OPTAIN

Optimal Strategies to Retain Water and Nutrients

D3.2: Solutions to overcome data scarcity

Authors: Brigitta Szabó (ATK), János Mészáros (ATK), Piroska Kassai (ATK), Péter Braun (ATK), Attila Nemes (NIBIO), Csilla Farkas (NIBIO), Natalja čerkasova (KU)

Co-Authors: Federica Monaco (UMIL), Enrico Antonio Chiaradia (UMIL), Felix Witing (UFZ)

Delivery Date: 27 June 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant agreement No. 862756.

Disclaimer

This document reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

Intellectual Property Rights

This document contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation, or both.

Project Consortium

Document Information

Program	EU Horizon 2020 Research and Innovation Action H2020-EU.3.2.1.1 (SFS-23-2019)
Grant agreement No.	862756
Project acronym	OPTAIN
Project full name	Optimal strategies to retain and re-use water and nutrients in small agricultural catchments across different soil-climatic regions in Europe
Start of the project	September 2020
Duration	60 months
Project coordination	Prof. Dr. Martin Volk Helmholtz-Centre for Environmental Research GmbH - UFZ www.optain.eu
Deliverable	D3.2: Solutions to overcome data scarcity Approaches and workflow to derive missing input data for the environmental and socio-economic models.
Work package	WP3: Retrieval of modelling data and solutions to overcome data scarcity
Task	Task 3.3: Overcoming data scarcity
Lead beneficiary	Agrártudományi Kutatóközpont
Author(s)	Brigitta Szabó (ATK, partner no. 4), János Mészáros (ATK, partner no. 4), Piroska Kassai (ATK, partner no. 4), Péter Braun (ATK, partner no. 4) Attila Nemes (NIBIO, partner no. 14), Csilla Farkas (NIBIO, partner no. 14), Natalja Čerkasova (KU, partner no. 11), Federica Monaco (UMIL), Enrico Antonio Chiaradia (UMIL), Felix Witing (UFZ, partner no. 1)
Contributor(s)	Michael Strauch (UFZ), Christoph Schürz (UFZ), Tatenda Lemann (UBERN), Annelie Holzkaemper (WBF), Joana Eichenberger (UBERN), Ágota Horel (ATK), Györgyi Gelybó (ATK), Péter Molnár (OVF), Marek Giełczewski (WULS), Wieslawa Kasperska-Wolowicz (ITP), Miha Curk (UL), Luka Honzak (UL), Gregor Kramberger (KGZ MB), Marie Anne Eurie Forio (UGent), Peter Goethals (UGent), Jovita Mėžinė (KU), Rasa Idzelytė (KU), Claudio Gandolfi (UMIL), Amro Sayed Ramadan Mohamed Negm (UMIL), Dominika Krzeminska (NIBIO), Alexander Engebretsen (NIBIO), Petr Fučík (VUMOP), Veronika Sítková (VUMOP), Štěpán Marval (VUMPO), Arturs Skute (DU), Juris Soms (DU), Martyn Futter (SLU), Emma Lannergard (SLU)
Quality check	Felix Witing (UFZ), Martin Volk (UFZ)
Planned delivery date	Month 22 (June 2022)
Actual delivery date	27/06/2022
Citation	Szabó, B., Mészáros, J., Kassai, P., Braun, P., Nemes, A., Farkas, Cs., Čerkasova, N., Monaco, F., Chiaradia, E. A., Witing, F. (2022): *Solutions to overcome data scarcity.*

	Deliverable D3.2 EU Horizon 2020 OPTAIN Project, Grant agreement No. 862756
Dissemination level*	PU

*PU = Public; PP = Restricted to other program participants (including the Commission Services; CO = Confidential, only for members of the consortium (including the Commission Services).

Deliverable status

Version	Date	Author(s)/Contributor(s)	Notes
0.1	31.05.2021	Brigitta Szabó (ATK), János Mészáros (ATK), Piroska Kassai (ATK), Attila Nemes (NIBIO), Natalja Čerkasova (KU)	First complete draft
0.2	08.06.2022	Felix Witing (UFZ), Martin Volk (UFZ)	Revision and first quality check
0.3	20.06.2022	Brigitta Szabó (ATK)	Revision
0.4	23.06.2022	Felix Witing (UFZ), Martin Volk (UFZ)	Second quality check
1.0	23.06.2021	Brigitta Szabó (ATK)	Final

Summary

An important aim of the OPTAIN project is to derive missing information on necessary model input variables in a harmonized way to allow for a sound cross-case study assessment of NSWRM effectiveness. Therefore, in this report we provide approaches applicable for all OPTAIN case studies (CS) to fill data gaps.

The specific objective of OPTAINs task 3.3 was to provide methods to cover missing input data that is required for the environmental modelling and socio-economic analysis. The deliverable includes guidelines with detailed explanations about the derivation of missing data for the CS leaders.

Based on the information provided by CS leaders in the OPTAIN milestone "MS7 Data inventory of input data for integrated modelling collected from all case studies", the following information had to be covered by approaches provided by WP3 to fulfil the input requirements of the models and analysis: 1) soil phosphorus content, 2) effective bulk density, 3) moist soil albedo of the top layer, 4) USLE soil erodibility (K) factor, 5) available water capacity, 6) saturated hydraulic conductivity, 7) time series crop data.

The mapping of soil phosphorus content is based on the LUCAS topsoil dataset. During the mapping the geometric mean phosphorus content by land use types – characteristic for the region of the CS – is applied. Further required data are the LUCAS Land Use / Cover Area Frame Survey, European agro-climate zone map and the land use or land cover map of the CS – a local one, if available.

For the calculation of soil physical and hydraulic properties we apply methods available from the literature.

The derivation of crop maps is based on remote sensing data. A crop classification model was trained on the cropland data of the LUCAS Land Use / Cover Area Frame Survey of the years 2015 and 2018, merged with the Sentinel-1A and -1B satellite radar images. The pixel based crop classification was carried out with a random forest algorithm on the Google Earth Engine platform. The method can be applied for 2015 and all following years. By adding a map of field boundaries, the pixel based crop prediction can be aggregated to field level using the majority of the predicted crop.

Regarding the socio-economic data, missing information is planned to be covered from official statistics. The EU database does not account properly for the Norwegian and Swiss sites, therefore required data will be retrieved ex novo from local sources or literature.

<u>Related deliverables:</u>

The content of this deliverable D3.2 is strongly related to the upcoming deliverable D3.3 ('Created data pre-processors successfully applied for input data restructuring', month 24).

Table of Contents

Abbreviations

AIC	Akaike Information Criterion
AL	Acid ammonium acetate Lactate extraction
AWC	Available Water Capacity
BD	Bulk Density
CS	Case Study
FC	Field Capacity
GEE	Google Earth Engine
HSG	Hydrologic Soil Group
IMD	Integral Mean Deviation
KS	Saturated hydraulic conductivity
LUCAS	Land Use/Land Cover Area Frame Survey
MRE	Mean Relative Error
NSWRM	Natural/Small Water Retention Measures
P	Phosphorus
PSD	soil Particle-Size Distribution
PTF	Pedotransfer Function
RMSE	Root Mean Square Error
SSE	Sum of Squared Error
SWAT+	Soil and Water Assessment Tool
USLE	Universal Soil Loss Equation
WLP	Wilting Point
WP	Work Package

1. Introduction

One of the aims in the OPTAIN project is to assess the effectiveness of Natural/Small Water Retention Measures (NSWRMs) by using environmental models (tasks 4.3, 4.4) and to analyse the influence of NSWRM implementation from a socio-economic point of view (task 4.5).

For the environmental and socio-economic models the following data is required: environmental parameters (topography, land cover and land use, hydromorphology, soil properties), calibration data (discharge, water quality, sediment, crop yield, soil water content, soil temperature, drainage, surface runoff), agricultural management data (crop rotation, soil tillage, fertilization, irrigation, tile drainage, etc.), water use and management (reservoirs, point sources, water withdrawals), location and properties of existing NSWRMs, their implementation and maintenance costs, economic metrics (costs related to agricultural production, subsidies, natural resources).

An important objective is to use as much locally available information for each case study (CS) as possible for the analysis, because the performance of the models are influenced by the accuracy of the input data as well. In work package 3 (WP3) we focused on the derivation of those input data, which can be retrieved with an accuracy that is appropriate for the modelling tasks.

2. Background

In OPTAIN, we focus on analysing the effectiveness of NSWRMs in retaining water and nutrients, therefore it is important to have field-based input data regarding the soil and crop information both for the catchment and field-scale models. The collection of these data can be hampered at the catchment scale due to missing measured information or no access to sensitive data for some part or the entire target area.

The SWAT+ model requires the soil properties for the hydrological response units (HRUs) defined based on slope, land use and soil maps. These soil properties are soil layering, maximum rooting depth, information on soil cracking, effective bulk density, available water capacity, saturated hydraulic conductivity, organic carbon content, sand, silt and clay content, rock fragment content, moist soil albedo, Universal Soil Loss Equation (USLE) soil erodibility factor, hydrologic soil group, and nutrient content of the surface soil layer. Most of the basic soil properties – e.g. soil organic carbon content, particle size distribution – are usually locally available, but information on soil hydraulic properties are often missing. The soil hydrophysical parameters can be computed with pedotransfer functions (PTFs), which are widely-used indirect techniques enabling the soil properties to be predicted by using easily-retrievable basic soil information. An alternative can be the use of open access international data if local or national information is lacking.

Regarding the nutrient content of the soil surface, the SWAT+ model can use information on the initial soil nutrient content as input (nutrient.sol). This input is optional for the model, else it uses default values if intital soil nutrient information are not provided (Arnold *et al.*, 2012). Since we aim at analysing the effectiveness of NSWRMs in retaining nutrients, it is important to derive approximate soil nutrient maps for the case studies (CSs) instead of using the model's default values.

In SWAT+ information on crop types is required to simulate plant growth (PLANTS.PLT), set general land use properties (LANDUSE.LUM), characteristics of crop rotation (PLANT.INI), management operations (MANAGEMENT.SCH), SCS curve number (CNTABLE.LUM) and USLE P value (HRU-LTE.HRU). The main generic land use/ land cover categories distinguished by the model are agricultural land (arable land), orchard, forest, wetlands, hay, pasture, range and water.

In order to meet project's aim to analyse the efficiency of NSWRMs, we need to further specify agricultural land categories (AGRL, AGRR, AGRC) – defined in the SWAT documentation (Neitsch *et al.*, 2009: [link]) – into field-level time series of changing crop types for each CSs for the modelling period. This data is not always available for an entire agricultural catchment and the whole modelling period. It can also happen that information on crops are only available for agricultural field blocks – which can include several fields – but not for single fields, which can hamper the analysis of nutrient and water retention in case of a field routing based simulation.

A promising method is to derive missing information on crop types from remote sensing data. Earth observation based crop mapping has been developed in many studies, such as by Phalke & Özdoğan (2018) and Teluguntla *et al.* (2018) who mapped based on Landsat data, Immitzer *et al.* (2016), Belgiu & Csillik (2018), Gumma *et al.* (2020) and Defourny *et al.* (2019) that used Sentinel-2 data input, Kenduiywo *et al.* (2018) and d'Andrimont *et al.* (2021) who considered Sentinel-1 data. Orynbaikyzy *et al.* (2019) reviewed those crop mapping approaches, which were based on both optical and radar data. They found that most frequently cereals, oilseed and sugar crops are targeted during the identification.

For the field-level crop mapping an appropriate image resolution as well as a short revisit period of the satellite sensor data is important. It is also preferable if the sensor has negligible dependency on atmospheric conditions, which can be obtained by using e.g. S1A and S1B polar-orbiting Synthetic Aperture Radar imagery (d'Andrimont *et al.*, 2021)in 10 m resolution.

For the field-scale modelling, the required input parameters are less demanding, therefore less data types had to be covered by indirect or open access datasets. Most often the soil hydraulic properties are the ones which are missing. Those can be derived with the above mentioned pedotransfer functions.

The socio-economic data is required to identify the factors which influence farmers to apply NSWRMs. This input will be used to estimate the impact of NSWRM implementation on e.g. agricultural gross margins at farm and catchment scale. For the evaluation of the economic sustainability of NSWRMs it is indispensable to collect socio-economic data for each CS. The data will be used for cost-benefit analysis, assessment of long-term benefits for farmers and non-monetary values related to environmental concerns.

3. Methods

In cooperation with WP4 a data inventory has been prepared with the following aims:

i. to overview what data is available for the catchment-scale and field-scale models,
ii. to see what data is missing,

iii. to decide which missing data is feasible to cover.

In the inventory detailed information was included about the required inputs of the catchment and field-scale model and was filled out by the CS leaders for the case study sites. The inventory was used to identify which input data is missing by CS and which of those were feasible to cover by prediction approaches or from open access datasets.

3.1. Soil properties

3.1.1. Physical and hydrological parameters

In the case of soil properties both open access datasets and prediction methods were considered to cover missing local data. For soil hydraulic properties it is recommended to compute them from the parameters of the van Genuchten model (van Genuchten, 1980). This way the parameters will be self-consistent, and rely on dynamic criterion based on soil internal drainage dynamics (Assouline & Or, 2014; Nasta *et al.*, 2021). The most frequently used pedotransfer functions, which can be used to predict soil water content and hydraulic conductivity from easily available soil information, were listed (Table 1 and 2) and tested on European datasets: in GRIZZLY, HYPRES and EU-HYDI datasets (Nasta *et al.*, 2021). The prediction performance of the selected PTFs were analysed by comparing measured and estimated soil water retention values – at 30 prescribed matric head values – and saturated hydraulic conductivity.

Table 1: List of eleven PTFs tested in three European databases for the prediction of soil water content.

PTF	Reference	Location of training data	Type of WRC model *	Type of the PTF**	Input variables of selected PTFs						
					sand	silt	clay	ρ_b	OM	soil depth	topsoil
					%	%	%	g cm^{-3}	%	cm	
SAX86	Saxton et al. (1986)	USA	-	pseudo-continuous, LR	+		+				
C&S92	Campbell and Shiozawa (1992)	Washington (USA)	BC	parametric, LR		+	+	+			
R&B85	Rawls and Brakensiek (1985)	USA	BC	parametric, LR	+		+	+			
O&C80	Oosterveld and Chang (1980)	Alberta, (Canada)	-	pseudo-continuous, LR	+		+	+		+	
WOS99	Wösten et al. (1999)	Europe	VG	parametric, LR		+	+	+	+		+
VER89	Vereecken et al. (1989)	Belgium	VG	parametric, LR	+		+	+	+		
euptfv2	Szabó et al. (2021)	Europe	VG	parametric, RF	+	+	+	+/-	+/-	+	
WEY09	Weynants et al. (2009)	Belgium	VG	parametric, LR	+		+	+	+		
ROSETTA	Schaap et al. (2001)	Europe and North America	VG	parametric, NN	+	+	+	+			
T&H98	Tomasella and Hodnett (1998)	Amazonia, Brazil	-	point (9 prescribed ψ-values), LR		+	+		+		
RAW82	Rawls et al. (1982)	USA	-	point (12 prescribed ψ-values), LR	+	+	+	+	+		

+/- denotes optional; *Type of model for the description of the water retention curve: BC: Brooks and Corey (1964) model, VG: van Genuchten (1980) model; **LR: linear regression, RF: random forest, NN: neural network, pseudo-continuous PTF: uses matric head as input variable, this way predicts the soil water content at any matric head value without applying any WRC models (Haghverdi et al., 2012).

Table 2: List of ten PTFs tested in three European databases for estimating hydraulic conductivity.

PTF	Reference	Location of training data	Type of PTF	Input variables of selected PTFs						
				sand	silt	clay	ρ_b	OM	soil depth	topsoil
				%	%	%	g cm^{-3}	%	cm	
WOS99	Wosten et al. (1999)	Europe	LR		+	+	+	+		+
euptfv2	Szabó et al. (2021)	Europe	RF	+	+	+	+/-	+/-	+	
ROSETTA	Schaap et al. (2001)	Europe and North America	NN	+	+	+	+			
A&G19	Araya and Ghezzehei (2019)	USA	BRT	+	+	+	+/-	+/-		
GUP20	Gupta et al. (2021)	global	LR	+		+	+			
COS84	Cosby et al. (1984)	USA	LR	+		+				
S&R06	Saxton and Rawls (2006)	USA	LR	+		+		+		
VER89 & GUA07	Vereecken et al. (1989), Guarracino (2007)	Belgium, USA	LR	+		+	+	+		
R&B85 & NAS13	Rawls and Brakensiek (1985), Nasta et al. (2013)	USA, Europe	LR	+		+	+			
WEY09 & GUA07	Weynants al. (2009), Guarracino (2007)	Belgium, USA	LR	+		+	+	+		

+/- denotes optional; *LR: linear regression, RF: random forest, NN: neural network, BRT: boosted regression tree

3.1.2. Soil nutrients

The availability of data on the soil nutrient content varies among in the CSs. Some of them have measured data at field-scale, and others lack this information. SWAT+ requires spatial data that completely cover the CS areas, thus the input soil nutrient maps must not contain missing values. As the LUCAS Topsoil Survey dataset (Tóth *et al.*, 2013a) contains measured Olsen-P content data of about 20,000 soil samples in Europe, we decided to use it to derive approximate maps of soil P content for data scarce areas. This way, a common method is provided for producing Olsen-P maps for any of the OPTAIN CSs based on the LUCAS dataset. For the prediction of the phosphorus content of the surface soil layer the geometric mean Olsen phosphorus values were calculated by land use/land cover categories using the LUCAS Topsoil Survey dataset. If a local measured dataset is available for some areas or fields of the catchment, those should replace the mean values. This method requires the delineation of land use/ land cover categories of the land use map available for the CS and land use/ land cover categories available in the LUCAS dataset.

The results of soil P mapping was analysed at the Felső-Válicka CS (CS3b). For the CS3b, measured P determined by the acid ammonium acetate lactate extraction method (AL) were available for some fields in P_2O_5. First it was given in AL-P, then the AL-P values were converted into Olsen-P with the equation published by Sárdi *et al.*, (2009).

3.2. Land use and crop classification

For crop mapping a prediction approach has been derived, which can classify the crops based on the time series reflectance data of Sentinel-1A and -1B satellite radar images. The backscattering registration of a vertically transmitted microwave signal in a vertical (VV) and horizontal (VH) receiver and VH/VV ratio index were considered as predictors based on d'Andrimont *et al.* (2021). For deriving the crop classification algorithm we used the harmonised version of the Land Use / Cover Area frame statistical Survey (LUCAS) dataset (d'Andrimont *et al.*, 2020) as in-situ data. The training of the crop classification was performed on LUCAS from years 2015 and 2018, overlapped with the 2015 and 2018 Sentinel-1 images. We used the random forest algorithm to derive the prediction algorithm. This approach provides a pixel-based crop classification. For the analysis of the model performance a 10-fold cross-validation was applied.

We analysed the optimal length of the time window, for which image data was aggregated. The VV, VH and VH/VV ratio index were averaged for 2, 4, 6 and 8 weeks and used as predictors.

Field boundary maps are required to avoid the salt-and-pepper effect of the pixel-based prediction of the crop types. After the crop type prediction, the crop maps are finalized by assigning only one crop by one agricultural field. That crop type is kept, which has the majority in terms of area within the field.

3.3. Socio-economic data

The socio-economic data relate to:

- the financial features of individual measures (e.g., implementation costs),
- the possible implications on the economics of agriculture at farm- and catchment-level (e.g., farm gross margin, agricultural net value added), and
- the societal impacts expected from NSWRMs' implementation, owing to a better quality of natural resources or to a higher resilience to climate and environmental risks.

The list of potential socio-economic input data provided the following information: unit, format, if data type is static or time series, required time frequency, data availability and sources.

3.4. Evaluation of prediction approaches

The most common metrics used to quantify the predictive capability of the tested PTFs are: root mean square error (RMSE), which combines both bias and lack of precision, the coefficient of determination (R^2), which measures how well the data pairs fit a straight line, and the mean relative error (MRE), which quantifies the average under-estimation (if positive) or over-estimation (if negative). These statistical indicators are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_i^N (O_i - P_i)^2} \qquad [1]$$

$$R^2 = \frac{\sum_i^N (O_i - P_i)^2}{\sum_i^N (O_i - \overline{O_i})^2} \qquad [2]$$

$$MRE = \frac{1}{N}\sum_{i}^{N}\left(1-\frac{P_i}{O_i}\right)\cdot 100 \qquad [3]$$

where O, $\bar{O}$, and P are the observed, mean of observed, and predicted values of a variable, respectively. Subscript i is the counter and N is the highest number of counter-points.

The prediction performance of the soil water retention was evaluated also by using the integral mean deviation (IMD), which reveals biases in predicting the shape of the observed curve (IMD>0 means systematic underprediction):

$$IMD = \frac{1}{(\xi_u-\xi_l)}\int_{\xi_l}^{\xi_u}[\theta(\xi)_{OBS}-\theta(\xi)_{PTF}]d\xi \qquad [4]$$

where $\xi=\ln(\psi)$, with "ln" denoting the natural logarithm, whereas the subscripts u and l denote the upper and lower bounds, respectively. We set the prefixed lower and upper suction head values at 10^0 cm and $10^{4.2}$ cm, respectively to cover the water retention curve from saturation up to the wilting point.

The Akaike information criterion (AIC) estimates the amount of information lost by a PTF and quantifies the trade-off between goodness of fit and model parsimony (Diks and Vrugt, 2010). The AIC is computed as follows:

$$AIC = lnNS\ ln\left(\frac{SSE}{NS-1}\right) + lnNS + 2p \qquad [5]$$

where NS is the number of soil samples, p is the number of predictors employed in a PTF and SSE is the sum of squared errors, given by:

$$SSE = \sum_{i}^{NS}(O_i - P_i)^2 \qquad [6]$$

For an optimal prediction, RMSE, MRE, and IMD should be close to zero, R^2 values should be close to 1 whereas AIC should be as high as possible (Romano & Palladino, 2002). The RMSE is widely used in the body of scientific literature and generally spans between 0.020 and 0.12 $cm^3\ cm^{-3}$, expressing very high and very low performances, respectively (Cornelis *et al.*, 2001; Schaap *et al.*, 2001). Ideally, perfect performance is obtained when all individual MRE and IMDs are zero, while in reality underestimation and overestimation are obtained when MRE and IMD values are positive or negative, respectively. The AIC penalizes PTFs using a large number of predictors and indicates that model complexity is not always associated with high performance.

The accuracy of the crop classification is described with its 'total accuracy' [Eq. 7] and Kappa coefficient, which measures the agreement between classification and observed values. Perfect agreement between the observed and predicted value results Kappa value of 1.

$$Total\ accuracy = \frac{Number\ of\ correctly\ classified\ values}{Total\ number\ of\ values} \qquad [7]$$

By using the crop classification, we analyzed the relevance of predictors on the response variable determined by computing the variable importance based on the mean decrease in impurity (Hastie *et al.*, 2009).

4. Results and discussion

The milestone entitled "MS7 Data inventory of input data for integrated modelling collected from all case studies" summarizes the input data availability by CSs for the catchment and field-scale models. Based on the inventories the following data has been defined as needed to cover by WP3 with an approach which derives these parameters from easily available information:

- effective bulk density,
- moist soil albedo,
- USLE soil erodibility factor,
- soil hydraulic properties, such as available water capacity and saturated hydraulic conductivity,
- soil phosphorus content,
- crop maps.

For the assessment of the socio-economic impacts of NSWRMs the potential input data (including data availability and sources) have been listed.

Hereinafter we summarize the results of the approaches suggested to overcome data scarcity by type of missing data. More detailed information about the proposed workflows are provided in the Annex in the form of 4 specific guidelines.

4.1. Soil properties

4.1.1. Physical and hydrological parameters

In the case of soil physical and hydrological properties, methods available from the literature are suggested to be applied. For the effective (moist) bulk density the method of Wessolek *et al.* (2009) could be applied. For the moist albedo and USLE soil erodibility factor, the equations presented by Abbaspour *et al.*, (2019) could be used.

Table 3: Performance of PTFs predicting soil water content at 30 prescribed matric head values in GRIZZLY (660 soil samples), HYPRES (522 soil samples) and EU-HYDI (4940 soil samples). NS: total number of soil samples; p=number of predictors; RMSE: root mean square error; R^2: coefficient of determination; IMD_a: average integral mean deviation; AIC: Akaike information criterion. Text in italic indicates when a part of or the entire data was used to train the PTF.

