Mapping soil properties from an existing national soil data set using freely available ancillary data

HENGL Tomislav (1)**, ROSSITER David G.**(1) **and HUSNJAK Stjepan** (2)

(1) Soil Science Division, International Institute for Aerospace Survey & Earth Sciences (ITC), P.O. Box 99, 7500AA Enschede, the Netherlands
(2) Soil Science Department, Faculty of Agriculture, Svetosimunska 25, 10000 Zagreb, Croatia

Abstract

The paper demonstrates how NOAA's 1x1 km NDVI images, downloaded from the web, together with free coarse resolution elevation and climatic data, can be used to improve spatial detail of the Croatian national soil-data set consisting of 2,349 profile observations. Two regression models were developed: to map pH (measured in H_2O) and organic matter (%) in topsoil. Environmental predictors used are standard landform parameters (elevation, slope, curvature, aspect, wetness index), climatic data (rainfall, temperature) and vegetation components derived from the annual NDVI time series. Results show that these two soil properties can be mapped using the CLORPT approach with equal or better precision than with using the existing 1:50,000 soil class map and averaging the properties per soil mapping unit. While the precision of prediction for pH did not improve significantly (residual standard error: 0.60 versus 0.61), the precision for OM was considerably better (residual standard error: 2.81 versus 3.85). The models accounted for 54% (pH) and 66% (organic matter) of the total variation. Principal components of the NDVI time series proved to be most significant predictors of the soil properties, showing clearly general vegetation types and their dynamics. The prediction of pH indeed seems to be more difficult then the prediction of OM. The achieved coefficient of variation for pH was 16.8%, while the model for OM it was 10.8%.

Keywords: soil survey, environmental regression, CLORPT, NOAA's AVHRR, Croatia

Introduction

The ‘pedometric’ soil survey is generally understood to be a quantified inventory of soil attributes expressed over a continuous spatial field. Much work has been done in this subject, almost all as detailed studies at field level. For decision-making, these attributes are needed over large areas (regions, counties, etc.) and at coarser resolution corresponding to medium and small scales. Here we demonstrate how the existing soil data from the national soil survey can be quantified at an acceptable cost, i.e. using cheap or completely free remote sensing, terrain and climatic data.

CLORPT approach to spatial prediction

In recent years, there has been increasing demand for quantitative information at increasingly fine resolutions. Specifically, there is a need for quantitative survey methods applicable at intermediate scales 1:50K to 1:500K (McKenzie *et al*., 1999).

One way of quantifying the soil information is based on Jenny's equation, where the distribution of a soil property is conceptualised as a function of the effect of climate and organisms, modified by relief over some parent material in time. This implies that a given single soil property can in principle be calculated by some predictive equation involving state variables, e.g. by using a simple or generalised linear model. With a number of continuous maps of soil-related variables and a calibrated regression equation from an area of interest, it should be possible to predict soil properties over the whole area. This approach to soil mapping is often called the "CLORPT" approach, using the well-known abbreviation of Jenny's soil-forming factors (McBratney *et al*., 2000). In contrast to the conventional crisp approach to mapping soil properties, which relies on a map of delineations of soil types, from which properties are then inferred (Burrough, 1993), the CLOPRT approach relies mostly on the predictive variables, which are considered to be physical cause of soil properties. At national levels we can also use freely available coarse resolution data to fit the model. Figure 1 shows some functional relations empirically expected to appear in the nature, based on empirical work by Jenny and others (Boul and Hole, 1980). At smaller scales, where we cover larger areas belonging often to different landscapes or climatic zones, we cannot expect that the distribution of a soil variable can be explained using only a single soil-forming factor. Moreover, we can also notice that the relations between the soil and environmental variables are usually non-linear, often quadratic, logarithmic or the inverse of these.

