Biosystems Engineering, Volume 152, December 2016, Pages 104-116,
DOI:10.1016/j.biosystemseng.2016.04.018

# Machine Learning based Prediction of Soil Total Nitrogen, Organic Carbon and Moisture Content by Using VIS-NIR Spectroscopy

A. Morellos[1], X-E. Pantazi[1]*, D. Moshou[1], T. Alexandridis[3], R. Whetton[2],G. Tziotzios[1], J. Wiebesohn[4], R. Bill[4] and A. Mouazen[2]

[1]Agricultural Engineering Laboratory, Aristotle University of Thessaloniki, Faculty of Agriculture, Univ. Box. 275, Thessaloniki 54124, GreeceThessaloniki 54124, Greece, Tel: +302310998264, Fax: +302130998729, e-mail: renepantazi@gmail.com

[2]Cranfield Soil and AgriFood Institute, Cranfield University, Bedfordshire MK43 0AL, United Kingdom.

[3]Laboratory of Remote Sensing and GIS, Aristotle University of Thessaloniki, Faculty of Agriculture, Univ. Box. 259, Thessaloniki 54124, Greece

[4] Professorship for Geodesy and Geoinformatics, Faculty of Agricultural and Environmental Sciences, Rostock University

## Abstract

It is widely known that the visible and near infrared (VIS-NIR) spectroscopy has the potential of estimating soil total nitrogen (TN), organic carbon (OC) and moisture content (MC) due to the direct spectral responses these properties have in the NIR region. However, improving the predication accuracy requires advanced modelling techniques, particularly when measurement is planned for fresh (wet and un-processed) soil samples. The aim of this work is to compare the predictive performance of two linear multivariate and two machine learning methods for TN, OC and MC. The two multivariate methods investigated included Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR), whereas the machine learning methods included Least – Squares Support Vector Machines (LS-SVM), and Cubist. A mobile, fibre type, VIS-NIR spectrophotometer was utilised to collect soil spectra (305–2200 nm) in diffuse reflectance mode from 140 wet soil samples collected from one field in Germany. The results indicate that machine learning methods are capable of tackling non-linear problems in the dataset. LS-SVMs and the Cubist method over-performed the linear multivariate methods for the prediction of all three soil properties studied. LS-SVM provided the best prediction for MC (root mean square error of prediction (RMSEP) = 0.457 % and residual prediction deviation (RPD) = 2.24) and OC (RMSEP = 0.062 % and RPD = 2.20), whereas the Cubist method provided the best prediction for TN (RMSEP = 0.071 and RPD=1.96).

**Keywords:** VIS-NIR spectroscopy, data mining, chemometrics, soil properties

## 1. Introduction

Soil is a heterogeneous natural resource, the processes and mechanisms of which are complex and difficult to understand. Laboratory analysis has been the main key to better understand the soil system and to assess its quality and functions (Viscarra Rossel, Walvoort, McBratney, Janik and Skjemstad, 2006). Accurate information on soil at regional and national scale is essential, since it enables improved soil management according to land potential (Odeh & McBratney, 2000). Spatial assessment of soil properties allows researchers to understand the dynamics of ecosystems (Hively, McCarty, Reeves, Land, Oesterling and Delwiche, 2011). Understanding the soil properties and how these affect agriculture can lead to the implementation of sustainable agricultural and environmental management (Viscarra Rossel, Cattle, Ortega and Fouad, 2009). In precision agriculture the scale of soil information required for land and crop management is much smaller, and normally relay on proximal soil sensing (Kuang, Mahmood, Quraishi, Hoogmoed, Mouazen and van Henten, 2012) to allow collecting high sampling resolution data. However, the traditional laboratory methods for soil analysis are not able to fulfil the requirement of high sampling resolution, since they are tedious, time consuming, expensive and require expert laboratory operator.

One of the most common proximal soil sensing techniques is the visible (VIS) and near infrared (NIR) spectroscopy methods used to estimate soil properties and can be considered as a complimentary to chemical laboratory analysis methods. They are adopted for laboratory and field (both portable and on-line) measurements. Detailed information about accuracy and performance under different application conditions is provided in an intensive review from Kuang, Mahmood, Quraishi, Hoogmoed, Mouazen and van Henten (2012). There is an increasing interest in VIS-NIR analysis techniques as they are non-destructive, fast, cost-effective and more importantly allow for high sampling resolution (Viscarra Rossel and Hicks, 2015; Tekin, Kuang and Mouazen, 2013), which is particularly necessary to the implementation of variable rate farm inputs (e.g., fertilisers) in precision agriculture.