		GRIZZLY (NS=660)				HYPRES (NS=522)				EU-HYDI (NS=4940)			
PTF	*p*	RMSE	R^2	IMD_a	AIC	RMSE	R^2	IMD_a	AIC	RMSE	R^2	IMD_a	AIC
		$cm^3\ cm^{-3}$		$cm^3\ cm^{-3}$		$cm^3\ cm^{-3}$		$cm^3\ cm^{-3}$		$cm^3\ cm^{-3}$		$cm^3\ cm^{-3}$	
SAX86	2	0.111	0.72	0.470	-18.0	0.146	0.57	0.457	-13.8	0.120	0.57	0.4528	-23.6
C&S92	3	0.148	0.43		-12.3	0.146	0.44		-11.8	0.059	0.51		-33.6
R&B85	3	0.071	0.87	0.333	-21.9	0.083	0.84	0.296	-18.9	0.059	0.76	0.3275	-33.6
O&C80	4	0.146	0.52	0.472	-10.5	0.174	0.37	0.465	-7.6	0.146	0.43	0.4502	-16.3
WOS99	5	0.056	0.91	0.048	-20.9	*0.066*	*0.89*	*0.040*	*-17.7*	0.060	0.85	0.0082	-29.3
VER89	4	0.088	0.81	0.336	-17.1	0.094	0.81	0.333	-15.3	0.080	0.78	0.3774	-26.4
euptfv2	6	0.079	0.80	0.080	-14.4	0.071	0.87	0.058	-14.9	*0.045*	*0.91*	*0.0207*	*-32.2*
WEY09	4	0.046	0.91	0.002	-25.4	0.061	0.88	-0.003	-20.7	0.069	0.80	-0.0332	-29.0
ROSETTA	4	0.102	0.80	0.135	-15.2	0.104	0.81	0.118	-14.0	0.094	0.75	0.0983	-23.8
T&H98	3	0.082	0.81		-20.0	0.126	0.63		-13.7	0.110	0.61		-23.1
RAW82	5	0.067	0.84		-18.6	0.077	0.79		-15.9	0.082	0.68		-24.0

Table 3 and 4 show the performance of the PTFs tested for the prediction of soil water content using the GRIZZLY, HYPRES and EU-HYDI datasets (Nasta *et al.*, 2021). The most recent vG-based PTFs (WEY09, WOS99, euptfv2) proved to be accurate enough for predicting the water retention function. ROSETTA, VER89, SAX86, T&H98, and R&B85 have acceptable performance and can be theoretically improved in some textural classes although they were calibrated and validated outside Europe under different climatic and environmental conditions.

The ten PTFs used to predict saturated hydraulic conductivity demonstrate a generally poor performance with uncertainties ranging over one or two orders of magnitude (Table 4). The main predictors (soil bulk density, organic carbon content, and texture) are cross-correlated and might play contrasting roles in predicting saturated hydraulic conductivity especially because of the lack of its standardized measurement method.

Table 4: Performance of ten PTFs predicting $\log_{10}(K_s)$ in GRIZZLY (62 soil samples), HYPRES (253 soil samples), and EU-HYDI (1811 soil samples). NS: total number of soil samples; p=number of predictors; RMSE: root mean square error; R^2: coefficient of determination; AIC: Akaike information criterion. Text in italic indicates when a part of or the entire data was used to train the PTF.

		GRIZZLY (NS=62)			HYPRES (NS=253)			EU-HYDI (NS=1811)		
PTF	*p*	RMSE cm day^{-1}	R^2	AIC	RMSE cm day^{-1}	R^2	AIC	RMSE cm day^{-1}	R^2	AIC
WOS99	5	1.350	0.28	16.67	*1.666*	*0.04*	*21.20*	1.686	0.078	25.34
euptfv2	6	1.347	0.02	18.65	1.120	0.14	18.81	*0.793*	*0.711*	*16.02*
ROSETTA	4	1.625	0.24	16.20	1.918	0.02	20.76	1.768	0.064	24.06
A&G19	5	1.145	0.24	15.31	1.732	0.13	21.63	1.685	0.019	25.33
GUP20	3	1.409	0.36	13.03	1.851	0.03	18.37	1.769	0.014	22.07
COS84	2	1.329	0.40	10.54	1.803	0.13	16.08	1.787	0.035	20.21
S&R06	3	1.437	0.26	13.19	1.891	0.07	18.60	1.811	0.006	22.42
VER89 & GUA07	4	2.574	0.18	20.00	1.814	0.08	20.15	2.229	0.067	27.53
R&B85 & NAS13	3	1.286	0.39	12.27	2.144	0.11	20.00	1.943	0.063	23.47
WEY09 & GUA07	4	1.057	0.27	12.65	1.404	0.15	17.31	1.458	0.037	21.16

If soil properties, which are expected to be available – such as soil organic carbon content, sand, silt and clay content, rock fragments content, bulk density – are missing, the use of SoilGrids dataset (Poggio *et al.*, 2021) is recommended to use, which is freely available from https://soilgrids.org/.

Different countries/institutes measure soil particle-size distribution (PSD) using different methods and recognizing different classification standards. In the SWAT+ model. The sand, silt and clay content are defined according to the USDA system, which means that the particle size limit is < 0.002 mm for the clay, 0.002-0.05 mm for the silt and 0.05-2 mm for the sand fraction. When conversions between different methods are required it is advised to apply k-nearest neighbour interpolation (formerly called: 'similarity technique'), which results in less uncertainty, lesser bias and shrinkage of resulting texture range compared to the simpler loglinear interpolation (Nemes *et al.*, 1999).

More detailed information about the proposed workflows to derive soil physical properties are included in Guideline 1 ("Derivation of soil physical and hydraulic properties") of the Annex. The related R scripts and supporting files, which enhance the calculations, are available on the public ZENODO repository of OPTAIN (https://doi.org/10.5281/zenodo.6656454).

4.1.2. Soil nutrients

For the Felső-Válicka CS, all Hungarian LUCAS samples were considered to compute the mean soil P content values by land cover/ land use categories. The 65% of the selected samples were from arable land, 20% from pasture, 13% from forest and 1% from permanent crops. The other land cover/ land use categories were less represented and typically not fertilized ones, therefore for those categories the mean P content computed on the whole LUCAS topsoil dataset was used.

Figure 1a and b show the Olsen-P content map for CS3b without and with measured soil P content values.

a) b)

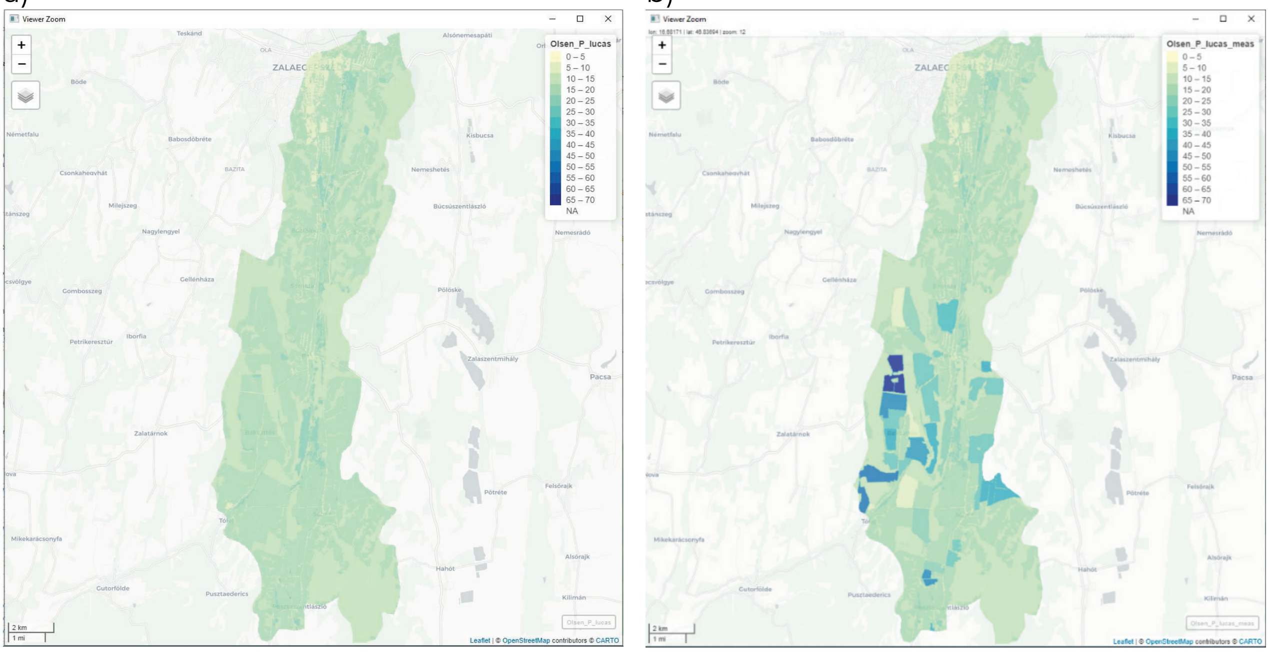


Figure 1: Derived soil Olsen-P content (ppm) map for the Felső-Válicka case study (CS3b) based on a) LUCAS Topsoil data and local land use map (20 m resolution) and b) further including locally measured soil P content.

For CS3b the mean value for arable land that was computed based on the Hungarian LUCAS samples was 20 ppm. When locally measured values were added from year 2009, the P content varied between 8 and 65 ppm. The mean error was -4.11 ppm for the arable land category.

The nitrate content is highly variable in space and time and the dynamic of its amount is significantly influenced by nitrogen fertilization (Zhu *et al.*, 2021). It is thus recommended to use the default value of the SWAT model in the case of missing data and use locally available information on nitrogen fertilization in the management table of SWAT+.

Guideline 2 ('Preparation of soil phosphorus content maps for SWAT+ modelling') of the Annex includes detailed information on how to derive soil P content maps. The related

R scripts (https://doi.org/10.5281/zenodo.6652572) have been uploaded to OPTAINs public repository on ZENODO.

4.2. Land use and crop classification

If local or national land use data is not available, the CORINE dataset can be used, which is accessible from https://land.copernicus.eu/pan-european/corine-land-cover.

For the crop classification the analysis of the length of the period used for the training resulted that the starting week needs to be set to 4 and the ending week to 40 in the case of CS4, CS8 and CS12. Setting the starting at week 4 and ending week at 44 for all other CS resulted in the most accurate crop type predictions. The optimal time window for image data aggregation was 4 weeks.

Based on the 10-fold cross validation, the mean total accuracy of applying the crop classification method was between 0.629 and 0.769 and the Cohen's kappa varied from 0.480 and 0.642, when no further local training data was added. The highest accuracy could be achieved for the CS7 - La Wimbe (BE), the lowest was found on CS4 - Upper Zglowiaczka (PL). The mean accuracy and kappa values by CSs is shown on Figure 2. For CS2 and CS10 no crop classification model could be derived because LUCAS dataset is not available for Switzerland and Norway.

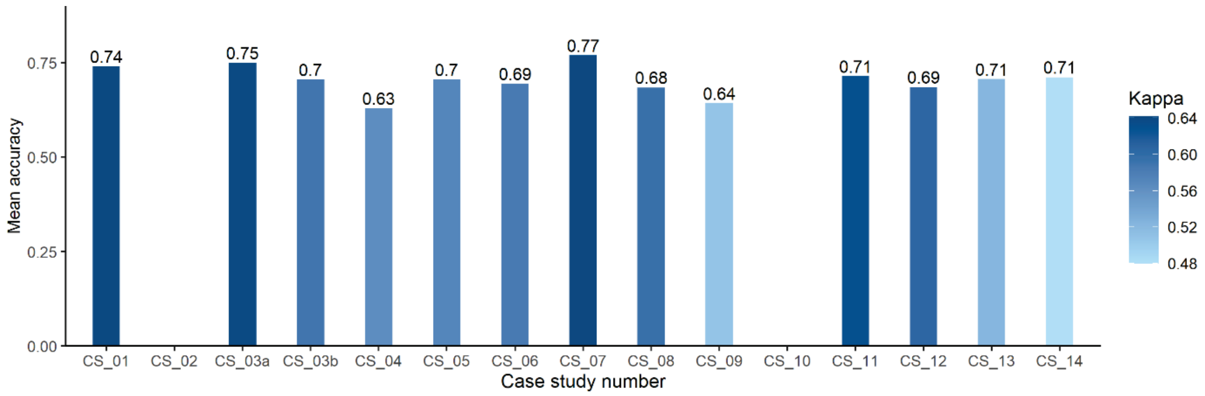


Figure 2: Mean accuracy and kappa values of the crop classification by case studies (CS). For CS2 and CS10 no crop classification model could be derived because LUCAS dataset is not available for Switzerland and Norway.

In case studies with a high variability in the crops grown, the performance of the prediction is lower than for areas with lower crop type diversity. During the classification crop types with a less number of samples in the training dataset, will have lower accuracy and are often mapped as grassland or wheat, such as for instance rye, and oats. Lucerne is often misclassified as grassland. There is a high degree of misclassification between barley and wheat, prediction accuracy of barley is less than wheat. The variability of crop types is lower on the predicted maps than in the training dataset. The classification accuracy of rape and sunflower is the highest, which is in line with other studies (d'Andrimont *et al.*, 2021).

The performance of the prediction is also influenced by the degree of how well the selected LUCAS dataset represents the crops grown in the CS. Figure 3 and 4 show for each CS the number of samples available for the training by crop types. The frequency of the crop types depends on the specific conditions of the pedoclimatic zones. The five most frequent crops in the training datasets are the following:

- in the boreal region: common wheat, rape and turnip rape, barley, dry pulses and other leguminous and mixtures for fodder;
- in the continental biogeographical region: maize, common wheat, rape and turnip rape, barley and rye;
- in the pannonian region: maize, common wheat, sunflower, barley and rape and turnip rape.

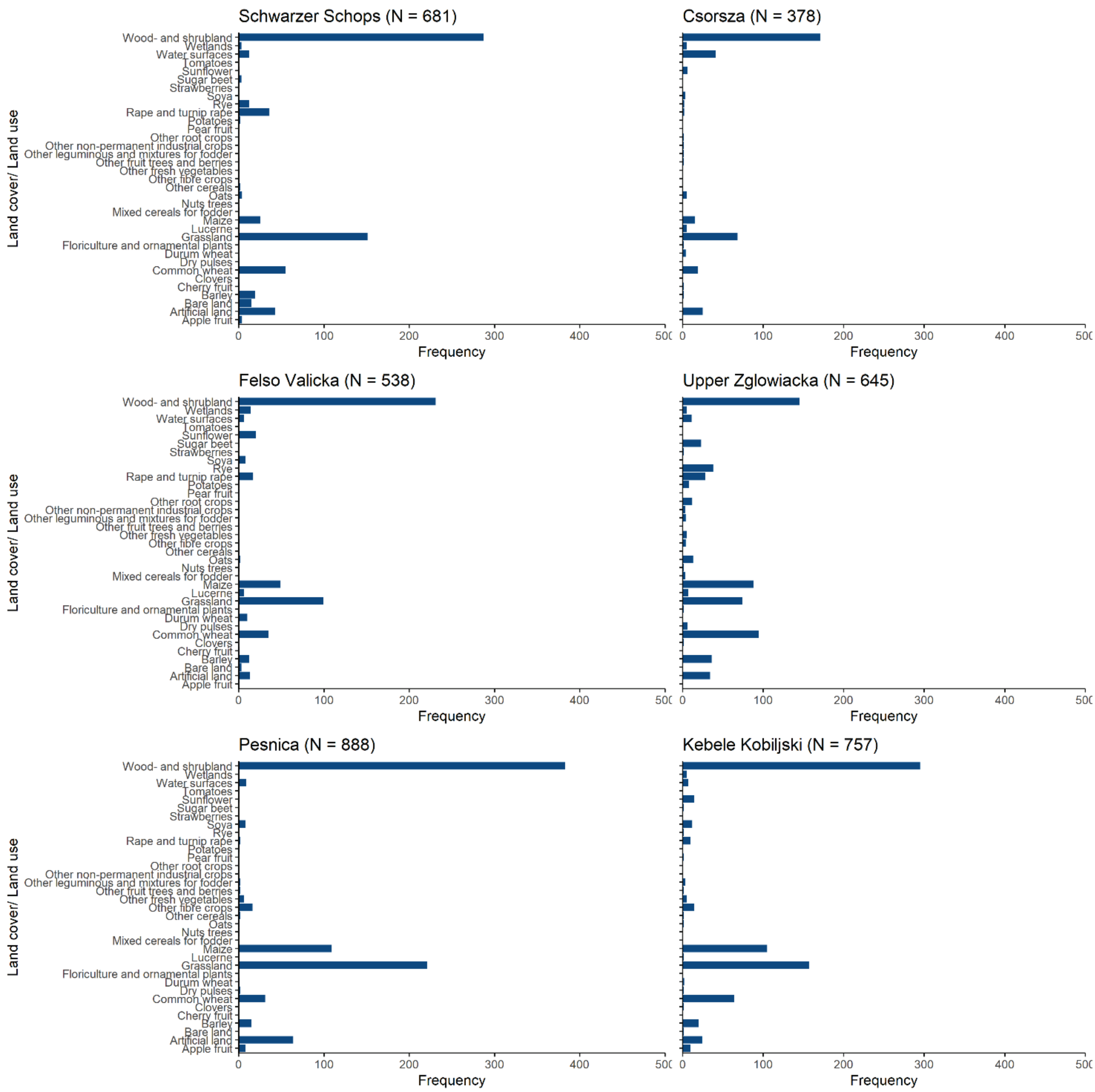


Figure 3: Frequency of crop data available from the LUCAS dataset for the case study areas plus 30 km buffer zone to derive crop classification for the case studies No. 1, 3a, 3b, 4, 5, 6.

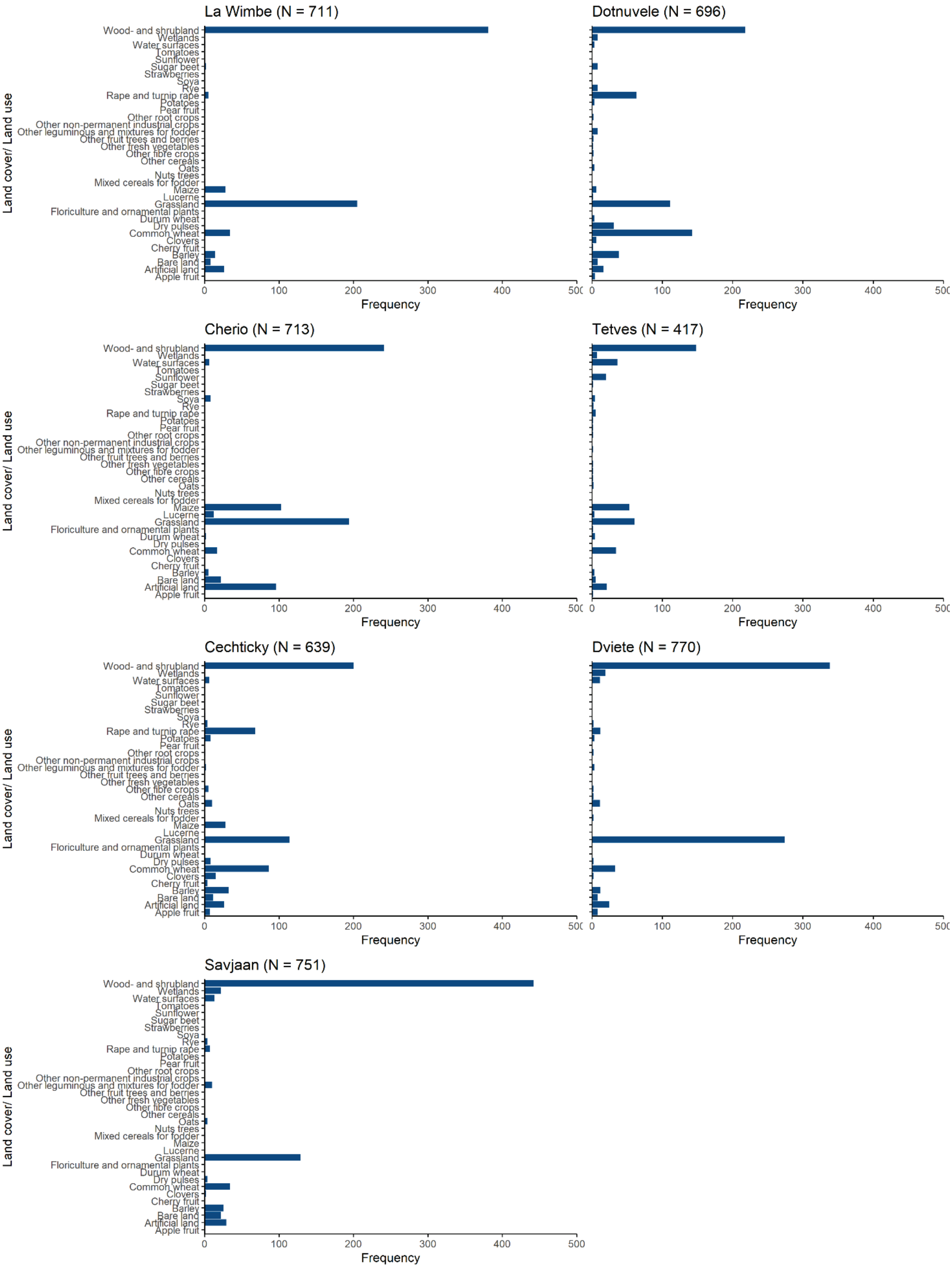


Figure 4: Frequency of crop data available from the LUCAS dataset for the case study areas plus 30 km buffer zone to derive crop classification for the case studies No. 7, 8, 9, 11, 12 13, 14.

Obviously, those crops, which are not included in the LUCAS dataset selected for a given CS will not be included on the map. If crops that are dominant at the area of the CS are not included in the training data, further local ground truth crop data has to be added to the selected LUCAS dataset, else that crop will not be represented in the derived crop maps. The addition of local data can significantly increase the performance of the crop classification, therefore it is highly recommended to use it if available.

Figures 5 and 6 show the variable importance by CS, none of the predictors have outstandingly higher importance than the others, spectral information from the entire vegetation period is required for the crop prediction, especially radar image from the middle of the vegetation period is important. VV, VH and VV/VH index from different time spans are all important.

The prediction accuracy can be improved if the mapping is not pixel-based but object-based (Orynbaikyzy *et al.*, 2019), i.e. if the average polarization band values of the crop field is used.

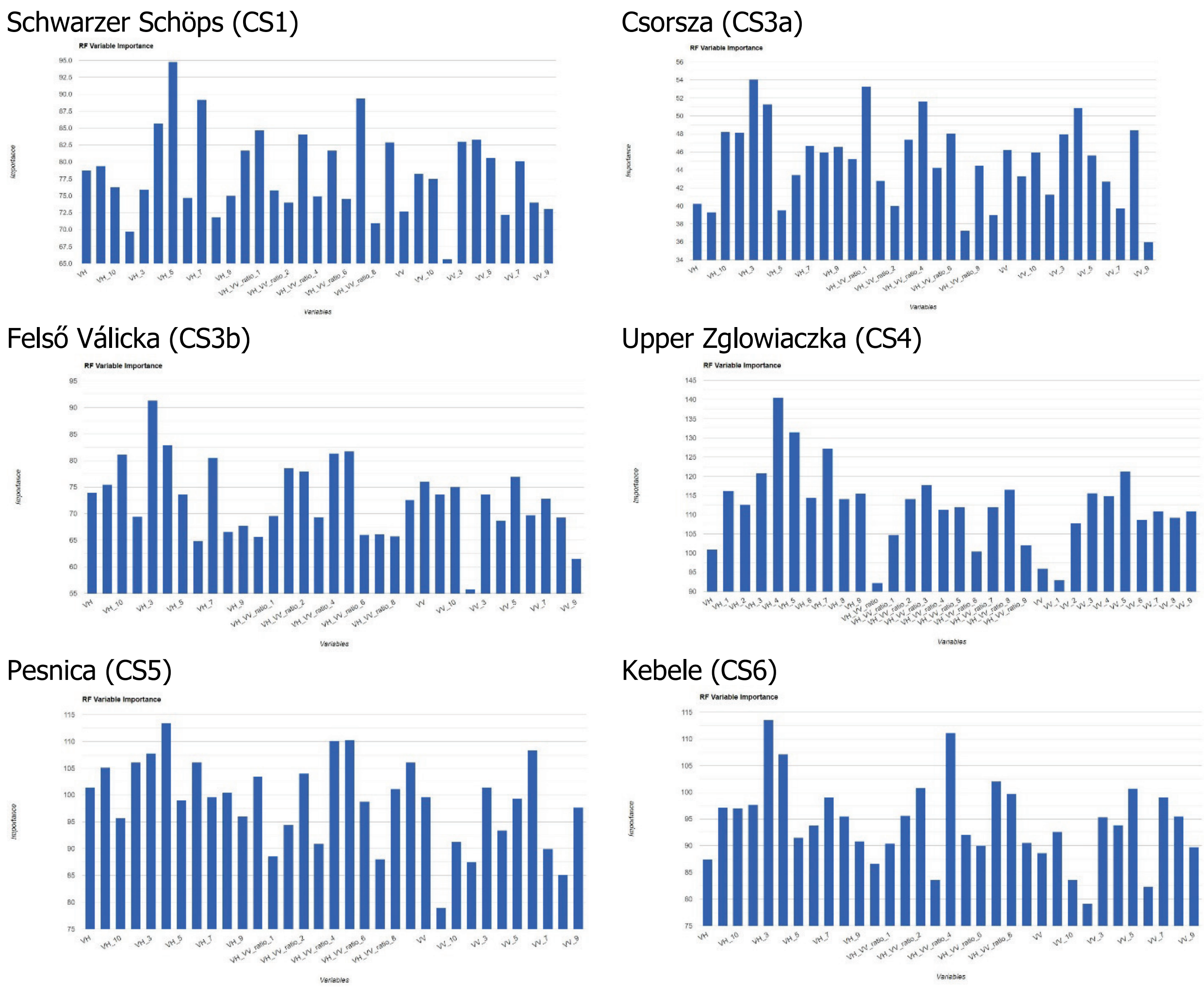


Figure 5: Variable importance to predict crop types for case studies (CS) 1, 3a, 3b, 4, 5, 6.

