https://doi.org/10.5194/soil-2019-83
Preprint. Discussion started: 7 November 2019

Oblique geographic coordinates as covariates for digital soil mapping

Anders Bjørn Møller[1], Amélie Marie Beucher[1], Nastaran Pouladi[1], Mogens Humlekrog Greve[1]

[1]Department of Agroecology, Aarhus University, Tjele, 8830, Denmark

Correspondence to: Anders Bjørn Møller (anbm@agro.au.dk)

Abstract. Decision tree algorithms such as Random Forest have become a widely adapted method for mapping soil properties in geographic space. However, implementing explicit geographic relationships into these methods has proven problematic. Using x- and y-coordinates as covariates gives orthogonal artefacts in the maps, and alternative methods using distances as covariates can be inflexible and difficult to interpret. We propose instead the use of coordinates along several axes tilted at oblique angles to provide an easily interpretable method for obtaining a realistic prediction surface. We test the method for mapping topsoil organic matter contents in an agricultural field in Denmark. The results show that the method provides accuracies on par with the most reliable alternative methods, namely kriging and the use of buffer distances to the training points. Furthermore, the proposed method is highly flexible, scalable and easily interpretable. This makes it a promising tool for mapping soil properties with complex spatial variation. We believe that the method will be highly useful for mapping soil properties in larger areas, and testing it for this purpose is a logical next step.

1 Introduction

Machine learning has become a frequently applied means for mapping soil properties in geographic space. The most common approach is to train models from soil observations and covariates in the form of geographic data layers. The models can often provide reliable predictions of soil properties. Many researchers have used decision tree algorithms, as they are computationally efficient, do not rely on assumptions about the distributions of the input variables, can use both numeric and categorical data and are immune to correlated and redundant covariates (Quinlan, 1996, Mitchell, 1997, Rokach and Maimon, 2005, Tan et al., 2014). Additionally, they effectively handle nonlinear relationships and complex interactions (Strobl et al., 2009).

However, a disadvantage of decision tree models is that they do not explicitly take into account spatial relationships. Unlike geostatistical methods, such as kriging, the predictions can therefore contain biases in the form of spatially autocorrelated residuals.

A number of studies have applied regression-kriging as a solution (Knotters et al., 1995, Odeh et al., 1995, Hengl et al., 2004). By kriging the residuals of the predictive model and adding the kriged residuals to the prediction surface, soil mappers have been able to reduce or remove spatial biases. A disadvantage of this approach is that the combination of two models hinders the combination of spatial relationships with the other covariates. Spatial relationships therefore remain disconnected from other statistical relationships in the analysis, leading to difficulties in interpreting the model and its associated uncertainties.

An obvious solution to this problem would be to use the x- and y-coordinates of the soil observations as covariates. However, results have shown that this approach can lead to unrealistic orthogonal artefacts in the output maps when used in conjunction with decision tree algorithms (Behrens et al., 2018, Hengl et al., 2018, Nussbaum et al., 2018). The cause of this problem lies in the splitting procedure of decision tree algorithms, as they use only one covariate for each split. Therefore, a dataset containing only the x- and y-coordinates will force the algorithm to make orthogonal splits in geographic space.

Several researchers have proposed solutions to this problem. Behrens et al. (2018) proposed the use of Euclidean distance fields (EDF) in the form of distances to the corners and middle of the study area as well as the x- and y-coordinates of the soil observations. Their results showed that this approach efficiently integrated spatial relationships and that accuracies were better than or on par with other methods for integrating spatial context.

On the other hand, Hengl et al. (2018) suggested an approach referred to as spatial Random Forest (RFsp). This method consists of calculating data layers with buffer distances to each of the soil observations in the training dataset. It then trains a Random Forest model, using the buffer distances as covariates, combined with auxiliary data or on their own. The authors demonstrated the method on a large number of spatial prediction problems and showed that it effectively eliminated spatial autocorrelation from the residuals.

Although these two methods are able to integrate spatial relationships in machine learning models, they are not without shortcomings. Firstly, the distances used in both methods depend either on the geometry of the study area, in the case of EDF, or on the locations of the soil samples, in the case of RFsp. The meaning and interpretation of the distances therefore varies depending on the study area and the soil observations.

Another shortcoming relating to EDF and RFsp is that both methods specify the number of geographic data layers a priori. For EDF, the number of distance fields is seven, and for RFsp, the number of buffer distances is equal to the number of soil observations. This means that there is no straightforward way to increase the number of spatially explicit covariates, if the number is insufficient to account for spatial relationships. In addition, vice versa, there is no way to decrease the number of spatially explicit covariates even if a smaller number would suffice. The latter is especially relevant for RFsp, as the method is computationally unfeasible for datasets with a large number of observations (Hengl et al., 2018).

In this study, we propose an alternative method for including spatially explicit covariates for mapping soil properties. With the method, we aim to address directly the cause of orthogonal artefacts arising from the use of x- and y-coordinates as covariates in decision tree models. Furthermore, we aim to improve upon the shortcomings of previous methods by developing a method that is both flexible and easily interpretable.

We refer to the method as Oblique Geographic Coordinates (OGC). In short, it works by calculating coordinates for the observations along a series of axes, tilted at various oblique angles relative to the x-axis. By including oblique coordinates as covariates, we enable the decision tree algorithm to make oblique splits in geographic space. As this is not possible with only x- and y-coordinates as covariates, this addition should allow the model to produce a more realistic prediction surface. Furthermore, as the number of oblique angles is adjustable, it should be possible to optimize it, both in terms of accuracy and computational efficiency.

We test the method for predicting soil organic matter contents in a densely sampled agricultural field in Denmark, located in northern Europe. We hypothesise that OGC can provide accuracies on par with previous methods for including explicitly spatial covariates. We also hypothesize that it is possible to adjust the number of oblique angles in order to optimize accuracy, and that the results allow meaningful interpretations.

2 Materials and methods

2.1 Study area

The study area is a 12-ha agricultural field located in Denmark in northern Europe (9.568°E; 56.375°N, ETRS 1989) (Figure 1). It lies in a kettled moraine landscape 55 – 66 m above sea level. The parent materials in the field include clay till, glaciofluvial sand and peat. The climate is temperate coastal, with mean monthly temperatures ranging from 1°C in January to 17°C in July and a mean annual precipitation of 850 mm (Wang, 2013). The field contains 285 measurements of soil organic matter (SOM) from the depth interval 0 – 25 cm, located in a 20 m grid.

