Geoderma 281 (2016) 69–82

Contents lists available at ScienceDirect

Geoderma

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

Mapping soil organic matter concentration at different scales using a mixed geographically weighted regression method

Canying Zeng [a,c], Lin Yang [b,*], A-Xing Zhu [a,b,c,d], David G. Rossiter [a,e], Jing Liu [d], Junzhi Liu [a], Chengzhi Qin [b], Desheng Wang [a]

[a] Key Laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University, 1 Wenyuan Road, Nanjing, Jiangsu 210023, China
[b] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[c] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, 1 Wenyuan Road, Nanjing, Jiangsu 210023, China
[d] Department of Geography, University of Wisconsin-Madison, Madison, USA
[e] School of Integrative Plant Sciences, Section of Soil & Crop Sciences, Cornell University, Ithaca, NY 14850, USA

ARTICLE INFO

Article history:
Received 21 January 2016
Received in revised form 22 April 2016
Accepted 27 June 2016
Available online 8 July 2016

Keywords:
Mixed geographically weighted regression (MGWR)
Geographically weighted regression (GWR)
Multiple linear regression (MLR)
Soil organic matter concentration (SOM)

ABSTRACT

The present regression models in digital soil mapping usually assume that relationships between soil properties and environmental variables are always fixed (as in MLR) or varying (as in GWR) in geographical space. In reality, some of the environmental variables may be fixed in affecting soil property variation and some are local varying. In this study, a mixed geographically weighted regression (MGWR) method which can deal with fixed and varying spatial relationships between a target variable and its environmental variables were proposed and used to predict topsoil soil organic matter (SOM) concentration in two study areas (Heshan, Heilongjiang province and Xuancheng, Anhui province, China) at two scales. Three groups of sample sets were created based on the total samples in the study areas to evaluate the robustness and stability of the model. Multiple linear regression (MLR), geographically weighted regression (GWR), GWR-kriging (GWRK), local regression-kriging (LRK), kriging with an external drift (KED), and ordinary kriging (OK) were used for comparison with MGWR. The validation results showed that the use of MGWR reduced the RMSE of GWR by 10.5% and 7.6% on average, reduced the RMSE of MLR by 12.8% and 9.9% on average for Heshan and Xuancheng study areas respectively. MGWR also showed a good competitiveness when compared with GWRK, LRK, KED and OK. In Heshan study area, the influence of flow length, relative position index, foot slope and distance to the nearest drainage were constant, whereas the elevation, topographic wetness index and valley index showed different influence in different regions. In Xuancheng study area, the fixed environmental variables were profile curvature, topographic wetness index and slope, whereas the varying environmental variables were precipitation, temperature, elevation, and limestone. The results indicate that the accuracy of predictions can be improved by adaptive coefficient according to the variation of environmental variables as implemented in MGWR compared with others considering only the local or global relationships. It was concluded that mixed geographically weighted regression model could be a potential method for digital soil mapping.

1. Introduction

Knowledge of soil spatial variation is essential for ecological processes modeling (Li, 2010; Du et al., 2015). Soil has long been considered as the result of the interaction of its formative environment, including climate, parent material, terrain, and vegetation conditions (Winklerprins, 1999; Mcbratney et al., 2003; Yang et al., 2008; Stoorvogel et al., 2009). Therefore, the relationships between soil and its environmental covariates can be used to map soil variations over space (Thompson et al., 2006; Sumfleth and Duttmann, 2008; Zhu et al., 2010).

Numerous methods have been developed to predict soil spatial distribution based on the relationships between soil and its environmental covariates (Odeh and Mcbratney, 2000; Zhang et al., 2012; Zhao et al., 2014). Multiple linear regression (MLR) is one of those commonly-used methods in early time (Odeh and Mcbratney, 2000; Zornoza et al., 2007; Qiu et al., 2003; Chung and Alexander, 2002; Moore et al., 1993), and is usually used as a basic model comparing with other mapping methods (Zhu et al., 2010; Lesch et al., 1995; Qin et al., 2012; Song et al., 2016). The desirability for linear regression methods lies in its simplicity, easy to apply. However, it assumes that the relationships are multivariate linear and the same for the whole area. This is a strong assumption, especially over large areas. Although the relationships generated using methods such as decision tree and random forest are not the

* Corresponding author.
E-mail address: yanglin@lreis.ac.cn (L. Yang).

http://dx.doi.org/10.1016/j.geoderma.2016.06.033
0016-7061/

same or linear, these methods ignore the local spatial autocorrelation of soil properties (Henderson et al., 2005; Reza Pahlavan Rad et al., 2014; Taghizadeh-Mehrjardi et al., 2014). To consider both the relationships between soil properties and environmental covariates and spatial autocorrelation effect of soil property itself, regression kriging (RK), Kriging with External Drift (KED) were developed and widely used (Hengl et al., 2004; Hengl et al., 2007; Mishra et al., 2012; Bishop and Mcbratney, 2001). RK and KED combine a regression (usually multiple linear regression, but can also be a random forest etc.) of a target soil attribute on its environmental co-variables with kriging of the regression residuals. Many studies show higher accuracy than feature space-only models, due to the inclusion of spatial autocorrelation of model residuals in the models (Brus and Heuvelink, 2007; Odeh and Mcbratney, 2000; Sumfleth and Duttmann, 2008; Sun et al., 2012). However, the model residuals can't always tally with the first order or second order stationarity and seldom of these methods can be adapted to fit data locally with varying coefficients over space for the regression (Kumar et al., 2012; Walter et al., 2001).