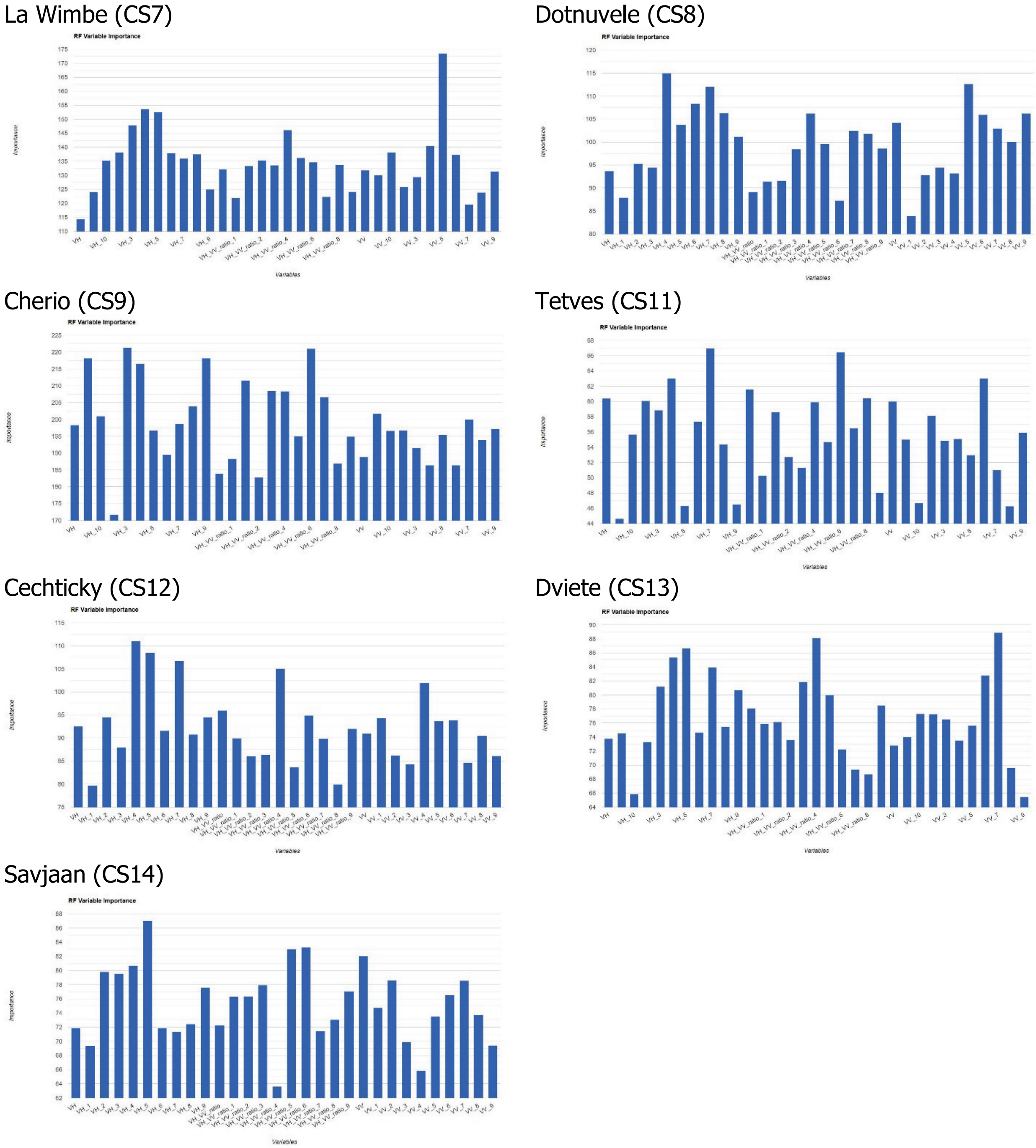


Figure 6: Variable importance to predict crop types for case studies (CS) 7, 8, 9,11, 12, 13 and 14.

The guideline for building a crop classification model and applying it is included in Guideline 3 ("Derivation of time series crop maps") in the Annex. The script to use this approach in Google Earth Engine Platform is available at ZENODO: https://doi.org/10.5281/zenodo.6669643.

4.3. Socio-economic data

The missing socio-economic data will be obtained from official statistics, such as from EC FADN public database and Eurostat, possibly at regional level when and where possible. EU databases do not account properly for the Norwegian and Swiss sites, therefore required data will be retrieved ex novo from local sources or literature. Table 5 shows the data needed to quantify the SPIs through economic modelling.

Table 5: List of potential data for the quantification of socio-economic indicators required for the analysis.

Required input		Unit	Format	Static/ time series (TS)	Required time frequency	Data availability and sources
AGRICULTURAL PRODUCTION AND ECONOMIC RESULTS						
Production costs (variable costs)		€/ha				
Consumables		€/ha	xls, csv	Time series	Annual, 3 or 5 yrs	**EC FADN public database**, regional level and farming type, all case studies except Switzerland and Norway
Labour		€/ha		Time series	Annual, 3 or 5 yrs	
Production factors		€/ha		Time series	Annual, 3 or 5 yrs	
	Seeds and plants	€/ha				
	Fertilizers	€/ha				
	Pesticides	€/ha				
	Water supply	€/ha				
	Energy	€/ha				
	Other variable costs (e.g., feed and fodder)	€/ha				
	Subcontracting	€/ha				
	Harvesting costs	€/ha				
Other costs and info						
Land rent / acquistion		€/ha	xls, csv	Time series	Annual, 3 or 5 yrs	**To be retrieved from regional and/or case-specific sources**
Livestock unit		LU		Time series		**FADN public database,** regional level and farming type, for all case studies except Switzerland and Norway; **EUROSTAT**, regional level, for all case studies
Investment costs for new machinery		€ [€/ha]		Static		**Literature** and/or standard data from other **(local) sources**
Plant material		€/plant		Static		**Literature** and/or standard data from other **(local) sources**
Maintenance costs		€/ha, €/m		Static		**Literature** and/or standard data from **other (local) sources**
Forestry rotation time		yrs		Static		**EUROSTAT**, national level, for all case studies

Forestry production value		€		Time series		(**FADN public database**, regional level, for all case studies except Switzerland and Norway;) **EUROSTAT**, regional level, for all case studies
Farm net value added						
Farm net value added		€ [€/ha]	xls, csv	Time series	Annual, 3 or 5 yrs	**FADN public database**, regional level and farming type, for all case studies except Switzerland and Norway
Market prices of agricultural production						
Crop market price	for each relevant crop included in the landuse pattern	€/t	xls, csv	Time series	Annual, 5-years	**EUROSTAT**, national level, for all case studies (annual selling prices and price indices)
Timber market price	forestry farms	€/t		Time series	Annual, 5-years	Standard data from **local sources**
Market prices of animal products (e.g., milk, meat, livestock head)	where relevant, depeding on the farming type	€/t or €/head		Time series	Annual, 5-years	**EUROSTAT**, national level, for all case studies (annual selling prices and price indices)
Subsidies, incentives and fundings						
CAP fundings	I pillar	€ [€/ha]	xls, csv	Static	Annual, 2020	**FADN public database**, regional level and farming type, for all case studies except Switzerland and Norway. More detailed info **can be retrieved** from CAP subsidy record systems
CAP fundings	II pillar rural dev.	€ [€/ha]		Static	Annual, 2020	**FADN public database**, regional level and farming type, for all case studies except Switzerland and Norway. More detailed info **can be retrieved** from CAP subsidy record systems
Other subsidies - water policy		tbd		Static		**To be retrieved from local sources**

Other subsidies - environmental policy		tbd		Static		**FADN public database,** regional level and farming type, for <u>all case studies except Switzerland and Norway</u>
Other subsidies - climate change mitigation policy		tbd		Static		To be retrieved from local sources
Other subsidies - other nat./reg./local policies		tbd		Static		**FADN public database,** @regional level, for <u>all case studies except Switzerland and Norway</u>
ENVIRONMENTAL IMPLICATIONS						
Quality of natural resources						
Pollutant uptake capacity of crops, shrubs and trees		t/kg d.m.	xls, csv	Static		To be retrieved from iterature and/or other local sources
Emission-control cost factors				Static		To be retrieved from literature and/or other local sources
C sequestration during the growing season		t/kg d.m.		Static		To be retrieved from iterature and/or other local sources
CO_2 emissions (avoided) from measure implementation		t/ha, t/m, …		Static		To be retrieved from literature and/or other local sources
EU Allowances (Emission Trading Scheme)		€/t CO_2 eq.		Static		
Defensive environmental costs (actual protection costs, expenditures to compensate or repair environmental degradation)		€/ha, €/m, …		Static		to be decided

5. Conclusion

For the derivation of soil hydraulic parameters it is recommended to use the pedotransfer functions euptfv1 and euptfv2, depending on what input soil properties are available for the prediction. For the computation of other physical soil properties, we suggested methods that are available from the literature.

Among the chemical soil properties, a workflow to map the soil Olsen phosphorus content is provided. The derived map provides mean values by land use / land cover categories and it is suggested to further specify it with locally measured data, especially for arable lands (if available).

In the case of lacking information on time series crop maps, those could be derived with the Sentinel-1 satellite radar image based approach. For reaching higher accuracy in crop mapping it is advisable to use field level crop data, if available for the CS. This ground truth data could be added to the LUCAS dataset, which is used as a basic training set. The use of local data in addition can secure that the training can be performed for those crop types not represented in the LUCAS dataset, e.g.: onion or parsley is highly underrepresented in LUCAS dataset, thus the model cannot be trained to predict the presence of those crops from remote sensing data.

Missing local socio-economic data will be covered by regional or national official statistics.

Detailed information about the workflow, required input data and data preparation are provided in the following guidelines, available from the Annexes of this report:

- "Guideline 1. Derivation of soil physical and hydraulic properties" , with R script and supporting files available at https://doi.org/10.5281/zenodo.6656454,
- "Guideline 2. Preparation of soil phosphorus content maps for SWAT+ modelling" with R script available at https://doi.org/10.5281/zenodo.6652572,
- "Guideline 3. Derivation of time series crop maps" with a script that is applicable at Google Earth Engine available at https://doi.org/10.5281/zenodo.6669643.

6. Annex

6.1. Guideline 1. Derivation of soil physical and hydraulic properties

6.1.1. Overview

According to the SWAT terminology, the model uses soil physical and chemical data (Neitsch *et al.*, 2009: [link]). Under physical data the following soil properties are required for each soil layer: soil name, soil hydrologic group, maximum rooting depth of the soil profile, depth from soil surface to bottom of layer, moist bulk density, available water capacity of the soil layer, saturated hydraulic conductivity, organic carbon content, clay, silt and sand content, rock fragment content, moist soil albedo, USLE equation soil erodibility (K) factor. These are considered to influence the movement of water and air in the soil profile, thus have impact on the soil hydrologic processes in the Hydrological Response Unit. Chemical properties are optional inputs and includes the following properties: initial nitrate, organic nitrogen, labile phosphorus, organic phosphorus concentration of the surface soil layer.

For most of the CS moist (effective) bulk density, available water capacity of the soil layer, saturated hydraulic conductivity, moist soil albedo and the USLE equation soil erodibility (K) factor was missing, therefore we provide workflow to compute those hereinafter and the definition of the soil hydrologic groups.

6.1.2. Compute soil physical properties

<u>Effective bulk density</u>

In OPTAIN we consider the moist bulk density to be synonym with the effective bulk density which can be computed with the method of Wessolek *et al.* (2009):

- for soils with organic carbon content > 1%:

$$BD_{eff} = BD_{dry} + 0.009 \cdot clay \qquad [1]$$

- for soils with organic carbon content <= 1%:

$$BD_{eff} = BD_{dry} + 0.005 \cdot clay + 0.001 \cdot silt \qquad [2]$$

where BD_{eff} (g cm^{-3}) is effective bulk density, BD_{dry} (g cm^{-3}) is the dry bulk density, clay is clay content (mass %), silt is silt content (mass %).

<u>Moist soil albedo of the top layer</u>

Any of the equations presented by Abbaspour *et al.*, (2019) (Table 1) could be used to calculate the moist soil albedo of the top layer. For the calculations field capacity (FC) is required, which is derived by the approach described in 2.2.1 paragraph.

Table 1: Equations to compute the moist soil albedo of the top layer based on water content at field capacity (θ33) presented in Abbaspour *et al.*, (2019).

Albedo = 0.1807 + 0.1019*exp(−3.53*θ33)
Albedo = 0.15 + 0.31*exp(−12.7*θ33)
Albedo = 0.26 + 0.1068*exp(−4.9*θ33)

USLE soil erodibility (K) factor

It is recommended to use the equation presented by Abbaspour *et al.*, (2019) (Table 2) for the prediction of the Universal Soil Loss Equation (USLE) soil erodibility (K) factor. The computation requires sand silt, clay and organic carbon content of the soil.

Table 2: Equations to compute the USLE soil erodibility (K) factor based on sand (S), silt (T), clay (C) and organic carbon (OC) content for the soil. Source: (Sharpley & Williams, 1990).

E_S = 0.2 + 0.3*exp[−0.0256*S*(1 − T/100)]
$E_{C\text{-}T}$ = [T/(C + T)]^0.3
E_{OC} = 1−(0.25*OC/(OC + exp(3.72 − 2.95*OC)]
E_{HS} = 1−{0.7*(1 − S/100)/[(1 − S/100) + exp(−5.51 + 22.9*(1 − S/100)]}
K_{USLE} = E_S*$E_{C\text{-}T}$*E_{OC}*E_{HS}

6.1.3. Compute soil hydraulic properties

It is recommended to derive the soil hydraulic parameters from the parameters of the van Genuchten model (van Genuchten, 1980). This way the parameters will be self-consistent, and rely on dynamic criterion based on soil internal drainage dynamics (Assouline & Or, 2014; Nasta *et al.*, 2021). Available water capacity and hydraulic conductivity are computed with the predicted van Genuchten parameters.

Available water capacity

Plant available water capacity (AWC) is defined by the water content at field capacity and at wilting point with the following equation:

$$AWC = FC - WLP \qquad [1]$$

where:

FC: water content at field capacity, can be at

-100 cm matric potential head, and

-330 cm matric potential head;

WLP: water content at wilting point, water content at -15,000 cm matric potential head.

Main steps to compute AWC:

1. *Predict parameters of the van Genuchten model*

   First predict parameters of the van Genuchten model that describes the full soil water retention curve (Figure 1):

   - θ_r , θ_s , α and n
   - equation:

$$\theta(\psi) = \theta_r + \frac{\theta_s - \theta_r}{[1+(\alpha\psi^n)]^m} \text{ with } \quad m = 1 - 1/n \qquad [2]$$

where θ_r (cm^3 cm^{-3}) and θ_s (cm^3 cm^{-3}) are the residual and saturated soil water contents, respectively, α (cm^{-1}) is a scale parameter, m (-) and n (-) are shape parameters.

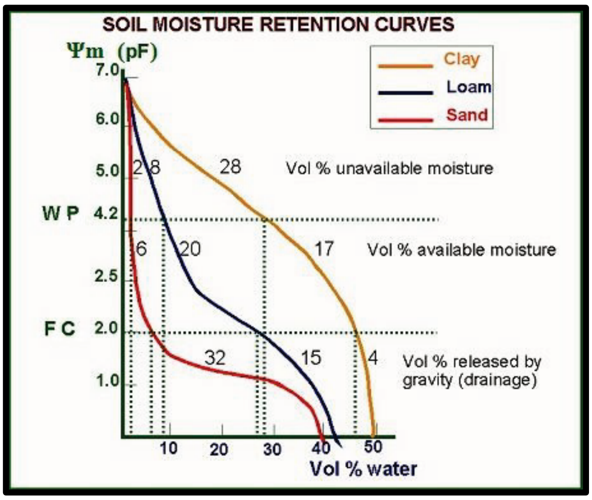


Figure 1: Soil water retention curve of clayey, loamy and sandy soil. Source: www.tankonyvtar.hu – Prof. Lajos Blasko.

The approach to predict parameters of the van Genuchten model depends on the type of input data available for the prediction. As a minimum requirement soil texture classes and topsoil/ subsoil distinction have to be available for the prediction. Hereinafter two main approaches are recommended regarding the input data availability.

1.1. If only soil texture classes and topsoil/subsoil distinction are available, the euptfv1 class pedotransfer function (look up table) (Tóth *et al.*, 2015) could be used to predict the parameters of the van Genuchten model. Tables 3 and 4 show the look up tables depending on the type of soil texture classification.

Table 3: Look up table to assign the parameters of the van Genuchten model to the FAO soil texture classes.

	Modified FAO texture classes	MVG parameters						
		θ_r (cm^3 cm^{-3})	θ_s (cm^3 cm^{-3})	α (cm^{-1})	n (-)	M (-)	K_0 (cm day^{-1})	L (-)
Topsoils	coarse	0.045	0.438	0.0478	1.3447	0.2563	17.30	-2.5587
	medium	0.000	0.459	0.0309	1.1920	0.1611	12.49	-3.8570
	medium fine	0.000	0.432	0.0094	1.2119	0.1749	1.68	-4.4460
	fine	0.000	0.478	0.0403	1.1176	0.1053	40.19	-4.7040
	very fine	0.000	0.522	0.0112	1.1433	0.1253	2.69	-5.0000
	organic	0.111	0.697	0.0069	1.4688	0.3192	1.42	0.3284
Subsoils	coarse	0.057	0.404	0.0426	1.5349	0.3485	9.68	-1.8191
	medium	0.000	0.428	0.0347	1.1725	0.1471	11.78	-4.9869
	medium fine	0.000	0.418	0.0066	1.2173	0.1785	1.87	-3.3761
	fine	0.000	0.430	0.0011	1.2290	0.1863	0.07	-1.8486
	very fine	0.000	0.511	0.0002	1.4048	0.2882	0.02	5.0000
	organic	0.000	0.835	0.0113	1.2256	0.1841	10.81	2.7337

Table 4: Look up table to assign the parameters of the van Genuchten model to the USDA soil texture classes.

	USDA texture classes	MVG parameters						
		θ_r (cm^3 cm^{-3})	θ_s (cm^3 cm^{-3})	α (cm^{-1})	n (-)	m (-)	K_0 (cm day^{-1})	L (-)
Topsoils	Sand	0.061	0.411	0.0258	1.8005	0.4446	8.33	-0.7306
	loamy sand	0.052	0.475	0.0341	1.4846	0.3264	8.95	-1.8749
	sandy loam	0.000	0.441	0.0750	1.1904	0.1599	44.88	-4.3523
	Loam	0.000	0.491	0.0347	1.1931	0.1618	14.17	-4.3000
	silt loam	0.000	0.424	0.0074	1.2545	0.2029	1.17	-3.5496
	Silt	0.009	0.465	0.0042	1.4853	0.3267	1.38	-2.6418
	sandy clay loam	0.000	0.409	0.0700	1.1335	0.1178	43.63	-5.0000
	clay loam	0.000	0.465	0.1284	1.1160	0.1040	195.15	-5.0000
	silty clay loam	0.000	0.463	0.0107	1.1892	0.1591	1.38	-2.6418
	sandy clay	0.192	0.523	0.0351	1.4455	0.3082	43.80	-1.6202
	silty clay	0.000	0.455	0.0309	1.1110	0.0999	0.01	5.0000
	Clay	0.000	0.499	0.0234	1.1200	0.1072	17.07	-5.0000
	Organic	0.111	0.697	0.0069	1.4688	0.3192	1.42	0.3284
Subsoils	Sand	0.034	0.368	0.0356	1.7767	0.4372	5.97	-1.4096
	loamy sand	0.037	0.423	0.0419	1.4222	0.2968	14.84	-1.9583
	sandy loam	0.000	0.437	0.0681	1.1966	0.1643	53.50	-3.7279
	Loam	0.000	0.432	0.0336	1.1701	0.1454	8.58	-5.0000
	silt loam	0.000	0.422	0.0077	1.2483	0.1989	1.76	-3.3247
	silt	0.009	0.465	0.0042	1.4853	0.3267	0.45	-5.0000
	sandy clay loam	0.000	0.384	0.0717	1.1206	0.1076	37.09	-5.0000
	clay loam	0.000	0.413	0.0227	1.1191	0.1064	12.35	-5.0000
	silty clay loam	0.000	0.408	0.0032	1.1993	0.1662	0.45	-5.0000
	sandy clay	0.000	0.365	0.0016	1.1812	0.1534	43.80	-1.6202
	silty clay	0.000	0.442	0.0003	1.3861	0.2786	0.01	5.0000
	clay	0.000	0.461	0.0004	1.3027	0.2323	0.04	1.1840
	organic	0.000	0.835	0.0113	1.2256	0.1841	10.81	2.7337

1.2. If information on mean soil depth, sand, silt and clay content is available, it is recommended to use euptfv2 (Szabó et al., 2021):

Minimum input requirements:

- mean soil depth (cm),
- percentages of clay (<2 μm), silt (2–50 μm) and sand (50–2000 μm) content,

Optional further inputs:

- organic carbon content (mass %), bulk density (g cm-3), calcium carbonate content (mass-%), pH in water (-), cation exchange capacity (cmol (+) kg−1).

Tools to use euptfv2:

- user friendly web interface: https://ptfinterface.rissac.hu (Szabó *et al.*, 2019),
- R package to use the pedotransfer functions, archived on Zenodo: https://doi.org/10.5281/zenodo.4281046 (Weber *et al.*, 2020).

2. *Compute FC and WLP from the van Genuchten parameters*
   For the computation of AWC, WLP and FC is required.
   - Computation of WLP: use the predicted van Genuchten parameters:

$$\text{WLP} = \theta_r + \frac{\theta_s - \theta_r}{[1+(\alpha \cdot 15000^n)]^{1-1/n}} \qquad [3]$$

   - Computation of FC: for sake of simplicity, it is often estimated at a fixed soil matric head (e.g. –500 cm, or –330 cm, or –100 cm), depending mainly on the dominant soil textural class in the soil profile, but the FC value can also be predicted through the physically-based analytical equation proposed by Assouline & Or (2014):

$$\text{FC} = \theta_r + (\theta_s - \theta_r)\left\{1 + \left[\frac{n-1}{n}\right]^{(1-2n)}\right\}^{\left(\frac{1-n}{n}\right)} \qquad [4]$$

     where θ_r ($cm^3\ cm^{-3}$) and θ_s ($cm^3\ cm^{-3}$) are the residual and saturated soil water contents, respectively, α (cm^{-1}) is a scale parameter, and n (-) is the shape parameter of the van Genuchten (1980) model.

   This computation of FC is self-consistent, dynamic criterion based on soil internal drainage dynamics.

3. *Compute Available soil water capacity (AWC)*
   - based on the FC and WLP computed in step 2 compute AWC with Eq. [1].

Saturated hydraulic conductivity (KS)

Compute from parameters of the van Genuchten model – which were predicted above for AWC from basic soil properties – by the equation of Guarracino (2007):

$$K_S = 4.65 \cdot 10^4 \theta_s \alpha^2 \qquad [5]$$

where K_s is expressed in units of $cm\ d^{-1}$.

Hydrologic Soil Group

The hydrologic soil groups (HSG) are based on the infiltration characteristic of the soil and include four groups having similar runoff potential. The groups are defined based on the saturated hydraulic conductivity, depth to high water table and depth to water impermeable layer. More details can be found in U.S. Department of Agriculture Natural Resources Conservation Service (2009).

For defining the HSG, the following input data is required in GeoTIFF format:
- depth to water table (m),
- map of saturated hydraulic conductivity (μm/s).

