Geo-statistic based sample size optimization for regional soil organic carbon mapping in North-East China

Chen Changhua, Guo Dongjing, Chen Xi Yun*

(School of Geography, Beijing Normal University, No. 19, XinJieKouWai St., HaiDian District, Beijing 100875, P. R. China)
*Author of coresponding, Email:chen.xiyun@bnu.edu.cn

1. Introduction

Regionl soil organic carbon investigation is crucial for both soil carbon sequestration and soil nutrition assessment(Wang. et al.， 1995). Traditionally, this work can be done by designed field soil sampling and classification based estimation. This is time-consuming, labour-intensive and costly. How to effectively mapping soil organic carbon with least sample point to gain reasonable accuracy, is a challenge. Many researches have revealed that spatial distribution of soil organic carbon is related to the soil formation processes, depending on landscape position, climate, as well as the biological conditions and land use history. Recent explorations on digital soil mapping using geostatistics and geographic information system technology have showed that soil organic carbon content within certain range has strong autocorrelation and good semivariance structure. Meanwhile, numberous methods of soil variable interpolation from discrete soil sampling points to continuous spatial distributing maps have been developed(Yang. et al.， 2011; Yang. et al.， 2010; Zhao. et al.， 2012). Among those methods, combined geostatistical and GIS is the widely used one. It provides a promising way of rational sampling based on spatial analysis(Lu. et al.， 2011).

Here we present primry results of sample size optimization for regional soil organic carbon mapping. In this study, datasets of black soils in North-east China from the second national soil survey is retrieved from the soil species of China, and geostatistics is combined with GIS to explore the relationship between the number of sampling points and mapping accuracy of soil organic carbon content(SOCC).

2. Material and methods

2.1 Research area and datasets

In order to include all the areas with black soil, here we define the Northeast China as eastern Heilongjiang, Jilin, Liaoning and Chifeng City, Tongliao City, Xinganmeng and Hulun Buir City of Inner Mongolia, located at 115°52’ to 135°09'E, 38°72' to 53°55’N(Fig 1). Total area is 1,242,600 km^2.

Totally, 417 soil profiles with detailed records, which includes geographic location, altitude, mean annual temperature, mean annual precipitation, soil depth, soil bulk density, soil organic matter content and land use are retrived from the second national

soil survey database. After calculation of soil organic carbon from organic matter content with a factor of 0.58, ten subsets of data with sample size (as show in table 1) are extracted to analysis SOCC spatial variability and digital mapping accuracy.

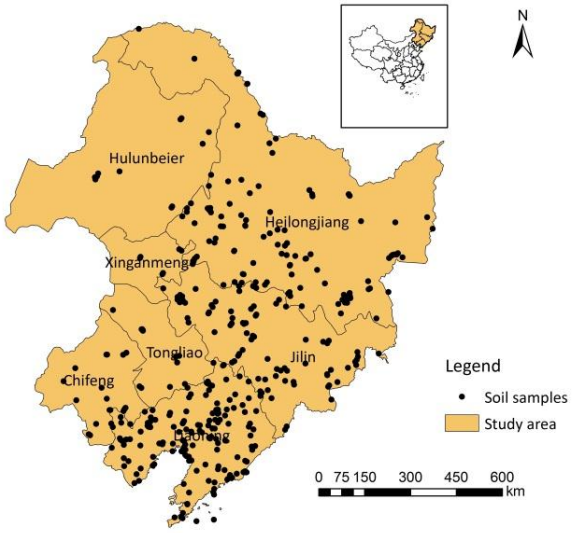

Figure 1. Location of the research area and the sample points distribution.

2.3 Mappig methods and acuracy test

Firstly, all the profile data are grouped according to soil species, horizon and land use type respectively to calculate a representative profile, and then the depth, SOCC and several other attributes for each representative profile is estimated. Land use is grouped into three major types, including agriculture, grassland and forest.

