Predicting soil organic carbon content in Cyprus using remote sensing and Earth observation data

Cristiano Ballabio[*a], Panos Panagos [a] and Luca Montanarella[a]
[a] European Commission, Joint Research Centre, Institute for Environment and Sustainability
Via E. Fermi 2749, I-21027 Ispra (VA), Italy

*cristiano.ballabio@jrc.ec.europa.eu; phone +39 0332 783021; fax +39 0332 783694; http://eusoils.jrc.ec.europa.eu

ABSTRACT

The LUCAS (Land Use/Cover Area frame Statistical Survey) database currently contains about 20,000 topsoil samples of 15 soil properties. It is the largest harmonised soil survey field database currently available for Europe. Soil Organic Carbon (SOC) levels have been successfully determined using both proximal and airborne/spaceborne reflectance spectroscopy. In this paper, Cyprus was selected as a study area for estimating SOC content from multispectral remotely sensed data.

The estimation of SOC was derived by comparing field measurements with a set of spatially exhaustive covariates, including DEM-derived terrain features, MODIS Vegetation indices (16 days) and Landsat ETM+ data. In particular, the SOC levels in the LUCAS database were compared with the covariate values in the collocated pixels and their eight surrounding neighbours. The regression model adopted made use of Support Vector Machines (SVM) regression analysis. The SVM regression proved to be very efficient in mapping SOC with an R^2 fitting of 0.81 and an R^2 k-fold cross-validation of 0.68. This study proves that the inference of SOC levels is possible at regional or continental scales using available remote sensing and Earth observation data.

Keywords: Soil Organic Carbon, LUCAS, Landsat ETM+, Cyprus, MODIS, Vector Regression, Support Vector Machines

1. INTRODUCTION

The decline in Soil Organic Carbon (SOC) is recognised as being one of the eight soil threats identified in the European Union's Thematic Strategy for Soil Protection (COM(2006) 231 final)[1]. One of the key goals of the strategy is to maintain and enhance SOC levels. A recent policy document known as the Roadmap to a Resource Efficient Europe (COM(2011) 571 final)[2], which is one of the building blocks of the Europe 2020 strategy, sets the objective of increasing levels of SOC in areas where less than 2% of SOC has been detected, by 2020.

Most of the food, fuel and fibre used by humans is produced on soil[3]. Soil resources are critical for future food security and conservation of biodiversity, as well as to secure other important ecosystem services such as carbon sequestration, water-retention capacity, and flood prevention. Soil functions are crucial both for society and the environment, and should be maintained. Nevertheless, soil resources in many parts of Europe are being overexploited, degraded and irreversibly lost due to inappropriate land management practices, industrial activities and land-use changes that lead to soil sealing, contamination, erosion, and loss of organic carbon. The European Commission recognises the value of soils, and has proposed a directive for the protection of soils[1].

Detailed spatial coverage of soil organic carbon is needed to guide sustainable management decisions on land use practices. The current pan-European estimates of soil organic carbon at 0-30cm are based on the Organic Carbon in Topsoils (OCTOP) model[4]. To generate soil organic carbon estimates, the OCTOP model takes into account soil type, land use/land cover and climatic conditions. In 2009, the European Soil Data Centre (ESDAC) conducted a SOC data collection exercise through the European Environment Information and Observation Network for soil (EIONET-SOIL)[5]. Cyprus was not among the countries that provided data in the EIONET SOC data collection[5] exercise, and is therefore not included in the modelled OCTOP dataset.

Second International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2014),
edited by Diofantos G. Hadjimitsis, Kyriacos Themistocleous, Silas Michaelides, Giorgos Papadavid,
Proc. of SPIE Vol. 9229, 92290F · © 2014 SPIE · CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2066406

In this paper, we present a novel methodology to deploy spatial coverage of SOC based on a ground observation data and remote sensing datasets using Support Vector Regression as an inference model. The objectives of this paper are to demonstrate how the LUCAS soil database allows us to:

- Develop a country-level Soil Organic Carbon (SOC) dataset with a minimum uncertainty;
- Downscale the SOC distribution model from larger to regional scales;
- Predict SOC content where data is scarce.

