Applied Geochemistry xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Applied Geochemistry

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

Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland

Chaosheng Zhang [a,*], Ya Tang [b], Xianli Xu [c], Ger Kiely [c]

[a] GIS Centre, Ryan Institute and School of Geography and Archaeology, National University of Ireland, Galway, Ireland
[b] Department of Environmental Sciences, Sichuan University, Chengdu, Sichuan 610065, China
[c] Department of Civil and Environmental Engineering, University College Cork, Ireland

ARTICLE INFO

Article history:
Received 22 November 2010
Accepted 12 April 2011
Available online xxxx
Editorial handling by Dr. R. Fuge

ABSTRACT

It is challenging to perform spatial geochemical modelling due to the spatial heterogeneity features of geochemical variables. Meanwhile, high quality geochemical maps are needed for better environmental management. Soil organic C (SOC) distribution maps are required for improvements in soil management and for the estimation of C stocks at regional scales. This study investigates the use of a geographically weighted regression (GWR) method for the spatial modelling of SOC in Ireland. A total of 1310 samples of SOC data were extracted from the National Soil Database of Ireland. Environmental factors of rainfall, land cover and soil type were investigated and included as the independent variables to establish the GWR model. The GWR provided comparable and reasonable results with the other chosen methods of ordinary kriging (OK), inverse distance weighted (IDW) and multiple linear regression (MLR). The SOC map produced using the GWR model showed clear spatial patterns influenced by environmental factors and the smoothing effect of spatial interpolation was reduced. This study has demonstrated that GWR provides a promising method for spatial geochemical modelling of SOC and potentially other geochemical parameters.

1. Introduction

It is challenging to perform spatial geochemical modelling due to the spatial heterogeneity feature of geochemical variables caused by multiple environmental factors such as rock type and soil type. Jordan et al. (2007) attempted to separate the influences of various factors and investigated the probability features of surface soil geochemistry in Northern Ireland, and acknowledged that "it remains a challenging task in geochemistry to separate all the factors and to model their influence at the regional scale."

One of the important geochemical variables is soil organic C (SOC) which is not only related to soil fertility (Tisdale et al., 1995) but also plays an important role in climate change. Small changes in SOC may have a major impact on CO_2 in the atmosphere (Lal, 2003). High quality maps of SOC are needed not only to provide guidance for practical soil management but also to enable more accurate calculations of C stocks. With the development of computer software packages, especially integration in a geographical information system, spatial interpolation techniques such as kriging (Webster and Oliver, 2007) and inverse distance weighted (IDW) are widely applied in soil geochemistry for the production of spatial distribution maps of soil parameters (Zhang, 2006). One of the main features of the traditionally used spatial interpolation techniques is that they generally focus on their mathematical power (e.g., how to keep the estimation errors to a minimum), but seldom include environmental factors in the modelling.

Many studies have shown relationships between environmental factors and soil properties (Odeh et al., 1995; McKenzie and Ryan, 1999). More specifically, some studies have tried to map SOC distribution using secondary information such as land use, soil type, lithology, topography and other environmental factors (Mueller and Pierce, 2003; D'Acqui et al., 2007; Rawlins et al., 2009; Schulp and Verburg, 2009).

Developments in kriging analyses have tried to incorporate auxiliary information such as co-kriging (McBratney and Webster, 1983; Odeh et al., 1995) and regression kriging for spatial interpolation for soil parameters (Odeh et al., 1995; Rawlins et al., 2009). Another possible way of incorporating environmental factors for spatial interpolation is to use the geographically weighted regression method (GWR) (Fotheringham et al., 2002) which is receiving increased attention. Strong spatial variation becomes a challenge for the estimation of SOC at un-sampled sites, e.g., a function at the global scale of a study area is not sufficient to address the spatially varying relationships at the local level. A GWR approach is useful when the assumption of spatial stationarity of the relationship between dependent variable and independent variables is invalid, which is demonstrated by Osborne et al. (2007) on species distribution. Scull (2010) showed that GWR performed better than the ordinary multiple linear regression (MLR) for the prediction of

* Corresponding author. Fax: +353 91 495505.
E-mail address: Chaosheng.Zhang@nuigalway.ie (C. Zhang).

0883-2927/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apgeochem.2011.04.014

several surface soil properties including SOC in the USA, using a climate parameter (precipitation or temperature) as the independent variable. Recently, Mishra et al. (2010) applied the GWR method for prediction of spatial variation of the SOC pool at a regional scale in the Midwestern USA.

The technique of GWR is a variant of MLR with a weight function included, which only takes the samples within a defined neighbourhood (band width or number of samples) into calculation and also may weigh the contributions of closer samples more than those farther away. In this case, it is possible to have separate input samples and output parameters from each case under calculation, honouring the feature of spatially varying relationships at the local level.

This paper explores the application of GWR for spatial geochemical modelling using SOC in Ireland as an example, and tries to demonstrate the procedures in an easy and concise way. To assess the performance of GWR, three other methods of ordinary kriging (OK), IDW and MLR were chosen for comparison. The OK method was chosen as it was previously used for production of SOC map using the National Soil Database of Ireland (Zhang et al., 2008a,b). The IDW is popularly used in environmental geochemistry and easy to implement. The reason why MLR was chosen was because of its "global" feature in comparison with GWR. It should be noted that this paper does not attempt to find out the best spatial modelling or interpolation methods which could be an endless task and beyond the scope of this study. The focus of this study was to explore the application of the GWR method for spatial geochemical modelling.

The objectives of this study included: (1) to integrate environmental variables in the GWR model for spatial geochemical modelling of SOC in Ireland; (2) to assess the performance of GWR; and (3) to discuss the features and uses of the GWR method in spatial geochemical modelling.

2. Materials and methods

2.1. Sampling and analysis for SOC

The Republic of Ireland has a total land area of 71,000 km^2. It is traditionally an agricultural country with 66.4% of the total land cover being used for agriculture and 18.69% being peat bogs and wetlands (EPA, 2009). Blanket peat is distributed in the mountain areas mainly along the western coastline of Ireland. There is also basin peat located in the lowland areas, mainly in the central part of midland Ireland.

The SOC data were retrieved from the National Soil Database of Ireland (Fay and Zhang, 2007). A total of 1310 surface (0–10 cm) soil samples were collected from predetermined positions based on a grid sampling scheme at a density of two samples per 100 km^2 unless the predetermined location was not accessible (see Fig. 2). A total of 1179 samples (90% of the total number) were randomly selected from the database for spatial analyses, and the remaining 131 samples were reserved for evaluation of the performances of the methods. Detailed information for sampling and SOC analyses is available in Fay et al. (2007).

2.2. Generation of rainfall map of Ireland

The map of the 30-a average annual rainfall of Ireland (1971–2000) was created for this study using a regression model (Sweeney et al., 2003) plus IDW interpolation for the regression residuals.

