IECAG 2021

MDPI

Proceedings

Spatial prediction of the properties of chernozem soils on a field scale using machine learning methods based on data from Landsat 8 OLI and Sentinel 2 images for precision farming †

Ilnas Sahabiev*, Elena Smirnova and Kamil Giniyatullin

Institute of Environmental Sciences, Kazan Federal University, Kremlevskaya str. 18, 420008, Russia, ilnassoil@yandex.ru

* Correspondence: ilnassoil@yandex.ru

† Presented at the 1st International Electronic Conference on Agronomy, 3–17 May 2021; Available online: https://sciforum.net/conference/IECAG2021

Abstract: We studied the possibility of using spectral parameters of open soil (Landsat 8 and Sentinel 2 data) and machine learning methods for using, on a single field scale, refined maps of organic carbon content, available forms of nitrogen, phosphorus, and potassium, silt and clay fractions. The accuracy of the obtained predictive maps of changes in soil properties was assessed in the aspect of their use for information support for the introduction of precision farming systems. It has been shown that the use of spectral reflectance data to refine digital maps provides a significant improvement in spatial prediction when using machine learning methods compared to traditional linear models. The content of SOC, available nitrogen, and available potassium is well predicted using the random forest (RF) and support vector regression (SVMr) models; the content of available phosphorus, silt and clay is somewhat worse. Refined digital maps based on Sentinel 2 data are characterized by a greater degree of detail in the spatial variability of soil parameters; at the same time, the use of Landsat 8 data can also be productive, since in some cases it provides higher accuracy of the spatial prediction.

Keywords: Precision Agriculture; digital maps; machine learning methods; remote sensing

Citation: Sahabiev, I.; Smirnova, E.; Giniyatullin, K. Spatial prediction of the properties of chernozem soils on a field scale using machine learning methods based on data from Landsat 8 OLI and Sentinel 2 images for precision farming. **2021**, *4*, x. https://doi.org/10.3390/xxxxx

Published: date

1. Introduction

The need to increase the production of crop products while ensuring the minimum negative impact on the environment is the most important problem of the modern development of society. One of the promising directions for solving this problem is considered the introduction of digital technologies for the variable application of mineral fertilizers in agricultural production [1]. These technologies are focused on the optimal satisfaction of the needs of cultivated plants in nutrients, taking into account the spatial heterogeneity of arable land in terms of agrochemical properties. Successful implementation of Precision Agriculture technologies requires an ultra-precise description of the spatial heterogeneity of soil cover properties, which must be scaled, in this case, at the level of a single crop rotation field. The solution to this problem requires, along with detailing the analytical study of soils, the use of modern methods of mathematical processing of the data, based on the prediction of the formation of the spatial distribution of soil indicators [1, 2].

Recently, in this aspect, machine learning methods have been widely used, which make it possible to efficiently process materials of remote sensing of the Earth (RS). The use of approaches based on these methods can significantly improve the accuracy of digital maps of soil properties. The use of these methods can make it possible in the future to provide, based on the prediction of changes in agrochemical indicators in space

and in time, the refinement of rough agrochemical maps created for traditional fertilization, as well as to update outdated cartographic material. The possibility of a productive solution to this problem on the scale of one field (or several fields) will significantly simplify and reduce the cost of obtaining digital maps of nutrients, which can be used in the development maps of variable rate application of fertilizers. Particularly attractive is the possibility of using remote sensing data available in open sources to refine agrochemical maps, which can also significantly reduce the cost and simplify the information support for the implementation of digital farming systems.

The purpose of this work is to assess the possibility of using remote sensing data obtained from the Landsat 8 and Sentinel 2 satellites as predictors of spatial prediction of soil properties using machine learning methods.

2. Materials and Methods

The object of study was a field (254 ha) located on the territory of the Republic of Tatarstan (Russia) in the zone of distribution of chernozem soils. The site is characterized by a significant elevation difference (up to 60 m) with steep slopes and high heterogeneity of the soil cover in terms of fertility. The field was divided into elementary squares of about 5 hectares each, from which point soil samples (20-40 pcs.) were taken to compile mixed samples. In total, 50 mixed soil samples were taken, in which the content of available nitrogen, available phosphorus, and available potassium, soil organic carbon (SOC), silt, and clay were determined. The nutrients available to plants were determined on the basis of national standards, the SOC content was determined by dry combustion, and silt and clay fractions were determined by laser sedimentography.