2. MATERIALS AND METHODS

2.1 Study area

Cyprus is an island with an area of 9,251 km^2, located in the far eastern corner of the Mediterranean basin. Its topography includes two mountain ranges, one in the north and the other in the south west part. Between them lies a plain. The cultivated area is estimated at 1,340 km^2. Cyprus is the third largest island in the Mediterranean Sea, after the Italian islands of Sicily and Sardinia, both in terms of area and population. It measures 240 km in length and 100 km in width at its widest point. It lies between latitudes 34° and 36° N, and longitudes 32° and 35° E.

The physical relief of the island is dominated by two mountain ranges, the Troodos Mountains and the smaller Kyrenia Range, and the central plain they encompass, the Mesaoria. The Mesaoria plain is drained by the Pedieos River, the longest on the island. The Troodos Mountains cover most of the southern and western portions of the island and account for roughly half its area. The highest point on Cyprus is Mount Olympus at 1,952 m, located in the centre of the Troodos range. The narrow Kyrenia Range, extending along the northern coastline, occupies substantially less area, and elevations are lower, reaching a maximum of 1,024 m.

Cyprus has the warmest climate (and warmest winters) in the Mediterranean part of the European Union. The average annual temperature on the coast is around 24°C during the day and 14°C at night. Its Mediterranean climate is characterised by a rainy season (November to mid-March) and a longer dry season (mid-March to October)[6]. The average annual precipitation is 363 mm yr^{-1} [7], with a great amount of geospatial variability as shown in the historical data from a number of recording stations[8].

The dominant soil types in Cyprus are Regosols, Leptisols, Cambisols, Luvisols, Vertisols, Solonchaks and Gypsisols[9]. The soils of Cyprus are degraded due to high erosion risk, landslides, increased levels of soil sealing in the coastal areas, and low organic carbon levels. The geology of Cyprus is dominated by the Troodos Ophiolite, which is a fragment of a fully developed oceanic crust, consisting of plutonic, intrusive and volcanic rocks and chemical sediments. Sedimentary formations cover the coastal plains in the south and the intermountain plain in the north[10].

Regarding land use/land cover, 7.76% of Cyprus is artificial land, 41.45% cropland, 3.7% grassland, 38.79% forests, and 8.3% plantations[11]. Soil sealing is an alarming problem in Cyprus, responsible for a loss of 0.84% of agricultural land in the period 2000-2006.

2.2 Land Use/Cover Area frame Statistical Survey (LUCAS) database

LUCAS (Land Use/Cover Area frame Statistical Survey) is an in-situ survey, which means that the data are gathered through direct field observations. The aim of the LUCAS survey is to gather fully harmonised data on land use/cover in the European Union Member States and their changes over time. In the LUCAS (2009) survey, 265,000 geo-referenced points were visited by more than 500 field surveyors. The survey points were selected from a standard 2 km x 2 km grid based on stratification information provided by Eurostat[12].

The LUCAS (2009) survey was the first to include a soil module. Topsoil samples (0-30 cm) were collected from 10% of the survey points, thus providing approximately 20,000 soil samples. While LUCAS soil samples were taken from all land use/land cover types, the survey focused mainly on agricultural areas. Each soil sample was taken from the topsoil zone (top 30 cm) with a weight of circa 0.5 kg. The objective of the soil module was to improve the availability of harmonised data on soil parameters in Europe. The 20,000 LUCAS soil samples were analysed in a single ISO-certified laboratory that used harmonised chemical and physical analytical methods (compliant with ISO standards, or their equivalent) in order to obtain a coherent and harmonised dataset with pan-European coverage. The analysis results

formed the LUCAS soil database, including, inter alia, SOC in topsoils (0-30cm), expressed in g kg^{-1}. The LUCAS points were stratified in order to be representative of individual land uses. Points above 1,000 m in altitude were excluded, with few exceptions.

In 2009, the LUCAS soil survey was carried out in the 23 EU Member States that had already taken part in the 2006 LUCAS Land Use/Cover survey, plus Cyprus and Malta. The latter two countries participated in the 2009 LUCAS soil survey using their own surveyors, but following the general sampling guidelines. In 2012, the LUCAS soil survey was carried out in Bulgaria and Romania.