Main steps to define HSG:

1. For soil layers 0-50 cm, 0-60 cm and 0-100 cm the minimum saturated hydraulic conductivity has to be set.

2. Set the depth of the layer which has KS < 0.01 µm/s.
3. Define HSG codes based on maximum depth to impermeable soil layer, minimum KS value for different soil depth ranges and water table depths according to Table 7-1 Criteria for assignment of hydrologic soil group (HSG) of the U.S. Department of Agriculture Natural Resources Conservation Service (2009) (Table 5.)

Table 5. Definition of soil hydrologic groups based on U.S. Department of Agriculture Natural Resources Conservation Service (2009)

Depth to waterimpermeable layer 1/	Depth to highwater table 2/	K_{sat} of least transmissivelayer in depth range	K_{sat} depth range	HSG 3/
<50 cm	—	—	—	D
50 to 100 cm	<60 cm	>40.0 µm/s	0 to 60 cm	A/D
		>10.0 to ≤40.0 µm/s	0 to 60 cm	B/D
		>1.0 to ≤10.0 µm/s	0 to 60 cm	C/D
		≤1.0 µm/s	0 to 60 cm	D
	≥60 cm	>40.0 µm/s	0 to 50 cm	A
		>10.0 to ≤40.0 µm/s	0 to 50 cm	B
		>1.0 to ≤10.0 µm/s	0 to 50 cm	C
		≤1.0 µm/s	0 to 50 cm	D
>100 cm	<60 cm	>10.0 µm/s	0 to 100 cm	A/D
		>4.0 to ≤10.0 µm/s	0 to 100 cm	B/D
		>0.40 to ≤4.0 µm/s	0 to 100 cm	C/D
		≤0.40 µm/s	0 to 100 cm	D
	60 to 100 cm	>40.0 µm/s	0 to 50 cm	A
		>10.0 to ≤40.0 µm/s	0 to 50 cm	B
		>1.0 to ≤10.0 µm/s	0 to 50 cm	C
		≤1.0 µm/s	0 to 50 cm	D
	>100 cm	>10.0 µm/s	0 to 100 cm	A
		>4.0 to ≤ 10.0 µm/s	0 to 100 cm	B
		>0.40 to ≤4.0 µm/s	0 to 100 cm	C
		≤0.40 µm/s	0 to 100 cm	D

1/ An impermeable layer has a K_{sat} less than 0.01 µm/s [0.0014 in/h] or a component restriction of fragipan; duripan; petrocalcic; orstein; petrogypsic; cemented horizon; densic material; placic; bedrock, paralithic; bedrock, lithic; bedrock, densic; or permafrost.

2/ High water table during any month during the year.

3/ Dual HSG classes are applied only for wet soils (water table less than 60 cm [24 in]). If these soils can be drained, a less restrictive HSG can be assigned, depending on the K_{sat}.

6.1.4. Harmonize size limits of sand, silt and clay content

Different countries/institutes measure soil particle-size distribution (PSD) using different methods and recognizing different classification standards. The SWAT+ model defines the sand, silt and clay content according to the USDA system, which means that the particle size limit is < 0.002 mm for the clay, 0.002-0.05 mm for the silt and 0.05-2 mm for the sand fraction.

When conversions between different methods are required it is advised to apply k-nearest neighbor interpolation (formerly called: 'similarity technique'), which results in less uncertainty, lesser bias and shrinkage of resulting texture range compared to the simpler loglinear interpolation (Nemes *et al.*, 1999).

If conversion of the particle size distribution data of the CS is needed, please contact Attila Nemes (attila.nemes@nibio.no) to get support.

If only info on texture classes exist, the class median points could be used as clay-silt-sand data triplets.

6.1.5. Remarks

The equations suggested in sections 2.1 and 2.2 are available at https://doi.org/10.5281/zenodo.6656454 and from an R script, called "**pred_soil_prop.R**", stored in the OPTAIN cloud[1] for project partners. The R script requires the SWAT+ table in CSV format, which is used to define the soil properties, an example file is available at the above folder of the OPTAIN cloud as "**soil_prop_in.csv**". Example output files are stored in the cloud[2] as "**soil_prop_out.xlsx**" and "**pred_FC_WP.xlsx**".

If there is any other missing basic soil property – such as organic carbon content or sand, silt and clay content, or bulk density or coarse fragments – those are suggested to be covered from the SoilGrids dataset (Poggio *et al.*, 2021).

Other pedotransfer functions to predict the parameters of the van Genuchten model and saturated hydraulic conductivity are available at Appendix A of Nasta et al. (2021), which is also available in the OPTAIN cloud[3]. Furthermore, the following publications are provided: Abbaspour *et al.* (2019); U.S. Department of Agriculture Natural Resources Conservation Service (2009) (filename: **USDA_ NRCS_define_HSG_2009.pdf**). If other approaches than the above recommended are used by the CS leaders, it has to be justified.

[1] Folder: 'WPs & Tasks/WP3/Task_3_3/templates/derive_soil_hydraulic_properties'
[2] Folder: 'WPs & Tasks/WP3/Task_3_3/templates/derive_soil_hydraulic_properties/sample_output'
[3] Folder: 'WPs & Tasks/WP3/Task_3_3/templates/derive_soil_hydraulic_properties/references'

6.2. Guideline 2. Preparation of soil phosphorus content maps for SWAT+ modelling

6.2.1. Overview

The SWAT+ model can use information on the initial soil nutrient content as input (nutrient.sol). This input is optional for the model, else the model uses default values if intital soil nutrient information are not provided (Arnold *et al.*, 2012). In OPTAIN, we focus on analysing the effectiveness of NSWRMs in retaining water and nutrients, therefore it is important to derive approximate soil nutrient maps for the case studies (CSs) instead of using the model's default values.

Data availability is diverse in the CSs, some of them have measured soil nutrient content data at field-scale, and some lack this information. As SWAT+ requires spatial data that completely cover the CS areas, the input soil nutrient maps must not contain missing values. All land use/ land cover categories (including artificial surfaces, water bodies) must have a non NA value for soil nutrient content. Hereinafter a guideline is provided for all CSs on how to produce a base map of mean soil phosphorus content of the surface soil layer and add locally measured values if those are available. The presented workflow uses open access software and datasets, but could be performed in any other software or use any relevant dataset that is accessible for the partners.

SWAT is able to model six different phosphorus pools in the soil (Figure 1). Three pools are inorganic forms of phosphorus while the other three pools are organic forms of phosphorus (P). Fresh organic P is associated with crop residues and microbial biomass while the active and stable organic P pools are associated with the soil humus. The organic phosphorus associated with humus is partitioned into two pools to account for the variation in availability of humic substances to mineralization. Soil inorganic P is divided into solution, active, and stable pools. The solution pool is in rapid equilibrium (several days or weeks) with the active pool. The active pool is in slow equilibrium with the stable pool.

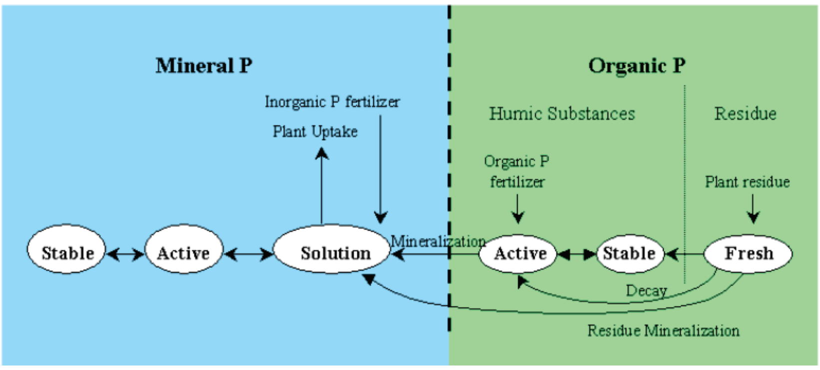


Figure 1: SWAT soil phosporus pools and processes that move phosphorus in and out of pools. Source: https://swat.tamu.edu/media/99192/swat2009-theory.pdf

A detailed description on how SWAT models the nutrient cycle for phosphorus can be found in the theoretical documentation of SWAT (Neitsch *et al.*, 2009: [link]). Particularly recommended sections are section 3.2 (page 206-219) on the terrestrial phosphorus cycle and related equations as well as sections 4.2.3 & 4.2.4 (page 272-275) on the movement of soluble P and transport of mineral/organic P that's attached to soil particles. Finally, also section 5.2.3.2 (page 335-338) is recommended for information how the model simulates the P uptake by plants.

Soil phosphorus content required for SWAT+

The SWAT+ model requires the labile phosphorus (P) content of the surface layer in ppm for initialization of the different P-pools. The model's default value is 5 ppm. Labile-P is the amount of P that is available for plants and microorganisms. Consequently, its the sum of inorganic and organic P absorbed in the soil in a way that it can easily enter the soluble phase (Costa *et al.*, 2016).

According to Chaubey *et al.* (2006) labile P is the P extracted by an anion exchange resin (Sharpley *et al.*, 1984) and therefore represents solution P plus weakly sorbed P. Within SWAT the pool "solution P" (Figure 1) is actually labile P in conformance with the original EPIC version of the P module as described in Jones *et al.* (1984) and Sharpley *et al.* (1984). The initial concentration of solution phosphorus in SWAT is of particular relevance, as it will be used for the model-internal initialisation of both mineral P pools (active & stable).

Several methods exist for the determination of different P formats, but the level of P analysed highly depends on the method used for its determination. Most commonly used P test methods are:

- Acid ammonium acetate lactate extraction (AL method; Egnér *et al.*, 1960), which is applied in Belgium (Flanders), Hungary, Lithuania, Norway, Slovenia and Sweden
- Sodium bicarbonate extraction (Olsen method; Olsen *et al.*, 1954), which is applied globally, it is the official soil P test in Denmark, England, France, Italy, Spain and this method is used in the LUCAS Topsoil survey (EC JRC).

Both, AL and Olsen P, are approximating plant available P. For the case studies (CSs) of the OPTAIN project measured soil P content is not available for most of the catchment areas. However, it is an important input data to analyse the nutrient retention in the CSs. As the LUCAS Topsoil Survey dataset contains measured Olsen-P content data of about 20,000 soil samples in Europe, we decided to use it to derive approximate maps of soil P content for data scarce areas. This way we provide an uniform method for producing Olsen-P maps for any of the OPTAIN CSs based on the LUCAS dataset.

Hereinafter we provide a guideline to derive Olsen-P content maps in order to fulfil the minimum agreed input data requirements of OPTAINs catchment scale modelling. Should any additional information be available on the listed SWAT+ soil nutrient properties for a CS, it is advantageous and recommended to use that as input data as well.

Theoretical background of mapping soil phosphorus content

Land use has a strong influence on soil P content. Most of the agricultural areas have higher levels of P compared to areas with natural land cover. The phosphorus level of agricultural soils mainly depends on the fertilization practices of the countries (Tóth *et al.*, 2014) and climate (Ballabio *et al.*, 2019). Fertilization practices are determined by several factors: economy of the country, climate, tillage practices, and the characteristic of the crop production. Based on the relationships mentioned above, the country, the land use and the agro-climate of the CSs are taken into account in the mapping process.

6.2.2. Required input datasets

Hereinafter input dataset required to derive the soil P maps are described. All the listed datasets are available in the OPTAIN cloud (Table 1) for the OPTAIN case study partners.

1. *LUCAS Topsoil Survey dataset*

The Topsoil Survey dataset of the Land Use/Land Cover Area Frame Survey (LUCAS) includes measured point data on, among others, Olsen-P content of topsoils (0-20 cm soil depth) for member states of the European Union from 2009 and 2015. In 2015, the LUCAS topsoil survey was also performed for Switzerland. The dataset is freely available and, besides land use and land cover, it includes information on basic soil properties such as coarse fragment content (%), particle size distribution (% clay, silt and sand content), pH (in $CaCl_2$ and H_2O), organic carbon (g kg^{-1}), carbonate content (g kg^{-1}), phosphorus content (mg kg^{-1}), total nitrogen content (g kg^{-1}), extractable potassium content (mg kg^{-1}), cation exchange capacity (cmol(+)kg^{-1}), multispectral properties (Tóth *et al.*, 2013b).

In the case of lacking information on soil nutrient content in CSs of EU countries, the LUCAS Topsoil Survey dataset could be used to map soil P contents, except for Norway, which is not covered by this dataset. Therefore national data on soil P content should be used there, if available. The open access dataset is also available at https://esdac.jrc.ec.europa.eu/projects/lucas .

2. *LUCAS Land Use / Cover Area Frame Survey*

The LUCAS survey provides information about land use and land cover. It can be merged with the LUCAS Topsoil Survey dataset. Within this guideline we use the LUCAS land cover codes because they can be easier aligned with the information available from local high resolution land use or land cover maps. For the analysis the harmonised version of the LUCAS 2009 and 2015 dataset is used (d'Andrimont *et al.*, 2020). The open access dataset is also available at https://figshare.com/articles/dataset/Harmonised_LUCAS_in-situ_land_cover_and_use_database_for_field_surveys_from_2006_to_2018_in_the_European_Union/9962765/2 .

3. *Local land use or land cover map*

For computing the mean Olsen-P content by land use or land cover categories, a national or local land use/ land cover map is required. If no national map is available, the CORINE dataset can be used. The open access dataset is also freely available at https://land.copernicus.eu/pan-european/corine-land-cover [link].

4. *European agro-climate zone map*

The observed agro-climate zones in the period of 1996-2016, which have been published by the European Environmental Agency (Ceglar *et al.*, 2019), are used to determine the agro-climate of the CSs. The open access dataset is also freely available at https://www.eea.europa.eu/data-and-maps/figures/observed-climate-zones-in-the .

6.2.3. Optional input datasets

If measurements of soil P content are available for a CSs, for example on agricultural fields, this data is a very important input for the mapping. This data can be used to:

1) assess the accuracy of the derived map (see section 1.6, step 7 "Steps automatically performed by the R script") and
2) overwrite the mean soil P content values – mapped in step 7 – with measured ones.

If additional locally measured soil P content data is available for a CS, the data needs to be converted into Olsen-P in ppm. Please be aware that locally measured P content might be determined with several different laboratory methods. A wide range of transfer functions to compute Olsen-P from P content measured by other methods is available in Steinfurth et al. (2021). For the OPTAIN case study partners this paper is available in the OPTAIN cloud[4].

Measured data on soil phosphorus content needs to be provided in a shape file format (Figure 2) and the path to the file must be added to "file_path_info.xlsx" (see section 1.5).

a) b)

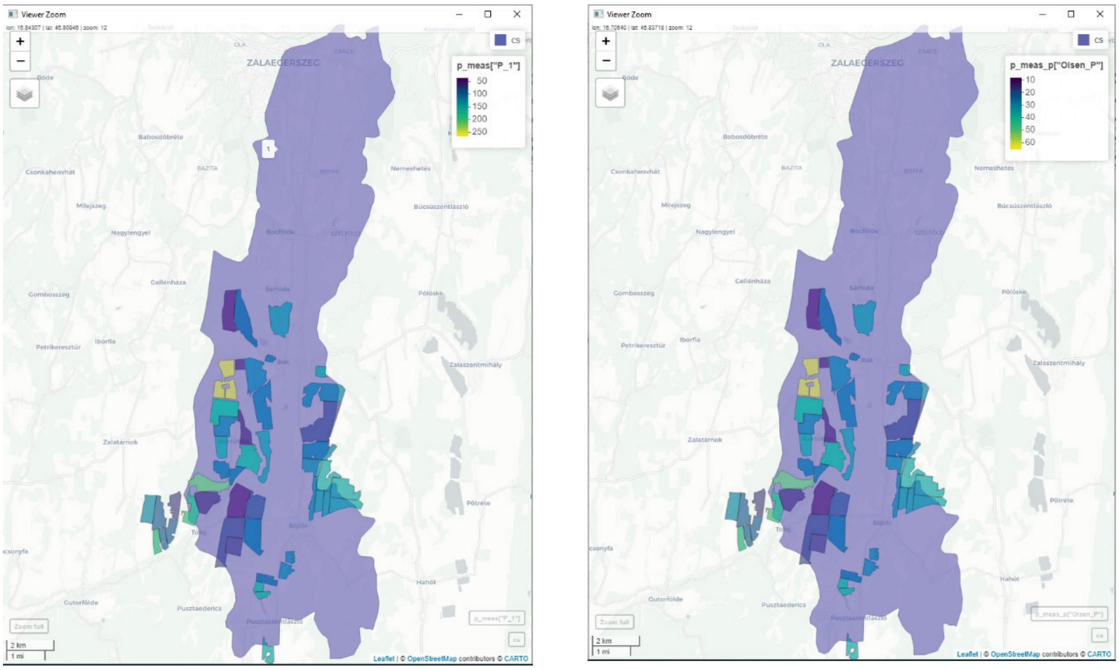


Figure 2: Shape file of the measured soil phosphorus content in ppm at the Felső-Válicka catchment (CS3b) a) based on acid ammonium acetate lactate extraction in P_2O_5 – originally available and b) converted into Olsen method.

6.2.4. Software to use

The preparation of soil phosphorus content maps for SWAT+ modelling can be performed in different programming environments. The authors of this deliverable D3.2 developed a script of the analysis and preparation of soil Olsen-P content in the R programming environment (R Core Team, 2021). The analysis can be adapted for each OPTAIN CS. Required R packages include 'raster' (Hijmans, 2021), 'rgdal' (Bivand *et al.*, 2021), 'sf' (Pebesma, 2018), 'mapview' (Appelhans *et al.*, 2021), 'fasterize' (Ross, 2020), 'dplyr' (Wickham *et al.*, 2021), 'tidyverse' (Wickham *et al.*, 2019) and 'readxl' (Wickham & Bryan, 2019).

[4] Folder: WPs & Tasks/ WP3/ Task_3_3/ templates/ map_soil_P/ references

6.2.5. Required data preparation

Before running the R script available as "**map_soil_P.R**", which is available in the OPTAIN cloud and at https://doi.org/10.5281/zenodo.6652572, each user needs to go through the following steps to prepare the required data.

1. *Download necessary input files*

The first step is to access the existing maps, datasets as well as a table to recode land use / land cover categories. The following files have to be downloaded (e.g. from the OPTAIN cloud[2]) into your local folder. These files can be used without any modifications.

Table 1: Input datasets, maps and tables accessible from the OPTAIN cloud[5].

Description	Name of object in R script	File name in the UFZ cloud
LUCAS Topsoil Survey dataset for 2009 and 2015	lucas_topsoil	LUCAS_topsoil_2009_EU.csv LUCAS_topsoil_2015_EU.csv
CORINE map for 2006, 2012 and 2018 (not needed if local land use map is available)	corine	CLC2006_V2020_20u1.tif CLC2012_V2020_20u1.tif CLC2018_V2020_20u1.tif
LUCAS land use and land cover dataset for 2009 and 2015	lucas	lucas_harmo_uf_2009.csv lucas_harmo_uf_2015.csv
Table to recode LUCAS land use and land cover categories into CORINE level 2 categories	lucas_lc_into_corine_2	recode_lucas_lc1_into_corine_2.csv
Agro-climate zones in Europe (EEA)	cz_map	CLIMATE_Zones_EEA_1996_2016.tif

2. *Provide meaning of local land use map categories according to CORINE level 2 categories*

A detailed land use raster map in GeoTIFF format from each CSs will be required to derive the soil P content map. Please use the most detailed land use map available. Information for the use of the local land use map:

- Required format: GeoTIFF
- Minimum spatial extent: boundaries of the CS
- Specific projection: not required, because the local land use map will be transformed into the projection of shape file of the CS within the R script
- Resolution: the resolution of the land use map will determine the resolution of the soil P map.

[5] The OPTAIN cloud is accessible for OPTAIN partners only. Relevant folder: WPs & Tasks/ WP3/ Task_3_3/ templates/ map_soil_P/input

Table 2: Codes and names of CORINE categories at each level. Level 2 in bold is used for the mapping of soil phosphorus content.

Level 1	Level 2	Level 3
1 Artifical surfaces	**11 Urban fabric**	111 Continuous urban fabric
		112 Discontinuous urban fabric
	12 Industrial, commercial and transport units	121 Industrial or commercial units
		122 Road and rail networks and associated land
		123 Port areas
		124 Airports
	13 Mine, dump and construction sites	131 Mineral extraction sites
		132 Dump sites
		133 Construction sites
	14 Artificial, non-agricultural vegetated areas	141 Green urban areas
		142 Sport and leisure facilities
2 Agricultural areas	**21 Arable land**	211 Non-irrigated arable land
		212 Permanently irrigated land
		213 Rice fields
	22 Permanent crops	221 Vineyards
		222 Fruit trees and berry plantations
		223 Olive groves
	23 Pastures	231 Pastures
	24 Heterogeneous agricultural areas	241 Annual crops associated with permanent crops
		242 Complex cultivation patterns
		243 Land principally occupied by agriculture, with significant areas of natural vegetation
3 Forest and semi natural areas	**31 Forests**	311 Broad-leaved forest
		312 Coniferous forest
		313 Mixed forest
	32 Scrub and/or herbaceous vegetation associations	321 Natural grasslands
		322 Moors and heathland
		323 Sclerophyllous vegetation
		324 Transitional woodland-shrub
	33 Open spaces with little or no vegetation	331 Beaches, dunes, sands
		332 Bare rocks
		333 Sparsely vegetated areas
		334 Burnt areas
		335 Glaciers and perpetual snow
4 Wetlands	**41 Inland wetlands**	411 Inland marshes
		412 Peat bogs
	42 Maritime wetlands	421 Salt marshes
		422 Salines
		423 Intertidal flats
5 Water bodies	**51 Inland waters**	511 Water courses
		512 Water bodies
	52 Marine waters	521 Coastal lagoons
		522 Estuaries
		523 Sea and ocean

If there is no local map, the CORINE 100 meter database from the relevant year (2006 or 2012 or 2018, which is closest to the sampling year of LUCAS dataset selected) should be downloaded and used as “local” land use map.

The local land use categories have to be converted into CORINE level 2 categories. For this a look-up table has to be created by the CSs modeller. The three levels of CORINE land cover categories are shown in Table 2. **The meaning of local land use categories has to be given with CORINE level 2 categories (level 1 and level 3 are not applicable).** The map does not have to be recoded (this step is included in the R script), but a recoding table in an .xls file needs to be provided. The template of this recoding table is available in the OPTAIN cloud for all project partners ('**recode_l_lu_map_corine_2.xls**') as well as at https://doi.org/10.5281/zenodo.6652572. Please adapt this file to match the local land use/ land cover categories (column A: l_lu_map_code) with CORINE level 2 categories (column C: corine_2_local) and add this .xls file and the local land use map to your folder of input data.

3. *Fill in "file_path_info.xlsx" file*

3.1 In the R script, information necessary to derive the soil P content map is read from the **"file_path_info.xlsx"** file. The template is available in the OPTAIN cloud for all project partners as well as at https://doi.org/10.5281/zenodo.6652572. The column E of this table has to be filled in for your CS. Cells in blue need to be filled in, cells in yellow are optional (can be left empty). Short explanation is inserted to each cells of this column. Columns C and D provide examples.

Do not delete any columns/rows from this table. If something is not needed, leave this cell empty (e.g. if you use a local land use map, the CORINE land cover map is not needed).

3.2 Add the selection criteria of the LUCAS dataset based on your expert knowledge in **"file_path_info.xlsx"** file. Do not use all LUCAS points to create the Olsen-P map of the CS, because then you would only get average P values for Europe, which would not be specific for your CS. In the cells of E11-18 of "file_path_info.xlsx" you have to specify a NUTS area (NUTS0, NUTS1, NUTS2) where the fertilization practices are similar to your CS area. It can be the NUTS code of the country or region of your CS.

Example: in the case of Hungarian CSs we used the code NUTS0 of the country (HU). It means that we used all points located in Hungary.

If the fertilization practices are different within the country of the CS, a second selection is required. There are two possible solutions to perform this second selection:

- keep the NUTS region name(s) selected for your country in E11-18 cells in "file_path_info.xlsx" and continue the selection based on agro-climate zones (Figure 3, Table 3). You can add the selected agro-climate zones in the cell E20-E24. If you select only one agro-climate zone leave the other cells empty.

Example: Italy lies in more than one agro-climate zone, in such a case only LUCAS data from the same agro-climate zone where the CS is located could be kept for the analysis.

- select NUTS1 or NUTS2 areas in your country where the fertilization practices are similar to the one characteristic for your CS and leave the cells of agro-climate zones empty.

Example: in Poland fertilization practices differ across NUTS regions. In this case LUCAS data with NUTS regions having similar fertilization practices could be retained for the analysis.

