universität innsbruck

eurac research

Program for mapping chemical soil properties: tested on grassland and arable land in South Tyrol, Northern Italy

MASTER THESIS

in oder to obtain the degree

MASTER OF SCIENCE (M.Sc.)

Submitted by

Elvis Burchia

Supervised by:

Priv.-Doz. Dr. Mag. Erich Tasser

Leopold-Franzens-Universität Innsbruck

Department of Ecology

Innsbruck, April 2021

Plagiarism Disclaimer

I hereby declare that this diploma thesis is my own and autonomous work. All sources and aids used have been indicated as such. All texts either quoted directly or paraphrased have been indicated by in-text citations. Full bibliographic details are given in the list of works cited, which also contains internet sources including URL and access date. This work has not been submitted to any other examination authority.

6th April 2021, 

Date, Signature

TABLE OF CONTENTS

Abstract 1

1. Introduction 2

2. Material and methods 4

2.1 Study area 4

2.2 Soil sample data 5

2.3 Predictor variables 7

2.4 Spatial interpolation and prediction of chemical soil properties 9

2.5 Validation 11

3. Results 12

3.1 Soil sample data – used for analysis 12

3.2 Predictor variables 12

3.2.1 Grassland 12

3.2.2 Grassland and arable land 12

3.3 Soil property maps 15

3.3.1 Soil Organic Matter (SOM) 16

3.3.2 pH 16

3.3.3 Phosphorus (P) 17

3.3.4 Potassium (K) 18

3.3.5 Nitrogen (N) 19

3.3.6 Carbon (C) 20

3.3.7 Carbon to Nitrogen ratio (C:N) 21

3.4 Validation 22

4. Discussion 23

4.1 Distribution of soil chemical properties 23

4.2 Deficiencies 24

4.2.1 Soil samples 24

4.2.2 Predictors 25

4.2.3 Method 25

4.3 Usage of the developed program 26

5. Conclusion and Outlook 26

Software 27

Acknowledgments 27

References 28

Appendix

ABSTRACT

Climate change and intensifying human activities lead increasingly to the degrading of many areas worldwide. Since information of chemical soil properties and their spatial variability is one of the best management tools to estimate soil condition and its degradation, a program based on digital soil mapping (DSM) methods was developed. DSM turns pointwise soil surveys into continuous maps through robust regression and interpolation methods, which selected are the Random Forest regression and the Ordinary Kriging interpolation. The aim of the study is to offer a general program for mapping any continuous soil parameters on any areas of interest. Its right functioning was tested by predicting SOM, pH, P, K, C, N and C:N ratio on South Tyrolean grassland and arable land using different combinations of the following predictors: elevation, slope, wetness index, livestock, aspect, land use and geological information, of which the most important was elevation. Other significant predictors were slope, aspect, livestock and wetness index. Geological and land use information had a significant influence only on certain soil parameters. The predictor describing the intensively use of hay meadows was important for the prediction of P and K, whereas the geological information about alkalinity of soil played a key role in the prediction of pH. The percent of explained variance ranged from 6.3% to 57.9% with a mean of 31.1% and standard deviation of 16.7%. This shows that some important predictors were still missing for an optimal prediction. Even though the sample densities and its distribution across altitudes were unbalanced, the maps are meaningful, since the validation shows a quite high accuracy. The R^2 values ranged from 0.43 to 0.79 with a mean of 0.60 and a standard deviation of 0.12, whereas the NRMSE ranged from 51.50 to 80.00 with a mean of 65.1 and a standard deviation of 8.3. The prediction of pH produced the best model, with the highest percent of explained variance (57.9% for grassland and 53.4% for grassland and arable land), the lowest NRMSE (53.80 and 51.50) and the second highest R^2 (0.74 in both areas of interest). These results confirm the right functioning of the developed program that could promote a long-term planning sustainable use of fertilizers on South Tyrolean grassland. It can also help in general soil investigations doing spatial predictions of several chemical soil properties that can be an important tool to estimate soil degradation and to identify areas where concrete interventions are needed. This applies especially for regions with intensive land use where availability of soil data is limited.

Keywords: digital soil mapping (DSM), software, prediction, chemical soil properties, grassland, South Tyrol

1. INTRODUCTION

The world is facing huge challenges including climate change, land degradation, biodiversity loss, food security, water resource management and ecosystem health (FAO and ITPS, 2015). The soil system and its functions that linking with biomass production, environmental buffering, water purification and climate mitigation play an important role for addressing the above-mentioned issues (McBratney *et al.*, 2014). Detailed spatial soil informations are needed, e.g. as a prerequisite for soil protection and sustainable cultivation. Despite this increasing demand for soil data, the cost and time required for traditional soil mapping means that often they are not available (Grimm and Behrens, 2010).

The ever growing computational capacities coupled with the development of data-mining algorithms and GIS tools, and the increased availability of spatial remote-sensing data has created a great potential for improvement in soil mapping techniques (Camera *et al.*, 2017; Minasny and McBratney, 2016). For this reason digital soil mapping (DSM) have been developed as an operational activity (Vaysse and Lagacherie, 2015), aiming to provide fast and accurate methods to predict soil variables spatially. Digital soil mapping can be also defined as sub-discipline of soil science which deals with computer-generated production of digital maps of soil types and soil properties (Naresh, 2020). It has experienced a continuous expansion in the last two decades, mainly due to its increased efficiency in comparison to conventional field soil mapping techniques (Kempen *et al.*, 2012). DSM relies on statistical relationships between measured soil observations and independent variables, also called predictors. In this background, the term predictive soil mapping has emerged (Scull *et al.*, 2003). DSM turns pointwise soil surveys into continuous maps through robust regression and interpolation methods (McBratney *et al.*, 2003; Minasny and McBratney, 2016; Robinson and Metternicht, 2006).

Of interest are modelling approaches for soil chemical parameters such as soil organic matter (SOM), pH, phosphorus (P), potassium (K), nitrogen (N), carbon (C) and C:N ratio. These chemical soil properties describe the availability of nutritional mineral elements for plants and the chemical parameters of soil in connection with their restoration (Janssens *et al.*, 1998; Bai *et al.*, 2020; Baer *et al.*, 2002; Tibbett *et al.*, 2019). Detailed knowledge about the storage of SOM in soils is essential with regard to the rising demand for agricultural land, ongoing soil degradation and requirements for sequestration of atmospheric CO_2 (Wiesmeier *et al.*, 2011). Especially grassland plays in the context of CO_2 sequestration an important role, because it contains at a global scale approximately 15% of total soil organic carbon (SOC) (Lal, 2009; Sanchez *et al.*, 2009). The pH controls mainly the solubility of plant nutrients in the soil, which in turn affects microbial and plant growth (Neina, 2019). This makes pH management important in controlling movement of heavy metals and potential groundwater contamination in soils. The pH also induces, like increasing aridity and temperature, soil C-N-P imbalance in grasslands (Jiao *et al.*, 2016). N,

P, and K are inorganic fertilizers widely used by farmers in order to improve plant productivity (Bahn, 2020), but also other components and functions of the ecosystem, including soil water storage and soil healthy (Pan *et al.*, 2014). On the other hand the C-N-P decoupling may reduce plant growth and production in grasslands ecosystems (Jiao *et al.*, 2016) but still more research is needed in this context.

With my work, I would contribute to the aforementioned development by developing a program (R script) based on DSM methods, that have been tested by various authors (Hengl *et al.*, 2018; Hengl *et al.*, 2004) and have been proven to have good predictive capability for continuous variables. The methods used to spatialize and predict the chemical soil properties in this study are the Random Forest (RF) regression (Breiman, 2001) and the Ordinary Kriging (OK) interpolation (Krige, 1951). The developed program is able to predict any continuous chemical parameters with the use of several predictors. I tested the model for soil organic matter (SOM), pH, phosphorus (P), potassium (K), nitrogen (N), carbon (C) and C:N ratio by using elevation, slope, wetness index, livestock unit, aspect, land use and geological information as predictors. My test area was South Tyrol. South Tyrol is predominantly mountainous, only a very small part of the area can be used for agriculture. Animal husbandry and milk production are practised in higher altitudes (88% of agricultural land), while in lower valley bottoms orchards and vineyards (10%) and arable land (2%) predominate (WIFO, 2019). South Tyrol is particularly well suited for the development of a DSM since soil data have been collected for many years. As a framework for soil data sourcing and management, involving farmers, public administrators and research scientists was installed in the 2006. A particularly large amount of data is available for apple orchards and vineyards. On these data basis, area-wide modelling of individual soil parameters has already been carried out (Della Chiesa *et al.*, 2019a; Della Chiesa *et al.*, 2019b). Such modelling approaches and high-resolution soil property maps are still lacking for grassland and arable land. The reason of this study is therefore given by these lacks and the “fertilization problem” given in South Tyrol, meaning a nutrient surplus due to a non-closed loop economy.

Modern cows feed not only on hay from the meadows but also on concentrated feed in order to be able to meet their nutrient and feed requirements. Concentrated feed, however, is mainly imported from areas outside South Tyrol, leading to large quantities of nutrients being used as fertilizers in grasslands. However, a closed cycle is only given if no additional nutrients are brought into the system from outside or if these nutrients are returned to the feed production sites as manure or slurry. Neither of these is the case in South Tyrol. There are directives, but they often do not reflect reality. In many cases it is not about the grassland yield at all, but about the disposal of the resulting farm manure. As excessive fertilization management can lead to detrimental environmental impact and indirect costs to ecosystem (King *et al.*, 2015) and many South Tyrolean farmers have nowadays to struggle with impoverished and partly weedy grassland areas, action on base of such DSM studies as here is needed.

The reason for this study is to fill knowledge gaps of detailed spatially distributed information of important chemical soil properties in South Tyrolean grassland and arable land, since many current grassland areas were once used for crop production or still used to be cultivated in alternation between arable farming and grassland. The aim is to turn pointwise soil surveys of SOM, pH, P, K, N, C and C:N ratio into continuous maps through DSM methods by using elevation, slope, wetness index, livestock unit, aspect, land use and geological information as predictors. The results should allow conciliating intensive agriculture production with profitability and environmental sustainability. Finally, the validation of the models should confirm the correct functioning of the developed program.

2. MATERIAL AND METHODS

2.1 Study area

This study covered grassland and arable land soils in South Tyrol in Northern Italy (**Fig. 1**). The autonomous province is located in the Alps, between 10°22'E and 12°30'E and 46°13'N and 47°4'N. It is characterized by its mountainous topography with elevations ranging from 200 m to more than 3902 m (Mt. Ortles) and diverse climatic conditions. It has a typical continental Alpine precipitation and temperature regime, with a mean annual precipitation of 700-800 mm and mean annual temperatures between 9-11°C (ASTAT, 2019). South Tyrol covers an area of almost 7,400 km^2 of which 40% is covered with woodland, 36% is agriculturally usable land, 22% is unproductive land and 2% is agriculturally non usable land (ASTAT, 2017). 88% of the agriculturally used area are meadows (27%) and pastures (61%), 10% are wood crops and 2% is covered by apple orchards and vineyards (WIFO, 2019). The land use of the agriculturally usable areas is mainly controlled by the degree of accessibility for vehicles. Easily accessible areas are increasingly used intensively, while the poorly accessible areas are being abandoned or used as pastures (Tasser and Tappeiner, 2002; Zimmermann *et al.*, 2010). In last decades, especially pastures run also the risk of being abandonment (**Fig. 2**) due to cost disadvantages (Busch *et al.*, 2018), so that they often become forests.

Since grassland comprises the second largest area after woodland in South Tyrol, it is the first area of interest taken into account in this study. Many of these grassland areas were once used for crop production (**Fig. 2**) or still used to be cultivated in alternation between arable farming and grassland, so that several chemical soil properties can be affected by these land use changes. For this reason arable land was considered together with grassland as second area of interest.

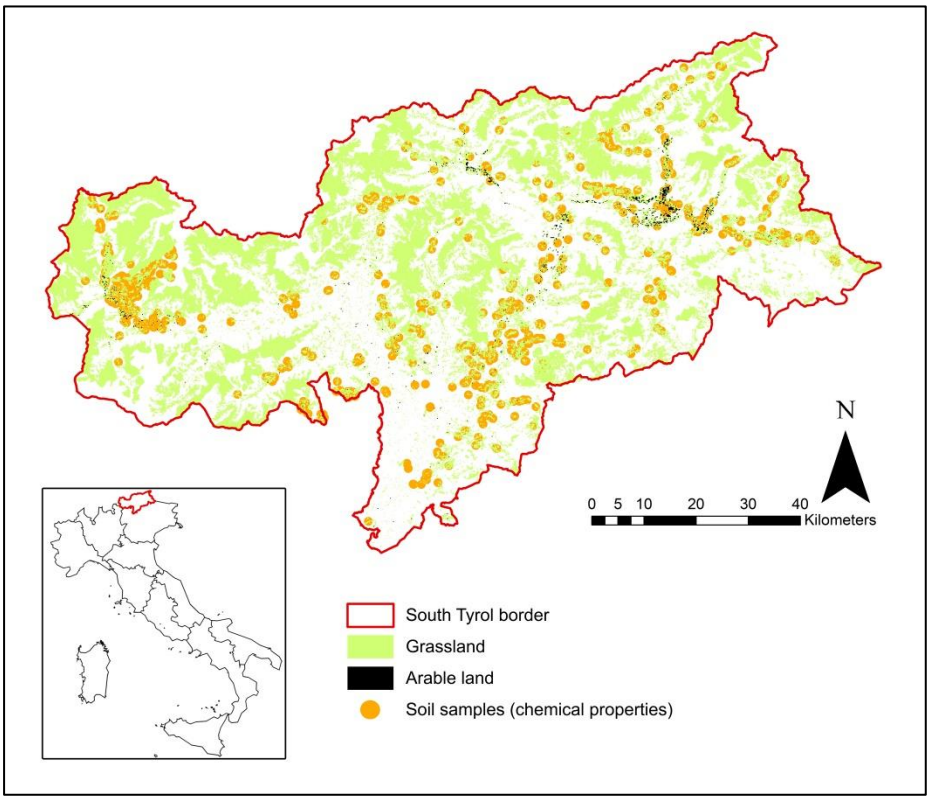


Figure 1: Study area in South Tyrol, Northern Italy including the used samples for DSM.

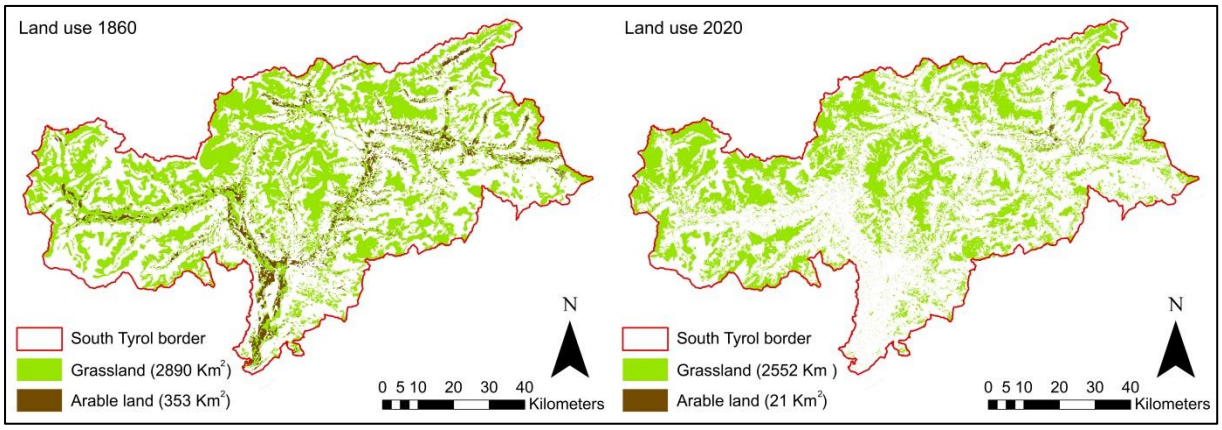


Figure 2: Historical (1860-2020) land use of South Tyrolean grassland and arable land (Jaufenthaler, 2019).

2.2 Soil sample data

In South Tyrol there are two important research centres that collect soil samples and do soil study. On the one hand there is the EURAC research in Bolzano and on the other hand there is the public chemical laboratory of the Research Centre for Agriculture and Forestry, Laimburg in Pfatten. The institutions analyse samples following common protocols and store digital soil data from across South Tyrol. The current study focuses on 719 georeferenced soil samples evenly collected all over the study area (**Fig. 1**)

and provided by the above-mentioned institutions. The soil samples were collected between 2006 and 2019, of which only the samples from actual grasslands and arable land areas were considered for analysis (**Table 1**). The sample locations were at different altitudes, with a mean of 1319 m and a standard deviation of 366 m for grassland and a mean of 954 m and a standard deviation of 132 m for arable land (**Fig. 3**). The most available samples were those of SOM and pH with 510 samples in grassland areas (mean sample density: 0.20 km^{-2}) and 17 samples in arable land (0.81 km^{-2}). P and K showed a smaller amount of samples from grasslands with 454 samples (0.18 km^{-2}) and the same amount like SOM and pH from arable land. At least data were about C, N and C:N ratio with 108 samples from grassland (0.04 km^{-2}) and 6 samples from arable land (0.29 km^{-2}).

Table 1: Soil parameters considered in the analysis: units, analysis methods and information about the samples in the areas of interest (grassland and arable land).
(Elmnt.: elemental analysis. CaCl2: CaCl2 glass electrode. CAL: Calcium Acetat-Lactate. n.a.: not available)

	Soil parameters	SOM	pH	P	K	N	C	C:N				
	Units	%	-	mg $100g^{-1}$	mg $100g^{-1}$	%	%	-				
	Analysis method	Elmnt.	CaCl2	CAL	CAL	Elmnt.	Elmnt.	Elmnt.		Altitude [m]		
No.									Year	Mean	St. Dev.	Reference
1	Tot. samples	531	531	531	531	-	-	-	2006-2016	1217	278	Research Centre Laimburg (www.laimburg.it)
	Grassland	374	374	374	374	-	-	-		1248	263	
	Arable land	11	11	11	11	-	-	-		988	150	
2	Tot. samples	34	34	34	34	-	-	-	2003-2004	1156	251	Rüdisser *et al.* , 2015
	Grassland	28	28	28	28	-	-	-		1189	252	
	Arable land	-	-	-	-	-	-	-		-	-	
3	Tot. samples	7	7	7	7	7	7	7	n.a.	1724	50	Tasser *et al.* , 2000
	Grassland	5	5	5	5	5	5	5		1749	14	
	Arable land	-	-	-	-	-	-	-		-	-	
4	Tot. samples	79	79	-	-	79	79	79	n.a.	1657	519	Kollmann, 2012
	Grassland	56	56	-	-	56	56	56		765	78	
	Arable land	-	-	-	-	-	-	-		-	-	
5	Tot. samples	17	17	17	17	17	17	17	2019	785	83	Moosgöller, ongoing
	Grassland	8	8	8	8	8	8	8		1714	482	
	Arable land	-	-	-	-	-	-	-		-	-	
6	Tot. samples	12	12	12	12	12	12	12	2019	1752	521	Ongoing project MICROVALU - Euregio IPN94
	Grassland	11	11	11	11	11	11	11		1736	543	
	Arable land	-	-	-	-	-	-	-		-	-	
7	Tot. samples	16	16	16	16	16	16	16	2019	981	199	Rottersteiner, 2020
	Grassland	9	9	9	9	9	9	9		1069	220	
	Arable land	4	4	4	4	4	4	4		895	73	
8	Tot. samples	15	15	15	15	15	15	15	2019	1187	482	Biodiversity monitoring South Tyrol (biodiversity.eurac.edu)
	Grassland	10	10	10	10	10	10	10		1382	455	
	Arable land	2	2	2	2	2	2	2		887	42	
9	Tot. samples	9	9	9	9	9	9	9	2013	2065	35	Ressi *et al.* , 2014
	Grassland	9	9	9	9	9	9	9		2065	35	
	Arable land	-	-	-	-	-	-	-		-	-	
1-9	Tot. samples	719	719	641	641	160	160	160	2006-2019	1271	375	See above 1-9
	Grassland	510	510	454	454	108	108	108		1319	366	
	Arable land	17	17	17	17	6	6	6		954	132	

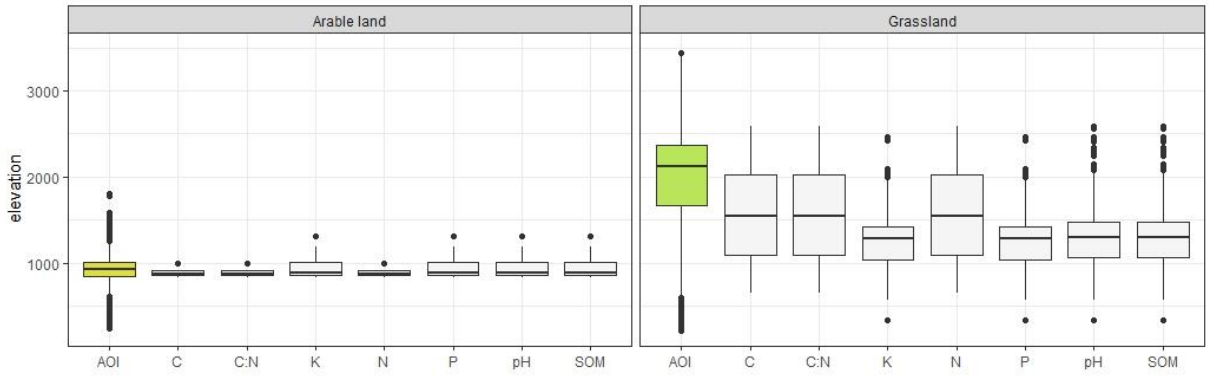


Figure 3: Boxplots of the elevation levels of the soil samples of each variable. The shaded plots describe the elevation of the areas of interest (AOI) where the chemical soil properties were predicted.

The sample densities were taken to define the raster resolution for this study. The denser the measurement points, the larger the scale of mapping, the smaller the resolution. The formula used to determine it, is the one described by (Hengl, 2006)

$$grid\ resolution = 0.0791 * \sqrt{\frac{A}{N}}$$

where *A* is the surface of the area of interest in m^2 taken into account and *N* is the total number of soil samples contained. The resolution used in this study is 150 m, which is a rounded value derived from the sample quantity of SOM and pH in the areas of interest.

2.3 Predictor variables

Elevation (Digital Terrain Model), slope, wetness index and livestock unit as continuous variables, and aspect, land use and geological information as categorical variables were used as predictors (**Table 2**) to predict the chemical soil properties.

The Digital Terrain Model (Autonomous Province of Bolzano - Geocatalog) was used to derive elevation, slope, aspect and topographic wetness index (ArcGIS - Spatial Analyst). The slope describes the steepness at each location and is given in degrees. The aspect indicates the compass direction from 0-360° that the surface faces. Flat areas having no downslope direction are given a value of -1 (Burrough and McDonell, 1998). As exposure does not have a linear but a degree scale, the values were reclassified in 5 classes: -1 = no exposition, flat, 0-45 and 315-360 = north, 45-135 = east, 135-225= south and 225-315 = west. For each class a binary map was generated.