The soil mapping and classification has been historically performed through various methods, including statistical techniques such as principal components regression (PCR) (Chang, Laird, Mausbach and Hurburgh , 2001; Islam, Singh, McBratney, 2003; Mouazen, Kuang and De Baerdemaeker, 2010), partial least squares regression (PLSR) (McCarty, Reeves III, Reeves, Follett and Kimple, 2002; Mouazen, Kuang, De Baerdemaeker and Ramon, 2010) and also the use of machine learning techniques such as different types of artificial neural networks, decision trees and support vector machines (SVM) (Brown, Shepherd, Walsh, Mays, Reinisch, 2006; Vasques, Grunwald and Sickman , 2008; Mouazen, Kuang and De Baerdemaeker, 2010; Viscarra – Rossel and Behrens, 2010; Minasny, McBratney, Stockmann and Hong, 2013; Kuang, Tekin and Mouazen , 2015). Stevens, Nocita,Tóth, Montanarella and van Wesemael (2013), have used support vector machines and Cubist to predict Organic Carbon (OC). Cubist is able to make very efficient spectral variable selection and the rule structure is transparent to the user regarding the association of the spectra to soil properties allowing useful conclusions to be made about this relationship. However, these authors have used processed (dried, grinded and sieved) soil samples in their analysis. Since processed soil samples are of different physical conditions of fresh samples under field spectroscopy analyses, calibration models need to be developed with fresh samples. Then the performance of advanced data mining techniques needs to be evaluated for improved prediction capability of the VIS-NIR spectroscopy of studied soil properties for field spectroscopy application.

The aim of this paper is to compare the performance of four different regression methods for the prediction of TN, OC and MC in fresh (wet and unprocessed) soil samples by means of a portable VIS-NIR spectrophotometer which is designed for field applications. These include two linear multivariate methods (e.g., Principal Components Regression (PCR) and Partial Least Squares Regression (PLSR), and two machine learning (e.g., Least Squares – Support Vector Machines (LS-SVM) and the Cubist).

## 2. Materials and Methods

### 2.1 Soil Sampling and chemical analyses

A total of 140 soil samples were collected from the top soil layer (0-20 cm) of an arable field with an area of 31, 020 ha in Premslin, Germany (**Error! Reference source not found.**) during August 2013, after harvest of winter

wheat. The soil type according to the Food and Agriculture Organization (FAO) is a Luvisol. The soil samples were analysed in the soil laboratory of Cranfield University for TN, OC and MC. Soil OC and TN were measured by a TrusSpecCNS spectrometer (LECO Corporation, St. Joseph, MI, USA), using the Dumas combustion method. Soil MC was determined by oven drying of the soil samples at 105 ºC for 24 h.

(Figure1 here)

## 2.2 Optical soil measurements

The preparation of soil samples for optical measurements was carried out, as described by Kuang, Tekin and Mouazen (2015). Fresh (wet and non-processed) soil samples were put into glass containers and mixed well, after large stones and plant residue were removed (Mouazen, Karoui, Deckers and De Baerdemaeker, 2007). The optical measurements were taken from the smooth surface of soil samples, in order to achieve a higher signal to noise ratio (Mouazen, Karoui, Deckers and De Baerdemaeker, 2007). The soil samples were scanned by the AgroSpec portable VIS-NIR spectrophotometer (Tec5 Technology for Spectroscopy, Germany) that provides spectral measurements in the range between 305-2200 nm. A 100% white reference was measured before scanning, which was repeated every 30 min. The 100% white reference was made from lime material to ensure 100% of light is reflected back. A total of 10 scans were collected from each glass container and these were averaged in one spectrum. The spectra from 305-370 nm and from 2150-2200 nm at the fringe of the active range of the spectrophotometer showed an excessive noisy pattern and were removed from further analysis.

## 2.3 Multivariate Regression Models

### 2.3.1 Principal component regression (PCR) and partial least squares regression (PLSR)

Both PCR and PLSR are linear chemometrics tools used for analysis of spectroscopic data for different applications. They are extensively explained in the literature (e.g., Martens and Naes, 1989). They are the most common modelling techniques for quantitative spectroscopy analyses in soils (Kooistra, Wehrens, Leuven and Buydens, 2001; Viscarra–Rossel, 2008). They both represent techniques that are based on the decomposition of the spectral data into features (called principal components for PCR (PCs) and latent variables for PLSR (LVs)) that

represent most of the variance exist in the raw VIS-NIR data and the creation of linear models between the sample scores of the selected features of the most correlated factors.

In the current study, both PCR and PLSR models are calibrated using 1-20 PCs and LVs and the optimal number of features is selected according to the venetian blinds cross validation method for both techniques, so that to avoid over-fitting during the calibration (Reeves, Watson, Osborne, Pounds, O'Brien, Short and Schartel, 2002).

**2.3.2 Least Square – Support Vector Machines**

Least squares support vector machines (LS-SVM) is a method that was recently developed by Suykens, Van Gestel De Brabanter, De Moor, Vandewalle and Van Gestel (2002), as an easy but robust approach for the classification and regression analysis of linear and nonlinear multivariate problems, using linear equations set and not quadratic programming as in the classical SVM. It has been widely used during the last few years in the sector of chemometrics (Chauchard, Cogdill, Roussel, Roger and Bellon-Maurel, 2004; Thissen, Üstün, Melssen and Buydens, 2004; Borin, Ferrao, Mello, Maretto and Poppi, 2006; Sá, Ferrão, Galdos, Bittar and Poppi, 2010; Balabin & Lomakina, 2011 among others). Chemometrics applications such as in the soil spectroscopy are highly non-linear, especially for OC (Stenberg, Rossel, Mouazen and Wetterlind, 2010). For this reason, a normal SVM that is usually utilised for linear classification can result in poor prediction capability, hence, it needs to be expanded for nonlinear regression by using a kernel function (Vapnik, 1995 and 1998).