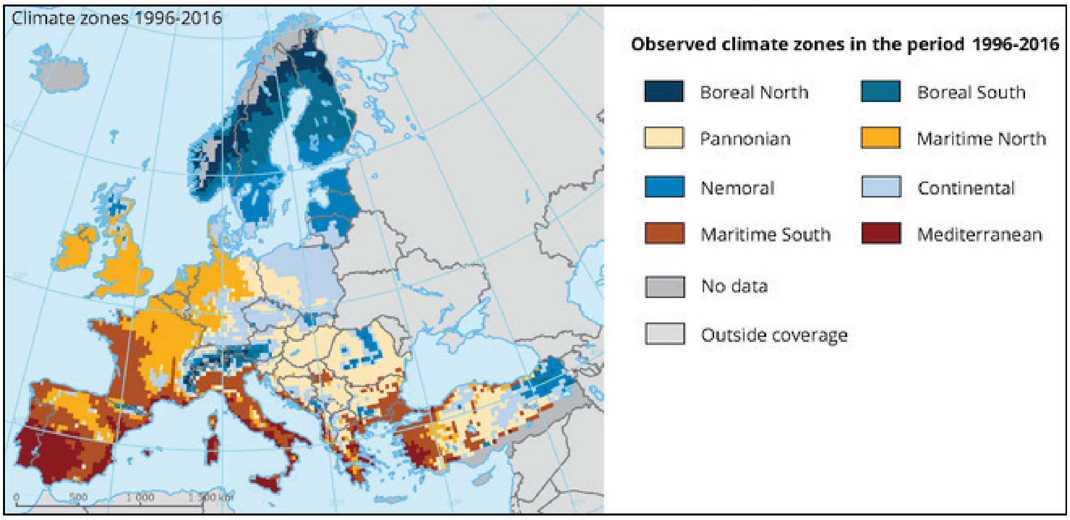


Figure 3: Agro-climate zones in Europe. Source: Ceglar *et al.*, (2019).

<u>IMPORTANT NOTE!</u> At least about 200 LUCAS points are required to run the script properly and prepare the map. If in your country (or in the selected area) the number of LUCAS points is <200, you have to expand the dataset with points from surrounding countries or areas. The number of LUCAS points per country can be seen in Table 4.

Example: in the case of Belgium only 74 LUCAS points are in the country. Points from other countries e.g. the Netherlands and North-France should be involved. Or points from the agro-climate zone of the CSs should be used from the surrounding countries.

Table 3: List of the agro-climate zones of the OPTAIN CSs and CS countries

Case Study name	Country	CS agro-climate zone	Agro-climate zones in the country
Schwarzer Schöps	DE	Pannonian, Continental	Pannonian, Continental, Nemoral
Petite Glane	CH	Continental, Pannonian	Continental, Pannonian, Nemoral, Boreal North, Boreal South
Csorsza	HU	Pannonian	Pannonian
Felső-Válicka	HU	Pannonian	Pannonian
Upper Zglowiaczka	PL	Continental	Continental, Pannonian, Nemoral
Pesnica	SI	Pannonian, Continental	Pannonian, Continental, Maritime South
Kebele-Kobiljski	HU	Pannonian	Pannonian
Kebele-Kobiljski	SI	Pannonian	Pannonian, Continental, Maritime South
La Wimbe	BE	Maritime North	Maritime North, Continental
Dotnuvele	LT	Continental	Continental, Nemoral
Cherio	IT	Maritime South, Continental	All
Krakstad (Hobol)	NO	Nemoral	Nemoral, Boreal North, Boreal South, Continental, Maritime North
Tetves	HU	Pannonian	Pannonian
Cechticky	CZ	Continental	Continental, Pannonian, Nemoral
Dviete	LV	Nemoral	Nemoral, Continental
Savjaan	SE	Nemoral	Nemoral, Boreal South, Boreal North, Continental

In the case of countries where no LUCAS points are available (e.g. Norway) a similar procedure could be followed. Swedish samples might be used or samples selected from the agro-climate zone similar to that of the CS. If national data on soil P content is available that could be used as well instead of the LUCAS Topsoil Survey dataset, but please convert the data to Olsen-P content for comparability between the OPTAIN CSs.

Table 4: Number of LUCAS points for some relevant countries by year of survey.

Country	N (2009)	N (2015)
AT	420	543
BE	74	146
CH	0	150
CZ	431	440
DE	1947	1687
DK	232	222
EE	220	194
ES	2696	4027
FI	1716	1149
FR	2952	3050
GR	491	0
HR	0	114
HU	497	412
IE	233	197
IT	1333	1642
LT	356	352
LU	3	13
LV	351	310
NL	211	172
PL	1648	1377
PT	476	447
SE	2256	1903
SI	112	147
SK	268	228
UK	942	744

The NUTS2 phosphorus map of the European cropland areas (Figure 3) might also be useful in the data selection (Tóth *et al.*, 2014).

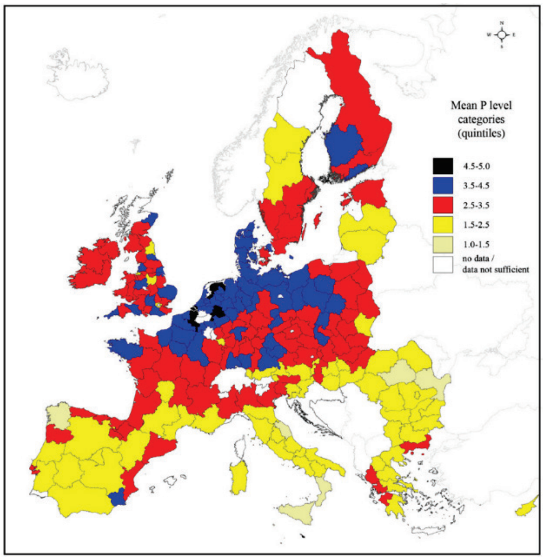


Figure 4. Mean Olsen P level of croplands (based on quintile categories).

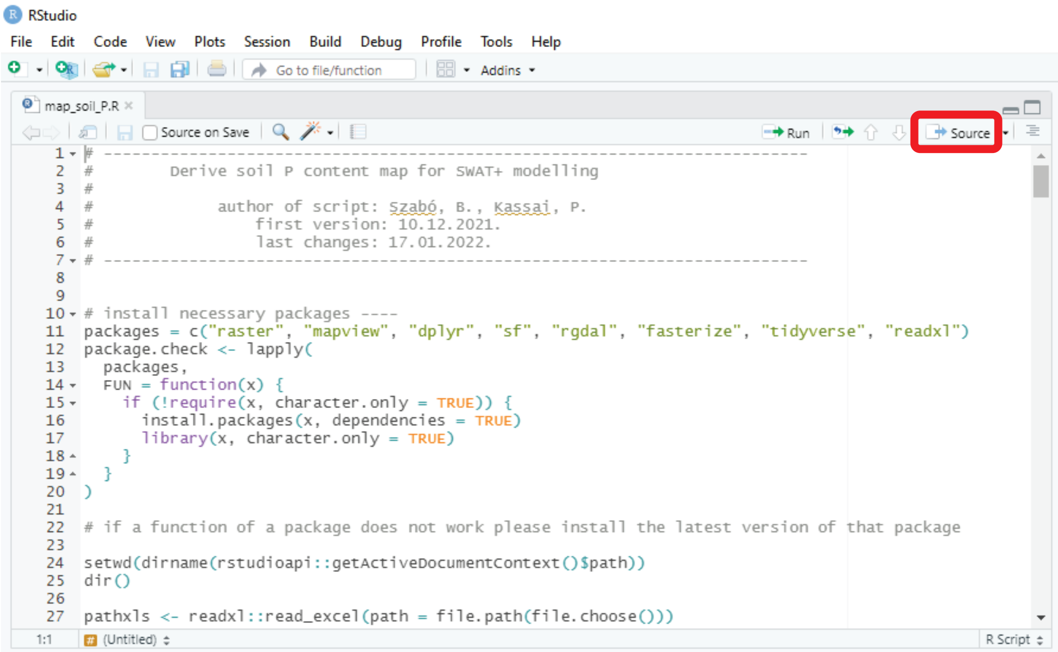
4. *Format locally measured soil phosphorus content data, if available*

If some additional locally measured soil P content data is available for the CS (for example on agricultural fields) **convert that data into Olsen-P** in ppm. Please be aware that locally measured P content might be determined with several different laboratory methods. Please find available transfer functions to compute Olsen-P from P content measured by other methods in Steinfurth et al. (2021). Please provide this measured data in a shape file and add its file location in **"file_path_info.xlsx"** (see also section 1.3).

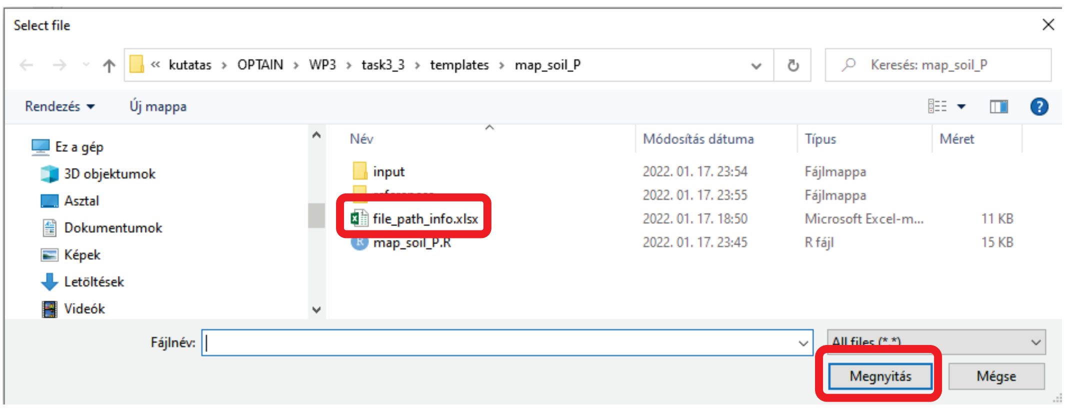
6.2.6. Main steps to map soil Olsen phosphorus content for the surface layer

I. Download **map_soil_P.R** file from the OPTAIN cloud or https://doi.org/10.5281/zenodo.6652572 and open the script in **RStudio** with **File/Open File...**

II. Click on the 'Source' button in the top right corner of the main window:



III. The script will prompt you a window where you have to browse for the **"file_path_info.xlsx"** Excel spreadsheet. **Select it** and click on **Open**:



If "file_path_info.xlsx" is properly filled in, the script runs and soil nutrient maps are created. The calculations need a few minutes.

List of steps that are automatically performed by the R script:

1. Input data and info loaded:
   1.1. Load shape of the catchment.
   1.2. Select and load that LUCAS Topsoil Survey data which is relevant for the modelling period (2009 or 2015).
   1.3. Load that LUCAS land cover data – 2009 or 2015 - , which is identical with the sampling year of LUCAS Topsoil Survey dataset loaded in step 1.2.
   1.4. Load agro-climate zone map.
   1.5. Load relevant CORINE, based on the SWAT modelling period, not needed, if local land use map is available.
   1.6. Load local land use map provided by CS partner. *Example: in the case of Felső-Válicka CS we used a 20 m resolution national land use map (NÖSZTÉP map, Tanács et al., 2019).*
   1.7. Load table to recode local land use map categories into CORINE level 2 categories, table provided by the CS partner.
   1.8. Load table to recode LUCAS land cover/ land use categories into CORINE level 2 categories.
2. LUCAS land use and land cover categories and agro-climate zone codes are added to the LUCAS Topsoil Survey dataset.
3. LUCAS land use codes are recoded into CORINE level 2 categories. Conversion table is "recode_lucas_lc1_into_corine_2.csv", shown in Table 5, too.

Table 5: Recoding table of LUCAS land cover categories into CORINE level 2.

LUCAS code	LUCAS name	CORINE code	CORINE name
A11	Buldings with one to three floors	11	Urban fabric
A12	Buldings with more than three floors	11	Urban fabric
A13	Greenhouses	11	Urban fabric
A21	Non built-up area features	11	Urban fabric
A22	Non built-up linear features	12	Industrial, commercial and transport units
B11	Common wheat	21	Arable land
B12	Durum wheat	21	Arable land
B13	Barley	21	Arable land
B14	Rye	21	Arable land
B15	Oats	21	Arable land
B16	Maize	21	Arable land
B17	Rice	21	Arable land
B18	Triticale	21	Arable land
B19	Other cereals	21	Arable land
B21	Potatoes	21	Arable land
B22	Sugar beat	21	Arable land
B23	Other root crops	21	Arable land
B31	Sunflower	21	Arable land
B32	Rape and turnip rape	21	Arable land
B33	Soya	21	Arable land
B34	Cotton	21	Arable land
B35	Other fibre and oleaginous crops	21	Arable land
B36	Tobacco	21	Arable land
B37	Other non-permanent industrial crops	21	Arable land
B41	Dry pulses	21	Arable land
B42	Tomatoes	21	Arable land
B43	Other fresh vegetables	21	Arable land
B44	Floriculture and ornamental plants	21	Arable land
B45	Strawberries	21	Arable land

B51	Clovers	21	Arable land
B52	Lucerne	21	Arable land
B53	Other leguminous and mixtures for fodder	21	Arable land
B54	Mix of cereals	21	Arable land
B55	Temporary grassland	23	Pastures
B70	PERMANENT CROPS: FRUIT TREES	22	Permanent crops
B71	Apple fruit	22	Permanent crops
B72	Pear fruit	22	Permanent crops
B73	Cherry fruit	22	Permanent crops
B74	Nuts trees	22	Permanent crops
B75	Other fruit trees and berries	22	Permanent crops
B76	Oranges	22	Permanent crops
B77	Other citrus fruit	22	Permanent crops
B81	Olive groves	22	Permanent crops
B82	Vineyards	22	Permanent crops
B83	Nurseries	22	Permanent crops
B84	Permanent industrial crops	22	Permanent crops
BX1		22	Permanent crops
C10	Broadleaved woodland	31	Forests
C20	Coniferous woodland	31	Forests
C30	Mixed woodland	31	Forests
D10	Shrubland with sparse tree cover	32	Shrub and/or herbaceous vegetation association
D20	Shrubland without tree cover	32	Shrub and/or herbaceous vegetation association
E10	Grassland with sparse tree/shrub cover	23	Pastures
E20	Grassland without tree/shrub cover	23	Pastures
E30	Spontaneously re-vegetated surfaces	23	Pastures
F00	BARE LAND AND LICHENS/MOSS	33	Open spaces with little or no vegetation
G10	Inland water bodies	51	Continental waters
G20	Inland running water	51	Continental waters
H11	Inland marshes	41	Inland wetlands
H12	Peatbogs	41	Inland wetlands
H21	Salt marshes	42	Coastal wetlands
H22	Salines	42	Coastal wetlands

4. Compute geometric mean of Olsen-P by CORINE level 2 categories (corine_2_lc1) on the total LUCAS Topsoil dataset.
   CORINE 11 (urban fabric), 12 (industrial, commercial and transport units) categories: multiply it with 0.49 because 51% of the artifical surface are sealed (impermeable) and 49% are not sealed at average in the EU (Prokop *et al.*, 2011), therefore soil P content is relevant only for the 49% of these land cover categories.
5. Select relevant samples from LUCAS Topsoil Survey dataset for the CS based on NUTS region (country) and agro-climate zone (optional) based on the selection criteria that was defined during “Data preparation for mapping process” step.
6. Compute geometric mean of Olsen-P by CORINE level 2 categories on P values of the selected points and then overwrite underrepresented categories as follow:

   - CORINE category 11 (urban fabric), 12 (industrial, commercial and transport units): low number of samples for them, therefore geometric mean of category 11 and 12 computed on whole LUCAS Topsoil Survey dataset is used.
   - CORINE category 13 (mine, dump and construction sites), 14 (Artificial, non-agricultural vegetated areas): no data for them either in whole LUCAS topsoil, therefore geometric mean of category 11 and 12 is used, sealed areas are not dominant, thus not multiplied by 0.49.

- CORINE category 24 (heterogeneous agricultural areas): no data for them either in whole LUCAS topsoil, use geometric mean computed on all arable categories of the LUCAS dataset selected for the catchment.
- CORINE category 32 (scrub and/or herbaceous vegetation associations), 33 (open spaces with little or no vegetation), 51 (inland waters), 52 (marine waters) have low number of samples within the category, therefore add geometric mean computed on the whole non arable, non-artificial LUCAS dataset for these CORINE categories.
- CORINE 41 (inland wetlands) and 42 (coastal wetlands) category: add geometric mean computed on the whole LUCAS dataset for these categories.

All land use/ land cover categories must have a non NA value for soil P content, even water bodies (51, 52) to let SWAT+ run. Waterbodies are now routed differently in SWAT+, hence they are "removed" from the land-phase, this way it will not cause the problem that these categories have soil P contents.

7. Add mean Olsen-P values to the local land use map. First for the crop local land use map with the shape of the CS, than recode local land use categories into CORINE level 2 categories. The recoding table defined by the CS partners in "Data preparation for mapping process" is used (recode_l_lu_map_corine_2.xls). Add the geometric mean Olsen-P values to the land use raster layer (Figure 5). Visually check the Olsen-P content map and if needed, revise the recoding of the local land use categories into CORINE level 2 categories in "recode_l_lu_map_corine_2.xls".

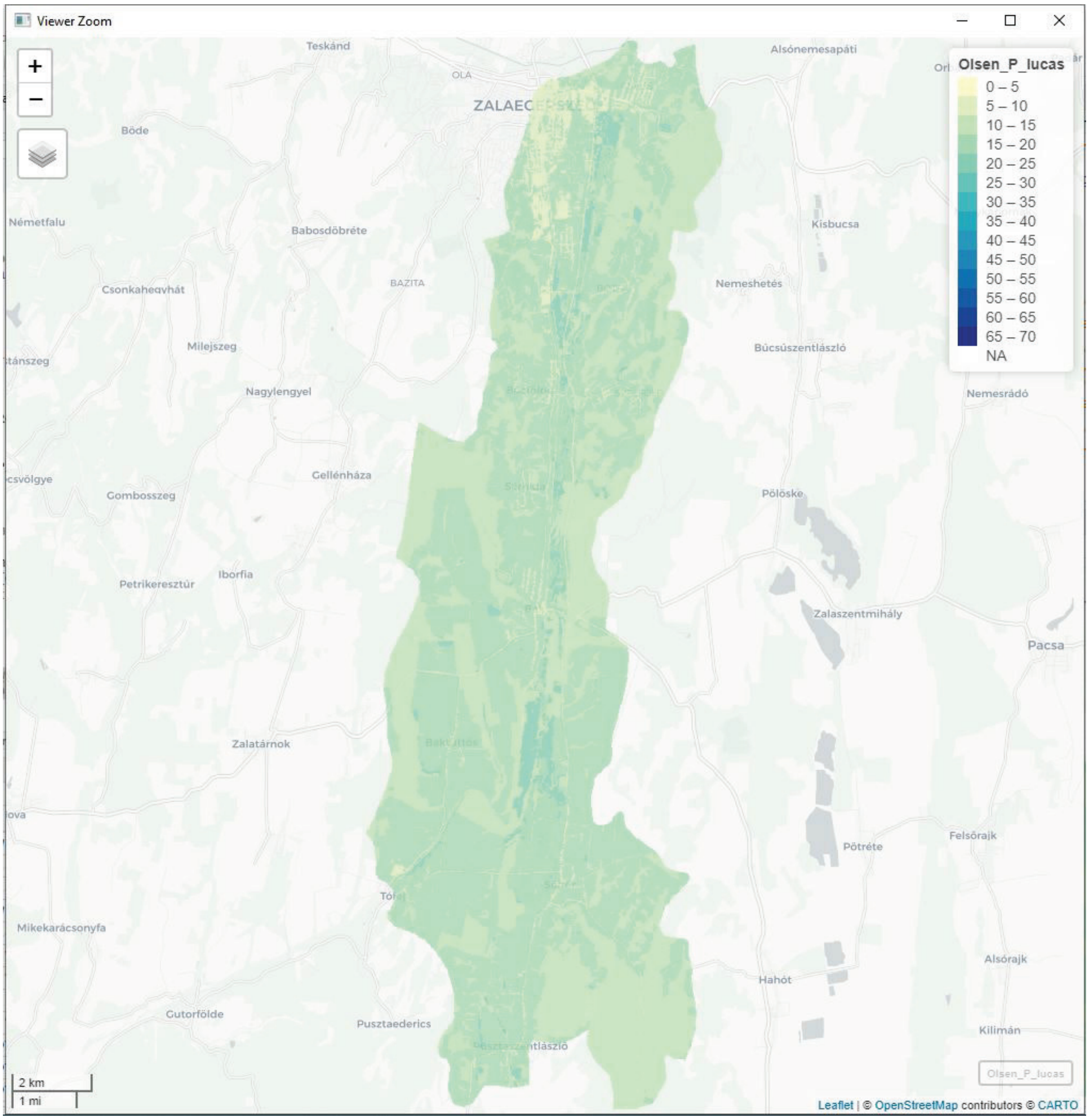


Figure 5: Map of soil Olsen-P content (ppm) for the Felső-Válicka case study (CS3b) based on LUCAS Topsoil data and local land use map (20 m resolution).

8. If some additional locally measured soil P content data is available for the CS (for example for agricultural fields), overwrite or replace the raster values of these fields to the locally measured data.

Example: we had measured soil P content for some crop fields at the Felső-Válicka CS. The data were available in a shape file, which was rasterized and merged with the Olsen-P content map of step 7 (Figure 6).

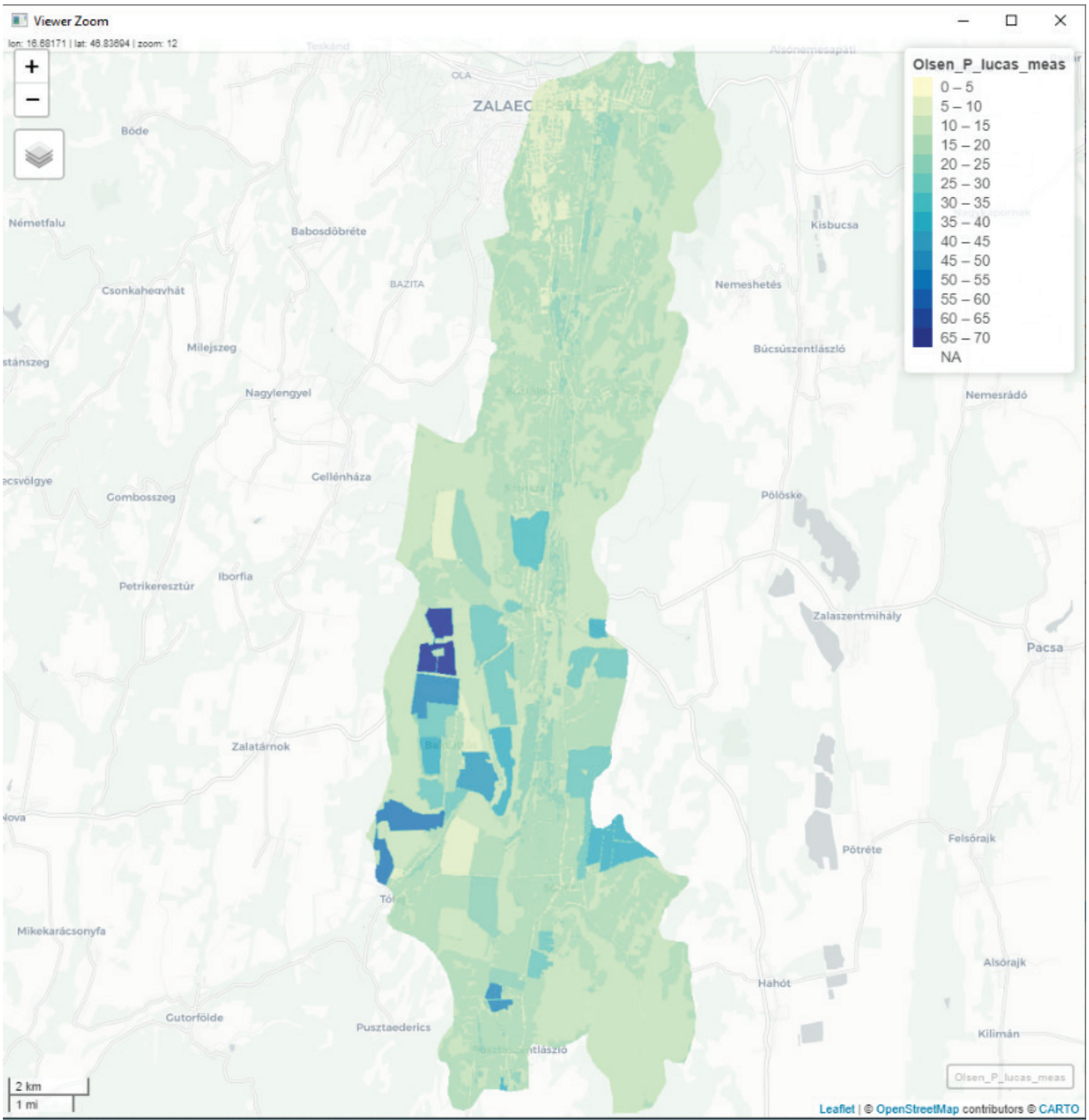


Figure 6: Final map of soil Olsen-P content (ppm) for the Felső-Válicka case study (CS3b) based on LUCAS Topsoil data, local land use map (20 m resolution) and locally measured soil P content.