Table 2: Overview and information about the independent variables (predictors) used for prediction. Detailed information about land use, geology and detailed geology is given in the Appendix 1, 2, 3 and 4. (G: Grassland. GA: Grassland and Arable land. St. Dev.: standard deviation)

	Predictor variables	Code	Unit	Mean	St. Dev.	Variable type	Source
	Elevation (digital terrain model)	10	m	1744.51	657.79	numerical	Geocatalog (Autonomous Province of Bolzano)
	slope	22	°	26.45	13.35	numerical	Generated in ArcGIS using the DTM
	wetness index	23	-	12.54	1.97	numerical	Generated in ArcGIS using the DTM
	livestock unit (LSU - GVE)	24	n	0.96	1.42	numerical	Tasser, 2012
Aspect	east (45° < x < 135°)	1	0/1	-	-	categorical	Generated in ArcGIS using the DTM
Aspect	flat (x = 1°)	2	0/1	-	-	categorical	
Aspect	north (0 °< x < 45° and 315° < x < 360°)	3	0/1	-	-	categorical	
Aspect	south (135° < x < 225°)	4	0/1	-	-	categorical	
Aspect	west (225° < x < 315°)	5	0/1	-	-	categorical	
Land use	pastures	6	0/1	-	-	categorical	Rüdisser *et al.* , 2015 (reclassified)
Land use	intensively used hay meadows	7	0/1	-	-	categorical	
Land use	extensively used hay meadows	8	0/1	-	-	categorical	
Land use	grassland special area	9	0/1	-	-	categorical	
Land use	arable land	30	0/1	-	-	categorical	
Geology	acidic soil (pH < 5)	25, 27	0/1	-	-	categorical	Geocatalog (Autonomous Province of Bolzano) Brandner, 1980 (reclassified)
Geology	acidic-basic soil (5 < pH < 6)	29	0/1	-	-	categorical	
Geology	basic soil (pH > 6)	26, 28	0/1	-	-	categorical	
Detailed geology	shist	11	0/1	-	-	categorical	Geocatalog (Autonomous Province of Bolzano) Brandner, 1980 (reclassified)
Detailed geology	vulcanite	12	0/1	-	-	categorical	
Detailed geology	vinschgau shear zone	13	0/1	-	-	categorical	
Detailed geology	quaternary depositions	14	0/1	-	-	categorical	
Detailed geology	sediment sequences	15	0/1	-	-	categorical	
Detailed geology	plutons	16	0/1	-	-	categorical	
Detailed geology	central gneiss	17	0/1	-	-	categorical	
Detailed geology	quartz phyllite	18	0/1	-	-	categorical	
Detailed geology	insignificant alpine metamorphosis	19	0/1	-	-	categorical	
Detailed geology	medium-low alpine metamorphosis	20	0/1	-	-	categorical	
Detailed geology	medium-high alpine metamorphosis	21	0/1	-	-	categorical	

The topographic wetness index (TWI), also known as the compound topographic index (CTI), is a steady state wetness index. It is defined as $\ln\ (a * \tan\beta^{-1})$ where *a* is the local upslope area draining through a certain point per unit contour length and $\tan\beta$ is the local slope. It is commonly used to quantify topographic control on hydrological processes (Sørensen *et al.*, 2006), thus to identify areas that contribute to runoff. Therefore landscapes with larger upslope drainage areas and shallower slopes will produce larger TWI values, indicating higher propensity for runoff.

The livestock unit (LSU) (Tasser, 2012) is an agricultural-ecological measure of land-use intensity. Livestock numbers are usually expressed in livestock units per hectare (LU ha^{-1}). The LSU is calculated on the basis of the mean live weight of the different livestocks and is equal to 500 kg.

Information about land use was taken from (Rüdisser *et al.*, 2015). The original land-use/cover types were reclassified in 5 categories (for details see Appendix 1). The category of "pastures" contains pastures near the farm and summer pastures in subalpine/alpine belt, "intensively used hay meadows" contains all fertilized hay meadows and mixed rotation meadows (rotation meadows between

grasslands and crops). The category “Extensively used hay meadows” includes hay meadows that are not fertilized and not cut regularly. The category “grassland special area” contains intensive grasslands or other areas that could not be assigned to the other classes. The last category “arable land” includes areas where different vegetables (cauliflower, head cabbage, radicchio, salad, maize and herb), cereals or forage herbs (clover, alfalfa) are cultivated. For each category, a binary map was generated.

Geological information is based on the Brandner map (Brandner, 1980) and geological knowledge of the Geology and Materials Testing Office of the Autonomous Province of Bolzano. The original geological classification was reclassified in 11 classes (for details see Appendix 2).

Finally, the information from the geological map (Autonomous Province of Bolzano – Geocatalog; (Brandner, 1980) was also used to classify the alkalinity of the soils. The soils above the different geological units were divided corresponding the alkalinity of the bedrocks into acidic soils (pH < 5), basic soils (pH > 6) and soil in neutral range (5 < pH < 6) (for details see Appendix 3 and 4).

Different combinations of the listed predictors were used in the analyses, in order that the program was able to find the best mathematical model for prediction. The selection of the independent variables can be made on the basis of a statistical evaluation or it can also be defined manually, as was done in this study. After computing regressions using different random predictors, those with the best results and meaningful combinations were used for spatial interpolation and prediction (**Table 3**). The selection was based on the exclusion of different predictors, except for the numerical variables (except LSU) that were always taken as predictors.

2.4 Spatial interpolation and prediction of chemical soil properties

An overview of the framework is described in **Figure 4**. Starting from 80% of the soil samples described in chapter 2.2, the program does a large-scale prediction (up scaling) of these chemical soil properties (SOM, pH, P, K, C, N and C:N) for other areas with the same land use (**Fig. 1**).

In a first step, the distribution of the original and transformed (logarithm, square, root, inverse) values of the parameters was statistically checked via the Shapiro-Wilk test (**1** – **Fig. 4**). This check was needed because geostatistical methods work best when the mean/variance of data do not vary significantly (Genova, 2017) and significant deviations from normality can also affect the kriging estimators (Lark, 2000). Only data with a Shapiro test statistic W greater than 0.90 was considered for spatial interpolation (for details see Appendix 5).

Table 3: Combinations of predictors tested in the analysis and those that produce the best prediction of each chemical soil property in the two areas of interest. The crosses with a blue background were only taken into account for grassland and arable land areas. The ones with a red background represent predictors that were not significant, as they have an %IncMSE of 0, therefore they are not shown in Figure 4.a,b,c. (G: Grassland. GA: Grassland and Arable land)

				Combinations with best model accuracy													
				SOM		pH		P		K		C		N		C:N	
	Predictor variables	Code	Tested combinations of predictors	G	GA	G	GA	G	GA	G	GA	G	GA	G	GA	G	GA
	Elevation (digital terrain model)	10	x x x x x x x x x x x x x x x x x x x x x x x x x x x x x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	slope	22	x x x x x x x x x x x x x x x x x x x x x x x x x x x x x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	wetness index	23	x x x x x x x x x x x x x x x x x x x x x x x x x x x x x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	livestock unit (LSU - GVE)	24	x x x x x x x x x x x x x x x x x x x x x x x x x x x x x	x		x	x	x				x	x	x	x	x	x
Aspect	east (45° < x < 135°)	1	x x x x x x x x x x x x x x x x x x x x x x x x x x x x x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Aspect	flat (x = 1°)	2	x x x x x x x x x x x x		x		x	x	x	x		x	x				x
Aspect	north (0 °< x < 45° and 315° < x < 360°)	3	x x x x x x x x x x x x x x x x x x x x x x x x x x x x x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Aspect	south (135° < x < 225°)	4	x x x x x x x x x x x x x x x x x x x x x x x x x x x x x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Aspect	west (225° < x < 315°)	5	x x x x x x x x x x x x x x x x x x x x x x x x x x x x x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Land use	pastures	6	x x x x x x x x x x x x x x x x x x	x	x	x		x		x	x	x	x	x		x	
Land use	intensively used hay meadows	7	x x x x x x x x x x x x x x x x x x	x	x	x		x		x	x	x	x	x		x	
Land use	extensively used hay meadows	8	x x x x x x x x x x x x		x			x		x	x		x	x			
Land use	grassland special area	9	x x x x x x x x x x x x x x x x x x	x	x	x		x		x	x	x	x	x		x	
Land use	arable land	30	x x x x x x x x x x x x x x x x x x		x						x		x	x			
Geology	acidic soil (pH < 5)	25	x x x x x x x x x			x							x	x			
Geology	acidic soil (pH < 5)	27	x x x x x x x x x	x		x				x		x	x	x			
Geology	acidic-basic soil (5 < pH < 6)	29	x x x x x x x x x	x		x				x		x	x	x			
Geology	basic soil (pH > 6)	26	x x x x x x x x x			x								x			
Geology	basic soil (pH > 6)	28	x x x x x x x x x	x		x				x		x	x	x			
Detailed geology	shist	11	x x x x x x x x x x x			x	x	x			x		x	x	x		
Detailed geology	vulcanite	12	x x x x x x x x x x x			x	x	x			x		x	x	x		
Detailed geology	vinschgau shear zone	13	x x x x x					x			x	x		x			
Detailed geology	quaternary depositions	14	x x x x x x x x x x x			x	x	x			x		x	x	x		
Detailed geology	sediment sequences	15	x x x x x x x x x x x			x	x	x			x		x	x	x		
Detailed geology	plutons	16	x x x x x					x			x		x	x			
Detailed geology	central gneiss	17	x x x x x x x x x x x			x	x	x			x		x	x	x		
Detailed geology	quartz phyllite	18	x x x x x x x x x x x			x	x	x			x		x	x	x		
Detailed geology	insignificant alpine metamorphosis	19	x x x x x x x x x x x			x	x	x			x		x	x	x		
Detailed geology	medium-low alpine metamorphosis	20	x x x x x x x x x x x			x	x	x			x		x	x	x		
Detailed geology	medium-high alpine metamorphosis	21	x x x x x x x x x x x			x	x	x			x		x	x	x		

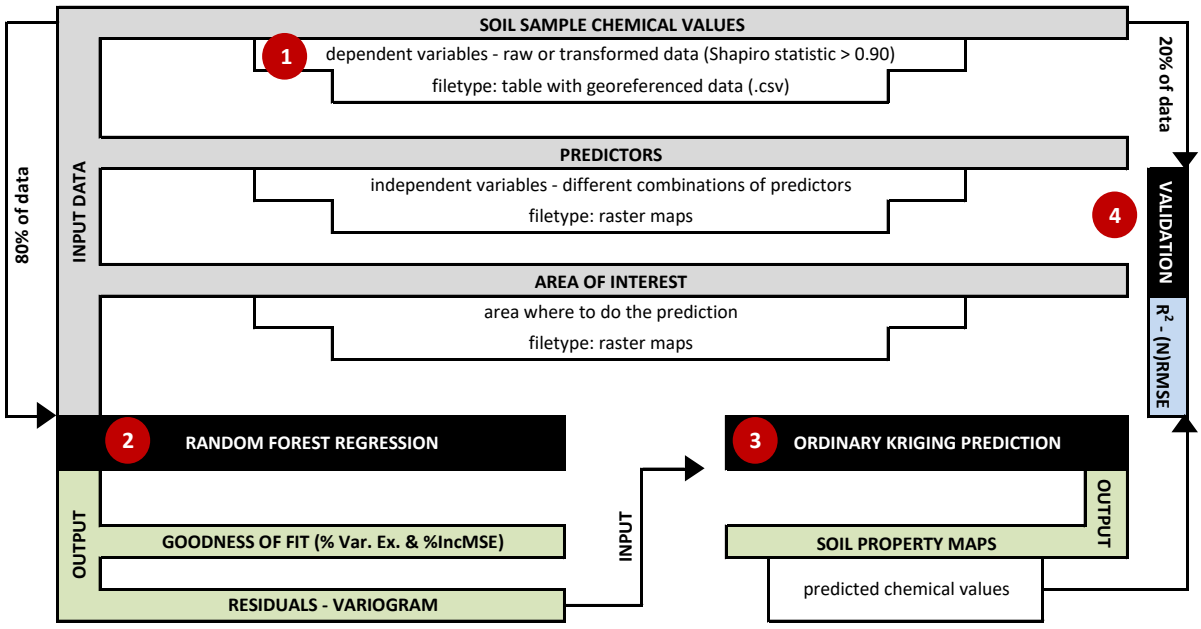


Figure 4: Overview of the framework used in this study – based on (Hengl, 2009). The steps are cited and explained in the text.

After the check of normality, the spatialization and prediction start with the Random Forest (RF) regression (**2** – **Fig. 4**), using 80% of the source data as training set. It is a powerful ensemble-learning method proposed by Breiman (Breiman, 2001), in which more decision trees are run in parallel and the mean prediction for the continuous dependent variables will be outputted (Chen *et al.*, 2017). The regression analysis outputs a residual variogram that gives an assessment of the variance of each variable. To compute a spatial interpolation such variogram is required (Genova, 2017). It is used in the next step of the analysis, the spatial prediction/interpolation, where the GSIF R package (Hengl *et al.*, 2016) is the main tool used.

The Ordinary Kriging (OK) (Krige, 1951) interpolation method is used to create the final map rasters of the chemical soil properties (Li and Heap, 2014) (**3** – **Fig. 4**). It is a widely adopted method that provides a solution to the problem of estimation based on a continuous model of stochastic spatial variation. The variation of the soil sample values is determined through the residual variogram from the regression part, which is connected to the kriging estimator to interpolate the value for each cell of a raster mask (Genova, 2017). So, this spatial model estimates a value at a point of an unsampled region for which a variogram is known.

The estimation is done by the following linear combination of values at sample locations

$$prediction\ at\ unsampled\ regions = \sum_{i=1}^{n} weight_{\ i} * value_{\ i}$$

where *i* is the running variable and *n* is the number of soil samples. Every location has a weight, an importance, which is given by the variogram. The sample locations that are further away will become less weight than the closer ones. More detailed information about is contained in Appendix 6.

2.5 Validation

In order to allow the model validation (**4** – **Fig. 4**), the spatial prediction/interpolation of the chemical soil properties was done with a training dataset containing 80% of the source data. The remaining randomly chosen 20% of data was used as validation dataset. The R-Squared (R^2) and the Root Mean Square Error (RMSE) were the most important values for model goodness considered. The RMSE describes the error of a model, so it is a direct measure of when a prediction is wrong. In addition it has the same unit of measure as the studied variable. For this reason the models of this study, from which the soil property maps were generated, are based on the parameter sets that returned the best RMSE or in some cases the best combination of the RMSE and R^2. For the comparison of models it is more suitable to look at the Normalized Root Mean Square Error (NRMSE).

3. RESULTS

3.1 Soil sample data – used for analysis

Descriptive statistics of the measured soil chemical properties SOM, pH, P, K, N, C and C:N ratio, its transformations and those that produced the best final models for each variable and area of interest are summarized in Appendix 5. The original data was not used for the spatial prediction model, since it was not normally distributed, with the exceptions of pH and N that nearly met the assumption. The following transformations were used for analysis: SOM (log – W = 0.99, sqrt – W = 0.94, inverse – W = 0.91), pH (log – W = 0.95, sqrt – W = 0.97, square – W = 0.99, inverse – W = 0.90), P (log – W = 0.99, sqrt – W = 0.93), K (log – W = 1, sqrt – W = 0.90), N (log – W = 0.99, sqrt – W = 0.96, inverse – W = 0.95), C (log – W = 0.98, sqrt – W = 0.90, inverse – W = 0.95), C:N (log – W = 0.94, sqrt – W = 0.90, inverse – W = 0.97).

3.2 Predictor variables

The predictors used for the modelling of the chemical soil properties and those combinations that had the best model accuracy are shown in **Table 3**. **Figure 5.a-b-c** shows the influence of the predictor variables on each soil parameter in the two areas of interest. The value that was taken into account is the percent increase in MSE of predictions (%IncMSE).

3.2.1 Grassland

The predictor with the most influence on all soil properties considered in the study is elevation. A second significant predictor is the livestock unit and in most cases also the aspect and slope. The predictors of pH show the highest percent of explained variance (57.9%) (**Fig. 5.a**), while those of N show the lowest (6.3%) (**Fig. 5.b**). The predictor about land use, the category “intensively used hay meadows”, influences significantly the predicted values of P and K (**Fig. 5.a,c**). Geology as predictor is especially important for the prediction of pH (**Fig. 5.a**).

3.2.2 Grassland and arable land

The most important predictor is the same as on grassland, the elevation. The livestock unit, the slope and aspect play also here an important role. The general patterns remain the same, but one predictor that seems to have a slightly higher influence on some chemical soil properties on grassland and arable land is the wetness index. The predictors of pH show the highest percent of explained variance (53.4%), while those of P show the lowest (7.5%) (**Fig. 5.a**).

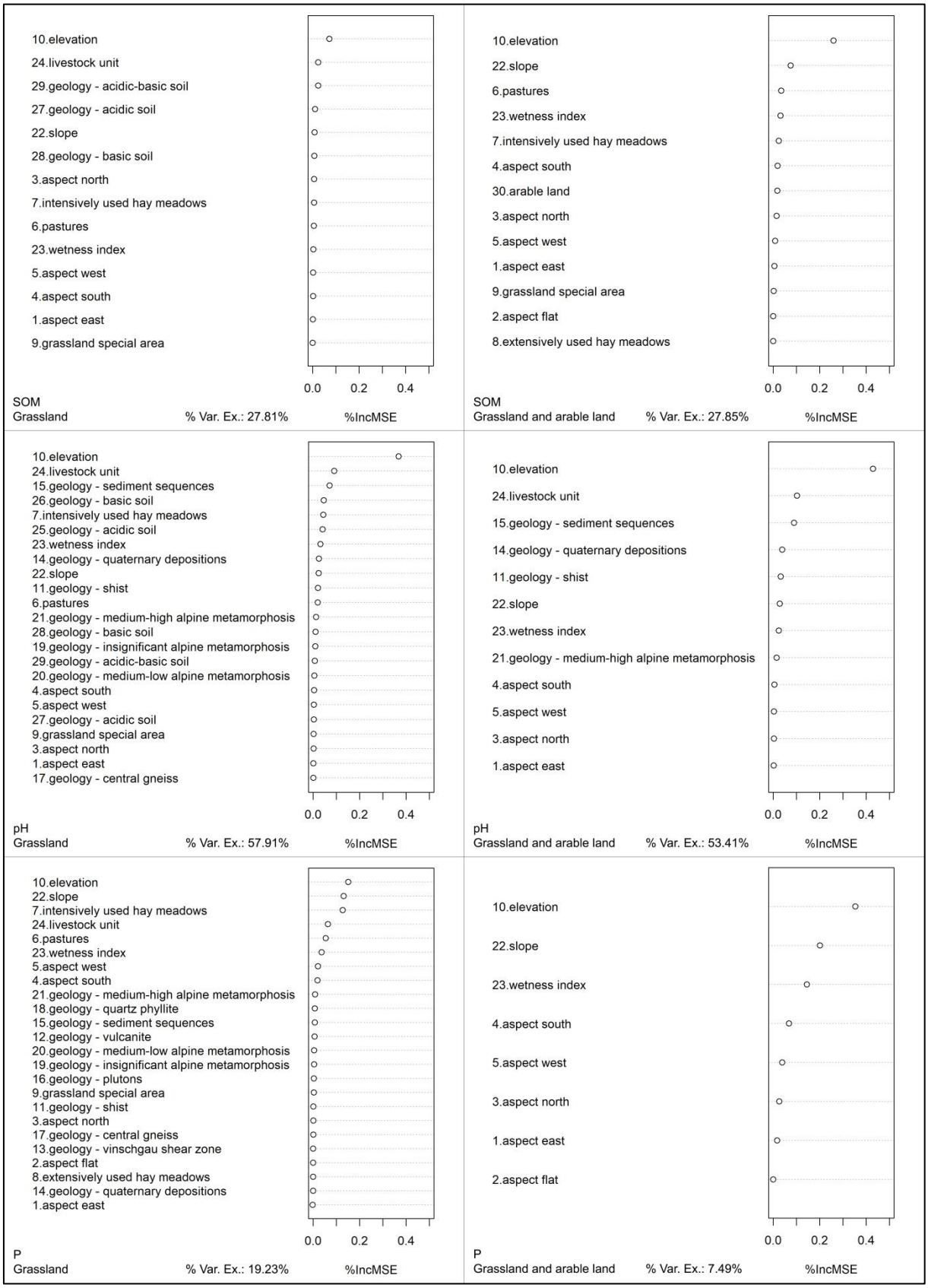


Figure 5.a: Influence of the predictors on SOM, pH and P in the Random Forest regression model.
(% Var. Ex.: percent of variance explained. %IncMSE: percent increase in MSE of predictions)

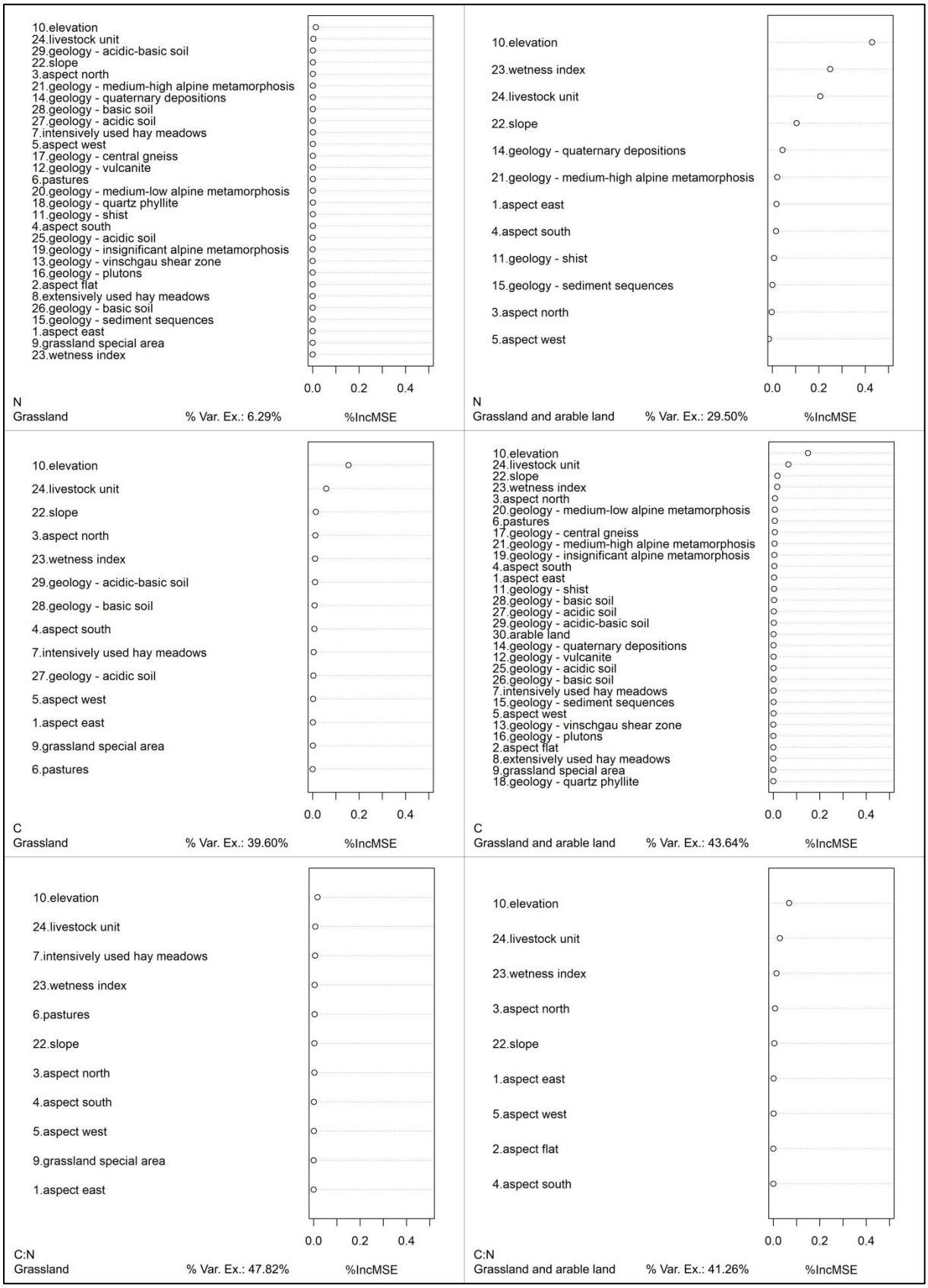


Figure 5.b: Influence of the predictors on N, C and C:N ratio in the Random Forest regression model.
(% Var. Ex.: percent of variance explained. %IncMSE: percent increase in MSE of predictions)

