Towards Detailed and Public Data on Croatian Soil as a Prerequisite for Sustainable Environmental Management and Nature Conservation

Josip Križan[1], Luka Antonić[1,], Tomislav Hengl[2], Oleg Antonić[3]*

[1] MultiOne j.d.o.o., Zagreb, Croatia, lantonic@multione.hr
[2] EnvirometriX Ltd, Wageningen, The Netherlands, tom.hengl@envirometrix.net
[3] Granum Salis Cooperative, Zagreb, Croatia, oantonic@zadrugagranumsalis.hr
* corresponding author

doi: 10.5281/zenodo.11585221

Abstract: Soil is an indispensable component of the terrestrial ecosystem. It is thus unimaginable to perform sustainable land management without information on the pedosphere, regularly collected by mapping pedocartographic units and/or sampling at typical locations for laboratory acquisition of physical and chemical parameters. Despite soil inventory having been done in Croatia for decades, complex interinstitutional relations with often poorly regulated jurisdiction, rights and obligations have led to the nonexistence of a public soil database to date, one that would assist in daily efforts of many stakeholders (farmers, foresters, conservationists, planners, consultants, authorities and various decision makers, among others), and would strongly improve environmental and natural resource management in our country. In a project implemented on behalf of the World Bank from 2020 to 2021, a 30 m resolution spatial dataset was produced with 16 standard pedological variables estimated by applying machine learning algorithms against satellite imagery alongside climatic, geological and geomorphometric indicators (for a total of 533 covariates), while unifying soil data from numerous sources, some of which are not publicly available, which limits the availability of the project results. However, subsequent analyses performed outside of the project and only on publicly available target data have yielded results of similar quality, providing new opportunity to many interested parties.

Keywords: soil mapping; machine learning; open data.

1 Introduction

Open access to high-resolution spatial data in general, and to soil data in particular, has the potential to be highly transformative to a wide palette of human activity, but perhaps most notably to environmental management (extending to agronomy and to most other forms of cultivation) and to nature conservation. Thus, effort is widely being made to provide open spatial data at a global (through Copernicus and similar initiatives) or at least continental scale (e.g. the HORIZON programmes of the European Commission). However, public funding (e.g. through various European Commission programmes) is also often spent on compiling datasets which end up having a singular purpose to meet the objectives of the project at hand, and shelved afterwards with no means of outside access, despite their potential usefulness for other purposes. Such cases are often lost opportunities to bring added value to work already completed, as well as to prevent future duplication of work (and, by extension, funding). This is particularly observable at the national level, where public institutions often claim ownership of underlying data (despite the questionability of such claims upon any data which have invariably also been procured through public funding), and therefore prohibit any distribution of the derivatives thereof.

One such occurrence was during the STARS-RAS project, the basis for Agro-Ecological Zoning of Croatian territories. Through the course of the project, we were hired to produce a complete digital soil map of Croatia, consisting of 16 soil physical and chemical properties and soil classification as per the World Reference Base taxonomical standard, mapped across the entire country at 30 m resolution. Due to various government agencies claiming ownership on some of the underlying data used to produce these layers (both soil samples and covariates), any use of the products outside the project was prohibited. With the goal of bringing any potential value of these products to the wider Croatian public for further application, we have reproduced the described data products at our own cost, with fully analogous methodology, but this time only using publicly available covariates and soil samples, and have made the data products publicly available via Zenodo.

Physical and chemical soil properties and classification were modeled through ensemble machine learning (using gradient boosting trees, gradient boosting linear models and random forests). While the physical and chemical property models displayed mixed capability for generalization, many of the properties were modeled with sufficient quality to demonstrate usefulness (validation adjusted R^2 above 0.8). Classification modeling, however, displayed poor generalization capacity (validation weighted $f1 < 0.5$), indicating a need for methodology revision.