Then, statistical distribution of SOCC is detected for original datasets. Afterwards, data is logarithm transformed and sorted according to longitude and latitude to meet the need of spatial analysis.

All the spatial analyses and graphing in this study are conducted by the GS^+ and ARCGIS9.3 software, including optimised semivariogram model selection, kriging interpolations and digital mapping. After interpolation, SOCC are re-sampled from the digital maps with different sample points to test the mapping accuracy by comparing model interpolations with actual samples results. The errors and fitting effciency are indicated by Root Mean Square Error (RMSE), average standard error (ASE) and fitting coefficient (R^2), respectively.

Finally, SPSS 11.5 software is used to detect the relationship between SOCC and auxiliary variables, including geographical location, mean annual air temperature, land uses, to test the possibility of using the auxiliary information to improve the accuracy of SOCC prediction. For all analyzes, the high significance is defined at the 0.05 level (2-tailed).

3. Results and discussion

Statistical exploration of whole dataset shows that SOCC follows the skewed distribution. All the 10 sub-datasets are logarithm transformed accordingly for semivariance fitting with GS^+ software. The best fitting semivariance models and parameters for each subset are as show in table 1. Best fitting model for data with different sample size is different, indicating the self-correlations of sample point and thus the spatial variability of SOCC presented in each subset are different. In addition, the ratios of nugget to sill(**C0/ Sill**) are in the range of 18.32%~38.04%, with average value below 25%, suggesting that the autocorrelations of SOCC in the whole research area are significant. Fitting coefficients for different subsets are in the range of 0.308 to 0.775. It is greater than 0.705 when the sample size greater than 153 and change slowly from sample size of 153 to 417. This suggests that the model fitting reveals about 70% of the semivariant structure of SOCC in the research region when the sample size larger than 153.

Table 1.Semivariance models and fitting parameters of SOCC under different sample size

Data sets	Numbers of sample	Model	Nugget (C_0)	Sill (C_1)	[C0/ Sill] %	Major range/m	R^2	RSS
N	417	Gaussian	0.38	1.04	37	11837	0.775	5.270
A	375	Exponential	0.28	1.02	28	26851	0.755	1.150
A1	300	Gaussian	0.36	0.94	38	11837	0.736	2.380
A2	240	Spherical	0.34	1.17	29	11837	0.771	3.320
A3	192	Gaussian	0.33	1.22	27	11837	0.769	4.060
A4	153	Spherical	0.28	1.09	25	16711	0.705	3.250
A5	122	Spherical	0.20	1.10	18	17871	0.669	5.440
A6	97	Spherical	0.21	0.59	36	7941	0.493	1.950
A7	77	Spherical	0.21	0.79	27	12317	0.461	2.590
A8	61	Exponential	0.20	0.62	32	13811	0.308	0.873
A9	48	Spherical	0.19	0.53	37	9192	0.416	0.101

Note: model denote the best fitting model used under different sample number. R^2 and RSS refer to the fitting coefficient and residual SS, respectively.

Spatial interpolations of SOCC by the simple kriging method based on different sample size are as show in Figure 2. We generate spatial distribution maps for all the 10 subsets. Here we only select four to illustrate the differences caused by sample size. In order to interpret mapping results, the space distribution of SOCC on the maps have divided into high-value area (> 25g/kg), transition zone (11.87 ~ 25 g / kg), median (6.01 ~ 11.87 g / kg) and low-value area (<6.01 g / kg). It is clear that all the interpolation show the spatial trends of SOCC in the research area, with higher values at north-east part and low values at southwest part. However, the high-value area is not included in the map with a sample size of 48, indicating that too few sample points may not represent the full range of the SOCC in the region. When the sample size greater than 153 the value areas keep stable. Differences between interpolation results are in the size and shape of the areas. Differences between value areas for sample size of 192 and 417 are relatively small and mainly in the size and places for high and low value areas. When sample size greater 300, SOCC interpolation results nearly show the same pattern as sample size of 417. This

suggests that the sample size of 192 to 300 may represent most of the spatial variability of SOCC in the region.