The average density of LUCAS soil points is circa 1 sample every 199 km^2, which corresponds approximately to a grid cell of 14 km x 14 km[13]. The Cypriot surveyors collected 90 soil samples from southern Greek Cypriot due partially to accessibility issues. The density of points is much higher (1 sample every 103 km^2) than the rest of LUCAS soil survey. Most of the points were sampled in areas with elevations of 0-99 m and 200-399 m (Fig. 1). Regarding the slopes, most of the points were sampled in slopes between 6-12 degrees (Fig. 1). Only 11 points have been sampled in flat areas (0-2 degrees).

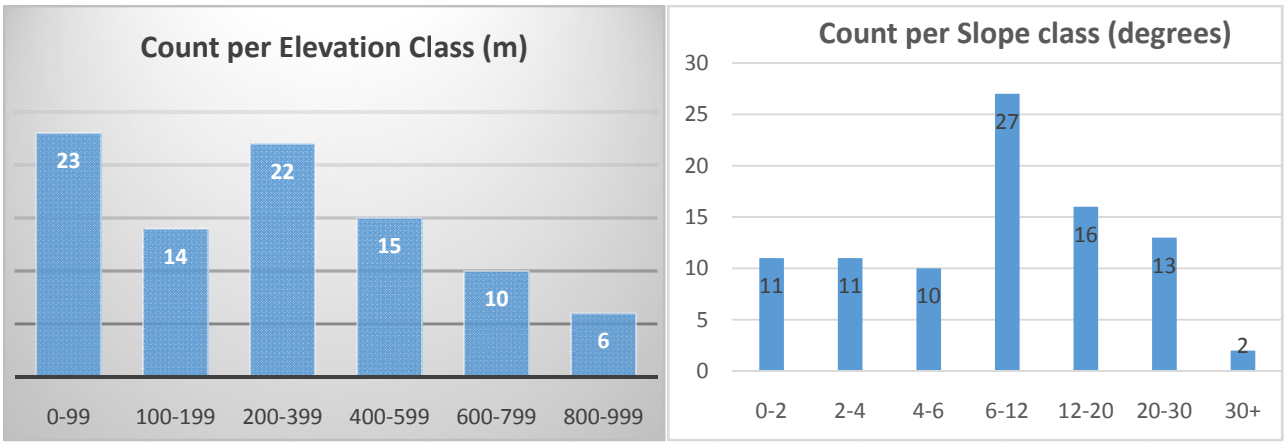

Figure 1: Number of LUCAS points in Cyprus according to elevation (metres) and slope (degrees)

The soil geochemical atlas of Cyprus is a recent addition to the series of national to continental-scale geochemical mapping programmes implemented over the past two decades for environmental and resource applications. The soil geochemical atlas of Cyprus was compiled at the high sampling density of 1 site per 1 km^2, with multi-element and multi-method analysis performed on samples of topsoil (0–25 cm) and subsoil (50–75 cm), from a grid of over 5,350 sites across a major portion of Cyprus[14]. However, there were not available enough soil organic carbon data in order to be used in the current study.

2.3 Landsat

Landsat remote sensing images facilitate the extraction of vegetation indexes such as the Normalized Difference Vegetation Index (NDVI), which is an important attribute for the correlation of Soil Organic Carbon (SOC) distribution. Since possible land-use/land cover changes may have occurred between 2003 (Landsat images) and 2009 (LUCAS survey), the correspondence between point measures, Landsat pixels and the actual state of the land might be lost. To reduce the occurrence of possible mismatches, neighbouring pixels were also considered as candidates for the regression procedure. Moreover, neighbouring pixel values provide some form of additional information about the variation of reflectance over larger areas, and provide information about reflectance patterns. As different kinds of land cover result in highly spatially correlated patterns[15], characterised by structured and defined shapes, their identification may improve the accuracy of the prediction of soil properties and, in particular, SOC.

2.4 Aster DEM

The Global Digital Elevation Model (GDEM) is a relatively high resolution (30-m) Digital Elevation Model (DEM) (Fig. 2) obtained from the ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) satellite images. The GDEM covers the planet from 83 degrees North to 83 degrees South, and is created from the stereoscopic correlation of visible and near-infrared images taken globally by ASTER. In spite of its relatively high resolution, the quality of the GDEM is rather low due to artifacts. In order to reduce the effect of artifacts on the derivation of covariates, the DEM was filtered using wavelet thresholds to remove most of the artifacts while preserving the general shape of land features. Elevation, slope, length of slope, Topographic Wetness Index and solar radiation are covariates derived from GDEM which have an important role in the proposed regression analysis.