2 Materials and methods

The methodology for predictive soil mapping was adopted from Hengl and MacMillan, 2019. An ensemble machine learning approach was used against an exhaustive set of spatial covariates to model a total of 16 soil physical and chemical properties: organic carbon content (oc), total NCS nitrogen (n_tot_ncs), Mehlich3 extractable calcium (ca_mehlich3), potassium (k_mechlich3), magnesium (mg_mehlich3) and phosphorus (p_mehlich3), summary cation exchange capacity (cec_sum), saturation extract electrical conductivity (ec_satp), carbonate content

(caco3), 1:1 soil-water (ph_h2o) and soil-KCl suspension pH (ph_kcl), total clay (clay_tot_psa), silt (silt_tot_psa) and sand content (sand_tot_psa), coarse fragment content (wpg2), < 2 mm fraction ovendry bulk density (db_od) and depth to bedrock up to 400 cm (dbr), as well as soil type classification as per the World Reference Base taxonomical standard (wrb_rsg). Additionally, soil texture classification was derived from the modeled clay, silt and sand content, using the method described in Radočaj et al. (2020).

2.1 Target dataset

Target soil property data was compiled from the following sources: 1) data on Croatian soils collected by the former State Department for Environment and Nature Protection, consisting of 2199 pedological profiles sampled from 1963 to 1966 (Martinović and Vranković 1997), further referenced as martinovic_1997, 2) data from the project *Spatial variability of trace and toxic metals in agricultural soils of Croatia*, provided by the Faculty of Agriculture, Zagreb, consisting of 811 samples acquired on a 8x8 km grid (Romić 2013), further referenced as agricultural_2013, 3) data from the project *Change in soil carbon stocks and calculation of soil total nitrogen and organic carbon trends and C:N ratios*, consisting of 2519 samples collected from 1994 to 2004 (for the purpose of compiling the Geochemistry Atlas of Croatia) and 742 additional samples from locations revisited from 2015 to 2016 (URL-1), further referenced as azo_2013 and azo_2016. The depth to bedrock target dataset was further enhanced with national piezometric observations (Croatian Waters 2016), considering measurements of a minimum of 4 m, totaling in 812 additional observations. Data from all sources were harmonized through unit consolidation and extraction of measurements equivalent to the topsoil (0–30 cm) depth horizon. As the source measurements were performed at varying depth horizons (of below 30 cm width), topsoil equivalent was obtained as a weighted average of measurements from all horizons which overlap with topsoil, with overlap fraction used as the weight. The total number of topsoil observations obtained per soil parameter is listed in Table 1. The spatial distribution of observations is shown in Figure 1.

Table 1. Number of observations obtained per soil parameter.

Parameter	Samples	Parameter	Samples
oc	3452	db_od	3148
n_tot_ncs	5074	p_mehlich3	2598
ca_mehlich3	726	ec_satp	685
k_mehlich3	3364	caco3	2624
mg_mehlich3	726	ph_h2o	2784
clay_tot_psa	3408	ph_kcl	2041
silt_tot_psa	3369	dbr	2576
sand_tot_psa	3407	wrb_rsg	5025
wpg2	2116		