A system of 500 m $\times$ 500 m grid was created using the coastline of Ireland, and this standard grid system formed the basis for raster mapping in this study. Each grid point received an elevation value from the nearest point of SRTM (Shuttle Radar Topography Mission, V. 4, Jarvis et al., 2008) data of Ireland using the spatial join function of a GIS. The spatial resolution of 500 m used this study was considered good enough to reveal the spatial patterns of variables under study.

The raw point data of rainfall were acquired from Met Éireann (191 stations for Republic of Ireland) and Met Office of the UK (49 stations for Northern Ireland). To estimate the rainfall value for each of the standard grid points, a regression model including trend surface and elevation was adopted. It was found that the 2nd order polynomial regression equation (including variables x, y, xy, x^2, y^2) and elevation was most applicable for the available data in Ireland (Sweeney et al., 2003). A stepwise regression analyses showed that all the variables were significant ($p < 0.001$, except for xy at $p = 0.023$). Due to a problem of collinearity, the variable xy was excluded from the regression model, and the following function was fitted and applied for the production of the trend map of rainfall in Ireland during 1971 and 2000:

$$\text{Rainfall} = 2432.713 - 8.859x - 3.128y + 0.014x^2 + 0.008y^2 + 2.092z \qquad (1)$$

where x is the easting coordinate in km, y is the northing coordinate in km, and z is the elevation above the sea level in m. The adjusted R-square for this regression analysis was 0.756. It is noted that this regression model captured the "trend" of rainfall. Other issues such as rainfall shadow cannot be well modelled in this way. Therefore, the residuals of the regression model for the 240 locations were utilized to create the residual map based on the IDW method with 12 points as the neighbourhood and the power of 1. The "trend" map and the "residual" map were combined using the "plus" calculation for each grid point to produce the final 30-a average annual rainfall map of Ireland during 1971 and 2000 for use in this study.

2.3. GWR and MLR methods

The regression analysis is used to model the relationship between one variable and one or more other variables. Based on Fotheringham et al. (2002), an ordinary (global) MLR model can be written as:

$$y_i = \beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i \qquad (2)$$

where y is the dependent variable with the value y_i at the ith location, x_k are the independent variables with the value x_{ik} at the ith location, k is the number of independent variables, β_0 and β_k are the parameters to be estimated, and ε_i is an error term.

The GWR takes the spatial locations of samples into consideration, permitting the estimated parameters to vary locally, thus better reflecting the spatially varying relationships between the dependent and independent variables. A GWR model can be written as (Fotheringham et al., 2002):

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \qquad (3)$$

where (u_i, v_i) denotes the coordinates of the ith location, and $\beta_0(u_i, v_i)$ and $\beta_k(u_i, v_i)$ are the estimated parameters for the ith location, whose values vary with the location. Therefore, compared with the ordinary regression model, the GWR is a local regression model, honouring local variation using spatially varying estimated parameters.

The parameters for the ordinary (global) MLR model can be estimated based on ordinary least square (OLS) as the following matrix format (Fotheringham et al., 2002):

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \qquad (4)$$

where $\boldsymbol{X}$ is the matrix formed by the values of variables x_k and $\boldsymbol{Y}$ is the vector formed by values of variable y.

For the GWR, the parameters can be estimated using a weighting function (Fotheringham et al., 2002):

$$\hat{\beta}(u_i, v_i) = (\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i)\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i)\boldsymbol{Y} \quad (5)$$

where $\boldsymbol{W}(u_i, v_i)$ are the weights which are chosen so that those observations near the point under study have more influence on the results than those farther away. The weight function for GWR in this study was selected as the adaptive spatial kernel type (Fotheringham et al., 2002), so that the spatial extent for samples included in the GWR varied based on sample density. This was useful for locations in the coastal and border areas where there was a border effect (without samples located outside the study area). The distance band width was determined using the Akaike Information Criterion (AICc) which was effective in finding the optimal band width in GWR (Fotheringham et al., 2002). The optimal number of samples was 194 for the dataset with 1179 samples, and it was 233 for the whole dataset. When the band width was too short or the number of samples was too few, one of the main problems for GWR called local multicollinearity became unavoidable, making the GWR results unreliable. The multicollinearity is a well-known problem in regression analyses, which happens when there is redundant information among two or more independent variables. In GWR, the dummy variables for a categorical variable (e.g., soil type and land cover) could become "redundant" locally when the categories are spatially clustered in some areas with some categories missing in other areas. A suggestion would be to avoid using categorical variables that have values spatially clustered, or to combine and reduce the number of categories, or to combine it with other variables.

2.4. Kriging and IDW methods

Kriging and IDW are perhaps the most popular spatial interpolation methods in environmental applications. Both methods take neighbouring known samples into consideration, and give different weights to different samples. The weights for samples in IDW decrease with the increase of distances between the known samples and the point to be estimated, and the rate of decrease is proportional to "inverse distance". The main problems for an IDW are that the decision for the "rate" of decreasing weight (defined as the power parameter of distance) and the number of neighbours to be included is to some extent arbitrary. On the other hand, the weights in kriging are decided based on the spatial structure parameters of a variogram which measures the relationships between squared differences between paired samples and their distances (Webster and Oliver, 2007). In this study, the OK was included as this method was applied previously for the same dataset (Zhang et al., 2008b).

2.5. Input parameters for GWR

The category variables used in a regression model were replaced by dummy variables with the values of either 1 or 0, showing the presence or absence of each category. To avoid the local multicollinearity problem and to make each category variable meaningful, it was necessary to simplify the land cover and soil type data to be used in this study.

The land cover data used in this study were the Coordination of Information on the Environment (CORINE) Land Cover 2006 data provided by the Irish Environmental Protection Agency as part of the EU programme (EPA, 2009). The highest level (level 1) of CORINE classification has five classes: artificial surfaces, agricultural areas, forest and semi-natural areas, wetlands and water bodies. The 1310 SOC values were classified into the first four classes, and their differences were tested using analysis of variances (ANOVA) following a normal score transformation, which was effective in pushing the data towards normality via ranking the raw data and assigning the corresponding normal scores to each of the raw data values. The four classes were further simplified using a multiple comparison of Duncan's method (Duncan, 1955) following the test for homogeneity of variances using the Levene's test, and classes in the same group were combined into a new category.

The soil type data was acquired via digitization of the Soil Association Map of Ireland (Gardiner and Radford, 1980) by Teagasc, Ireland. There were a total of 44 soil associations which were simplified to 9 soil types: Podzols, Brown Podzolics, Grey Brown Podzolics, Acid Brown Earths, Gleys, Brown Earths, Rendzinas, Lithosols and Peat (Gardiner and Radford, 1980). The same procedures of ANOVA and a multiple comparison were performed for the soil type data for further simplification. It was observed that "peat" was regarded as a separate group in both classifications of land cover (with the term of "wetlands") and soil type, but their spatial distributions were different. Since the soil type "peat" was further combined with other soil types, "peat" or "non-peat" was not separately treated in this study. Furthermore, harmonizing the "peat" distribution in the two different maps from different sources could be a challenging task, which is beyond the scope of this study.