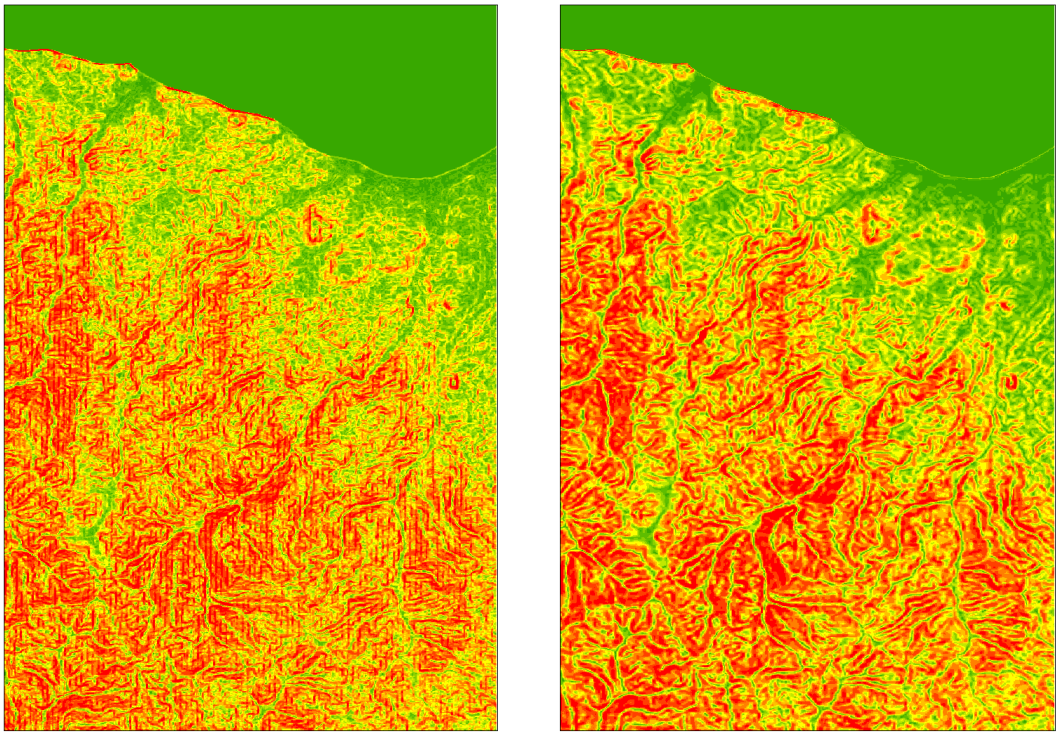

Figure 2: Maps of slope derived from the ASTER DEM. Left, slope obtained from original DEM. Right, slope derived from wavelet-filtered DEM.

2.5 Support Vector Regression

Support Vector Regression (SVR)[16] is an extension of the Support Vector Machines (SVM)[17] approach to regression. In the case of a linear function *f*, this has the form:

$$f(x) = \langle w, x \rangle \tag{1}$$

where $\langle .,. \rangle$ denotes the dot product. It is evident how this formulation is similar to that of linear regression. However, in SVR, or more specifically in ε-SVR[18], an important difference is that the loss function has an ε-insensitivity zone defined as L_ε:

$$L_\varepsilon = |y - f(x_i)|_\varepsilon = \begin{cases} 0 & if\ |y - f(x_i, w)| < \varepsilon \\ |y - f(x_i)| - \varepsilon & otherwise \end{cases} \quad (2)$$

Where ε defines the radius of the hyper-tube within which the regression function must lie. Thus, loss is equal to 0 if the difference between the predicted *f(x,w)* and the measured y_i is less than *ε*. In this way, SVR uses only the data pairs lying on the edges of or outside the tube. The only requirements for *f* are that the function must be as flat as possible, and that the errors must deviate less than ε from the targets. The empirical risk is then minimised, minimising the function:

$$R^{\varepsilon}_{emp}(w, b) = \frac{1}{l}\sum_{i=1}^{l}|y_i - \langle w, x_i\rangle - b|_\varepsilon \quad (3)$$