9. Check the following objects in R:

   a) number of samples: check_table (Figure 7, see console at the bottom left),
   b) location of samples used for the computation: lucas_sel_map (Figure 7),
   c) created maps: Olsen_P_lucas_map; and if locally measured P was available: Olsen_P_lucas_meas_map.

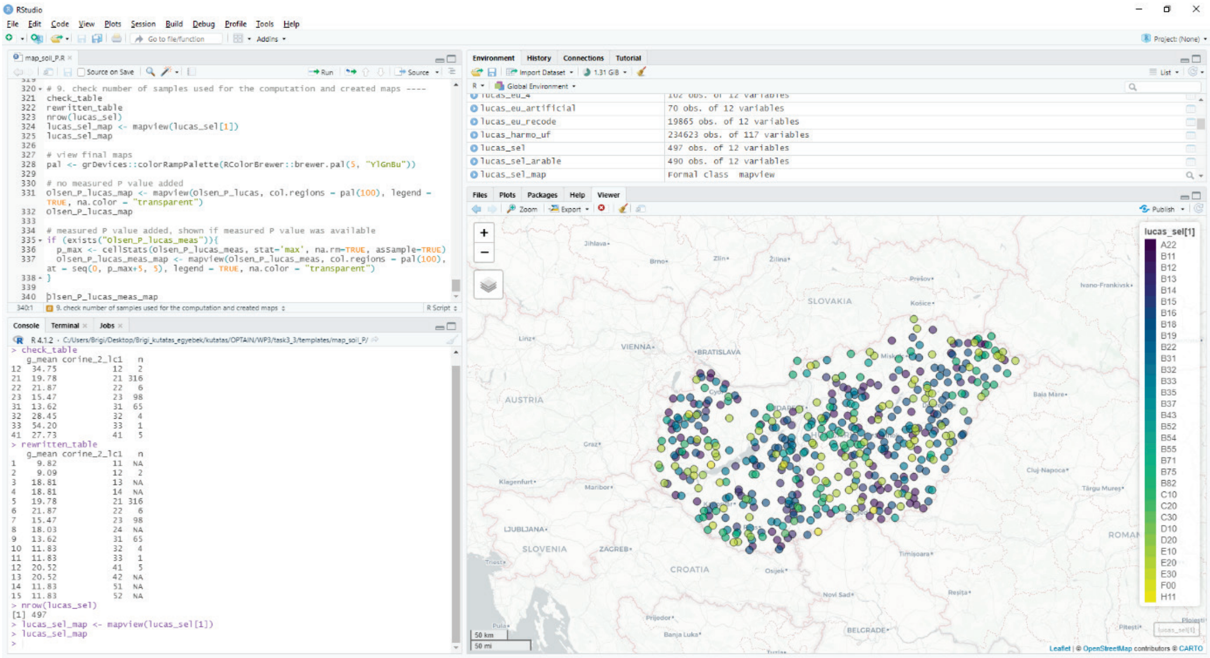
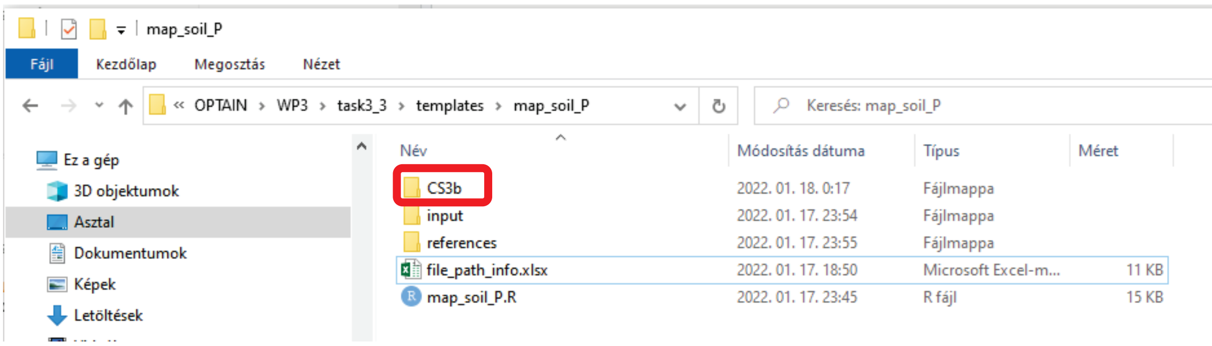


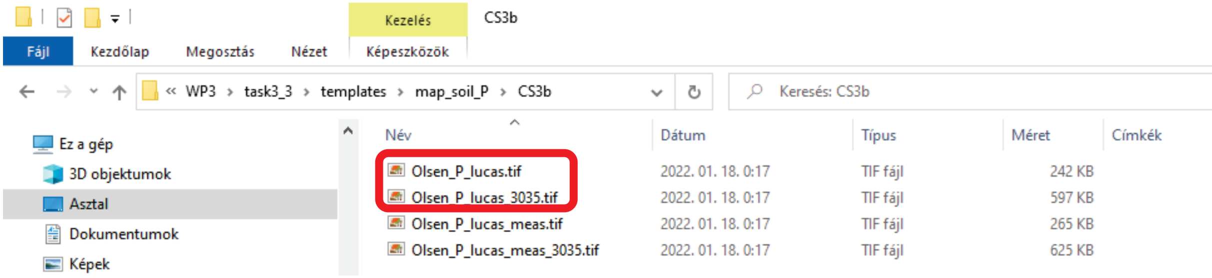
Figure 7: Number of samples with location and computed geometric mean of the soil Olsen-P content in the selected LUCAS Topsoil dataset by CORINE level 2 categories in the case of Felső-Válicka case study (CS3b).

Location of the derived maps

The derived soil Olsen-P content maps of the surface layer, i.e. 0-20 cm soil depth, are saved into a folder named after your CS, located under the folder of the map_soil_P.R file:



The derived maps are saved in the coordinate system (CRS) of the local land use map and in EPSG:3035, too.



If measured soil P content was available, the map including those values are saved as well as Olsen_P_lucas_meas.tif (CRS is the same as the local land use map) and Olsen_P_lucas_meas_3035.tif (CRS is EPSG:3035).

6.2.7. Remarks

If the number of samples within a land use/ land cover category seems to be low, please consider to revise the selection criteria of the LUCAS samples, correct cells of E11-18 and/or E20-24 of “file_path_info.xlsx”, rerun all the computations and contact WP3 for further support if needed.

Visually check the derived maps, if there are NA values, please check if all land use categories are correctly coded and translated into LUCAS land use/ land cover categories.

6.3. Guideline 3. Derivation of time series crop maps

6.3.1. Overview

For the analysis of the efficiency of Natural/Small Water Retention Measures (NSWRMs), time series about changing crop types and crop rotations on agricultural fields are required. This input data is not available for each case study (CS) and / or not for each modelling year. Therefore a remote sensing based approach has been derived to prepare crop maps for the target areas.

Hereinafter a guideline is provided for all CSs to produce a time series base map of crop types and add locally available crop data to the training dataset if those are available. The presented workflow uses open access software and datasets, but could be performed in any other software or use any relevant dataset that is accessible for the partners.

In SWAT+ information on crop types is required to simulate plant growth (PLANTS.PLT), set general land use properties (LANDUSE.LUM), characteristics of crop rotation (PLANT.INI), management operations (MANAGEMENT.SCH), SCS curve number (CNTABLE.LUM) and USLE P value (HRU-LTE.HRU). The main generic land use/ land cover categories distinguished by the model are agricultural land (arable land), orchard, forest, wetlands, hay, pasture, range and water (Table 1).

Table 1: SWAT Input/Output documentation, version 2016, Appendix A. Model databases.

Table A-2: Generic Land Covers included in database.

Name	Plant Code	Origin of Plant Growth Values	Plant Type
Agricultural Land-Generic	AGRL	use values for Grain Sorghum	warm season annual
Agricultural Land-Row Crops	AGRR	use values for Corn	warm season annual
Agricultual Land-Close-grown	AGRC	use values for Winter Wheat	cool season annual
Orchard	ORCD	use values for Apples	trees
Hay[‡]	HAY	use values for Bermudagrass	perennial
Forest-mixed	FRST	use values for Oak	trees
Forest-deciduous	FRSD	use values for Oak	trees
Forest-evergreen	FRSE	use values for Pine	trees
Wetlands	WETL	use values for Alamo Switchgrass	perennial
Wetlands-forested	WETF	use values for Oak	trees
Wetlands-nonforested	WETN	use values for Alamo Switchgrass	perennial
Pasture[‡]	PAST	use values for Bermudagrass	perennial
Summer pasture	SPAS	use values for Bermudagrass	perennial
Winter pasture	WPAS	use values for Fescue	perennial
Range-grasses	RNGE	use values for Little Bluestem (LAI_{max}=2.5)	perennial
Range-brush	RNGB	use values for Little Bluestem (LAI_{max}=2.0)	perennial
Range-southwestern US	SWRN	use values for Little Bluestem (LAI_{max}=1.5)	perennial
Water[*]	WATR		not applicable

In order to meet project's aim to analyse the efficiency of NSWRMs, we need to further specify agricultural land categories (AGRL, AGRR, AGRC, Table 1) into field-level time series of changing crop types for each CSs for the modelling period. This data is not always readily available for an entire agricultural catchment and the whole modelling period. It can also happen that information on crops are only available for agricultural blocks – which can include several fields – but not for single fields, which can hamper the analysis of nutrient and water retention in case of a field routing based simulation.

A promising method is to derive missing information on crop types from remote sensing data. Earth observation based crop mapping has been developed in many studies, such as by Phalke & Özdoğan (2018) and Teluguntla *et al.* (2018) who mapped based on Landsat data, Immitzer *et al.* (2016), Belgiu & Csillik (2018), Gumma *et al.* (2020) and

Defourny *et al.* (2019) that used Sentinel-2 data input, Kenduiywo *et al.* (2018) and d'Andrimont *et al.* (2021) who considered Sentinel-1 data. Orynbaikyzy *et al.* (2019) reviewed those crop mapping approaches, which were based on both optical and radar data. They found that most frequently cereals, oilseed and sugar crops are targeted during the identification.

For the field-level crop mapping the appropriate image resolution and short revisit period of the satellite sensor data is important. It is also preferable if the sensor has negligible dependency on atmospheric conditions, which can be obtained by using e.g. S1A and S1B polar-orbiting Synthetic Aperture Radar imagery (d'Andrimont *et al.*, 2021), which has 10 m resolution.

6.3.2. Required datasets to derive crop maps

The minimum datasets required for the crop classification are open access and listed hereinafter. However further adding local crop information can improve the prediction of crop types.

1. *LUCAS Land Use / Cover Area Frame Survey*

The Land Use / Cover Area frame statistical Survey (LUCAS) provides detailed in-situ land use/ land cover information of 270,000 ground truth points across all 28 EU member states. The land use and land cover information is structured according to a hierarchical classification scheme (EUROSTAT, 2015) and under cropland category it includes information about crop types as well. All together 84 sub-classes are differentiated in the dataset.

For deriving the crop classification algorithm the harmonised version of LUCAS 2015 and 2018 dataset (d'Andrimont *et al.*, 2020) is used as in-situ data.

2. *Sentinel-1 data*

For the identification of crop phenology we used Sentinel-1A and -1B satellite radar images based on d'Andrimont et al. (2021). The advantage of this remote sensing data is its cloud penetration capability, which enables its use on CSs with frequent cloud cover. The backscattering registration of a vertically transmitted microwave signal in a vertical (VV) and horizontal (VH) receiver and VH/VV ratio index were considered as predictors in the crop classification. According to the data catalogue of Google Earth Engine (GEE), the radar images are available from 3 October 2014. We acquired the images from 2015 spring.

3. *Field boundary map*

The field boundary map is required to avoid the salt-and-pepper effect of the pixel-based prediction of the crop types. After the crop type prediction, the crop maps are finalized by assigning only one crop by one agricultural field. That crop type is kept which has the majority in terms of area within the field.

6.3.3. Optional dataset to derive crop maps

Local crop data

For reaching higher accuracy in crop mapping it is advisable to use field level crop data if available for the CS. This ground truth data could be added to the LUCAS dataset, which is used as a basic training set. The use of local data in addition can secure that the training can be performed for those crop types not represented in the LUCAS dataset, e.g.: onion or parsley is highly underrepresented in LUCAS dataset, thus the model cannot be trained to predict the presence of those from remote sensing data. If further training data is added, it should be in CSV format having four columns, including:

- year – when the crop was grown, should be between 2014 and 2021 – ,
- latitude,
- longitude,
- lc1, which is the code of the crop grown, please find meaning of crop codes in "OPTAIN_cropmap_codes_lc1.xlsx" in Table 2 and the OPTAIN cloud ([link]).

Table 2: LUCAS and OPTAIN-optimized crop codes legend.

OPTAIN cropmap code	Name	LUCAS codes
1	Artificial land	A11, A12, A13, A21, A22
3	Wood- and shrubland	C10, C21, C22, C23, C31, C32, C33, D10, D20
5	Grassland	E10, E20, E30 + B55 (temporary grassland)
6	Bare land	F10, F20, F30, F40
7	Water surfaces	G10, G11, G12, G20, G21, G22, G30, G50
8	Wetlands	H11, H12, H21, H22, H23
11	Common wheat	B11
12	Durum wheat	B12
13	Barley	B13
14	Rye	B14
15	Oats	B15
16	Maize	B16
17	Rice	B17
19	Other cereals	B19
21	Potatoes	B21
22	Sugar beet	B22
23	Other root crops	B23
31	Sunflower	B31
32	Rape and turnip rape	B32
33	Soya	B33
34	Cotton	B34
35	Other fibre crops	B35
36	Tobacco	B36
37	Other non-permanent industrial crops	B37
41	Dry pulses	B41
42	Tomatoes	B42
43	Other fresh vegetables	B43
44	Floriculture and ornamental plants	B44
45	Strawberries	B45
51	Clovers	B51
52	Lucerne	B52
53	Other leguminous and mixtures for fodder	B53
54	Mixed cereals for fodder	B54
71	Apple fruit	B71
72	Pear fruit	B72
73	Cherry fruit	B73
74	Nuts trees	B74
75	Other fruit trees and berries	B75
76	Oranges	B76
77	Other citrus fruit	B77

Sample formatted input data can be accessed in the OPTAIN cloud as well ([link]; filename: "crop_data_samplefile.csv"). The formatted training data has to be sent to János Mészáros (meszaros.janos@atk.hu), to be incorporated in the Google Earth Engine calculations for your CS. A guideline to build the training dataset from vector data is provided in Annex Guideline 4 ("Generating local training dataset for crop classification").

6.3.4. Method for crop classification

For the selection of training data a 30 km buffer around the area of the CSs was applied. The buffer area was used to select the LUCAS 2015 and 2018 point data – relevant for each CSs – and Sentinel-1 images from year 2015 and 2018. The subclasses of the cropland category were kept in genuine LUCAS categories. Other than cropland categories were aggregated into broad land cover classes, such as 1-artifical lands, 3-woodland+shrubland, 5-grassland, 6-bare land, 7-water, and 8-wetlands. A detailed legend table can be accessed in the OPTAIN cloud ([link]; filename: "OPTAIN_cropmap_codes_lc1.xlsx"). Sentinel-1 data was extracted for the selected LUCAS points by each CS and used as a training dataset for the crop classification.

The crop classification analysis was performed in GEE. We acquired the Sentinel-1 data in the cloud computing platform, derived the prediction algorithm and applied it for the requested time period there. Time-series analysis was performed on the Sentinel-1 data between week 4th and 44th of each year, in periods of 4 weeks. We used the random forest algorithm to predict the crop types from the Sentinel-1 data. For the analysis of the model performance 10-fold cross-validation was applied. This approach provides a pixel-based crop classification. This prediction approach can be applied from 2015 up to now. By adding a field boundary map, the pixel based prediction can be overwritten by the majority of the predicted crop.

Accuracy metrics

On the user interface of the GEE platform, accuracy metrics are represented by the mean accuracy of the test, mean kappa of the tests, consumer accuracy, producer accuracy, total accuracy and confusion matrix. Information on variable importance can be accessed, too.

Software to use

The analysis can be performed in different programming environments. The script of deriving crop classification algorithm and mapping crops for the required modelling period is provided for the GEE programming environment (Gorelick *et al.*, 2017).

Main steps to derive time-series crop maps

I. Setup and data pre-processing
   1. Import required CS data, such as shape of the catchment, LUCAS dataset, field boundary map.
   2. Define starting and ending week of the year for the crop season, local buffer radius (in meter) around CS polygon to define area where LUCAS points are collected for training data, length of period in weeks to collect satellite image for and slice whole period and year of training data.
   3. Collect Sentinel-1 image for defined years.

4. Pre-process Sentinel-1 image and add polarization ratio bands (VV, VH, VH/VV).
5. Clip satellite images for the CS plus buffer area.

II. Training data collection

6. Select training data from the LUCAS for the CS.
7. Compute median of the image bands for the set time-spans.
8. Overlay selected LUCAS points with the satellite images.

III. Building and testing Random Forest model

9. Build random forest-based crop classification model with 10-fold cross validation.
10. Compute and export error metrics, confusion matrix on the test sets and variable importance derived during the training.

IV. Classification on defined period

11. Generate crop maps for the defined years
12. Combine crop maps into a single multiband image

V. Visualization

VI. Export the outputs

6.3.5. Remarks

The GEE script to build and apply the crop classification model is available at https://doi.org/10.5281/zenodo.6669643 and at [Google EarthEngine link] after registration at the GEE. The field boundary map is required to derive more accurate crop maps.

Saved output of the crop classification:

- detailed accuracy metrics in CSV format,
- multiband image (CRS: EPSG 3035) of crop maps in GeoTIFF format (band names are corresponding years for easier identification)

For Petite Glane (CS2), La Wimbe (CS7) and Cherio (CS9) CSs Sentinel-1 data is available only for 2017-2021 period, thus the crop maps can be derived only for those years. For Petite Glane (CS2) and Hobol (CS10) no local data is available in the LUCAS dataset. Local observation and/or data from an EU member state with similar pedoclimatic conditions could be used, possible solution is under preparation. For the other CSs the period 2015-2021 can be considered, adding further training data is optional.

For further specifying agricultural land (AGRL, AGRR and AGRC) into crop types the following option could be considered by the CS Leaders:

a) use available crop maps,
b) derive crop maps based on remote sensing data based on workflow provided by WP3,
c) combine option a) and b).

6.4. Guideline 4. Generating local training dataset for crop classification.

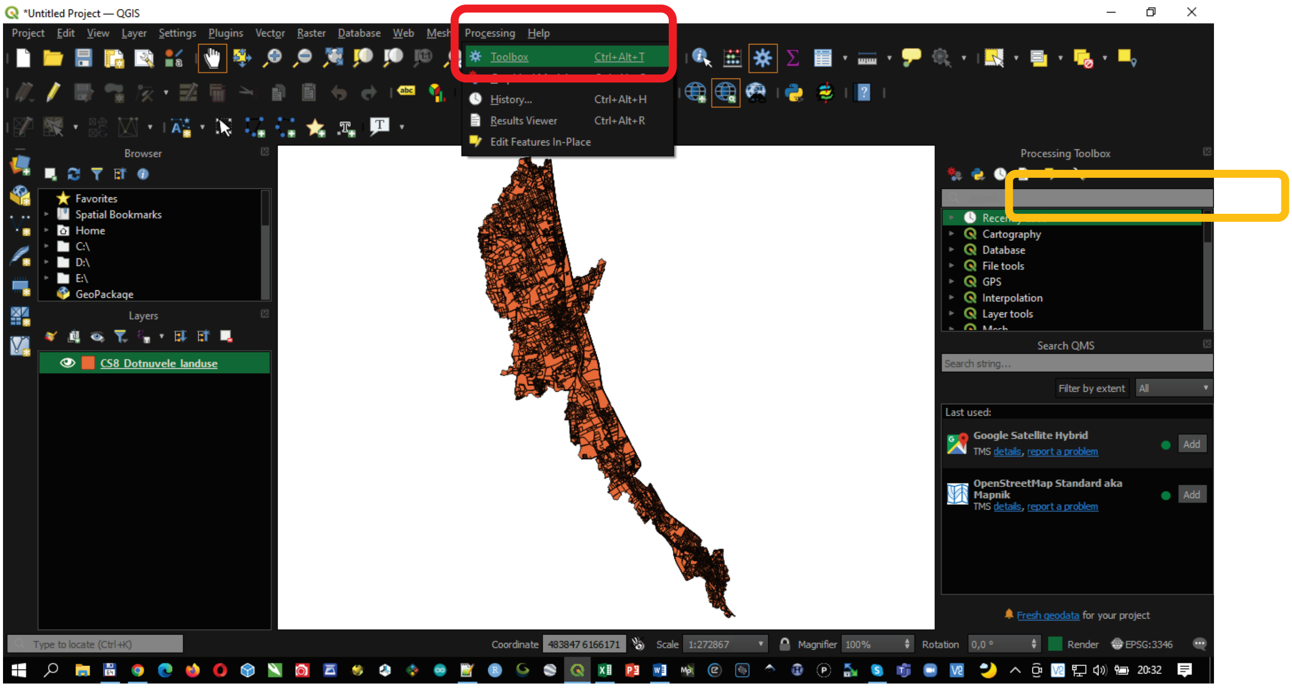
Generating random sample points for crop type polygons

Step-by-step guide using QGIS – we assume that users have a general knowledge of QGIS and we only focus on the specific tasks and steps to generate additional training data to LUCAS crop data.

In this tutorial we will use the Dotnuvele CS08 land use layers as example.

1. Initial state – QGIS with the loaded polygon layer containing crop types.

For further steps the “**Processing**” toolbox will be used, so ensure it is activated (“**Processing**” menu item -> “**Toolbox**” or Ctrl+Alt+T). It appears on the right side of the main frame of QGIS.