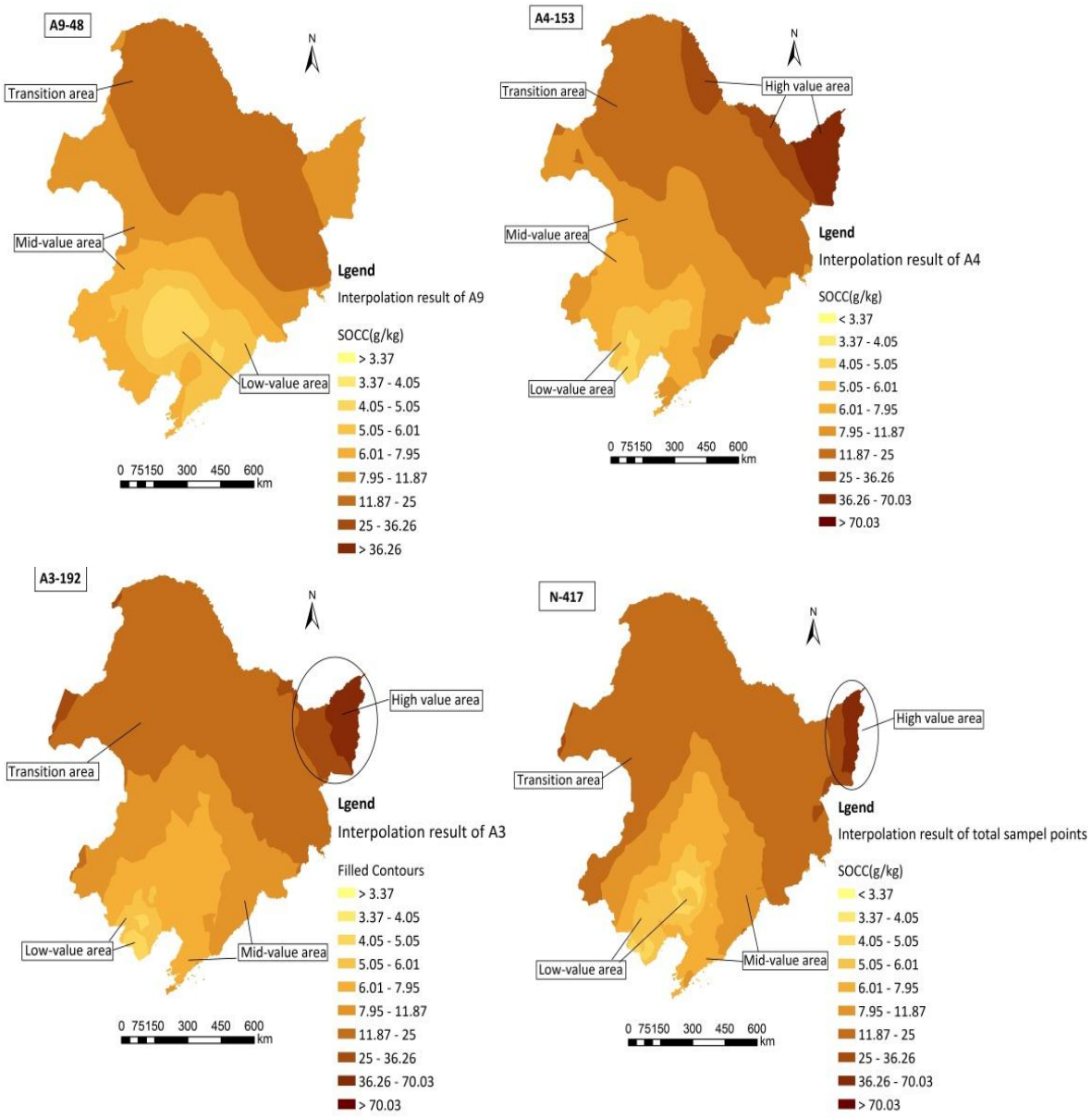

Figure 2. Selected digital maps of soil organic carbon in North-east China based on different sample points. Among these, A9,A4,A3 and N, refer to the result for sample points of 48,153,192 and 417, respectively.