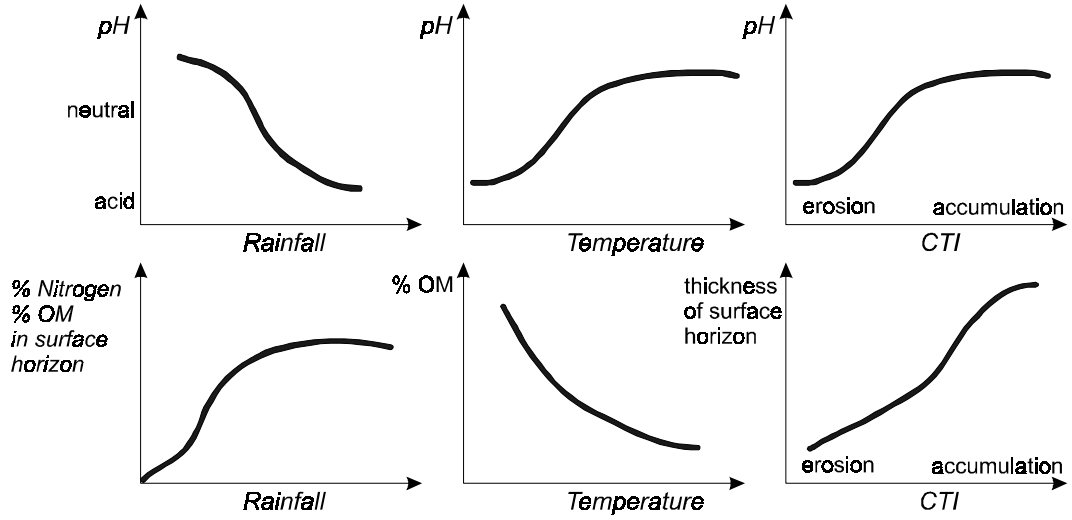


Figure 1 Some empirical models expected to appear in the nature: influence of rainfall and temperature on the OM content and pH of soil.

Remote sensing data and mapping of soil properties

Apart from some specific cases, such as using radar images to map soil moisture content, it has not proved possible to directly use single bands of multispectral sensors to map a soil property. However, compound indices such as NDVI that generally reflects biomass status, have been shown to correlate well with the distribution of the organic matter or epipedon thickness. This clearly has a physical logic and agrees with generally accepted results in soil genesis (McKenzie *et al*., 1999). Recently Odeh and McBratney got a positive result when predicting the clay content with the help of coarse

(1x1 km) AVHRR data (Odeh and McBratney, 2000). They also achieved significant correlation between filtered NDVI values and CEC, EC and pH. Another step was to integrate use of DEM and radiometric data i.e. AVHRR data that is complemented with the DEM parameters: curvature, slope, aspect and the potential drainage density layers (Dobos *et al*., 2000). The results from this case study in Hungary showed that the use of multi-spectral and multi-temporal databases together with the digital elevation and terrain descriptor data improved the classification performance significantly.

Soil data at national levels (Croatia)

The extensive survey with large number of profiles described, analysed and classified in Croatia during the 70's, 80's and 90's (National soil inventory) have still not been used spatially for soil prediction. The original survey covered about 56K km^2 of land with many thousands of profiles (1-5 per 10 km on average). These were full profile descriptions with selective lab data analysis. The planned working scale was medium (1:50K), however, due to the technical limitations of that time, the map had to be generalised to an effective scale of 1:300K. In 1990's the Soil Science Department of Zagreb University developed a GIS database by digitising the 4,500 polygons of the national map and linking to it some 303 representative profiles with 65 taxonomic units (Bogunović and Rapaić, 1993). A part of this large group of observations was inputted and organised in a database format (Martinović and Vranković, 1997). In recent years there have been several attempts to increase the effective scale of the Basic soil map of the Croatia, i.e. to provide spatially explicit information on soil properties at some reasonably high resolution. The soil-profile data in Croatia is extensive and of good quality. It cannot however be interpolated using pure geostatistical methods, due to the fact that the points belong to different landscapes and climatic zones and there is obviously a strong trend or multiple trends in the data. The use of ancillary data that aims at explaining the trend in the data at this level of interpolation is a more preferred method of interpolation since it has physical logic and therefore is more confident (Goovaerts, 1999; Wackernagel, 1998). If the methodology proves to be successful, soil geoinformation could be improved in detail and given in form that can be directly used in land use planning. Similarly, there is a large amount of high quality soil field data in the World that could be quantified without making significant additional investments.