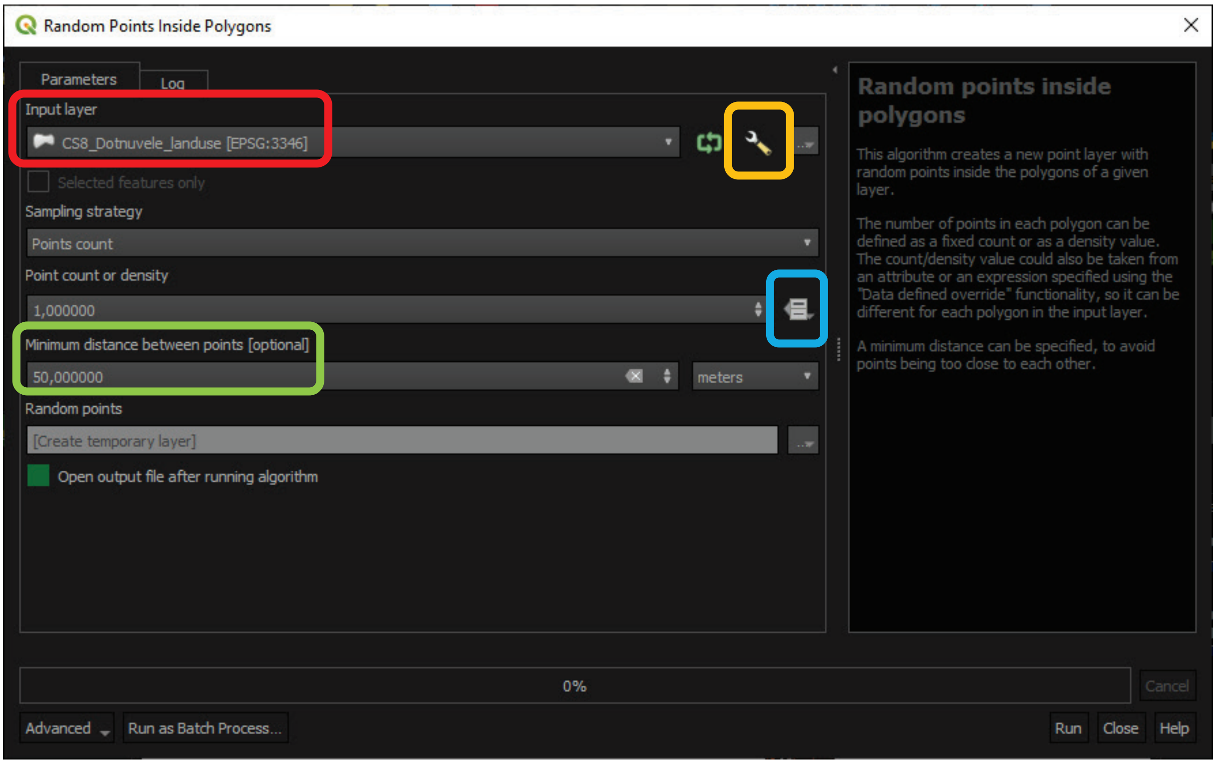
2. In the algorithm search frame of the “**Processing Toolbox**” window search for the “**Random points inside polygons**” algorithm to generate random points inside polygons with minimal distance and size constraints. Initiate it and define parameters this way:

Red rectangle – input polygon layer (shapefile)

Orange rectangle – input geometry error settings (details in 2a section)

Blue rectangle – point count rule (details in 2b section)

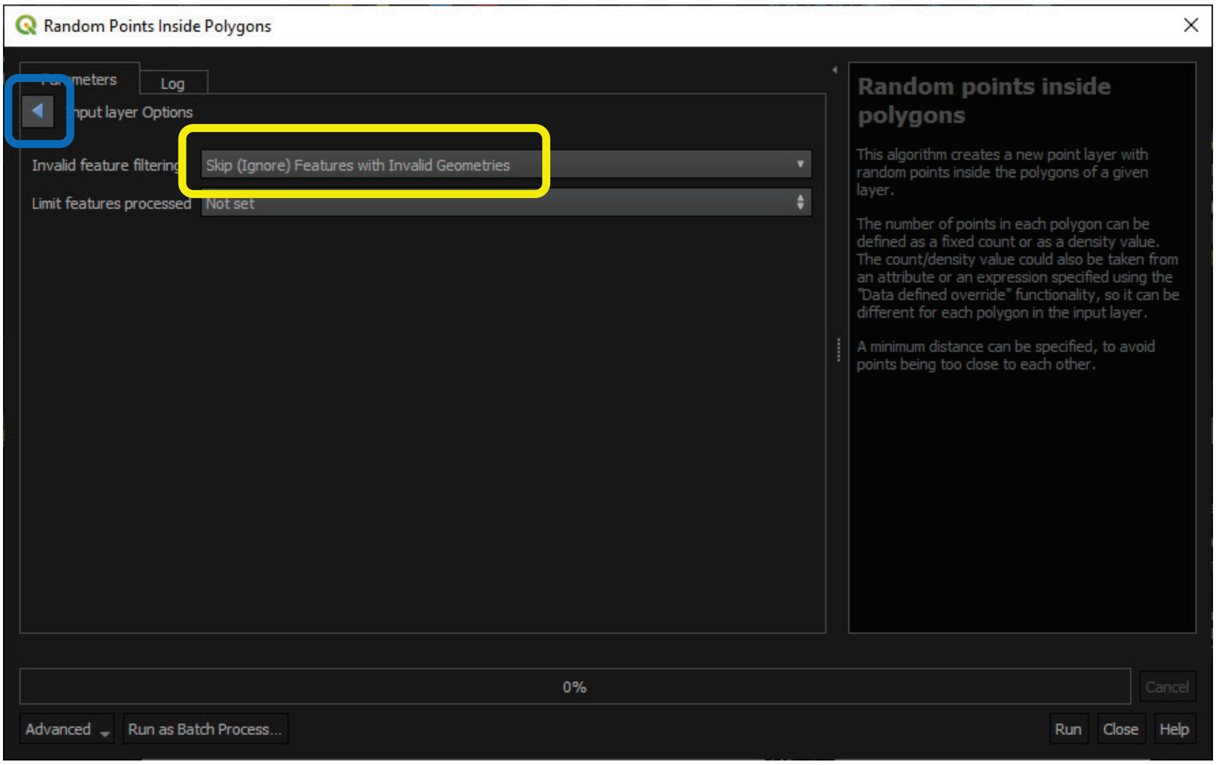
Green rectangle – minimum distance between points, please take into consideration that spatial resolution of satellite images used for crop mapping is 10 m, values higher than this value are all good (e.g. 20 or 50 m).



2.a Fixing geometry frame

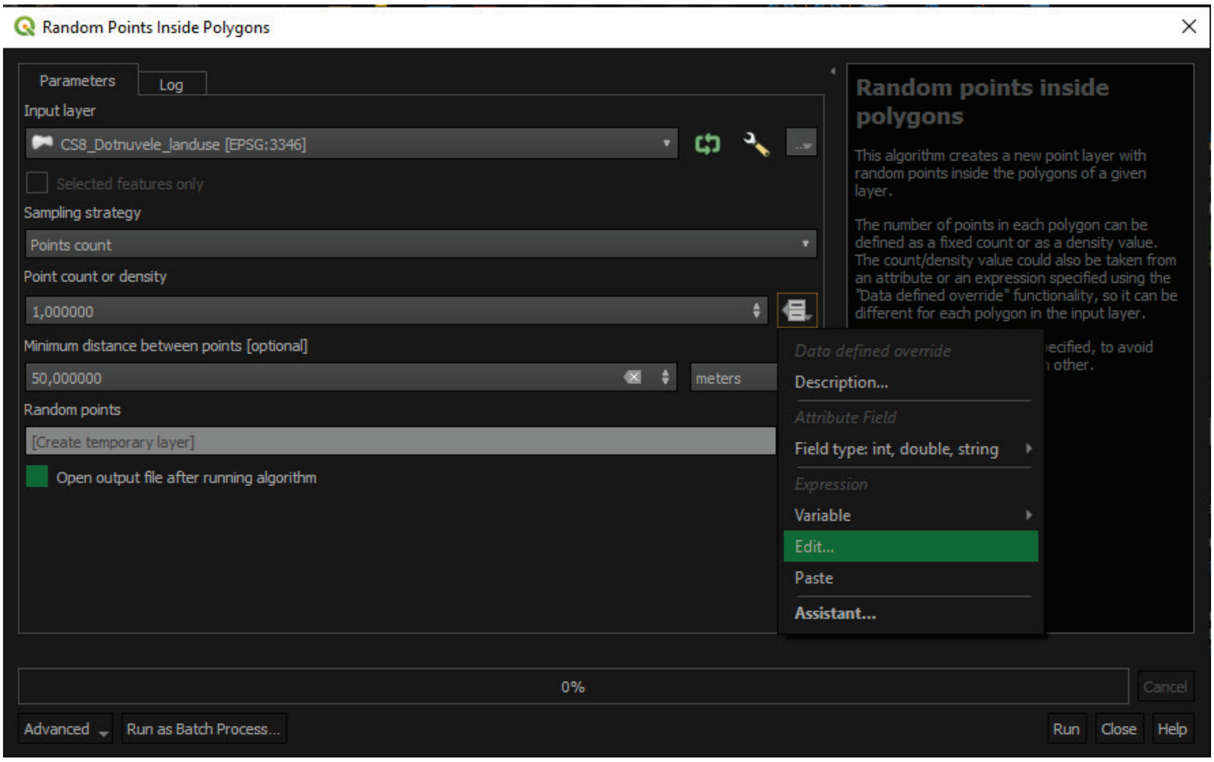
Set “**Invalid feature filtering**” parameter to “**Skip (ignore) features with invalid Geometries**” (or edit your geometries with classical geometry editing tools in case of problematic polygons prior to generating random points).

Go back to regular parameters with the blue arrow button:



2.b Definition of ruleset how to place random points inside a polygon

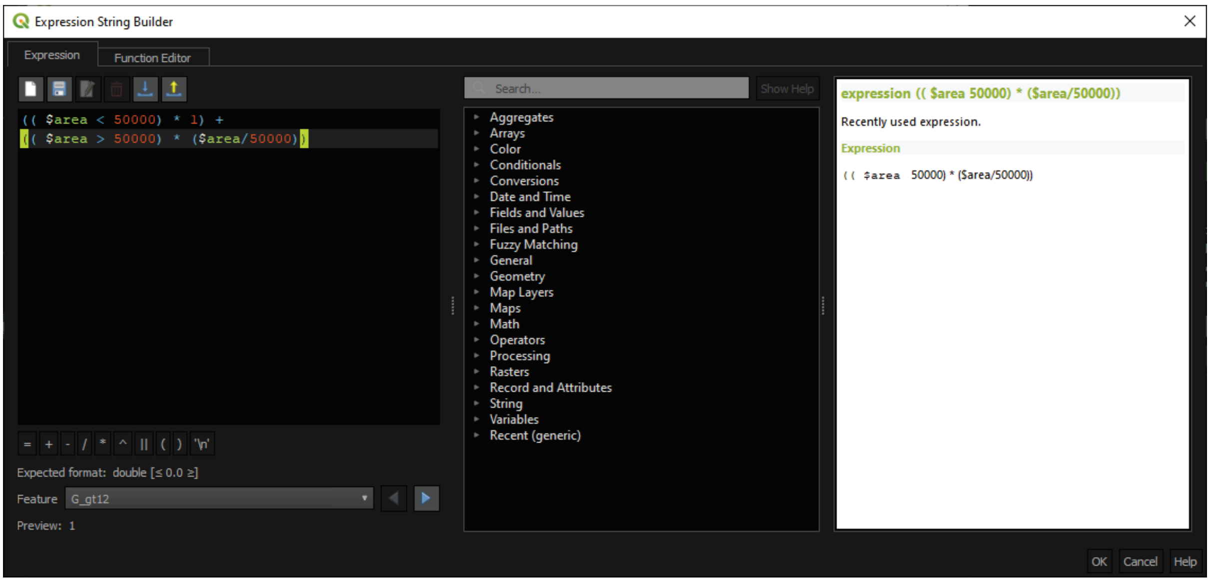
If you click on the  button, drop-down list will appear -> choose "**Edit**" and the "**Expression String Builder**" window will pop up.



2.c Copy (and edit if needed) the below formulae into the "**Expression**" frame on the left-most side of the window:

(($area < 50000) * 1) +

(($area > 50000) * ($area/50000))

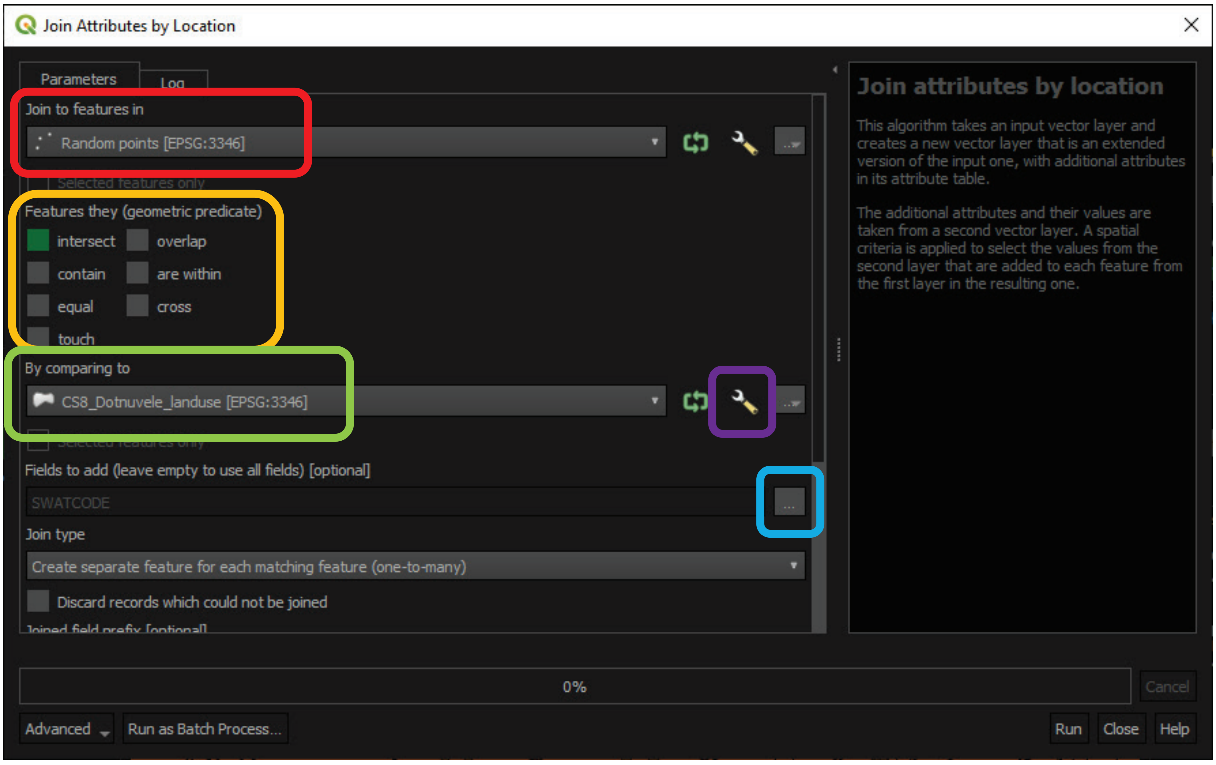


The formulae defines a rule set that if a polygon is smaller than 50.000 m^2, then only one point will be generated. However, as a second case if polygon is bigger than given threshold, the number of generated points depends on the area of the polygon (one point/5 ha). Feel free to change the threshold if needed in your case.

Click on "**OK**" to close "**Expression string builder**" and on "**Run**" to initiate the random point generation, resulting points will be seen in the main map frame of the QGIS.

3. Sampling the crop type in the random points with an intersection tool

Again, in the "**Processing toolbox**" search frame, look for the "**Join attributes by location**" algorithm and start it. In the pop-up window set its parameters as in the example below:

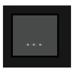


Red rectangle – input random points layer generated in the previous step

Orange rectangle – geometry constraint (intersect)

Green rectangle – crop type polygons layer to sample

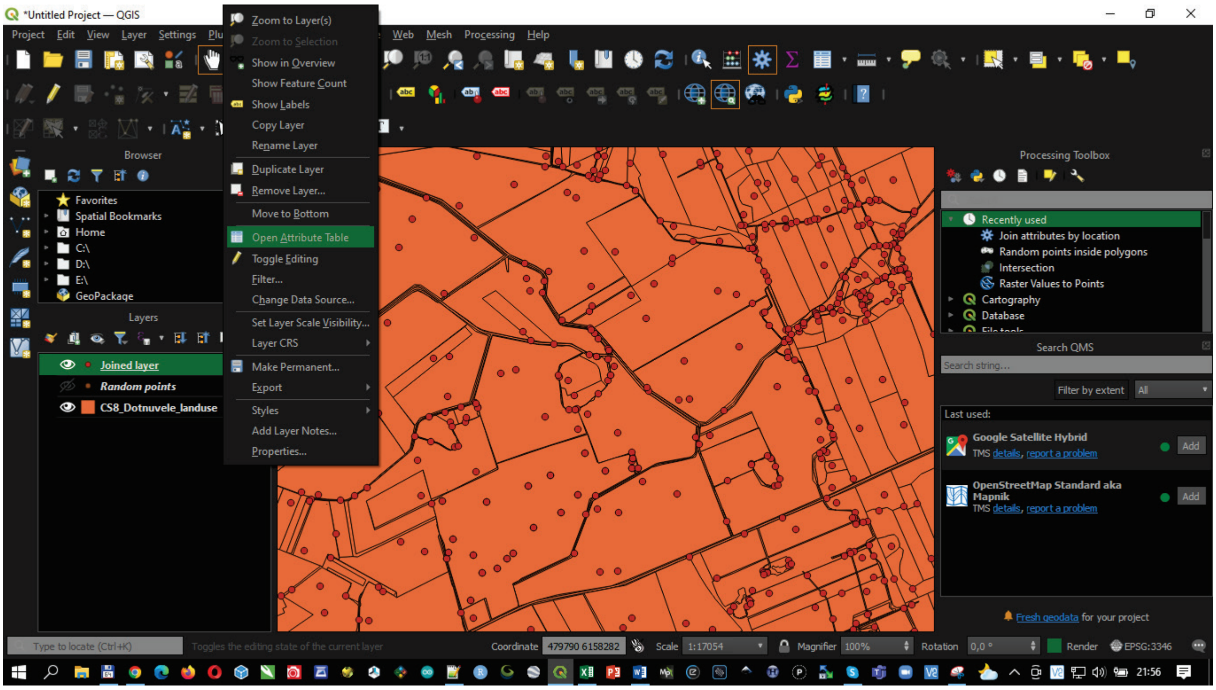
Purple rectangle – same as in the 1. step to skip polygons with bad geometry

Blue rectangle – clicking on the  button will show the list of existing attributes of the crop type polygons layer, choose that needed to sample and transfer to the random points.

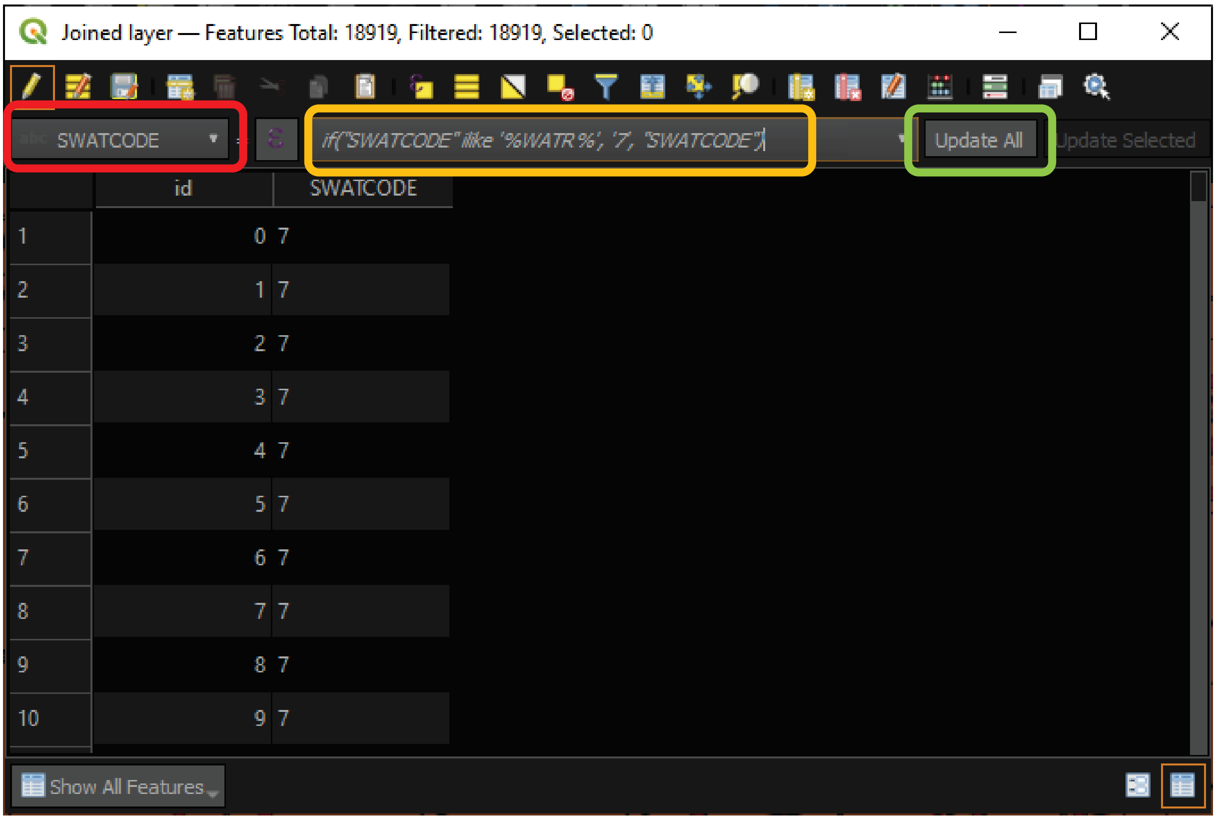
Clicking on "**Run**" will generate a second layer with the same points as the first layer but with different attributes.

4. Filtering and changing the crop types or their codes to the OPTAIN-optimized codes (see Appendix B for detailed list of codes)

Right-click on the layer name generated in the previous step (probably "**Joined layer**") and choose the "**Open attribute table**" option in the pop-up list.



Make the layer "**Editable**" with the yellow pencil button () and change the attribute name (red rectangle) to the crop type attribute in the list (we will modify its values) [please don't be confused with the "SWATCODE" attribute, it is just an example now].



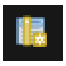
In the frame (orange rectangle) copy the expression string below, modify it according to your needs and finally click on the "**Update all**" button:

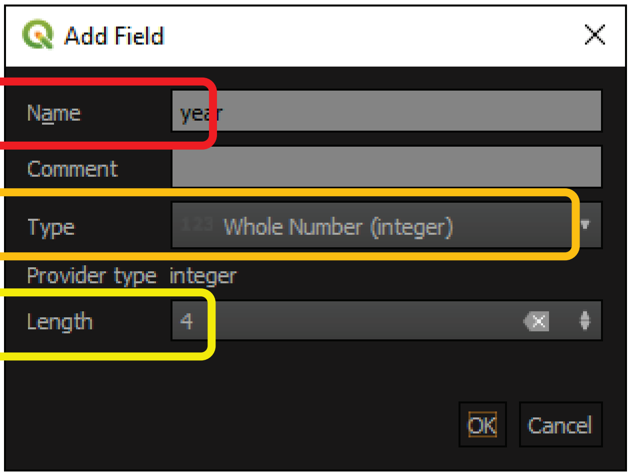
if("SWATCODE" ilike 'WATR', '7', "SWATCODE")

This expression check each rows in the "SWATCODE" attribute (in the example), if the value string is equal to the given string, it will be changed to "7" (code for waterbodies in the OPTAIN-optimized crop type/land use table); if not, it remains the original.

Please repeat this step with all your codes to transform them according to the OPTAIN-optimized codes, changing the string (e.g. WATR) to the OPTAIN-code (e.g. 7).

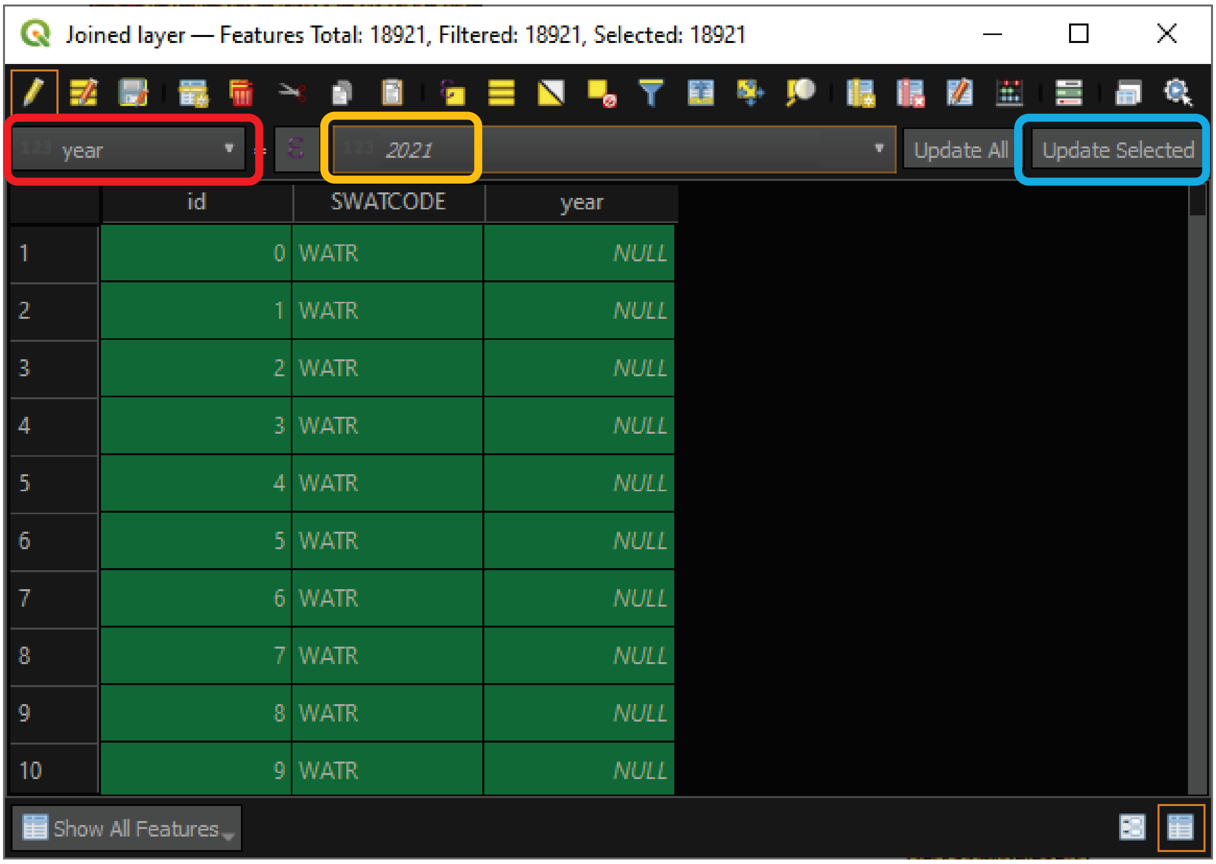
5. Add year as new attribute

A new attribute will be added to "**Attribute table**" in this step. Clicking on the  button a pop-up window will show up:



Please fill in the “**Name**” field as ‘**year**’, “**Type**” as ‘**Whole Number (integer)**’ and finally the “**Length**” as **4** (it tells QGIS to store values on four digits). Finally, click on “**OK**” and you will get a new column in the attribute table with NULL values.