The remote sensing sources were data from publicly available satellites (Landsat 8 OLI and Sentinel 2). Satellite images were obtained from the sites of the US Geological Survey and the European Space Agency. The selection of space satellites as potential data sources for spatial prediction of agrochemical properties of soils was based on their availability, openness, differences in the resolution, and the presence of a wide range of bands of the electromagnetic spectrum in space images. For the work, we used images with open soil, i.e. with minimal influence of vegetation. For Landsat 8 OLI, such conditions corresponded to the image from 31/05/2019, for the Sentinel 2 satellite - the image from 12/05/2019. Space images were selected taking into account the minimum influence of atmospheric disturbances, however, all images were atmosphericallycorrected using the DOS 1 method. Then, spectral indices were calculated, which are represented by the ratios of individual bands and indices characterizing the open surface (NDVI, Grain size index, Clay index, MIR index, Bare soil index, Redness index, Saturation index, Coloration index, etc.). The data of individual bands and spectral indices were extracted and averaged over the elementary sampling sites. Working with raster images and modeling was carried out in the environment of the object-oriented language R [3].

Linear models (MLR), support vector regression models (SVMr) and random forest (RF) were used as models. RF and SVMr models have been tuned. The models were validated using the bootstrap procedure, taking into account performance optimism [4]. The performance values of the models were calculated for individual samples of the bootstrap and then the performance of the model fitted to the original data was calculated. The value of optimism of predict was calculated by subtracting the average performance values of the models of individual bootstrap samples and models based on the original data. The performance totals were considered values without optimism. RMSE, MAE, and R2 criteria were used to evaluate the models. The best models were those with a minimum value of RMSE, MAE, and a maximum value of R2. Subsampling of data for each model was carried out using Recursive Feature Elimination (RFE) from the caret package. For this, the importance scores were iteratively determined, ranked, and subsamples with the minimum importance scores were selected. For each type of model, the corresponding evaluation functions were used.

3. Results and discussion

According to the data in Table 1, of the three types of models, the MLR models that had the minimum values of the coefficient of determination were recognized as the worst. The MLR model is inferior in the accuracy of spatial prediction for all indicators of soil properties, both when using remote sensing data from the Landsat 8 OLI satellite and the Sentinel 2 satellite.

The RF and SVMr models best predict available nitrogen, potassium, and organic carbon. In the case of using the Landsat 8 OLI satellite data for available nitrogen for the RF model, RMSE = 9.21, and for the SVMr model, RMSE = 4.51, for the RF model, R2 = 0.83, for the SVMr model, R = 0.95. A similar situation is observed for available potassium, in which the SVMr model has lower RMSE and MAE values, as well as higher R2 values than the RF model. For SOC for RF and SVMr models, the R2 has similar values (R2RF = 0.83 and R2SVMr = 0.82). When using more detailed Sentinel 2 satellite imagery, the RF model for available nitrogen and SOC shows the best results. For the RF model, available nitrogen has R2 = 0.85, and for SOC - R2 = 0.75. For available potassium the RF and SVMr models have close R2 values (R2RF = 0.74 and R2SVMr = 0.75).

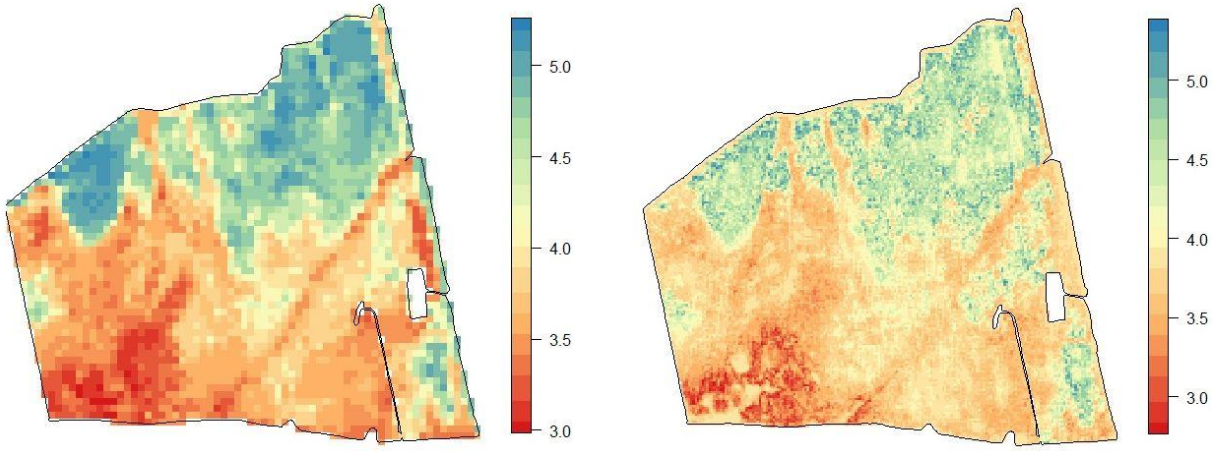
Table 1. Estimates of model performance.