In the current study, a LS-SVM is used with the Gaussian radial basis function (RBF) kernel as a training algorithm (Eq. (1)). There have also been test runs with polynomial kernels, but they were omitted because the results were not satisfactory. The RBF kernel algorithm requires two parameters for tuning, namely, the gamma ($\gamma$), which is the regularization parameter that determines the trade-off between the training error minimization and smoothness (De Brabanter, Karsmakers, Ojeda, Alzade, De Bradanter, Pelckmans and Suykens, 2011) and the $\sigma^2$ (Eq. (1)), which is the squared bandwidth of the Gaussian curve. For the tuning of these parameters, leave-one-out cross validation is used for choosing the initial random parameters (Stone, 1974) to be optimised by means of performing the standard simplex method (Suykens, Van Gestel, De Brabanter, De Moor, Vandewalle and Van Gestel, 2002).

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right) \qquad \text{Eq. (1)}$$

The input parameters used for the training of the LS-SVM are the VIS-NIR features that will be derived from the LVs calculated from the PLS regression model. Mouazen, Kuang, De Baerdemaeker and Ramon (2010) adopted a similar approach, where the latent variables obtained from PLSR were used as input to a back propagation artificial neural network (BPNN), not to SVM as done in the current work.

#### 2.3.3 Cubist regression model

The Cubist model is a data mining technique that works in a similar way to that of the Decision Tree regression models. It is based on the M5 algorithm, developed by Quinlan (1992) and has been successfully used in VIS-NIR soil spectroscopy analyses, achieving very successful results with small errors, hence it is considered to be competitive with other methods of multivariate regression in terms of accuracy (Bui, Henderson and Viergever 2009; Viscarra Rossel and Webster, 2012; Stevens, Nocita, Tóth, Montanarella and van Wesemael , 2013; Lacoste, Minasny, McBratney, Michot, Viaud, & Walter ., 2014; Malone**,** Minasny, Odgers, and McBratney, 2014; Miller Koszinski, Wehrhan, and Sommer,2015; Viscarra Rossel and Hicks, 2015).

The Cubist model is based on the construction of an unconventional type of regression tree (Minasny & McBratney, 2008), where the prediction is based on the intermediate linear models at each step. It creates subsets of sample of the original data set that have similar attributes and creates multi-linear regression rules by selecting the optimum predictor variables to be used as regression variables among all of the spectral variables. These rules are connected to each other with an "**if** [condition is true], **then** [regression rule], and **else** [go to next rule]" condition sequence. If the tested sample falls into the restrictions of the first subset, it performs the regression rule that was chosen for that subset, or else it moves to the next rule as described by Viscarra Rossel and Webster (2012). The main advantages of the Cubist regression method is its ability to handle non-linear relationships between dependent and independent variables and the ability to use both discrete and continuous variables as inputs (Im, Jensen, Coleman and Nelson, 2009).

In the presented work, it is assumed that the Cubist model is capable of discovering spectra features that contribute highly to a specific soil property and it is able to construct a multivariate regression model to predict this property.

## 2.4 Spectral data pre-treatment

When the data pre-treatment techniques that are applied to the raw spectral data prior to the regression analysis are successfully implemented, it is possible that various problems such as noise, light scattering and external effects

can be reduced and this in turn can improve the accuracy of the models that will be created (Stevens, van Wesemael, Bartholomeus, Rosillon, Tychon, and Ben-Dor, 2008). In this study, the raw VIS-NIR reflected data (R) were first transformed into log(1/R), in order to reduce the nonlinearities that probably exist in the spectra (Viscarra Rossel, 2008). The transformed data were then pre-treated so that they became mean centred, with a standard deviation equal to 1 (autoscaling). Following this, soil spectra were subjected of the Savitzky – Golay (1964) filter for smoothing, using a first derivative transformation with 31 smoothing points. The first derivative transformation enhances small spectral absorptions and eliminates the background effect (Viscarra Rossel, Walvoort, McBratne, Janik and Skjemstad , 2006). Scatter removal from the transformed data was succeeded by implementing the standard normal variate technique (SNV) (Barnes, Dhanoa and Lister, 1989), which centres each spectrum by its mean and then scales it by its standard deviation in order to remove the path length variations.

The outliers were omitted from inclusion in further modelling steps. The outliers were decided after performing PCA and checking the Hotelling's $T^2$ test and taking 95% confidence intervals, as well as the Q-residuals, also in 95% confidence intervals that were derived by the PCA MANOVA (Constantinou, Papakonstantinou, Benaki, Spraul, Shulpis, Koupparis and Mikros, 2004).

## 2.5 Model evaluation

The accuracy of the models was assessed with the root mean squared error (RMSE) for the cross validation (RMSECV) and prediction (RMSEP), the coefficient of determination ($R^2$) and the residual prediction deviation (RPD). Before running the analysis the entire data set of 140 samples were randomly divided into training set (100 samples) and prediction set (40 samples). The data for the prediction set were selected using the Venetian blinds method, according to total acidity (Chauchard, Codill, Roussel, Roger and Bellon Maurel, 2004). The same testing sets were used for all the different models that were developed. The cross – validation methods that gave the best results was the Venetian blinds cross-validation for the PCR and PLSR methods and the leave-one-out cross-validation for the LS-SVM and Cubist methods. The RMSE, shown in (Eq. (2)), represents the mean absolute error of the time-series that was calculated by the model between the observed estimators and the measured values (Stone, 1993). The disadvantage of using the RMSE is that its value is sensitive to heavily weighted outliers and it can give a false estimation simply by taking some of the outliers into account (Bermejo and Cabestany, 2001).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_{obs,i} - x_{mod,i})^2}{n}} \qquad \text{Eq. (2)}$$

Where $x_{obs}$ is the observed value, $x_{mod}$ is the simulated value at observation *i* and *n* is the number of observations.

