Grid Soil Sample Interpolation Using Geographicaly Weighted Regression And Random Forest

E. G. Matcham[1], S. Subburayalu[2], [3]J. Fulton, E. Hawkins[3], P. Paul[4], and L. E. Lindsey[1]
[1]Dept. of Horticulture and Crop Science, The Ohio State University, Columbus, OH; [2]Dept. of Agricultural Sciences, Central State University, Wilberforce, OH; [3]Dept. of Food, Agricultural and Biological Engineering, The Ohio State University, Columbus, OH; [4]Dept. of Plant Pathology, The Ohio State University, Wooster, OH
matcham.3@osu.edu

ABSTRACT

Soil sampling is useful in agriculture for setting fertilizer application rates. High density soil samples can also be used for variable rate seeding and other precision agriculture applications. Half-acre grid soil samples were collected from 6 soybean fields, and phosphorous (P), potassium (K), and organic matter (OM) were measured. Each soil parameter was interpolated for each field, with terrain attributes as covariates, using two different methods: geographically weighted regression (GWR) and random forest (RF). Global error for each interpolation method was measured using root mean squared error (RMSE). In 14 out of 18 instances, GWR had lower RMSE, indicating that it had lower error in the interpolation. Random forest did occasionally out-perform GWR, but GWR had significantly lower error in the study overall ($U=143$, $p=0.010$) and is an appropriate method for interpolating grid soil samples at the field scale. There was not a pattern with regard to site or soil property for when RF out-performed GWR. The percentage of the field that was below the critical level was the same or higher for RF in 5 out of 6 fields, which combined with the high global error rate indicates that RF interpolation could over-apply fertilizer.

INTRODUCTION

Soil sampling is important for determining appropriate fertilizer rates for agriculture. Sampling methods usually fall into one of two categories: zone or grid. Interpolation is not necessary for zone sampling because the boundaries that each soil sample represents are defined before samples are collected. In grid sampling, the soil test value could be used as the fertility estimate for the entire grid cell, but many farmers and retailers interpolate the soil test results because application equipment can utilize soil fertility information at a higher resolution than the grid cell size of most sampling regimes. Interpolation turns the point data from soil test results into a surface of discrete values across the entire field at a higher resolution than was actually sampled. This study focused on P, K, and OM because they are soil properties commonly mapped for variable rate prescriptions.

Interpolation accuracy can greatly impact how soil test results are interpreted because it can change whether the field is modeled as having adequate or inadequate fertility levels overall, and the spatial distribution of the soil fertility. If the interpolation underestimates the fertility levels of the field, more fertilizer will be applied. There is financial risk to under or over applying fertilizer, and significant environmental risk to over application. In the case of modeling OM, seeding rate prescriptions based off the interpolated map could be inaccurate.

One way to improve the accuracy of soil maps based on grid samples is incorporating elevation and other terrain derivatives as a covariate, and selecting an appropriate modeling method. Auxiliary information, such as elevation and slope, was useful for interpolating soil OM at a regional scale using geographically weighted regression and regression kriging. Geographically weighted regression (GWR) had lower root mean square error (RMSE) than regression kriging, indicating that the GWR interpolation had a higher accuracy (Wang et al., 2012). A similar regional-scale study was performed using elevation as a covariate to predict soil organic carbon at a regional scale. Among the methods GWR, multiple linear regression, and regression kriging, GWR had the lowest RMSE, indicating that it was the most accurate model (Mishra et al., 2010).

At the regional scale, GWR has been successfully used to interpolate soil test values with terrain covariates, but it has not been fully investigated at the field scale. Geographically weighted regression is sensitive to multicollinearities between covariates at both a global and local scale (Wheeler and Tiefelsdorf, 2005). The scale-dependency of local multicollinearities may impact GWR performance at the field scale.

Random forest (RF) is a machine learning algorithm that can be used to interpolate soil samples, and is a viable alternative to GWR due to its robustness against multicollinearities. Random forest creates an ensemble of decision trees using a random subset of observations and covariates, preventing overfitting issues that are common in other decision tree algorithms (Grimm et al., 2008; Guio Blanco et al., 2018). Another benefit of RF is that it ranks the importance of covariates. In a regional-scale soil organic matter mapping project, terrain attributes were the most important predictors of soil OM (Grimm et al., 2008).

A potential downside of RF is its sensitivity to the distribution of the training data set. It should not be used to extrapolate outside of the range of covariates observed in the training data (Guio Blanco et al., 2018). In grid sampling where points have not been strategically placed to capture the complete range of covariates in the study area, this could limit RF performance.