So a linear regression hyperplane f(x) =⟨ w,x⟩ + b, is found by minimising:

$$R = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}|y + \langle w, x_i\rangle|_\varepsilon \quad (4)$$

where *C* is a term which acts as a trade-off between the two terms in the equation. As some data can still lie outside the *ε*-tube, it is possible add slack variables *(ξ,ξ*)* [19]. These variables measure the distance between the observation and the surface of the ε tube, leading to large deviations in the observations. The value of the two slack variables is linked, since one of the two variables always equals zero, so the slack can only be positive *(ξi* = 0)* or negative *(ξi = 0)*.

$$\begin{aligned} |y - f(w, x)| - \varepsilon &= \xi \\ |y - f(w, x)| - \varepsilon &= \xi^* \end{aligned} \quad (5)$$

Thus a general formulation to minimise both the complexity of the model (measured by $||w||^2$) and its prediction error *(max(ξi,ξi*) = ξi + ξi*)*, is:

$$R_{w,\xi,\xi^*} = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i{}^*) \quad (6)$$

$$subject\ to \begin{cases} y_i - \langle w, x_i\rangle - b \le \varepsilon + \xi_i & i = 1, \dots, m \\ \langle w, x_i\rangle + b - y_i \le \varepsilon + \xi_i^* & i = 1, \dots, m \\ \xi_i, \xi_i^* \ge 0 & i = 1, \dots, m \end{cases}$$

where *C* determines the trade-off between the flatness of the hyperplane and the amount of data pairs that are tolerated outside the ε-tube. The above formulation corresponds to an *ε*-insensitivity loss function $|\xi|_\varepsilon$ described as:

$$|\xi|_\varepsilon := \begin{cases} 0 & if\ |\xi| \le \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \quad (7)$$

The SRM principle addresses the problem of over-fitting by balancing the model's complexity against its success at fitting the training data. After calculation it is possible to find the optimal weight vector of the regression hyperplane as:

$$\mathbf{w_0} = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)x_i \quad (8)$$

and the best regression hyperplane found is expressed as:

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\langle x, x_i\rangle + b \quad (9)$$

The SVR formulation has been extended to perform non-linear regression. The non-linear approximation is approached by using basis expansion, projecting the original data into a higher dimensionality feature space F:

$$x \in Rn \rightarrow \Phi(x) = T[\phi 1(x), \phi 2(x), \ldots, \phi n(x)] \in Rf \quad (10)$$

where $\Phi(x)$ is the mapping function.

The kernel transformation approach is not exclusive to SVR, but is shared by other techniques such as splines or Gaussian processes. The advantage of kernel transformation is that the mapping task could be performed using simple kernel functions[20]. A kernel function $K(x_i,x_j)$ is a function in the input space, which has the advantage of avoiding a complete mapping; instead scalar products in feature space $\Phi^T(x_i)\Phi(x_j)$ are calculated directly computing kernels. In this way it is possible to avoid the problems connected with extreme dimensionality, and to construct an SVR which operates in a virtually infinite dimension space.

3. RESULTS

3.1 Soil Organic Carbon Map of Cyprus

The topsoil Soil Organic Carbon (SOC) map of Cyprus (Fig. 3) has a cell size resolution of 30 m x 30 m. The soil organic carbon content ranges from 0 to 43.3%, with a mean value of 1.53% and a standard deviation of 0.5%. The western part of the island shows the highest concentrations of SOC, especially in the Paphos forest.

Past assessments of soil organic carbon in Cyprus are in agreement with the current study estimates. Dunan et al.[21] have estimated SOC concentration at between 0.99 and 1.66% in northern Cyprus. A recent application of the CENTURY model at the pan-European scale estimated SOC content at between 0.8 and 1.2% for arable lands[22]. Since the LUCAS soil survey has been performed only in the western part of the island, the data has been extrapolated to the non-accessible part (the northeastern part of Cyprus). As a result, the prediction of SOC for northeastern Cyprus has a high level of uncertainty.