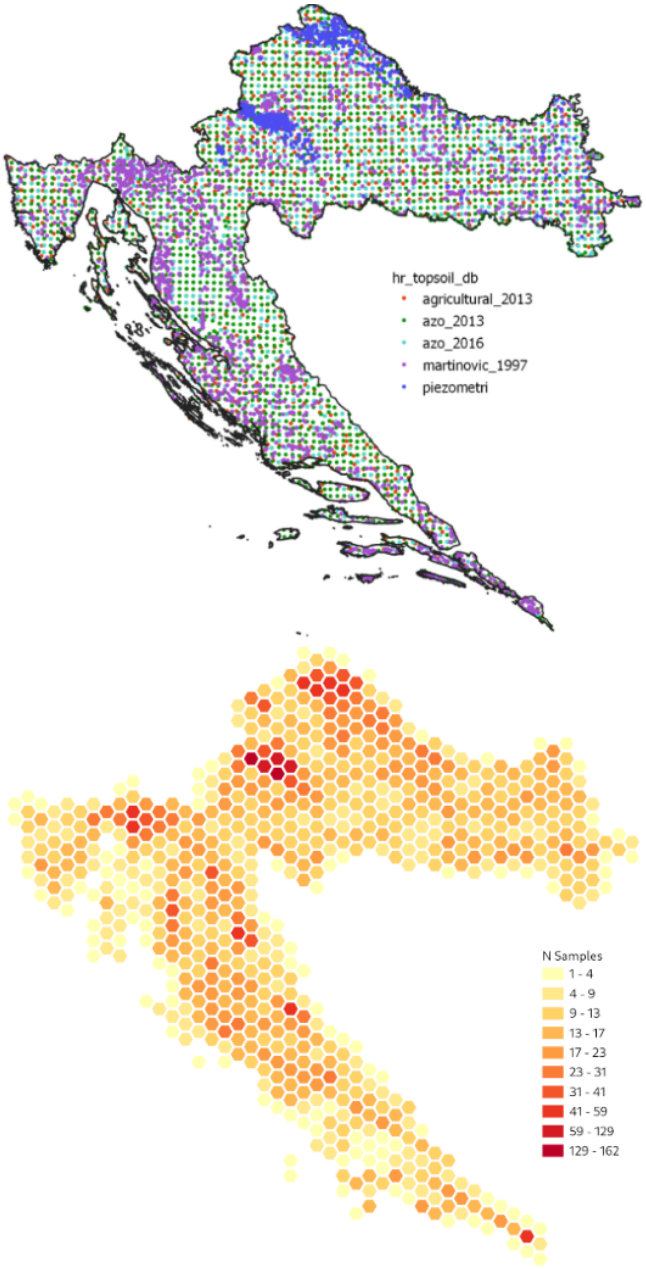

Figure 1. Target observation coverage per source dataset (top), total sample density (bottom).

2.2 Covariate set and preprocessing

The covariates used for modeling soil properties and taxonomy were as follows: 1) a total of 16 terrain property layers derived from the official Croatian elevation dataset (URL-2), produced using GRASS GIS (GRASS Development Team 2017) and WhiteboxTools (Lindsay 2016), 2) Sentinel-2 seasonal cloudless mosaics of 6 bands (blue, green, red, NIR, SWIR1, SWIR2) at 3 quantiles (P25, median and P75) from 2018 to 2020 for a total of 180 layers, 3) Sentinel-1 monthly mosaics for 2018 and 2019 of the VV and VH bands, mosaicked as monthly mean over pixel, totaling 48 layers, 4) Global surface water (GSW) dataset for 2019 (Pekel et al. 2016), 6 layers, 5) Surface soil moisture (SSM) dataset for 2018 and 2019 (URL-3), 6 layers, 6) CHELSA climate and bioclimatic data (URL-4), 50 layers, 7) 15-year global 1 km cloud cover (Wilson and Jetz 2016), 13 layers, 8) Long-term MODIS LST day-time and night-time temperatures (URL-5),

100 layers, 9) Long-term Aerosol thickness, monthly means and annual mean and standard deviation for period of 2000 to 2017 (URL-6), 14 layers, 10) Monthly precipitation at 1 km resolution based on SM2RAIN-ASCAT 2007–2018, IMERGE, CHELSA Climate and WorldClim (URL-7), 25 layers, 11) Monthly mean water vapor content for period of 2000 to 2017 (URL-8), 12 layers, 12) Geological map of Croatia (Croatian geological institute 2019), reclassified to 54 classes representing lithological composition, resulting in 54 layers representing pixel distance from nearest occurrence of each lithological class. All covariates were reprojected to ETRS89-extended / LAEA Europe and resampled in 30m resolution,