On the other hand, $R^2$ is the percentage of the total variation in the dependent values, which can be explained by the regression equation (Nagelkerke, 1991). $R^2$ ranges from 0 to 1 and the higher the value of $R^2$, the better fit the model is for the purpose. It is defined as follows (where $\overline{x}$ is the mean of the observed data):

$$R^2 = 1 - \frac{\sum_n(x_{obs,i} - x_{mod,i})^2}{\sum_n(x_{obs,i} - \overline{x})^2} \qquad \text{Eq. (3)}$$

RPD is also used for the evaluation of the models' accuracy and is the standard deviation divided by RMSEP as proposed by Saeys, Mouazen and Ramon (2005) and described by Eq. (4). According to this indicator, if RPD is below 1.5, the model performance is considered to be very poor and can't be used for prediction. If it is between 1.5 and 1.8, the model can give fair results, but it has a margin for improvement. For values of RPD in the range of 1.8 to 2 the prediction is considered to be good. Finally, if it is higher than 2.0, the model performance is considered to be very good.

$$\mathrm{RPD} = \frac{s_y}{\mathrm{RMSEP}} \qquad \text{Eq. (4)}$$

Where $s_y$ is the standard deviation of the observed values.

Different regression analyses were developed for each soil parameter. The LS-SVM analysis was carried out using the LS-SVMlab toolbox for Matlab (Mathworks, Natick, MA., USA), which was also used by Suykens, Van Gestel, De Brabanter, De Moor, Vandewalle and Van Gestel (2002). PLS regression and PCA have been performed using Unscrambler X10 (Camo Software, Oslo, Norway), whereas the Cubist analysis was performed using R toolbox (R-project, 2015) developed by Kuhn, Weston, Keefer, and Coulter (2011).

The spatial distribution of the models' performance was evaluated with residual maps of each parameter. The residuals of each model's prediction of the soil parameter on the 140 locations were interpolated spatially using radial basis function / completely regularised spline (RBF/CRS). RBF/CRS is an exact deterministic interpolator that works well for large number of data points with most soil parameters, such as organic matter (Robinson and

Metternicht 2006), MC (Western, Grayson, Blöschl, Willgoose and McMahon, 1999) and TN (Bruland, Grunwald, Osborne, Reddy and Newman, 2006).

# 3. Results and discussion

## 3.1. Overview of reference and optical measurement of soil samples

From the statistics of the values of the soil properties that resulted from laboratory analyses it was concluded that there were no outliers to be excluded from further analysis, as the median and mean are only slightly different to each other and there are no extreme values in the minimum and maximum values (**Error! Reference source not found.**

(Table 1 here)

Although the soil variables had quite a wide range of values, the reflectance spectra for the different samples had a similar pattern without significant deviations, due to the fact that all samples belong to one soil type (e.g. luvisols soil) and may also indicate small variability in different soil properties (Figure 2).

(Figure 2 here)

## 3.2 Prediction performance of principal component regression (PCR) and partial least squares regression (PLSR) models

The results of the regression (cross-validation) with PCR and PLSR and the performance of the models during the validation phase (prediction) are shown in **Error! Reference source not found.** and **Error! Reference source not found.**. From the results one can easily observe that PLSR outperforms the PCR method for the prediction of the TN and MC, whereas PCR performs slightly better than PLS for the OC. The PLSR models can be considered to have a good prediction ability, as far as the prediction of TN and OC is concerned, as the RPD values are >1.8, while for the MC the prediction is considered to be very good, as the RPD = 2.17. A similar prediction performance

can be observed with the PCR models, where prediction results were better for MC (RPD = 2.19, very good prediction) as compared to those for TN (RPD = 1.72, fair prediction) and OC (RPD = 1.90, good prediction) (Saeys, Mouazen and Ramon, 2005). This can be also seen in Figure 3, which clearly shows that for TN and MC, only a few readings are highly diverging from the 1:1 line and are responsible for the lower model performance. PCR was expected to demonstrate lower performance than the PLSR model for the prediction of TN and OC, according to the independent literature reports by Viscarra Rossel (2006) and Mouazen, Kuang, De Baerdemaeker and Ramon (2010).

(Table 2 here)

(Figure 3 here)

## 3.3 Prediction performance of least – squares support vector machines (LS-SVMs) models

During the development of the LS-SVM models, both scores of PLSR and PCR were tested as input features. However, the PCs obtained from the PCR did not give satisfactory results as compared to the LVs, derived by PLSR, which is in line with findings of similar research performed by Mouazen, Kuang, De Baerdemaeker and Ramon (2010).In the latter work researchers compared the performance of artificial neural networks using PCs and LVs as input variables for the network. For this reason, only the LVs were used for the calibration and validation of the LS-SVM. The $\gamma$ and $\sigma^2$ values used for the calibration of the LS-SVM in the current work are shown in **Error! Reference source not found.**.