Materials and Methods

Data sources

Soil explanatory, environmental variables used in this study are all of coarse resolution but were collected with almost no cost, i.e. downloaded from web servers or taken from the remote sensing organizations or taken from existing sources. Predictors selected in this study can be grouped as follows, using Jenny's notation:

CL) mean annual rainfall; mean annual surface temperature;

O) long-term normal difference vegetation index (NDVI) and its seasonal components;

R) elevation, slope, aspect, wetness index (CTI), tangent and profile curvature;

P) geological strata, in particular the nature of carbonatic rocks, which are dominant in much of the country;

Long-term normal difference vegetation index (NDVI) and its seasonal components were calculated from NOAA's Advanced Very High Resolution Radiometer (AVHRR)

1x1km images. The mean annual NDVI was calculated for the year 1995 from series of 36 decadal images, downloaded from a public server (US Geological Survey-NASA Distributed Active Archive Centre, 2001). From 36 decadal images we first produced 12 monthly images, taking only the maximum value per pixel. This effectively filtered the images for inaccuracies such as clouds, haze etc. (D'Souza *et al*., 1996). From the 12 monthly summaries we calculated the principal components maps. These were interpreted as long-term biomass (PC1), vegetation region especially taking into account seasonality (PC2), and vegetation type within region (PC3); this corresponds to the analysis by Eastman and Faulk. These are considered to be important soil-forming factors, both as indicators of overall climate and of the vegetation type. Landform parameters-slope, aspect, profile and tangent curvature, were calculated from the 1x1 km resolution DEM also distributed freely. The wetness index was calculated using 10 iterations and multiple flow directions (Quinn *et al*., 1991). The mean annual temperature and rainfall maps were inputted in GIS from the climatic atlas of Croatia (Oppitz and Makjanić, 1988). These were both based on the long-term meteorological observations. All data was georeferenced and processed in the ILWIS 3.0 GIS package at 1x1 km resolution (Unit Geo Software Development, 2001). Statistical analysis was done S-Plus 4.5 statistical package (Lam, 1999).

Study area

Croatia was mapped during the 1970's and 1980's in a large national project named "Basic Soil map of Croatia", using the Yugoslav methodology for soil mapping and description (Bogunović *et al*., 1998; Škoric *et al*., 1985). The total area of Croatia is 56,610 km² (without sea surface). Here we do calculations within an area contained in a rectangle of 501 x 491 km (same size in pixels). The elevations range from the sea level up to 1,831 m. The general climate and the biogeography can be grouped in two main groups: Mediterranean and moderate continental. See the study area, soil profiles used and bounding coordinates in Figure 2.