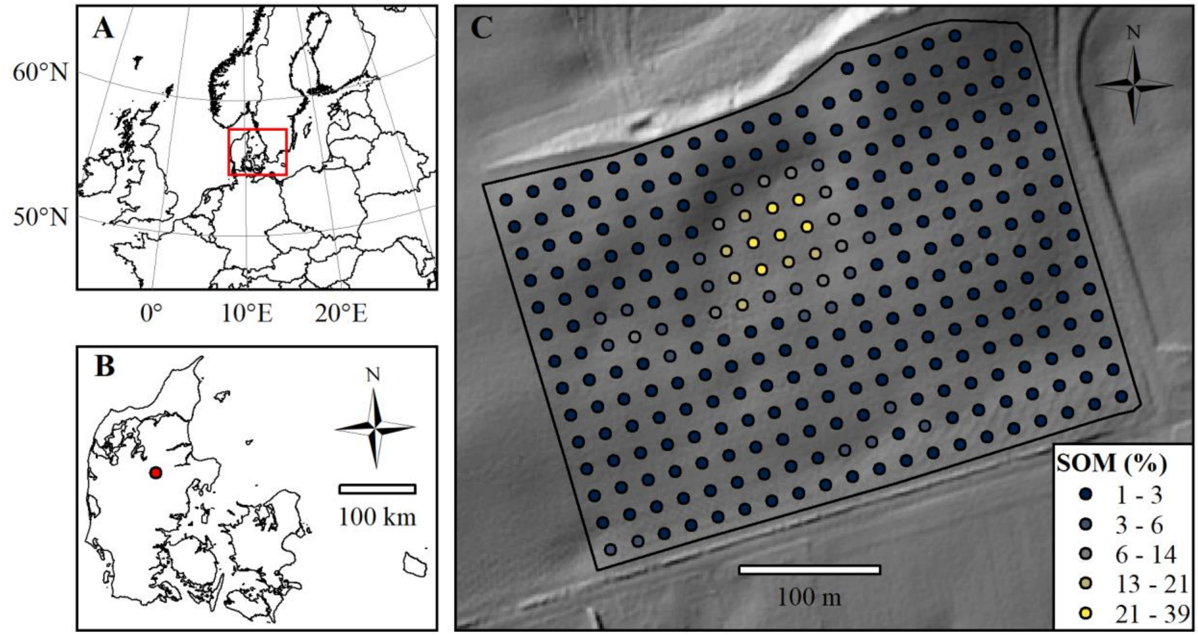


Figure 1: A: Location of Denmark in northern Europe. B: Location of the study area within Denmark. C: Map of the study area, including locations of the samples extracted for soil organic matter (SOM) measurements. The background shows hill shade (northwest, 45° altitude) based a digital elevation model (DEM) in 1.6x1.6 m resolution (National Survey and Cadastre, 2012).

The SOM contents of the topsoil in the field range from 1.3% to 38.8% with a mean value of 3.5% and a median of 2.2%. The values have a strong positive skew of 4.7 and are leptokurtic with a kurtosis of 26.9. Logarithmic transformation reduces skewness (2.9) and kurtosis (11.1). Pouladi et al. (2019) described that spatial structure of the data with a stable variogram with 139 m range, nugget of 0 and sill of 23.8.

2.2 Oblique geographic coordinates

The method that we propose consists of calculating coordinates along a number of axes titled at various oblique angles, relative to the x-axis. In the following, we show that it is possible to calculate the coordinate of a point *(b_1, a_1)* along an axis tilted at an angle θ relative to the x-axis, based on θ and the x- and y-coordinates of *(b_1, a_1)*. We also show that it is possible to derive the calculation using basic trigonometry. Equations (1), (2), (3) and (4) show the derivation of the calculation of the oblique coordinate, using Figure 2 for illustration.

$$b_2 = c * \cos A_2 \tag{1}$$

$$c = \sqrt{{a_1}^2 + {b_1}^2} \tag{2}$$

$$A_2 = \theta - A_1$$

$$A_1 = \tan^{-1}\frac{a_1}{b_1} \tag{3}$$

$$b_2 = \sqrt{{a_1}^2 + {b_1}^2} * \cos\left(\theta - \tan^{-1}\frac{a_1}{b_1}\right) \tag{4}$$

, where θ is the angle of the titled axis relative to the x-axis; A_1 is the angle between the x-axis and the line *c* between the origin of the coordinate system and the point *(b_1, a_1)*; A_2 is the difference between θ and A_1; a_1 is the y-coordinate of *(b_1, a_1)*; b_1 is the x-coordinate of *(b_1, a_1)*; b_2 is a line with the angle θ between the origin of the coordinate system its intersection with a_2; a_2 is a line perpendicular to b_2 going from *(b_1, a_1)* to its intersection with b_2. The length of b_2 is equal to the coordinate of *(b_1, a_1)* along an axis tilted with the angle θ relative to the x-axis.

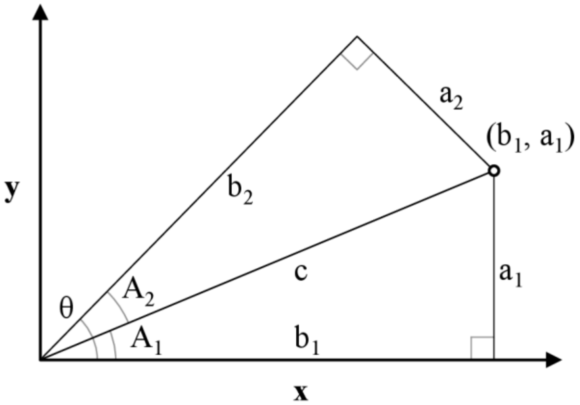


Figure 2: Illustration for the derivation of the coordinate for the point *(b_1, a_1)* along an axis tilted with the angle θ from the x-axis. The coordinate is equal to the length of b_2. Triangles *a_1b_1c* and *a_2b_2c* are right triangles with the same hypotenuse *c*. The sides *a_1*

and b_1 are the x- and y-coordinates of the point (b_1, a_1), respectively. A_1 is the angle between the x-axis and the line c between the origin of the coordinate system and the point (b_1, a_1); A_2 is the difference between θ and A_1.

As the x- and y-coordinates of soil observations are known, and θ is given, it is possible to calculate coordinates at oblique angles for all soil observations in a dataset. Likewise, as the x- and y-coordinates of the cells in a geographic raster layer are known, it is possible to calculate oblique coordinates for the cells. Our approach relies on calculating coordinates along n axes tilted at angles ranging from 0 to $\pi((n - 1)/n)$ with increments of π/n between the angles. θ should not be π or greater, as coordinates along axes tilted at these angles will correlate with coordinates along axes tilted at angles of 0 to $\pi((n - 1)/n)$. For example, coordinates along an axis with $\theta = 0.25\pi$ (northeast) perfectly correlate with coordinates along an axis with $\theta = 1.25\pi$ (southwest). Figure 3 shows coordinates along axes tilted at six different angles relative to the x-axis for the study area. The coordinate rasters A and D are equivalent to the x- and y-coordinates, respectively, while the coordinate rasters B, C, E and F show coordinates at oblique angles.

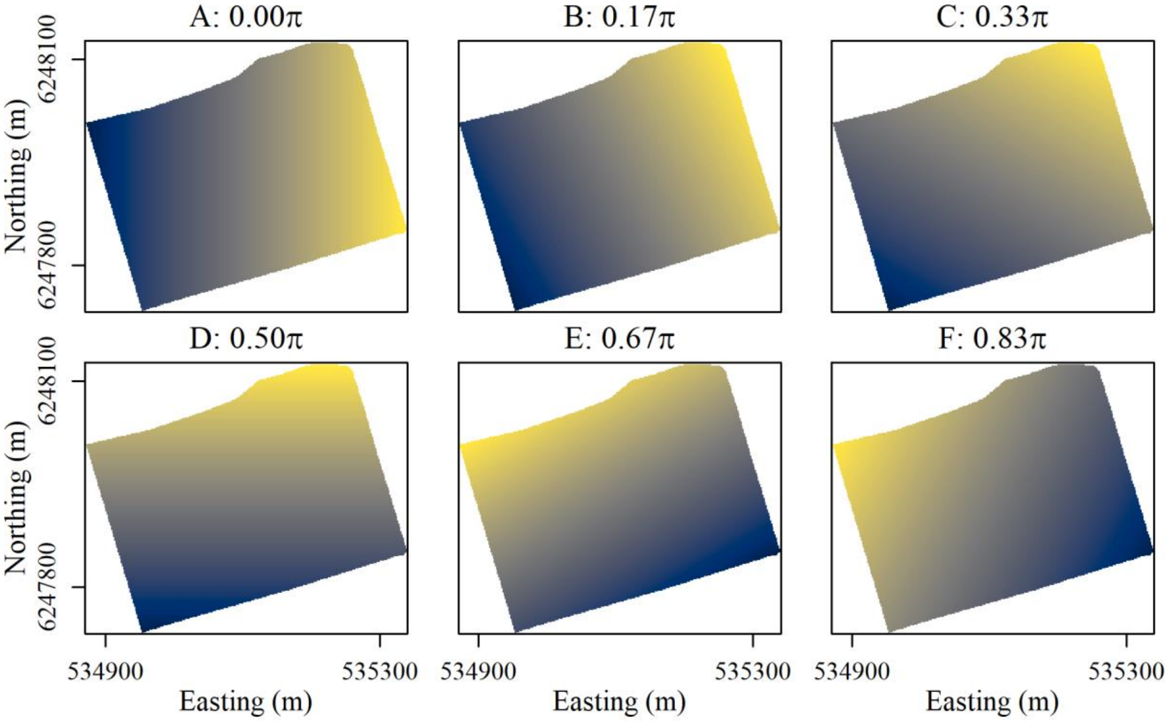


Figure 3: Examples of rasters with coordinates tilted at six different angles for the study area. Easting and northing for UTM Zone 32N, ETRS 1989.