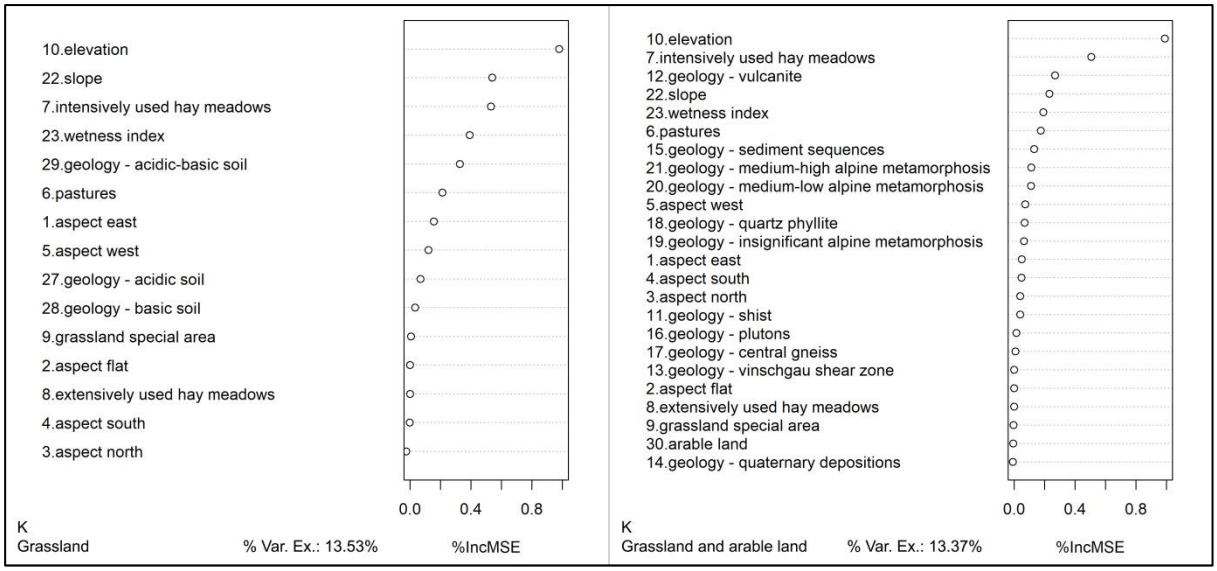


Figure 5.c: Influence of the predictors on K in the Random Forest regression model.
(% Var. Ex.: percent of variance explained. %IncMSE: percent increase in MSE of predictions)

3.3 Soil property maps

The result of the spatial prediction are the maps of SOM, pH, P, K, C, N and C:N ratio on grasslands and on grassland and arable land areas (**Fig. 6, 9, 11, 12, 14, 16, 18**). **Table 4** shows a descriptive overview of the predicted values of each chemical parameter on the two areas of interest. Since the two areas of interest almost coincide (**Fig. 1**), the chemical property maps show similar patterns.

Table 4: Descriptive statistic of the spatial predicted values (raster values of the property maps in Figure 6, 9, 11, 12, 14, 16, 18) of the soil chemical properties in the two areas of interest.
(Min.: minimum value. 1st Qu.: first quartile value. 3rd Qu.: third quartile value.
Max.: maximum value. St. Dev.: standard deviation)

	Area of interest	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std Dev.
SOM	grassland	2.40	9.10	12.10	12.50	15.20	41.00	12.50
	+ arable land	2.40	9.40	12.20	12.20	4.60	34.90	3.60
pH	grassland	3.50	4.30	4.60	4.90	5.50	7.40	0.70
	+ arable land	3.70	4.40	4.70	5.00	5.60	7.30	0.70
P	grassland	0.60	3.50	4.30	6.10	6.70	205.50	4.30
	+ arable land	2.20	6.30	8.10	9.40	11.30	107.70	4.40
K	grassland	2.90	12.90	17.30	19.00	23.70	225.70	9.00
	+ arable land	5.00	12.50	16.40	18.40	22.90	192.20	8.00
N	grassland	0.30	0.40	0.50	0.50	0.50	0.90	0.10
	+ arable land	0.20	0.40	0.50	0.50	0.50	0.90	0.10
C	grassland	2.40	4.70	6.50	6.70	8.20	15.90	2.30
	+ arable land	1.70	4.50	6.00	6.20	7.60	19.30	2.00
C:N	grassland	8.90	11.30	12.50	12.40	13.40	19.60	1.60
	+ arable land	8.60	11.10	12.80	12.70	13.90	22.70	1.70

3.3.1 Soil Organic Matter (SOM)

The SOM spatial distribution is shown in **Figure 6**. The predicted values range from the same minimum of 3.5% for both areas of interest to a maximum of 7.4% for grassland and 7.3% for grassland and arable land. The mean and median are similar in both areas, but in the second area most of the predicted values tend to be slightly higher, as shown in **Table 4**. The predicted values slightly decrease at an altitude within 500 and 1000m and then they increase with increasing altitude (**Fig. 7**). So, lower SOM values are predicted in lowlands and higher values in higher altitudes. An interesting west-east gradient appears, where in the western part of South Tyrol the SOM values are higher than in the east.

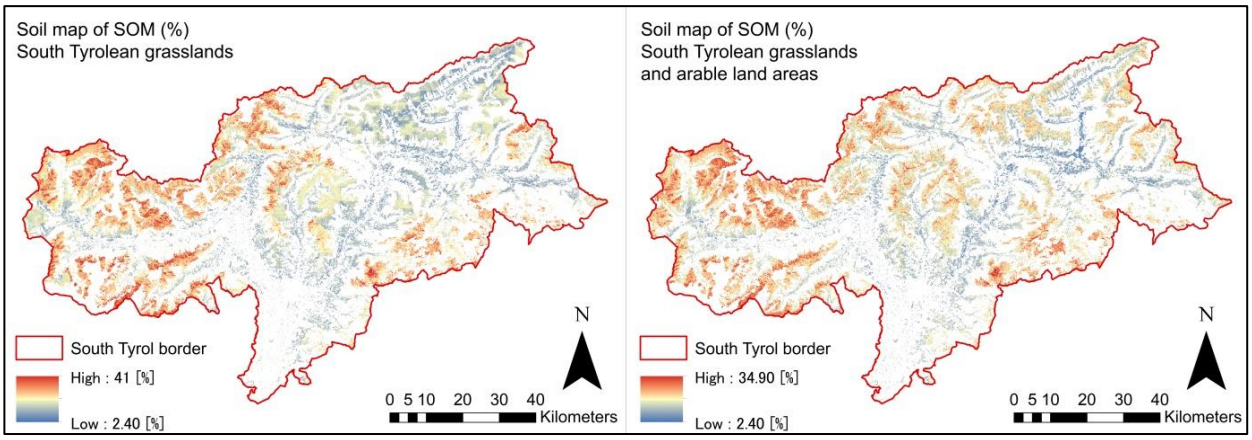


Figure 6: Distribution of SOM based on soil samples of grasslands (on the left) and samples of grasslands and arable land areas (on the right). (Validation values: RMSE = 3.0 (4.0) and R^2 = 0.5 (0.6))

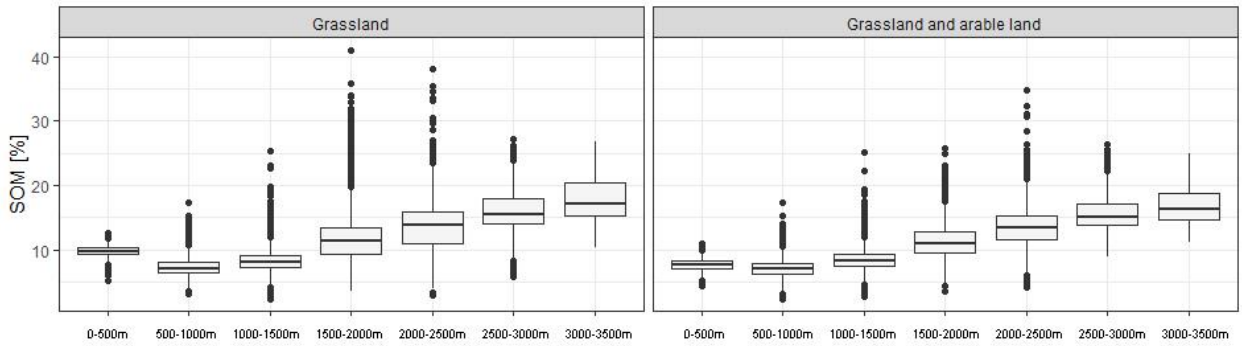


Figure 7: Boxplots of predicted SOM (%) values on grassland (on the left) and on grassland and arable land (on the right) classified according to altitude.

3.3.2 pH

The most predicted pH values range from strong acidic (pH < 5.5) with a minimum of 3.50 in grasslands and 3.70 in the second area of interest, to moderate acidic (5.5 < pH < 6.0). There are also areas in the slightly acidic (6.0 < pH < 6.5), neutral (6.5 < pH < 7.3) and near slightly alkaline range (7.3 < pH < 7.8) with a maximum of 7.40 in grasslands and 7.30 in grasslands and arable land areas

(**Table 4**). The entire pH spectrum is covered. Higher pH values are predicted in lower altitudes, with the highest values at an altitude within 500 and 1000m, as **Figure 8** shows. Moderate acidic soil is present especially in the southeastern part of the Salten-Schlern district. Gadertal and the Vinschgau Valley stand out, because of their neutral-slightly alkaline pH values (**Fig. 9**).

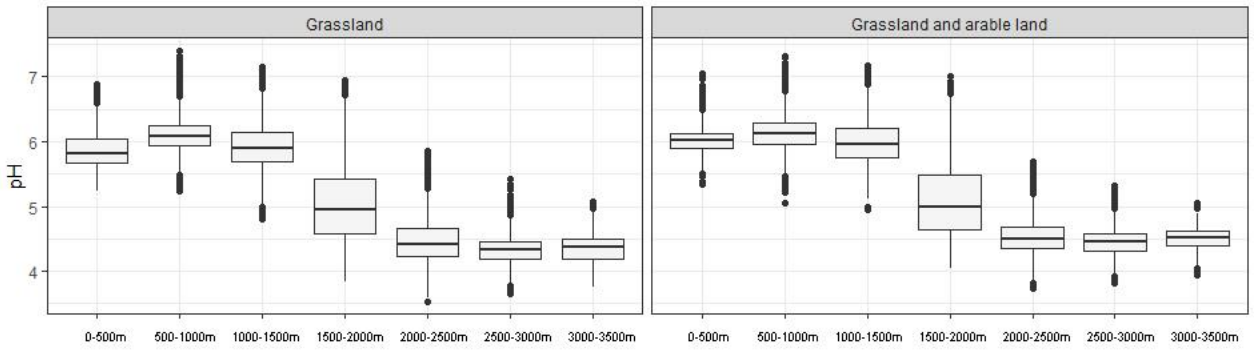


Figure 8: Boxplots of predicted pH values on grassland (on the left) and on grassland and arable land (on the right) classified according to altitude.

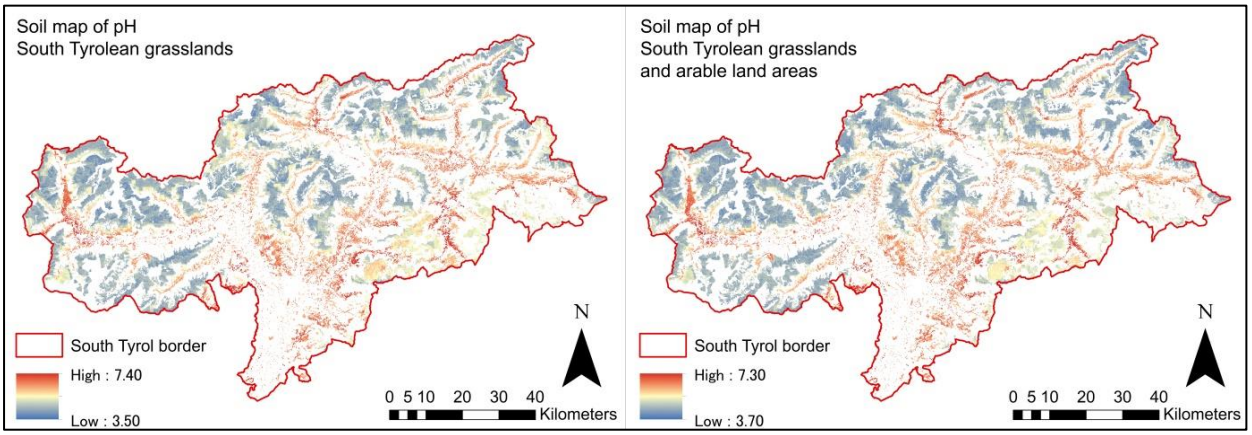


Figure 9: Distribution of pH based on soil samples of grasslands (on the left) and samples of grasslands and arable land areas (on the right). (Validation values: RMSE = 0.4 (0.4) and R^2 = 0.7 (0.7))

3.3.3 Phosphorus (P)

The trend for phosphorus is that the predicted values are lower in high altitudes and higher in the flat valley bottoms (**Fig. 10**). The highest P values are predicted at an altitude within 500 and 1000m on grassland and at an altitude within 0 and 500m on grassland and arable land. Overall, the predicted P values on South Tyrolean grasslands and arable land areas are rather low (**Fig. 11**), especially on high areas. They range from a minimum of 0.60 mg $100g^{-1}$ and a maximum of 205.50 mg $100g^{-1}$ with a mean of 6.10 mg $100g^{-1}$ and a median of 4.60 mg $100g^{-1}$ for the first area of interest. Even if the maximum is lower, the predicted values in grasslands and arable land areas are higher. They range from a minimum of 2.20 mg $100g^{-1}$, a maximum of 107.70 mg $100g^{-1}$, a higher median of 8.10 mg $100g^{-1}$ and also a higher mean of 9.40 mg $100g^{-1}$ (**Table 4**).

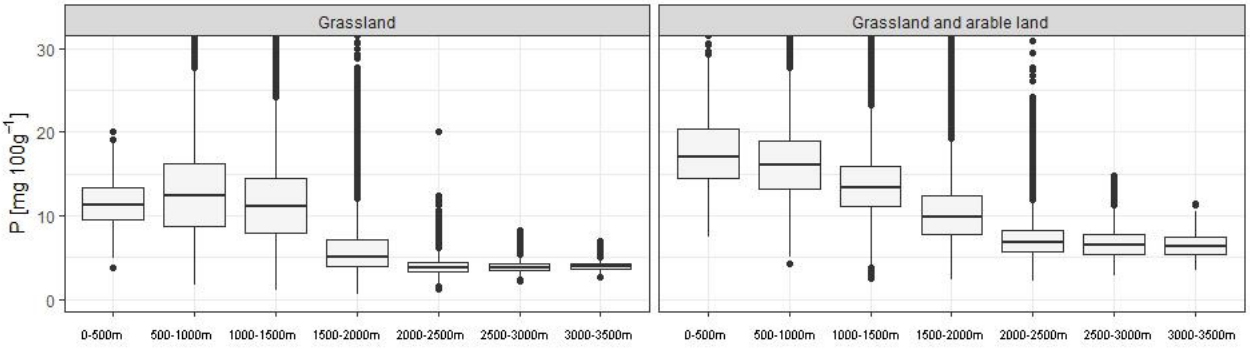


Figure 10: Boxplots of predicted P values on grassland (on the left) and on grassland and arable land (on the right) classified according to altitude. Just a part of the Y-axis is illustrated. There were outliers up to 205.5 m for grassland and 107.7 m for grassland and arable land (Table 4).

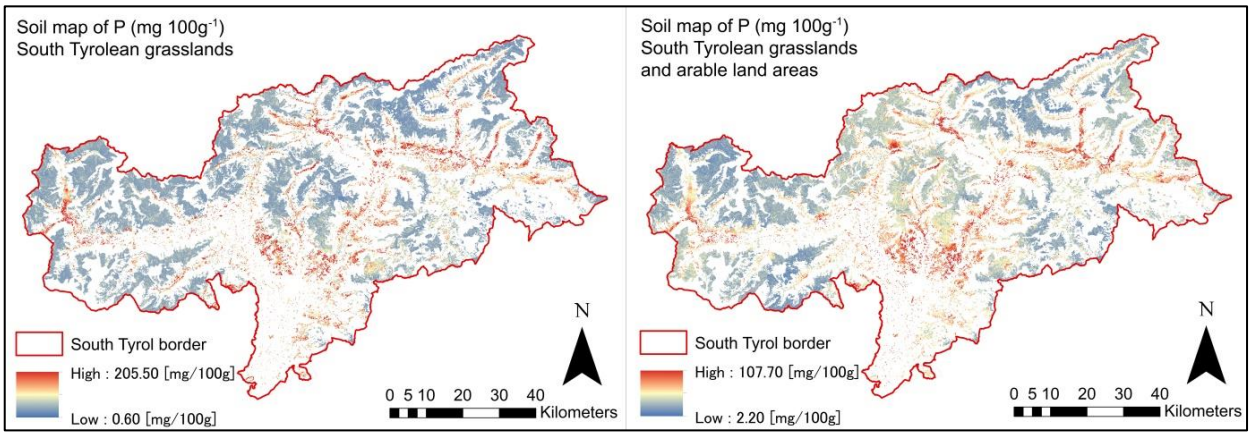


Figure 11: Distribution of P based on soil samples of grasslands (on the left) and samples of grasslands and arable land areas (on the right). (Validation values: RMSE = 10.6 (10.6) and R^2 = 0.5 (0.5))

3.3.4 Potassium (K)

Figure 12 shows the heterogeneous spatial pattern of potassium. The predicted values on grassland range from a minimum of 2.90 mg/100g to a maximum of 225.70 mg 100g^{-1} with a mean of 19 mg 100g^{-1}. On grassland and arable land the values range from a minimum of 5 mg 100g^{-1} to a maximum of 192.20 mg 100g^{-1} with a mean of 18.40 mg 100g^{-1} (**Table 4**). Here the predicted values tend to be in average slightly lower (**Fig. 13**). The fact that the northeastern part of South Tyrol (Wipptal, Eisacktal, Pustertal) and the Vinschgau district show the lowest K values trend, applies for both areas of interest. A municipality where particularly high K values are predicted is Kastelruth.

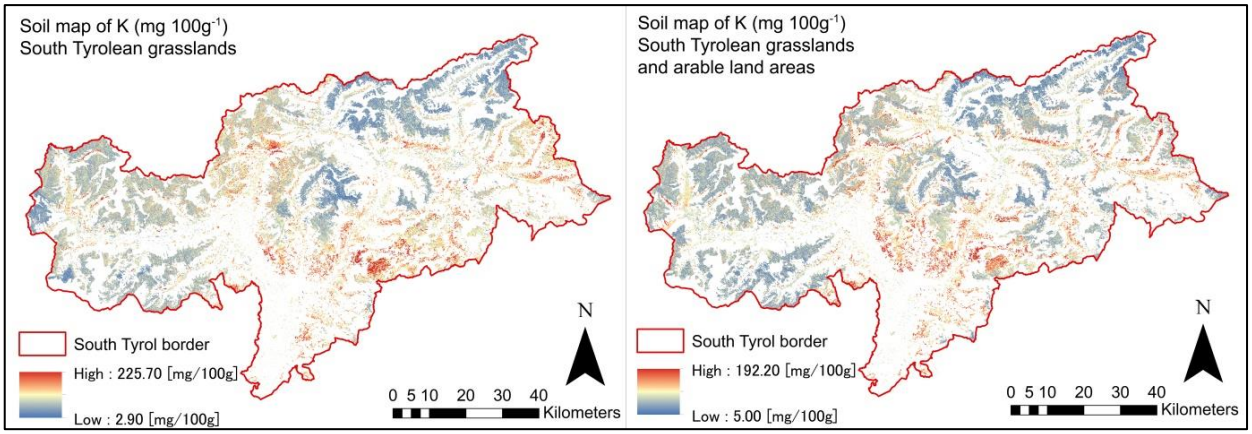


Figure 12: Distribution of K based on soil samples of grasslands (on the left) and samples of grasslands and arable land areas (on the right). (Validation values: RMSE = 20.7 (19.4) and R^2 = 0.6 (0.7))

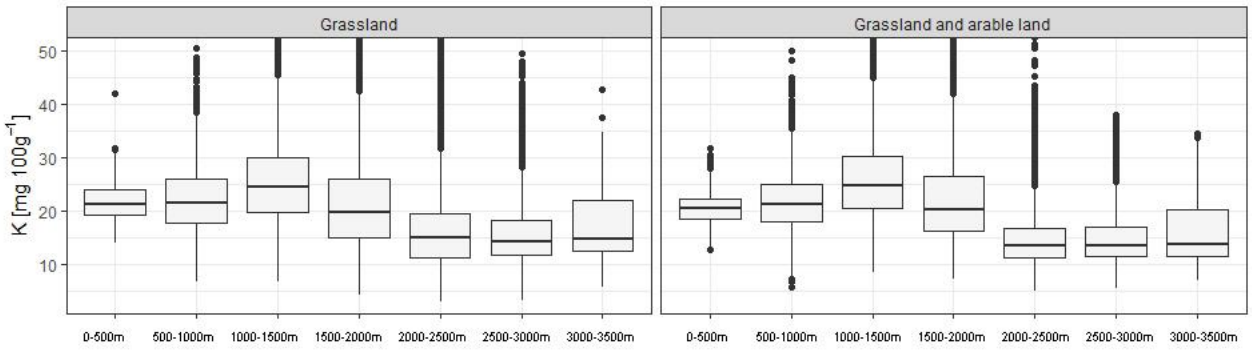


Figure 13: Boxplots of predicted K values on grassland (on the left) and on grassland and arable land (on the right) classified according to altitude. Just a part of the Y-axis is illustrated. There were outliers up to 225.70 m for grassland and 192.2 m for grassland and arable land (Table 4).

3.3.5 Nitrogen (N)

The spatial distribution of N is shown in **Figure 14**. The descriptive statistics in **Table 4** show that the distribution of the predicted N values is almost the same for both areas of interest. The values ranged from a minimum of 0.30% for grasslands and 0.20% for grasslands and arable land areas, to a maximum of 0.90% with a mean/median of 0.50% in both cases. On grassland and arable land most of the predicted values tend to be slightly lower. The predicted values slightly decrease at an altitude within 500 and 1000m on both areas of interest and then they increase with increasing altitude (**Fig. 15**). So, lower N values are predicted in lower as in higher altitudes. Also in this case as SOM, an interesting west-east gradient appears, where N decreases towards the east.