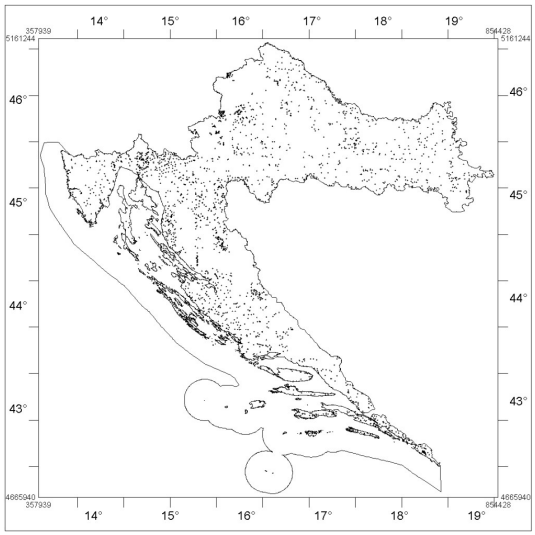


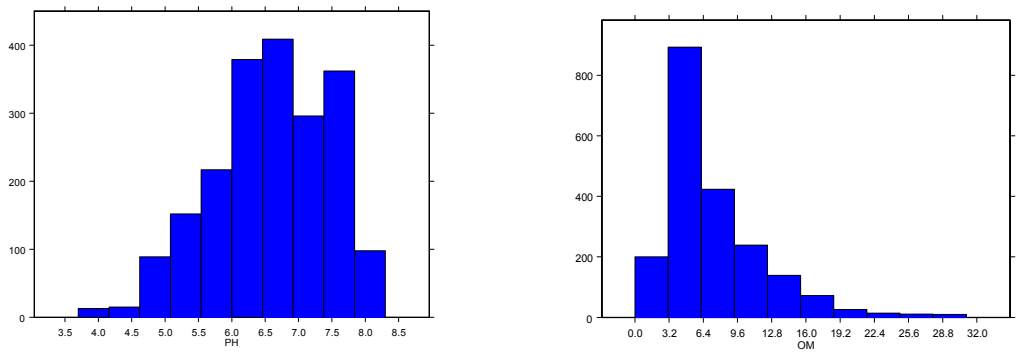
Figure 2 Study area-Croatia and 2,350 profiles used to interpolate the data. The co-ordinates are in UTM system-zone 33, northern hemisphere.

Regression analysis

The original profile observations (2,349) had to first be resampled to the 1x1 km resolution to be able to integrate the data. We examine in this work regression models to predict the content of organic matter (OM in %) and pH measured (in H_2O) in topsoil. The profile locations were selected as representative profiles, but for polypedons of much smaller than 1 km^2. Therefore, we needed first to calculate appropriate values of properties at the 1x1 km resolution. We used block kriging (Burrough and McDonnell, 1998), to average local profile observations per square kilometre. This resulted in smoothed values of pH and OM. The OM content had to be first log-transformed to achieve normality (see descriptive statistics in Table 1). Both pH and OM content showed spatial correlation also at this scale, and were interpolated using a simple spherical model (pH: sill=0.45; nugget=1.0; range=14,000; ln(OM): sill=0.30; nugget=0.52; range=35,000). The total number of pixels was reduced to 2077 since some 10% of the profiles were located at distance of less than 1,000 m.

Table 1 Descriptive statistics for pH and OM in top-horizon (N=2349).

Variable	pH in H_2O	OM in topsoil (%)
Average	6.58	7.53
Range (95%)	4.2 – 8.1	1.2 – 28.7
Stdev	1.15	7.49



To fit the soil data (pH and OM in this case), we used stepwise regression method where the system selects most significant predictors using an iterative procedure (Cook and Weisberg, 1999). We have also used non-linear transformations (X^2-transformation), to improve the correlation coefficient. The results of regression analysis (CLORPT approach) were then compared to the values calculated from the soil map as attribute maps. In this case, the profile data was averaged per each of the 65 Soil Mapping Units (SMU). Both methods were evaluated using residuals calculated at the profile locations. The comparison can be seen in Figure 5 The two approaches were compared on the basis of the proportion of the variance in the calibration dataset accounted for, and by the properties of the residuals from the model fit.

Results and Discussion

Time series analysis of NDVI multi-temporal images

We used only first four principal components of 12 monthly NDVI maps that explained almost 99.9% of total variation (NDVI_PC1 = 98.7%; NDVI_PC2 = 0.63%;

NDVI_PC3 = 0.33%; NDVI_PC4 = 0.16%). When examined only visually, the results seem to coincide with the previous results for the African dataset (Eastman and Fulk, 1993) where the PC1 represents characteristics or mean NDVI value over an area, while the second (NDVI_PC2) and third (NDVI_PC3) principal component express the seasonal changes i.e. winter/summer dichotomy. Moreover, in this case we noticed that the NDVI_PC3 appears to be good segregator of evergreen vegetation (mountainous forests) from the broad-leave species. Similarly, the NDVI_PC2 seems to show a clear boundary between the Mediterranean and Continental climatic region (Figure 4). In some cases, there were still inaccuracies in NDVI PC components. That was due to the fact that some marginal pixels of first four months differed in highly contrasting areas – junction from the land to water surface. The source of error is unknown but it could be that the authors used different masks for water bodies. This was hard to detect on the original images, but showed well after the principal component analysis. These pixels were then corrected with approximate values taken from neighbouring months. This should be, however, taken into account when using the principal component analysis of NOAA data in future.