Three environmental factors were included in the GWR in this study: rainfall, land cover and soil type. Rainfall was a scale variable and was used as an independent variable directly. Land cover classes were simplified into three categories: (1) wetlands, (2) forest and semi-natural areas, and (3) the others. Soil type classes were grouped into two main categories: (1) Podzols, Rendzinas, Lithosols and peat, (2) the others. Detailed justification for such grouping is provided later.

Therefore, two dummy variables were needed for land cover, and one dummy variable was used for soil type. The input parameters for GWR were: rainfall, land cover dummy 1, land cover dummy 2, and soil type dummy. The 1310 soil samples and all the 500 m grid points were given rainfall values from their nearest grid points of the rainfall map. They received the values of "1" for land cover dummy 1 if they were located in wetlands, and "0" otherwise. The values for land cover dummy 2 were "1" if the points were located in forest or semi-natural areas, and "0" otherwise. The values for soil type dummy were "1" if the points were located in Podzols, Rendzinas, Lithosols or peat areas, and "0" otherwise.

2.6. Data transformation and computer software

To alleviate the problems of non-normality and skewness of the raw data (Zhang et al., 2008a), a Box–Cox transformation (Box and Cox, 1962) was applied to SOC data for kriging, GWR and MLR analyses. The transformed data were also used for IDW for consistency. All the spatial analyses were carried out and all maps were produced using ArcGIS® (version 10) software, and conventional statistical analyses were performed using SPSS® (version 18) and Microsoft Excel® (version 2007). The basic GIS data were acquired from Ordinance Survey Ireland (OSI).

3. Results and discussion

3.1. SOC in soils of Ireland and its relationships with land cover and soil type

The basic statistics for SOC in Ireland (n = 1310) are available in Zhang et al. (2008a). It varied from 1.40% to 55.80%, with a median value of 7.00%. Meanwhile, the histogram of SOC exhibited two clear peaks (Fig. 1), indicating two distinct soil types: the mineral

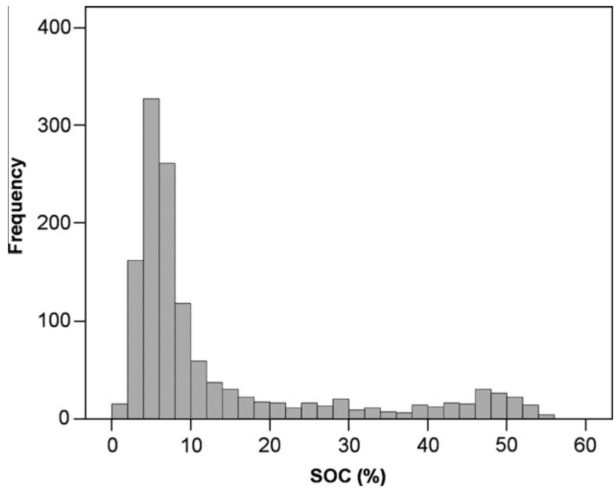

Fig. 1. Histogram of SOC in Ireland.

soils and organic soils (peat) (Zhang et al., 2008a), and implying the influence of land cover and soil type.

The symbol map for SOC in Ireland (Fig. 2) showed clear spatial patterns: high values in the west and midland and low values in the east. Western Ireland, especially along the west coastal areas, is mountainous with extensive cover of blanket peat. In the midland areas, scattered basin peat exists in the lowland areas. In eastern Ireland, soils are mainly mineral where agricultural activities are more intensive, except for some small areas such as the Wicklow Mountains (see Fig. 4). The good spatial relationship between SOC and peat is clearly shown. In this map, the spatial locations of 131 (10% of the total) samples randomly selected for the following validation of models are shown as crosses.

The influence of land cover and soil type on SOC can be demonstrated via box-plots (Fig. 3). Wetlands showed the highest SOC content, followed by forest. For soil type, elevated SOC values were observed in Lithosols, peat, Podzols and Rendzinas. The high SOC values in areas of Lithosols and Rendzinas could be related to the sampling procedure: soils were collected from rock cracks where it was possible to collect soil samples. There were quite a few outliers in several groups, especially in agricultural areas, Acid Brown Earths, Brown Podzolics, Gleys and Grey Brown Podzols. Meanwhile, those which did not exhibit outliers had quite long box lengths (wide inter-quartile ranges), such as forest, wetlands, Lithosols and peat. These features demonstrated the strong variation and heterogeneity of SOC within each group, causing large uncertainties in modelling SOC using environmental factors. Such strong variability of SOC could be related to variation caused by other environmental factors within each group, the uncertainties in classification of the samples using maps, as well as possible errors during sampling and laboratory analyses.

While both land cover and soil type showed their influences on SOC, some of their groups shared similar central tendencies, e.g., the medians for agricultural areas and artificial surface groups were close. To simplify the grouping, ANOVA was applied following the normal score transformation. Significant differences ($p < 0.05$) were found in both land cover groups and soil type groups. Results of the Levene's test for homogeneity of variances for land cover groups and soil type groups were both insignificant ($p > 0.05$), with significance values of 0.185 and 0.064, respectively. The multiple comparison simplified the land cover into three groups with wetlands and forest areas in separate groups, and agriculture and artificial surfaces in the other group (detailed statistical results not shown here). Soil types were simplified into two main groups, with peat, Lithosols, Rendzinas and Podzols in one group, and the others in the other group. Such new grouping results were transformed into dummy variables for use in GWR. This provides justification for the input parameters in the GWR described earlier.

3.2. Relationship between SOC and rainfall

The 30-a (1971–2000) annual average rainfall map for Ireland is shown in Fig. 4. It needs to be mentioned that the recent 30-a average annual rainfall data were used here for exploration of the statistical relationship between SOC and rainfall, while rainfall has been affecting SOC for a much longer period. The rainfall largely reflected the topography of Ireland with high values (>1400 mm a^{-1}) in the mountain areas, mainly along the coastal areas of the west. In the midland and east (except for the Wicklow Mountains), rainfall was relatively low (<1000 mm a^{-1}).

The spatial distribution of rainfall map coincided well with that of SOC (Fig. 2), both of which had high values along the coastal areas in the west. However, differences existed in the midland and SE areas. In the midland lowland areas, there were scattered patches of basin peat causing high values of SOC where rainfall was not high. In the SE, intensive agricultural activities and lack of peat caused the generally low values of SOC, even though rainfall was relatively high. Some scattered high SOC values in the SE and east could also be attributed to scattered peat in elevated areas. Such spatially varying relationships between SOC and rainfall required spatial statistics at the local level, not at the global scale, making it reasonable to consider the methodology of GWR.

Another factor that could be considered is the elevation (Rawlins et al., 2009). Since the generation of the rainfall map of Ireland already included elevation, only one of them should be considered in GWR. The Spearman's correlation coefficients with SOC showed both were significant at the level of <0.01. The specific coefficients were 0.348 between rainfall and SOC, and 0.135 between elevation and SOC. Therefore, rainfall was included in the GWR modelling. Since the correlation coefficient between SOC and rainfall was still low, other influencing factors should be included in the regression model to achieve better results.