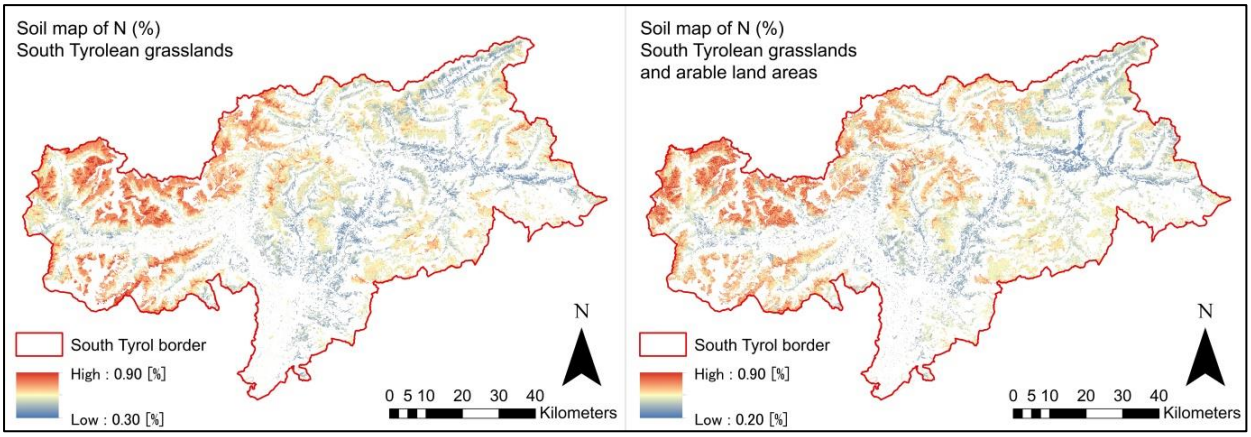


Figure 14: Distribution of N based on soil samples of grasslands (on the left) and samples of grasslands and arable land areas (on the right). (Validation values: RMSE = 0.2 (0.8) and R^2 = 0.1 (0.6))

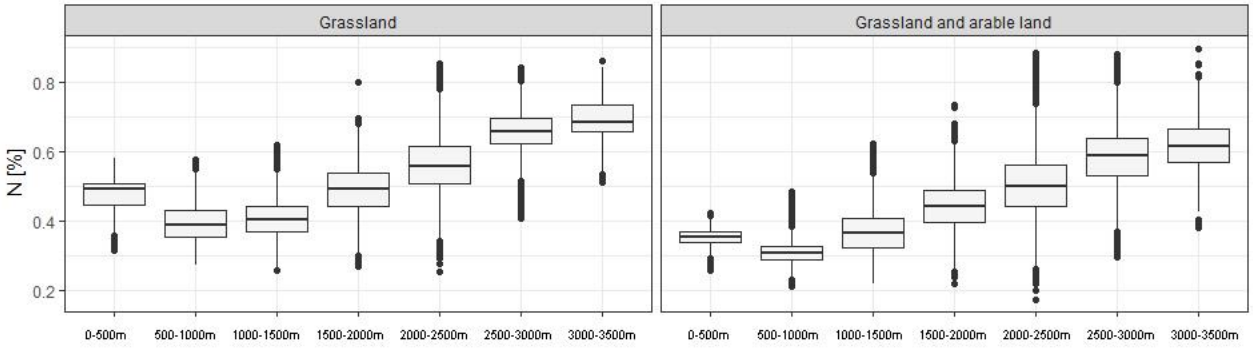


Figure 15: Boxplots of predicted N values on grassland (on the left) and on grassland and arable land (on the right) classified according to altitude.

3.3.6 Carbon (C)

The spatial distribution of C on grassland, and on grassland and arable land is shown in **Figure 16**. Here the same west-gradient as in the SOM and N models appears and the trend is similar to that of nitrogen. The predicted values are lower in lowlands and higher in higher altitudes (**Fig. 17**). The lowest C values are predicted at an altitude within 500 and 1000m in both areas of interest and then they increase with increasing altitude. The descriptive statistics of C are summarized in **Table 4**. C values range from 2.40% to 15.90% with a mean of 6.70% on grassland. These values tend to decrease slightly if the arable land areas are also taken into account. Here the values range from a minimum of 1.70% to a maximum of 19.30% with a lower mean of 6.20%.

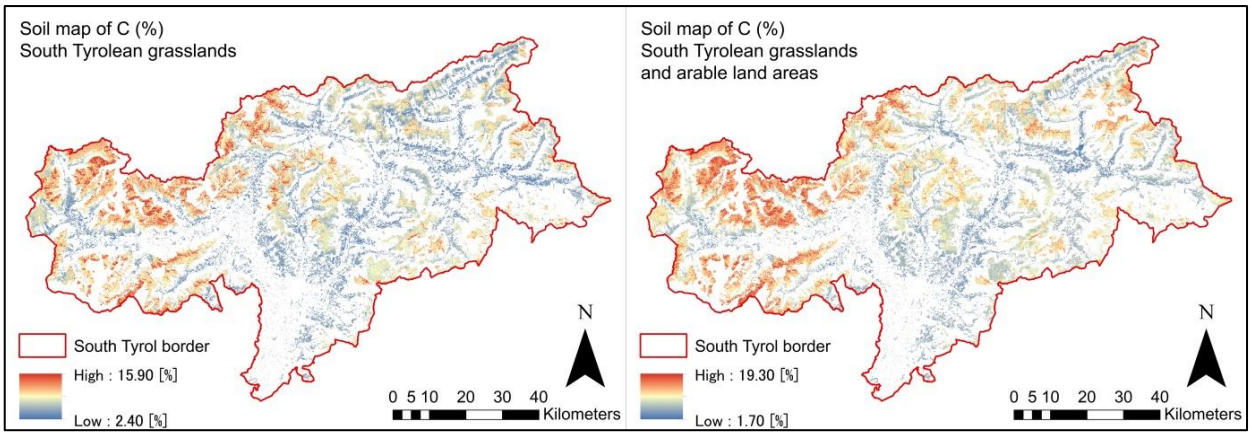


Figure 16: Distribution of C based on soil samples of grasslands (on the left) and samples of grasslands and arable land areas (on the right). (Validation values: RMSE = 1.6 (2.3) and R^2 = 0.4 (0.7))

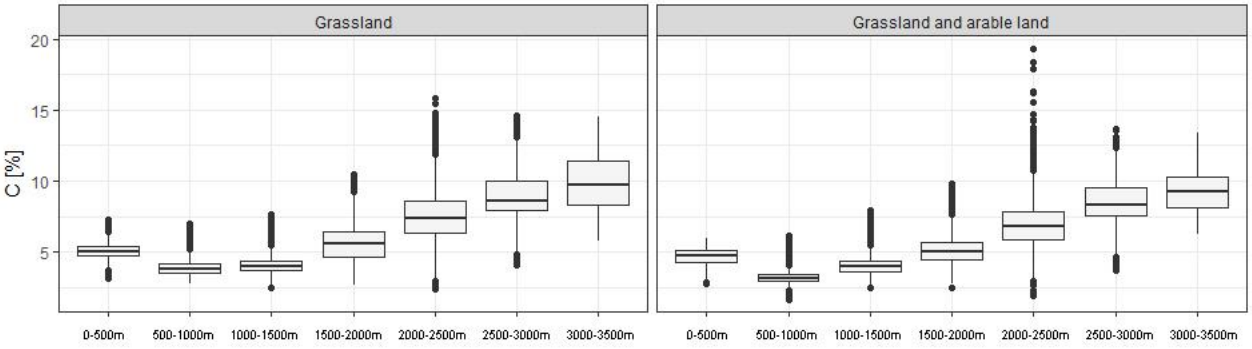


Figure 17: Boxplots of predicted C values on grassland (on the left) and on grassland and arable land (on the right) classified according to altitude.

3.3.7 Carbon to Nitrogen ratio (C:N)

The map of the C:N ratio is shown in **Figure 18**. On grassland, the predicted values range from a minimum of 8.90 to a maximum of 19.60 with a mean of 12.40.

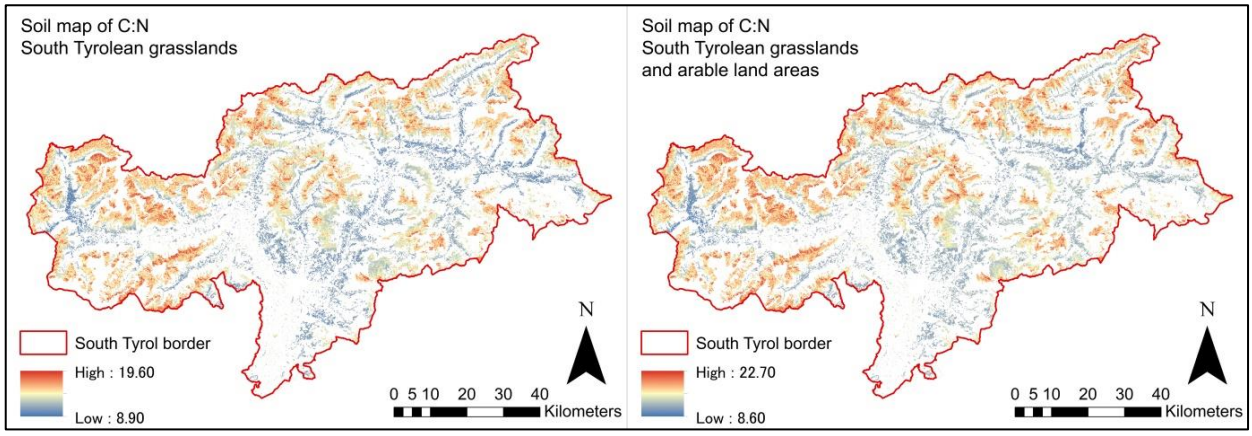


Figure 18: Distribution of C:N based on soil samples of grasslands (on the left) and samples of grasslands and arable land areas (on the right). (Validation values: RMSE = 2.0 (1.3) and R^2 = 0.4 (0.7))

On grassland and arable land they range from a minimum of 8.60 to a maximum of 22.70 with a mean of 12.70 (**Table 4**). Since C and N are both lower in lowlands and higher in higher altitudes (**Fig. 15, 17**), C:N shows also this trend, where a slight increase in C:N with increasing altitude can be observed (**Fig. 19**).

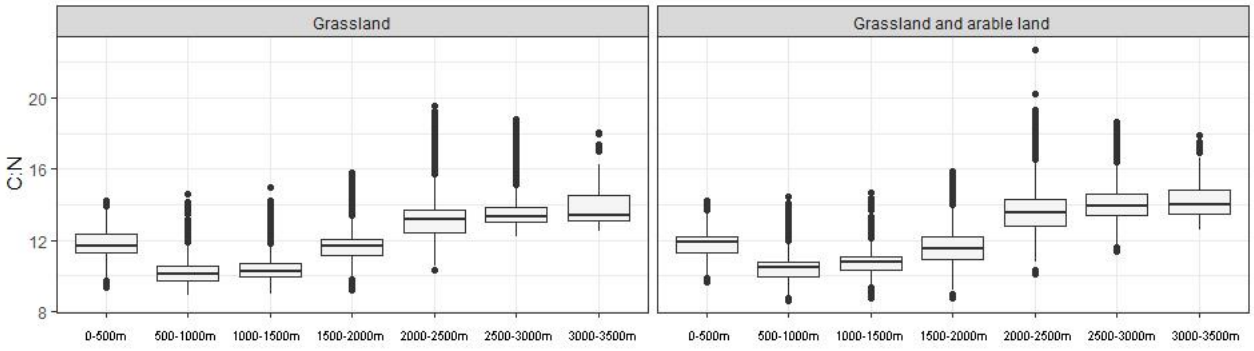


Figure 19: Boxplots of predicted C:N values on grassland (on the left) and on grassland and arable land (on the right) classified according to altitude.

3.4 Validation

The presented soil property maps are based on the best mathematical models found by the spatial prediction process that lead to the best validation results described in **Table 5**. The R^2 values range from 0.43 to 0.79 with a mean of 0.60. The highest R^2 was achieved by the model that predicted P on grasslands and the lowest R^2 by the model that predicted the C:N ratio on grassland. The NRMSE range from a minimum of 51.50 to a maximum of 80.00 with a mean of 65.12. The model with the lowest NRMSE is the one predicting pH on grassland. The highest NRMSE value was achieved by the model that predicted C on grassland. The NRMSE and R^2 values for the predictions on grassland and arable land are higher than those for the prediction models on grassland. There is one exception and that is N. Overall, the models that show a lower NRMSE show also a lower R^2.

Table 5: Validation results of the predictive models for each chemical parameter and area of interest. (G: Grassland. GA: Grassland and Arable land. RMSE: Root Mean Square Error. NRMSE: Normalized RMSE. R^2: R-Squared)

	SOM		pH		P		K		N		C		C:N	
	G	GA	G	GA	G	GA	G	GA	G	GA	G	GA	G	GA
RMSE	3.02	4.00	0.44	0.42	10.57	10.59	20.66	19.37	0.16	0.12	1.59	2.25	2.02	1.29
NRMSE	71.70	70.00	53.80	51.50	69.10	68.50	62.60	59.30	64.00	70.10	80.00	59.20	74.70	57.20
R^2	0.49	0.60	0.74	0.74	0.53	0.53	0.61	0.70	0.79	0.55	0.44	0.70	0.43	0.66

4. DISCUSSION

4.1 Distribution of soil chemical properties

The most significant predictor modelling the stocks of the chemical soil properties considered in this study was elevation. The effect of elevation is visible in chemical soil value distribution at lower and at higher elevations. It has a significant influence on the temperature and precipitation regimes (Liljequist and Cehak, 1979). Temperature and precipitation play together an important role, so also for the distribution of SOM (Nianpeng *et al.*, 2013; Lützow and Kögel-Knabner, 2009), that increases with increasing elevation. At lower altitudes it is warmer and this increase SOM turnover (Najera *et al.*, 2020), making the storage of SOM difficult (Torn *et al.*, 2009). Since increased precipitation accelerates SOM turnover (Han *et al.*, 2017), in contrast with storage, and the fact that the district of Vinschgau is characterized by dry climatic conditions (Della Chiesa *et al.*, 2014), explain the high predicted SOM values at higher altitudes, especially in the western part of South Tyrol. This interesting west-east gradient, which is also visible by the N and C maps, can be also explained by the climate-induced soil leaching. The western part of South Tyrol is characterized by in average higher altitudes and, as aforementioned, by dry climatic conditions. The Vinschgau valley is surrounded by high mountain regions with two important glaciers, the Ortler (3.902 m) and the Similaun (3.607 m). Since climate change leads to an increased snowmelt (Schlögel *et al.*, 2020) and it is claimed that in relatively dry climatic conditions mountain areas generally act as “water towers” (Bertoldi *et al.*, 2014), a large amount of water is released, especially at higher altitudes when temperature rises. Warm temperatures would normally allow an optimal chemical turnover (Najera *et al.*, 2020), but too wet conditions can compromise this (Lundin, 1995) and a storage follows. The spatial match of high SOM and high C can be explained by the fact that organic matter is the largest carbon reservoir in rapid exchange with atmospheric CO, and is thus important as a potential source and sink of greenhouse gases (Fischlin and Gyalistras, 1997). On the other hand the prediction of higher N values at higher altitudes is surprising. N, as well as P and K, are in this study to be understand as stocks, this means that higher N, P or K values do not mean more available nutrients for plants. Most of the soil nutrients are often unavailable for plants, so they must first be made available by soil microorganisms that play therefore a key role in enhancing soil efficiency and plant growth (Miransari, 2013). N, P and K belong to the most used fertilizers (Bahn, 2020), mainly used at lower altitudes, because upland grassland is more difficult to reach with machinery and so fertilization processes required more effort (Tasser and Tappeiner, 2002). Thus, the N values would have to be predicted higher at lower altitudes and not at higher, like P and K. A past arable land use of grassland areas could explain the lower N values (Quemada *et al.*, 2020), therefore it would be important to add historical land use information as predictor in order to evaluate its influence. N addition reduces pH (Tian and Niu, 2015), so the pH was predicted lower where the N

values were predicted higher, at higher altitudes. This especially in the Dolomites, Wipptal and in the upper Vinschgau due to the very high carbonate contents of the bedrock (Stimpfl *et al.*, 2006). Even though the property maps show meaningful general patterns, with the exception of the N distribution, there are still some deficiencies.

4.2 Deficiencies

4.2.1 Soil samples

Since digital soil mapping activities have strongly increased in recent years, there are many soil maps produced by quantitative (statistical) methods. In order to achieve good map quality a high number of soil samples and effective validation models are required (Brus *et al.*, 2011). Even though a validation was done in this study with good results, the amount and the distribution of the chemical soil samples across altitudes can be put into discussion, since the predicted chemical properties correspond only partly with the true soil composition. The areas of interest on which the predictions were made, are on average at higher altitudes as the soil samples. A mean of around 357 samples from grassland (510 for SOM and pH, 454 for P and K, 108 for N, C and C:N) and 13 samples from arable land (17 for SOM, pH, P, and K, 6 for N, C and C:N) was available for the modelling and prediction of area-with chemical soil properties. This amount of source data is significantly less than the standard sampling of 1 sample ha^{-1} to 1 sample every 5 ha (mean of 0.02 samples every 5 ha) (Nanni *et al.*, 2011). For comparison, the remarkably high number of about 16,000 soil samples (mean of 3.25 samples every 5 ha) is available for the prediction of chemical soil properties on South Tyrolean apple orchards and vineyards (Della Chiesa *et al.*, 2019a). These areas require particular attention, as many fertilizers and pesticides are used. Since such agricultural practices have a strong impact on ecosystem functions and consequently on ecosystem services (Garcia *et al.*, 2018; Demestihas *et al.*, 2017), the European Commission pushes for sustainability in agriculture across the EU through the common agricultural policy (CAP) (European Commission, 2021). That is why owners have to collect regularly soil information. As a consequence, a cooperation between farmers (as data sources) and soil scientists of local organisations (Research Centres: Eurac and Laimburg) was established, in order to have a large amount of data to promote a successful framework for extensive and long-term soil monitoring in these areas. The same routine would be needed for grassland, where intesive fertilization management can also lead to serious environmental impacts and indirect costs to society (King *et al.*, 2015). Grassland soil health plays a key role for livestock food quality and the related quality of the milk products. This is a reason to conduct more research studies on grassland and do a more intensive soil sampling to reduce negative impacts on ecosystem and consequently also to contribute to a possible improvement in the quality of the property maps presented in this study.

4.2.2 Predictors

The most important predictor of the chemical soil properties in this study was elevation and the resulting elevation-related climate changes. Other significant predictors were slope and aspect. Livestock influenced SOM, pH, P, but especially N and C, since both N and C cycles are closely positive connected to livestock's role in land use and land use change (Steinfeld and Wassenaar, 2007). Geological and land use information had a significant influence only on P, K and pH. The intensively use of hay meadows was important for the prediction of P and K, since fertilization processes are linked to intensive land use, whereas the geological information about alkalinity of soil played a key role in the prediction of pH. These predictors were almost the same on both areas of interest, with one exception. The predictor describing wetness had a slightly higher influence on the prediction of the chemical soil properties on grassland and arable land. Grassland areas are variably sloped, while arable land areas are generally flat, so water can therefore run off more slowly and can influence differently several chemical soil properties. The percent of explained variance by the several predictors ranged from a minimum of 6.3% (prediction of N on grassland) to a maximum of 57.9% (prediction of pH on grassland) with a mean of 31.1% and a standard deviation of 16.7%. These results show that some important predictors are still missing. To say which ones, further research is needed. As aforementioned the historical land use information of grassland areas could for example be an additional predictor to take into account.

4.2.3 Method

The DSM methods used in this study have been tested by various authors (Hengl *et al.*, 2018; Hengl *et al.*, 2004) and have been proven to have good predictive capability for continuous variables. The final models were choosen taking into account the validation values, R^2 and the RMSE. Both are important measures of the goodness of the model. The RMSE describes the error of the model, so it is a direct measure of when a prediction is wrong and it has the same unit of measure as the studied variable. This is why it was used to choose the best model. Since the aforementioned percent of variance was not always so high, there is the possibility to choose the best model in two steps. Since different combinations of predictors were tested, the first step would be to find the Random Forest models and the predictor selections, for example the first five that explained the highest percent of variance. After that, the second step would be to do with these models, the Ordinary Kriging estimation and then to choose the best final model, according to the RMSE or R^2. This two-step approach could improve the developed program.

4.3 Usage of the developed program

The R^2 validation values, ranging from 0.43 (prediction of C:N on grassland) to 0.79 (prediction of N on grassland) with a mean of 0.60, confirm the right functioning of the developed program that could promote long-term planning sustainable use of fertilizers on South Tyrolean grassland. This in combination with better soil sample data, since the sample densities and its distribution across altitudes in this study were not the best. Since climate change and intensifying human activities lead to the degrading of many areas worldwide, it can also help in general soil investigations doing spatial predictions of several chemical soil properties. They can be an important tool to estimate soil degradation and to identify areas where concrete interventions are needed. This applies especially for regions with intensive land use and where availability of soil data is limited.

5. CONCLUSION AND OUTLOOK

Even though the results of the validation were good and the property maps show meaningfully general pattern, with the exception of the N distribution, there are still some areas of possible improvement, starting from the soil sample data sourcing.

The samples were evenly distributed across South Tyrol, but on average at lower altitudes as the areas of interest on which the predictions were made. The Matschertal in the district of Vinschgau was, with around 10% of all available soil samples, optimally sampled. This unbalanced amount and distribution of the chemical soil samples across altitudes should serve as motivation to carry out more soil investigations on grassland, where intensive fertilization management can lead to serious environmental impact and indirect costs to society. Grassland plays a key role for livestock food and the related quality of the milk products, therefore an intensive soil monitoring on grassland would be justified.

Another point that would contribute to reduce negative impacts on ecosystem and also probably improve the quality of the property maps presented in this study is the integration of several predictors that were not considered so far. The mean percent of explained variance, around 30%, shows that some important predictors are still missing. Since many grassland areas were once used for crop production or still used to be cultivated in alternation between arable farming and grassland, the historical land use information could be an additional predictor to explain probably the low N values at lower altitudes.

In order to improve the developed program, already based on methods with good predictive capability for continuous variables, a two-step approach could help. The final models in this study were choosen considering just the validation values R^2 and the RMSE. Since the aforementioned percent of explained variance was not so high, there is the possibility to choose firstly the Random Forest models explaining the highest percent of variance, then to do with them the Ordinary Kriging estimation and finally to choose the best final model according to the RMSE or R^2.

SOFTWARE

The softwares used in this study are Microsoft Excel and SPSS 26® (www.ibm.com) to organize and prepare the source data, Esri ArcGIS 10.7.1® (www.esri.com) to produce the raster maps of the areas of interest and those of the predictor variables, and R-Studio 1.3.959® (www.cran.r-project.org) to write the script containing the mathematical models that provide the spatial prediction. The maps and layout were produced in Esri ArcGIS 10.7.1®.

ACKNOWLEDGMENTS

This work was supported by the University of Innsbruck and the Eurac research based in Bolzano. I would like to thank especially my supervisor Priv.-Doz. Dr. Erich Tasser for giving me the opportunity to follow this project and for his constant supervision and guidance throughout the work. I thank Giulio Genova M.Sc. for his continuous support and tireless dedication. I would like to thank the Laimburg Research center for providing most of the source data and Stefano Della Chiesa M.Sc. for the precious advices. Last but not least, I would like to thank my family for their support during my academic years, which were not always easy.

REFERENCES

ASTAT (2017), "Südtirol in Zahlen 2017", *Autonome Provinz Bozen - Südtirol (Landesinstitut für Statistik)*.

ASTAT (2019), "Südtirol in Zahlen 2019", *Autonome Provinz Bozen - Südtirol (Landesinstitut für Statistik)*.