Empirically expected relationships were more or less achieved in all cases. Especially clear correlations were achieved between OM and rainfall and between pH and temperature. Also the PC2 and PC3 of the NDVI time series showed strong correlation with the distribution of the pH and OM (see Figure 3). The climatic and vegetation variables prove to be more significant predictors then the landform parameters at this scale of work. This is probably due to the very coarse resolution of the DEM, which thus cannot provide very accurate estimation of the landform parameters (Lagacherie *et al*., 1996).

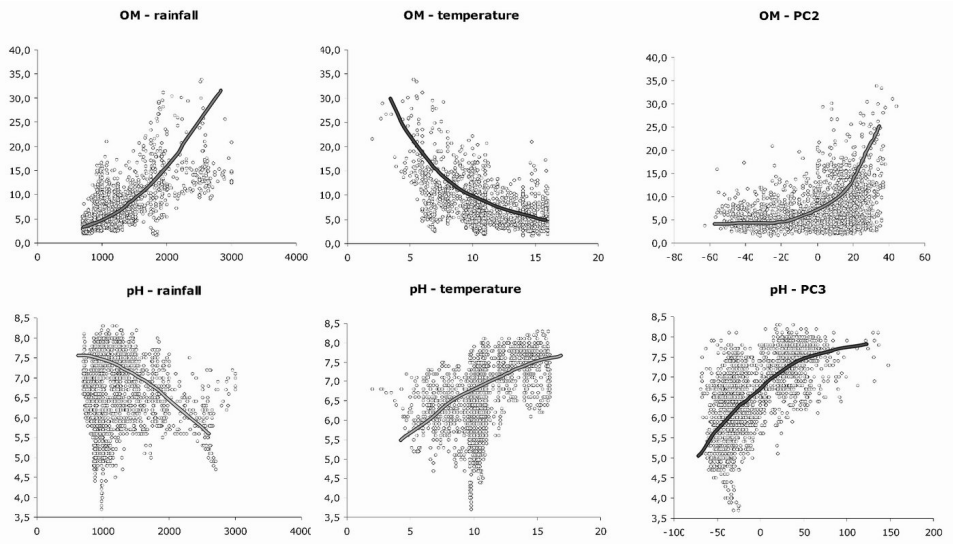


Figure 3 Achieved correlations: OM and pH in topsoil in relation to the climatic and vegetation variables. Compare with the empirical models.

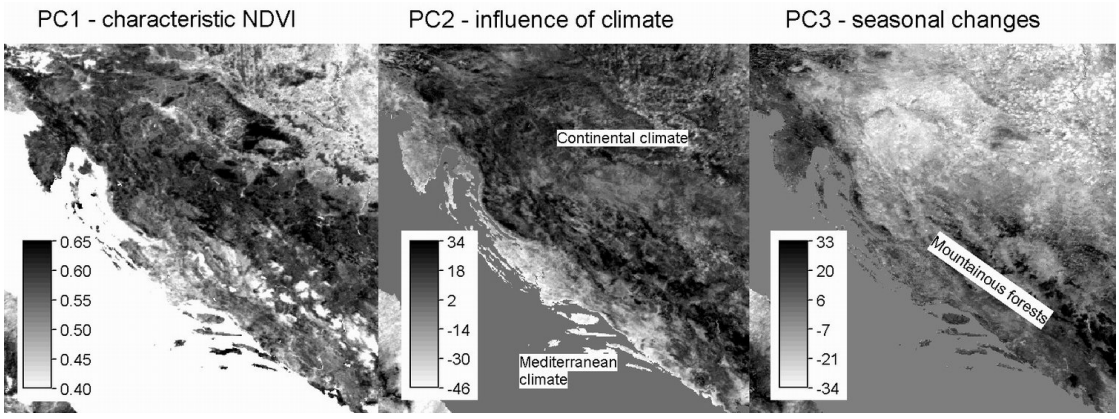


Figure 4 Results of principal component analysis of 12 monthly AVHRR NDVI images.

