Statistička analiza podataka o tlu

Habjanec, Petra

Undergraduate thesis / Završni rad

2024

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Electrical Engineering and Computing / Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:168:116406

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2025-03-14**

Repository / Repozitorij:

FER Repository - University of Zagreb Faculty of Electrical Engineering and Computing repozitory

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

BACHELOR THESIS No. 1569

STATISTICAL EXPLORATION AND INFERENCE ON THE SOIL DATASET

Petra Habjanec

Zagreb, June 2024

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

BACHELOR THESIS No. 1569

STATISTICAL EXPLORATION AND INFERENCE ON THE SOIL DATASET

Petra Habjanec

Zagreb, June 2024

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Zagreb, 04 March 2024

BACHELOR THESIS ASSIGNMENT No. 1569

Student:	**Petra Habjanec (0036537979)**
Study:	Electrical Engineering and Information Technology and Computing
Module:	Computing
Mentor:	assoc. prof. Marina Bagić Babac

Title: **Statistical exploration and inference on the soil dataset**

Description:

Soil is a non-renewable resource that requires constant monitoring to prevent its degradation and promote its sustainable management. The Land Use/Cover Area frame statistical Survey Soil (LUCAS Soil) is an topsoil survey conducted across the European Union and the largest harmonized open-access dataset of topsoil properties. The purpose of this thesis is to conduct a statistical descriptive and inferential analysis of the LUCAS dataset, exploring all measured soil parameters and performing regression on selected variables. This approach aims at better understanding how various factors influence soil properties and how this can impact land management and environmental sustainability.

Submission date: 14 June 2024

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

ZAVRŠNI ZADATAK br. 1569

Pristupnica:	**Petra Habjanec (0036537979)**
Studij:	Elektrotehnika i informacijska tehnologija i Računarstvo
Modul:	Računarstvo
Mentorica:	izv. prof. dr. sc. Marina Bagić Babac

Zadatak: **Statistička analiza podataka o tlu**

Opis zadatka:

Tlo je neobnovljivi resurs koji zahtijeva stalno praćenje kako bi se spriječilo njegovo propadanje i promoviralo održivo upravljanje. Land Use/Cover Area frame statistical Survey Soil (LUCAS Soil) je istraživački projekt tla provođen diljem Europske unije te predstavlja najveći usklađeni otvoreni skup podataka o svojstvima površinskog tla. Svrha ovog rada je provesti statističku deskriptivnu i inferencijsku analizu ovog skupa podataka, istražujući izmjerene parametre tla i provodeći regresiju za odabrane varijable. Ovim pristupom teži se boljem razumijevanju kako različiti čimbenici utječu na svojstva tla i kako to može utjecati na upravljanje zemljištem i održivost okoliša.

Rok za predaju rada: 14. lipnja 2024.

Thanks to everyone who endured me whilst I was writing this paper. :)

Thank you, Mare, for your greatly appreciated revisions.

Table of Contents

Introduction

This paper represents a comprehensive and in-depth research study into the fascinating world of soil properties, their distribution across geographical regions, and their impact on the types of land cover. The LUCAS (Land Use/Cover Area frame Statistical survey) Soil Database, established in the EU, provides a comprehensive data on soil properties across Europe, including a variety of measurements such as pH levels, electrical conductivity, organic carbon content etc. Soil science is a field that focuses on the study as a natural resource, it encompasses the formation, classification, and mapping of soils, and it examines their physical, chemical, biological, and fertility properties.

A key component of this research is the investigation into the relationship between these soil properties and the various land cover types including cropland, grassland, woodland and bare land. To successfully understand and predict these relationships, statistical tests and regression models are employed, each with its own set of strengths and limitations that have influenced the outcomes of the study.

This paper more clearly explains methods adopted and the results derived in predicting soil properties and land cover types. Various models are used, such as the K-Neighbours Regressor, Ridge Regression, and Random Forest Regression, each of which have contributed with unique insights into our research.

This research has not been limited to the soil properties alone. Spatial data is explored, using geo-spatial data to further enhance predictions and provide a wider view of the soil properties and their impact on land cover.

The findings of this research could serve as a valuable resource for further environmental research and can potentially guide the development of effective policies related to land management and environmental conservation. By gaining a deeper understanding of the soil and its properties, we can work towards more sustainable agricultural practices, better land use planning, and ultimately, a healthier planet.

1. Related work

Ballabio, et al. used [2] Gaussian process regression to map out LUCAS topsoil chemical properties. They made their study on the LUCAS data set from 2009-2012, selecting Gaussian processes regression mostly for its capacity for modelling uncertainty and the possibilities of adding prior knowledge in form of covariance to the model. Their goal was to make maps that would establish baselines that could help monitor soil quality and provide guidance for environmental research and help develop policies in the EU.

Ballabio, et al. in 2015[3] made a paper discussing how the LUCAS dataset can be used for mapping soil properties on continental scale. They describe predictions of soil texture and derived physical properties. Among other variables, their models used MODIS sensor data, that allowed them to monitor changes in vegetation depending on soil properties. They also explored predictions of collinear variables like soil texture.

In paper exploring Copper distribution in European soils, Ballabio, et al. [4] used Generalized Linear Models to investigate factors impacting copper distribution in EU soils. In regression analysis they found how important topsoil properties, land cover and climate are in predicting copper concentration in soil. They also found that, besides the traditional use of copper as fungicide, effects of soil properties, high pH, organic carbon and clay, combines with humid and wet conditions favoured copper accumulation in soil of vineyards and tree crops. Ballabio, et al. [4] also used Gaussian Process Regression (GPR) combined with kriging for mapping out concentrations of copper in topsoil. Their GPR model in combination with kriging accounted for 66% of Cu deviance.

In 2014, Panagos et al. [5] explored the soil erodibility in Europe. They explored the K-factor in soil erosion model, which is used for evaluating sensibility of soil to erode. Using the LUCAS soil survey from 2009, they calculated soil erodibility for LUCAS data points with nomograph of Wischemeier and Smith. With Cubist regression model, they tried correlating spatial data – latitude, longitude, and terrain features, they developed a high-resolution soil erodibility map. Their produced data set compared well with the local and regional data, but the protective effect of surface stone cover resulted in an average 15% decrease of the K-factor. By not including this effect in their calculations, their results are likely to overestimate soil erosion, especially in Mediterranean countries (where there are observed high percentages of surface stone cover).

Also, using the LUCAS dataset in 2020 Gao Y. et al. [6] evaluated global land-cover products (GLC) for understanding the differences between these products. Evaluating three 30-m GLC products based on the areal and spatial consistency using the LUCAS data set, they got that the GlobeLand30-2010 product accuracy was the best at the value of 88.90 ± 0.68%, followed by GLC_FCS30-2015(84.33 ± 0.80%) and FROM_GLC2015(65.31 ± 1.0%). They also discovered that the consistency between the GLC_FCD30-2015 and GlobeLand30-2010 is higher than the consistency between other products, also across the EU dominant land-cover types of forestry and cropland are more consistent across the three products, compared to the consistency for the bareland, grassland, shrubland, and wetland, which is relatively low.

Karydas et al. [7] used the LUCAS data frame form 2009 as a reference for validating a Land Cover Map of Greece from 2007. They decomposed critical vegetation parameters – type, height, density, and composition, to investigate unconformities between the maps. Using the "automated" and "supervised" processes, they made non-square error matrix for both processes. For the "supervised" process, they designed a decision-tree with the critical vegetation parameters, allowing objective labelling of both systems. In the end they concluded that the LUCAS point database was found to be supportive, but not entirely efficient, for identifying various sources of error in land cover maps derived with remote sensing. Concluding that the high-resolution satellite images and air photos are absolutely necessary for validating accuracy, specifically in heterogenous environments.

The study of Weigand et al. [8] evaluates the use of LUCAS in-situ reference data for classifying high-resolution Sentinel-2 imagery on a large scale, proposing a new pre-processing scheme for automated national-level classification. Comparing different positioning and semantic selection approaches, their study that the positional correlation significantly enhances classification accuracy, with an average improvement of 3.7%. Newly developed pre-processing scheme achieves the highest overall accuracy of 93.1%, while other pre-processing schemes achieve over 80% of accuracy. It was concluded that the LUCAS in-situ data is suitable for reference information on a large-scale high-resolution LC mapping using Sentinel-2 imagery.

In the paper *Automatic classification of land cover from LUCAS in-situ landscape photos using semantic segmentation and a Random Forest model* [9] they aimed to develop computer vision methodology to extract land cover information form the photos from the LUCAS data base. Using the selected representative 1120 photos covering different land

cover types, and for each figure they used the LUCAS land cover type, segmented objects and pixel count for each ADE20k class, using those as input features they trained Random Forest model to predict the land cover type of photos. Their model shows a mean F1 score of 89%, and if wetlands are not included the mean F1 score increases to 93%.

Showing the importance of spatial data during the GIS summer school in Zagreb, Kliment et al. [10], groups have collected point features with data attributes for HILUCS land use and LUCAS land cover codes, together with photos representing the area. The product was a land use data set for Zagreb from 2014 stored on the GIS database, showing that INSPIRE and LUCAS approaches can be used to make a harmonized land use dataset from topological and fieldwork data.

Also, demonstrating the importance of spatial data d'Andrimont et al. [11] harmonized the data collected during five LUCAS surveys, making the most comprehensive in-situ dataset on land cover and use in European Union, making it valuable for geo-spatial and statistical analysis though years.

1.1. Research Questions

There are several key research questions explored in subsequent sections. These questions include investigating differences in soil properties distribution across various land cover types, geographical regions, and countries within the same region. Research questions explored in this paper are:

- *RQ1*. Are there differences in soil properties distribution between soil properties for different Land Cover types?
- *RQ2.* Are there differences in soil properties distribution between different geographical regions?
- *RQ3.* Prediction of electrical conductivity based on soil properties and geospatial data.
- *RQ4.* Prediction of pH measured in calcium chloride based on soil properties and geospatial data.
- *RQ5.* Prediction of pH measured in water based on soil properties and geospatial data.
- *RQ6*. Predicting land cover type based on soil properties.

From these research questions we explore the following hypotheses:

- ***Ha.*** There is no significant difference in pH measured in calcium chloride for different land cover types.

- ***Hb*** There is no significant difference in pH measured in water for different land cover types.
- ***Hc*** There is no significant difference in calcium carbonate content for different land cover types.
- ***Hd*** There is no significant difference in electrical conductivity for different land cover types.
- ***He*** There is no significant difference in extractable potassium for different land cover types.
- ***Hf*** There is no significant difference in total nitrogen for different land cover types.
- ***Hg*** There is no significant difference in organic carbon content for different land cover types.
- ***Hh*** There is no significant difference in total phosphorus for different land cover types.
- ***Hi.*** There is no significant difference in pH measured in calcium chloride for different geographical regions.
- ***Hj*** There is no significant difference in pH measured in water for different different geographical regions.
- ***Hk*** There is no significant difference in calcium carbonate content for different geographical regions.
- ***Hl*** There is no significant difference in electrical conductivity for different geographical regions.
- ***Hm*** There is no significant difference in bulk density in 0-10cm depth for different geographical regions.
- ***Hn*** There is no significant difference in extractable potassium for different geographical regions.
- ***Ho*** There is no significant difference in total nitrogen for different geographical regions.
- ***Hp*** There is no significant difference in organic carbon content for different geographical regions.