Property	Model	RMSE	MAE	R^2	RMSE	MAE	R^2
		Landsat 8 OLI			Sentinel 2		
Hydrolysable nitrogen	MLR	11.00	0.86	0.69	12.38	0.92	0.60
	RF	9.21	0.28	0.83	8.51	0.47	0.85
	SVMr	4.51	0.64	0.95	8.91	0.23	0.79
Available phosphorus	MLR	47.62	4.19	0.12	48.96	4.29	0.07
	RF	33.71	1.39	0.66	34.62	3.10	0.65
	SVMr	33.39	6.80	0.57	35.26	7.83	0.52
Available Potassium	MLR	28.80	2.27	0.57	31.59	2.56	0.48
	RF	23.89	1.72	0.77	25.34	1.48	0.74
	SVMr	19.23	2.31	0.81	22.01	4.28	0.75
SOC	MLR	0.53	0.04	0.51	0.49	0.04	0.58
	RF	0.35	0.02	0.83	0.35	0.02	0.84
	SVMr	0.32	0.04	0.82	0.37	0.07	0.77
Silt	MLR	6.86	0.55	0.18	6.95	0.56	0.16
	RF	4.23	0.17	0.75	5.55	0.21	0.57
	SVMr	2.64	0.34	0.87	2.84	0.46	0.76
Clay	MLR	3.02	0.26	0.09	2.99	0.26	0.11
	RF	2.24	0.15	0.61	2.23	0.21	0.61
	SVMr	2.13	0.24	0.54	1.93	0.25	0.62

Available phosphorus in soils is predicted significantly worse than other agrochemical properties, using both Landsat 8 OLI and Sentinel 2 data. In both cases, the best model for available phosphorus is the RF model, for the model with Landsat data R2 = 0.66, with Sentinel 2 - R2 = 0.65. In general, acceptable results of spatial prediction of the content of available phosphorus by spectral reflectance can only be obtained using machine learning methods.

Insufficient prediction is also typical for particle size distribution (PSD) of soils. The worst predicted of all indicators is the clay content, for which with Landsat data, R2 = 0.61 in the RF model, and with Sentinel 2 data, R2 = 0.62. Silt content is predicted better than clay content, for which the highest R2 is observed in the SVMr model for both Landsat (R2 = 0.87) and Sentinel 2 (R2 = 0.76) data. As in the case of assessing the spatial heterogeneity of the content of available phosphorus, an acceptable prediction of the content of PSD fractions can be provided only with the use of special procedures for using machine learning methods.

Machine learning models are increasingly being used in studies of the spatial distribution of soil properties. For example, a review by Wadoux et al. indicates that a large number of machine learning algorithms and their variants are used in the digital soil mapping (DSM) literature. The most common DSM currently used is the RF algorithm [5]. However, SVMr algorithms also find application for DSM. Stevens et al. along with other regression models used spectral-based SVMr models to analyze cropland SOC, which also showed high determination coefficients for at different scales [6]. Kovačević et al also pointed out the value of regional scale SVMr models for assessing clay and physical sand in soils, which were assessed based on the classification of soils. The coefficients of determination ranged from 0.36 to 0.76 depending on the model. The SVMr models had higher R2 values for SOM (R2 = 0.96), nitrogen (R2 = 0.85), and pH (R2 = 0.90) [7]. Deiss et al. noted that when using spectral data, nonlinear models (SVMr) outperformed linear models of sand, clay, pH, total carbon in soils in Tanzania and the US Midwest. The performance of the SVMr (R2) models varied within 0.57-0.94 depending on the soil property [8]. A similar pattern for all soil parameters was observed in our study.



a) Landsat 8 OLI b) Sentinel 2

Figure 1. Example of predictive maps of SOC content obtained using satellite data from Landsat 8 OLI (a) and Sentinel 2 (b).