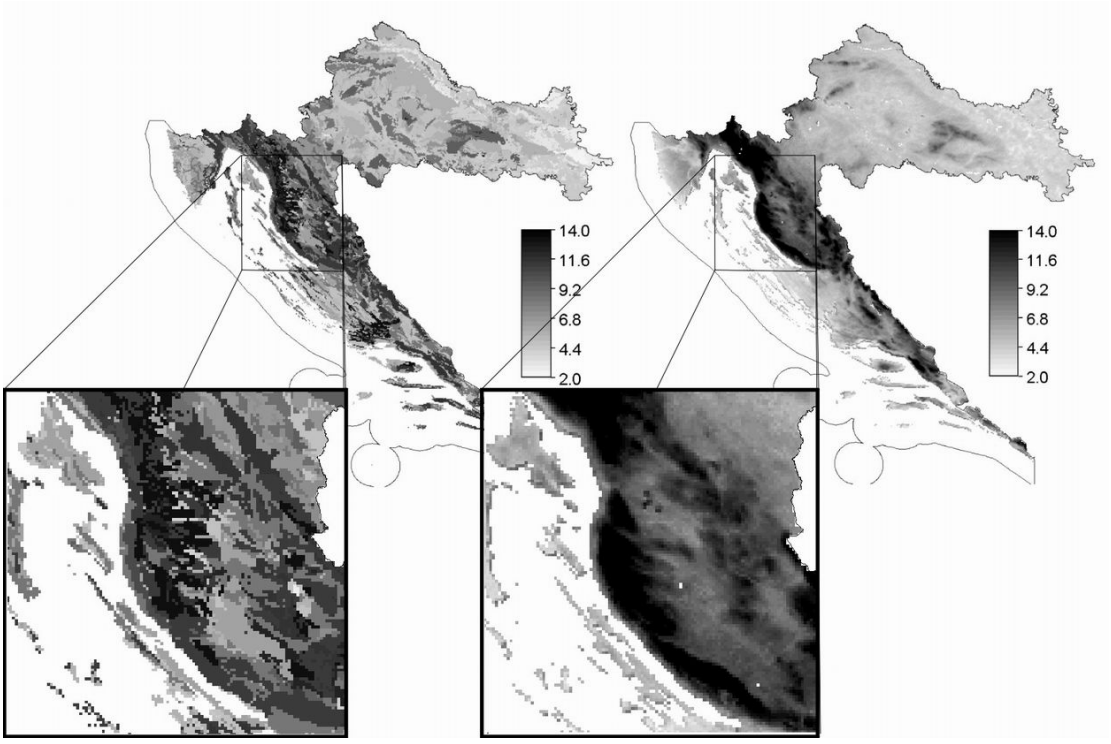


Figure 5 OM in topsoil horizon: comparison of the map produced using the regression equation and based on the soil map.

This free source also had some minor artefacts that were discovered only after the slope or curvature maps were calculated. The strongest correlation was however achieved between the OM and elevation. This probably reflects the fact that forests are denser at higher elevations; temperatures are cooler and rainfall higher.