Comparing performance of GWR and RF at the field scale would improve interpolation of soil test results for fertilizer application and other variable rate technology. Therefore, the objectives of this study were to evaluate (1) whether GWR or RF were more accurate, as measured by global RMSE, at a field scale, and (2) whether interpolation method impacts P application rates.

METHODS

Half-acre grid soil samples were collected across 2 fields in 2017 and 4 fields in 2018. After locating each sample point using Google Maps on an LG G6 Android phone, three 20 cm deep, 1.9 cm wide cores were collected within 10 feet of the identified point. Soil samples were air-dried in at ambient temperature and humidity. Samples were shipped to A&L Great Lakes Laboratories for soil test P, available K,, and OM analysis. Phosphorus and potassium was measured using Mehlich-3 extraction, OM was measured based on loss on ignition. To compare P results back to the critical level as published in the Tri-State Fertility guide, Bray-P extractant values were estimated using the conversion provided by OSU Extension (Ohio State University Extension, 2018; Vitosh et al., 1995).

Digital elevation models with 0.76 m resolution were downloaded from the Ohio Geographically Referenced Information Program database as TIFF files. SAGA GIS (2.3.2) was used to generate the following terrain derivatives: slope, aspect, topographic wetness index, and relative slope position (Conrad et al., 2015).

Geographically weighted regressions were calculated at the same resolution as the original elevation models using the Geographically Weighted Regression for Multiple Predictor Grids functionality in SAGA GIS. SAGA kernel defaults were used (Gaussian kernel shape and bandwidth determined by grid size). Actual and predicted values from each GWR were exported from SAGA as a shapefile, and the attribute table from this shapefile was copied into a CSV file for RMSE calculation in R.

Terrain factors were measured at soil sampling locations using the "Add raster values to points" tool with the SAGA plugin for QGIS (2.18.19) (QGIS Core Development Team 2018). Soil test values and terrain factors were exported from QGIS as a CSV file for interpolation via random forest in R (3.4.2) (R Core Team 2017). Random forest models were built using the package *randomForest* with parameters *ntree*=500 and *mtry*=2 (Liaw and Wiener, 2002). Out-of-bag predictions were used for RMSE calculation.

Root mean square error was calculated using the package *Metrics* for both random forest and GWR interpolation (Hamner et al., 2018). To determine if RMSE was significantly different between the two interpolation methods, a pairwise Wilcox rank test was performed in R (alpha=0.05). Beyond RMSE, the predicted values were also analyzed to see what percentage were below the phosphorous critical level for Ohio, 15 ppm Bray (Vitosh et al., 1995).

RESULTS

Overall, RMSE was lower for the GWR model in 14 out of 18 instances, indicating that global error for GWR was significantly lower than for RF (U=143, p=0.010). See Figure 1 for a visual representation of RMSE across each interpolation.