To deal with the spatial non-stationarity of regression coefficients between a target variable and explanatory variables, geographically weighted regression (GWR) was developed to estimate varying coefficients of explanatory variables locally (Brunsdon and Fotheringham, 1999). Coefficients of GWR at each prediction point are estimated using a weighting matrix in which observations around a sampling point are weighted using a distance decay function, meaning that closer observations have a greater effect on the resulting localized regression coefficients. The spatial dependence of soil is accounted for by using the distance decay function. Many applications of GWR have shown good results for spatial non-stationarity modeling of soil variation (Mishra and Riley, 2012; Song et al., 2016; Wang et al., 2013; Zhang et al., 2011). However, the relationships between soil and some environmental covariates may be constant, not always varying as modeled in GWR in a given study area. For example, the effect of climate variables such as precipitation would be consistent for soil texture variation at a small watershed. Attempting to fit varying relations when they are not in fact present will be fitting random noise and thus result in poorer models. Further, model interpretation in terms of related or causative factors will be misleading.

Mixed geographically weighted regression (MGWR), an extension of GWR, was proposed to determine which predictors are fixed and varying over space (Fotheringham et al., 2002). It has been successfully applied in economics and agriculture (Mei et al., 2004; Pecci and Sassi, 2008; Qin et al., 2007; Wei and Qi, 2012). Qin et al. (2007) developed an iterative algorithm to estimate fixed and varying coefficients in MGWR and further tested it by using average prices of house blocks in Shanghai. MGWR gave superior model fits to GWR. In this test case the population density and unemployment rate showed fixed relations to the house prices over geographic space, whereas the distance to subway stations, greening rate and etc. varied.

MGWR has not been applied in soil mapping, although it would seem to be a promising approach due to its ability to discern fixed or constant independent variables. The relationships between soil and its environmental co-variables are known to be scale-dependent and varying in different soil landscapes. It is clear that not all environmental co-variables affect the soil variable at the same scale. Think for example of regional climate, compared to local topographic effects. For a regional scale area, it is possible that the impact of even climate variables for soil is varying over the area. While for a small watershed area, climate variables would affect soil variation consistently for the area. As for topographic variables, the relationships between soil and topographic variables are usually varying. While for low relief areas with small coverage, the relationship between soil and some topographic variables may be consistent. Therefore, it is not appropriate always use local coefficients to describe the relationships as in GWR and an environmental variable will be not always fixed or varying in different regions depend on the characteristics of the study area. MGWR is able to determine which variables are fixed or varying over a given study area. It can be a potential method in digital soil mapping. Such model may lead to a new interpretation of the phenomena soil variation by considering both spatially stationary and non-stationary effects.

This paper aims to introduce MGWR into digital soil mapping. The method was used to predict the A-horizon soil organic matter (SOM) concentration in two study areas with different sampling densities, one with gentle terrain at watershed scale and the other with complex environmental conditions at regional scale. Fixed variables and varying variables for SOM in the two study areas were detected using MGWR. And to test the effectiveness of MWGR, GWR, MLR, GWRK (GWR-kriging), LRK (local regression-kriging), KED (kriging with an external drift) and OK (ordinary kriging) were also applied to compare with MGWR.

2. Study areas and data

2.1. Study areas

Two study areas were selected with environmental conditions of different complexities at different scales (Fig. 1). The first study area is Heshan farm at a watershed scale with area about 60 km^2, located in Nenjiang County of Heilongjiang province, China (48°43′ to 49°03′ N and 124°56′ to 126°21′ W). This study area have a gentle environmental gradient with slope gradient mainly under 4° and elevation varying from 276 m to 363 m above sea level. The land use is mainly cropland including soybean and wheat cropland. The parent materials are mainly silt loam loess over the whole area except in the valley bottom, which is mainly occupied by fluvial deposits. The average annual precipitation is about 500–600 mm and the annual temperature varies from −38 °C to 36 °C with very cool climate in winter and warm climate in summer (Zhu et al., 2015). Land use, climate and parent materials in this study area have been fairly uniform (Yang et al., 2013; Zhu et al., 2010).

The second study area is Xuancheng county at a regional scale, located in Anhui province of China with an area of 5900 km^2 (29°57′N-31°19′N and 117°58′W-119°40′W). The average annual precipitation in this area is 1200–1800 mm and the annual temperature is 11–16 °C, with cool and dry climate in winter and warm and humid climate in summer. The northwest of the area is low-relief and other areas are mountainous. “Land use mainly involves cultivated land with irrigated rice as the dominant crop, and secondary or planted forest land covered by bamboo, fir, shrub and other evergreen coniferous or deciduous broad-leafed trees. The soil parent materials in the study area are complex, including shale, sandstone, pyroclastic rocks, granite and granodiorite, limestone, conglomerate, quaternary clay-silt-gravel, quaternary vermicule boulder and grave clay” (Yang et al., 2016). The soil forming environment of this study area is complex mainly because of the complicated soil parent materials and variable landforms including plains, hills, and low mountains etc.

A-horizon SOM concentration in g/kg fine soil (<2 mm particle size) was the target soil property.

2.2. Environmental data

In the Heshan study area landform is the main environmental factor controlling the soil formation. Seventeen topographic factors were generated from the DEM using a terrain analysis software SimDTA (Qin et al., 2009a, 2009b) or Arcgis 10.1 (Table 1). A 10 m horizontal resolution Digital Elevation Model (DEM) was created from the 1:10,000 scale topographic map published by Chinese Bureau of Surveying and Mapping (1987).

In the Xuancheng study area, topographical, parent materials, vegetation and climate play important roles for soil formation (Yang et al., 2016). Twenty-four environmental factors were generated in this study area (Table 1). The topographic factors were derived from the DEM which obtained from the Shuttle Radar Topographic Mission