Based on the above analyses, the three factors of rainfall, land cover and soil type were included in the GWR modelling. Other factors, if found to play an important role on SOC, could also be considered in future studies. However, there will be an increasing possibility of multicollinearity when more independent variables are included in a regression model.

3.3. Comparison of performance of GWR with other methods

To investigate how useful GWR was for SOC spatial interpolation, its performance was compared with OK, IDW and MLR. The methods chosen for comparison were not supposed to be a complete list of available, advanced or complicated methods, as it was not the objective of this study to prove that the GWR method was better than the others. The aim was to determine if the GWR provided reasonable results, and if so to evaluate its features. The 1179 samples randomly selected from the database were used for spatial interpolation, and the remaining 131 samples were reserved for evaluation of the performances of these methods.

As with the whole dataset (Zhang et al., 2008b), a Box–Cox transformation was performed for the 1179 SOC values. The best power value for the 1179 values was −0.372, which was similar to the value of −0.376 for the whole dataset (Fay et al., 2007). For consistency, the power of −0.376 was also applied in this study. Specific variogram parameters were studied for the 1179 samples, and it was found that they were similar to the parameters for the whole dataset (Zhang et al., 2008b). For consistency, the variogram parameters used for the whole data set were applied

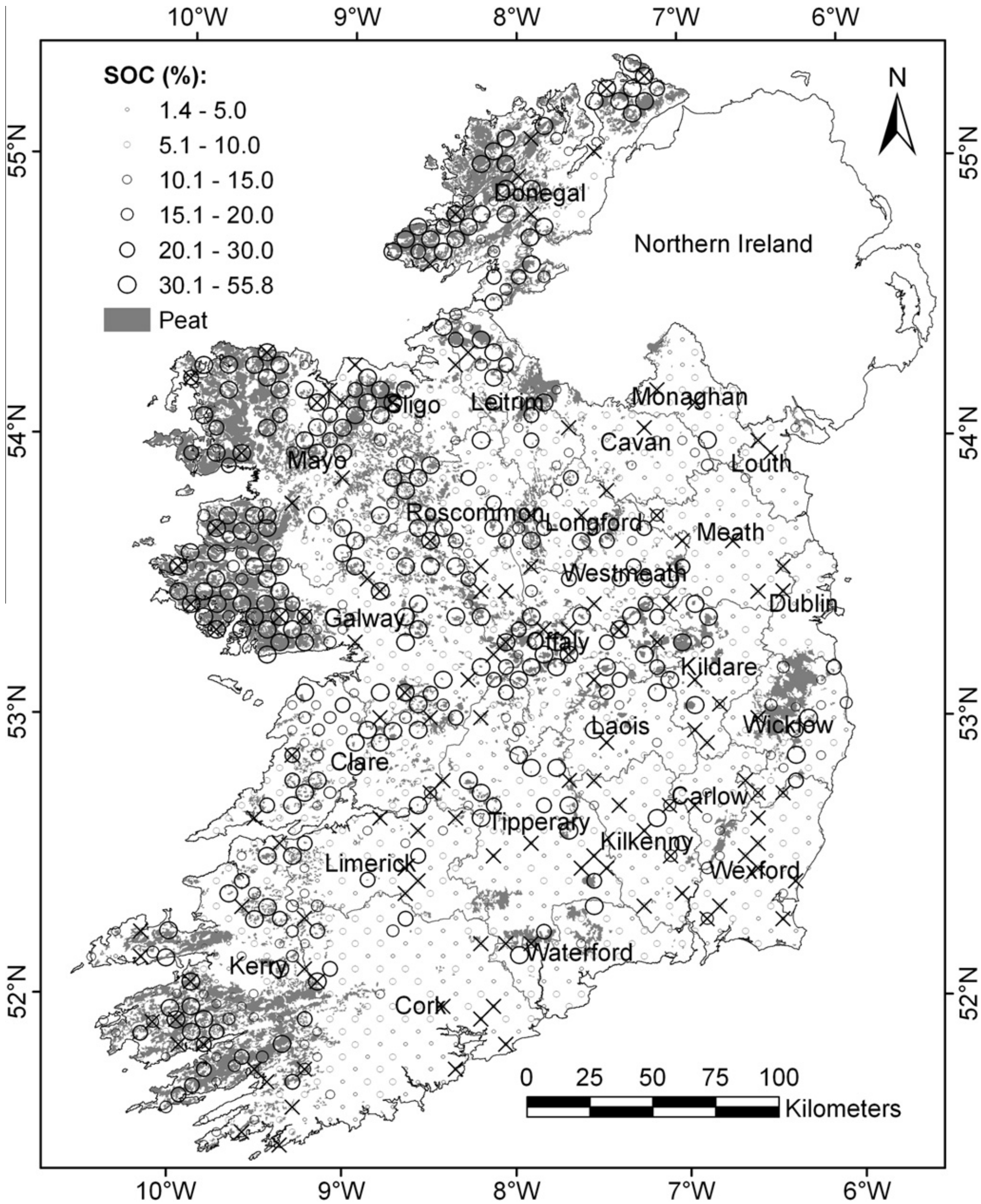

Fig. 2. Symbol map for SOC in Ireland overlaid on peat distribution (crossed samples are reserved for validation).

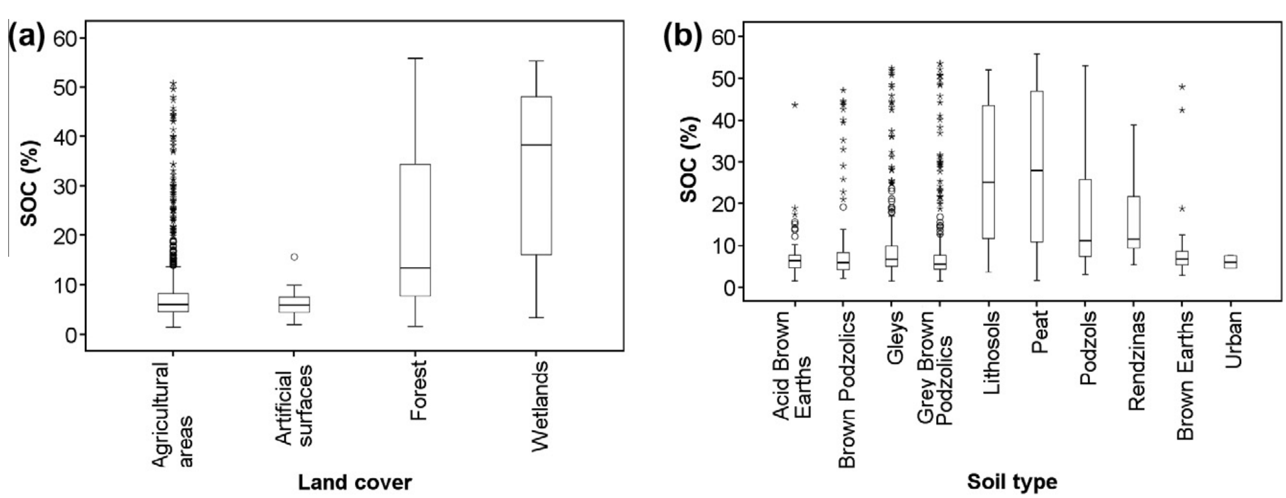

Fig. 3. Box-plots for SOC in different groups (a) land cover and (b) soil type.