(Table 3 here)

From **Error! Reference source not found.** it becomes obvious that LS-SVM has a very good prediction performance for all three studied soil properties, as demonstrated by the low RMSE, high $R^2$ and RPD values that varies from 1.90 to 2.24. **Error! Reference source not found.** depicts the scatter plots of the LS-SVM model prediction performance with the prediction sample set, showing only few points that significantly deviate from the 1:1 line for MC (**Error! Reference source not found.**B) and OC (**Error! Reference source not found.**C).

For the training of the LS-SVM for the TN, 10 LVs were used as they showed to have the best performance during the training and the prediction phase, whereas for the prediction of MC and OC, 2 and 3 LVs, respectively, were sufficient to obtain the best model performance for both properties. It is worth confirming that both OC and MC

have direct spectral responses in the NIR range, while it was not clearly confirmed this to be the case for TN (Stenberg Rossel, Mouazen and Wetterlind, 2010). This may explain why a significantly smaller number of LVs was needed for OC and MC, as compared to TN.

(Table 4 here)

(Figure 4 here)

## 3.4 Prediction performance of Cubist model

The outcome of the simulations for the Cubist model is shown in **Error! Reference source not found.** in the cross-validation and prediction phases. A closer look to this table shows that Cubist model has a very high accuracy both in terms of cross-validation and prediction. This can be also confirmed by the low RMSE and high RPD values (higher than or at least very close to 2) in the prediction set, which confirm these predictions can be used with good model performance (Saeys, Mouazen and Ramon, 2005). **Error! Reference source not found.** depicts the scatter plots of measured versus predicted soil properties using the prediction samples sets.

(Table 5 here)

Only one rule with one variable was enough for the Cubist method to find the best fit for the TN data. The spectral wavelength that was chosen by the Cubist model to predict the TN was 475 nm, which is one of the band listed in literature as sensitive to TN content (Viscarra Rossel, Walvoort, McBratney, Janik and Skjemstad, 2006; Viscarra Rossel and Webster, 2012). This wavelength variable may also associate with the blue absorption band that is reported to be around 450 nm (Mouazen, Karoui, Deckers, De Baerdemaeker and Ramon, 2007). For the prediction of MC, two rules have been created that are associated with 1862 nm wavelength. However, more spectral variables (wavelengths) were highlighted to provide good prediction of MC including 616, 684, 823, 1402, 1715, 1715, 1862, 1864 and 1867nm, with 1867nm being the most important spectral variable for model fitting, which is also in-line with findings of other researchers (Bowers and Hanks, 1969; Lobel and Asner, 2002, Mouazen, Karoui, De Baerdemaeker and Ramon, 2006; Viscarra Rossel and Webster, 2012). Wavebands at 1402 nm and 1860 nm can be attributed to the second and first overtones on O-H absorption in the NIR range (Tekin, Tumsavas, Mouazen and (2012). Both 616 and 684 may be attributed to absorption of red colour around 680 nm (Viscarra Rossel, Walvoort, McBratney, Janik and Skjemstad, 2006; Mouazen, Karoui, De Baerdemaeker and Ramon, 2007).

Finally, for the OC there was again one rule for the best fitting of the data, which was based on two wavelength regression variables of 476 and 808 nm. While the former wavelength can be linked with the blues colour absorption band at 450 nm (Mouazen, Karoui, Deckers, De Baerdemaeker and Ramon, 2007), the latter waveband can be attributed to aromatics (C-H) band at 825 nm according to Viscarra Rossel and Behrens (2010). The significant wavelengths identified for the investigated properties are shown in Table 6.

(Table 6 here)

(Figure 5 here)

## 3.5. Models Comparison

From the results and based on RPD values, it can be concluded that all four calibration methods provided prediction results ranging from 'fair' to 'very good', according to the evaluation criteria of Saeys, Mouazen and Ramon (2005). The LS-SVM method though, shows the best performance, in comparison with the rest of the models for the prediction of MC and OC, whereas the Cubist model has slightly over-performed LS-SVM for TN. Despite this, the RMSE (the most important parameter to evaluate accuracy) values for both models were as the same. For MC prediction, although the Cubist model has a higher fitting $R^2$ than LS-SVM, it also has lower RMSE and RPD values. For this reason, LS-SVM is assumed to have better prediction performance ability than the Cubist model.

The spatial distribution of the models' performance is displayed in Figures 6 (parameters) and 7 (residuals). The spatial distribution of the parameters shows higher values along the north-south central axis of the field, which is consistent across the models tested (Fig. 6). In red colour are displayed areas where the model has overestimated the respective parameter, while in green are displayed the areas with underestimations (Fig. 7). All models tested show a similar spatial pattern of residuals across the three spatial parameters. In specific, PLSR shows an area of overestimation of TN in the north end of the field and another in the southwest, while LS-SVM and Cubist show an additional area in the eastern and central parts of the field. For MC, all models show an area of overestimation in the northern and central parts of the field and an area of underestimation exactly to the south of the latter, which is more pronounced for Cubist. For OC, all models show an overestimation in the central part of the field, with an additional area in the north for Cubist. PLSR results in a large area of underestimation in the western side of the field.