The change of mapping accuracy with sample size is show in figure 3. Based on the standard of geostatistics, lower values of RMSE and ASE as well as closer of them indicate better prediction. RMSE and ASE are closest at a sample size about 240. The fitting coefficients increase rapidly as sample size below 122 and change weakly when

sample size greater than 153. In the meanwhile, fitting coefficient vary in the range of 0.71 to 0.78 and RMSS is around 1 when sample size greater than 153. When we include both the mapping accuracy and sampling costs, a sample size of about 240 may be suggested for SOCC mapping in the research area.

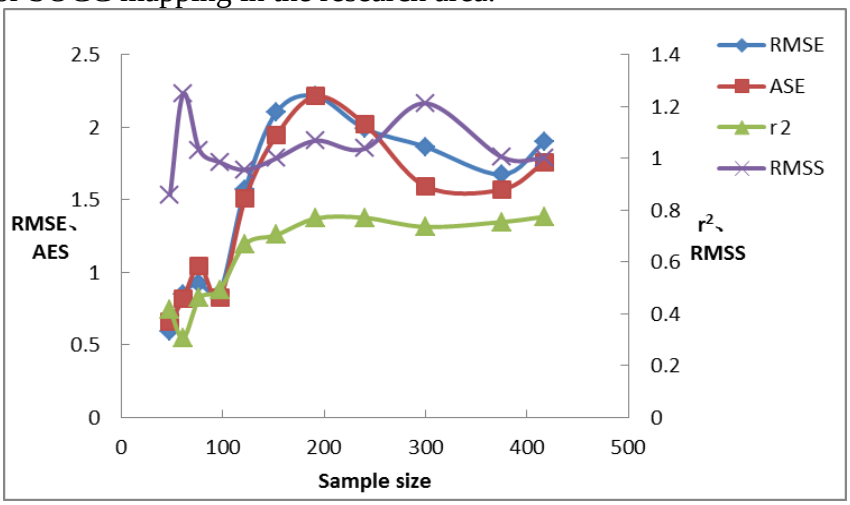

Figure 3. Change of interpolation parameters with sample size. RMSE, ASE, RMSS and R^2 are the Root Mean Square Error, Average Standard Error, Root Mean Square Standardized and fitting coefficients for different models under different sample size, respectively.

Stepwise regression show that SOCC is related to auxiliary variables, including the longitude, latitude, CEC, pH, mean annual temperature, soil bulk density and land use types of the sampling points (results not show). Among these variables, mean annual temperature, CEC and pH are not readily obtainable for sample points by field work, but others could be measured at the same field campaign. It provides the possibility of using this auxiliary information to improve the accuracy of SOCC prediction.

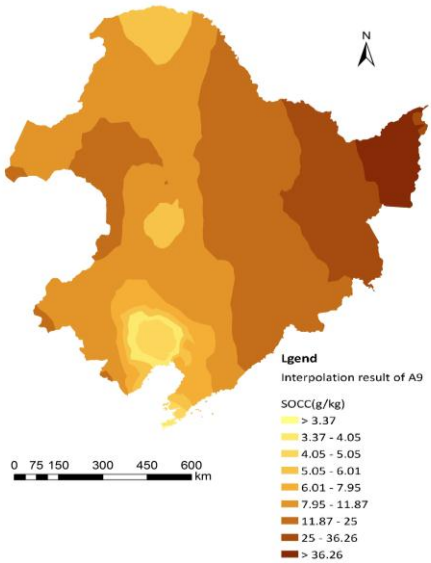

Figure 4. SOCC distribution based on the regression model prediction.