with the considered data mask consisting of the landmask dictated by the coverage of the terrain layers (corresponding to land on Croatian territory), with pixels classified as built-up areas and water bodies (Pflugmacher et al. 2019) left out of the mask. The total set of covariates resulted in over 400 individual layers, a very high number relative to the training sample size (as low as ~700 for some target variables). Furthermore, layers extracted from the same dataset were highly intercorrelated. To account for this, principal component analysis (PCA) was performed on the covariates to reduce the dimensionality of the input space. Predictors were divided into 10 groups and a separate PCA was performed on each group, such that the number of components accounted for 90% of cumulative variability of the group, resulting in under 70 total covariate layers across all groups. The PCAs were fitted on predictor values sampled at one million randomly chosen pixels within the data mask.

2.3 Model specification

Physical and chemical soil properties were modeled with two-level stacked regressor ensembles. The bottom level of the ensemble stack consisted of three groups of models: 1) gradient boosting trees, 2) gradient boosting linear models and 3) random forest models. For each target property, five models from each group were selected for the ensemble, through a random search over their respective hyperparameter space (consisting of learning rate, conservativity constraints and tree topology), spatially cross-validated with five folds consisting of samples grouped into discrete 10x10 km tiles over the spatial domain. At the top stack level, another gradient boosting tree model was fitted to predict the final output from the predictions of the fifteen bottom-level models, with output constrained to the value range of the training data. In order to normalize target distributions, all properties were modeled as the natural logarithms of the actual values, with the exceptions of sand_tot_psa, silt_tot_psa, clay_tot_psa, ph_h2o, ph_kcl and db_od, and were converted back during postprocessing.

Additionally, uncertainty of the predictions was estimated through two additional top-level models, fitted to predict the 5% and 95% quantiles of the output. Soil type classification was modeled with a voting classifier ensemble, with individual estimators trained in the same manner as with the regression models. The final class probability was obtained as a weighted soft vote between the fifteen selected estimators, with the estimator weight equal to its mean *f1* score over the validation fold, multiplied by the ratio of test to train scoring in order to additionally penalize overfitting. In addition to pixel-wise probability of each class, uncertainty of the classification was estimated as pixel-wise relative entropy (ratio of prediction entropy to maximum possible entropy for the given number of classes). Modeling was performed in a Python environment with XGBoost (Chen and Hestrin 2016) and scikit-learn (Pedregosa et al. 2011), on a workstation equipped with a 12-core AMD Ryzen Threadripper 2920X CPU, a single NVIDIA RTX 2060 GPU and 64 GiB of memory.

Table 2. Regression metrics for physical and chemical soil properties, including R-squared, R-squared adjusted for number of predictors and observations, and concordance correlation coefficient (CCC), mean over training folds and validation folds.

Target	R^2 train	R^2 val	R^2_adj train	R^2_adj val	CCC train	CCC val
oc	0.98	0.92	0.98	0.91	0.99	0.96
ln(oc)	0.96	0.91	0.95	0.91	0.98	0.95
n_tot_ncs	0.86	0.75	0.86	0.74	0.92	0.85
ln(n_tot_ncs)	0.85	0.74	0.85	0.74	0.91	0.85
ph_h2o	0.94	0.85	0.94	0.85	0.97	0.92
ph_kcl	0.93	0.81	0.93	0.80	0.96	0.89
clay_tot_psa	0.93	0.86	0.93	0.86	0.96	0.93
silt_tot_psa	0.74	0.52	0.74	0.51	0.83	0.68
sand_tot_psa	0.87	0.74	0.86	0.74	0.92	0.85
db_od	0.79	0.70	0.78	0.70	0.88	0.82
dbr	0.99	0.94	0.99	0.94	0.99	0.97
ln(dbr)	0.98	0.93	0.98	0.93	0.99	0.96
caco3	0.95	0.74	0.95	0.73	0.97	0.85
ln(caco3)	0.93	0.81	0.93	0.80	0.96	0.90
wpg2	0.97	0.77	0.97	0.76	0.99	0.88
ln(wpg2)	0.98	0.87	0.98	0.87	0.99	0.93
p_mehlich3	0.96	0.70	0.96	0.69	0.98	0.84
ln(p_mehlich3)	0.90	0.78	0.90	0.77	0.94	0.87
k_mehlich3	0.94	0.82	0.94	0.81	0.97	0.90
ln(k_mehlich3)	0.92	0.83	0.92	0.83	0.95	0.91
ca_mehlich3	0.99	0.80	0.98	0.78	0.99	0.89
ln(ca_mehlich3)	0.99	0.88	0.99	0.86	0.99	0.93
mg_mehlich3	0.98	0.67	0.98	0.63	0.99	0.81
ln(mg_mehlich3)	0.97	0.76	0.97	0.73	0.99	0.86
ec_satp	0.99	0.73	0.99	0.69	1.00	0.83
ln(ec_satp)	0.99	0.78	0.98	0.75	0.99	0.87
cec_sum	0.88	0.73	0.88	0.72	0.93	0.84
ln(cec_sum)	0.89	0.74	0.88	0.73	0.94	0.85