2.3 Experiments

We use the 285 SOM observations from the study area in order to test the accuracy of predictions made by Random Forest models using OGC as covariates. In addition to OGC, we also employed 19 data layers with auxiliary data, which Pouladi et al. (2019) derived from a 1.6 m DEM, satellite imagery and electromagnetic induction. Topographic variables included the

https://doi.org/10.5194/soil-2019-83
Preprint. Discussion started: 7 November 2019

sine and cosine of the aspect, depth of sinks, plan and profile curvature, elevation, flow accumulation, valley bottom flatness, mid-slope position, standard and modified topographic wetness index, slope gradient, slope length and valley depth. Satellite imagery included normalized difference, absolute difference, ratio and soil-adjusted vegetation indices. Lastly, we used the apparent electrical conductivity from a DUALEM 1 sensor in perpendicular mode.

Table 1: Auxiliary data variables used as covariates in the study, including name, description, the mean value and the range. Pouladi et al. (2019) describe the derivation of the variables.

Predictor variable	Description	Mean (range)
DEM		
Cos aspect	Cosine of surface aspect	-0.1 (-1.0 - 1.0)
Sin aspect	Sine of surface aspect	0.32 (-1.0 - 1.0)
Depth of sinks	Depth of sinks (m)	0.1 (0.0 - 1.1)
Plan curvature	Shape of the surface in the horizontal plane	0 (-34 - 15)
Profile curvature	Shape of the surface in the vertical plane	0.00 (-0.06 - 0.04)
Elevation	Elevation from DEM; m above sea level	60.8 (54.6 - 66.2)
Flow accumulation	Number of upslope cells	74 (3 - 8969)
MRVBF	Multiresolution index of valley bottom flatness	1.5 (0.0 - 4.9)
Mid-slope position	Covers the warmer zones of slopes	0.5 (0.0 - 1.0)
SAGA wetness index	SAGA GIS modified topographic wetness index	4.0 (2.2 - 8.6)
Slope gradient	Local slope gradient (degrees)	4.9 (0.0 - 17.5)
SL	Slope length factor	0.4 (0.0 - 2.3)
TWI	Topographic wetness index	6.6 (3.7 - 14.6)
Valley depth	Depth of valleys (m)	1.4 (0.1 - 8.1)
Sentinel 2		
DVI	Difference vegetation index	1735 (1202 - 3294)
NDVI	Normalized difference vegetation index	0.5 (0.3 - 0.7)
RVI	Ratio vegetation index	2.8 (2.0 - 6.4)
SAVI	Soil-adjusted vegetation index	0.7 (0.5 - 1.1)
DUALEM 1mPRP		
ECa	Apparent electrical conductivity	8.9 (4.9 - 16.0)

In order to optimize the number of raster layers for OGC, we generated datasets with 2 – 100 coordinate rasters. We then trained Random Forest models from each dataset, both with and without auxiliary data. In order to assess predictive accuracy, we used 100 repeated splits on the SOM observations, each using 75% of the observations for model training and a

25% holdout dataset for accuracy assessment. We trained models using the R package *ranger* (Wright and Ziegler, 2015) and parameterized the models using the R package *caret* (Kuhn, 2008). For each split, we tested five different values for *mtry*, minimum node sizes of 1, 2, 4 and 8, and two different splitting rules *variance* and *extratrees*. We mainly adjusted *mtry* and the minimum node size in order to avoid overfitting. The *extratrees* splitting rule allows suboptimal splits, which can increase randomization relative to the default *variance* splitting rule (Geurts et al., 2006). We selected the setup that provided the lowest RMSE for the out-of-bag predictions on the training data, and used this setup for predictions on the 25% holdout dataset.

We used the same 100 repeated splits for each number of coordinate rasters, with and without auxiliary data. We calculated accuracy based on R^2, RMSE and Lin's concordance criterion (ccc), and subsequently used the number of coordinate rasters that yielded the lowest RMSE. We selected a different number of coordinate rasters with and without auxiliary data.

We then compared the accuracies obtained with the optimal numbers of coordinate rasters, with and without auxiliary data, to the accuracies obtained with other methods. We tested kriging, Random Forest models trained only on the auxiliary data and Random Forest models trained using EDF and RFsp, with and without auxiliary data. We trained the Random Forest models using the same procedure outlines above. For kriging, we used variograms automatically fitted on logarithmic-transformed SOM observations using the *autofitVariogram* function of the R package *automap* (Hiemstra, 2013).

We used the same 100 repeated splits for assessing the accuracies of all methods. This allowed us to carry out pairwise t-tests between the accuracies of the methods. We used the results of the pairwise t-tests to rank the methods according to their accuracies according to each of the metrics. If there was no statistical difference ($p > 0.05$) between the accuracies of two or more methods, these methods received the same rank. We calculated separate ranks for the methods for each accuracy metric, resulting in three different sets of ranks.

In order to illustrate the results, we produced maps of SOM with each method, using models trained from all the data. Furthermore, we investigated the covariate importance of models trained with OGC and tested the results for spatially autocorrelated residuals using experimental variograms.

3 Results and discussion

3.1 Optimal number of coordinate rasters

Without auxiliary data, accuracies of predictions obtained with OGC increased with the number of coordinate rasters up to an optimum at seven coordinate rasters (Figure 4). However, with more than seven coordinate rasters, accuracies deteriorated slightly with the number of coordinate rasters. This pattern was the same for all three metrics. On the other hand, with OGC in combination with auxiliary data, accuracies generally increased with the number of coordinate rasters. The increase was greatest when the number of coordinate rasters was small, while the effect of more coordinate rasters decreased for larger numbers of coordinate rasters. With auxiliary data, the optimal number of coordinate rasters was 94 for R^2, 80 for RMSE and 89 for ccc. Accuracies with auxiliary data were almost invariably higher than accuracies achieved without auxiliary data.