Baer, S.G., Kitchen, D.J., Blair, J.M. and Rice, C.W. (2002), "Changes in ecosystem structure and function along a chronosequence of restored grasslands", *Ecological applications a publication of the Ecological Society of America*, Vol. 12 No. 6, pp. 1688–1701.

Bahn, M. (2020), *Functional Ecology*, Lecture (VU+SE), University of Innsbruck.

Bai, Y., Ma, L., Degen, A.A., Rafiq, M.K., Kuzyakov, Y., Zhao, J., Zhang, R., Zhang, T., Wang, W., Li, X., Long, R. and Shang, Z. (2020), "Long-term active restoration of extremely degraded alpine grassland accelerated turnover and increased stability of soil carbon", *Global change biology*, Vol. 26 No. 12, pp. 7217–7228.

Bertoldi, G., Della Chiesa, S., Notarnicola, C., Pasolli, L., Niedrist, G. and Tappeiner, U. (2014), "Estimation of soil moisture patterns in mountain grasslands by means of SAR RADARSAT2 images andhydrological modeling", *Journal of Hydrology*, Vol. 516 No. 2, pp. 245–257.

Brandner, R. (1980), "Tektonische Uebersichtskarte von Tirol 1:600.000 (tectonic map of Tyrol, scale, 1:600.000).", *Universitätsverlag Wagner, Innsbruck (in German)*.

Breiman, L. (2001), "Random Forests", *Machine Learning*, Vol. 45 No. 1, pp. 5–32.

Brus, D.J., Kempen, B. and Heuvelink, G.B.M. (2011), "Sampling for validation of digital soil maps", *European Journal of Soil Science*, Vol. 62 No. 3, pp. 394–407.

Burrough, P.A. and McDonell, R.A. (1998), "Principles of Geographical Information Systems", *Oxford University Press, New York*, 190 pp.

Busch, G., Kühl, S. and Gauly, M. (2018), "Consumer expectations regarding hay and pasture-raised milk in South Tyrol. Konsumentenerwartungen an Heu- und Weidemilch in Südtirol", *Austrian Journal of Agricultural Economics and Rural Studies*, Vol. 27.11.

Camera, C., Zomeni, Z., Noller, J.S., Zissimos, A.M., Christoforou, I.C. and Bruggeman, A. (2017), "A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization", *Geoderma*, Vol. 285 No. 6, pp. 35–49.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D.T., Duan, Z. and Ma, J. (2017), "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility", *CATENA*, Vol. 151 No. 1, pp. 147–160.

Della Chiesa, S., Bertoldi, G., Niedrist, G., Obojes, N., Endrizzi, S., Albertson, J.D., Wohlfahrt, G., Hörtnagl, L. and Tappeiner, U. (2014), "Modelling changes in grassland hydrological cycling along an elevational gradient in the Alps", *Ecohydrology*, Vol. 7 No. 6, pp. 1453–1473.

Della Chiesa, S., Genova, G., La Cecilia, D. and Niedrist, G. (2019a), "Phytoavailable phosphorus (P 2 O 5) and potassium (K 2 O) in topsoil for apple orchards and vineyards, South Tyrol, Italy", *Journal of Maps*, Vol. 15 No. 2, pp. 555–562.

Della Chiesa, S., La Cecilia, D., Genova, G., Balotti, A., Thalheimer, M., Tappeiner, U. and Niedrist, G. (2019b), “Farmers as data sources: Cooperative framework for mapping soil properties for permanent crops in South Tyrol (Northern Italy)”, *Geoderma*, Vol. 342, pp. 93–105.

Demestihas, C., Plénet, D., Génard, M., Raynal, C. and Lescourret, F. (2017), “Ecosystem services in orchards. A review”, *Agronomy for Sustainable Development*, Vol. 37 No. 2, p. 581.

European Commission (2021), “Sustainable agriculture in the EU”.

FAO and ITPS (2015), *Status of the world's soil resources: Main report*, FAO; ITPS, Rome.

Fischlin, A. and Gyalistras, D. (1997), “Assessing Impacts of Climatic Change on Forests in the Alps”, *Global Ecology and Biogeography Letters*, Vol. 6 No. 1, p. 19.

Garcia, L., Celette, F., Gary, C., Ripoche, A., Valdés-Gómez, H. and Metay, A. (2018), “Management of service crops for the provision of ecosystem services in vineyards: A review”, *Agriculture, ecosystems & environment*, Vol. 251, pp. 158–170.

Genova, G. (2017), *Spatial Distribution Assessment of Cu, Zn, PH and Soil Organic Matter in South Tyrolean Permanent Crops*, Master Thesis dissertation, *Università degli studi della Tuscia*.

Grimm, R. and Behrens, T. (2010), “Uncertainty analysis of sample locations within digital soil mapping approaches”, *Geoderma*, Vol. 155 No. 3-4, pp. 154–163.

Han, C., Wang, Z., Si, G., Lei, T., Yuan, Y. and Zhang, G. (2017), “Increased precipitation accelerates soil organic matter turnover associated with microbial community composition in topsoil of alpine grassland on the eastern Tibetan Plateau”, *Canadian journal of microbiology*, Vol. 63 No. 10, pp. 811–821.

Hengl, T. (2006), “Finding the right pixel size”, *Computers & Geosciences*, Vol. 32 No. 9, pp. 1283–1298.

Hengl, T., Heuvelink, G.B.M. and Stein, A. (2004), “A generic framework for spatial prediction of soil variables based on regression-kriging”, *Geoderma*, Vol. 120 No. 1-2, pp. 75–93.

Hengl, T., Kempen, B. and Heuvelink, G. (2016), *GSIF: global soil information facilities: R package, Version 0.5-3*.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M. and Gräler, B. (2018), “Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables”, *PeerJ*, Vol. 6, e5518.

Hinojosa, L., Tasser, E., Rüdisser, J., Leitinger, G., Schermer, M., Lambin, E.F. and Tappeiner, U. (2019), “Geographical heterogeneity in mountain grasslands dynamics in the Austrian-Italian Tyrol region”, *Applied Geography*, Vol. 106 No. 9, pp. 50–59.

Janssens, F., Peeters, A., Tallowin, J.R.B., Bakker, J.P., Bekker, R.M., Fillat, F. and Oomes, M.J.M. (1998), “Relationship between soil chemical factors and grassland diversity F.”, *Plant and Soil*, Vol. 202 No. 1, pp. 69–78.

Jiao, F., Shi, X.-R., Han, F.-P. and Yuan, Z.-Y. (2016), “Increasing aridity, temperature and soil pH induce soil C-N-P imbalance in grasslands”, *Scientific reports*, Vol. 6, p. 19601.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M. and Vries, F. de (2012), "Efficiency Comparison of Conventional and Digital Soil Mapping for Updating Soil Maps", *Soil Science Society of America Journal*, Vol. 76 No. 6, pp. 2097–2115.

King, K.W., Williams, M.R., Macrae, M.L., Fausey, N.R., Frankenberger, J., Smith, D.R., Kleinman, P.J.A. and Brown, L.C. (2015), "Phosphorus transport in agricultural subsurface drainage: a review", *Journal of environmental quality*, Vol. 44 No. 2, pp. 467–485.

Kollmann, K. (2012), "Klima- und landnutzungsbedingte Bodenverteilung im Matschertal, Südtirol", M.Sc. Thesis, *Universität Innsbruck*.

Krige, D.G. (1951), "A statistical approach to some basic mine valuation problems on the Witwatersrand", *Journal of the Southern African Institute of Mining and Metallurgy*, Vol. 52 No. 2, pp. 119–139.

Lal, R. (2009), "Challenges and opportunities in soil organic matter research", *European Journal of Soil Science*, Vol. 60 No. 2, pp. 158–169.

Lark, R.M. (2000), "A comparison of some robust estimators of the variogram for use in soil survey", *European Journal of Soil Science*, Vol. 51 No. 1, pp. 137–157.

Li, J. and Heap, A.D. (2014), "Spatial interpolation methods applied in the environmental sciences: A review", *Environmental Modelling & Software*, Vol. 53 No. 9, pp. 173–189.

Liljequist, G.H. and Cehak, K. (1979), "Der Niederschlag", in Liljequist, G.H. and Cehak, K. (Eds.), *Allgemeine Meteorologie*, Vieweg+Teubner Verlag, Wiesbaden, pp. 155–174.

Lundin, L. (1995), "Soil water chemistry dependence on water pathways and turnover", *Water, Air, & Soil Pollution*, Vol. 85 No. 3, pp. 1695–1700.

Lützow, M. von and Kögel-Knabner, I. (2009), "Temperature sensitivity of soil organic matter decomposition—what do we know?", *Biology and Fertility of Soils*, Vol. 46 No. 1, pp. 1–15.

McBratney, A., Field, D.J. and Koch, A. (2014), "The dimensions of soil security", *Geoderma*, Vol. 213 No. 1, pp. 203–213.

McBratney, A.B., Mendonça Santos, M.L. and Minasny, B. (2003), "On digital soil mapping", *Geoderma*, Vol. 117 No. 1-2, pp. 3–52.

Minasny, B. and McBratney, A.B. (2016), "Digital soil mapping: A brief history and some lessons", *Geoderma*, Vol. 264 No. 1, pp. 301–311.

Miransari, M. (2013), "Soil microbes and the availability of soil nutrients", *Acta Physiologiae Plantarum*, Vol. 35 No. 11, pp. 3075–3084.

Moosgöller, B. (ongoing), "Comparison of biodiversity of soil macrofauna of two differing management systems (apple orchards, meadow orchards) in South Tyrol (I)", M.Sc. Thesis, *Universität Innsbruck*.

Najera, F., Dippold, M.A., Boy, J., Seguel, O., Koester, M., Stock, S., Merino, C., Kuzyakov, Y. and Matus, F. (2020), "Effects of drying/rewetting on soil aggregate dynamics and implications for organic matter turnover", *Biology and Fertility of Soils*, Vol. 56 No. 7, pp. 893–905.

Nanni, M.R., Povh, F.P., Demattê, J.A.M., Oliveira, R.B.d., Chicati, M.L. and Cezar, E. (2011), "Optimum size in grid soil sampling for variable rate application in site-specific management", *Soil Science Society of America Journal*, Vol. 68 No. 3, pp. 386–392.

Naresh, D.R.K. (2020), *Advances in Agriculture Sciences*, AkiNik Publications.

Neina, D. (2019), "The Role of Soil pH in Plant Nutrition and Soil Remediation", *Applied and Environmental Soil Science*, Vol. 2019 No. 3, pp. 1–9.

Nianpeng, H., Ruomeng, W., Yang, G., Jingzhong, D., Xuefa, W. and Guirui, Y. (2013), "Changes in the temperature sensitivity of SOM decomposition with grassland succession: implications for soil C sequestration", *Ecology and evolution*, Vol. 3 No. 15, pp. 5045–5054.

Pan, Y., Cassman, N., Hollander, M. de, Mendes, L.W., Korevaar, H., Geerts, R.H.E.M., van Veen, J.A. and Kuramae, E.E. (2014), "Impact of long-term N, P, K, and NPK fertilization on the composition and potential functions of the bacterial community in grassland soil", *FEMS microbiology ecology*, Vol. 90 No. 1, pp. 195–205.

Quemada, M., Lassaletta, L., Jensen, L.S., Godinot, O., Brentrup, F., Buckley, C., Foray, S., Hvid, S.K., Oenema, J., Richards, K.G. and Oenema, O. (2020), "Exploring nitrogen indicators of farm performance among farm types across several European case studies", *Agricultural Systems*, Vol. 177 No. 1554, p. 102689.

Ressi, W., Posch, K., Gruber, A., Melcher, D., Bogner, D., Aigner, S., Obweger, A., Hasler, S., Rieder, C., Klein, R., Frieß, T., Holzinger, W., Schlosser, L. and Tasser, E. (2014), "AlmWaal: Endverwendungsnachweis zum Projekt SPA04/24", *Umweltbüro Klagenfurt, Klagenfurt*.

Robinson, T.P. and Metternicht, G. (2006), "Testing the performance of spatial interpolation techniques for mapping soil properties", *Computers and Electronics in Agriculture*, Vol. 50 No. 2, pp. 97–108.

Rottersteiner, M. (2020), "Bodenmakrofauna auf Ackerflächen in Südtirol", Dipl. Thesis, *Universität Innsbruck*.

Rüdisser, J., Tasser, E., Peham, T., Meyer, E. and Tappeiner, U. (2015), "The dark side of biodiversity: Spatial application of the biological soil quality indicator (BSQ)", *Ecological Indicators*, Vol. 53 No. 2, pp. 240–246.

Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M.d.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A. and ZHANG, G.-l. (2009), "Environmental science. Digital soil map of the world", *Science (New York, N.Y.)*, Vol. 325 No. 5941, pp. 680–681.

Schlögel, R., Kofler, C., Gariano, S.L., van Campenhout, J. and Plummer, S. (2020), "Changes in climate patterns and their association to natural hazard distribution in South Tyrol (Eastern Italian Alps)", *Scientific reports*, Vol. 10 No. 1, p. 5022.

Scull, P., Franklin, J., Chadwick, O.A. and McArthur, D. (2003), "Predictive soil mapping: a review", *Progress in Physical Geography: Earth and Environment*, Vol. 27 No. 2, pp. 171–197.

Sørensen, R., Zinko, U. and Seibert, J. (2006), “On the calculation of the topographic wetness index: evaluation of different methods based on field observations”, *Hydrology and Earth System Sciences*, Vol. 10 No. 1, pp. 101–112.

Steinfeld, H. and Wassenaar, T. (2007), “The Role of Livestock Production in Carbon and Nitrogen Cycles”, *Annual Review of Environment and Resources*, Vol. 32 No. 1, pp. 271–294.

Stimpfl, E., Aichner, B., Thaler, C., Vidoni, A., Andreaus, O. and Cassar, A. (2006), “The state of grassland soils in South Tyrol (Italy)”, *Laimburg Journal*, No. 3, pp. 2–73.

Tasser, E. (Ed.) (2012), *Wir Landschaftmacher: Vom Sein und Werden der Kulturlandschaft in Nord-, Ost- und Südtirol*, Verl.-Anst. Athesia, Bozen.

Tasser, E. and Tappeiner, U. (2002), “Impact of land use changes on mountain vegetation”, *Applied Vegetation Science*, Vol. 5 No. 2, pp. 173–184.

Tasser, E., Tappeiner, U. and Cernusca, A. (2000), “Südtirols Almen im Wandel”, *Athesia, Bozen*.

Tian, D. and Niu, S. (2015), “A global analysis of soil acidification caused by nitrogen addition”, *Environmental Research Letters*, Vol. 10 No. 2, p. 24019.

Tibbett, M., Gil-Martínez, M., Fraser, T., Green, I.D., Duddigan, S., Oliveira, V.H. de, Raulund-Rasmussen, K., Sizmur, T. and Diaz, A. (2019), “Long-term acidification of pH neutral grasslands affects soil biodiversity, fertility and function in a heathland restoration”, *CATENA*, Vol. 180 No. 7–8, pp. 401–415.

Torn, M.S., Swanston, C.W., Castanha, C. and Trumbore, S.E. (2009), “Storage and Turnover of Organic Matter in Soil”, in Senesi, N., Xing, B. and Huang, P.M. (Eds.), *Biophysico-Chemical Processes Involving Natural Nonliving Organic Matter in Environmental Systems*, John Wiley & Sons, Inc, Hoboken, NJ, USA, pp. 219–272.

Vaysse, K. and Lagacherie, P. (2015), “Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France)”, *Geoderma Regional*, Vol. 4 No. Supplement 1, pp. 20–30.

Wiesmeier, M., Barthold, F., Blank, B. and Kögel-Knabner, I. (2011), “Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem”, *Plant and Soil,* 2011, pp. 7–24.

WIFO (2019), “Economy in figures - Die Südtiroler Wirschaft - Aktuelle Daten, Indikatoren und Entwicklungen”, *Institut für Wirtschaftforschung der Handelskammer,* 2019.

Zimmermann, P., Tasser, E., Leitinger, G. and Tappeiner, U. (2010), “Effects of land-use and land-cover pattern on landscape-scale biodiversity in the European Alps”, *Agriculture, ecosystems & environment*, Vol. 139 No. 1-2, pp. 13–22.

PROGRAM FOR MAPPING CHEMICAL SOIL PROPERTIES:
tested on grassland and arable land in South Tyrol, Northern Italy

universität innsbruck
eurac research

Master thesis of Elvis Burchia https://orcid.org/0000-0001-8925-2009
Software and documentation are available online: https://github.com/elvisburchia/DigitalSoilMappingSoftware

APPENDIX

1. Table: Predictor reclassification (land use) 2
2. Table: Predictor reclassification (detailed geology) 3
3. Table: Predictor reclassification (geology – alkalinity of soil) 4
4. Table: Predictor reclassification (geology – alkalinity of soil) 5
5. Table: Descriptive statistics of the source data (original and transformed) 6
6. Spatial interpolation and prediction method 7
7. Program: instruction manual 10
    a. Input data 10
    b. Program changes in R needed 12
    c. Tips 16
    d. Output explanation 18
8. References 28
9. Figures and tables 27

INTRODUCTION

Digital soil mapping (DSM) is extremely important, because knowledge of soil condition is a prerequisite for appropriate soil protection and sustainable cultivation. For this reason a program in R, based on DSM methods was developed. The program does, starting from chemical soil samples, a large-scale prediction for areas around the sampled locations (**Figure 1**). It can predict any continuous chemical soil properties (e.g. SOM, pH, etc.) with the use of several predictors (e.g. aspect, slope, etc.). High resolution soil property maps, which are the final output, can fill knowledge gaps of detailed spatially-distributed information of chemical soil parameters in any areas.

1 TABLE: Predictor reclassification (land use)

Table 1: Binary reclassification of land use information (land use).
The mixed rotation meadows ("gemischte Wechselwiese") are counted to grassland if only grassland is considered as area of interest. In the second area of interest, where arable land is also taken into account, these meadows are rated as crops.
(Org. Code: original code. G: Grassland. GA: Grassland and Arable land)

Code		
6	pastures	Org. Code
	Alpe (bestockt 20%)	AL2
	Alpe (bestockt 50%)	AL3
	Alpe (ohne Tara)	AL1
	Alpe (Tara 70%)	AL9
	Alpe (versteint 20%)	AL4
	Alpe (versteint 50%)	AL5
	Potenziell beweidbare Almfläche (ohne Tara)	AL6
	Potenziell beweidbare Almfläche (Tara 20%)	AL7
	Potenziell beweidbare Almfläche (Tara 50%)	AL8
	Weitere Flächen	ANA
	Weide	PA1
	Weide (Tara 20%)	PA2
	Weide (Tara 50%)	PA3
	Brachflache - EFA	AA10
7	intensively used hay meadows	
	Wiese (Dauerwiese)	AP2
	Wiese (Dauerwiese Tara 20%)	AP4
	Gemischte Wechselwiese	AF1
8	extensively used hay meadows	
	Wiese (halbschürig)	AP3
	Wiese (halbschürig Tara 20%)	AP5
9	grassland special area	
	Wiese Sonderfläche	AS
	Wiese Sonderfläche (Tara 20%)	AS1
	Wiese Sonderfläche (Tara 50%)	AS2
30	arable land	
	Blumenkohl	AA9
	Feldgemüsebau	AA3
	Getreide	AA1
	Kopfkohl	AA8
	Kräuterbau	AA4
	Radicchio	AA7
	Salate	AA6
	Klee	AF4
	Luzerne	AF3
	Mais	AF2
	Gemischte Wechselwiese	AF1

2 TABLE: Predictor reclassification (detailed geology)

Table 2: Binary reclassification of detailed geological information.
(G: Grassland. GA: Grassland and Arable land)

Code	
11	shist
	Bündnerschiefer mit Ophiolithen
	Altes Dach und Untere Schieferhülle
12	vulcanite
	Triassische Vulkanite und Vulkanoklastite
	Permische Vulkanite: Etschtaler Vulkanit-Gruppe
	Permische Vulkanite: Orobie Vulkanite
13	vinschgau shear zone
	Vinschgauer Scherzone
14	quaternary depositions
	Quartäre Ablagerungen
15	sediment sequences
	Permo-jurassische Sedimentabfolge
	Permomesozoische Sedimentabfolge der Nördlichen Kalkalpen
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Brenner Mesozoikum
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Pennes Schuppe
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Tarntaler Permomesozoikum
	Permomesozoische Sedimentabfolge der Nördlichen Kalkalpen
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Lienzer Dolomiten
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Drauzug
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Lienzer Dolomiten
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Kalkstein Schuppe
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: S-charl
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Jaggl
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: S-charl
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Ortler Decke
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Quettervals Decke
	Kretazische und paleogene Sedimentabfolge
16	plutons
	Permische Plutone: Ifinger-Brixen
	Permische Plutone: Brixen
	Permische Plutone: Ifinger
	Permische Plutone: Marteller Granit
	Permische Plutone: Kreuzberg
	Permische Plutone: Cima d'Asta
	Permische Plutone: Klausen
	Permische Plutone: Tirano Gabbro
	Permische Plutone: Sondalo Gabbro
	Permische Plutone: Dos del Sabion
	Tertiäre Plutone: tonalitische Lamelle
	Tertiäre Plutone
	Tertiäre Plutone: Rieserferner
	Tertiäre Plutone: Adamello Tonalit
17	central gneiss
	Zentralgneis
18	quartz phyllite
	Niedriggradig metamorphes varisziches Grundgebirge: Brixen Quartzphyllit
	Niedriggradig metamorphes varisziches Grundgebirge: Agordo Quartzphyllit
	Niedriggradig metamorphes varisziches Grundgebirge: Val Sugana Quartzphyllit
19	insignificant alpine metamorphosis
	Tonale-Decke - Einheit mit hochgradiger variszischer und unbedeutender alpiner Metamorphose
	Tektonische Einheiten Passaier - Einheit mit mittel-hochgradiger variszischer und unbedeutender alp. Met.
	Antholz-Einheit - Einheit mit mittel-hochgradiger variszischer und unbedeutender alpiner Metamorphose
	Sesvenna-Einheit - Einheit mit mittel-hochgradiger variszischer und unbedeutender alpiner Metamorphose
	Innsbrucker Phyllit - Einheit mit niedriggradiger variszischer und unbedeutender alpiner Metamorphose
	Landecker Phyllit - Einheit mit niedriggradiger variszischer und unbedeutender alpiner Metamorphose
	Zebrù-Schuppe - Einheit mit niedriggradiger variszischer und unbedeutender alpiner Metamorphose
	Thurntaler Einheit - Einheit mit mittelgradiger variszischer und unbedeutender alpiner Metamorphose
20	medium-low alpine metamorphosis
	Marlinger Schuppe - Einheit mit mittel-hochgradiger variszischer und niedriggradiger alpiner Metamorphose
	Campo-Decke - Einheit mit mittel-hochgradiger variszischer und niedriggradiger alpiner Metamorphose
	Marlinger Schuppe - Einheit mit mittel-hochgradiger variszischer und niedriggradiger alpiner Metamorphose
	Bernina-Decke - Einheit mit mittel-hochgradiger variszischer und niedriggradiger alpiner Metamorphose
	Ötztal-Einheit - Einheit mit mittel-hochgradiger variszischer und mittel-niedriggradiger alpiner Metamorphose
	Ötztal-Einheit (Umbrail-Chavalatsch Schuppe) - Einheit mit mittel-hochgr. Varisz. und mittel-niedrig. Alp. Met.
21	medium-high alpine metamorphosis
	Taufers-Einheit - Einheit mit mittel-hochgradiger variszischer und mittel-hochgradiger Metamorphose
	Schneebergerzug - Einheit mit mittel-hochgradiger alpiner Metamorphose
	Texel-Einheit - Einheit mit hochgradiger variszischer und alpiner Metamorphose
	Steinacher Decke - Einheit mit niedriggradiger variszischer und alpiner Metamorphose