- ***Hr*** There is no significant difference in total phosphorus for different geographical regions.
- ***Hs*** There is no significant difference in bulk density in 10-20cm depth for different geographical regions.

2. Descriptive Analysis

LUCAS data frame is an area frame statistical survey organized and managed by Eurostat, the Statistical office of the EU, to monitor the changes in land use (LU) and Land Cover (LC) over time.

In 2018, soil sampling was carried out in all EU member states using the same set of 25,947 locations that were targeted in 2015. In 65% of these locations, samples were to be taken following the standardized sample from a depth of 20 cm. In the remaining 35% of the locations (approximately 9,000 points), metallic rings were used to collect soil cores to determine bulk density from the depth of 0-10 and 10-20 cm^2. Also, 1,000 fresh samples were also collected to assess biodiversity. In the end there are 18,984 soil samples in the LUCAS Soil frame from 2018. In this round of surveys, the number of sampled locations was lower compared to the previous surveys due to a range of issues – land ownership, meteorological conditions during the survey and difficulties in reaching the locations. Also, compared to previous years, the survey in 2018 included not only the physio-chemical properties of the soil, on some data points additional analysis of soil biodiversity, bulk density, and field measurements were obtained.

At each LUCAS point, the surveyors documented agro-environment observations by filling out a field form and by taking photographs. The 2018 LUCAS Soil module comes with multiple files providing various aspects of soil data:

- Basic LUCAS Soil – providing an insight into depths at which the samples were taken, basic soil data, oxalate extractable aluminium and iron, land use and land cover
- LUCAS Soil Erosion
- LUCAS Soil Organisms
- LUCAS Soil Bulk Density

This paper is focused on statistical analysis of LUCAS Soil and LUCAS Soil Bulk Density. In Tables 1.1 and 1.2 we can see the descriptions of data given in the data sets.

Table 2.1 Fields in the file of basic soil properties in the 2018 LUCAS Soil Module

Filed	Description
Depth	Based on sample collected (e.g. 0-20cm, 0-10cm, 20-30cm)

POINTID	LUCAS Point Identifier – link to Eurostat LUCAS Microdata
pH_CaCl2	pH – measured in calcium chloride
pH_H2O	pH – measured in water
EC	Electrical conductivity (milli Siemens per meter – mS m^{-1})
OC	Organic carbon content (g kg^{-1})
CaCO3	Calcium carbonate content (g kg^{-1})
P	Total phosphorus (g kg^{-1})
N	Total nitrogen (g kg^{-1})
K	Extractable potassium (g kg^{-1})
OC (20-30cm)	Organic carbon content (g kg^{-1})
CaCO3 (20-30cm)	Calcium carbonate content (g kg^{-1})
Ox_Al	Oxalate extractable Al (mg kg^{-1})
Ox_Fe	Oxalate extractable Fe (mg kg^{-1})
NUTS_0, ..., NUTS_3	NUTS 0 Code, ..., NUTS 3 Code
TH_LAT	LUCAS POINT Theoretical Latitude
TH_LONG	LUCAS POINT Theoretical Longitude
SURVEY_DATE	Date of Survey
Elev	Elevation in meters from surveyor GPS
LC	Primary land cover
LU	Primary land use
LC0_Desc	Description of primary land cover
LC1_Desc	Description of secondary land cover
LU1_Desc	Description of primary land use

Table 2.2 Fields in the file of Bulk Density of LUCAS Soil 2018

POINT ID	LUCAS Point Identifier
BD 0-10	Measured Bulk Density for the depth 0-10 cm (g cm^{-3})
BD 10-20	Measured Bulk Density for the depth 10-20 cm (g cm^{-3})
BD 20-30	Measured Bulk Density for the depth 20-30 cm (g cm^{-3}) - only Portugal

BD 0-20	Bulk Density for the depth 0-20 cm (g cm^{-3}) - arithmetic mean based on values for 0-10 cm and 10-20 cm

2.1. Sample Collection

For **common sampling procedure**, samples of approximately 500g were taken at each LUCAS point. For each sample taken the subsamples of a geo-referenced location, and then four other subsamples each 2m in the cardinal directions and all of them mixed in a bucket. Before taking the samples, stones, vegetation residue, grass and litter were removed from the soil surface. Approximately 500g of the mixed soil was taken from the bucket, placed in a plastic bag, and labelled. Soil samples were allowed to air dry in the bags before they were sealed.[1]

For determining **bulk density**, from the depths of 0 to 10 cm and 10 to 20 cm soil cores were collected. In Portugal, the soil cores were taken from the depths of 20 to 30 cm to test the implications of extending the sampling depths in the LUCAS Soil Module. Again, before taking the samples, vegetation residues, grass and litter were removed from the soil surface. Five coils were taken from 0 to 10 cm depth with a metallic ring od 100 cm^3 at each LUCAS point. Again, the initial coil was taken from geo-referenced location and the other four coils were taken at 2m in each cardinal direction.[1]

Lastly, assessment of **soil biodiversity** was carried out on the same subset of locations as for bulk density. Field moist samples were taken from a depth of 20 cm using the standard sampling procedure. The final sample was placed in a labelled jar and stored in a polystyrene box that was had been cooled with freezer packs, then the samples were sent to the JRC to preserve the biological characteristics. Samples were then frozen and stored at –20C at the JRC until their shipment to the laboratory for analysis.[1]

2.2. Data Frame Description

In this data analysis, two previously mentioned data frames have been used. When analysing the bulk density of the soil, the two data frames were merged by the point ids. Also, the outliers were removed based on the interquartile range, they were replaced with *Nan* values. In the Basic Soil Properties data frame, we have 18,984 data points, and in the Bulk Density

data frame, we have 6,172 data points. In Table 1.3, dispersion of each value in the data frame of Basic Soil Properties can be seen, and the same information for the merged Bulk Density with Basic Soil Properties in Table 1.4.

Table 2.3 Descriptions of numerical data in Basic Soil Properties data frame

	pH_CaCl2	pH_H20	EC	OC	CaCO3	P	N	K
count	18,983	18,983	17,781	16,941	9,804	13,303	17,332	17,955
mean	5.706	6.259	14.239	25.379	42.398	30.451	2.232	172.384
std	1.398	1.319	8.005	17.857	75.332	17.928	1.311	115.193
min	2.600	3.340	0.240	2.100	1.000	0.300	0.200	6.200
25%	4.500	5.120	7.770	12.400	1.000	16.200	1.300	82.200
50%	5.800	6.290	13.170	19.600	3.000	25.400	1.900	144.900
75%	7.100	7.500	18.860	33.200	44.000	40.500	2.900	237.500
max	9.800	10.43	39.340	87.100	306.000	85.100	6.500	525.000
	CaCO3 (20-30 cm)		Ox_Al	Ox_Fe	Elev			
count	14		2,313	2,374	18,459			
mean	288.857		0.928	2.128	397.843			
std	236.021		0,470	1.494	366.193			
min	1.000		0,000	0.100	-55.000			
25%	64.500		0,600	1.000	121.000			
50%	287.500		0,800	1.700	251.000			
75%	417.750		1.200	2,900	600.000			
max	702.000		2.300	6.700	1,478.000			

Table 2.4 Descriptions of numerical data in Bulk Denstity data frame

	pH_CaCl2	pH_H20	EC	OC	CaCO3	P	N	K
count	6,171	6,171	5,793	5,484	2,866	4,224	5,599	5,832
mean	5.749	6.309	13.982	24.735	111.846	29.429	2.134	172.209
std	1.404	1.354	7.526	17.717	152.626	16.652	2.248	115.135
min	2.800	3.340	0.240	2.200	1.000	0.300	0.200	6.200
25%	4.600	5.180	7.880	12.100	1.000	16.300	1.200	83.275

50%	5.800	6.350	13.210	19.000	20.000	24.700	1.800	144.250
75%	7.100	7.590	18.470	31.800	195.750	39.000	2.700	236.225
max	8.400	9.620	37.350	86.300	574.000	80.300	6.200	516.700
	CaCO3 (20-30 cm)		Ox_Al	Ox_Fe	Elev	BD 0-10	BD 10-20	BD 20-30
count	14		794	814	6,059	5,908	5,390	189
mean	288.857		0.974	2.054	422.287	1.052	1.164	1.201
std	236.021		0.487	1.411	385.0.51	0.333	0.284	0.207
min	1.000		0.100	0.100	-50.000	0.147	0.329	0.715
25%	64.500		0.600	0.900	128.000	0.863	0.978	1.058
50%	287.500		0.900	2.700	260.000	1.102	1.196	1.231
75%	417.750		1.200	2.800	677.50	1.284	1.366	1.326
max	702.000		2.500	6.500	1,569.00	1.950	1.972	1.661

2.3. Cartographic Representation of Land Cover Types

In Figure 1.1, different Land Cover types mapped out across the continent can be seen. It is clear from the figure that most of points were taken on the cropland surfaces, with 7430 points. Second most points had the woodland land cover type, with 6092 points, next was the grassland type with 3988 points. There were 720 shrubland points, 638 bare land points, 71 artificial land points, 40 wetland points, and lastly only 5 water points.

It is observed that Scandinavian countries have a high concentration of woodland. Spain, France, and Italy have a high concentration of cropland, with Bulgaria, and Romania also having patches of high density of cropland. Also, worth mentioning is the high density of shrubland and bare land points in Spain, and high density of grassland points in Ireland.

