Geoderma 261 (2016) 11–22

Contents lists available at ScienceDirect

Geoderma

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

Mapping soil organic carbon content by geographically weighted regression: A case study in the Heihe River Basin, China

Xiao-Dong Song [a], Dick J. Brus [a,b], Feng Liu [a], De-Cheng Li [a], Yu-Guo Zhao [a], Jin-Ling Yang [a], Gan-Lin Zhang [a,*]

[a] State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China
[b] Alterra, Wageningen University and Research Centre, PO Box 32, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 13 March 2015
Received in revised form 9 June 2015
Accepted 30 June 2015
Available online xxxx

Keywords:
Digital soil mapping
Kriging with an external drift (KED)
Restricted maximum likelihood (REML)
Map accuracy
Cross-validation
Effective sample size

ABSTRACT

In large heterogeneous areas the relationship between soil organic carbon (SOC) and environmental covariates may vary throughout the area, bringing about difficulty for accurate modeling of the regional SOC variation. The benefit of local, geographically weighted regression (GWR) coefficients was tested in a case study on soil organic carbon mapping across a 50,810 km^2 area in northwestern China. This area is composed of an alpine ecosystem in the upper reaches and oases in the middle reaches. The benefit was quantified by comparing the quality of the maps obtained by GWR and geographically weighted ridge regression (GWRR) on the one side and multiple linear regression (MLR) on the other side. In these methods spatial dependence of model residuals is ignored. The root mean squared error (RMSE) of predictions of natural log-transformed SOC obtained with GWR was smaller than with MLR: 0.565 versus 0.618 g/kg. The use of a local ridge parameter in GWRR did not lead to an increase in accuracy. Besides we compared the quality of maps obtained by geographically weighted regression followed by simple kriging of model residuals (GWRSK) and kriging with an external drift (KED) with global regression coefficients. In these methods the spatial dependence of model residuals is incorporated in the model. The RMSE with KED was smaller than with GWRSK: 0.515 versus 0.546 g/kg. We conclude that fitting regression coefficients locally as in GWR only paid when no spatial random effect was included in the model. When a spatial random effect was included, the flexibility of local, geographically weighted regression coefficients was not needed and even undesirable as it led to less accurate predictions than KED with global regression coefficients. In comparing the accuracy of prediction methods by leave-one-out cross-validation (LOOCV) of a non-probability sample it is important to account for possible autocorrelation of pairwise differences in the prediction errors. The effective sample sizes were substantially smaller than the total number of sampling points, so that most pairwise differences in MSE were not significant at a significance level of 10% in a two-sided paired *t*-test.

1. Introduction

Soil is one of the most important carbon stocks globally and maps showing soil organic carbon (SOC) can be used to guide practical soil management (Meersmans et al., 2008). Cost-efficient methods for mapping SOC content are therefore indispensable (Kheir et al., 2010).

A recent review of 90 papers on digital soil mapping and modeling revealed that 31% (28) of these papers focused on SOC (Grunwald, 2009). Amid the plethora of mapping methods exploiting the relation between SOC and covariates were regression kriging (Piccini et al., 2014), (boosted) regression trees (Vasques et al., 2008, 2009), random forest (Grimm et al., 2008), and neural network (Li et al., 2013).

Despite the significant progress based on these methods, there are still methodological challenges, especially in large, highly variable areas, with spatially varying relationships between soil properties like soil organic carbon and environmental covariates. Recently GWR has received increased attention because of its ability to account for local relationships between the study variable and covariates (Brunsdon et al., 1998; Fotheringham et al., 2002). In GWR modeling it is assumed that neighboring observations have a stronger effect on the regression at a target point than observations at a greater distance. In GWR a distance decay function is applied to obtain local estimates of the regression coefficients (Tu, 2011).

The potentials of GWR for SOC mapping have been explored in various regional studies. Mishra et al. (2010) compared GWR with multiple linear regression (MLR) and regression kriging (RK). In the latter two methods the regression coefficients were global, i.e., these were assumed constant throughout the area. GWR led to a reduction in RMSE of 22% over MLR, but only 2% over RK. Zhang et al. (2011) and Wang et al. (2013) compared GWR with MLR. In both studies GWR outperformed MLR: in the former study the RMSE as obtained with MLR was reduced by 5%, in the latter study by 11%.

[*] Corresponding author.
E-mail address: glzhang@issas.ac.cn (G.-L. Zhang).

http://dx.doi.org/10.1016/j.geoderma.2015.06.024
0016-7061/© 2015 Published by Elsevier B.V.

Geographically weighted regression kriging (GWRK) (Harris et al., 2010; Kumar et al., 2012) and simple kriging with GWR-derived local means (GWRSK) (Harris and Juggins, 2011; Lloyd, 2010) are extensions of the GWR approach. In these approaches the spatial variation is modeled as the sum of a deterministic trend modeled by GWR, and spatially correlated residuals. Kumar et al. (2012) compared GWRK and RK, and found that RMSE as obtained by GWRK was reduced by 43% due to the local regression coefficients.

Other studies have demonstrated that GWR models not always outperformed geostatistical models assuming global regression coefficients (Harris and Juggins, 2011; Harris et al., 2010; Lloyd, 2010). Also, we are not aware of studies of SOC mapping in which GWR and GWRSK are compared with kriging with an external drift (KED) using restricted maximum likelihood (REML) estimation of the model parameters. In RK the model parameters are estimated in two separate steps. In the first step the regression coefficients are estimated by ordinary least squares assuming independent data. In the second step the variogram parameters are estimated by method-of-moments from the regression model residuals. Ideally these steps are repeated until convergence, with generalized least squares estimation of the regression coefficients. This estimation procedure is known to be suboptimal; the model parameters can best be estimated by REML (Lark and Webster, 2006). Suboptimal estimates of the model parameters may lead to suboptimal predictions. Therefore we prefer to compare GWR and GWRSK with KED–REML (hereafter shortly denoted as KED) in order to assess the benefit of local regression coefficients in mapping SOC.