3 TABLE: Predictor reclassification (geology – alkalinity of soil)

Table 3: Binary reclassification of geological information.
(G: Grassland. GA: Grassland and Arable land)

Code	
25	Acidic soil
	Bündnerschiefer mit Ophiolithen
	Zentralgneis
	Altes Dach und Untere Schieferh++lle
	Quartäre Ablagerungen
	Permische Plutone: Ifinger-Brixen
	Tektonische Einheiten Passaier - Einheit mit mittel-hochgradiger variszischer und unbedeut. alp. Metam.
	Tertiäre Plutone: tonalitische Lamelle
	Steinacher Decke - Einheit mit niedriggradiger variszischer und alpiner Metamorphose
	Ötztal-Einheit - Einheit mit mittel-hochgradiger variszischer und mittel-niedriggradiger alpiner Metamorph.
	Tertiäre Plutone
	Permische Plutone: Brixen
	Niedriggradig metamorphes varisziches Grundgebirge: Brixen Quartzphyllit
	Antholz-Einheit - Einheit mit mittel-hochgradiger variszischer und unbedeutender alpiner Metamorphose
	Thurntaler Einheit - Einheit mit mittelgradiger variszischer und unbedeutender alpiner Metamorphose
	Taufers-Einheit - Einheit mit mittel-hochgradiger variszischer und mittel-hochgradiger Metamorphose
	Tertiäe Plutone: Rieserferner
	Sesvenna-Einheit - Einheit mit mittel-hochgradiger variszischer und unbedeutender alpiner Metamorphose
	Texel-Einheit - Einheit mit hochgradiger variszischer und alpiner Metamorphose
	Permische Plutone: Ifinger
	Marlinger Schuppe - Einheit mit mittel-hochgradiger variszischer und niedriggradiger alpiner Metamorphose
	Campo-Decke - Einheit mit mittel-hochgradiger variszischer und niedriggradiger alpiner Metamorphose
	Tonale-Decke - Einheit mit hochgradiger variszischer und unbedeutender alpiner Metamorphose
	Kretazische und paleogene Sedimentabfolge
	Triassische Vulkanite und Vulkanoklastite
	Permische Vulkanite: Etschtaler Vulkanit-Gruppe
	Zebrù-Schuppe - Einheit mit niedriggradiger variszischer und unbedeutender alpiner Metamorphose
	Permische Plutone: Marteller Granit
	Ötztal-Einheit (Umbrail-Chavalatsch Schuppe) - Einheit mit mittel-hochgr. Varisz. & mittel-niedrig. alp.Metam.
	Vinschgauer Scherzone
	Permische Plutone: Kreuzberg
26	Basic soil
	Schneebergerzug - Einheit mit mittel-hochgradiger alpiner Metamorphose
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Brenner Mesozoikum
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Pennes Schuppe
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Drauzug
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Kalkstein Schuppe
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: S-charl
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Jaggl
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Ortler Decke
	Permo-jurassische Sedimentabfolge

4 TABLE: Predictor reclassification (geology – alkalinity of soil)

Table 4: Binary reclassification of geological information.
(G: Grassland. GA: Grassland and Arable land)

Code	
27	Acidic soil [ph < 5]
	Altes Dach und Untere Schieferhülle
	Niedriggradig metamorphes varisziches Grundgebirge: Brixen Quartzphyllit
	Permische Plutone: Brixen
	Permische Plutone: Marteller Granit
	Taufers-Einheit - Einheit mit mittel-hochgradiger variszischer und mittel-hochgradiger Metamorphose
	Tertiäre Plutone
	Tertiäre Plutone: Rieserferner
	Tertiäre Plutone: tonalitische Lamelle
28	Acidic-basic soil [5 < pH < 6]
	Antholz-Einheit - Einheit mit mittel-hochgradiger variszischer und unbedeutender alpiner Metamorphose
	Campo-Decke - Einheit mit mittel-hochgradiger variszischer und niedriggradiger alpiner Metamorphose
	Kretazische und paleogene Sedimentabfolge
	Marlinger Schuppe - Einheit mit mittel-hochgradiger variszischer und niedriggradiger alpiner Metamorphose
	Ötztal-Einheit - Einheit mit mittel-hochgradiger variszischer und mittel-niedriggradiger alpiner Metamorphose
	Ötztal-Einheit (Umbrail-Chavalatsch Schuppe) - Einheit mit mittel-hochgr. Varisz. und mittel-niedrig. Alp.Metam.
	Permische Plutone: Ifinger
	Permische Plutone: Ifinger-Brixen
	Permische Plutone: Klausen
	Permische Plutone: Kreuzberg
	Permische Vulkanite: Etschtaler Vulkanit-Gruppe
	Permo-jurassische Sedimentabfolge
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Brenner Mesozoikum
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Drauzug
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Jaggl
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Pennes Schuppe
	Quartäre Ablagerungen
	Schneebergerzug - Einheit mit mittel-hochgradiger alpiner Metamorphose
	Steinacher Decke - Einheit mit niedriggradiger variszischer und alpiner Metamorphose
	Tektonische Einheiten Passaier - Einheit mit mittel-hochgradiger variszischer und unbedeutender alp. Metam.
	Texel-Einheit - Einheit mit hochgradiger variszischer und alpiner Metamorphose
	Thurntaler Einheit - Einheit mit mittelgradiger variszischer und unbedeutender alpiner Metamorphose
	Tonale-Decke - Einheit mit hochgradiger variszischer und unbedeutender alpiner Metamorphose
	Triassische Vulkanite und Vulkanoklastite
	Vinschgauer Scherzone
	Zebrù-Schuppe - Einheit mit niedriggradiger variszischer und unbedeutender alpiner Metamorphose
29	Acidic soil [pH < 6]
	Bündnerschiefer mit Ophiolithen
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Kalkstein Schuppe
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: Ortler Decke
	Permomesozoische Sedimentabfolge verschiedene Deckeneinheiten: S-charl
	Sesvenna-Einheit - Einheit mit mittel-hochgradiger variszischer und unbedeutender alpiner Metamorphose
	Zentralgneis

5 TABLE: Descriptive statistics of the source data (original and transformed)

Table 5: Descriptive statistics of the original dataset (original and transformed data) for SOM, pH, P, K, N, C, N and C:N. The transformations that produced the best models have a shading (green: grassland. grey: grassland and arable land). (Min.: minimum value. 1st Qu.: first quartile value. 3rd Qu.: third quartile value. Max.: maximum value. Skew.: Skewness. Kurt.: Kurtosis. W.: Shapiro-Wilk test statistic)

	Transformation	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Skew.	Kurt.	W
SOM (719 points)	raw	2.10	6.20	8.10	9.34	11.10	52.94	2.73	14.98	0.81
	log	0.74	1.82	2.09	2.13	2.41	3.97	0.23	0.55	0.99
	sqrt	1.45	2.49	2.85	2.98	3.33	7.28	1.21	3.53	0.94
	square	4.41	38.44	65.61	111.09	123.21	2802.92	8.61	108.82	0.43
	inverse	0.02	0.09	0.12	0.13	0.16	0.48	1.46	4.06	0.91
pH (719 points)	raw	3.20	5.40	5.90	5.89	6.40	7.80	-0.42	0.03	0.98
	log	1.16	1.69	1.77	1.76	1.86	2.05	-0.86	0.89	0.95
	sqrt	1.79	2.32	2.43	2.42	2.53	2.79	-0.63	0.38	0.97
	square	10.21	29.16	34.81	35.35	40.96	60.84	-0.03	-0.30	0.99
	inverse	0.13	0.16	0.17	0.17	0.19	0.31	1.37	2.61	0.90
P (641) points)	raw	0.50	5.00	11.00	15.87	21.00	246.00	4.96	51.84	0.67
	log	-0.69	1.61	2.40	2.33	3.04	5.51	-0.26	-0.15	0.99
	sqrt	0.71	2.24	3.32	3.59	4.58	15.68	1.23	3.72	0.93
	square	0.25	25.00	121.00	549.44	441.00	60516.00	20.07	455.80	0.14
	inverse	0.00	0.05	0.09	0.16	0.20	2.00	4.62	30.38	0.56
K (641 points)	raw	1.50	12.00	21.00	28.91	35.00	272.00	3.83	21.94	0.66
	log	0.41	2.48	3.04	3.01	3.56	5.61	-0.04	0.28	1.00
	sqrt	1.22	3.46	4.58	4.92	5.92	16.49	1.52	4.21	0.90
	square	2.25	144.00	441.00	1736.90	1225.00	73984.00	8.76	87.93	0.25
	inverse	0.00	0.03	0.05	0.07	0.08	0.67	3.88	20.80	0.63
N (160 points)	raw	0.15	0.30	0.41	0.45	0.56	1.25	1.25	1.77	0.91
	log	-1.90	-1.20	-0.89	-0.88	-0.58	0.22	0.17	-0.42	0.99
	sqrt	0.39	0.55	0.64	0.66	0.75	1.12	0.69	0.25	0.96
	square	0.02	0.09	0.17	0.25	0.31	1.56	2.50	7.67	0.73
	inverse	0.80	1.79	2.44	2.64	3.33	6.67	0.80	0.43	0.95
C (160 points)	raw	1.35	3.36	4.60	5.34	6.30	33.10	3.71	23.40	0.71
	log	0.30	1.21	1.53	1.54	1.84	3.50	0.49	0.64	0.98
	sqrt	1.16	1.83	2.15	2.23	2.51	5.75	1.67	5.85	0.90
	square	1.82	11.26	21.18	40.24	39.67	1095.61	9.24	101.73	0.30
	inverse	0.03	0.16	0.22	0.24	0.30	0.74	0.98	1.47	0.95
C:N (160 points)	raw	8.41	9.72	11.00	11.52	12.90	26.48	1.84	6.26	0.86
	log	2.13	2.27	2.40	2.42	2.56	3.28	0.90	1.20	0.94
	sqrt	2.90	3.12	3.32	3.38	3.59	5.15	1.31	3.05	0.90
	square	70.78	94.48	121.00	139.24	166.40	701.36	3.44	20.15	0.72
	inverse	0.04	0.08	0.09	0.09	0.10	0.12	-0.34	-0.44	0.97

6 SPATIAL INTERPOLATION AND PREDICTION METHOD

The program starts by checking if the source chemical soil data is normally distributed, because geostatistical methods work best when this condition is given and when the mean / variance of data do not vary significantly. After computing data transformations (logarithm, square, root, inverse) of each dependent variable, they are checked for the listed requirements, because significant deviations from normality can affect the kriging estimators (Lark, 2000). Only transformations with a Shapiro statistic greater than 90% are considered for the analysis (**Table 6**).

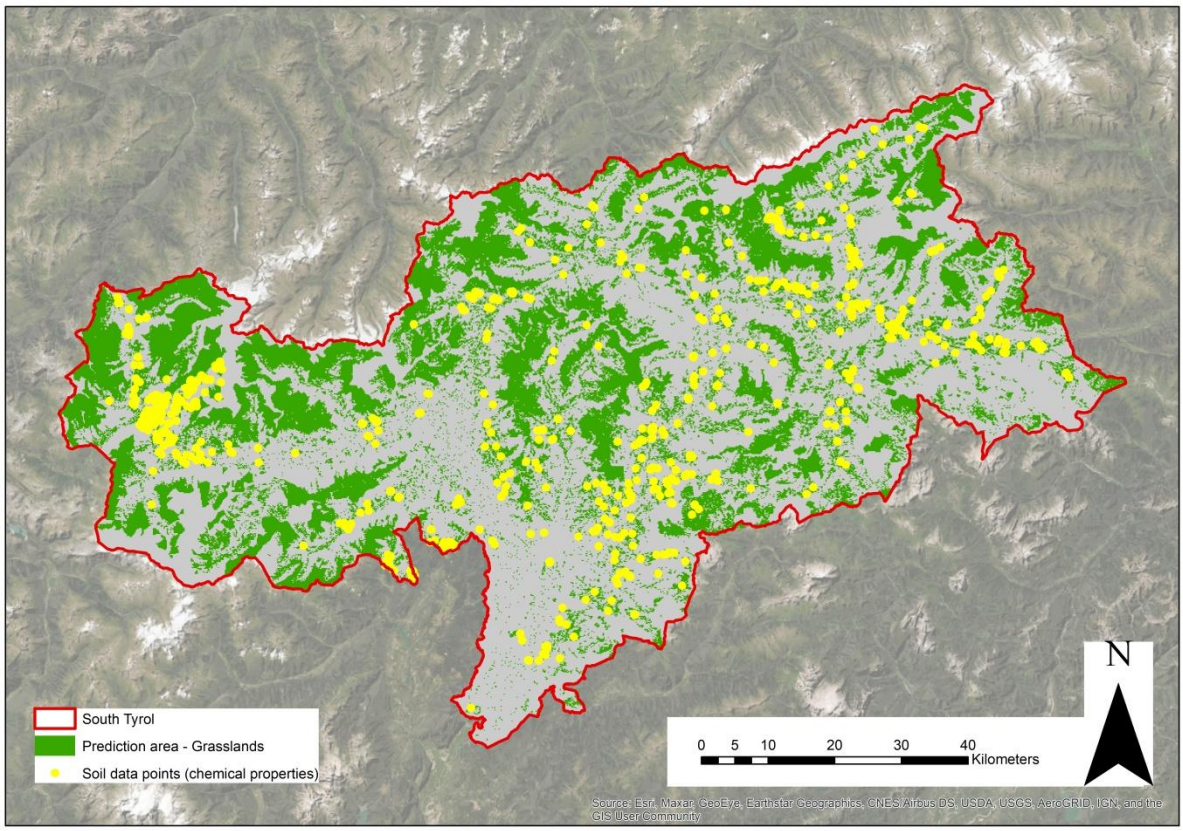


Figure 1: Yellow points with known coordinates and chemical soil properties (e.g. SOM, pH, P, K, C, N) used for prediction in the green area of interest.

Table 6: Check normality and stationary of different source data transformations.

	transformation	skewness	kurtosis	shapiro_p.value	shapiro_statistic	Normal_distributed_data
1	raw	2.7261235	14.9771617	5.400101e-28	0.8107730	NO
2	log	0.2284843	0.5531135	1.562020e-02	0.9947544	YES
3	sqrt	1.2059857	3.5348276	2.629747e-16	0.9399253	YES
4	quad	8.6059409	108.8193115	1.611069e-42	0.4256776	NO
5	inverse	1.4615519	4.0646882	6.796240e-20	0.9113354	YES

The program is based on two DSM methods.

Random Forest (RF) is the regression method used (Liaw and Wiener, 2002). It is a powerful ensemble-learning method proposed by Breiman (Breiman, 2001), in which more decision trees are run in parallel and the mean prediction for the continuous dependent variables will be outputted (Chen *et al.*, 2017). The regression analysis outputs two plots for each variable and data transformation (**Figure 2**). One plot is the goodness of fit, which describes how similar the observed and the predicted values are. The second plot is the residual variogram that gives an assessment of the variance of each variable. This is necessary to compute a spatial interpolation (Genova, 2017).

The crosses in **Figure 2** correspond to the "experimental variogram" on the residuals, which explains the semivariance (gamma) computing the distance between each couple of points' values of the dataset and regrouping the distances in bins. The green curve corresponds to the "theoretical variogram" that describes how semivariance changes. It is fitted on the experimental variogram in order to find the model parameters for the variable to be interpolated.

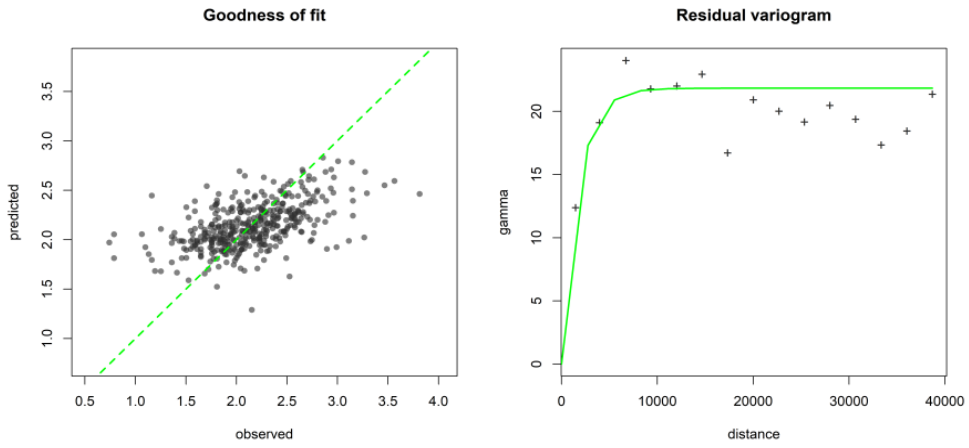


Figure 2: Output plots of the Random Forest regression analysis.
The goodness of fit describes the accuracy of the regression and the residual variogram gives an assessment of the variance of each variable. The crosses are the "experimental variogram" and the green curve is the "theoretical variogram".

The second method is required for the prediction part. The Ordinary Kriging (OK) (Krige, 1951) interpolation method is used to create the final map rasters of the chemical soil properties (Li and Heap, 2014). It is a widely adopted method that provides a solution to the problem of estimation based on a continuous model of stochastic spatial variation. The variation of the chemical properties is determined through the variogram model. The theoretical variogram model from the first part is connected to the kriging estimator to interpolate the value for each cell of a raster mask (Genova, 2017). So this spatial model gives the possibility to estimate a value at a point of an unsampled region for which a variogram is known.

The estimation is done by the following linear combination of values at known locations

$$unsampled\ region\ prediction = \sum_{i=1}^{n} weight_{\ i} * value_{\ i}$$

where *i* is the running variable and *n* is the number of known chemical soil data points (**Figure 3**). Every location has a weight, an importance, which is given by the theoretical variogram. The known locations that are further away will become less weight than the closer ones.

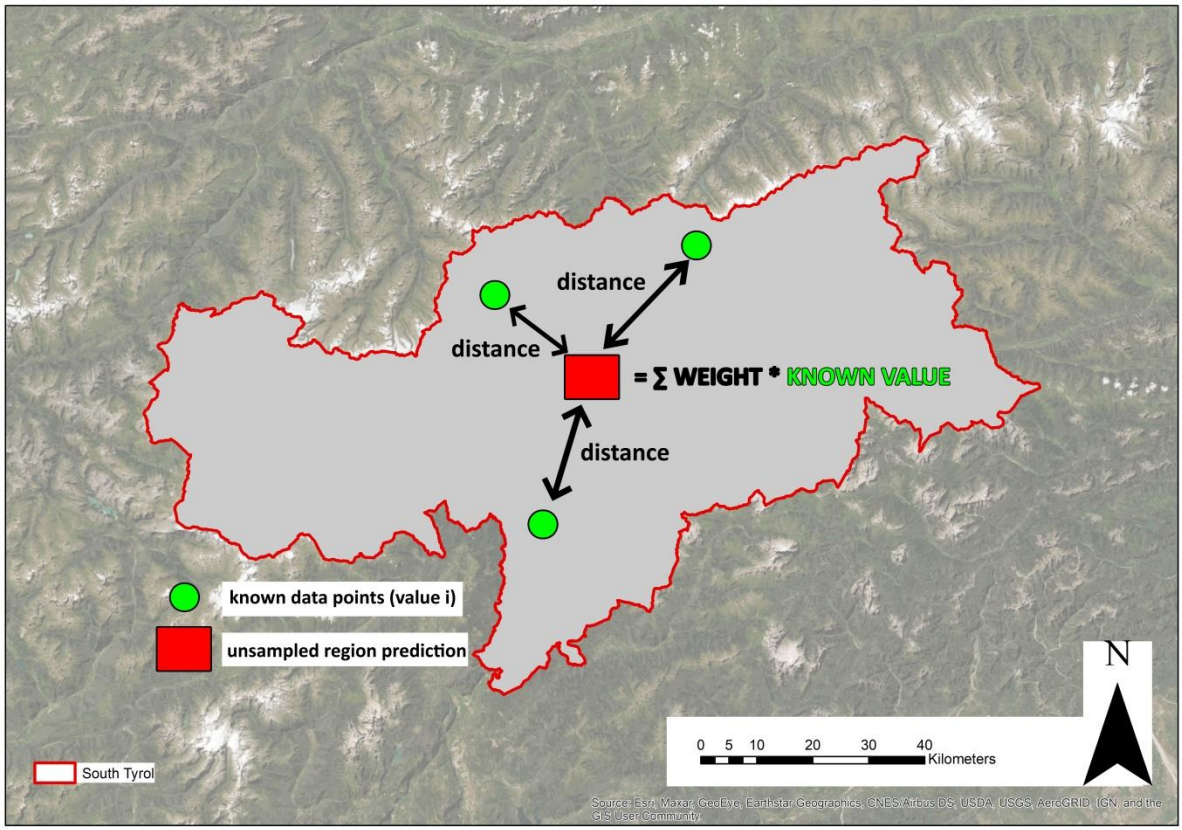


Figure 3: Simple graphic illustration of how Ordinary Kriging (OK) interpolation works.

In order to allow the model validation the source data (chemical soil properties) is split into two parts, a training dataset with 80% of the source data and a validation dataset with the remaining randomly chosen 20% of the source data. The training dataset is used to fit the models (RF and OK). The validation dataset is not used for the analysis, but it is overlayed with the predicted values from the OK interpolation. The R-Squared (R^2) and the Root Mean Square Error (RMSE) are calculated to help choosing the best model (**Table 7**).

Table 7: Possible overview table with the most important output values of each program run.

No.Run	Transformation	Cutoff	Predictors	Mean_Of_Squared_Residuals	%VarExplained	RMSE	NRMSE	NRMSEMAXMIN	R_SQUARED
122	log	30000	1/3/4/5/6/7/9/10/22/23/27/28/29	0.1467528	31.94	2.751960	71.9	13.6	0.4927847
107	log	32000	1/2/3/4/5/6/7/8/9/10/11/12/13/14/15/16/17/18/19/20/21/22/23	0.1398088	35.16	2.775192	72.5	13.7	0.4856251
126	log	29000	1/3/4/5/6/7/9/10/22/23	0.1614357	25.13	2.784648	72.8	13.8	0.4682315
109	log	30000	1/2/3/4/5/6/7/8/9/10/22/23/27/28/29	0.1483269	31.21	2.796911	73.1	13.8	0.4786584
116	log	29000	1/2/3/4/5/10/22/23	0.1700704	21.13	2.815992	73.6	13.9	0.4540478

The structure of the main program is as described here (**cf. p. 17**).