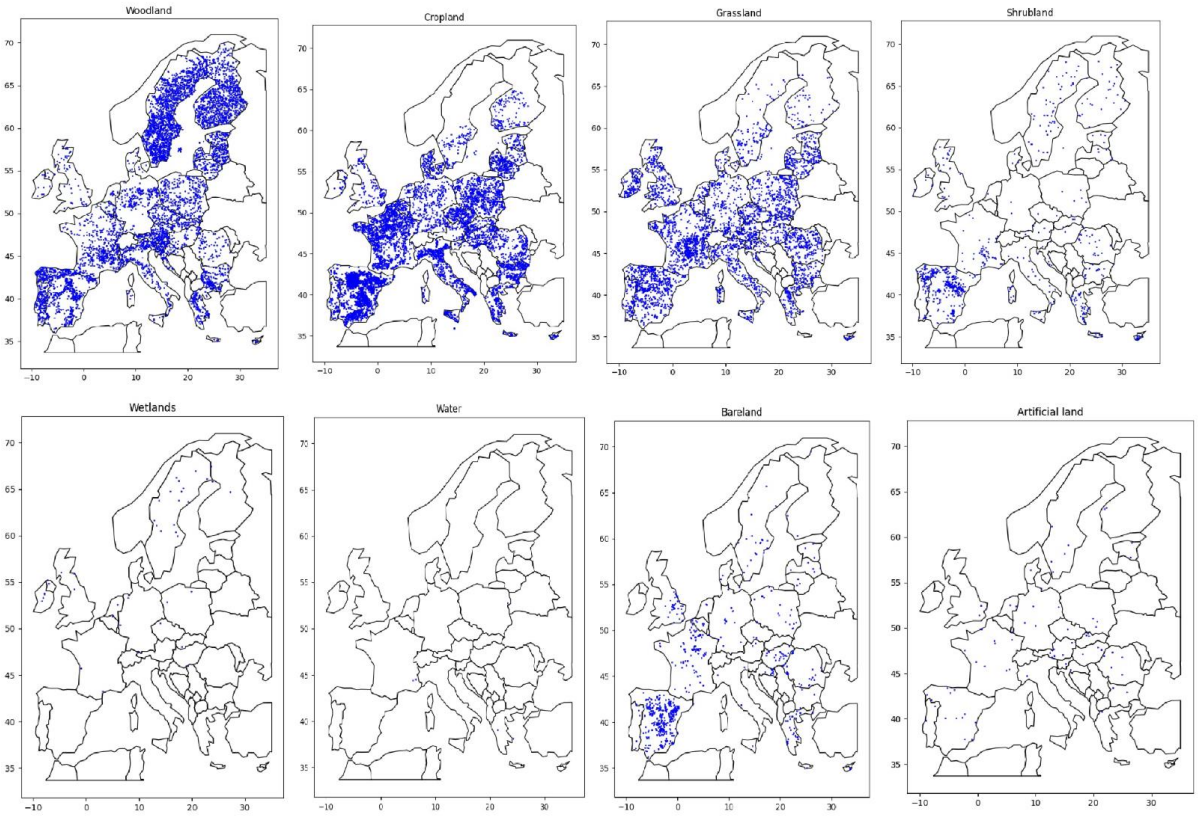

Figure 2.1 Maps of description of Land Cover

2.4. Cartographic Representation of Land Use Types

In Figure 1.2, spatial distribution of Land Use types with more than 100 points can be seen, mostly because of many diverse types of Land Cover with low point count. Firstly, Land Cover type with most points is Agriculture (excluding fallow land and kitchen gardens), with 10,931 points, then Forestry with 5,603 points, Semi-natural and natural areas not in use with 1,284 points, Fallow land with 737 points, and Other abandoned areas with 123 points.

Following types are not included in the map representation: Amenities, museum, leisure (e.g. parks, botanical gardens) (66 points), Electricity, gas and thermal power distribution (56 points), Residential (54 points), Road transport (35 points), Kitchen gardens (23 points), Mining and quarrying (12 points), Community services (8 points), Sport(8 points), Energy production (6 points), Abandoned residential areas (6 points), Protection infrastructures (6 points), Construction (5 points), Other primary production (4 points), Railway transport (4 points), Commerce (4 points), Water supply and treatment (2 points), Logistics and storage (2 points), Abandoned industrial areas (2 points), Water transport (1 point), Financial, professional and information service (1 point) and Abandoned transport areas (1 point).

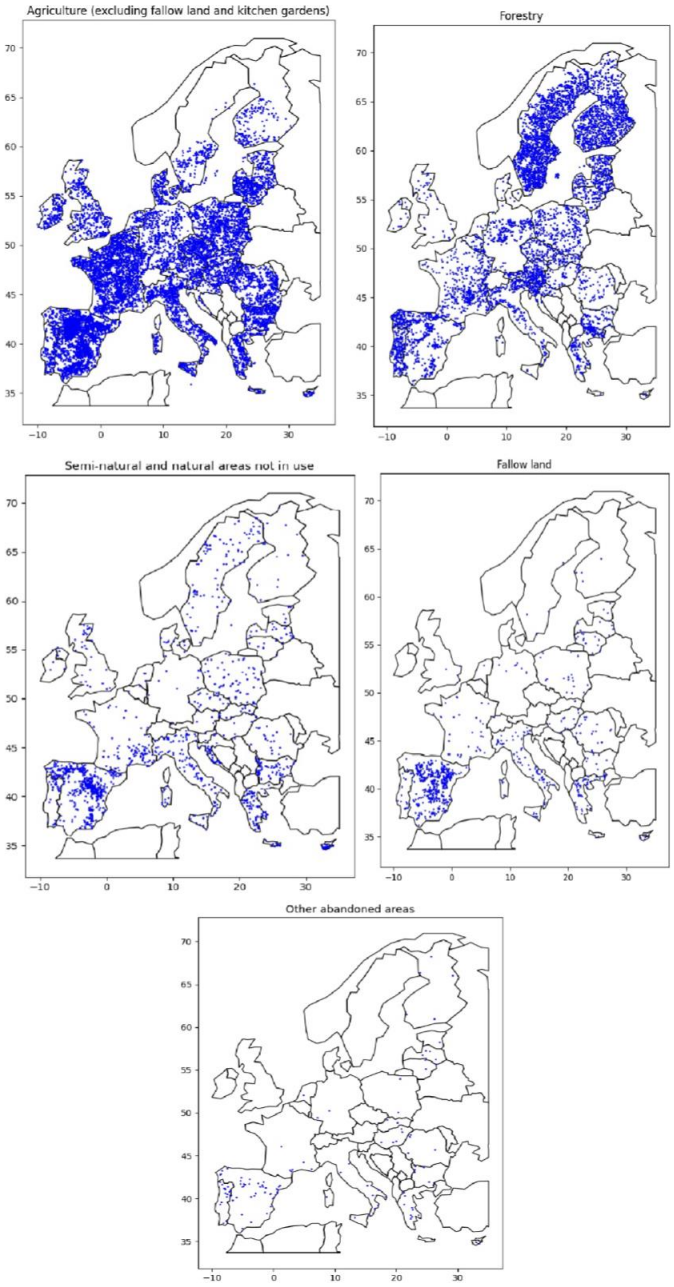

Figure 2.2 Figure Maps of description of Land Use

2.5. Mapping and Visualizing Soil Properties

Photos 1.3-1.11 explain distributions of oxalate extractable aluminum, oxalate extractable iron, electrical conductivity, extractable potassium, total nitrogen, total organic carbon content, total phosphorus, pH measured in calcium chloride, and pH measured in water.

Photos 2.3 and 2.4 show how oxalate aluminum and oxalate iron were taken in only a little bit over 10% of points, about 2,500 points out of 19,000 points. For oxalate aluminum it can be seen how most of the higher values reside in Scandinavian countries, southern Italy, and northwest France. Most of the lower values are in central Spain, Poland, and Germany. The spatial distribution of oxalate iron looks quite like that of oxalate aluminum.

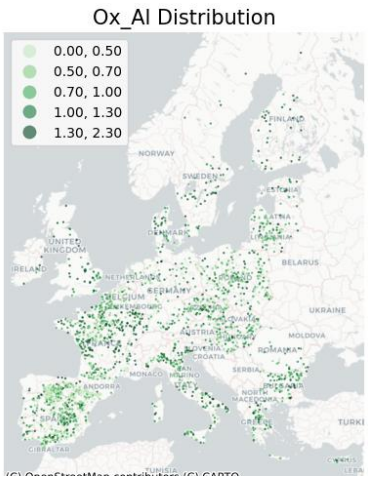

Figure 2.3 Map of oxalate aluminium distribution across a map

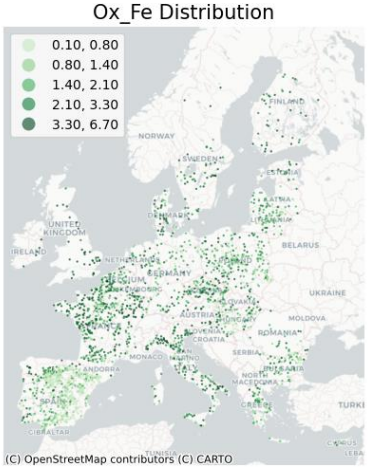

Figure 2.4 Map of oxalate iron distribution across a map

Figure 2.5 explains distribution of electrical conductivity through Europe. Some of the highest values lie in the UK, some parts of Spain, and in an area between Italy and Slovenia. The lowest values of electrical conductivity lie in Scandinavian countries.

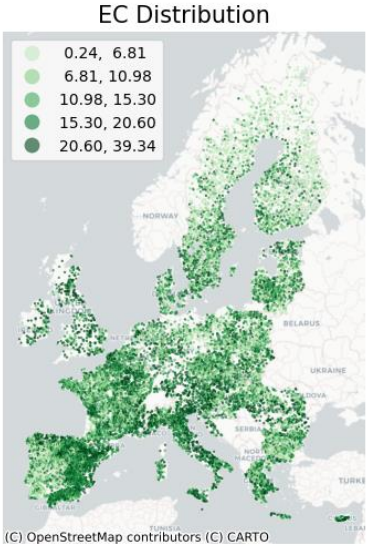

Figure 2.6 Map of electrical conductivity distribution across a map

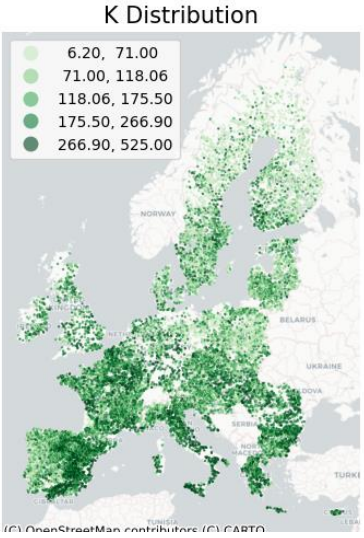

Figure 2.5Map of total potassium distribution across a map

Figure 2.6 shows that some of the highest values of extractable potassium are in Mediterranean parts of Spain and southern Italy, and the lowest values are in Scandinavia, northern parts of Germany, and Poland.

Figure 2.7 shows distribution of total nitrogen. Some of the higher nitrogen values are in northern coastal parts of Spain, and southern part of Sweden, and Slovenia. Lower nitrogen values are in Spain, Poland, and northern parts of Scandinavia.

Figure 2.8 is presenting distribution of total organic carbon, with higher values being in Slovenia, northern seaside part of Spain and southern parts of Scandinavia, lower values being in central Spain, and central Poland.

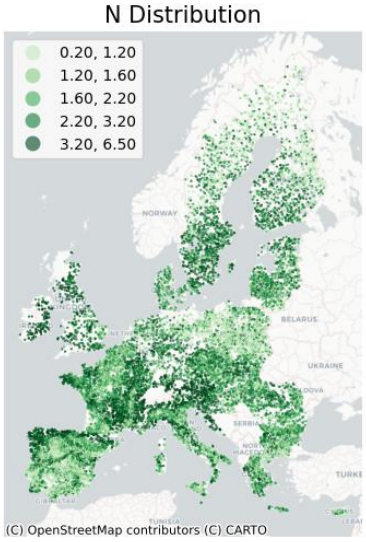

Figure 2.8 Map of total nitrogen distribution across a map

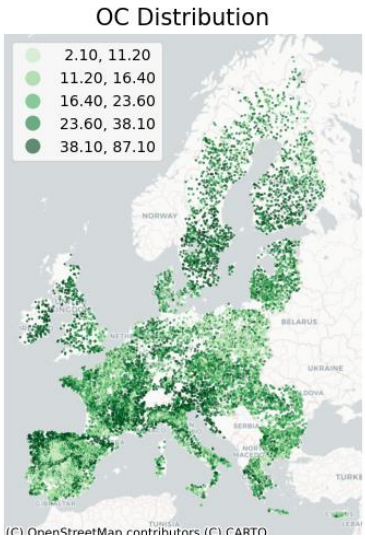

Figure 2.7 Map of organic carbon content distribution across a map

Figure 2.9 shows distribution of total phosphorus, with high values being in Denmark, the Netherlands, Poland, and northeastern parts of France, and low values in Romania and central Italy.