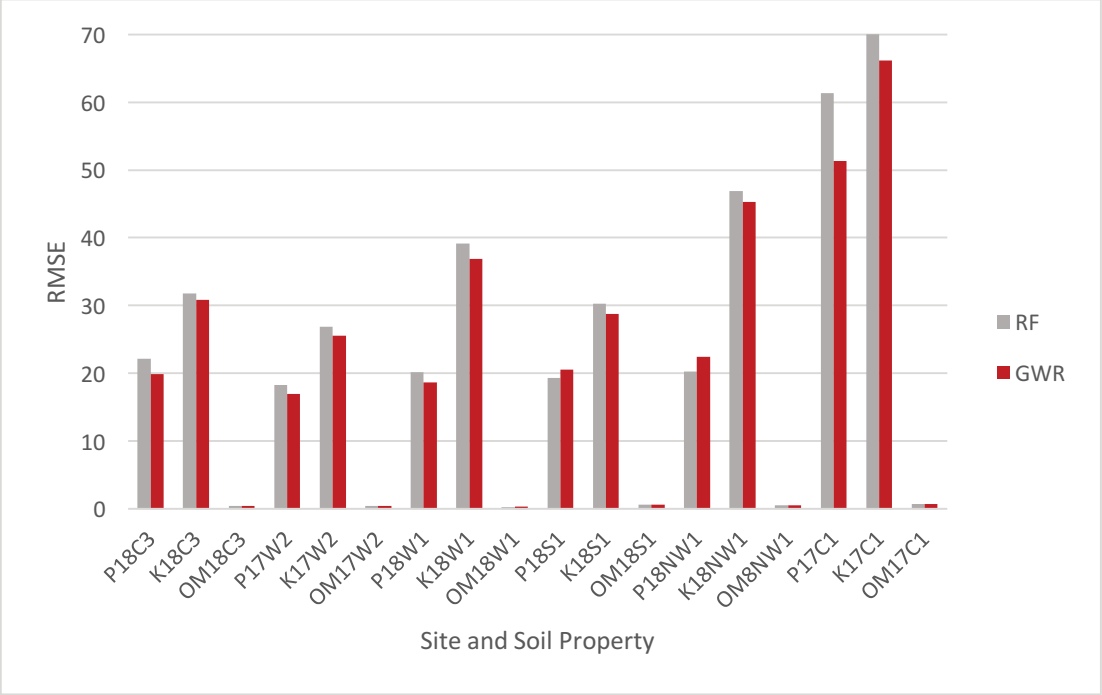

Figure 1: Comparison of RMSE for P, K, and OM interpolation across 6 sites using geographically weighted regression and random forest

At 4 out of 6 sites, RF outperformed GWR for one soil property. In all other cases, GWR had lower RMSE. The RMSE was lower using GWR in all K interpolations, in 4 of 6 P interpolations, and 4 of 6 OM interpolations. RF outperformed GWR occasionally without regard to site or soil property, but GWR was overall the method with lower global error. GWR can be used to interpolate soil samples at the field scale.

In 5 fields out of 6, the GWR interpolation would require the same amount of P fertilizer or less than the RF interpolation. A summary of the percentage of the field below the critical level as indicated by the interpolation is available in Table 1.

Table 1: Summary of what percentage of each field is below the P critical level when interpolated using random forest or geographically weighted regression

Site	RF(%)	GWR(%)
18C3	63	47
17W2	62	52
18W1	17	3
18S1	39	8
18NW1	0	0
17C1	44	45

SUMMARY

For high-density grid soil sampling, different interpolation methods produce different results. Geographically weighted regression had significantly lower global error rates overall and is a valuable method for field scale interpolation.

When possible, it is preferable to calculate both RF and GWR interpolations and compare before selecting a method, since in some cases RF may give a lower global error. In instances where the spatial distribution of error is meaningful, such as when waterways are near fields, direct comparison using residual mapping would be a useful additional tool to determine an appropriate method. Overall, GWR was significantly better than RF for field scale interpolation and can be used to estimate soil fertility levels across fields with grid soil test results.

Using error-prone interpolations can cause over- or under-application of P fertilizer. The GWR interpolation indicates that a lower P level would be needed than the RF interpolation. That combined with the high global error for the RF interpolation indicates that RF is likely over-estimating fertilizer need, which could increase P runoff from fields. Higher fertilizer rates also cost farmers more money, without providing a yield benefit.

WORK CITED

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J. 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev. 8: 1991-2007. doi:10.5194/gmd-8-1991-2015.

Grimm, R., T. Behrens, M. Märker, and H. Elsenbeer. 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis. Geoderma 146(1–2): 102–113. doi: 10.1016/j.geoderma.2008.05.008.

Guio Blanco, C.M., V.M. Brito Gomez, P. Crespo, and M. Ließ. 2018. Spatial prediction of soil water retention in a Páramo landscape: Methodological insight into machine learning using random forest. Geoderma 316(July 2017): 100–114. doi: 10.1016/j.geoderma.2017.12.002.

Mishra, U., R. Lal, D. Liu, and M. Van Meirvenne. 2010. Predicting the Spatial Variation of the Soil Organic Carbon Pool at a Regional Scale. Soil Sci. Soc. Am. J. 74(3): 906–914. doi: 10.2136/sssaj2009.0158.

Ohio State University Extension. 2018. Understanding Soil Test Reports for Phosphorous Values. URL https://agbmps.osu.edu/faq/understanding-soil-test-reports-phosphorus-values

Vitosh, M.L., J.W. Johnson, and D.B. Mengel. 1995. Tri-state Fertilizer Recommendations for Corn, Soybeans, Wheat and Alfalfa. Ext. Bull. E-2567 (New), July 1995 2567(July): 1–4. https://www.extension.purdue.edu/extmedia/AY/AY-9-32.pdf.

Wang, K., C. Zhang, and W. Li. 2012. Comparison of Geographically Weighted Regression and Regression Kriging for Estimating the Spatial Distribution of Soil Organic Matter. GIScience Remote Sens. 49(6): 915–932. doi: 10.2747/1548-1603.49.6.915.

Wheeler, D., and M. Tiefelsdorf. 2005. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. J. Geogr. Syst. 7(2): 161–187. doi: 10.1007/s10109-005-0155-6.