(Figures 6 and 7 here)

The superior performance of the LS-SVM, in comparison with the rest of the models tested can be explained by its high ability to deal with the nonlinear pattern (Chauchard, Cogdill, Roussel, Roger, and Bellon-Maurel, 2004), which is reported during modelling the VIS-NIR spectra for soil properties, particularly for OC (Stenberg, Rossel, Mouazen and Wetterlind , 2010). Cubist models over-perform both PCR and PLSR models, for the prediction of all three soil attributes. Stevens, Nocita, Tóth, Montanarella and van Wesemael (2013) also found that a classic SVM performs competitively compared to the Cubist model for the prediction of OC. Viscarra Rossel and Behrens (2010) and Mouazen, Kuang, De Baerdemaeker and Ramon (2010) found machine learning techniques to provide better results than PLSR and PCR in soil spectroscopy. The results found in the current work confirm this finding but for fresh (wet and unprocessed) soil samples, while findings of previous studies were based on processed (dried, crushed and sieved) soil samples.

3.6 Research limitations

Although from the comparison of the different models it becomes obvious that machine learning techniques lead to very satisfactory results for the prediction of the formentioned soil parameters (TM, MC and OC), there is a limitation concerning the generalisation of the results of this study.

The models were calibrated and tested according to the samples collected from one field of one soil type in a specific region in Germany. That means that the specific models can't be generalised for the prediction of the same parameters in any soil type given. This happens because for the machine learning techniques, the generalisation of the models is strongly dependent on the calibration and testing samples variability. Therefore, there is a need to use samples from a big diversity of soil types and from different fields, in order for these models to be generalised to a larger scales e.g., regional, country, continental of global.

From the models developed, it is obvious that such techniques as Cubist and LS-SVM can have very good predictive ability for TN, MC and OC and for this reason it can be assumed that those methods could also give comparably good results for different soil types. Further calibration-validation procedure is needed for other soil types to validate this assumption..

## 4. Conclusions

The comparison between principal component regression (PCR), partial leas squares regression (PLSR), least-squares support vector machine (LS-SVM) and Cubist methods for the prediction of total nitrogen (TN), moisture content (MC) and organic carbon (OC) with the visible and near infrared spectroscopy based on 140 fresh soil samples collected from one field in Germany have led to the following conclusions:

- For the given dataset, LS-SVM outperformed all the other methods for the prediction of the MC and OC, but TN was best predicted by the Cubist method.
- Machine learning techniques such as Cubist and LS-SVM showed a better explanatory power than the classic multivariate regression techniques such as PCR and PLSR. For this reason they are recommended for soil spectroscopy analyses.
- Although, for the MC, the Cubist showed the best model fitting, it had higher RMSE and lower RPD than LS-SVM and for this reason it was chosen as the best performing model.
- The machine learning techniques investigated in this study can be used in field spectroscopy for off-line and on-line prediction of the soil parameters studied in fields with similar soil type and variability.

**Acknowledgements**

The research presented was carried out in the framework of project FARMFUSE of ICT AGRI 2 ERANET, funded through GSRT (Greece), DEFRA (UK) and (Germany).

## References

Balabin, R. M., & Lomakina, E. I. (2011). Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. *Analyst*, *136*(8), 1703-1712.

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra.*Applied spectroscopy*, *43*(5), 772-777.

Bermejo, S., & Cabestany, J. (2001). Oriented principal component analysis for large margin classifiers. *Neural Networks*, *14*(10), 1447-1461.

Borin, A., Ferrao, M. F., Mello, C., Maretto, D. A., & Poppi, R. J. (2006). Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. *Analytica Chimica Acta*, *579*(1), 25-32.

Bowers, S. A., & Hanks, R. J. (1965). Reflection of radiant energy from soils.*Soil Science*, *100*(2), 130-138.

Brown, D. J., Shepherd, K. D., Walsh, M. G., Mays, M. D., & Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, *132*(3), 273-290.

Bruland, G. L., Grunwald, S., Osborne, T. Z., Reddy, K. R., & Newman, S. (2006). Spatial distribution of soil properties in Water Conservation Area 3 of the Everglades. *Soil Science Society of America Journal*, 70(5), 1662-167.

Bui, E., Henderson, B., Viergever, K., 2009. Using knowledge discovery with data miningfrom the Australian Soil Resource Information System database to inform soil carbonmapping in Australia. Global Biogeochem. Cycles 23 6.

Chang, C. W., & Laird, D. A. (2002). Near-infrared reflectance spectroscopic analysis of soil C and N. *Soil Science*, *167*(2), 110-116.

Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh Jr., C.R., (2001). Near-infrared reflectance spectroscopy—principal components regression analysis of soil properties. Soil Science Society of America Journal 65, 480 – 490.

Chauchard, F., Cogdill, R., Roussel, S., Roger, J. M., & Bellon-Maurel, V. (2004). Application of LS-SVM to non-linear phenomena in NIR spectroscopy: development of a robust and portable sensor for acidity prediction in grapes.*Chemometrics and Intelligent Laboratory Systems*, *71*(2), 141-150.