Figure 2.10 shows distribution of pH measured in $CaCl_2$. Some of the higher values of pH lie in south and southeast parts of the Spain, Greece, and Adriatic parts of Italy. Some of the lowest values lie in Scandinavian countries.

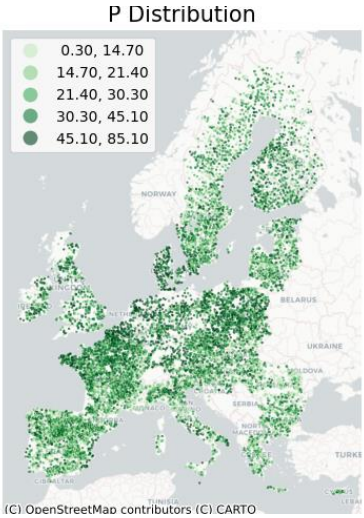

Figure 2.10 Map od total phosphorus distribution across a map

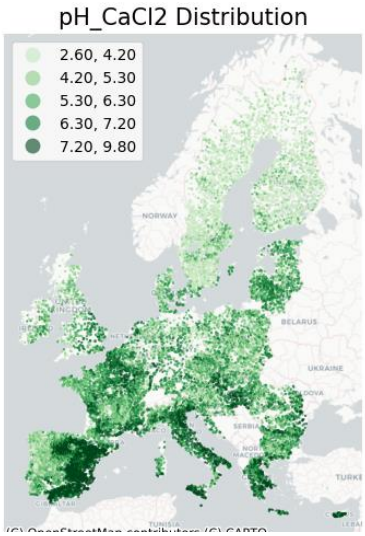

Figure 2.9 Map of pH measured in calcium chloride distribution across a map

Figure 2.11 shows distribution of pH measured in H_2O, where the distribution is similar to the distribution of pH $CaCl_2$.

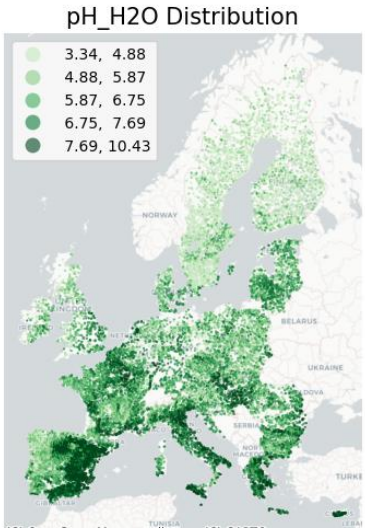

Figure 2.11 Map of pH measured in water distribution across a map

2.6. Boxplots by Countries

This part examines differences between the countries in measured values. Regarding the boxplots for oxalate extractable aluminum (Figure 5.1), there is a noticeable variability in oxalate extractable aluminum levels across countries. Some countries like Finland and Ireland exhibit higher median values, indicating richer aluminum oxalate content. Countries like Latvia, Belgium, and Slovakia show narrow IQRs, indicating low variability of values in these countries. Contrary to that, countries like Cyprus, the UK, and the Netherlands have wide IQRs showing great variability of values within those countries.

Regarding boxplots for oxalate extractable iron (Figure 5.2), levels of oxalate extractable iron show substantial differences among countries. Ireland and the UK stand out with higher median values, reflecting higher iron oxide content. A considerable number of outliers are present in countries like Spain and Poland, indicating a wide range of iron oxalate concentrations. Countries such as Latvia, Lithuania, Ireland, Slovenia, and Spain show narrow IQRs, indicating low variance of values in those countries, while others like Finland, Italy, and Sweden have wider IQRs, signifying diverse iron oxide levels.

Turning to the values of electrical conductivity across European countries (Figure 5.3). Countries like the UK, Italy, and Ireland have higher median values, indicating higher values of EC in their soil. There are numerous outliers in all countries, with Finland, Spain, and Sweden having especially high counts of outliers, suggesting considerable variation in electrical conductivity of soil. For IQRs, countries such as Finland and Luxembourg have narrower IQRs, whereas Ireland, Austria, Romania, and the UK exhibit broader IQRs indicating greater variation of EC across those countries.

Looking at the distribution of pH measured in H_2O (Figure 5.4), a substantial variability across countries is revealed. Sweden and Finland stand out as countries with by far most outliers, suggesting considerable variability. Countries like Hungary, and Czech Republic have wide IQRs, showing a wide variation, contrary to that Cyprus stands out with quite narrow IQR, showing small to no variation in soil pH. Looking at the pH measured in $CaCl_2$(Figure 5.5), similar distribution of pH can be seen, with some of the biggest differences showing up for Hungary and Greece. Hungary for this measure of pH has a bit narrower IQR for $CaCl_2$, and Greece having one major outlier lying in the pH of 10.

Looking at the distribution of calcium carbonate content, in Figure 5.6, there are quite a few outliers, showing that $CaCO_3$ distribution in all countries is quite various. Countries like Spain and Italy have a wide IQRs, showing the wide variation of calcium carbonate content in those countries. Many countries show narrow IQRs and many outliers, showing great variation of values within those countries.

Analysing the graph in Figure 5.7, distribution of extractable potassium is examined. Most of the countries have a wide spread of IQRs and many outliers, indicating a great variation of extractable potassium across most European countries. Only country jumping out is Malta because of only 2 readings in that country.

Similar occurrence is observed in Figure 5.8 - widespread IQRs and many outliers with distribution of total nitrogen in soil. In this distribution, Cyprus has narrower IQR and a small number of outliers, showing a small variation of total nitrogen within the country's soil.

For distribution of organic carbon content (Figure 5.9), there is a similar pattern of countries having wide IQRs and many outliers, with Cyprus and Belgium standing out with a small number of outliers and narrow IQRs. Observing the graph in Figure 5.10, there is a similar pattern in the distribution of total phosphorus, many outliers and wide IQRs across many countries. The ones that stand out the most are Belgium, Luxemburg, Slovakia, and Netherlands with wide IQRs but not many outliers. This suggests that, although significant variation is present, values remain within a consistent range without extreme deviation.

Lastly, bulk density for the depth of 0-10cm in Figure 5.11 is examined. Median values of bulk density values vary across countries, indicating differences in the central tendency of the data. Countries like Slovenia, Latvia, Estonia, Finland, and Sweden have wide IQRs and no outliers, indicating a wide variation of values, but still within certain limits. On the other hand, countries like Czech Republic, France, and Spain still have wide IQRs, but also outliers on each side indicating the presence of extreme values outside of a certain range. In Figure 5.12, for bulk density on depth of 10-20cm there is a continuation of a similar pattern.

2.7. Dependencies of Values

Looking at Figure 2.12 it can be seen that there are correlations of basic soil properties from the database. Some of the highest values worth mentioning are correlation between pH measured in calcium chloride and pH measured in water with the correlation of 0.99, and correlation between organic carbon content and total nitrogen with the correlation of 0.83.

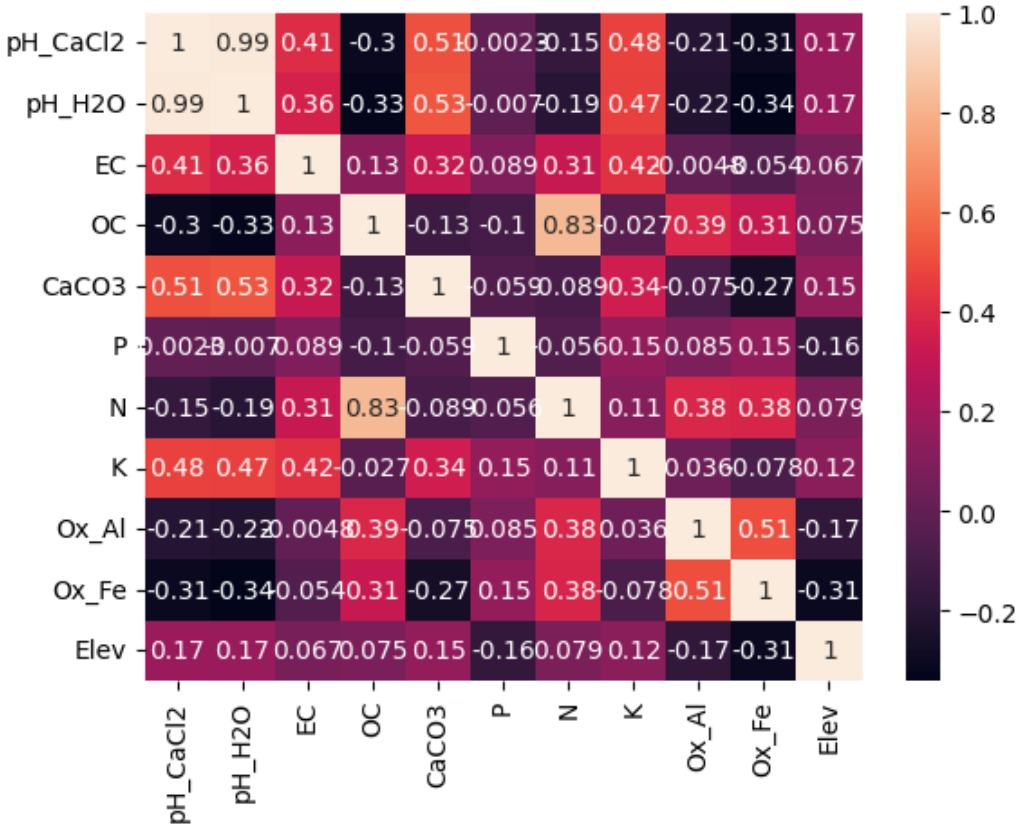

Figure 2.12 Correlation of basic soil properties

3. Methodology and Theoretical Framework

3.1. Kruskal-Wallis Test

The Kruskal-Wallis test [12] is a nonparametric procedure for testing if the k independent samples are from identical populations. For testing the null hypothesis, we compute the formula (1) where r_i is the assumed value of R_i, for $i = 1, 2, \dots, k$. If h falls in the critical region $H > \chi_\alpha^2$ with $v = k + i$ degrees of freedom, we reject the H_0 at the α-level of significance, otherwise we fail to reject the H_0.