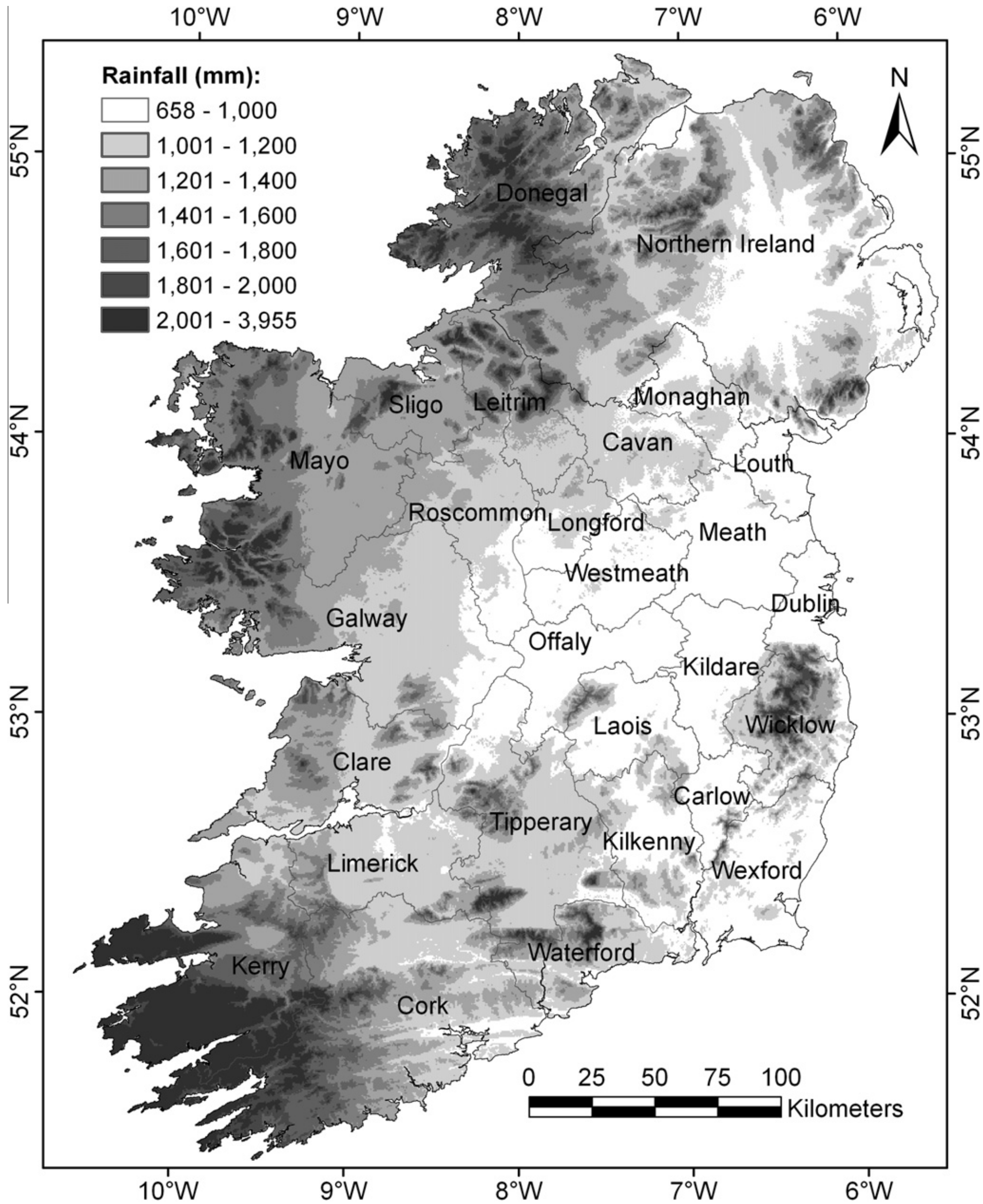

Fig. 4. Rainfall map of Ireland.

in this study. Then, a trans-Gaussian kriging (Cressie, 1993) was performed.

The performances of the spatial interpolation methods were evaluated using the 131 reserved SOC values. The predicted values were compared with the observed values. The error values were calculated as "predicted value – observed value" for all the 131 samples. Basic statistics for the error, absolute error and squared error are shown in Table 1. These statistics provide a robust measurement of the performances. Meanwhile, the commonly used root mean square error (RMSE) values are also listed.

All the minimum and maximum errors were high and similar, showing that all the methods over-estimated some small values and under-estimated some high values, and thus no method was perfect. The maximum error value of 42.74 was observed for GWR which could be related to the misclassification of the sampling location into "peat" using maps while it was mineral soil based on the observed data. This highlights the importance of quality of maps used in GIS analyses and subsequent environmental geochemical modelling. The IDW, GWR and MLR had the median

Table 1
Comparison of performances of GWR with other chosen methods (error, absolute error and RMSE units in % SOC, squared error unit in %-square SOC).

Method	Min.	25%	Median	75%	Max.	RMSE
Error						
GWR	−41.41	−3.40	0.14	1.84	42.74	NA
OK	−40.17	−3.81	1.87	5.41	19.44	NA
IDW	−43.91	−5.67	−0.07	2.67	17.02	NA
MLR	−42.15	−5.03	−0.14	2.11	20.04	NA
Absolute error						
GWR	0.01	0.97	2.15	7.54	42.74	NA
OK	0.03	2.28	5.18	8.90	40.17	NA
IDW	0.07	1.51	3.31	9.37	43.91	NA
MLR	0.02	1.19	2.37	9.40	42.15	NA
Squared error						
GWR	0.00	0.94	4.61	56.83	1826.31	10.99
OK	0.00	5.20	26.83	79.28	1613.50	11.50
IDW	0.00	2.29	10.98	87.82	1927.67	12.07
MLR	0.00	1.42	5.63	88.41	1776.21	11.57

errors of −0.07, 0.14 and −0.14 respectively, which were close to "0". For the absolute errors and squared errors, the GWR had the lowest 25th percentiles, median values and 75th percentiles. MLR also had a relatively good overall performance, as auxiliary information was used. Meanwhile, GWR also showed the lowest value of RMSE. Overall, these results demonstrated that the GWR did provide good and reasonable results in comparison with the other methods.

Another analysis on the validation results was the correlation coefficients between the measured and estimated SOC values for GWR, kriging, IDW and MLR. They were 0.646, 0.591, 0.593, and 0.619 using Pearson's method, and 0.629, 0.516, 0.519, and 0.579 using the Spearman's method, respectively. Even though all were significant at the level of $p < 0.01$, both methods demonstrated that GWR had the highest correlation coefficients among the four methods compared.