3 Results and discussion

Soil property regression yielded mixed results overall, with the worst performing model achieving mean validation adjusted R^2 of 0.51 (soil silt content), while the top performer scored as high as 0.91 (organic carbon content), indicating that the methodology described, while capable of sufficient of even very good generalization for some soil properties, may not be uniformly applicable to the entire parameter space, at least with respect to the order of magnitude of the number of available target samples. Regression metrics for all targets are listed in Table 2.

Classification modeling yielded poor overall generalization, with a weighted *f1* score over the validation fold of 0.46, while the score over the entire sample space was 0.81, indicating high overfitting. Classification metrics are shown in Table 3, while the relative confusion matrix of the classifier, scatter plots for two of the regressors, as well as the soil texture map produced from the modeled silt, sand and clay contents are shown in Figure 2.

Table 3. Classification metrics over the validation fold and across the entire target dataset, including precision, recall, and f1 score, weighted with class sample size.

	precision	**recall**	**f1**
validation	0.47	0.50	0.46
All	0.85	0.81	0.81

A notable deficiency in the described methodology is the static nature of modeling soil properties without regard for the temporal dimension. Soil changes slowly relative to human perception, but it does indeed change, mainly through human activity and increasing change in climate. However, it is difficult to account for time variation when modeling soil at a local (national) level as described, due to low total number of available samples (which is why only a single, static layer was produced for each soil property, even though the target soil samples were collected over a span of 53 years). Consequently, efforts have been made (or are underway) at a continental or wider level (such as the soil layers listed at URL-9) in order to take advantage of