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{r_i^2}{n_i} - 3(n+1) \qquad (1)$$

3.2. Conover's test

The Conover's test [13] performs the post hoc pairwise multiple comparisons procedure appropriate to follow the rejection of Kruskal-Wallis test. Conover tests make $m = \frac{k(k-1)}{2}$ multiple pairwise comparisons based on the Conover-Iman t-test statistic for the rank-sum differences (2), where $\widehat{H} *$ is defined like in (1), $\overline{R_i}$ and $\overline{R_j}$ are the mean ranks of groups i and j, s^2 the variance of ranks defined like in (3), n is the total number of observations, k is the number of groups, n_i and n_j are the sizes of groups i and j. Here we have null hypothesis that the distribution of group i is equal to the distribution of group j.

$$\left|\overline{R_i} - \overline{R_j}\right| > t_{1-\frac{\alpha}{2};n-k} \sqrt{s^2 \left[\frac{n-1-\widehat{H}*}{n-k}\right] \left[\frac{1}{n_i} + \frac{1}{n_j}\right]} \qquad (2)$$

$$s^2 = \frac{1}{n-1} \left[\sum R_i^2 - n\left(\frac{n+1}{2}\right)^2\right] \qquad (3)$$

Here using the Holm p-value adjustment with the formula (4) where k is the rank of the p-value, and m is the total number of comparisons.

$$p - value = \min\left(p \times (m - k + 1), 1\right) \qquad (4)$$

3.3. R-Squared – Coefficient of Determination

R^2 is a coefficient of determination [14], it is a measure of the proportion of variability explained by the fitted model. Using the sum of squares (5) and the total corrected sum of squares (6). SSE values is the variation due to an error, or variation explained. Then R^2 is values is calculated by the formula (7). If the fit is perfect, all residuals are zero, and then $R^2 = 1.0$, but if SSE is only slightly smaller than SST, then we have $R^2 \approx 0.0$. A value of $R^2 \approx 1.0$ illustrates a good fit, and $R^2 \approx 0$ a poor fit.

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (5)$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \qquad (6)$$

$$R^2 = 1 - \frac{SSE}{SST} \qquad (7)$$

3.4. Mean Squared Error

Mean squared error (MSE) [15] measures the amount of error in statistical models. It addresses the average squared difference between the observed and predicted values. The formula is as it states in (8), where y_i is the i-th observed value, $\hat{y}_i$ is the corresponding predicted value, and n as the number of observations.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n} \qquad (8)$$

The calculations for the mean squared error are similar to the variance. Squaring increases the impact of larger errors. These calculations disproportionately penalize larger errors more than smaller errors.

3.5. K-Nearest Neighbours Regressor Model

The K-Nearest Neighbours (KNN) algorithm [16] is a non-parametric method used for regression. The model operates on the principle that similar instances exist in close proximity in the feature space. It predicts the value of a new data point based on the values of its k-nearest neighbours in the training dataset.

In this study, the Haversine metric (9) [17] is employed to measure the distance between points. It is used for calculating the great circle distance between two points on the Earth's surface, providing an accurate measure of distance over spherical geometry. This is

particularly useful considering we are using distances measured along the earth surface, rather than the straight line.

$$d = 2r \sin^{-1}\left(\sqrt{\sin^2(\frac{\phi_2-\phi_1}{2}) + \cos(\phi_1)\cos(\phi_2)\sin^2(\frac{\lambda_2-\lambda_1}{2})}\right) \quad (9)$$

Algorithm works by first calculating the distance, then the algorithm identifies the k-nearest neighbours to the query point on the calculated distance. For regression tasks, it calculates the predicted value for the query based on the mean of the values of its k-nearest neighbours.

It takes two parameters, on we already mentioned is the distance metric where we use the Haversine metric. It also takes the k, number of neighbours considered for calculating.

3.6. Radius Neighbours Regressor

The Radius Neighbours Regressor [18] is an extension of the KNN model, but instead of fixed number of neighbours, it considers all neighbours within a specified radius r.

For each test data point, all training data points within a specified radius r are identified using the Haversine distance (9) [17]. For predicting, the output value is computed as the mean of the target values of the neighbours within the radius.

The radius within which neighbours are considered is critical parameter. Too small of a number might result in too few neighbours, while a large radius could include irrelevant neighbours.

3.7. Ridge Regression

Ridge Regression [19] is a technique used to address multicollinearity in linear regression models by adding a penalty term to the ordinary least squares (OLS) method, which helps in reducing the model's complexity and preventing overfitting.

It modifies the cost function of the linear regression model by adding a penalty proportional to the square of the magnitude of the coefficient. The goal is minimizing the sum of the squared residuals (RSS) while also shrinking the coefficients (10) where λ is the regularization parameter, and β_j are the coefficients. $\lambda \sum_{j=1}^{p} \beta_j^2$ is a regularization term which introduces bias into the model. This bias helps to reduce the variance, leading to better performance on new data by avoiding overfitting.

$$Ridge\ Regression\ Cost\ Function = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (10)$$

Process is, fitting a linear regression model to the data by minimizing the RSS (11), where y_i are the observed values and $\hat{y}_i$ are the predicted values. Then adding the regularization term (12) to modify the cost function to include the penalty term.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (11)$$

$$Cost\ Function = R \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (12)$$

Ridge Regression introduces bias into the model to reduce variance, this trade-off is crucial in preventing overfitting, as a model with high variance in sensitive to small fluctuations in the training data.

3.8. Random Forest Regression

Random Forest Model [20] works by creating many decision trees, each built on a randomly chosen subset of the data. It uses an ensemble of decision trees to predict continuous target variables.

When building a random forest regression model, firstly we make a number of randomly chosen bootstrapped subsets of the data, then from making decision trees from each one. Each subset can contain the same data row more times. So, then when the model is making predictions, it gets the outputs of each tree it created and based on the average of all the outputs from the forest of decision trees.

This process has many advantages, it is less prone to overfitting than other linear models. These models are computationally efficient and require fitting fewer parameters compared to other algorithms.

3.9. Multinominal Logistic Regression

Multinominal Logistic regression [21] is an extension of logistic regression that allows for the inclusion of more than one independent variable. This methodology is particularly useful when the research involves a nominal dependent variable and two or more measurement independent variables. The primary goal of multiple logistic regressor is to model the relationship between these variables and to predict the probability of certain outcomes based on the values of the independent variables.

The model given the predictors $X_1, X_2, \ldots, X_n$ models the probability of each j of Y by (13), it implies that $\sum_{j=1}^{J} p_j(\boldsymbol{x}) = 1$ and that there are $(J-1) \times (p+1)$ coefficients. In the formula β_{0j} is the intercept term for class j, β_{ij} are the coefficients for predictor X_i for class k, n is the number of predictors, and $J$is the number of classes.

$$p_j(\boldsymbol{x}) = \mathbb{P}[Y = j|\boldsymbol{X}] = \frac{e^{\beta_{0j}+\sum_{i=1}^{n}\beta_{ij}X_i}}{1+\sum_{l=1}^{J-1} e^{\beta_{0l}+\sum_{i=0}^{n}\beta_{il}X_i}} \qquad (13)$$

When the model predicts the outcome the class with the highest predicted probability is chosen as the predicted class. This process allows logistic regression to be extended to handle multiple classes of data classification effectively.

4. Experimental Results

The data was taken from a LUCAS Soil Database in 2018. Both basic soil properties and bulk density were used in this data analysis. Only for instances where the bulk density was used in the analysis, the merged data frame was used. Based on all the previous data. The aim was to answer research questions (*RQ1 – RQ6*).

4.1. Differences in Soil Properties Distribution for Different Land Cover Types

The purpose of this part is to answer if there is any difference in soil properties between different land cover types (*RQ1*). It is important to mention that this analysis only analyse the land cover types *woodland*, *cropland*, *grassland* and *bare land,* because the rest of the land cover types, including *Shrubland*, *Artificial land*, *Wetland* and *Water*, do not have enough data points for analysis.

To explore if there are any significant differences in soil properties for land cover types, Kruskal-Wallis statistical test is performed, and data is shown in the Table 4.1. Considering small p-values, somewhere as small as 0(*), the differences are investigated more in-depth, with post-hoc Conover's tests. In this part, only the Conover's tests with noteworthy results are explored.

Table 4.1 Results of Kruskal Wallis test for different land cover types

property	pH_CaCl2	pH_H2O	CaCO3	EC
statisic	6877.241	6941.469	1990.063	1543.069
p-value	0.000*	0.000*	0.000*	0.000*
Hypothesis	Ha	Hb	Hc	Hd
Conclusion	Reject	Reject	Reject	Reject
property	K	N	OC	P
statisic	2597.263	1763.355	3753.436	473.035

p-value	0.000*	0.000*	0.000*	3.32E-102
Hypothesis	He	Hf	Hg	Hh
Conclusion	Reject	Reject	Reject	Reject

Figure 4.1 shows the results for Conover's test pairwise comparisons for electrical conductivity. The notable part is the value for a pair of *cropland* and *bareland* types with the p-value of 0.068, indicating notable similarity between types with the significance level of 0.05.

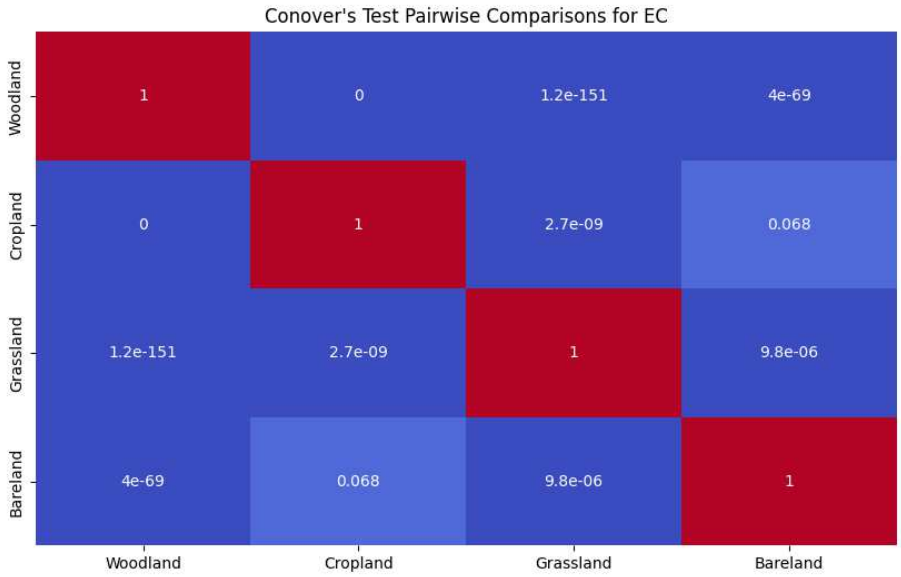

Figure 4.1 Conover's test pairwise comparisons for EC

Examining the results of Conover's test in Figure 4.2, small values between almost all pairs are seen, confirming the differences between the pairs. Pair of *grassland* and *bareland* being the outliers, with the p-value of 0.34 showing moderate similarity between the land cover types.