After including the information of longitude, latitude, soil bulk density and land use types, a new spatial distribution pattern of SOCC is generated (Figure 4). Together with the interpolated digital maps of SOCC (Figure 2, A3 and N), it is clear that Liaoyang, Fuxin, Panjin, Jinzhou are located in the low value areas of SOCC from both interpolation and regression model predication results. These areas also have lower uncertainty during model fitting and mapping. Thus, the easily obtainable auxiliary factors can be employed to reduce the points of soil sampling while keep mapping accuracy. Eastern part of Sanjiang Plain, including Jiamusi, Qitaihe, Jixi, Hegang and ShuangYashan are in the area with abundant soil organic matter. This area is also in the high value area both from interpolation and regression model prediction. Because the Songnen Plain region is China's major grain producing area and SOCC is affected by human activities, for instance, land use change, and further study for reasonable sample size and frequent field survey is strongly needed to guarantee an updated SOCC mapping. Differences between results of interpolation and regression model predictions are relatively large for Northeastern part of Inner Mongolia and northwestern Heilongjiang province. Sample points during soil survey are relatively sparse in this area (Figure 1). Grassland and woodland are the main land use types in this area. SOCC might be influenced strongly by environmental factors, such as climate, vegetation and topography. So, study on rational sample size, use of auxiliary information and model prediction should be strengthed for further improvement of SOCC mapping accuracy effectively.

4. Conclusion

Regional soil organic carbon content mapping is a novel way for large scale estimation. However the accuracy of mapping strongly depends on sample size. How to improve mapping accuracy is still a challenge. Combined geostatistical and GIS technique is a promising way of optimizing the sampling process.

Our results from black soils in North-east China show that:

1) A sample size of around 240 may give a relatively reasonable spatial estimation of soil organic carbon distribution for area of 1,242,600 km^2 in this region.
2) Use of readily obtainable auxiliary variables, including longitude, latitude, soil bulk density and land use types, may help to improve the sample size optimization.
3) Using of the national soil survey datasets show that low soil carbon content areas locate in Liaoyang, Fuxin, Panjin, Jinzhou, where have relative high density sampling points and easily obtainable auxiliary factors can be employed to reduce the points of soil sampling while keep mapping accuracy. High value areas are in Eastern part of Sanjiang Plain accompanied by much frequent land use change. Further study for reasonable sample size and frequent field survey is strongly needed to guarantee an updated SOCC mapping. Sample points during soil survey are relatively sparse in Northeastern part of Inner Mongolia and northwestern Heilongjiang province. Study on rational sample size, use of auxiliary information and model prediction should be strengthened to improve SOCC mapping effectively.

Acknowledgements

This work is supported by the National Science Foundation of China (41171360, 4093074 0). We thank undergraduate students Niu Jianli, Xu Ying, Zhang Tianyu, He Ziyun, Mi Chunlei and Li Fuhua from School of Geography, Beijing Normal University, for their contribution on database construction.

References

Lin Yang, A-xing Zhu., Chengzhi Qin et al., 2011. A soil sampling method based on representativeness grade of sampling points. *Acta Pedologica Sinica*, (5): 938-946.

Lin Yang, A-xing Zhu, Chengzhi Qin et al., 2010. A purposive sampling design method based on typical points and its application in soil mapping. *Progress in Geography*, (3): 279-286.

Qianqian Zhao, Gengxing Zhao, Huailong Jiang et al., 2012. Study on spatial variability of soil nutrients and reasonable sampling number at county scale. *Journal of Natural Resources*, (8): 1382-1391.

Xuefeng Wang., Heng Zhang, 1995. The spatial variability of soil organic matter. *Soils*, (2): 85-89.

Yanjun Lu., Chengzhi Qin, Weili Qiu et al., 2011. Sensitivity analysis of soil property parameter in typical-sample-based prediction model of digital soil mapping. *Scientia Geographica Sinica*, (12): 1549-1554.