Constantinou, M.A., Papakonstantinou, E., Benaki, D., Spraul, M., Shulpis, K., Koupparis, M.A., Mikros, A. (2004). Application of nuclear magnetic resonance spectroscopy combined with principal component analysis

in detecting inborn errors of metabolism using blood spots: a metabonomic approach. *Analytica Chimica Acta 511* (2), 303–312.

De Brabanter, K., Karsmakers, P., Ojeda, F., Alzate, C., De Brabanter, J., Pelckmans, K. & Suykens, J. A. K. (2011). LS-SVMlab toolbox user's guide. *ESAT-SISTA Technical Report*, 10-146.

Hively W.D., McCarty G.W., Reeves J.B., Land M.W., Oesterling R.A., Delwiche S.R. (2011). Use of Airborne Hyperspectral Imagery to Map Soil Properties in Tilled Agricultural Fields. *Applied and Environmental Soil Science*.

Islam, K., Singh, B., McBratney, A.B., 2003. Simultaneous estimation of various soil properties by ultra-violet, visible and near-infrared reflectance spectroscopy. Australian Journal of Soil Research 41, 1101 – 1114.

Im J., Jensen J.R., Coleman M., Nelson E. (2009). Hyperspectral remote sensing analysis of short rotation woody crops grown with controlled nutrient and irrigation treatments. Geocarto International 24, No.4, 293-312.

Johnson, D. M. (2014). An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Environment*, *141*, 116-128.

Kooistra, L., Wehrens, R., Leuven, R. S. E. W., & Buydens, L. M. C. (2001). Possibilities of visible–near-infrared spectroscopy for the assessment of soil contamination in river floodplains. *Analytica Chimica Acta*, *446*(1), 97-105.

Kuang, B., Mahmood, H. S., Quraishi, M. Z., Hoogmoed, W. B., Mouazen, A. M., & van Henten, E. J. (2012). 4 Sensing Soil Properties in the Laboratory, In Situ, and On-Line: A Review. *Advances in Agronomy*, *114*(1), 155-223.

Kuang, B., Tekin, Y., & Mouazen, A. M. (2015). Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. *Soil and Tillage Research*, *146*, 243-252.

Kuhn, M., Weston, S., Keefer, C., & Coulter, N. (2012). Cubist Models For Regression. URL: http://cran.icesi.edu.co/CRAN/web/packages/Cubist /vignettes/ cubist.pdf.

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., & Walter, C. (2014). High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. *Geoderma*, *213*, 296-311.

Lobell, D. B., & Asner, G. P. (2002). Moisture effects on soil reflectance. *Soil Science Society of America Journal*, *66*(3), 722-727.

Malone, B. P., Minasny, B., Odgers, N. P., & McBratney, A. B. (2014). Using model averaging to combine soil property rasters from legacy soil maps and from point data. *Geoderma*, *232*, 34-44.

Martens, H., and T. Naes. 1989. Multivariate calibration. 2nd ed. John Wiley & Sons, Chichester, UK.

McCarty, G.W., Reeves III, J.B., Reeves, V.B., Follett, R.F., Kimble, J.M. (2002). Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurements. Soil Science Society of America Journal 66, 640 – 646.

Miller, B. A., Koszinski, S., Wehrhan, M., & Sommer, M. (2015). Comparison of spatial association approaches for landscape mapping of soil organic carbon stocks. *SOIL*, *1*(1), 217-233.

Minasny B., McBratney A.B., Stockmann U., Hong S.Y. (2013). *Cubist a regression rule approach for use in calibration of NIR spectra.* 16th International Conference on Near Infrared Spectroscopy.

Mouazen, A.M.; Karoui, R.; De Baerdemaeker, J.; Ramon, H., (2006). Characterization of soil water content using measured visible and near infrared spectra. Soil Science Society of America Journal, 70, 1295-1302.

Mouazen, A.M., (2006). Soil Survey Device. International publication published under the patent cooperation treaty (PCT). World Intellectual Property Organization, International Bureau. International Publication Number: WO2006/015463; PCT/BE2005/000129; IPC: G01N21/00; G01N21/00.

Mouazen, A.M.; Karoui, R.; Deckers, S.; De Baerdemaeker, J.; Ramon, H., (2007). Potential of visible and near infrared spectroscopy to derive colour groups utilising the Munsell soil colour charts. Biosystems Engineering, 97(2): 131-143.

Mouazen, A. M., Maleki, M. R., De Baerdemaeker, J., & Ramon, H. (2007). On-line measurement of some selected soil properties using a VIS–NIR sensor. *Soil and Tillage Research*, *93*(1), 13-27.

Mouazen, A. M., Kuang, B., De Baerdemaeker, J., & Ramon, H. (2010). Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma*, *158*(1), 23-31.

Nagelkerke, N.J. (1991). A note on a general definition of the coefficient of determination. *Biometrica*, 78(3): 691-692.

Odeh I.O.A., McBratney A.B. (2000). Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia. Geoderma 97, 237-254.

Quinlan, J. R. (1992, November). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence* (Vol. 92, pp. 343-348).

Reeves III, J.B., McCarty, G.W. (2001). Quantitative analysis of agricultural soils using near infrared reflectance spectroscopy and fibre-optic probe. Journal of Near Infrared Spectroscopy 9, 25 – 34.