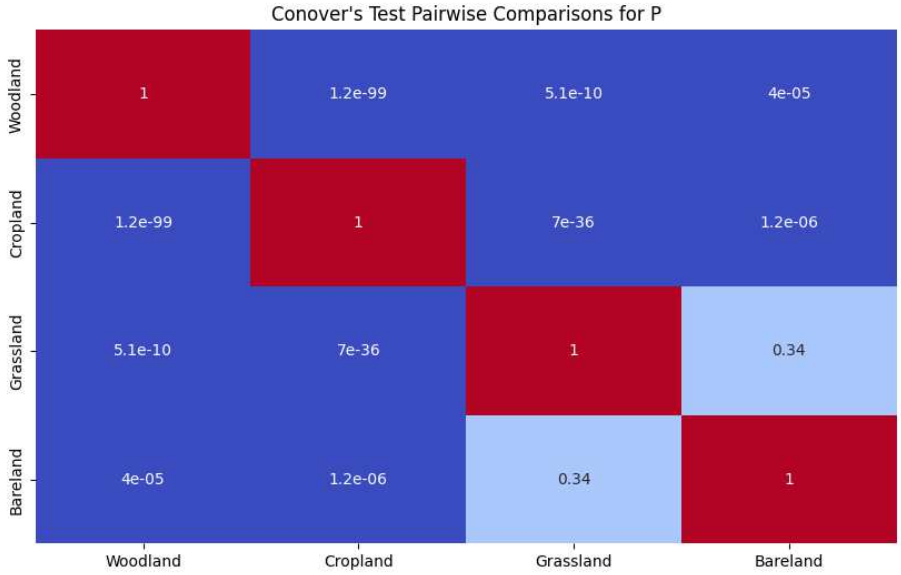

Figure 4.2 Conover's test pairwise comparisons for total

Looking at the results in Figure 4.3, small p-values between most of the pairs are observed, confirming the significant differences between land cover types. With p-values of 0.32, pair of *bareland* and *cropland* stand out, showing lack of significant difference and suggesting moderate similarity between the groups for measure of extractable potassium.

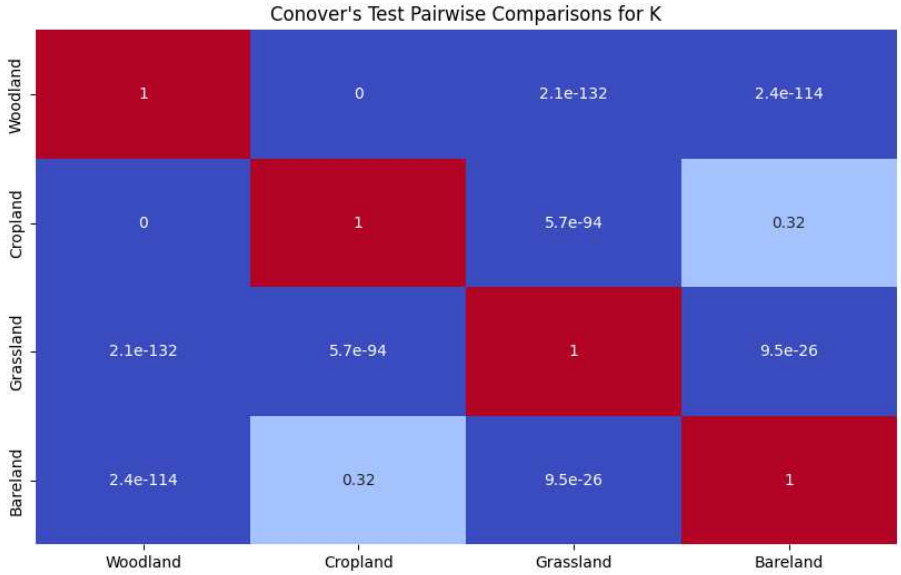

Figure 4.3 Conover's test pairwise comparisons for extractable potassium

4.2. Differences in Soil Properties Distribution for Different Geographical Regions

This part aims to explore differences in soil properties in different geographical regions (*RQ2)*. To divide the data into different regions we have used the geographical division, meaning data set has: the UK, Ireland, Luxembourg, Netherlands, Belgium, and France in Western Europe, Denmark, Sweden, Finland, Lithuania, Estonia, and Latvia in Northern Europe, Germany, Poland, Austria, Czech Republic, Slovenia, Croatia, Hungary, and Slovakia in Central Europe, Portugal, Spain, Italy, Cyprus, and Greece in Southern Europe, and Bulgaria, Romania and Malta in Southeastern Europe.

To explore if there are any significant differences in soil properties trough geographical regions of Europe, Kruskal-Wallis statistical test is performed, and the following results are presented in the Table 4.2. Based on the low p-values for each of the soil properties, it can be concluded that there are differences in each soil property trough different geographical regions of Europe. Considering quite small p-values, it was decided to explore the differences between soil properties in each region a bit further with post-hoc Conover's test. Next, some of the more interesting findings with the Conover's test will be explored.

Table 4.2 Results of Kruskal Wallis test for different geographical regions

property	pH_CaCl2	pH_H2O	CaCO3	EC	BD 0-10
statisic	6007.982	5725.906	2598.151	1673.47	469.708
p-value	0.000*	0.000*	0.000*	2.219E-284	2.381E-100
Hypothesis	Hi	Hj	Hk	Hl	Hm
Conclusion	Reject	Reject	Reject	Reject	Reject
property	K	N	OC	P	BD 10-20
statisic	2612.642	1037.538	1344.514	808.526	274.0535
p-value	0.000*	2.614E-223	7.425E-290	1.093E-173	3.359E-58
Hypothesis	Hn	Ho	Hp	Hr	Hs
Conclusion	Reject	Reject	Reject	Reject	Reject

Firstly, regarding the pairwise comparison for pH measured in calcium chloride shown in the Figure 4.4, it can be seen that for most of the pairs the p-values is quite low, confirming significant difference between the regions. The standout p-value is the one between

Southeast and Western Europe, with the p-values of 0.43, indicating no significant difference between regions.

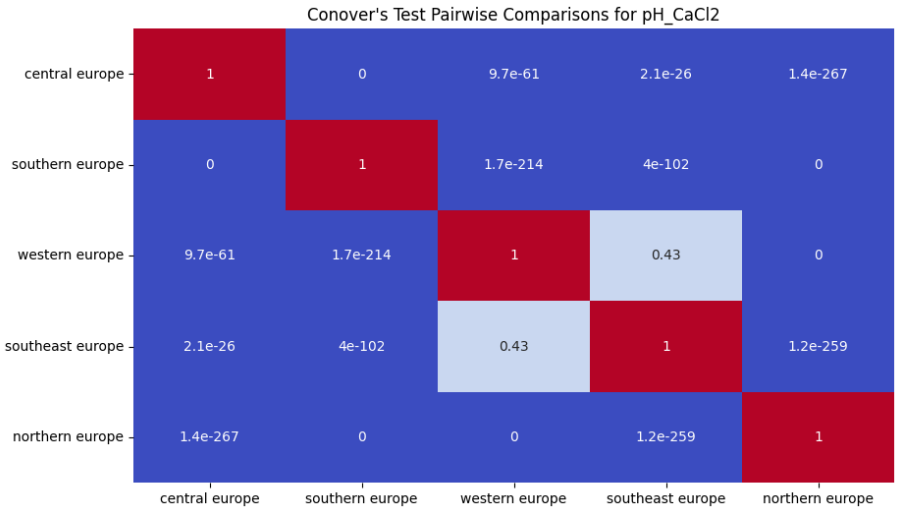

Figure 4.4 Conover's test pairwise comparison for pH measured in calcium chloride

Secondly, regarding the results of the Conover's test for pH measured in water (Figure 4.5), it can be seen that most pairs have a small p-values, confirming that there is a significant difference between pairs. We also have a standout p-value of 0.086 between Southeast and Western Europe, showing no significant difference between regions.

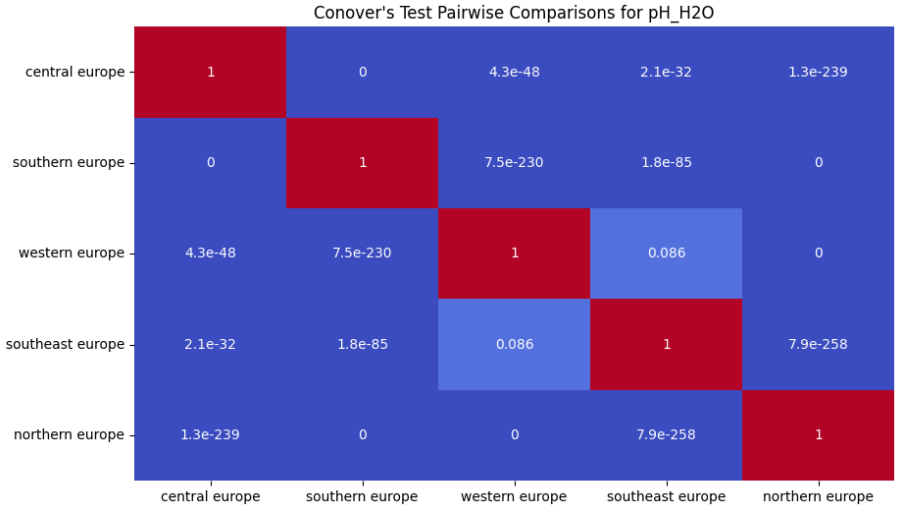

Figure 4.5 Conover's test pairwise comparison for pH measured in calcium chloride

Continuing, regarding the Conover's test for pairs based on the electrical conductivity, shown in Figure 4.6, the small values for all pairs can be seen, except for the pair of Western and Southern Europe with the p-value of 0.84. Where values lower than 0.05 indicate significant difference between pairs, and values higher than 0.05 indicating no significant difference between pairs.

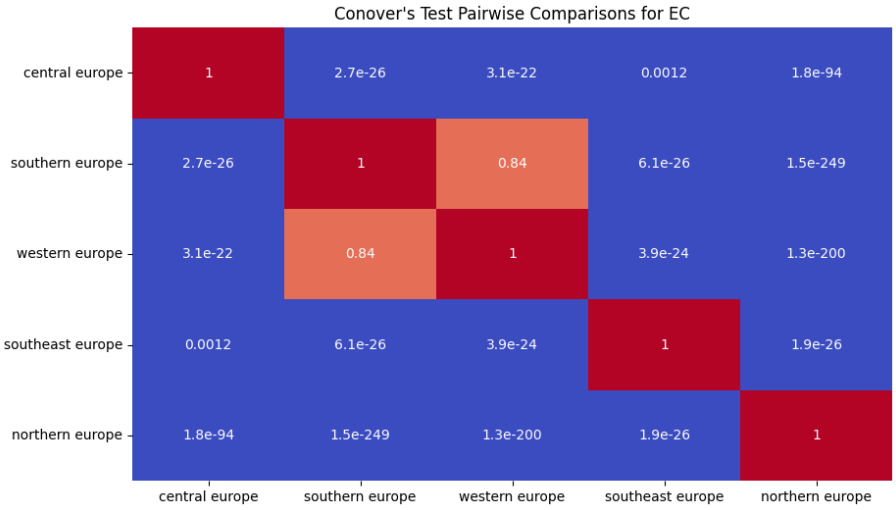

Figure 4.6 Conover's test pairwise comparison for electrical conductivity

Regarding the results of Conover's test for pairs depending on the organic carbon content of soil (Figure 4.7), it is noticed that most pairs have a low p-values indicating significant differences between pairs, excluding the p-value od 0.089 between pairs of Southeast and Central Europe, with the values indicating no significant difference between pairs.