Mapping soil properties at fine grain of detail

The regression analysis gave significant regression models in both cases. The models accounted for 54% (pH) and 66% (organic matter) of total variation (see Table 2). When compared to the traditional way of producing soil attribute maps (averaging per SMU), the environmental regression gives better detail and more realistic

continuous change of property. The residuals in both conventional maps and the CLORPT approach show normal distributions with no systematic bias in prediction. While the prediction of pH did not improve significantly (residual standard error: 0.60 versus 0.61), the CLORPT model for the prediction of the OM achieved better precision of the prediction (residual standard error: 2.81 versus 3.85). The prediction of pH indeed seems to be more difficult then the prediction of OM. The achieved coefficient of variation, calculated as the ratio between the residual standard error and the range of variation, i.e. precision of prediction for pH was 16.8%, while the model for OM gave a precision of 10.8%.

Table 2 Regression analysis statistics: pH and OM as a function of environmental variables.

	pH in H_2O	Organic matter
N	2016	2013
Total number of predictors	15	15
Most significant predictors	$NDVI^2$; $NDVI\_PC3^2$; NDVI, NDVI_PC2	NDVI_PC2, NDVI2; DEM, Intercept
Multiple R-Squared	0.54	0.66
Residual standard error	0.603	2.81
Range (min. – max.)	4.54 – 8.13	1.1 – 26.9
Precision s_x/Range (%)	±16.8	±10.9

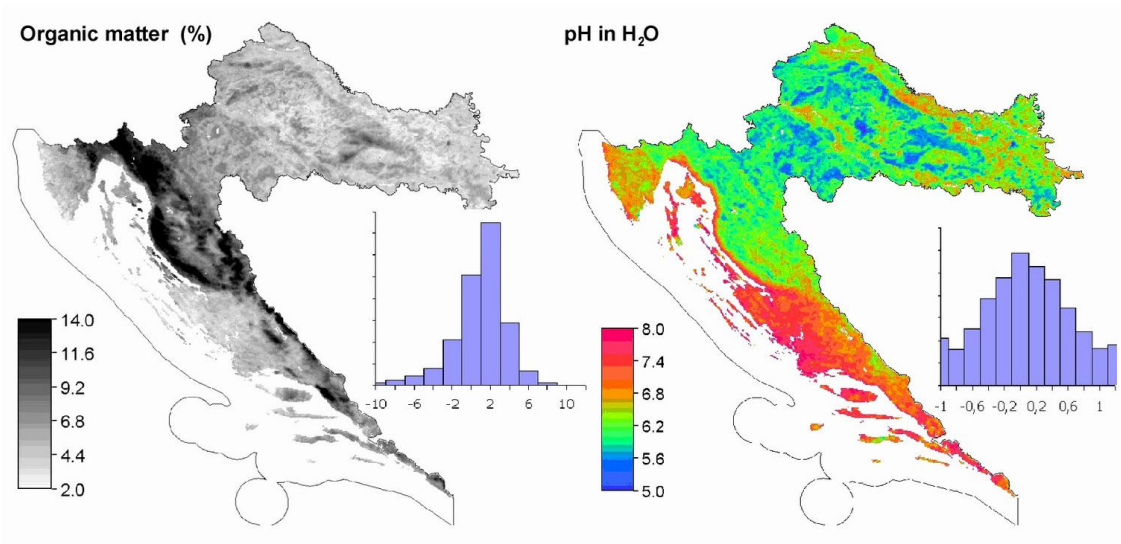


Figure 6 Predicted pH and OM, regression models and histograms of residuals.

Conclusion

This study shows that a regression model developed based on the CLOPRT approach can be used to map selected soil properties with equal or higher precision than by inference from a map of soil classes. We developed here reasonably detailed maps of two selected soil properties: pH measured in H_2O and OM in the topsoil. Interesting discovery is the importance of the annual NDVI time series PC components with high

multi-temporal resolution. We were able to extract several environmental factors from a single variable (NDVI) based only on its multi-temporal resolution, using principal component analysis. Some of these components proved to be the most significant predictors of the soil properties, as they reflect the vegetation types and their dynamics.
The CLOPRT approach has proven to be effective way of mapping soil properties at fine grain of detail. In this case we produced two maps at resolution of 1x1 km that probably belongs to finer scales in-between 1:300 K and 1:1 M. We can also notice that the Croatian data set used here have also provided information on soil properties geostatistically outside the range of spatial variation, but actually inside the state-space. The logic behind this model strongly suggests that it could be applied to estimate the modelled variables (pH and OM) based on the environmental predictors outside the study area, that is, in neighbouring countries, even without calibration points, as long as the soil-forming environment is not too different.