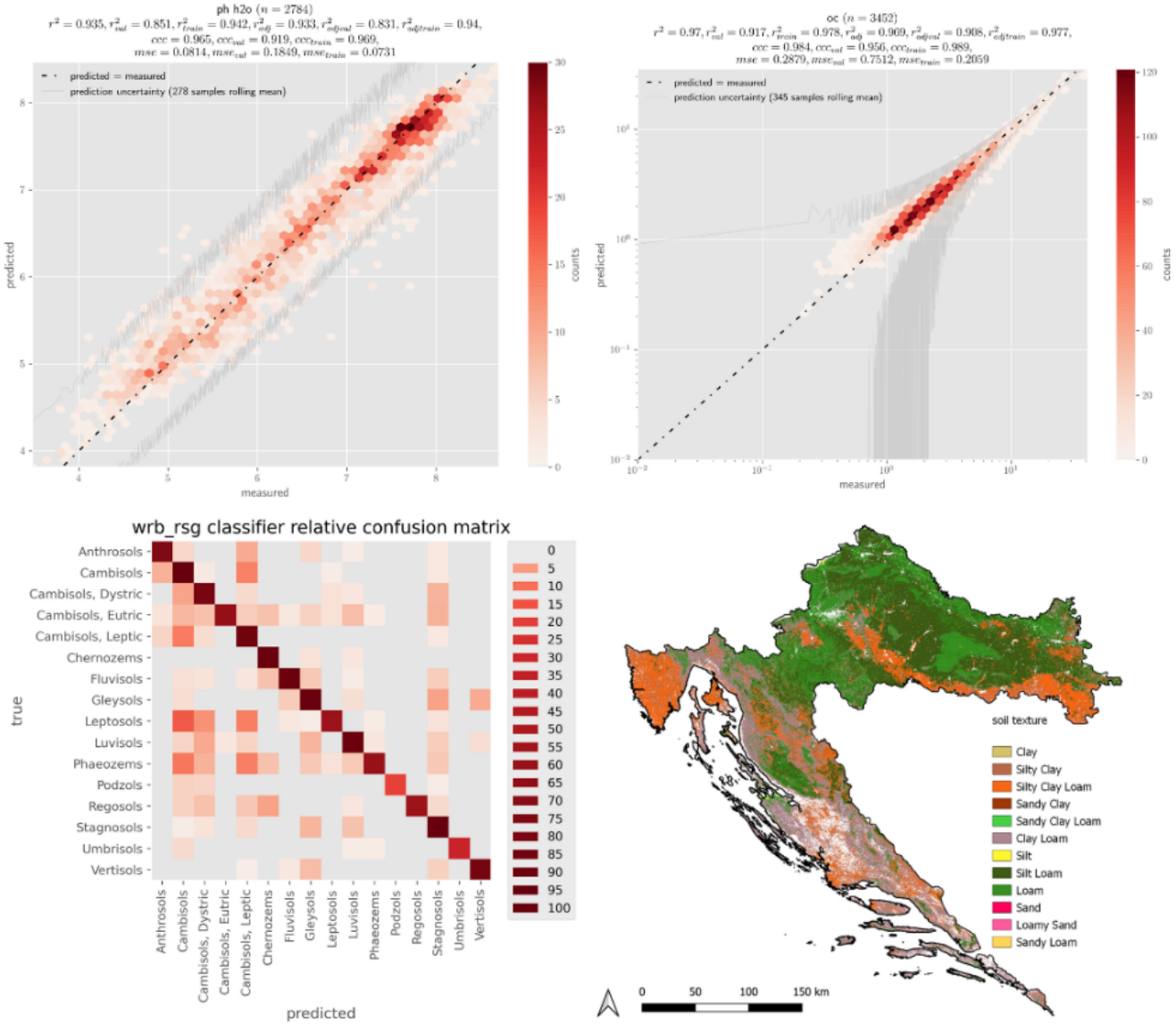

Figure 2. Soil pH prediction scatter and metrics (top left), organic carbon content prediction scatter and metrics - values unpacked from modeled logarithmic target (top right), soil taxonomy prediction relative confusion matrix (bottom left), soil texture classification map (bottom right).

the higher amount of available soil samples over such areas of interest. Additionally, there is notable temporal mismatch between the target data (collected from 1966 to 2016) and covariates (spanning 2000 onward) even in the context of producing a single, current-state dataset. While this was unavoidable due to availability of public soil sampling data in Croatia and the goal of the work to describe the current state of soils nationally, it is reasonable to assume that this affects the overall quality of the results. Further improvement, however, in contexts of both producing a current-state dataset and incorporating temporal dynamics, might come from utilising public soil samples collected in pedological and climatological conditions similar to those in Croatia. Nevertheless, we believe that this approach, which yielded the first public Croatian soil dataset in high spatial resolution, represents a step towards building national capacity for increasingly challenging environmental management.

4 Annex 1: Source code and data products

Source code used to produce the described layers is publicly available at: https://gitlab.com/jkrizan/dsmcroatia, all of the described data products are publicly available at: https://zenodo.org/records/10065971, and a small map portal was built for quick overview of the data products and made available at: http://hrsoil.multione.hr.