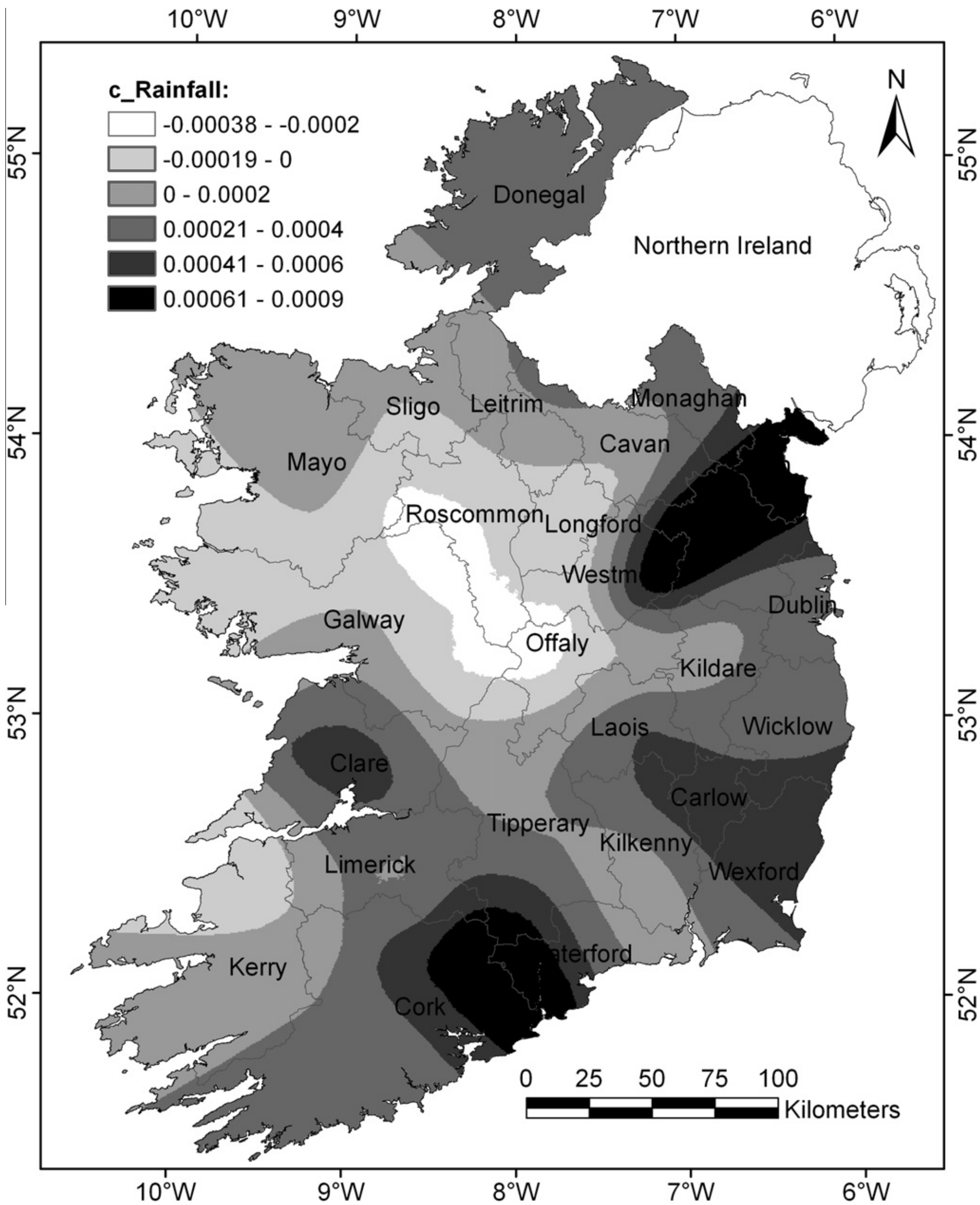

Fig. 5. Spatially varying coefficients for rainfall in the GWR model.

Table 2
Basic statistics for GWR model coefficients.

	Min.	5%	25%	Median	75%	95%	Max.
Intercept	0.176	0.528	0.882	1.041	1.344	1.555	1.822
C1_Rainfall	−0.00038	−0.00018	0.00002	0.00021	0.00035	0.00067	0.00090
C2_CLCDum1	−0.046	0.047	0.248	0.376	0.511	0.630	0.697
C3_CLCDum2	−0.017	0.008	0.093	0.217	0.278	0.382	0.446
C4_SoilDum	−0.071	0.001	0.098	0.198	0.234	0.289	0.343

Please cite this article in press as: Zhang, C., et al. Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Appl. Geochem. (2011), doi:10.1016/j.apgeochem.2011.04.014

3.4. Creation of spatial distribution map of SOC in Ireland using GWR

As explained earlier, a standard grid system at an interval of 500 m was created based on the boundary of Ireland. Each of the grid points acquired the rainfall, land cover and soil type information from GIS maps, and the dummy variables were created for GWR. Points with unmatched values obtained values from their nearest points with matched values.

To make use of all the available SOC data, all the 1310 values including the 131 values reserved for validation earlier were used for GWR modelling following the Box–Cox transformation. To justify the "spatially varying" relationship between SOC and the independent variables, the spatial distribution map showing coefficients for rainfall for the grid points was produced (Fig. 5) and the summary statistics for the GWR model coefficients for all the independent variables are listed in Table 2.

The spatially varying feature of coefficients for rainfall was clearly demonstrated (Fig. 5). The coefficients could be negative and positive, showing that the relationship between SOC and rainfall was different at different locations. Generally in the relatively small mountainous areas in the east, SE and NW, the relationship was positive. However, it was obvious that in the midland area where basin peat was scattered (rather than concentrated) the relationship between SOC and rainfall was negative: basin peat was located in low elevation areas with low rainfall, but the SOC concentrations were high. In the peat areas of the west, the coefficients for rainfall were close to "0", which should be related to the dominant role played by "peat" found in large mountainous areas.

For the other independent variables used in the GWR model (Table 2), they also varied from negative to positive values, and their strong variation was depicted by the differences between the minimum and maximum values. The results provided justification of using GWR for estimation of SOC in this study.

The GWR parameters were applied for estimating SOC values at the points of the 500 m grid system, and the values were back-transformed and converted to an ArcGIS GRID data for creation of the spatial distribution map of SOC in Ireland (Fig. 6).