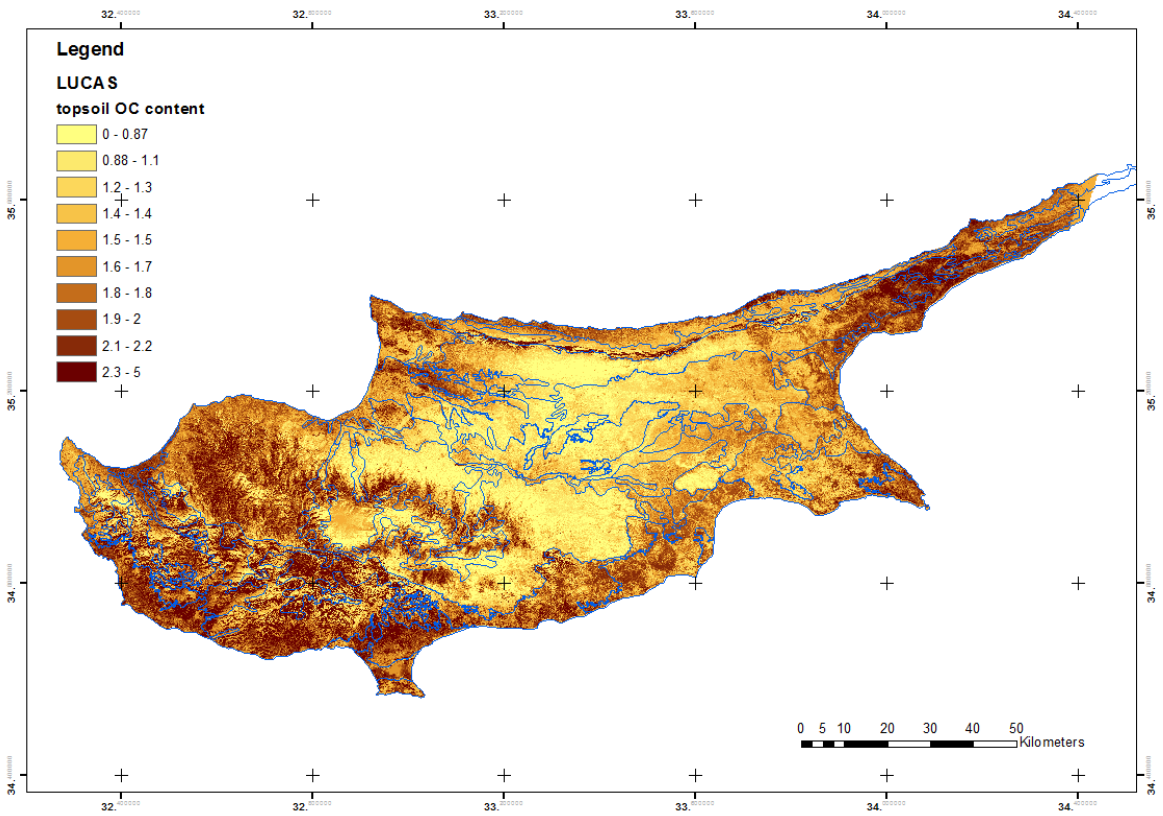

Figure 3: Topsoil organic carbon content (%) in Cyprus

Based on a literature review and a comparison with past studies, the topsoil Soil Organic Carbon (SOC) map of Cyprus has been cross validated using a Leave-One-Out (LOO) procedure (Fig. 4). Cross-validation is often used to estimate the extent to which a prediction model can be generalised by subsequently checking the model's capability of predicting a single sample that has been left out of the fitting set. The LOO cross validation allows for a relatively unbiased verification of the model. In this study, the LOO model was refitted at each iteration, and the best combination of the SVR parameters found. The final model utilises the best combination of SVR parameters found across all the LOO iterations.