The aim of this study was to quantify the benefit of using local, geographically weighted regression coefficients instead of global coefficients in mapping SOC in a study area with complex topographical conditions, under two modeling assumptions. In the first modeling assumption spatial dependence of data is ignored (data are assumed independent), whereas in the second assumption the spatial dependence of data is part of the model. The benefit of local regression coefficients assuming independent data is quantified by comparing various quality indices among which the RMSE of GWR (local coefficients) and MLR (global coefficients), whereas the benefit when accounting for spatial dependence of data is quantified by comparing the quality indices of GWRSK (local coefficients) and KED (global coefficients).

2. Study area and data

2.1. Study area

As the second largest inland river of China, the Heihe River is 821 km long, originating from the Qilian Mountains and flowing into the western Inner Mongolian Plateau (Lu et al., 2009; Wu, 2011). The study area, consisting of the upper and middle reaches of the Heihe River Basin, stretches for 340 km from the northwest to the southeast (Fig. 1), with coordinates of about 97°20′–101°51′ E and 37°41′–39°59′N. The major part of the study area is in Gansu Province and a small part in Qinghai Province. The upper reaches are in the southern part of this area (Fig. 1) with an average elevation of 3556 m a.s.l. Most peaks are higher than 4000 m a.s.l. The average elevation of middle reaches is about 1811 m a.s.l. This study area covers about 50,810 km^2, accounting for about 36% of the total area of the Heihe River Basin.

The mean temperature of this area during winter and summer is −3 and 7 °C, respectively. Annual precipitation varies spatially from 82.8 mm to 425.6 mm, with rainfall occurring mainly from June to September. The land covers of this area are heterogeneous due to the wide range of elevation and strong anthropogenic activities in terms of irrigation farming. The upper reaches are characterized by a humid and cold climate, whereas the climate of the middle reaches is a typical temperate arid environment with low precipitation and high evaporation (Li et al., 2012). While the annual precipitation is very limited in the middle section of the Heihe River Basin, water shortage is the major obstacle to the crops that rely mainly on irrigation (Kang et al., 2004).

2.2. Soil sampling

Part of the soil data (2010–2013), 144 soil points, were provided by “Heihe Plan Science Data Center, National Natural Science Foundation of China” (http://www.heihedata.org). To collect additional data we selected 404 sampling locations by purposive sampling, using the method of Zhu et al. (2008). In this method representative soil sites are selected from soil-scape units constructed by fuzzy *c*-means classification of pixels on the basis of the soil forming factors (covariates). The additional soil sampling was conducted in July 2012 and July 2013. All the sampling sites were located by handheld global positioning system (GPS) receivers. A 100–150 cm deep soil pit was dug at each site. In total 548 topsoil (0–20 cm) samples were collected and stored in a digital database (Fig. 1). The samples were subsequently dried, sieved at 2 mm and analyzed using the Walkley–Black procedure for SOC.

The observed SOC contents in surface soils varied from 0.70 to 132.19 g/kg, with a mean value of 31.33 g/kg (Table 1). The frequency distribution of SOC showed strong positive skew; the skewness was 1.50 (Fig. 2, Table 1). For all prediction methods we therefore transformed SOC measurements by taking the natural logarithms (LnSOC). The skewness dropped to 0.15 (Table 1).

2.3. Predictors

Predictors utilized in this study for the mapping of SOC content of 0–20 cm topsoil (g/kg) were slope, aspect, elevation, profile curvature, plan curvature, topographic wetness index (TWI), mean annual air temperature (MAAT), mean annual precipitation (MAP), solar radiation, soil type maps, land use maps, and normalized difference vegetation index (NDVI).

The terrain attributes slope, aspect, profile curvature, plan curvature and TWI were derived from a digital elevation model (DEM) which was obtained from the CIAT (International Centre for Tropical Agriculture) SRTM (Spaceshuttle Radar Topographical Mission) website (http://srtm.csi.cgiar.org). The SRTM DEM data (Jarvis et al., 2008) were georeferenced from three arc second resolution to 30 m × 30 m resolution. Slope, aspect and SAGA TWI (System of Automated Geoscientific Analyses Topographic Wetness Index) were extracted in SAGA GIS (SAGA Development Team, 2008). SAGA TWI was calculated based on the following equation (Moore et al., 1993):

$$\mathrm{TWI} = \ln\left(\frac{\alpha}{\tan\beta}\right) \qquad (1)$$

where α is the accumulative upslope area per unit contour length (or specific catchment area) computed with the D8 algorithm (O'Callaghan and Mark, 1984), and β the local slope gradient. SAGA TWI will assign a more realistic, higher potential soil wetness than the TWI$_{\mathrm{D8}}$ to grid cells situated in valley floors with a small vertical distance to a channel (Boehner et al., 2002). The cosine function was selected to transform aspect data to the range from −1 to 1 indicating the degrees of north. Potential insolation (incoming solar radiation), expressed in kW m^{-2}, mainly depends on elevation, slope and aspect, and thus can be derived directly from DEMs. Average annual potential insolation was calculated for the study area by means of the Solar Radiation function in ArcGIS 10.0 (ESRI, 2012). One year instead of a long-term average potential insolation was calculated, as the potential insolation only changes based on the scales of obliquity (Kunkel et al., 2011).

Beijing-1 multispectral data were obtained from Watershed Allied Telemetry Experimental Research (Li et al., 2009) at 32 m spatial