7 PROGRAM: INSTRUCTION MANUAL

Before starting the program the input data needs to be checked. Based on the source data, a few changes in the R-Script are needed. There are also some tips to consider. They might be useful.

a Input data

The input data (1-5) has to be stored in the following folders (**Figure 4**):

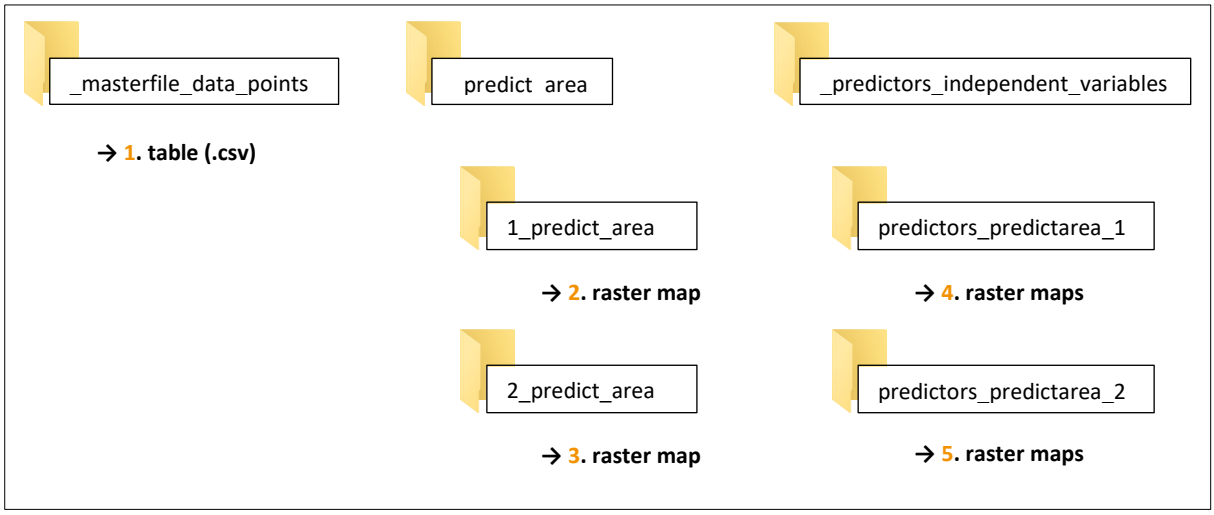


Figure 4: Overview of the input data folders. An explanation of the single files (1-5) follows.

1. Masterfile with the soil data points (**dependent variables**)

   Microsoft Excel or SPSS can be used to organize and prepare the source data, so to become at the end a table (.csv) containing coordinates with information about chemical soil properties. These coordinates correspond to the yellow points in **Figure 1**. The table has to correspond to the following layout (**Table 8**). The chemical values must start in the ninth column and the last two columns must contain the geographic coordinates.

Table 8: Layout of the table containing the source data of the chemical values (dependent variables).

FID	Source	Code	GIS_ID	Sample_Type	Habitat_Type	Town	Year	SOM	pH	P_mg100g	K_mg100g	C	N	CN	x	y
0	1	61001967	61001967	Soil	Grasslands	LAAS	2006	7.60	7.30	23.0	42.0	-9999.00	-9999.00	-9999.00	628836	5165114
1	1	61001998	61001998	Soil	Grasslands	ANTHOLZ	2006	2.20	6.40	7.0	26.0	-9999.00	-9999.00	-9999.00	737420	5195407
2	1	61005017	61005017	Soil	Grasslands	TOBLACH	2006	7.60	5.90	27.0	53.0	-9999.00	-9999.00	-9999.00	747596	5180101
3	1	61005302	61005302	Soil	Grasslands	ALDEIN	2006	6.30	5.50	17.0	50.0	-9999.00	-9999.00	-9999.00	678866	5135892
4	1	61005306	61005306	Soil	Grasslands	ALDEIN	2006	7.20	5.40	5.0	24.0	-9999.00	-9999.00	-9999.00	679167	5135742
5	1	61005307	61005307	Soil	Grasslands	ALDEIN	2006	5.00	5.20	5.0	12.0	-9999.00	-9999.00	-9999.00	679172	5136002

2. First area of interest

The area of interest is where the prediction of the soil chemical properties has to be done. It is possible to enter two prediction/spatial interpolation areas simultaneously, because the program can treat them separately. This means that it provides an output for each dependent variable and for each area of interest. It has to be a raster file with a specific resolution that is determined as follows. The raster resolution is related to the density of samples per area of interest. The denser the measurement points, the larger the scale of mapping, the smaller the resolution. The formula used to determine it, is the one described by (Hengl, 2006)

$$grid\ resolution = 0.0791 * \sqrt{\frac{A}{N}}$$

where *A* is the surface of the area of interest in m^2 and *N* is the total number of the points with chemical measurements. Programs for preparing these raster maps and the following ones, always with the same resolution, can be ArcGIS (closed source) or QGIS (open source).

3. Second area of interest

The second raster can be saved here. If there is no second area of interest, the folder "2_predict_area" must be deleted.

4. Predictors (**independent variables**) of the first area of interest

The raster maps of the predictors have to be saved here. It is important to distinguish between categorical and numerical data. For numerical data (e.g. elevation, slope, etc.), there must be one raster map per predictor. In the case of categorical data (e.g. geology, land use, etc.), the information with similar characteristics are grouped. For this type of data a binary raster for each category of each predictor is required. This means that only 0 and 1 (not two other values) can be included in the single rasters, it applies or it doesn't. An example is the following. If the predictor "geology" is taken into account, there can be more categories, so for example acidic soil types and basic soil types. For each category of the categorical predictors a binary raster is needed. In this case for example one for acidic soil types and another for basic soil types.

5. Predictors (**independent variables**) of the second area of interest

It can be that different predictors are taken into account for the second area of interest. These predictors can be saved here, following the same explanations as above. If there is no second area of interest the folder "predictors_predictarea_2" must be deleted.

The naming of the predictors' rasters (**cf.** 4 and 5) must be as follows:

- ⇨ **P_1_***Name of the first predictor*_**resolution**
- ⇨ **P_2_** *Name of the second predictor* **_resolution**
- ⇨ **…**

The parts in bold must be exactly the same as here. Instead of "resolution", the calculated grid resolution (**cf. p. 11**) must be entered here, for example "150". As you can see the predictors have to be sequentially numbered. These numbers are very important, because in the result tables they will appear instead of the predictors' names. The program can consider more predictors combinations that are determined with the predictors' numbers (**cf. p. 14**).

b Program changes in R needed

In the line 62 the calculated grid resolution (**cf. p. 11**) must be entered. Since the program considers more combinations of predictors, the code in line 64 offers the possibility to determine the mode of combination, "personalized" or "automatic" (**Figure 5**). If the number of the predictors is less than or equal to five, the automatic mode can be choosen. Otherwise the program takes too much time to process the data (**cf. p. 16**). A detailed explanation of the combinations' settings follows (**cf. p. 14**).

```
62  res <- 150
63  resolution <- paste0("",res,"")
64  predictors_combination <- "personalized"  # or "automatic"
```

Figure 5: Program changes in lines 62 and 64 ("res" is an abbreviation of resolution).

The next coming changes are for the labels of the dependent variables in graphs and plots. In line 298 ff. there are several if-else statements whose number depends on the number of the dependent variables. The dependent variables are those included in the masterfile (**cf.** 1). The general syntax of an if statement is the following (**Figure 6**):

```
if (text_expression) {
      statement
} else if (text_expression) {
      statement
} else {
      statement
}
```

Figure 6: General syntax of an if statement.

The different "text_expressions" and "statements" that have to be modified are

Line 298 ff.

if – text_expression 1: variables[j] == "masterfile's column name of the first variable"

statement 1: einheit = "unit of the first variable"

else if - text_expression 2: variables[j] == "masterfile's column name of the second variable"

statement 2: einheit = "unit of the second variable"

…

else if - text_expression n: variables[j] == "masterfile's column name of the n-th variable"

statement n: einheit = "unit of the n-th variable"

else – text_expression n+1: /

where *n* is the number of dependent variables. The orange parts have to be substituted by the right name of each variable that corresponds to the masterfile's column names (**Table 8**) and the right unit of the variables. At the end the number of "else if statements" must be *n-1*. The other parts remain unchanged. Here an example with the same dependent variables from **Table 8** (**Figure 7**).

```
297     # variable description for graphics
298     if (variables[j]=="SOM"){
299       varr = variables[j]
300       einheit = "%"
301       varbez = paste0(varr," ","[",einheit,"]")
302       } else if (variables[j]=="pH"){
303         varr = variables[j]
304         einheit = " "
305         varbez = varr
306       } else if (variables[j]=="P_mg100g"){
307         varr = "P"
308         einheit = "mg/100g"
309         varbez = paste0(varr," ","[",einheit,"]")
310       } else if (variables[j]=="K_mg100g"){
311         varr = "K"
312         einheit = "mg/100g"
313         varbez = paste0(varr," ","[",einheit,"]")
314       } else if (variables[j]=="C"){
315         varr = variables[j]
316         einheit = "%"
317         varbez = paste0(varr," ","[",einheit,"]")
318       } else if (variables[j]=="N"){
319         varr = variables[j]
320         einheit = "%"
321         varbez = paste0(varr," ","[",einheit,"]")
322       } else if (variables[j]=="CN") {
323         varr = variables[j]
324         einheit = " "
325         varbez = variables[j]
326       } else {
327         varbez = variables[j]
328       }
```

Figure 7: Example of how to lable the dependent variables in graph and plots (changes in line 298 ff.).

In the lines 1095 ff. and 1181 ff. the denomination of the variables for the correlation coefficients' plots/tables has to be changed. The naming must be as follows:

```
Line 1095 ff. (for the first area of interest)
        names(corr_parameters1)[names(corr_parameters1)
                == " P_1_Name of the first predictor_resolution (cf. p. 14) "] <- "P1_your name"
        names(corr_parameters1)[names(corr_parameters1)
                == " P_2_Name of the second predictor_resolution (cf. p. 14) "] <- "P2_your name"
        …

Line 1181 ff. (for the second area of interest)
        names(corr_parameters2)[names(corr_parameters2)
                == " P_1_Name of the first predictor_resolution (cf. p. 14) "] <- "P1_your name"
        names(corr_parameters2)[names(corr_parameters2)
                == " P_2_Name of the second predictor_resolution (cf. p. 14) "] <- "P2_your name"
        …
```

The orange parts have to be substituted by the right raster files' name of each predictor (**cf. 4** and **5**). Instead of the blue parts, any name for each predictor, which will then appear in the correlation coefficients' plots/tables, can be choosen. If there is no second area of interest, the changes in line 1181 ff. can be ignored.

The pre-last changes, in line 1274 ff. and line 1322 ff., refer to the different predictors' combinations executed by the program. This part should only be considered if the mode in line 64 is set to "personalized" (**Figure 5**). Before defining the combinations it can be useful to check autocorrelation between numeric variables and between the single predictors (**cf. X**). The amendments are

```
Line 1274 ff. (1st area of interest) and Line 1322 ff. (2nd area of interest)
nrofcombinations <- "number of combinations inputted in the following lines"
predictors_combi_1 <- c(1,2,3,…)
predictors_combi_2 <- c(1,2,3,…)
…
predictors_combi_n <- c(1,2,3,…)
```

where *n* must correspond to the number of combinations (`nrofcombinations`) in the line 1274. The predictors are sequentially numbered (**cf. p. 12**). These numbers are inputted instead of the orange parts to decide what predictors must be taken into account for the different combinations. Remember that the more combinations, the more time the program needs for processing (**cf. p. 16**).

The last changes, in lines 1699 ff., 1792 ff., 1885 ff., 2170 ff., 2263 ff., 2356 ff., 2641 ff., 2734 ff., 2827 ff., 3159 ff., 3252 ff., 3345 ff., 3631 ff., 3724 ff., 3817 ff., 4146 ff., 4239 ff., 4332 ff., are about the plots and tables that describe the influence of the individual predictors on the dependent variables, which results from the Random Forest regression. These changes

Line 1699 ff. / 2170 ff. / 2641 ff. / 3159 ff. / 3631 ff. / 4146 ff.

```
if (i==1) {   # [for the first area of interest]
    names(incmse)[names(incmse) ==
      " P_1_Name of the first predictor_resolution (cf. p. 14) "] <- " P1_your name  "
    names(incmse)[names(incmse) ==
      " P_2_Name of the second predictor_resolution (cf. p. 14) "] <- "P2_your name"
    …

} else if (i==2) {   # [for the second area of interest]
    names(incmse)[names(incmse) ==
      " P_1_Name of the first predictor_resolution (cf. p. 14) "] <- " P1_your name  "
    names(incmse)[names(incmse) ==
      " P_2_Name of the second predictor_resolution (cf. p. 14) "] <- "P2_your name"
    …

} else {}
```

Line 1792 ff. / 2263 ff. / 2734 ff. / 3252 ff. / 3724 ff. / 4239 ff.

```
if (i==1) {   # [for the first area of interest]
    names(IncNodePurity)[names(IncNodePurity) ==
      " P_1_Name of the first predictor_resolution (cf. p. 14) "] <- " P1_your name  "
    names(IncNodePurity)[names(IncNodePurity) ==
      " P_2_Name of the second predictor_resolution (cf. p. 14) "] <- "P2_your name"
    …

} else if (i==2) {   # [for the second area of interest]
    names(IncNodePurity)[names(IncNodePurity) ==
      " P_1_Name of the first predictor_resolution (cf. p. 14) "] <- " P1_your name  "
    names(IncNodePurity)[names(IncNodePurity) ==
      " P_2_Name of the second predictor_resolution (cf. p. 14) "] <- "P2_your name"
    …

} else {}
```

Line 1885 ff. / 2356 ff. / 2827 ff. / 3345 ff. / 3817 ff. / 4332 ff.

```
if (i==1) {   # [for the first area of interest]
    names(importancesd)[names(importancesd) ==
      " P_1_Name of the first predictor_resolution (cf. p. 14) "] <- " P1_your name "
    names(importancesd)[names(importancesd) ==
      " P_2_Name of the second predictor_resolution (cf. p. 14) "] <- "P2_your name"
    …

} else if (i==2) {   # [for the second area of interest]
    names(importancesd)[names(importancesd) ==
      " P_1_Name of the first predictor_resolution (cf. p. 14) "] <- " P1_your name "
    names(importancesd)[names(importancesd) ==
      " P_2_Name of the second predictor_resolution (cf. p. 14) "] <- "P2_your name"
    …

} else {}
```

The orange parts have to be substituted by the right raster files' name of each predictor (**cf. 4** and **5**). Instead of the blue parts, any name for each predictor, which will then appear in the plots and tables, can be choosen. If there is no second area of interest, the else if statement can be ignored.

c Tips

Program execution time

It is important to know about the time that the program takes to process the data. The execution time depends from:

- the number of dependent variables to predict (**cf. 1**),
- the number of soil data points of each variable (**cf. 1**),
- the number of predictors' combinations that have been processed (**cf. p. 14**),
- the number of the areas of interest s (one or two) and
- the area size of the areas of interest (**cf. 2** and **3**).

Since the program checks several combinations of the predictors, the mode of defining them must be selected at the beginning of the program, in line 64 (**Figure 5**). The program executes each predictors' combination three times for every data transformation that provides a Shapiro statistic greater than 90% (**Table 6**). The structure of the program, based on nested for loops, is

```
1) Exploratory data analysis
2) Main code – nested for loops
for (j in 1:lenvariables) {}            # (dependent variables)
   for (i in 1:nrpredictareas) {}         # (areas of interest)
      for (k in 1:lange) {}                 # (data transformations)
         for (s in 1:diffcutoff) {}           # (runs with different cutoffs)
            for (l in 1:nrofcombinations) {}    # (predictors' combinations)
```

where "**lenvariables**" is the number of dependent variables, "**nrpredictareas**" is the number of areas of interest, "**lange**" is the number of data transformations with a Shapiro statistic greater than 90%, "**diffcutoff**" is the number of runs done with different cutoff settings for the Random Forest regression and "**nrofcombinations**" is the number of the predictors' combinations. The variable "diffcutoff" is three, because the Random Forest regression is done one time with the minimum cutoff, a second time with the average cutoff and a third time with the maximum possible cutoff.

The time that the program takes to process the input data can be estimated with the second R script "*test_execution_time*" that is also contained in the folder "*DSM*". In line 10 the number of the predictors' combinations (*no.combi*) that have to be taken into account has to be specified. It must be executed before the main R script "*dsm*". The file "*estimated_execution_time.txt*" which is outputted contains all details. You cannot exactly rely on these times, because there can be deviations. It gives you only an idea of how long the program will take to process the input data.

Geographic coordinate system (GCS)

It is important to have all input data rasters (**cf.** 2, 3, 4 and 5) with the same reference framework that defines the locations of features.

File names and file paths' options

A helpful advice is to keep file names short, but meaningful. Abbreviations often help create such names that should be easy to read and understand. Do not use umlauts and follow the guidelines naming the predictors' rasters (**cf. p. 12**). The file paths cannot be too long, therefore it is a good idea to store the folder "dsm" on the desktop. It is also important to check the amount of free space on the hard disk. The necessary storage space depends on the source data and cannot be determined in advance.

Define area of interest 1 and area of interest 2

From the beginning it should be clear which the first area of interest and which the second area of interest is. Sometimes the area of interests file names appear in the output plots and tables and sometimes it simply says "area of interest 1" or "area of interest 2". It is also very important to know what predictors are exactly taken into account for the individual areas and its sequential numbers. For this reason it might be helpful to make a list of the predictors of each area of interest containing the name and the sequential number of the single predictors. The results will be then easier to interpret and hopefully no wrong interpretations will happen.

Plot pane in Rstudio

The plot window in Rstudio must have a certain size (half screen size) so that no errors occur.

Avoid computer overload

When the program is running all available resources should be accessible for it, so that it can performed optimally. In order to allow this, several open applications should be stopped.

Split up the run of the program

The estimated execution time (**cf. p. 16**) of the program can be long. It may be necessary to split the process. In this case the computer can set into sleep mode. As soon as the computer is restarted, the program resumes its activity automatically.

d Output explanation

The program creates two output folders: "**output**" and "**output_overview**".

The folder "**output**" contains all outputs of all runs of the program. The number of the runs of the program (*no. runs*) for each area of interest can be calculated with the formula

$$no.runs = \sum_{j=1}^{lenvariables} x_j * nrofcombinations * 3$$

where an explanation of the single variables is given above (**cf. p. 17**). The variable "x_j" corresponds to the number of data transformations (logarithm, square, root, inverse) of the dependent variable *j* that provides a Shapiro statistic greater than 90% (**Table 6**). This number corresponds to the amount of folders contained in the following file paths: *"…/DSM/output/prediction area*/dependent variable j"*. If there are two areas of interest the result *no. runs* must be doubled.

If there are too few available soil data points it may also be that the Random Forest regression may not work. In this case no data is outputted by the program. As a consequence also the number of runs decreases and the formula above is no longer valid. The correct number of runs can be taken from the output based on the last variable processed by the program. It can be found in the file path *"…/DSM/output/prediction area 1 /last dependent variable processed by the program/data transformation that begins with the latest letter in the alphabet*" if there is one area of interest. If there are two, the file path to consider is the following: *"…/DSM/output/prediction area whose name begins with the latest letter in the alphabet/last dependent variable processed by the program/data transformation that begins with the latest letter in the alphabet*". The order of which variables are processed first is based on **Table 8**. The columns' number determines the order. So the variables that have a smaller number are processed first. In the given paths there are the folders of all program runs' outputs. They are named as follows: "*No.Run_n*", where *n* corresponds to the run number. The highest *n* in the path to consider is the right number of all runs of the program. The file "*estimated_execution_time.txt*" (**cf. p. 17**) also contains a number, but it is not 100% reliable. It is possible that the source data of some dependent variables is not good enough, therefore no prediction and output data can be generated and so there can be less program runs. In this case the source data must be better and a TXT file will appear in the output folders. It explains that the source data is not good enough to do a prediction analysis.

The most important outputs are stored in the folder "**output_overview**". **Figure 8** on the next page describes an overview of its content. The variable **n** corresponds to the number of dependent variables. In this folder there are only outputs of the runs that could provide models with a high Root Mean Square Error (**RMSE**) and a high R-squared (**R^2**). The **runs with the five lowest RMSE values and the five highest R^2 values are picked out**. It is up to you whether you give more weight to RMSE or R^2. Both are important measures of the goodness of the model. The RMSE describes the error of the model, so it is a direct measure of when a prediction is wrong. In addition it has the same unit of measure as the studied variable. This information is very important in practical terms, for example for farmers or soil scientists who want to use the prediction models outputted by the program. The R^2 tells how much of the natural variability of the studied variables is captured by the model. Therefore it is a more general measure and it is also often affected by outliers. However, there is no right or wrong between RMSE and R^2. For the purpose of soil maps perhaps it is better to prioritise measuring how much predictions are wrong than measuring how much variability is accounted. The most important output of the program to take into account is a multi-page PDF file (**cf. IX**).

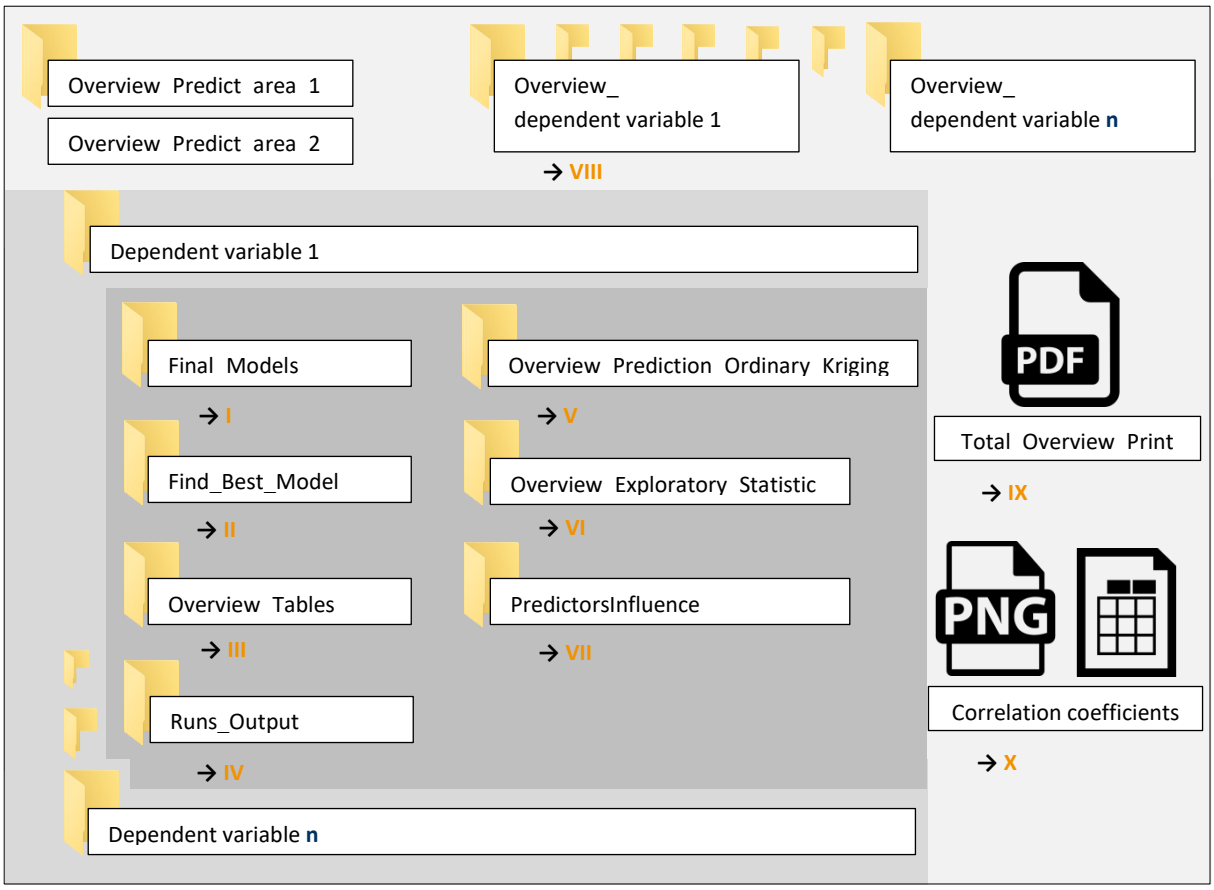


Figure 8: Overview of the content of the folder "output_overview". The explanation of the outputs I-X follows.