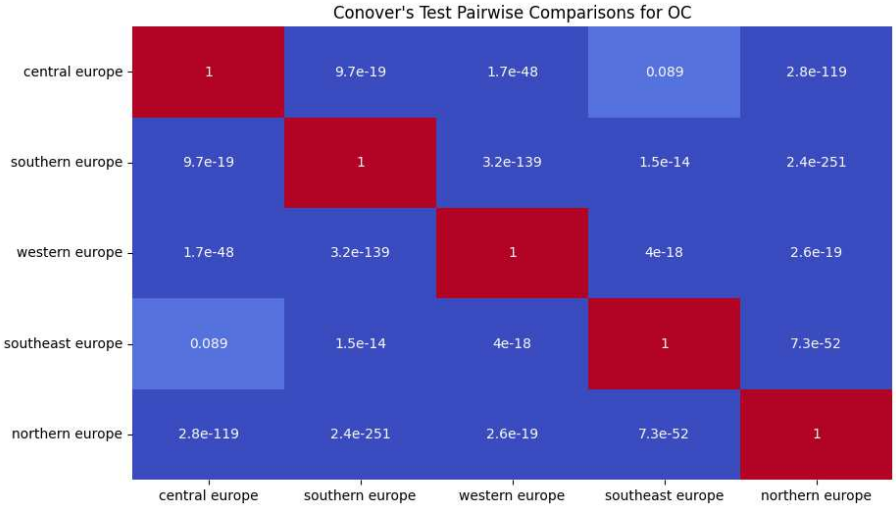

Figure 4.7 Conover's test pairwise comparison for organic carbon content

Looking at the results for the pairs depending on the calcium carbonate content, from Conover's test (Figure 4.8), it can be seen once again that many p-values are lower than 0.05 indicating significant differences between pairs in terms of the calcium carbonate content in soil. The p-value of 0.13 stands out for the pair of Southeast and Central Europe, indicating no significant difference in soils calcium carbonate content.

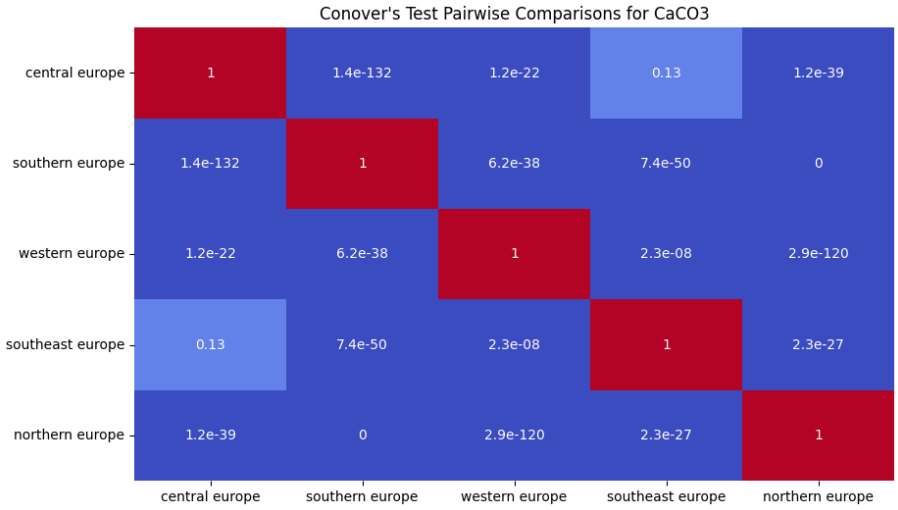

Figure 4.8 Conover's test pairwise comparison for calcium carbonate content

In terms of Conover's test for total phosphorus in soil (Photos 4.9), it can be seen that most of the values indicate no significant difference between pairs, having values lower than 0.05, except for the pair of Western and Central Europe. Pair of Western and Central Europe has a p-value of 0.046 indicating sufficient evidence to say that there is a significant difference between pairs, but with a lower level of significance contrary could be concluded.

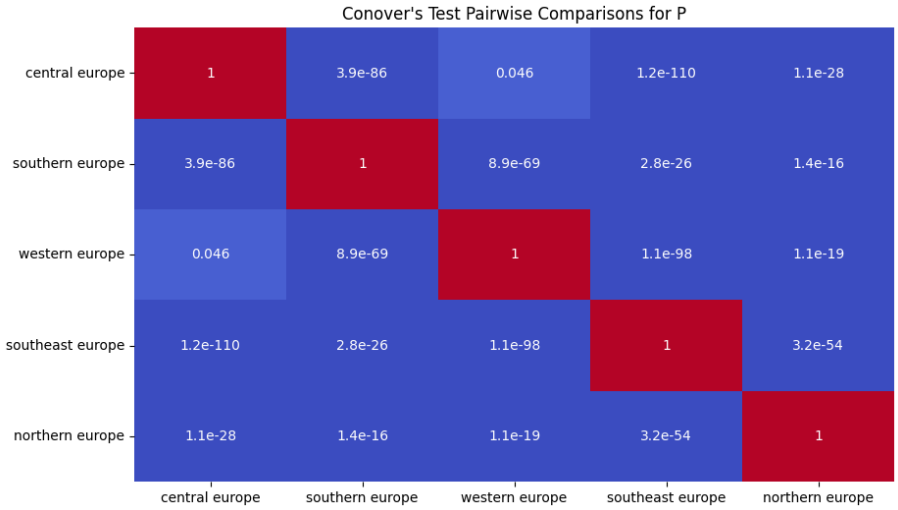

Figure 4.9 Conover's test pairwise comparison for total phosphorus

Examining Figure 4.10, results of Conover's test pairwise comparison for total nitrogen are explored. We see that most of pairs have small p-values indicating significant difference between pairs, excluding the pair of Southeast and Central Europe with the p-value of 0.088, showing that there is no significant difference between Southeast and Central Europe in terms of total nitrogen in soil. Also, the pair of Southeast and Northern Europe has a p-value of 0.13, also indicating that there's no significant difference between regions.

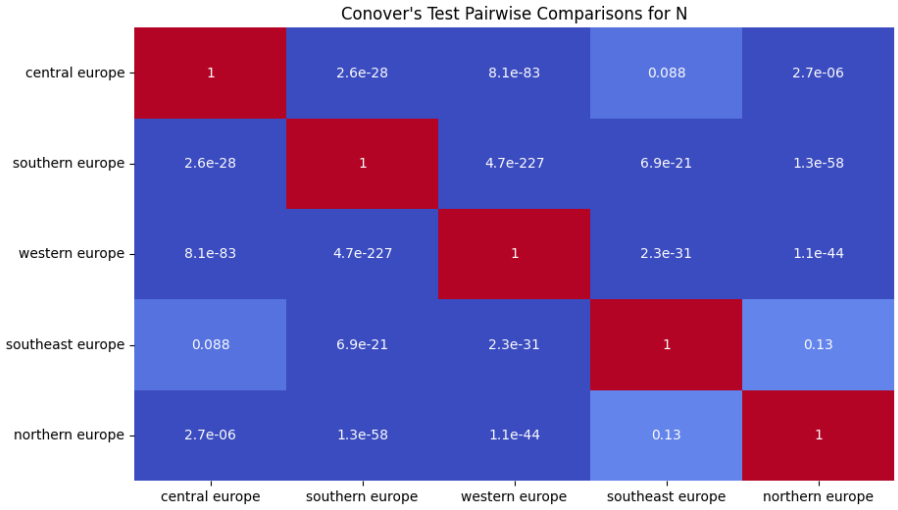

Figure 4.10 Conover's test pairwise comparison for total nitrogen

Lastly, Figure 4.11 shows the results of Conover's test for bulk density in depths of 10-20cm. It can be seen that regions of Southeast and Northern Europe have low p-values for each of their parings, showing significant differences in bulk density between those regions all other regions. Looking at values between Western, Southern and Central Europe we can see that there is no significant difference between those regions for the significance level of 0.05. For pairs Central and Southern Europe having a p-value of 0.54, Central and Western Europe having a p-value of 0.38.

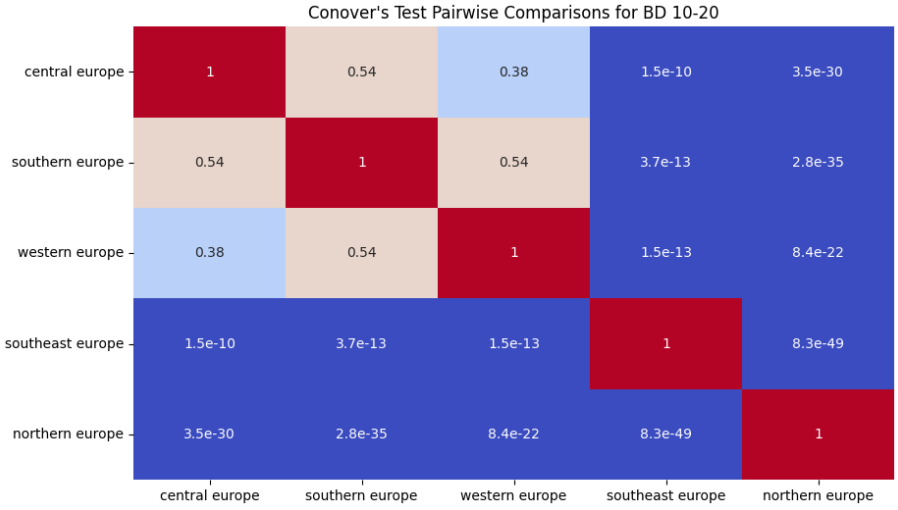

Figure 4.11 Conover's test pairwise comparison for bulk density 10-20cm depth

4.3. Predicting Electrical Conductivity of Soil Based on Soil Properties and Geographical Data

To predict electrical conductivity depending on the soil properties, few models, varying with levels of success were tried out, including K-Neighbours Regressor model, Radius Neighbours Regressor model, Ridge Regression and Random Forest Regression. First two models were only used with geographical data, and the other two with soil properties and geographical data.

K-Neighbours Regressor model, which uses only geographical data, has an MSE of 0.296 and R^2 value of 0.250, meaning that approximately 25.0% of the variance in the electrical conductivity can be explained by geographical features alone. This relatively low R-squared value suggests that geographical data alone might not be sufficient for accurately predicting electrical conductivity. To make model's predictions better we predicted logarithmic values of EC, which made the R^2 much higher than with ordinary EC.

Similarly, Radius Neighbours Regressor model, also utilizing only geographical data, performs slightly worse than the K-Neighbours model with an MSE of 0.276 and R^2 value of 0.301, indicating that the radius-based approach might capture the underlying patterns in data a little more effectively than the K-Neighbours model. Here too we used logarithmic values of electrical conductivity to make the model better.