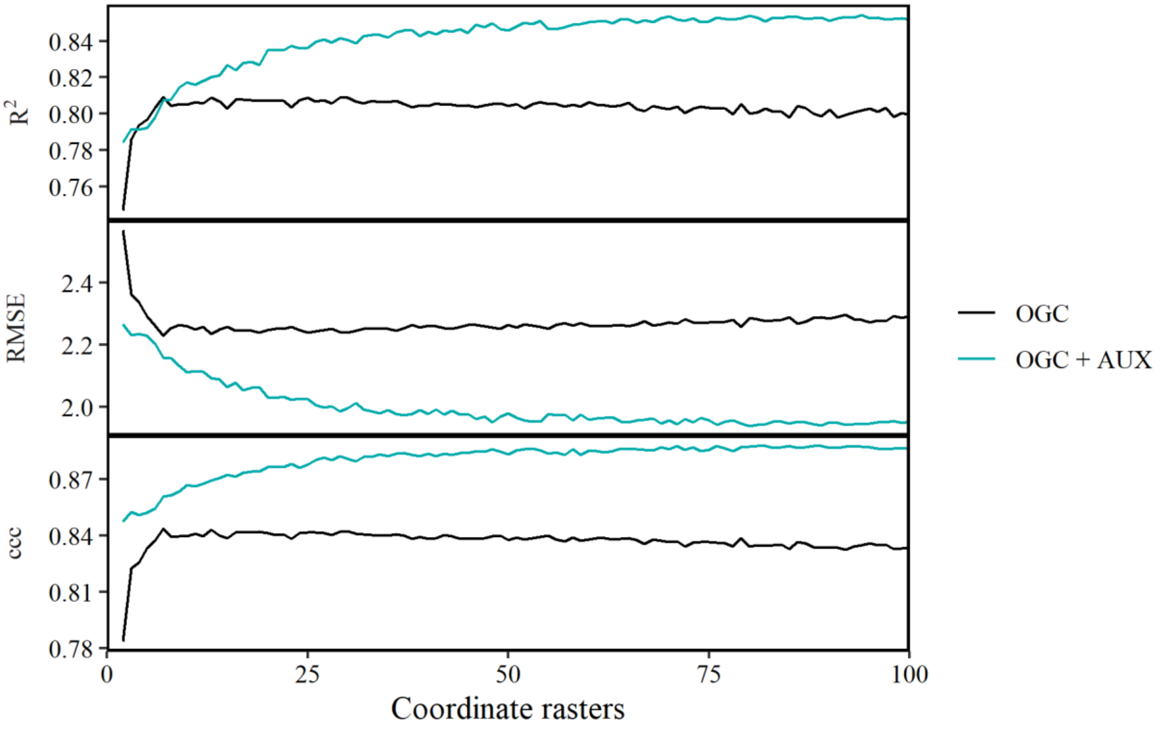


Figure 4: Effects of the number of coordinate rasters on the accuracy of SOM predictions, calculated as R^2, root mean square error (RMSE) and Lin's concordance criterion (ccc). We calculated effects for Random Forest models trained on only coordinate rasters (OGC) and with coordinate rasters in combination with auxiliary data (OGC + AUX). The lines represent mean values obtained from 100 repeated splits (75% training dataset, 25% test dataset) for each number of coordinate rasters.

Figure 5 shows the effect of increasing the number of coordinate rasters without auxiliary data. The predictions with only two coordinate rasters show a pattern very typical of predictions with x- and y-coordinates with very visible orthogonal artefacts. As the number of coordinate rasters increases, the patterns of the artefacts change. With coordinate rasters at three different angles, the artefacts have a hexagonal pattern, and with coordinate rasters at four different angles, the artefacts gain an octagonal pattern. Furthermore, as the number of coordinate rasters increases, the artefacts become less pronounced. Although some artefacts are visible with coordinate rasters at seven different angles, they are much less visible than the artefacts in the map produced with only two coordinate rasters.

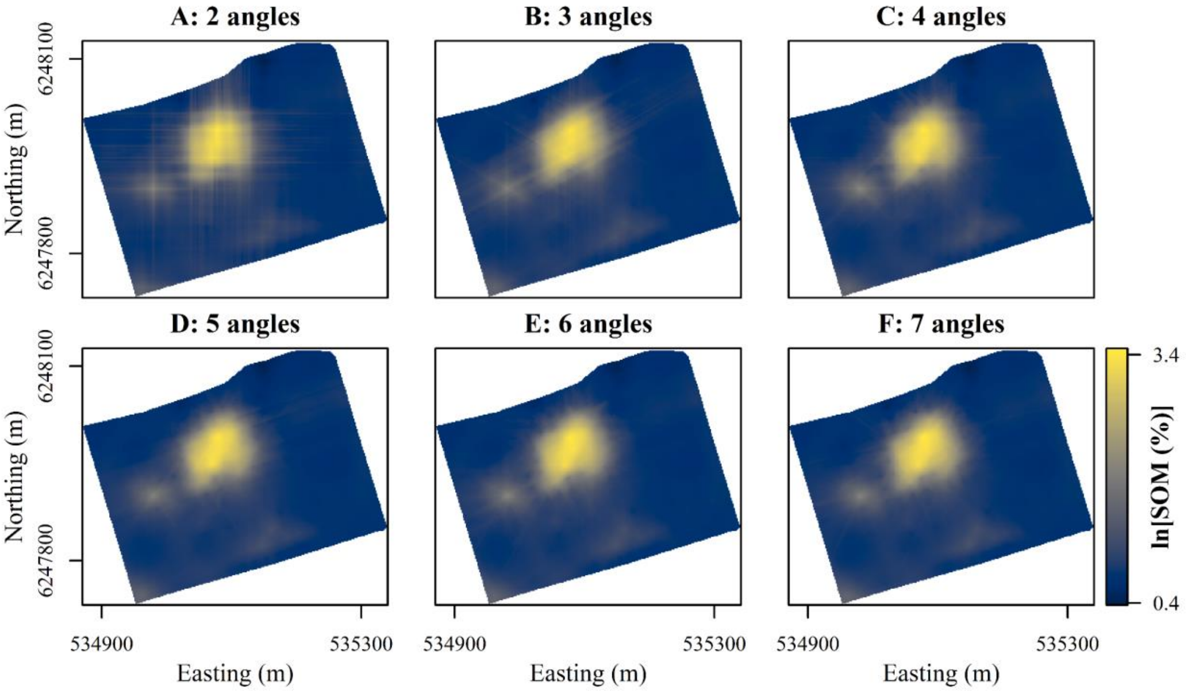


Figure 5: Maps of soil organic matter (SOM) contents in the topsoil predicted using Random Forest models trained with coordinate rasters at two to seven different angles as covariates. Easting and northing for UTM Zone 32N, ETRS 1989. 180

With auxiliary data, the effect of increasing the number of coordinate rasters was less clearly visible (Figure 6). Even with only two coordinate rasters, the predictions had no orthogonal artefacts. However, they contained noisy patterns and sharp boundaries in some areas. This is most likely an artefact from the auxiliary data. For example, using a high-resolution DEM may have created noise in the predictions. However, with coordinate rasters at 80 different angles, the spatial pattern of the predicted SOM contents became substantially smoother, with a reduction both in noise and in sharp boundaries. Furthermore, some areas with moderately high SOM contents became more clearly visible and coherent, for example in the area approximately one third of the way from the western to the northern corner of study area. The predicted patterns with a higher number of coordinate rasters were therefore not only more accurate, but also more realistic.