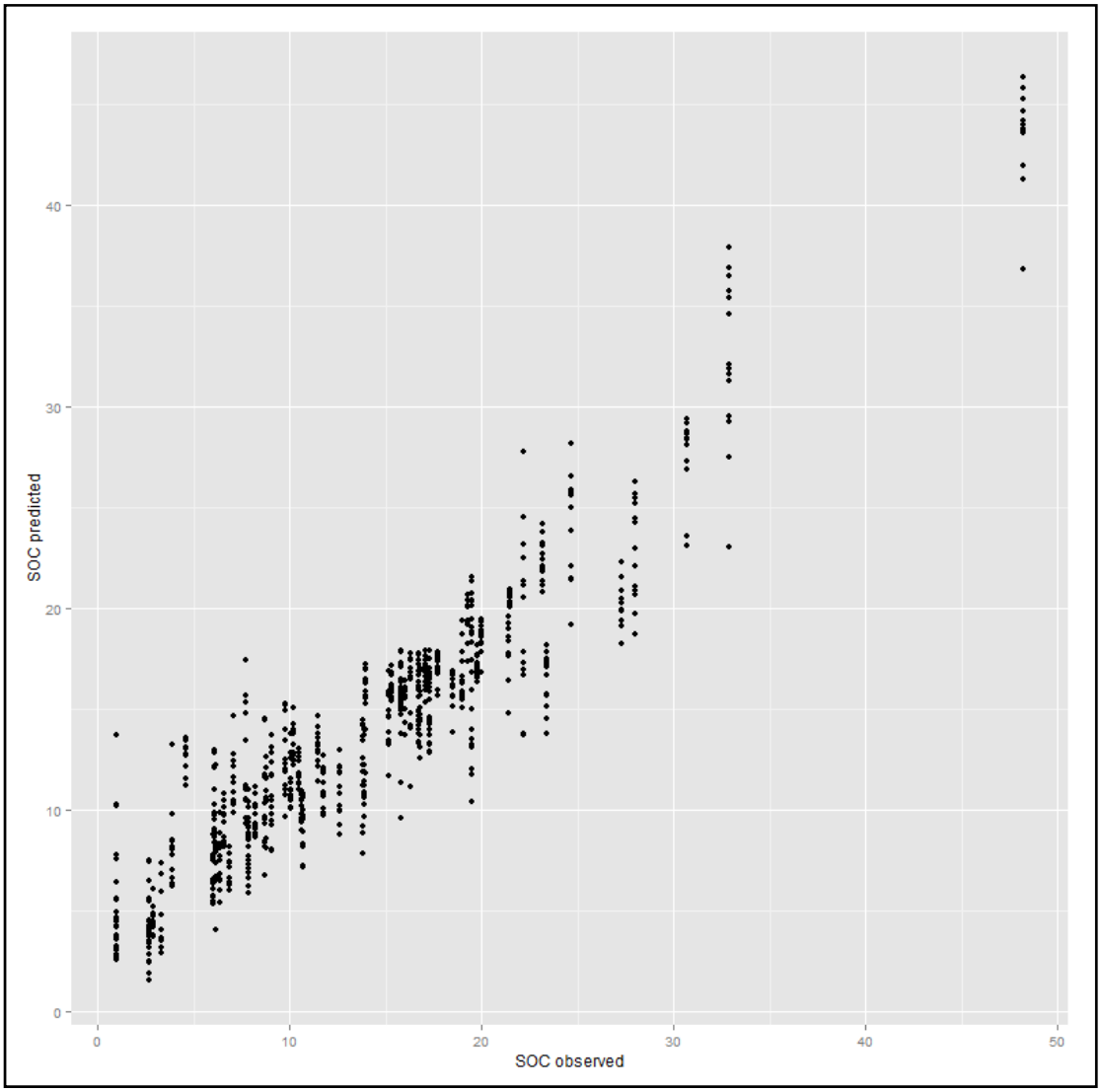

Figure 4: Leave-one-out (LOO) cross validation outcome. Each point represents the relation between an estimated sample and its measured value. Multiple points represent different estimations of the same point made by different models.

3.2 Discussion and conclusions

The estimation of SOC was derived by relating field measurements with a set of spatially exhaustive covariates, including DEM-derived terrain features, MODIS Vegetation indices (16 days, including reflectance) and Landsat ETM+ data. In particular, the SOC levels in the LUCAS database were related with the covariate values in the collocated pixels and their eight surrounding neighbours. The regression model adopted made use of Support Vector Machines (SVM) regression, as this technique is resistant to collinearity and very robust to noise. The SVM regression proved to be very efficient in mapping SOC, with an R^2 fitting of 0.81 and a, R^2 k-fold cross-validation of 0.68.

The remarkable outcome of the SVM model is the fact that the high temporal resolution of the MODIS data greatly enhances the prediction of SOC. This is probably related to the seasonal variation of vegetation cover and its influence on SOC dynamics. This study proves that the inference of SOC levels is possible at regional scales using a limited number of Earth observations. The proposed model can produce data at very high resolution (30 m x 30 m pixels) using covariates. The model, which needs high processing power, can be applied at the regional scale.

The results were also verified both with literature findings and using cross-validation techniques such as Leave-one-out (LOO).

