j.geoderma.2018.04.004
Received 5 October 2017; Received in revised form 29 March 2018; Accepted 3 April 2018

1997; Ryan et al., 2000). Consequently, soil scientists inevitably shifted from qualitative subjective modeling of soil properties and/or classes to quantitative objective modeling ("soil science under uncertainty") (Goovaerts, 2001).

These developments lead to the unifying modeling of soil spatial variation formalized by the regionalized variable theory (RVT) with the following equation (after Burrough, 1986)

$$Z(x) = \mu(x) + \varepsilon'(x) + \varepsilon''(x) \quad (1)$$

where:

- x: location in one, two or three dimensions,
- $Z(x)$: the random variable Z at location x,
- $\mu(x)$: deterministic structural component, trend (drift),
- $\varepsilon'(x)$: stochastic component, spatially dependent residual from μ(x) [the regionalized variable] but locally varying in both lateral and vertical direction,
- $\varepsilon''(x)$: nonspatially-correlated component, noise, unexplained variability.

Spatial variability in soil forms a spectrum of variation ranging from microscopic to megascopic scale (Wright and Wilson, 1979) as a function of many possible factors, including target area of extent, spatial resolution, specific soil properties or processes, spatial location and time (Lin et al., 2005). Altogether, these factors may form a trend at multiple scales depicted with a deterministic function ($\mu(x)$ in Eq. (1)). However, the processes responsible for soil variation are generally unknown and given the current expertise soil variability is unlikely to be captured analytically at multiple scales in both space or time (Heuvelink and Webster, 2001). Typically, the values for a soil property from samples taken at close geographic spacing is similar or spatially correlated (Oliver, 1987). This is the premise of the spatially dependent random component ($\varepsilon'(x)$ in Eq. (1)). Semivariograms have been used to characterize the stochastic structural component as a function of distance between two adjacent points under the stationarity assumption. Nonspatially-correlated component of the variation, noise, is the unexplained variability ($\varepsilon''(x)$ in Eq. (1)) which is present in the model having a mean zero and variance σ^2 (Webster, 2000).

The soil factorial model, an empirical-deterministic model of soil formation developed by V.V. Dokuchaev (Glinka, 1927), that was popularized by Jenny (1941), has been widely utilized to quantitatively describe the relationship between soil and its forming factors. By contrast RVT (Matheron, 1971) has allowed researchers to predict the values of various soil properties at unknown locations (Webster, 1994). Many statistical and purely geostatistical methods used since the 1960s have been collectively categorized under the new branch in soil science called "pedometrics". Pedometrics can be defined as the application of probability and statistical methods to soil science (Webster, 1994) or the application of mathematics and statistics to study the distribution and genesis of soil (McBratney et al., 2000). Deterministic and stochastic variation of soil attributes and classes have been systematically studied under the discipline of pedometrics since the 1990s.

Two main generic approaches that are representative of these two distinct model paradigms address soil variation and predict soil properties or classes at an unvisited location: (1) non-geostatistical techniques; e.g., simple and multiple linear regression (MLR), generalized additive model (GAM), classification and regression tree (CART) and (2) geostatistical techniques; e.g., ordinary kriging (OK), simple kriging (SK), universal kriging (UK) (Webster and Burgess, 1980, Moore et al., 1993; Odeh et al., 1994, McBratney et al., 2000). Non-geostatistical techniques have been used to quantify the relationship between soil properties and state factors accounting for the deterministic portion of the total variation "μ(x)" (Fig. 1). Geostatistical methods, on the other hand, have been used to quantify changes in soil properties over various distances unraveling the spatially dependent stochastic portion of the total variation "ε′(x)" (Fig. 1). These two generic approaches were combined to create hybrid techniques (i.e., non-stationary geostatistical methods) (Wackernagel, 2003), in the mid-1990s. While the non-geostatistical part detects the deterministic part of the total variation, the geostatistical part quantifies the spatially dependent stochastic part of the total variation.

A number of hybrid techniques have been developed including universal kriging (UK) (or kriging with internal drift) (Webster and Burgess, 1980) and kriging with external drift (KED) (Goovaerts, 1997). Both UK and KED have the same formulation, the trend and residuals are estimated in a system in which the prediction variance is jointly estimated. UK is a special case of kriging where the trend is modeled only by spatial position. In KED the trend is externally modeled from auxiliary variables. The key point is that with UK/KED there is one kriging system to solve, whereas with RK the regression (R) can be independent of the kriging (K). Therefore, with KED there is a joint estimation of the prediction variance, but with RK the variance parts from R and K must be summed. The biggest advantage of RK over its formal counterpart UK/KED is that the trend does not have to be defined by linear models; it can be defined by an array of nonlinear mathematical models such as regression trees, random forest (RF), and neural networks (Hengl et al., 2007a).

Odeh et al. (1994, 1995) coined the term RK and introduced RK type A, B and later RK type C. We provide a brief overview of RK as a key hybrid modeling approach in pedometrics and discern nuanced, yet critical, differences in the rather confusing naming conventions found in the literature. According to Odeh et al. (1995) in RK type A the regression residuals represent uncertainty which are incorporated into the kriging systems. First a regression is performed to derive a target variable and then kriging is performed with introduction of regression errors into the kriging system as prediction uncertainty. The aim is that kriging after regression may improve (with the introduction of the uncertainty due to regression errors into kriging equations) prediction performance in comparison to when regression or kriging are done separately. RK type A, which is also called "kriging combined with regression" by Knotters et al. (1995), involves kriging of the regressed values after regression is performed. Two variances are assessed, the first is the variance of the regression model and the second the variance of the kriging system which is minimized under the assumption that the errors are not correlated with the variable of interest (Knotters et al., 1995). In RK type B, which is called "kriging with guess field" (Ahmed and De Marsily, 1987), a regression model is fitted to compute a secondary variable and residuals are derived that are then kriged by ordinary kriging. The final estimate is derived by the sum of the kriged secondary variable and the kriged residuals. The variance of the estimation error of the final estimate is the sum of the variances of the estimation error of these two ordinary kriging system (Ahmed and De Marsily, 1987). RK type C (Odeh et al., 1995), which is called "kriging after detrending" (Goovaerts, 1999), is defined as the sum of the regressed values and kriged residuals from the regression. The difference of RK type C to type B is that it only uses the kriging of the residual to obtain the final prediction. For an extensive review of the hybrid kriging techniques, a full discussion of RK can be found elsewhere (Hengl et al., 2007a; Knotters et al., 1995). RK type C is one of the most widely used hybrid spatial interpolation method used in soil science to predict soil properties (Minasny and McBratney, 2007). An example of the steps to execute RK is provided in Fig. 2.

First, soil and ancillary environmental data are collected for a given study region. The next step is to compute a regression between the state factors and the target soil property. Then the trend model, identified by the regression equation, is subtracted from Z(x) and residuals are quantified. The residuals from the trend are treated as spatially correlated stationary random variables. Finally, the regression estimates and the kriged residual values are summed together to create the final prediction map. İt should be noted here that simple kriging or ordinary kriging of residual can be performed to execute RK. While some authors