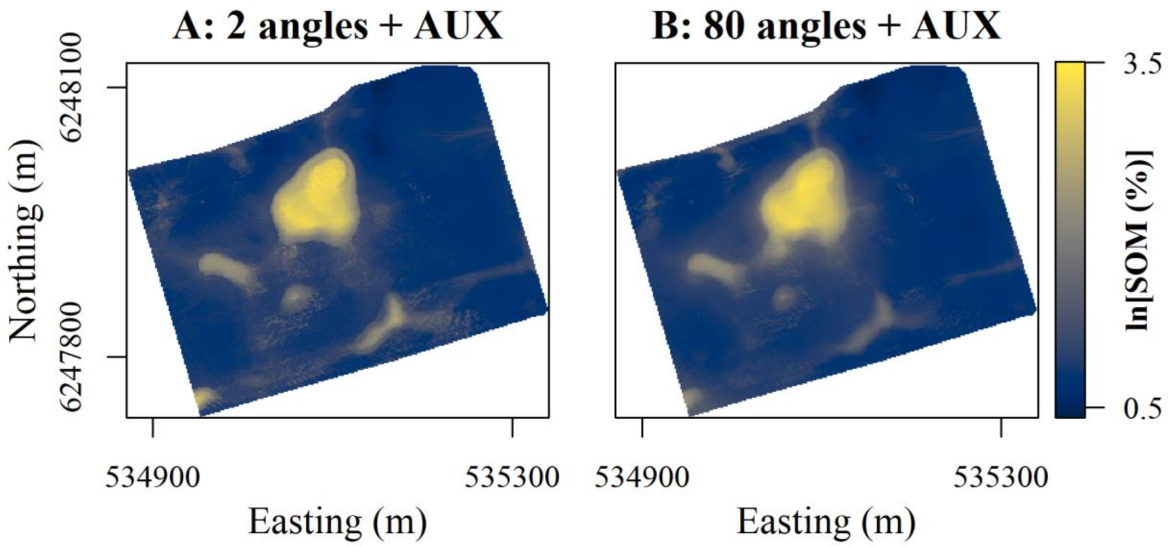


Figure 6: Maps of soil organic matter (SOM) contents in the topsoil predicted using Random Forest models trained using auxiliary data in conjunction with coordinate rasters at (A) two and (B) 80 different angles as covariates. Easting and northing for UTM Zone 32N, ETRS 1989.

3.2 Comparison with other methods

There were large overlaps in the accuracies of the methods, as accuracies varied across the 100 repeated splits (Figure 7), especially for RMSE. However, accuracies generally correlated between the methods across the repeated splits. The mean correlation coefficient (Pearson's R) was 0.52 (0.19 – 0.88) for R^2, 0.71 (0.65 – 0.71) for RMSE and 0.65 (0.41 – 0.89) for ccc. This shows that some holdout datasets yielded consistently high accuracies, while others yielded consistently low accuracies. Furthermore, especially for R^2 and ccc, a few holdout datasets yielded much lower accuracies than the other holdout datasets, leading to long negative tails.

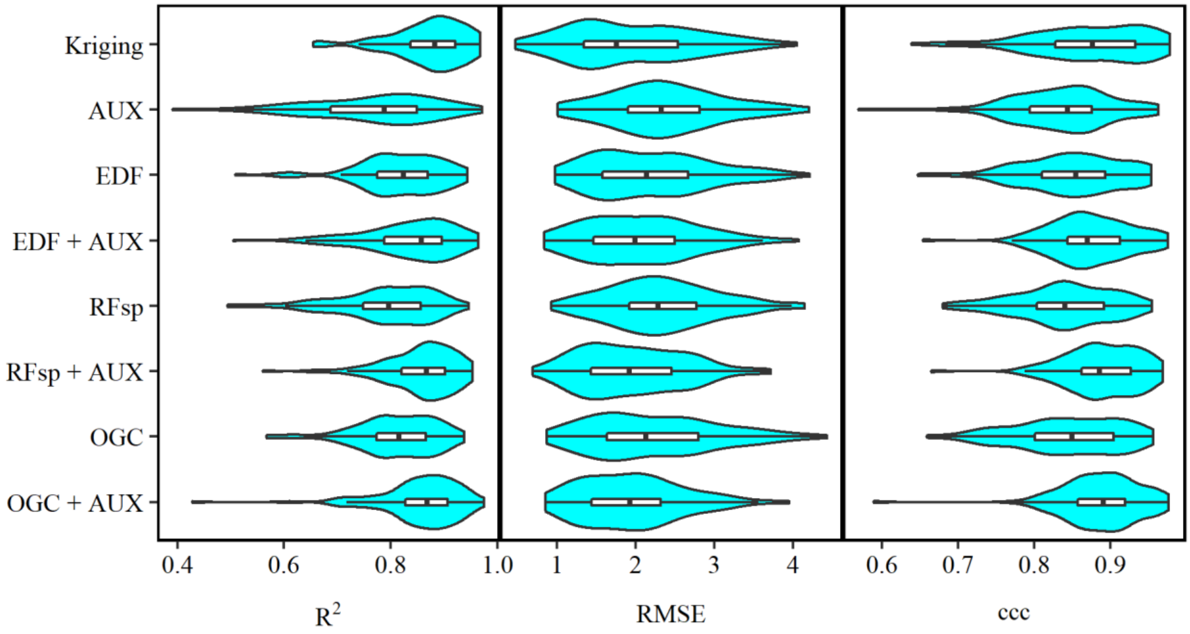


Figure 7: Violin plots showing accuracies of soil organic matter predictions with kriging, and Random Forest models trained using either auxiliary data (AUX), Euclidean distance fields (EDF), distances to observations (RFsp), oblique geographic coordinates (OGC) or EDF, RFsp or OGC in conjunction with AUX. The figure shows R^2, room mean square error (RMSE) and Lin's concordance obtained from 100 repeated splits (75% training dataset, 25% test dataset).

Kriging achieved the highest rank for R^2 (Table 2). For RMSE, kriging shared the highest rank with EDF, RFsp and OGC in combination with auxiliary data. Lastly, OGC and RFsp in combination with auxiliary data shared the highest rank for ccc. In short, kriging, RFsp with auxiliary data and OGC with auxiliary data all had the highest rank for two accuracy metrics out of three. We therefore regard these three methods as most accurate.

Auxiliary data used on their own, as well as RFsp without auxiliary data had the lowest rank for all three accuracy metrics. Furthermore, OGC without auxiliary data had the same rank as EDF without auxiliary data for all three accuracy metrics.

Table 2: Ranks for the accuracies of the methods, calculated as R^2, RMSE and ccc, respectively. Methods for which a pairwise t-test did not give a significant difference in accuracy ($p > 0.05$) received equal ranks for the metric in question. Ranks for the methods therefore differ between the three metrics. AUX: Auxiliary data. EDF: Euclidean distance fields. OGC: Oblique geographic coordinates. RFsp: Spatial Random Forest.

	R^2		RMSE		ccc	
Rank	**Method**	**Mean**	**Method**	**Mean**	**Method**	**Mean**
1	Kriging	0.87	EDF + AUX	2.0	OGC + AUX	0.89
			Kriging	2.0	RFsp + AUX	0.89
			OGC + AUX	1.9		
			RFsp + AUX	1.9		
2	OGC + AUX	0.85	EDF	2.2	EDF + AUX	0.87
	RFsp + AUX	0.86	OGC	2.2	Kriging	0.87
3	EDF	0.82	AUX	2.4	AUX	0.84
	EDF + AUX	0.83	RFsp	2.3	EDF	0.85
	OGC	0.81			OGC	0.84
					RFsp	0.84
4	AUX	0.77				
	RFsp	0.79				

Pouladi et al. (2019) tested several methods for predicting SOM within the field, including kriging and the machine learning algorithms Cubist and Random Forest, with and without kriged residuals. The authors found that kriging provided the most accurate predictions of SOM. The results in this study affirm the high accuracy of kriging predictions, but they also show that Random Forest models combining auxiliary data with spatial relationships can achieve similar accuracies.

Kriging produced a smooth prediction surface, which is very common for this method (Figure 8A). The prediction surface with EDF was mostly smooth, but it also contained a distinct "rings in the water" artefact caused by the use of the distance to the middle of the study area as a covariate (Figure 8B). The prediction surface with RFsp was smoother than the prediction surface produced by kriging (Figure 8C). The predictions with only auxiliary data were very similar to the predictions made with x- and y-coordinates in combination with auxiliary data (compare Figure 8C to Figure 6A). In combination with auxiliary data, both EDF and RFsp produced smoothing effects similar to the effect seen with OGC in combination with auxiliary data (compare Figure 8E and Figure 8F to Figure 6B). However, for EDF the smoothing was less visible than with OGC, and for RFsp it was more visible than with OGC.