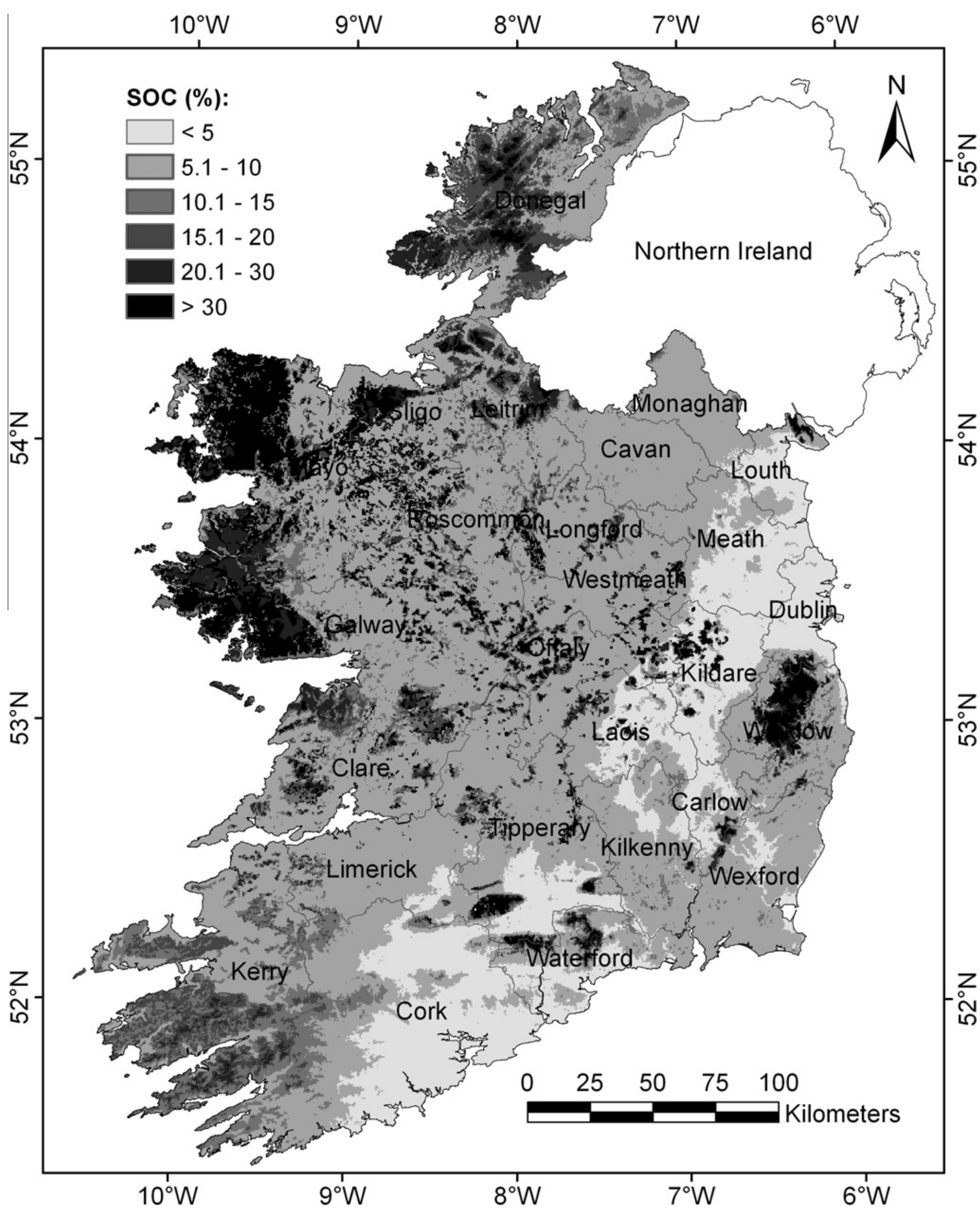

Fig. 6. Spatial distribution map of SOC in Ireland created using GWR.

Please cite this article in press as: Zhang, C., et al. Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Appl. Geochem. (2011), doi:10.1016/j.apgeochem.2011.04.014

The SOC maps showed elevated values in western Ireland where organic soils (or mainly blanket peat) are widespread, as well as the areas with high rainfall. Southwest Ireland and the Wicklow mountains in the East also exhibited high SOC. These areas are of high elevation and high rainfall, with upland blanket peats. In the midland of Ireland, there were scattered patches of high SOC areas, which were in line with the distribution of basin peat. The spatial distribution map of SOC in Ireland created using the GWR showed clear SOC relationships with rainfall and distribution of peat. Compared with the map created using a trans-Gaussian kriging (Zhang et al., 2008b), the smoothing effect was obviously reduced in the map created using GWR. This feature should also be regarded as an improvement.

3.5. Discussion of GWR method

Since GWR takes environmental factors into consideration, there are possibilities for further improvement of spatial interpolation using this method: when the influences of environmental factors are better understood; when the environmental factor maps are improved; or when the classification for sampling locations and locations to be predicted is improved, it is expected that better performance for GWR can be achieved. The currently used land cover map in the study area can be replaced by the land use map when it becomes available, as it is expected that land use may have a better relationship with SOC. The soil type map of Ireland in use was simplified based on a "soil associations" map (Gardiner and Radford, 1980), and it is expected that an updated version of the soil type map of Ireland will become available in a few years. Based on the box-plots (Fig. 3), misclassification of land cover and soil type for sampling locations did occur, e.g., some of the outliers in agricultural area had SOC values as high as 50%, and some of the wetland samples had SOC values less than 5%. Such misclassifications can hardly be avoided using maps. However, they can be corrected using field investigation, as well as "manual manipulation", e.g., soils with SOC >15% could be arbitrarily classified as "organic soils" (Zhang et al., 2008a). All of these possibilities provide ways for a further improved spatial interpolation using GWR which can be explored in the future.

Another factor that could affect the performance of GWR is the sample size of the dataset (SOC here) itself. When geochemical data are collected at a higher density, GWR models could be established at smaller spatial scales, and the performance can be further improved.

While this study has demonstrated that GWR provided at least comparable and reasonable results in comparison with the other chosen methods, there are some issues related to GWR that need attention. Like any regression analyses, there must be good relationships between the dependent variable (SOC here) and the independent variables (rainfall, land cover and soil type here). Such relationships can be complicated and spatially varying. GWR is good at dealing with the "spatially varying" aspect of the relationships, but the true relationships may still not be captured by the model, e.g., missing other important factors. Some studies have included the factor of geology (rock type) (e.g., Mishra et al., 2010) in GWR modelling. In Ireland, the spatial distribution maps between rock type and SOC did not exhibit a clear correlation. Therefore, rock type was not considered in the GWR in this study. The selection of appropriate environmental factors for GWR needs careful consideration and justification.

Another problem related to GWR is local multicollinearity. This problem may occur when there is a large area with the same value of an independent category variable (e.g., same land use of grassland), and lack of other values of the same independent variable (e.g., absence of one category of land use). One way to partly solve this problem is to use a more general and small number of categories, which was why the soil type and land cover type were generalized in this study. Environmental factors tend to have similar regional values with the absence of some categories, and thus this problem can be serious. In this case, the neighbourhood used in GWR has to be enlarged to include the other values of the same independent variable. When the neighbourhood is enlarged to the whole study area, GWR becomes MLR, thus loosing its power of modelling spatially varying relationships.