5 References

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794, New York, NY, USA: ACM.

Croatian Geological Institute, 2009. Geological map of the Republic of Croatia M 1:300 000, Croatian Geological Institute, Department of Geology, Zagreb.

Croatian Waters, 2016. Monitoring coordination programme, Zagreb, HR.

GRASS Development Team, 2017. Geographic Resources Analysis Support System (GRASS) Software, Version 7.2., Open Source Geospatial Foundation, https://grass.osgeo.org.

Hengl, T., MacMillan, R.A., 2019. Predictive Soil Mapping with R. OpenGeoHub foundation, Wageningen, NL, 370 pages, www.soilmapper.org.

Lindsay J.B., 2016. Whitebox GAT: A case study in geomorphometric analysis, Computers & Geosciences 95, 75-84.

Martinović, J., Vranković, A. (Editors), 1997. Database of Croatian soils, I. State Directorate for Environment and Nature protection, Zagreb, HR.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python, JMLR, 12, 2825-2830.

Pekel, J. F., Cottam, A., Gorelick, N., Belward, A. S., 2016. High-resolution mapping of global surface water and its long-term changes, Nature 540 (7633), 418-422.

Pflugmacher D, Rabe A, Peters M, Hostert P., 2019. Mapping pan-European land cover using Landsat spectral-temporal metrics and the European LUCAS survey, Remote Sensing of Environment 221, 583-595.

Radočaj, D., Jurišić, M., Zebec, V., Plaščak, I., 2020. Delineation of Soil Texture Suitability Zones for Soybean Cultivation: A Case Study in Continental Croatia, Agronomy 10, 823.

Romić, M. (Project leader), 2013. Spatial variability of trace and toxic metals in agricultural soils of Croatia, Faculty of Agriculture, Zagreb, HR.

Wilson A.M., Jetz W., 2016. Remotely Sensed High-Resolution Global Cloud Dynamics for Predicting Ecosystem and Biodiversity Distributions, PLoS Biol 14 (3), e1002415.

URL-1: Change in soil carbon stocks and calculation of soil total nitrogen and organic carbon trends and C:N ratios, https://envi-metapodaci.azo.hr/geonetwork/srv/hrv/catalog.search#/metadata/c7980264-97af-41ec-8853-35c7de2f55e6, (Accessed 19 May, 2024).

URL-2: Digital terrain model - Republic of Croatia, National geodetic administration, https://dgu.gov.hr/proizvodi-i-usluge/podaci-topografske-izmjere/digitalni-model-reljefa/180, (Accessed 19 May, 2024).

URL-3: Surface Soil Moisture 2014-present (raster 1 km), Europe, daily - version 1, https://land.copernicus.eu/global/products/ssm, (Accessed 19 May, 2024).

URL-4: Climatologies at high resolution for the earth's land surface areas, https://chelsa-climate.org/, (Accessed 19 May, 2024).

URL-5: Long-term MODIS LST day-time and night-time temperatures, sd and differences at 1 km based on the 2000–2017 time series, https://zenodo.org/records/1435938#.X3Lxm2gzYUE, (Accessed 19 May, 2024).

URL-6: Aerosol Optical Depth - NASA Earth Observatory, https://earthobservatory.nasa.gov/global-maps/MODAL2_M_AER_OD, (Accessed 19 May, 2024).

URL-7: Monthly precipitation in mm at 1 km resolution based on SM2RAIN-ASCAT 2007-2018, IMERGE, CHELSA Climate and WorldClim, https://zenodo.org/records/3256275#.X3L1NmgzYUE, (Accessed 19 May, 2024).

URL-8: Water Vapor - NASA Earth Observatory, https://earthobservatory.nasa.gov/ global-maps/MYDAL2_M_SKY_WV, (Accessed 19 May, 2024).

URL-9: Open Environmental Data Cube Europe, https://ecodatacube.eu, (Accessed 19 May, 2024).