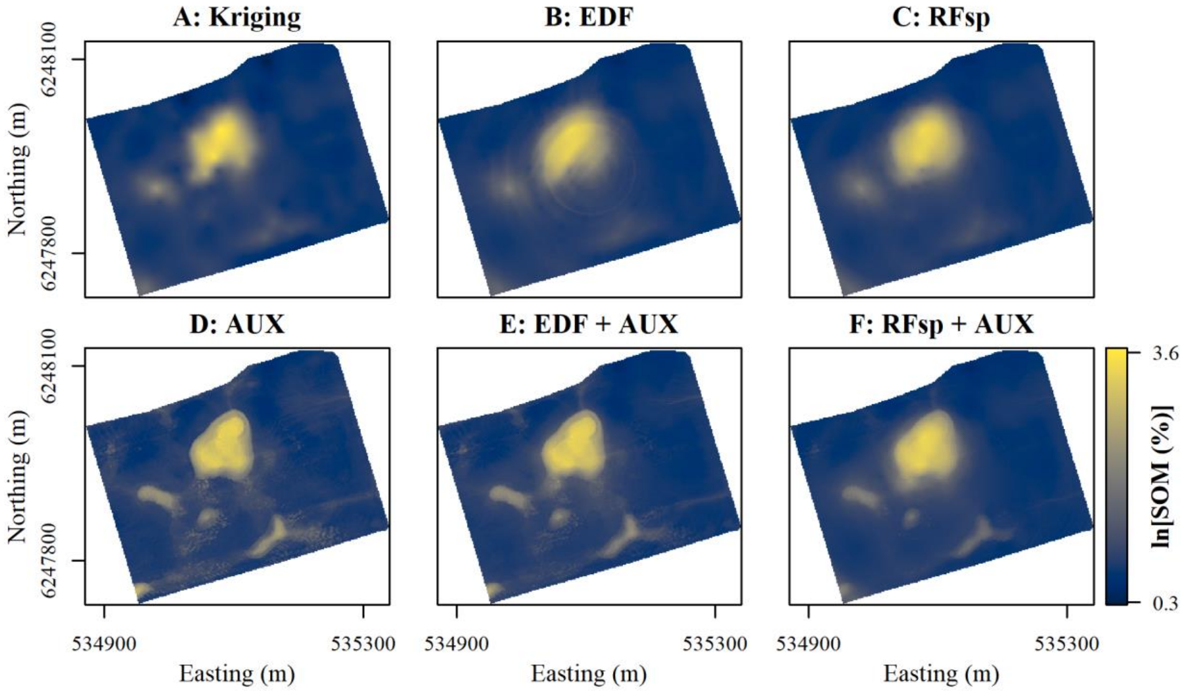
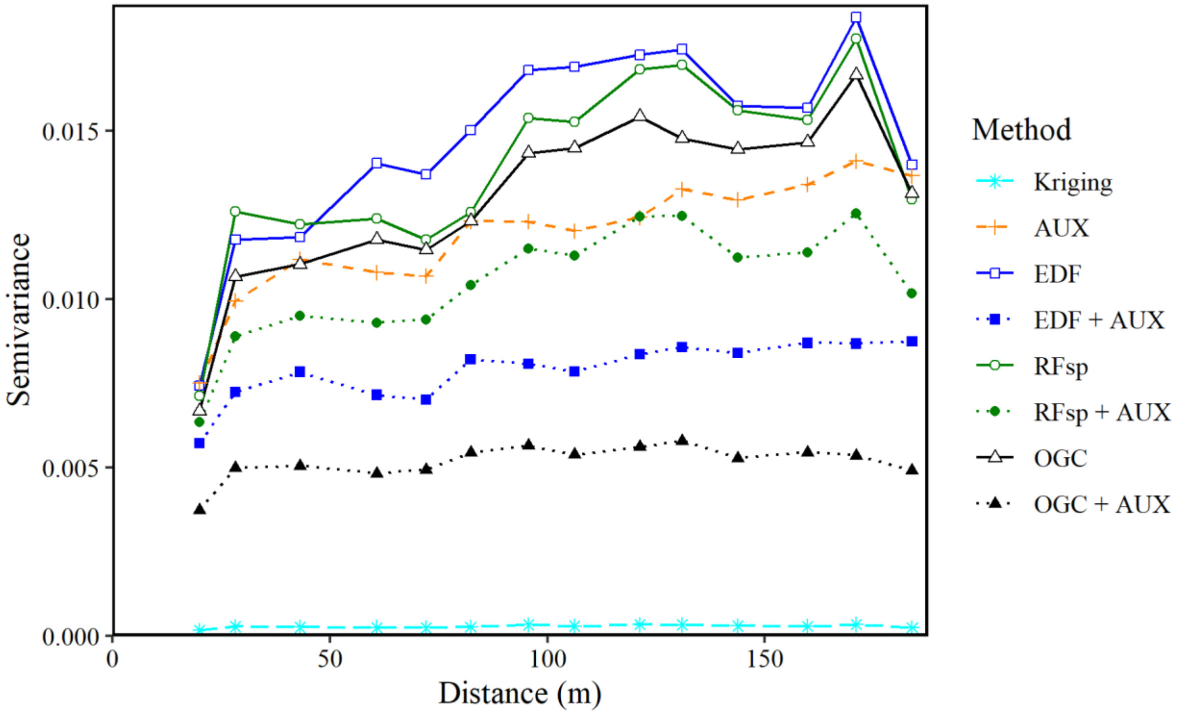


Figure 8: Prediction of soil organic matter (SOM) contents for the topsoil using A: Kriging, or Random Forest models trained with B: Euclidean distance fields (EDF), C: Distances to observations (RFsp), D: Auxiliary data (AUX), E: EDF in conjunction with AUX, or F: RFsp in conjunction with AUX. Easting and northing for UTM Zone 32N, ETRS 1989.

235 For all methods except kriging, the residuals of the SOM predictions had some degree of spatial dependence (Figure 9). EDF, RFsp and OGC used without auxiliary data had the most spatially dependent residuals, with nugget-to-sill ratios of 0.40. On the other hand, EDF and OGC in combination with auxiliary data had the least spatially dependent residuals after kriging, with nugget-to-sill ratios of 0.65 and 0.64, respectively.



240 **Figure 9: Experimental variograms for the residuals of the SOM predictions made with each method. The variograms use residuals from natural logarithmic-transformed SOM measurements and predictions. AUX: Auxiliary data. EDF: Euclidean distance fields. RFsp: Spatial Random Forest. OGC: Oblique geographic coordinates.**

3.3 Covariate importance

The most important covariate from the auxiliary data was the depth of sinks (Table 3). The most likely reason for its high 245 importance is the presence of a large sink with very high SOM contents northwest of the middle of the study area (Figure 1). As sinks trap surface runoff, they often have wet conditions, which give rise to peat accumulation.

Table 3: Covariate importance for the model using OGC in combination with auxiliary data. The importance for OGC represents the sum of the importance of the coordinate rasters at 80 different angles.

Covariate	Importance (variance)
OGC	2689
Depth of sinks	2003
MRVBF	476
SAGA wetness index	170
Elevation	166
Valley depth	157
EC_a	123
Slope gradient	101
Mid-slope position	84
NDVI	76
Plan curvature	74
SL	64
SAVI	58
Cos aspect	44
DVI	42
TWI	38
RVI	34
Flow accumulation	32
Sin aspect	32
Profile curvature	21

When used in combination with the auxiliary data, the importance of the individual coordinate rasters varied from 0.6% to 3.1% of the importance of the depth of sinks, with mean value of 1.7%. The most important coordinate raster had $\theta = 0.48\pi$ (close to a north-south axis) and was the 12th most important covariate. The sum of the importance of the coordinate rasters was equal to 134.3% of the importance of the depth of sinks (Table 3). Therefore, with coordinate rasters at 80 different angles, the effect of the individual rasters on the predictions was subtle, but their combined effect was strong.