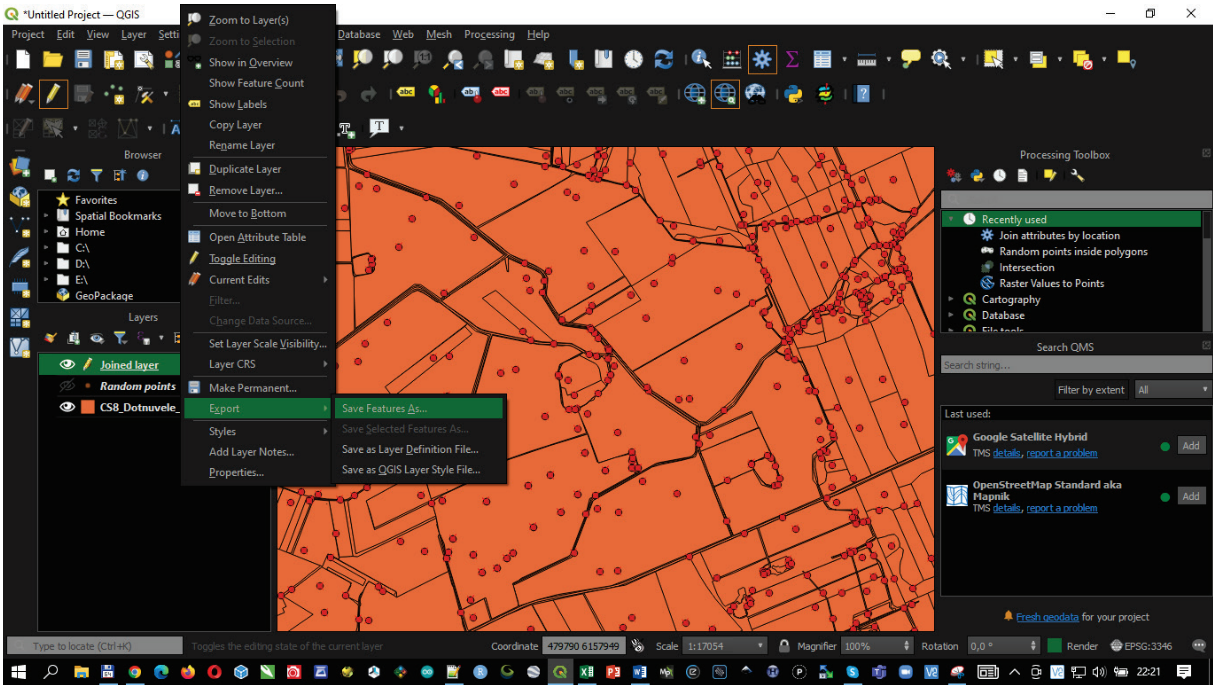
All points from the same year should be selected manually if you work on several different years data, or if all points belong to the same year, by ‘**CTRL+A**’ hotkey. In the top part of “**Attribute table**” choose from the drop-down list the ‘**year**’ field, type in the actual year number (e.g. 2021) and click on “**Update Selected**” button:



After all those changes, save editing with the  button and close the “**Attribute table**” window.

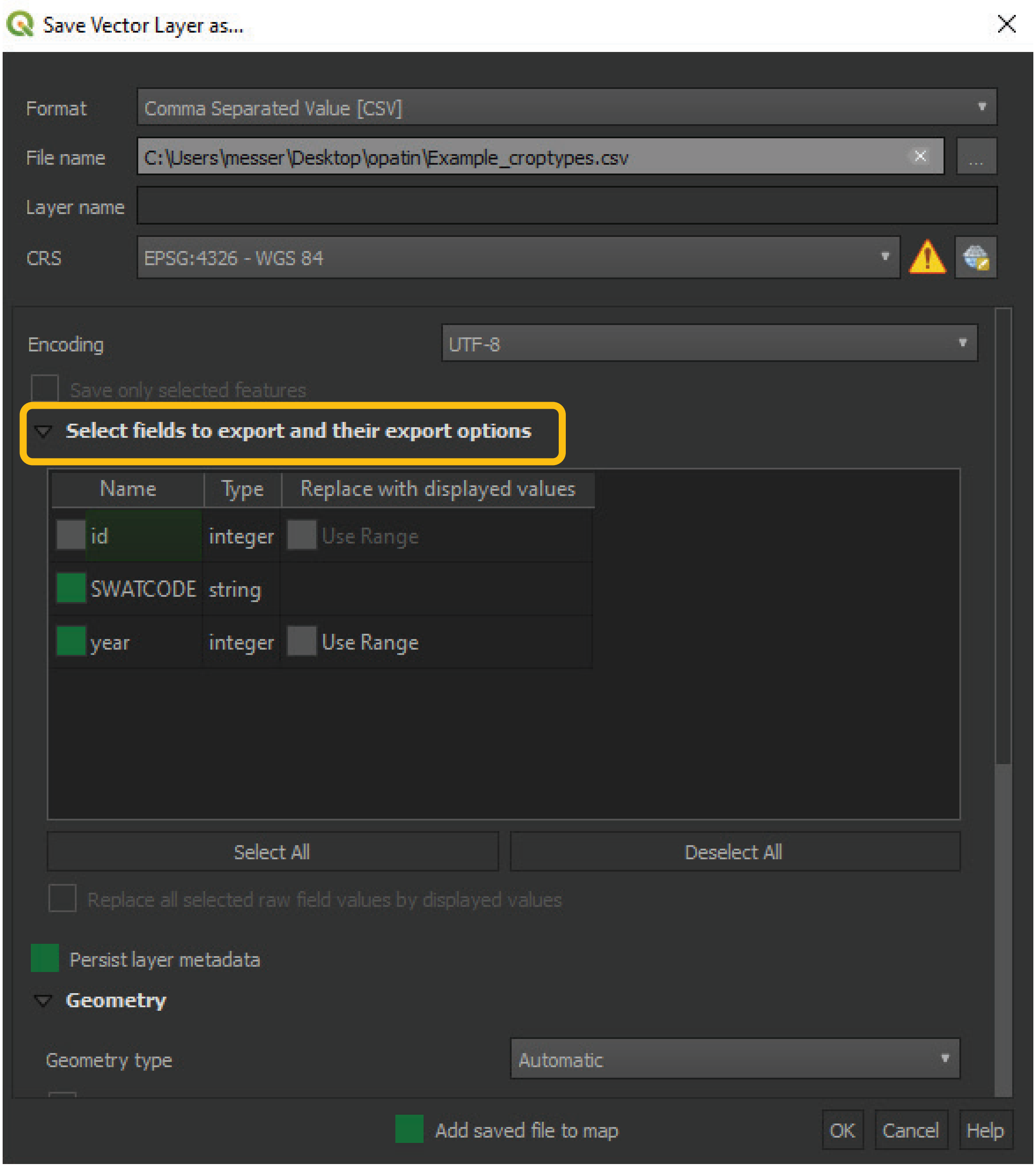
6. Save the layer as text file

Right-click again on the layer name and choose "**Export**" -> "**Save features as…**" options.

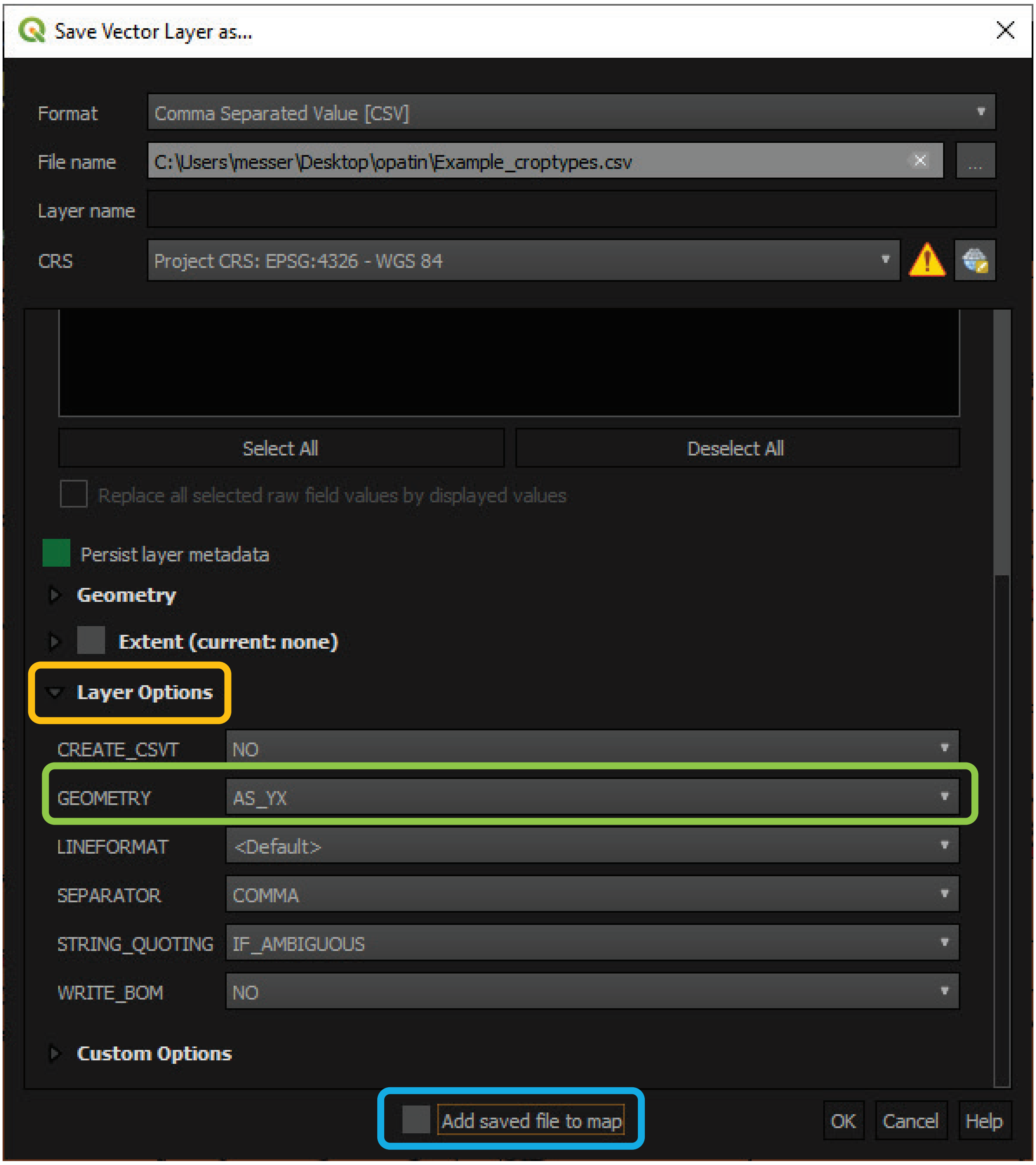


In the pop-up window define the file format ("**Comma Separated Value [CSV]**"), the new file name and its path on your machine (use … button to browse folders) and Coordinate Reference System as "**WGS84 – EPSG: 4326**".

Open down the "**Select fields to export and their export options**" menu and choose the attributes storing the crop type and year information.



Open down the "**Layer options**" menu, in the "**GEOMETRY**" parameter set it "**AS_YX**" and in the bottom of the window **switch off** the "**Add saved file to map**" option to not import the file into QGIS, only to save on your hard drive.



As a final step, please send the created CSV file to **János Mészáros** – **meszaros.janos@atk.hu**

7. References

Abbaspour, K.C., AshrafVaghefi, S., Yang, H. & Srinivasan, R. 2019. Global soil, landuse, evapotranspiration, historical and future weather databases for SWAT Applications. *Scientific Data*, **6:263**.

Appelhans, T., Detsch, F., Reudenbach, C. & Woellauer, S. 2021. mapview: Interactive Viewing of Spatial Data in R. R package version 2.10.0. (At: https://cran.r-project.org/package=mapview.).

Arnold, J.G., Kiniry, J.R., Srinivasan, R., Williams, J.R., Haney, E.B. & Neitsch, S.L. 2012. Soil & Water Assessment Tool, Input/Output Documenatation. 654.

Assouline, S. & Or, D. 2014. The concept of field capacity revisited: Defining intrinsic static and dynamic criteria for soil internal drainage dynamics. *Water Resources Research*, **50**, 4787–4802.

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L. & Panagos, P. 2019. Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. *Geoderma*, **355**, 113912, (At: https://doi.org/10.1016/j.geoderma.2019.113912.).

Belgiu, M. & Csillik, O. 2018. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. *Remote Sensing of Environment*, **204**, 509–523, (At: https://doi.org/10.1016/j.rse.2017.10.005.).

Bivand, R., Keitt, T. & Rowlingson, B. 2021. rgdal: Bindings for the "Geospatial" Data Abstraction Library. R package version 1.5-27. (At: https://cran.r-project.org/package=rgdal.).

Ceglar, A., Zampieri, M., Toreti, A. & Dentener, F. 2019. Observed Northward Migration of Agro-Climate Zones in Europe Will Further Accelerate Under Climate Change. *Earth's Future*, **7**, 1088–1101.

Cornelis, W.M., Ronsyn, J., Van Meirvenne, M. & Hartmann, R. 2001. Evaluation of Pedotransfer Functions for Predicting the Soil Moisture Retention Curve. *Soil Science Society of America Journal*, **65**, 638–648.

Costa, M.G., Gama-Rodrigues, A.C., Gonçalves, J.L. de M., Gama-Rodrigues, E.F., Sales, M.V. da S. & Aleixo, S. 2016. Labile and non-labile fractions of phosphorus and its transformations in soil under Eucalyptus plantations, Brazil. *Forests*, **7**, 1–15.

d'Andrimont, R., Verhegghen, A., Lemoine, G., Kempeneers, P., Meroni, M. & van der Velde, M. 2021. From parcel to continental scale – A first European crop type map based on Sentinel-1 and LUCAS Copernicus in-situ observations. *Remote Sensing of Environment*, **266**.

d'Andrimont, R., Yordanov, M., Martinez-Sanchez, L., Eiselt, B., Palmieri, A., Dominici, P., Gallego, J., Reuter, H.I., Joebges, C., Lemoine, G. & van der Velde, M. 2020. Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European Union. *Scientific Data*, **7**, 1–15.

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Nicola, L., Rabaute, T., Savinaud, M., Udroiu, C., Valero, S., Bégué, A., Dejoux, J.F., El Harti, A., Ezzahar, J., Kussul, N., Labbassi, K., Lebourgeois, V., Miao, Z., Newby, T., Nyamugama, A., Salh, N., Shelestov, A., Simonneaux, V., Traore, P.S., Traore, S.S. & Koetz, B. 2019. Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. *Remote Sensing of Environment*, **221**, 551–568, (At: https://doi.org/10.1016/j.rse.2018.11.007.).

Egnér, H., Riehm, H. & Domingo, W.R. 1960. Untersuchungen über die chemische Bodenanalyse als Grundlage für die Beurteilung des Nährstoffzustandes der Böden. II. Chemische Extraktionsmethoden zur Phosphor- und Kaliumbestimmung. *Lantbr. Ann.*, **26**, 199-215.

EUROSTAT. 2015. *LUCAS 2015 (Land Use/Cover Area Frame Survey) – Technical Reference Document: C3 Classification (Land Use and Land Cover). Technical Report.*

van Genuchten, M.T. 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, **44**, 892–898.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. & Moore, R. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, **202**, 18–27, (At: https://doi.org/10.1016/j.rse.2017.06.031.).

Guarracino, L. 2007. Estimation of saturated hydraulic conductivity Ks from the van Genuchten shape parameter α. *Water Resources Research*, **43**, 15–18.

Gumma, M.K., Tummala, K., Dixit, S., Collivignarelli, F., Holecz, F., Kolli, R.N. & Whitbread, A.M. 2020. Crop type identification and spatial mapping using Sentinel-2 satellite data with focus on field-level information. *Geocarto International*, **0**, 000, (At: http://dx.doi.org/10.1080/10106049.2020.1805029.).

Hastie, T., Tibshirani, R. & Friedman, J. 2009. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Second edi. Springer. (At: https://web.stanford.edu/~hastie/ElemStatLearn//printings/ESLII_print10.pdf. Accessed: 19/11/2018).

Hijmans, R.J. 2021. raster: Geographic Data Analysis and Modeling. R package version 3.5-2. (At: https://cran.r-project.org/package=raster.).

Immitzer, M., Vuolo, F. & Atzberger, C. 2016. First experience with Sentinel-2 data for crop and tree species classifications in central Europe. *Remote Sensing*, **8**.

Jones, C.A., Sharpley, A.N. & Williams, J.R. 1984. A Simplified Soil and Plant Phosphorus Model: III. Testing. *Soil Science Society of America Journal*, **48**, 810–813, (At: https://onlinelibrary.wiley.com/doi/full/10.2136/sssaj1984.03615995004800040022x. Accessed: 28/1/2022).

Kenduiywo, B.K., Bargiel, D. & Soergel, U. 2018. Crop-type mapping from a sequence of Sentinel 1 images. *International Journal of Remote Sensing*, **39**, 6383–6404, (At: https://doi.org/10.1080/01431161.2018.1460503.).

Nasta, P., Szabó, B. & Romano, N. 2021. Evaluation of pedotransfer functions for predicting soil hydraulic properties : A voyage from regional to field scales across Europe. *Journal of Hydrology: Regional Studies*, **37**, 100903.

Neitsch, S.L., Arnold, J.G., Kiniry, J.R. & Williams, J.R. 2009. Soil and Water Assessment Tool Theoretical Documentation—Version 2009. 618.

Nemes, A., Wösten, J., Lilly, A. & Voshaar, J.O. 1999. Evaluation of different procedures to interpolate particle-size distributions to achieve compatibility within soil databases. *Geoderma*, **90**, 187–202, (At: http://www.sciencedirect.com/science/article/pii/S0016706199000142. Accessed: 7/5/2013).

Olsen, R., Cole, C.V., Watanabe, F.S. & Dean, L.A. 1954. *Estimation of available phosphorus in soils by extraction with sodium bicarbonate.* Washington DC.

Orynbaikyzy, A., Gessner, U. & Conrad, C. 2019. Crop type classification using a combination of optical and radar remote sensing data: a review. *International Journal of Remote Sensing*, **40**, 6553–6595, (At: https://doi.org/10.1080/01431161.2019.1569791.).

Pebesma, E. 2018. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, **10**, 439–446, (At: https://doi.org/10.32614/RJ-2018-009.).

Phalke, A.R. & Özdoğan, M. 2018. Large area cropland extent mapping with Landsat data and a generalized classifier. *Remote Sensing of Environment*, **219**, 180–195.

Poggio, L., De Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E. & Rossiter, D. 2021. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, **7**, 217–240.

Prokop, G., Jobstmann, H. & Schönbauer, A. 2011. *Report on best practices for limiting soil sealing and mitigating its effects.* Technical. European Commission, European Communities. (At: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Overview+of+best+practices+for+limiting+soil+sealing+or+mitigating+its+effects+in+EU-27#0.).

R Core Team. 2021. R: A language and environment for statistical computing. (At: https://www.r-project.org/.).

Romano, N. & Palladino, M. 2002. Prediction of soil water retention using soil physical data and terrain attributes. *Journal of Hydrology*, **265**, 56–75.

Ross, N. 2020. fasterize: Fast Polygon to Raster Conversion. R package version 1.0.3. (At: https://cran.r-project.org/package=fasterize.).

Sárdi, K., Csathó, P. & Osztoics, E. 2009. Evaluation of soil phosphorus contents in long-term experiments from environmental aspects. In: *Proceedings of the 51st Georgikon Scientific Conference*, pp. 807-815. Keszthely, Hungary.

Schaap, M.G., Leij, F.J. & van Genuchten, M.T. 2001. Rosetta: a Computer Program for Estimating Soil Hydraulic Parameters With Hierarchical Pedotransfer Functions. *Journal of Hydrology*, **251**, 163–176, (At: http://linkinghub.elsevier.com/retrieve/pii/S0022169401004668.).

Sharpley, A.N., Jones, C.A., Gray, C. & Cole, C. V. 1984. A Simplified Soil and Plant Phosphorus Model: II. Prediction of Labile, Organic, and Sorbed Phosphorus. *Soil Science Society of America Journal*, **48**, 805–809, (At: https://onlinelibrary.wiley.com/doi/full/10.2136/sssaj1984.03615995004800040021x. Accessed: 28/1/2022).

Sharpley, A.N. & Williams, J.R. 1990. *EPIC — Erosion / Productivity Impact Calculator: 1. Model Documentation.*

Steinfurth, K., Hirte, J., Morel, C. & Buczko, U. 2021. Conversion equations between Olsen-P and other methods used to assess plant available soil phosphorus in Europe – A review. *Geoderma*, **401**, 115339.

Szabó, B., Gyurkó, D., Weynants, M. & Weber, T.K.D. 2019. Web interface for European hydraulic pedotransfer functions (euptfv2). (At: https://ptfinterface.rissac.hu/. Accessed: 23/4/2020).

Szabó, B., Weynants, M. & Weber, T.K. 2021. Updated European Hydraulic Pedotransfer Functions with Communicated Uncertainties in the Predicted Variables (euptfv2). *Geosci. Model Dev.*, **14**, 151–175, (At: https://doi.org/10.5194/gmd-14-151-2021.).

Teluguntla, P., Thenkabail, P., Oliphant, A., Xiong, J., Gumma, M.K., Congalton, R.G., Yadav, K. & Huete, A. 2018. A 30-m landsat-derived cropland extent product of Australia and China using random forest machine learning algorithm on Google Earth Engine cloud computing platform. *ISPRS Journal of Photogrammetry and Remote Sensing*, **144**, 325–340, (At: https://doi.org/10.1016/j.isprsjprs.2018.07.017.).

Tóth, G., Guicharnaud, R.-A., Tóth, B. & Hermann, T. 2014. Phosphorus levels in croplands of the European Union with implications for P fertilizer use. *European Journal of Agronomy*, **55**, 42–52, (At: http://linkinghub.elsevier.com/retrieve/pii/S1161030113001950. Accessed: 5/3/2014).

Tóth, G., Jones, A. & Montanarella, L. 2013a. *LUCAS Topsoil Survey. Methodology, data and results.* Publications Office of the European Union, Luxembourg: (At: https://esdac.jrc.ec.europa.eu/content/lucas-topsoil-survey-methodology-data-and-results.).

Tóth, G., Jones, A. & Montanarella, L. 2013b. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. *Environmental monitoring and assessment*, (At: http://www.ncbi.nlm.nih.gov/pubmed/23371251. Accessed: 7/5/2013).

Tóth, B., Weynants, M., Nemes, A., Makó, A., Bilas, G. & Tóth, G. 2015. New generation of hydraulic pedotransfer functions for Europe. *European Journal of Soil Science*, **66**, 226–238, (At: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4386477&tool=pmcentrez&rendertype=abstract. Accessed: 17/5/2016).

U.S. Department of Agriculture Natural Resources Conservation Service. 2009. Part 630 Hydrology, Chapter 7 Hydrologic Soil Groups. In: *National Engineering Handbook*.

Weber, T.K.D., Weynants, M. & Szabó, B. 2020. R package of updated European hydraulic pedotransfer functions (euptf2). (At: https://doi.org/10.5281/zenodo.4281045.).

Wessolek, G., Kaupenjohann, M. & Renger, M. 2009. Bodenphysikalische Kennwerte und Berechnungsverfahren für die Praxis. *Bodenökologie und Bodengenese*, **40**, 1–80, (At: https://www.boden.tu-berlin.de/fileadmin/fg77/_pdf/Rote_Liste/Rote_Reihe_Heft_40.pdf.).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. 2019. Welcome to the Tidyverse. *Journal of Open Source Software*, **4**, 1686.

Wickham, H. & Bryan, J. 2019. readxl: Read Excel Files. R package version 1.3.1.

Wickham, H., François, R., Henry, L. & Müller, K. 2021. dplyr: A Grammar of Data Manipulation. R package version 1.0.7.

Zhu, Y., Chen, Y., Ali, M.A., Dong, L., Wang, X., Archontoulis, S. V., Schnable, J.C. & Castellano, M.J. 2021. Continuous in situ soil nitrate sensors: The importance of high-resolution measurements across time and a comparison with salt extraction-based methods. *Soil Science Society of America Journal*, 85, 677–690.