The figure 1 shows, as an example, predictive maps of SOC content obtained using the Landsat 8 OLI and Sentinel 2 satellite data. A visual analysis of the refined maps shows that the maps based on Sentinel 2 are characterized by a greater degree of detail in the spatial variability of the studied soil property. Approximately the same pattern is observed in the refined maps of other soil properties. It is known that the Sentinel 2 data are more sensitive to local changes in soil parameters, in contrast to the Landsat 8 data, which have a coarser resolution. Similar conclusions were made in other works, for example, when studying saline soils in China [9]. However, at the same time, it can be noted that when applying refined digital maps based on Earth remote sensing data from the Landsat satellite, the use of machine learning methods makes it possible to obtain a material with the required spatial prediction accuracy and detail that satisfies the information support of the variable rate application of mineral fertilization

4. Conclusions

The use of spectral reflectance data to refine digital maps of SOC content, available forms of nitrogen, phosphorus, potassium, PSD fractions on a single field scale provides a significant improvement in prediction when using machine learning methods (RF and SVMr) compared to traditional linear models (MLR). When using the RF and SVMr models, the SOC, available nitrogen, and available potassium is well predicted; the content of available phosphorus and PSD fractions is somewhat worse. Refined digital maps based on Sentinel 2 data are characterized by a greater degree of detail in the spatial

variability of soil parameters; at the same time, the use of Landsat data can using machine learning methods, make it possible to obtain maps that are sufficient in terms of spatial prediction accuracy for the requirements of digital farming.

Author Contributions: I. S., E.V., and K.G. conceived and designed the experiments, performed the ones, analyzed the data and wrote the paper. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: This work was supported in part by the Russian Foundation for Basic Research, research project № 19-29-05061-mk.

Conflicts of Interest: The authors declare no conflicts of interestReferences

References

1. Oliver M.A. An Overview of Geostatistics and Precision Agriculture // in: Geostatistical Applications for Precision Agriculture. Springer Science+Business Media B.V. 2010. P. 1-34. DOI 10.1007/978-90-481-9133-8 1
2. Kerry R.; Oliver V.A., Frogbrook Z.L. Sampling in Precision Agriculture // in: Geostatistical Applications for Precision Agriculture. Springer Science+Business Media B.V. 2010. P. 35-64. DOI 10.1007/978-90-481-9133-8 1
3. R Development Core Team, 2018. R : a language and environment for statistical computing. R Foundation for Statistical Computinghttp://www.R-project.org.
4. Harrell F. E. Jr. Resampling, Validating, Describing, and Simplifying the Model. Regression Modeling Strategies. 2001. 87-103
5. Wadoux A. M.J.-C.; Minasny B.; McBratney A. B.; Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews. 2020. 210. 103359. https://doi.org/10.1016/j.earscirev.2020.103359
6. Stevens A; Udelhoven T; Denis A; Tychon B; Lioy R; Hoffmann L; Van Wesemael B. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. Geoderma. 2010. 158(1). 32–45. https://doi.org/10.1016/j.geoderma.2009.11.032
7. Kovačević M.; Bajat B.; Gajić B. Soil type classification and estimation of soil properties using support vector machines. Geoderma 2010. 154. 340–347 https://doi.org/10.1016/j.geoderma.2009.11.005
8. Deiss L.; Margenot A. J.; Culman S. W.; Demyan M. S. Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. Geoderma. 2010. 365. 114227 https://doi.org/10.1016/j.geoderma.2020.114227
9. Wang J.; Ding J.; Yu D.; Teng D.; He B.; Chen X.; Ge X.; Zhang Z.; Wang Y.; Yang X.; Shi T.; Su F. Machine learning-based detection of soil salinity in an arid desert region, Northwest China: A comparison between Landsat-8 OLI and Sentinel-2 MSI. Science of the Total Environment. 2020. 707136092. DOI: 10.1016/j.scitotenv.2019.136092