Figure 10 shows the importance of the coordinate rasters relative to θ, in a way similar to a wind rose. The plots repeat the bars for $\theta \geq \pi$, as the importance for a given angle is directionless. For example, the importance of $\theta = 0$ (East) is equal to the importance of $\theta = \pi$ (West).

Without auxiliary data, the most important coordinate rasters had a general northwest to southeast angle (Figure 10). On the other hand, the coordinate rasters with angles between a north-south and a northeast-southwest axis had low importance. The most likely reason for this pattern is the location of the sink with very high SOM contents to the northwest of the middle of the study area. This creates a large difference in the SOM contents of the northwestern and southeastern parts of the study area, giving large importance to covariates that can explain this difference. Additionally, the northwest side of the sink has a very steep slope, creating a steep gradient in SOM contents in this direction. A stable variogram showed anisotropy along a north-north-east to south-south-west axis (θ = 0.34π) with a major range of 136 m and a minor range of 118 m. The direction of the anisotropy therefore coincided with the direction of the least important coordinate rasters.

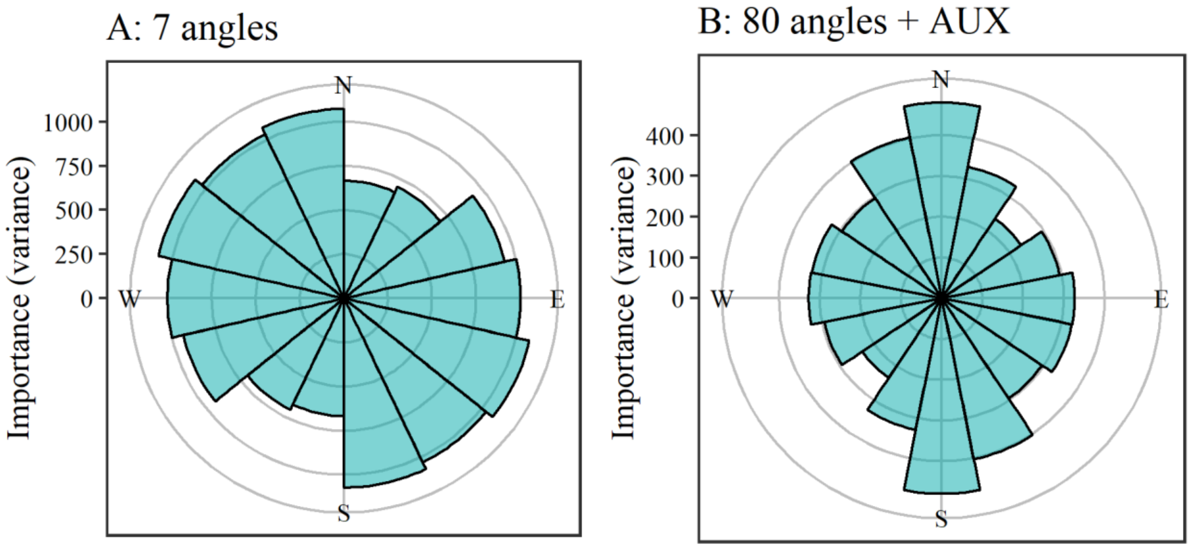


Figure 10: Covariate importance of the coordinate rasters at various angles. A: Importance of coordinate rasters at seven different angles. B: Importance of coordinate rasters at 80 different angles used with auxiliary data (importance of auxiliary data not shown). The sizes of the bars show the importance of the coordinate rasters at a given angle. Bars in B show the sum of the importance for coordinate rasters aggregated into 0.125π intervals.

On the other hand, with OGC in combination with auxiliary data, the most important coordinate rasters had tilt angles close to a north-south axis (θ = 0.5π). At the same time, the least important coordinate rasters had tilt angles close to a northeast-southwest axis (θ = 0.25π). The residuals from the predictions with auxiliary data only also displayed a degree of anisotropy. A stable variogram showed anisotropy along a northeast to southwest axis (θ = 0.21π), with a major range of 52 m and a minor range of 38 m. Again, the angle of the anisotropy coincided with the angle of the least important coordinate rasters. The spatial pattern of the residuals therefore differed from the spatial pattern of the SOM contents in the study area. Apparently, there are unaccounted-for processes decreasing the spatial variation along a northeast-southwest axis relative to other angles.

A possible cause of the anisotropy in the residuals may be the ploughing direction. The main ploughing direction in the study area is along an east-north-east to west-south-west axis (θ = 0.18π). This angle is nearly parallel to the angle of the least important coordinate rasters (Figure 11). The ploughing direction, combined with the topography, has a large impact on soil movement, as ploughing displaces soil both along and across its direction (Lindstrom et al., 1990, De Alba, 2003, Heckrath et al., 2006). Most of the study area has the same ploughing direction, irrespective of the topography, resulting in up-, down- and cross-slope ploughing in various parts of the field. This creates in a complex pattern of soil redistribution, which likely

https://doi.org/10.5194/soil-2019-83
Preprint. Discussion started: 7 November 2019

affects the SOM contents of the topsoil. As downslope soil movement is strongest in the ploughing direction, variation in soil properties parallel to this direction is likely to be smaller than the variation perpendicular to the ploughing direction. This corresponds to the low importance of coordinate rasters with angles close to the ploughing direction. However, none of the auxiliary data accounted for the ploughing direction. This indicates that OGC can add information on the most likely processes affecting soil properties in an area.

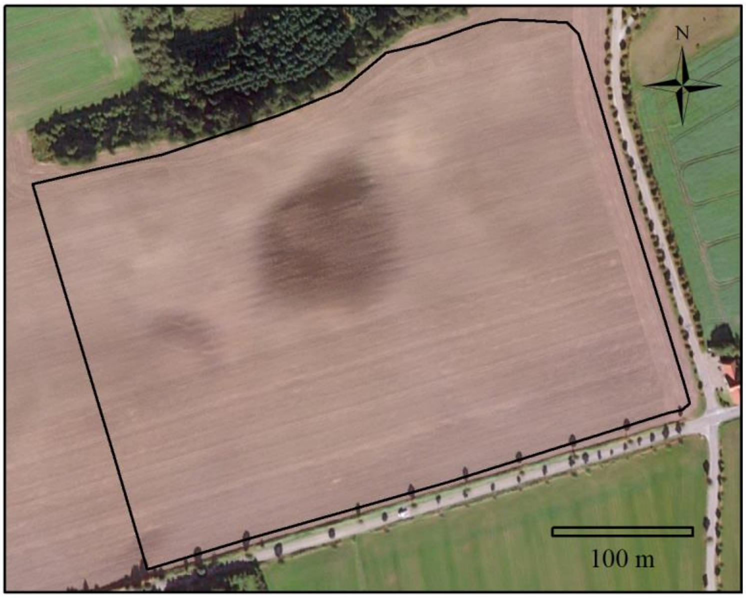


Figure 11: Orthophoto of the study area from September 27, 2016 (Esri, 2019). Sources: Esri, DigitalGlobe, Earthstar Geographics, CNES/Airbus DS, GeoEye, USDA FSA, USGS, Aerogrid, IGN, IGP, and the GIS User Community.