It was noted in this study that GWR may produce a small number of predicted values outside the range of the observed data, the same as an ordinary regression analysis. The issue became more serious for the power transformed data, as the back-transformation may end up in extreme values. A possible solution may be to replace the extreme values using the observed maximum and minimum values.

There are a lot of technical details and factors that need to be considered when using GWR. This paper neither attempts to find the "best" way for using various methods nor demonstrates that GWR is superior to other techniques. Further studies on the influences of various technical parameters on the performance of GWR could be considered. Nevertheless, GWR takes environmental factors into consideration and models the spatially varying relationships, thus providing a promising way for spatial geochemical modelling for SOC and potentially other soil properties. When the map quality of environmental factors is high, or when the environmental factors are properly assigned to sampling and estimation locations, it is expected that good results from GWR can be obtained.

4. Conclusions

Rainfall, land cover and soil type all influence SOC in Irish soils. GWR is able to include environmental factors into its spatial geochemical modelling, and the performance of GWR for spatial modelling of SOC in Ireland was good and reasonable in comparison with the chosen methods of OK, IDW and MLR. The GWR interpolated spatial distribution map of SOC of Ireland showed clear influences of the environmental factors, and it reduced the smoothing effect problem of other methods. With better understanding of environmental factors, availability of better environmental factor maps, and better classification of sampling and estimation locations, further improvement in the spatial geochemical modelling using GWR is possible.

Acknowledgments

This study was funded by Irish Environmental Protection Agency (EPA) STRIVE Programme 2007-2013 under the National Development Plan (Project No.: 2007-S-LS-1-S1), and the Chinese "111" Project (No. B08037) awarded to Sichuan University, China. The authors thank the editor Dr. Clemens Reimann and the reviewers for their helpful comments and suggestions.

References

Box, G.E.P., Cox, D.R., 1962. An analysis of transformations. J. Roy. Statist. Soc. Ser. B 26, 211–252.
Cressie, N., 1993. Statistics for Spatial Data, revised ed. John Wiley, New York.
D'Acqui, L.P., Santi, C.A., Maselli, F., 2007. Use of ecosystem information to improve soil organic carbon mapping of a mediterranean Island. J. Environ. Qual. 36, 262–271.
Duncan, D.B., 1955. Multiple range and multiple F tests. Biometrics 11, 1–42.
EPA (Environmental Protection Agency), 2009. CORINE Land Cover – IRELAND Land Cover Update for 2006 - FINAL REPORT. Environmental Protection Agency, Johnstown Castle Estate, Wexford, Ireland.
Fay, D., Zhang, C., 2007. Towards a National Soil Database. Associated Datasets and Digital Information Objects Connected to this Resource are Available at: Secure Archive for Environmental Research Data (SAFER) managed by Environmental

Protection Agency Ireland <http://erc.epa.ie/safer/iso19115/displayISO19115.jsp?isoID=7> (last accessed 05.02.11).

Fay, D., McGrath, D., Zhang, C., Carrigg, C., O'Flaherty, V., Carton, O.T., Grennan, E., 2007. EPA Report: Toward a National Soil Database (2001-CD/S2-M2). <http://www.epa.ie/downloads/pubs/research/land/> (last accessed 05.02.11).

Fotheringham, A.S., Brunsdon, C., Charlton, M.E., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester.

Gardiner, M.J., Radford, T., 1980. Soil Associations of Ireland and Their Landuse Potential – Explanatory Bulletin to Soil Map of Ireland 1980. Agricultural Institute, Dublin.

Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008, Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m. <http://srtm.csi.cgiar.org> (last accessed 05.02.11).

Jordan, C., Zhang, C.S., Higgins, A., 2007. Using GIS and statistics to study influences of geology on probability features of surface soil geochemistry in Northern Ireland. J. Geochem. Explor. 93, 135–152.

Lal, R., 2003. Global potential of soil carbon sequestration to mitigate the greenhouse effect. Crit. Rev. Plant Sci. 22, 151–184.

McBratney, A.B., Webster, R., 1983. Optimal interpolation and isorhythmic mapping of soil properties. 5. Co-regionalization and multiple sampling strategy. J. Soil Sci 34, 137–162.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67–94.

Mishra, U., Lal, R., Liu, D.S., Van Meirvenne, M., 2010. Predicting the spatial variation of the soil organic carbon pool at a regional scale. Soil Sci. Soc. Am. J. 74, 906–914.

Mueller, T.G., Pierce, F.J., 2003. Soil carbon maps: enhancing spatial estimates with simple terrain attributes at multiple scales. Soil Sci. Soc. Am. J. 67, 258–627.

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

Osborne, P.E., Foody, G.M., Suarez-Seoane, S., 2007. Non-stationarity and local approaches to modelling the distributions of wildlife. Divers. Distribut. 13, 313–323.

Rawlins, B.G., Marchant, B.P., Smyth, D., Scheib, C., Lark, R.M., Jordan, C., 2009. Airborne radiometric survey data and a DTM as covariates for regional scale mapping of soil organic carbon across Northern Ireland. Eur. J. Soil Sci. 60, 44–54.

Schulp, C.J.E., Verburg, P.H., 2009. Effect of land use history and site factors on spatial variation of soil organic carbon across a physiographic region. Agric. Ecosyst. Environ. 133, 86–97.

Scull, P., 2010. A top-down approach to the state factor paradigm for use in macroscale soil analysis. Ann. Assoc. Am. Geograph. 100 (1), 1–12. doi:10.1080/00045600903362279.

Sweeney, J., Brereton, T., Byrne, C., Charlton, R., Emblow, C., Fealy, R., Holden, N., Jones, M., Donnelly, A., Moore, S., Purser, P., Byrne, K., Farrell, E., Mayes, E., Minchin, D., Wilson, J., Wilson, J., 2003. Climate Change Scenarios & Impacts for Ireland (2000-LS-5.2.1-M1) Environmental RTDI Programme 2000–2006 Final Report. Environmental Protection Agency, Wexford, Ireland. <http://www.epa.ie/downloads/pubs/research/climate/> (last accessed 05.02.11).

Tisdale, S.L., Nelson, W.L., Beaton, J.D., 1995. Soil Fertility and Fertilizers. MacMillian, New York.

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists, second ed. John Wiley & Sons Ltd., Chichester.

Zhang, C.S., 2006. Using multivariate analyses and GIS to identify pollutants and their spatial patterns in urban soils in Galway, Ireland. Environ. Pollut. 142, 501–511.

Zhang, C.S., Fay, D., McGrath, D., Grennan, E., Carton, O.T., 2008a. Statistical analyses of geochemical variables in soils of Ireland. Geoderma 146, 378–390.

Zhang, C.S., Fay, D., McGrath, D., Grennan, E., Carton, O.T., 2008b. Use of trans-Gaussian kriging for national soil geochemical mapping in Ireland. Geochem. Explor. Environ. Anal. 8, 255–265.

Please cite this article in press as: Zhang, C., et al. Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Appl. Geochem. (2011), doi:10.1016/j.apgeochem.2011.04.014