The output folders **for each variable and area of interest** are the following (**I-VII**):

I. Final models

The final raster maps of the predicted dependent variables in each area of interest are stored here. There is a PNG file (.png) of the map, an R Workspace file (.RData) that contains the most important values outputted by the program and a GeoTIFF file (.tif) of the map. The R Workspace file can be used to load important values (RMSE, R^2, etc.) back by R. The GeoTIFF file contains georeferencing information which is necessary to establish an exact spatial reference for the file. It can be loaded in geospatial processing programs like ArcGIS (closed source) or QGIS (open source) to finalise the layout. The resolution of the final maps corresponds to that calculated as described here (**cf. p. 11**).

II. How to find the best model?

As already explained the program's runs with the five lowest RMSE values and the five highest R^2 values are picked out (**cf. p. 19**). It is up to you whether to give more weight to RMSE or R^2. The number of the run which provides the model with the best prediction values can be read from the two PNG files contained in the folder "*Find_Best_Model*". The overview tables (**cf. III**) and the goodness of fit of the Ordinary Kriging model from **V** are included. The two PNG files can be also found in the outputs **VIII** and **IX**.

III. Overview tables

There are more overview tables. The files "*8_Output_Overview_Total_predictarea*.xlsx (csv)*" contain the most important values outputted by each run of the program (RMSE, R^2, etc.). The same table is contained in the files "*8_Output_Overview_Total_predictarea* _HighestR2.xlsx (csv)*" and "*8_Output_Overview_Total_predictarea*_HighestR2.xlsx (csv)*", but the values are ordered in ascending order of RMSE in the first case and in descending order of R^2 in the second case. Then the first five lines of each table are picked out and written in two new tables: "*8_Output_Overview_5HighestR2_predictarea*.xlsx (csv/png)*" and "*8_Output_Overview_5LowestRMSE_predictarea*.xlsx (csv/png)*". These tables are also contained in **II**.

IV. Runs' outputs

There are all outputs of the five runs which have the lowest RMSE and R^2 values. It is not necessary to look at them, because all other important outputs are based on the files contained in this folder.

V. Prediction part – Ordinary Kriging

The output plots of the prediction part of the five runs of the program which have the lowest RMSE and highest R^2 values are stored here. The most important plot is the one showing the goodness of fit / accuracy of the model (in the upper right corner) (**Figure 9**). This means how similar the observed soil chemical values and the predicted values are. The straight line describes the optimal situation, where the observed values are equal to the predicted values. The further away the points are from the line, the greater the uncertainty. These plots are contained in the outputs **VIII** and **IX**, because they can be useful to choose the best prediction model, together with the RMSE and R^2 values.

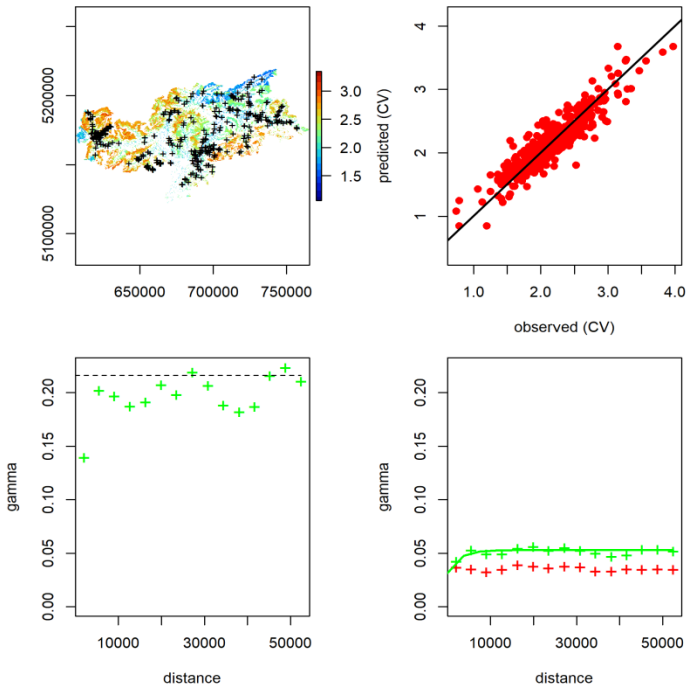


Figure 9: Example of the output plots of the prediction part (Ordinary Kriging).

VI. Exploratory / Descriptive data analysis

The files stored here give information about the descriptive/exploratory analyses done by the program (**Figure 10**). The file "*Descriptive_Statistic_*.png*" contains the minimum, maximum, median and quartile values of each transformation (raw, logarithm, square, root, inverse) of the sample data. The last column of the table says if the data is normal distributed or not. This information is given by the file "*Choose_Transformation_*.png*" where the sample data is tested for normality. Only transformations with a Shapiro statistic greater than 90% are considered for the analysis (**Table 6**), because geostatistical methods work best when the data is normally distributed and its mean and variance do not vary significantly. A Normal Q-Q plot ("*Overview_QQPlot.png*"), as visual check, can also be used to check if the assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation. It is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the normal distribution, the points should form a line that's roughly straight (Ford, 2015). The file "*Descriptive_Statistic_*.png*" is used to plot a boxplot, the sampling density and the histogram of each transformation of the sample data. The plots are contained in the files "*Overview_BoxplotDensity.png*" and "*Overview_Histogram.png*". They can tell about outliers and if the data is symmetrical.

transformation	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Normal_distributed_data
raw	2.10000000	6.20000000	8.1000000	9.3441915	11.1000000	52.9426516	NO
log	0.74193734	1.82454929	2.0918641	2.1289484	2.4069451	3.9692093	YES
sqrt	1.44913767	2.48997992	2.8460499	2.9751084	3.3316662	7.2761701	YES
quad	4.41000000	38.44000000	65.6100000	111.0924191	123.2100000	2802.9243627	NO
inverse	0.01888836	0.09009009	0.1234568	0.1312897	0.1612903	0.4761905	YES

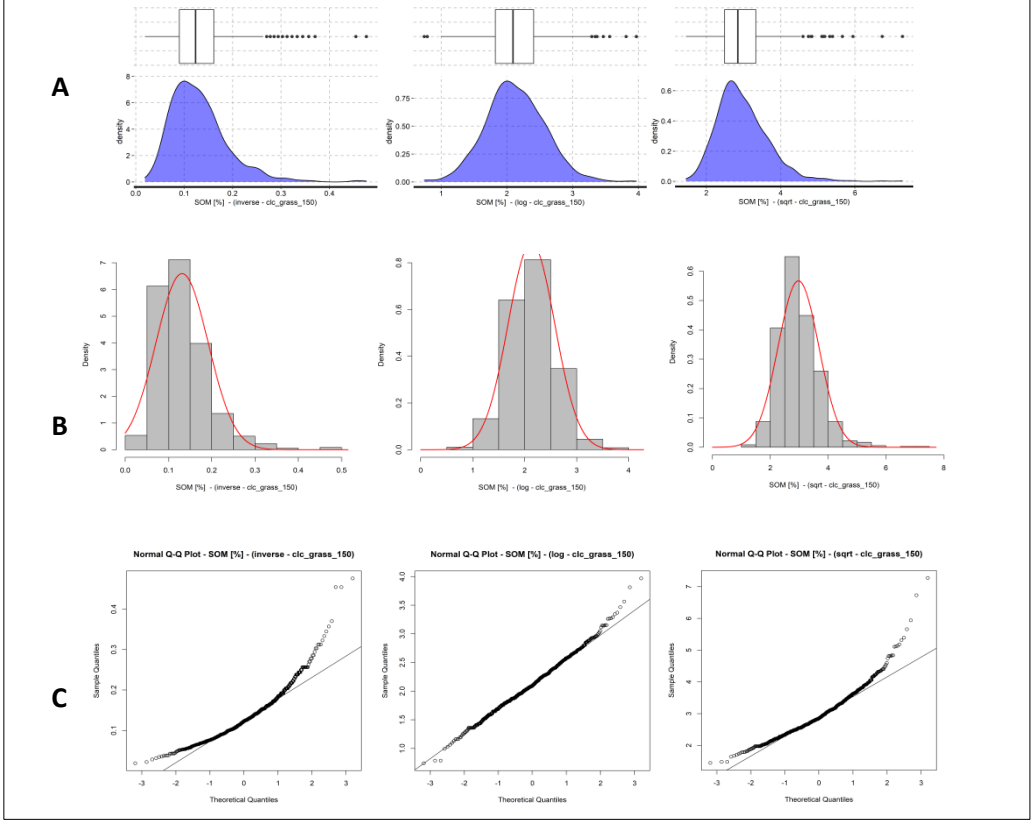
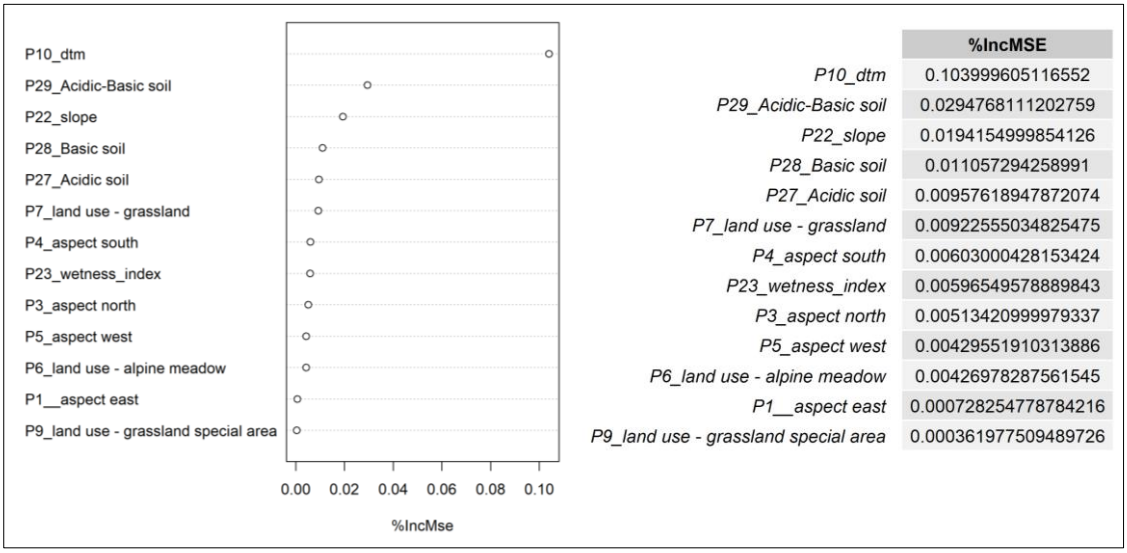


Figure 10: Example of the output of the exploratory / descriptive data analysis
(A: boxplot and sampling density, B: histogram, C: normal Q-Q plot)

VII. Influence of the individual predictors on the dependent variables

The output files describing the influence of the single predictors on the dependent variables are stored here. There are just the outputs of the five runs of the program which have the lowest RMSE and the highest R^2 values. The parameters that describe the influence are the %IncMse ("*4_IncMse_Plot (Table)_*.png*" and "*4_IncMse_ValuesTable_*.xlsx (.csv)*"), and IncNodePurity ("*4_IncNodePurity_Plot (Table)_*.png*" and "*4_ IncNodePurity _ValuesTable_*.xlsx (.csv)*"). They are all outputted by the first part of the program, which computes the regression based on the Random Forest method (**cf. p. 8**). The %IncMSE is the most robust and informative measure (**Figure 11**). It describes the increase of the mean square error (MSE) of the predictions as a result of the permutation of the variables, where the values are randomly shuffled. The higher the %IncMSE number, the more important is that predictor and its influence on the dependent variable taken into account. The same interpretation applies to the IncNodePurity. The importanceSD is also calculated by the program but does not need to be looked at more closely.



	%IncMSE
P10_dtm	0.103999605116552
P29_Acidic-Basic soil	0.0294768111202759
P22_slope	0.0194154999854126
P28_Basic soil	0.011057294258991
P27_Acidic soil	0.00957618947872074
P7_land use - grassland	0.00922555034825475
P4_aspect south	0.00603000428153424
P23_wetness_index	0.00596549578889843
P3_aspect north	0.00513420999979337
P5_aspect west	0.00429551910313886
P6_land use - alpine meadow	0.00426978287561545
P1__aspect east	0.000728254778784216
P9_land use - grassland special area	0.000361977509489726

Figure 11: Example of outputs of the Random Forest regression describing the influence of the single predictors on a specific dependent variable (%IncMSE value).

VIII. Find the best model that predicts a specific dependent variable

Here you find the most important files that help you to choose the best prediction model of each dependent variable. Only the PDF files are important. They contain the outputs **II**, **V** and the descriptive/exploratory statistics' tables of **VI** (**Figure 12**). They give an overview of the best outputs of the program. Only the values of the five runs of the program, which produce the lowest RMSE and the highest R^2 values, are listened.

IX. Find the best models for each dependent variable – Print multi-page document

This document is a merge of all PDF files from **VIII**, of each dependent variable and area of interest. This is **the most important output of the program** that can be printed. It gives you the possibility to find the best prediction model of each dependent variable and each area of interest. You have to decide whether to give more weight to RMSE or R^2 (**cf. p. 19**). After this decision you can **choose the best prediction model** of a specific dependent variable by its number ***x*** of program run. The final raster maps are contained in the folder called "*No_Run_**x***", which is itself contained in the folder "*Final_Models*" (**cf. I**).

X. Correlation coefficients

Before defining the predictors' combination to take into account (**cf. p. 14**) it can be useful to check autocorrelation between numeric variables and between the single predictors. The Pearson's ("*Pearson_Correlation_Predictarea*.png (xlsx, csv)*") and Spearman ("*Spearman_Correlation_Predictarea*.png (xlsx, csv)*") correlation coefficients are

calculated. The values, contained in the table files, range between -1 and 1. A correlation of -1 shows a perfect negative correlation while a correlation of 1 shows a perfect positive correlation. A correlation of 0 shows no linear relationship between the two variables taken into account. The PNG files do not contain the correlation coefficients directly, but they are grafically described with points (**Figure 13**). The color blue stands for a positive correlation and the color red for a negative correlation. The radius and the color strength describe the strength of the correlation. The larger and more intense the background of the points, the stronger the correlation between two specific variables. The Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable. The Spearman correlation coefficients measure the monotonic relationship between two variables. Here the variables tend to change together and the Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data (Minitab, 2019).

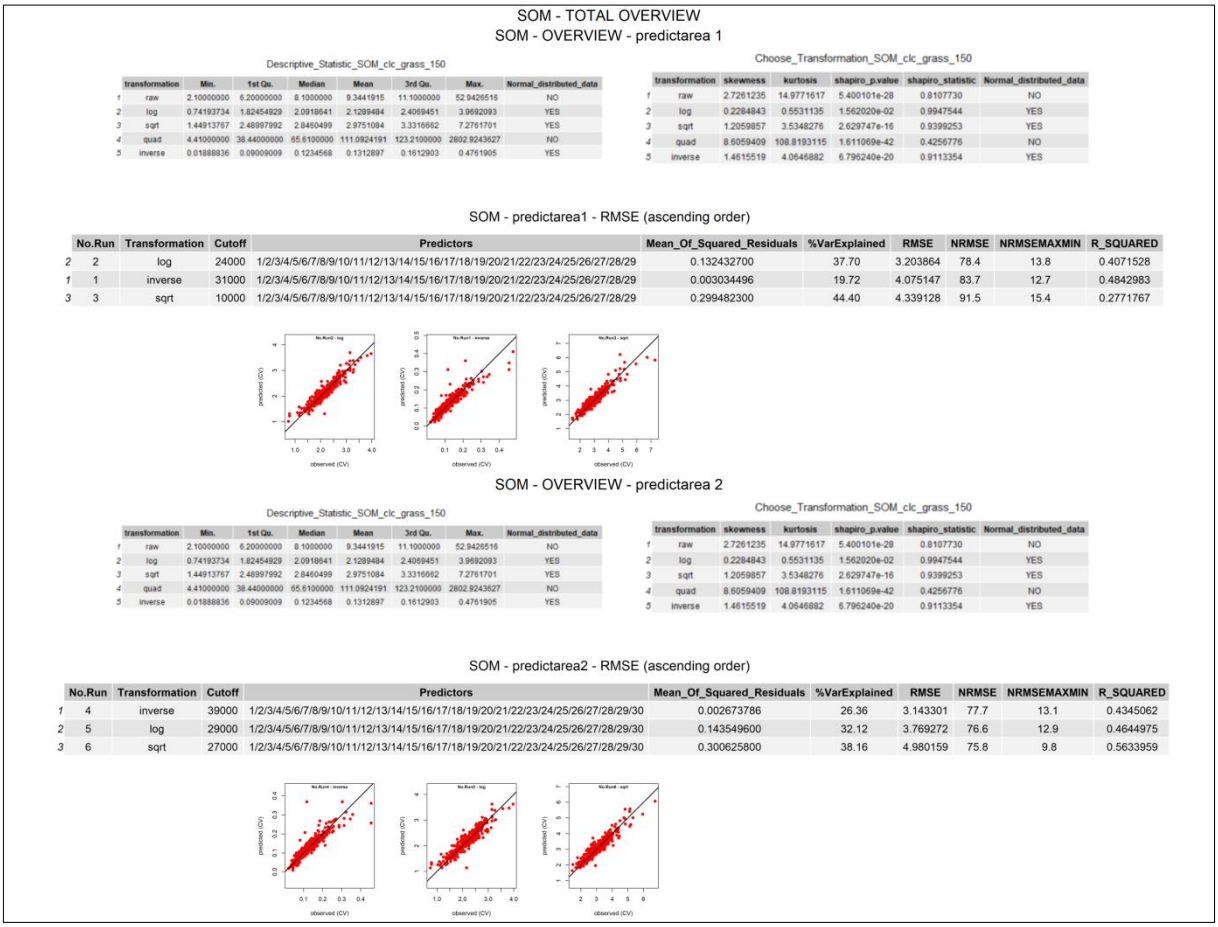


Figure 12: Example of the output that allows to choose the best prediction model of a specific dependent variable and area of interest.

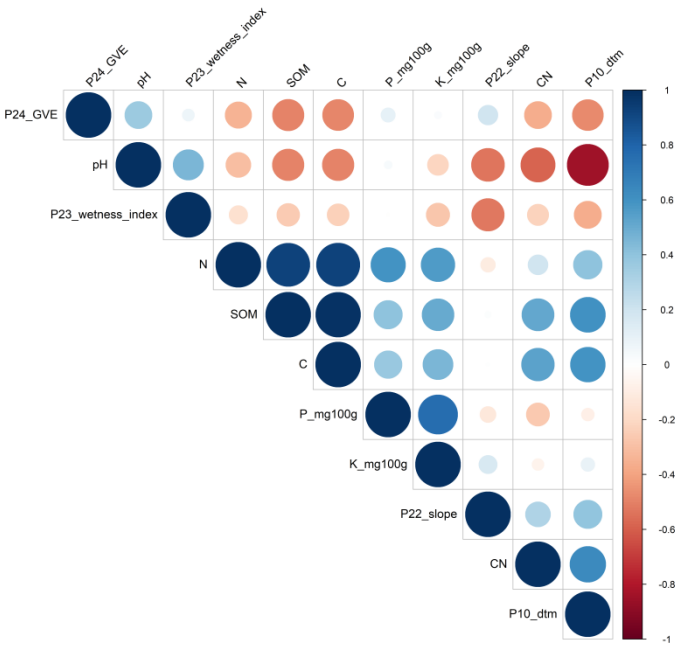


Figure 13: Example of a possible rapresentation of correlation coefficients.

8 REFERENCES

Breiman, L. (2001), "Random Forests", *Machine Learning*, Vol. 45 No. 1, pp. 5–32.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D.T., Duan, Z. and Ma, J. (2017), "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility", CATENA, Vol. 151 No. 1, pp. 147–160.

Ford, C. (2015), "Understanding Q-Q Plots", *University of Virginia Library*, https://data.library.virginia.edu/understanding-q-q-plots/ (access on 3rd February 2021)

Genova, G. (2017), "Spatial Distribution Assessment of Cu, Zn, PH and Soil Organic Matter in South Tyrolean Permanent Crops", Master Thesis dissertation, *Università degli studi della Tuscia*.

Hengl, T. (2006), "Finding the right pixel size", *Computers & Geosciences*, Vol. 32 No. 9, pp. 1283–1298.

Krige, D.G. (1951), "A statistical approach to some basic mine valuation problems on the Witwatersrand", *Journal of the Southern African Institute of Mining and Metallurgy*, Vol. 52 No. 6, pp. 119-139(21)

Lark, R.M. (2000), "A comparison of some robust estimators of the variogram for use in soil survey", *European Journal of Soil Science*, Vol. 51 No. 1, pp. 137–157.

Li, J. and Heap, A.D. (2014), "Spatial interpolation methods applied in the environmental sciences: A review", *Environmental Modelling & Software*, Vol. 53 No. 9, pp. 173–189.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. *R news*, 2(3), 18-22.

Minitab (2019), "A comparison of the Pearson and Spearman correlation methods", *Minitab,* https://bit.ly/39MWmod (access on 4th February 2021)

Krige, D.G. (1951), "A statistical approach to some basic mine valuation problems on the Witwatersrand", *Journal of the Southern African Institute of Mining and Metallurgy*, Vol. 52 No. 6, pp. 119-139(21)

9 FIGURES AND TABLES

Figure 1: Yellow points with known coordinates and chemical soil properties (e.g. SOM, pH, P, K, C, N) used for prediction in the green area of interest..... 7
Figure 2: Output plots of the Random Forest regression analysis. The goodness of fit describes the accuracy of the regression and the residual variogram gives an assessment of the variance of each variable. The crosses are the "experimental variogram" and the green curve is the "theoretical variogram". 8
Figure 3: Simple graphic illustration of how Ordinary Kriging (OK) interpolation works. 9
Figure 4: Overview of the input data folders. An explanation of the single files (1-5) follows. 10
Figure 5: Program changes in lines 62 and 64 ("res" is an abbreviation of resolution)..... 12
Figure 6: General syntax of an if statement..... 12
Figure 7: Example of how to lable the dependent variables in graph and plots (changes in line 298 ff.)..... 13
Figure 8: Overview of the content of the folder "output_overview". The explanation of the outputs I-X follows..... 20
Figure 9: Example of the output plots of the prediction part (Ordinary Kriging). 20
Figure 10: Example of the output of the exploratory / descriptive data analysis (A: boxplot and sampling density, B: histogram, C: normal Q-Q plot)..... 20
Figure 11: Example of outputs of the Random Forest regression describing the influence of the single predictors on a specific dependent variable (%IncMSE value)..... 20
Figure 12: Example of the output that allows to choose the best prediction model of a specific dependent variable and area of interest. 20
Figure 13: Example of a possible rapresentation of correlation coefficients. 20

Table 1: Binary reclassification of land use information (land use)..... 2
Table 2: Binary reclassification of detailed geological information..... 3
Table 3: Binary reclassification of geological information 4
Table 4: Binary reclassification of geological information 5
Table 5: Descriptive statistics of the original dataset (original and transformed data) for SOM, pH, P, K, N, C, N and C:N..... 6
Table 6: Check normality and stationary of different source data transformations. 7
Table 7: Possible overview table with the most important output values of each program run..... 10
Table 8: Layout of the table containing the source data of the chemical values (dependent variables)..... 11

Elvis Burchia

https://orcid.org/0000-0001-8925-2009

Software and documentation are available online:

https://github.com/elvisburchia/DigitalSoilMappingSoftware

DOI 10.5281/zenodo.4683096