3.4 Choice of method

The three most accurate methods were kriging, RFsp with auxiliary data and OGC with auxiliary data. Many soil mappers would probably choose kriging for mapping SOM in this field, given its computational efficiency and conceptual simplicity. The advantage of the methods based on machine learning instead lies in their interpretability. Kriging in itself does not provide information on the processes that control spatial variation in soil properties, but researchers can interpret machine learning models in order to discover the most likely processes affecting the spatial distribution of a soil property. With spatial approaches such as EDF, RFsp and OGC, researchers can incorporate feature space and geographic space in a machine learning model. The benefit is that researchers can interpret local and spatial effects at once. In this regard, OGC has an advantage over EDF and RFsp, as it is clear what the coordinate rasters represent. On the other hand, it is less clear how researchers should interpret distances to the corners of the study area or the distance to a specific observation. We have also shown that it is straightforward to illustrate covariate importance for OGC.

While kriging, RFsp with auxiliary data and OGC with auxiliary data yielded equally accurate predictions, it is likely that it is due to the high sampling density of the study area. For larger, less densely sampled areas, OGC and RFsp with auxiliary data are likely to provide higher relative accuracies.

Furthermore, an advantage of OGC relative to RFsp is that OGC required fewer covariates to achieve the same accuracy. In fact, without auxiliary data, OGC achieved a higher accuracy with a smaller number of covariates. This demonstrates a clear advantage of OGC, as it is possible to adjust the number of coordinate rasters. EDF and RFsp do not presently have similar options.

We will stress that soil mappers should not use machine learning models relying only on spatial relationships, as EDF, RFsp and OGC all yielded low accuracies without auxiliary data. Moreover, surprisingly, these methods had the most spatially autocorrelated residuals, although they relied exclusively on spatial relationships. The results therefore suggest that these methods are best suited for integrating spatial relationships with auxiliary data. If the purpose requires a purely spatial method, kriging is a better option.

4 Conclusions

We have shown in this study that the use of oblique geographic coordinates is a reliable method for integrating auxiliary data with spatial relationships for modelling and mapping soil properties. It is more interpretable than previous similar approaches, and more flexible, as it is possible to adjust the number of coordinate rasters. This should allow soil mappers to find a good compromise between accuracy and computational efficiency for mapping soil properties, as the optimal number of coordinate rasters may vary depending on the study area and the soil property in question.

At this point, we have only tested the method for one soil property in one area, and it will therefore be highly relevant to test the method for other soil properties and areas. It will especially be relevant to test the method in larger, less densely sampled areas, as it may have its greatest relative advantages in these cases. Furthermore, the results also suggest that the method can be useful for predicting properties with anisotropic spatial distributions, and it will therefore be relevant to test it on datasets with a high degree of anisotropy.

We call upon researchers within digital soil mapping to aid us in this endeavour, and we have therefore made the function for generating oblique geographic coordinates available as an R package. Moreover, to allow other researchers to test methods on the dataset that we used, we have made it available as well.

5 Code and data availability

The function for generating oblique geographic coordinates is available as an R package at https://bitbucket.org/abmoeller/ogc/src/master/rPackage/OGC/. The package also contains the SOM observations and auxiliary data used in this study.

Furthermore, we have made the R code used in this study available in a public repository at http://dx.doi.org/10.5281/zenodo.3496935.

6 Author contribution

Anders Bjørn Møller and Nastaran Pouladi prepared the data. Anders Bjørn Møller carried out the analyses and prepared the manuscript with inputs from all co-authors.

7 Competing interests

The authors declare that they have no conflict of interest

References

Behrens, T., Schmidt, K., Viscarra Rossel, R., Gries, P., Scholten, T. and MacMillan, R. Spatial modelling with Euclidean distance fields and machine learning. Eur. J. Soil Sci. 69(5), 757-770. http://dx.doi.org/10.1111/ejss.12687, 2018.

De Alba, S. Simulating long-term soil redistribution generated by different patterns of mouldboard ploughing in landscapes of complex topography. Soil Tillage Res. 71(1), 71-86. http://dx.doi.org/10.1016/s0167-1987(03)00042-4, 2003.

Esri. World Imagery. Scale not given. September 27, 2016. https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9 (accessed 19-06-2019), 2019.

Geurts, P., Ernst, D. and Wehenkel, L. Extremely randomized trees. Mach. Learn. 63(1), 3-42. http://dx.doi.org/10.1007/s10994-006-6226-1, 2006.

Heckrath, G., Halekoh, U., Djurhuus, J. and Govers, G. The effect of tillage direction on soil redistribution by mouldboard ploughing on complex slopes. Soil Tillage Res. 88(1-2), 225-241. http://dx.doi.org/10.1016/j.still.2005.06.001, 2006.

Hengl, T., Heuvelink, G.B.M. and Stein, A. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120(1-2), 75-93. http://dx.doi.org/10.1016/j.geoderma.2003.08.018, 2004.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B. and Gräler, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518. http://dx.doi.org/10.7717/peerj.5518, 2018.

Hiemstra, P. automap: Automatic interpolation package. R package version 1.0-14. https://cran.r-project.org/web/packages/automap/index.html (accessed 15-08-19), 2013.

Knotters, M., Brus, D.J. and Oude Voshaar, J.H. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma 67(3-4), 227-246. http://dx.doi.org/10.1016/0016-7061(95)00011-c, 1995.

Kuhn, M. Building predictive models in R using the caret package. J. Stat. Softw. 28(5), 1-26. http://dx.doi.org/10.18637/jss.v028.i05, 2008.

Lindstrom, M.J., Nelson, W.W., Schumacher, T.E. and Lemme, G.D. Soil movement by tillage as affected by slope. Soil Tillage Res. 17(3-4), 255-264. http://dx.doi.org/10.1016/0167-1987(90)90040-k, 1990.

Mitchell, T. Decision tree learning, in: Machine Learning. McGraw Hill, New York, 52-80, 1997.

National Survey and Cadastre. Danmarks Højdemodel 2007, DHM-2007/Terræn. National Survey and Cadastre, 2012.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E. and Papritz, A. Evaluation of digital soil mapping approaches with large sets of environmental covariates. SOIL 4(1), 1-22. http://dx.doi.org/10.5194/soil-4-1-2018, 2018.

Odeh, I.O.A., McBratney, A.B. and Chittleborough, D.J. Further results on prediction of soil properties from terrain attributes: Heterotopic cokriging and regression-kriging. Geoderma 67(3-4), 215-226. http://dx.doi.org/10.1016/0016-7061(95)00007-b, 1995.

Pouladi, N., Møller, A.B., Tabatabai, S. and Greve, M.H. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. Geoderma 342, 85-92. http://dx.doi.org/10.1016/j.geoderma.2019.02.019, 2019.

Quinlan, J.R. Learning decision tree classifiers. ACM Comput. Surv. 28(1), 71-72. http://dx.doi.org/10.1145/234313.234346, 1996.

Rokach, L. and Maimon, O. Decision trees, in: Data Mining and Knowledge Discovery Handbook. Springer, 165-192, 2005.

Strobl, C., Malley, J. and Tutz, G. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Method 14(4), 323-348. http://dx.doi.org/10.1037/a0016973, 2009.

Tan, P.-N., Steinbach, M. and Kumar, V. Classification: Basic concepts, decision trees, and model evaluation, in: Introduction to Data Mining. Pearson Education, Limited, 2014, 145-205, 2014.

Wang, P.R. Referenceværdier: Døgn-, måneds- og årsværdier for regioner og hele landet 2001 - 2010, Danmark for temperatur, relativ luftfugtighed, vindhastighed, globalstråling og nedbør. Teknisk Rapport 12-24. Danish Meteorological Institute, 2013.

Wright, M.N. and Ziegler, A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77(1). http://dx.doi.org/10.18637/jss.v077.i01, 2015.