References

Bogunović, M. and M. Rapaić. 1993. Digitalisation of basic soil map of Republic of Croatia. Bilt. Dalj. Istr. Fotoint. 12:65-76.

Bogunović, M., e. Vidaèek, S. Husnjak and M. Sraka. 1998. Inventory of Soils in Croatia. Agriculturae Conspectus Scientificus. 63(3):105-112.

Boul, S.W. and F.D. Hole. 1980. Soil Genesis and Classification. Iowa State Univ. Press, Ames. 404 p.

Burrough, P.A. 1993. The technologic paradox in soil survey: new methods and techniques of data capture and handling. ITC Journal(No):15-22.

Burrough, P.A. and R.A. McDonnell. 1998. Principles of Geographical Information Systems. Oxford University Press, Oxford. 327 p.

Cook, R.D. and S. Weisberg. 1999. Applied Regression Including Computing and Graphics. Probability and Statistics. John Wiley and Sons, New York. 583 p.

Dobos, E., E. Michelib, M.F. Baumgardnerc, L. Biehlc and T. Helt. 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. Geoderma 97(3-4).

D'Souza, G., A.S. Belward and J.P. Malingreau (eds.). 1996. Advances in the Use of NOAA AVHRR Data for Land Applications. Kluwer Academic Publishers, Dodrecht, Boston, London. 462 p.

Eastman, J.R. and M. Fulk. 1993. Long sequence time series evaluation using standardized principal components. Photogrammetric Engineering and Remote Sensing 59(8):1307-1312.

Goovaerts, P. 1999. Using elevation to aid the geostatistical mapping of rainfall erosivity. Catena 34(3-4):227-242.

Škorić, A., G. Filipovski and M. Čirić. 1985. Classification of Yugoslav Soils. Academy of Sciences and Arts of Bosnia and Hercegovina, Sarajevo. 71 p.

Lagacherie, P., R. Moussa, D. Cormary and J. Molenat. 1996. Effects of DEM data source and sampling pattern on topographical parameters and on a topography-based hydrological model. Application of geographic information systems in

hydrology and water resources management. Proc. HydroGIS'96 conference, Vienna. 235:191-199.

Lam, L. 1999. An Introduction to S-PLUS for Windows. Candiensten, Amsterdam. 350 p.

Martinović, J. and A. Vranković. (eds.). 1997. Baza podataka o hrvatskim tlima, I. Drž avna uprava za zaštitu prirode i okoliša, Zagreb. 365 p.

McBratney, A.B., I.O.A. Odeh, T.F.A. Bishop, M.S. Dunbar and T.M. Shatar. 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97(3-4):293-327.

McKenzie, N.J., P.J. Ryan and J.J. de Gruijter. 1999. Spatial prediction of soil properties using environmental correlation. Pedometrics 97,89(1-2):67-94.

Odeh, I.O.A. and A.B. McBratney. 2000. Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia. Geoderma 97(3-4):237-254.

Oppitz, O. and B. Makjanić. 1988. Croatia-Climate. Enciklopedija Jugoslavije, 5. JLZ, Zagreb. 150 p.

Quinn, P., K. Beven, P. Chevallier and O. Planchon. 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrological processes 5:59-79.

U.S. Geological Survey-NASA Distributed Active Archive Centre. 2001. FTP access to Global AVHRR 10-day composite data. http://edcdaac.usgs.gov/.

Unit Geo Software Development. 2001. ILWIS 3.0 Academic User's Guide. http://www.itc.nl/ilwis/. ITC, Enschede. 520 p.

Wackernagel, H. 1998. Multivariate Geostatistics: An Introduction with Applications. Springer-Verlag. 285 p.