ACKNOWLEDGMENTS

The authors would like to thank Gráinne Mulhern for revision of the article from a linguistic point of view.

REFERENCES

[1] EC, 2006. Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee of the Regions, Thematic Strategy for Soil Protection, COM (2006) 231 final.
[2] EC, 2011. Communication COM(2011) 571 final. Roadmap to a Resource Efficient Europe.
[3] Lal, R., "Soil science and the carbon civilization", Soil Science Society of America Journal, 71 (5), 1425-1437 (2007).
[4] Jones, R.J.A., Hiederer, R., Rusco, E., Montanarella, L., "Estimating organic carbon in the soils of Europe for policy support", Eur. J. Soil Sci., 56, 655–671 (2005).
[5] Panagos, P., Hiederer, R., Van Liedekerke, M., Bampa, F., "Estimating soil organic carbon in Europe based on data collected through an European network" Ecological Indicators 24, 439-450 (2013) .
[6] Tymvios, F., Savvidou, K., Michaelide,s S.C., "Association of geopotential height patterns with heavy rainfall events in Cyprus", Advances in Geosciences, 23 , 73-78 (2010).
[7] Charalambous, K., Bruggeman, A., Lange, M.A., "Assessing the urban water balance: The urban water flow model and its application in Cyprus", Water Science and Technology, 66 (3) , 635-643 (2012).
[8] Michaelides, S.C., Tymvios, F.S., Michaelidou, T., "Spatial and temporal characteristics of the annual rainfall frequency distribution in Cyprus", Atmospheric Research, 94 (4) , 606-661 (2009).
[9] Hadjiparaskevas C. European Soil Bureau — Research Report No. 9, Soil Survey and Monitoring in Cyprus; 2005.
[10] Zomeni, Z. Bruggeman., A. Soil Resources of Cyprus. In Soil Resources of Mediterranean and Caucasus Countries (Eds. Yigini. Y., Panagos, P., Montnanarella, L.) 2013. EUR 25988. JRC Technical Report.
[11] Toth, G., "Impact of land-take on the land resource base for crop production in the European Union", Science of the Total Environment, 435-436, pp. 202-214 (2012).
[12] Martino, L., Fritz, M., 2008. New insight into land cover and land use in Europe. In: Statistics in Focus, vol. 3. Eurostat, Luxembourg.
[13] Panagos, P., Ballabio, C., Yigini, Y., Dunbar, M.B., "Estimating the soil organic carbon content for European NUTS2 regions based on LUCAS data collection", Science of the Total Environment, 442 , 235-246 (2013).
[14] Cohen, D.R., Rutherford N.F., Morisseau, E., Zissimos, A.M., "Geochemical patterns in the soils of Cyprus", Science of the Total Environment, 420 , 250-262 (2012).
[15] Ballabio, C., Fava, F., Rosenmund, A., "A plant ecology approach to digital soil mapping, improving the prediction of soil organic carbon content in alpine grasslands", Geoderma, 187–188, 102-116 (2012).
[16] Smola A.J., Scholkopf B. A tutorial on support vector regression (2004) Statistics and Computing, 14 (3) , pp. 199-222.
[17] Chen P.-H., Lin C.-J., Scholkopf B. A tutorial on v-support vector machines (2005) Applied Stochastic Models in Business and Industry, 21 (2), pp. 111-136.
[18] Vapnik, V., [The Nature of Statistical Learning], Springer (1995)
[19] Vapnik, V., [Statistical Learning Theory], Wiley (1998)
[20] Aizerman, M.A., Braverman, E.M., Rozonoér, L.I., "Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control", 25, 821–837 (1964).
[21] Duman M., Kucuksezgin F., Atalar M., Akcali B.. Geochemistry of the northern Cyprus (NE Mediterranean) shelf sediments: Implications for anthropogenic and lithogenic impact (2012) Marine Pollution Bulletin, 64 (10), pp. 2245-2250.
[22] Lugato E., Panagos P., Bampa F., Jones A., Montanarella L. A new baseline of organic carbon stock in European agricultural soils using a modelling approach (2014), Global Change Biology, 20 (1), pp. 313-326.
[23] Cawley G.C., Talbot N.L.C. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers (2003) Pattern Recognition, 36 (11), pp. 2585-2592.