Reeves, J. N., Watson, D., Osborne, J. P., Pounds, K. A., O'Brien, P. T., Short, A. D. T., & Schartel, N. (2002). The signature of supernova ejecta in the X-ray afterglow of the γ-ray burst 011211. *Nature*, *416*(6880), 512-515.

Robinson, T. P., & Metternicht, G. (2006). Testing the performance of spatial interpolation techniques for mapping soil properties. *Computers and electronics in agriculture*, 50(2), 97-108.

SáA, S. O., FerrãoB, M. F., GaldosC, M. V., BittarD, C. M. M., & PoppiE, R. J. (2010). Application of LS-SVM-NIR spectroscopy for carbon and nitrogen prediction in soils under sugarcane. In *Proceedings of the 19th World Congress of Soil Science: Soil solutions for a changing world, Brisbane, Australia, 1-6 August 2010. Working Group 1.5 Soil sense: rapid soil measurements* (pp. 17-20). International Union of Soil Sciences (IUSS), c/o Institut für Bodenforschung, Universität für Bodenkultur.'

Saeys, W., Mouazen, A.M., Ramon, H. (2005). Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy. *Biosystems Engineering*, *91*(4), 393-402.

Savitzky, A., & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, *36*(8), 1627-1639.

Sinnaeve, G., Herman, J. L., Baeten, V., Dardenne, P., & Frankinet, M. (2001). Performances of an on board diode array NIR instrument for the analysis of fresh grass. *Journée Thématique AFMEX, Appareils Embarqués de Mesure de la Biomasse. November, Rennes, France*.

Stenberg, B., Rossel, R. A. V., Mouazen, A. M., & Wetterlind, J. (2010). Chapter five-visible and near infrared spectroscopy in soil science. *Advances in agronomy*, *107*, 163-215.

Stevens, A., van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., & Ben-Dor, E. (2008). Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. *Geoderma*, *144*(1), 395-404.

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., & Van Wesemael, B. (2010). Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma*, *158*(1), 32-45.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., & van Wesemael, B. (2013). Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. *PloS one*, *8*(6), e66409.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 111-147.

Stone, R. (1993). Improved statistical procedure for the evaluation of solar radiation estimation models. *Solar Energy* 51 (4): 289-291.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J., & Van Gestel, T. (2002). *Least squares support vector machines* (Vol. 4). Singapore: World Scientific.

Tekin, Y.; Tumsavas, Z.; Mouazen, A.M., (2012). Effect of moisture content on prediction of organic carbon and pH using visible and near infrared spectroscopy. Soil Science Society of America Journal, 76(1): 188-198.

Tekin, Y., Kuang, B., & Mouazen, A. M. (2013). Potential of on-line visible and near infrared spectroscopy for measurement of pH for deriving variable rate lime recommendations. *Sensors*, *13*(8), 10177-10190.

Tekin, Y., Tümsavas, Z., & Mouazen, A. M. (2014). Comparing the Artificial Neural Network with Parcial Least Squares for prediction of Soil Organic Carbon and pH at different moisture content levels using Visible and Near Infrared spectroscopy. *Revista Brasileira de Ciência do Solo*, *38*(6), 1794-1804.

Thissen, U., Üstün, B., Melssen, W. J., & Buydens, L. M. (2004). Multivariate calibration with least-squares support vector machines. *Analytical Chemistry*,*76*(11), 3099-3105.

Vapnik, V. (1995). The nature of s*tatistical learning theory*. New York: Springer-Verlag.

Vapnik, V. (1998). *Statistical learning theory* (Vol. 1). New York: Wiley.

Vasques, G. M., Grunwald, S. J. O. S., & Sickman, J. O. (2008). Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. *Geoderma*, *146*(1), 14-25.

Viscarra – Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J., & Skjemstad, J. O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties.*Geoderma*, *131*(1), 59-75.

Viscarra – Rossel, R. A. (2008). ParLeS: Software for chemometric analysis of spectroscopic data. *Chemometrics and intelligent laboratory systems*, *90*(1), 72-83.

Viscarra – Rossel, R. A., Cattle, S.R., Ortega, A., Fouad, Y. (2009). In situ measurements of soil colour, mineral composition and clay content by visNIR spectroscopy. Geoderma 150, 253-266.

Viscarra Rossel, R. A., and Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*. doi: 10.1016/ j.geoderma.2009.12.025.

Viscarra – Rossel, R. A., & Webster, R. (2012). Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. *European Journal of Soil Science*, *63*(6), 848-860.

Viscarra Rossel, R. A., & Hicks, W. S. (2015). Soil organic carbon and its fractions estimated by visible–near infrared transfer functions. *European Journal of Soil Science*, *66*(3), 438-450.

Western, A. W., Grayson, R. B., Blöschl, G., Willgoose, G. R., & McMahon, T. A. (1999). Observed spatial organization of soil moisture and its relation to terrain indices. *Water resources research*, 35(3), 797-810.

Wynn, J. G., Bird, M. I., Vellen, L., Grand-Clement, E., Carter, J., and Berry, S. L. (2006). Continental-scale measurement of the soil organic carbon pool with climatic, edaphic, and biotic controls, *Global Biogeochem. Cycles*, 20, GB1007,