Ridge Regression model, utilizing both soil properties and geographical data, has an R^2 value of 0.446 and mean squared value of 33.165. R^2 indicates that approximately 44.6% of the variance in electrical conductivity can be explained by the combined effects of soil properties and geographical features. The coefficients provide insight into the relationship between each feature and the electrical conductivity, considering the regularization effect imposed by Ridge Regression, coefficients can be seen in the Table 3.3.

Table 4.3 Coefficients of features in Ridge Regression for electrical conductivity

Feature	OC	$CaCO_3$	N	K	TH_LAT	TH_ LONG	Intercept
Coefficient	-3.481	1.747	5.436	1.959	-0.650	0.362	14.402

The Random Forest Regression model, which incorporates both soil properties and geographical data, achieves mean square error of 26.902 and a higher R^2 of 0.551. For predictions in this model, the best R^2 measure was achieved using the organic carbon content,

calcium carbonate content, total nitrogen, extractable potassium, and geospatial data (longitude and latitude). This suggests that the Random Forest model provides a much better fit to the data than other models.

4.4. Predicting pH Measured in Calcium Chloride of Soil Based on Soil Properties and Geographical Data

In predicting pH measured in calcium chloride, the same approach to answer the research question *RQ3* was used. Inspecting the predictions with K-Neighbours Regressor model, Radius Neighbours Regressor model, Ridge Regression and Random Forest Regressor model. In the first two we used only geospatial data for predictions, and geospatial data and soil features in second two approaches.

In K-Neighbours Regressor model, using only geospatial data, value of mean square error is 0.762 and R-squared is 0.622, meaning the model predicts about 62.2% of variance in pH measured in calcium chloride by only geospatial data.

Using Radius Neighbours model, the value of mean squared error is 0.781, and R^2 value is 0.612. These values indicate that the model captures only about 61.2% of the actual pH values, meaning our model might not be capturing the whole picture.

With Ridge Regression for predicting pH measured in calcium chloride, this time not using only geospatial data, but also soil properties, the value of R-squared is 0.639, and MSE is 0.673. R^2 value shows that the model correctly predicts about 63.9% of variance of pH. The coefficients provide us with an insight into impact of each feature on pH measured in calcium chloride, the specific coefficients can be seen in the Table 4.4, although they are regulated by the standard scaler necessary for Ridge Regression.

Table 4.4 Coefficients of features in Ridge Regression for pH measured in $CaCl_2$

Feature	OC	$CaCO_3$	P	N	K	TH_LAT	TH_ LONG	Intercept
Coefficient	-0.880	0.347	0.008	0.560	0.415	-0.287	0.043	6.095

Lastly, Random Forest Regressor model was implemented, with the aim of predicting pH measured in calcium chloride based on organic carbon content, calcium carbonate content, total nitogren, extractable potassium, and geospatial data (longitude and latitude). With this model R^2 value is 0.862 and mean MSE is 0.269, meaning the model predicts about 86.2%

in variance of pH. This makes this model best fit for predicting our pH measured in calcium chloride from our features.

4.5. Predicting pH Measured in Water of Soil Based on Soil Properties and Geographical Data

Trying to predict values of pH measured in water, several models were used, including K-Neighbours Regressor model, Radius Neighbours Regressor model, Ridge Regression and Random Forest Regressor model. In predicting for first two only geospatial data was used, but in the second two both soil features and geospatial data for predicting pH values was used.

Using the K-Neighbours Regressor model, the values for MSE is 0.691 and for R^2 is 0.613, meaning that with only spatial data we can predict about 61.3% values correctly. In other instance of using only geospatial data – Radius Neighbours Regressor we get the values of 0.700 as mean square error and 0.607 as R-squared, indicating that this model predicts only about 60.7% of variance in pH measured in water.

Using Ridge Regression, the values of 0.630 and 0.615, for R-square and mean squared error respectively are received. R-squared error indicates that the model predicts about 61.3% of variance. In Table 4.5 we can see the coefficients but scaled by the standard scaler required for the Ridge Regressor model.

Table 4.5 Coefficients of features in Ridge Regression for pH measured in water

Feature	OC	$CaCO_3$	P	N	K	TH_LAT	TH_ LONG	Intercept
Coefficient	-0.774	0.359	0.003	0.420	0.407	-0.241	0.0387	6.611

In the end, the implementation with Random Forest Regressor for predicting the pH values measured in water was tried out. As features for predicting we got the best values using organic carbon content, calcium carbonate content, total nitrogen, extractable potassium, and geospatial data (longitude and latitude) as features. We got the value R^2 of 0.870 and MSE value of 0.231. This model then predicts about 87.0% in variance of pH measured in water. This makes this model the best one made for predicting pH.

4.6. Predicting Land Cover Type Based on Soil Properties

The logistic regression model was developed to predict land cover type based on soil properties. In Table 4.6, the performance metric of the model can be seen. An overall accuracy of this model is 78.2%. This model, which considers organic carbon content, calcium carbon content, total nitrogen, and extractable potassium as predictors, effectively discriminates between *cropland*, *grassland,* and *woodland.* The model exhibits high precision of 0.79 and recall of 0.92 for *cropland*, indicating its robustness in correctly identifying this land cover type. For *woodland*, recall and precision values are also quite good at 0.83 and 0.81, respectively. On the other side, model's performance is less optimal for grassland, with a precision of 0.56 and recall of 0.30, suggesting room for improvement in this category. Still, despite this, the macro-average F1-score of 0.78 and the weighted-average F1-score of 0.76 highlight the model's overall efficiency. These results explain the potential of using soil properties in logistic regression models to predict land cover types.

Table 4.6 Performance metrics of the logarithmic regression model for predictions of land cover type

	Precision	Recall	f1-score	Support
Cropland	0.79	0.92	0.85	466
Grassland	0.56	0.30	0.39	149
Woodland	0.83	0.81	0.82	313
Accuracy			0.78	928
Macro average	0.73	0.68	0.69	928
Weighted average	0.77	0.78	0.76	928

5. Appendix

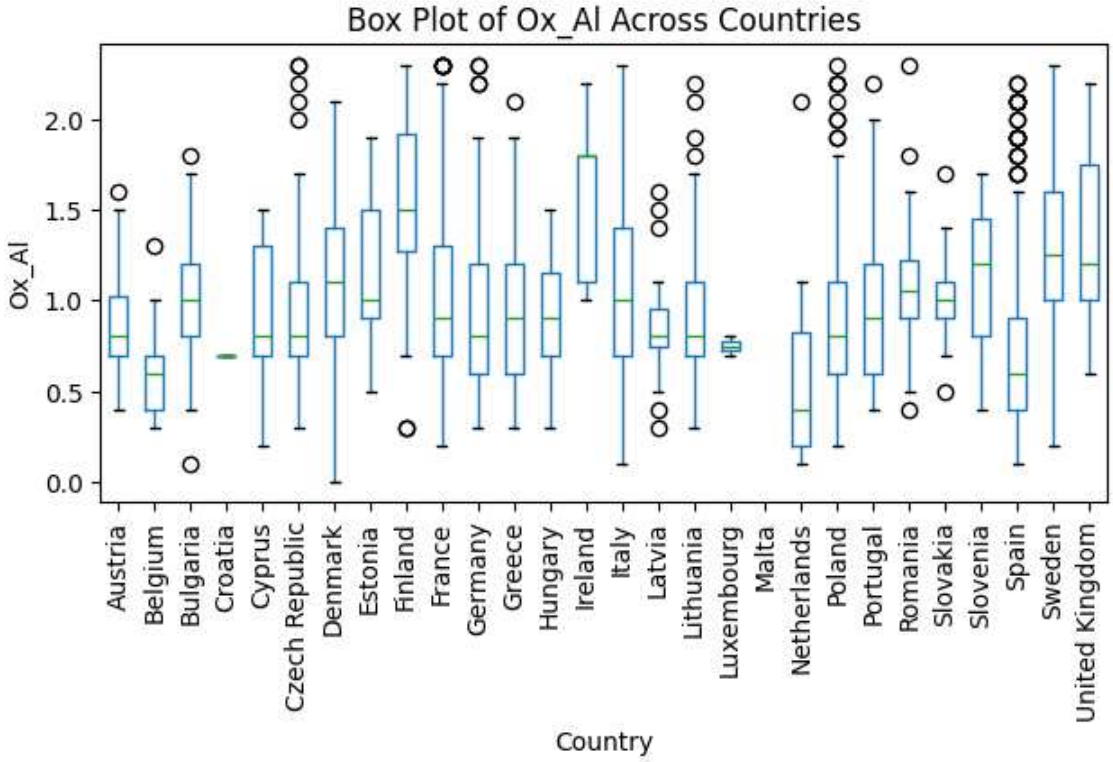

Figure 5.1 Box plots of oxalate extractable iron across countries

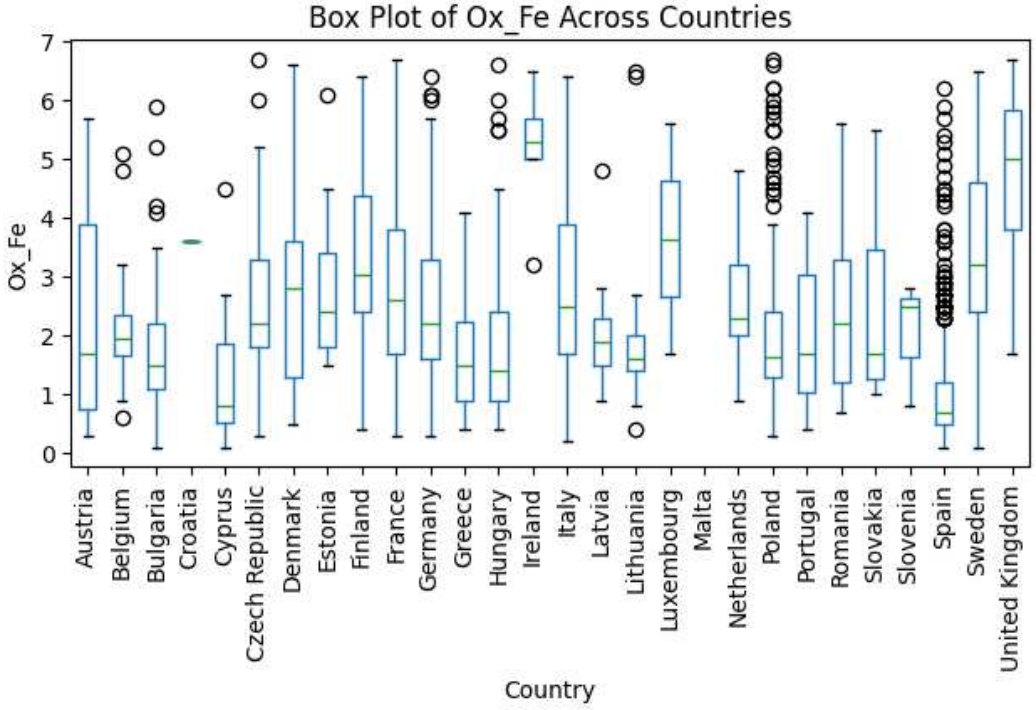

Figure 5.2 Box plots of oxalate extractable iron across countries

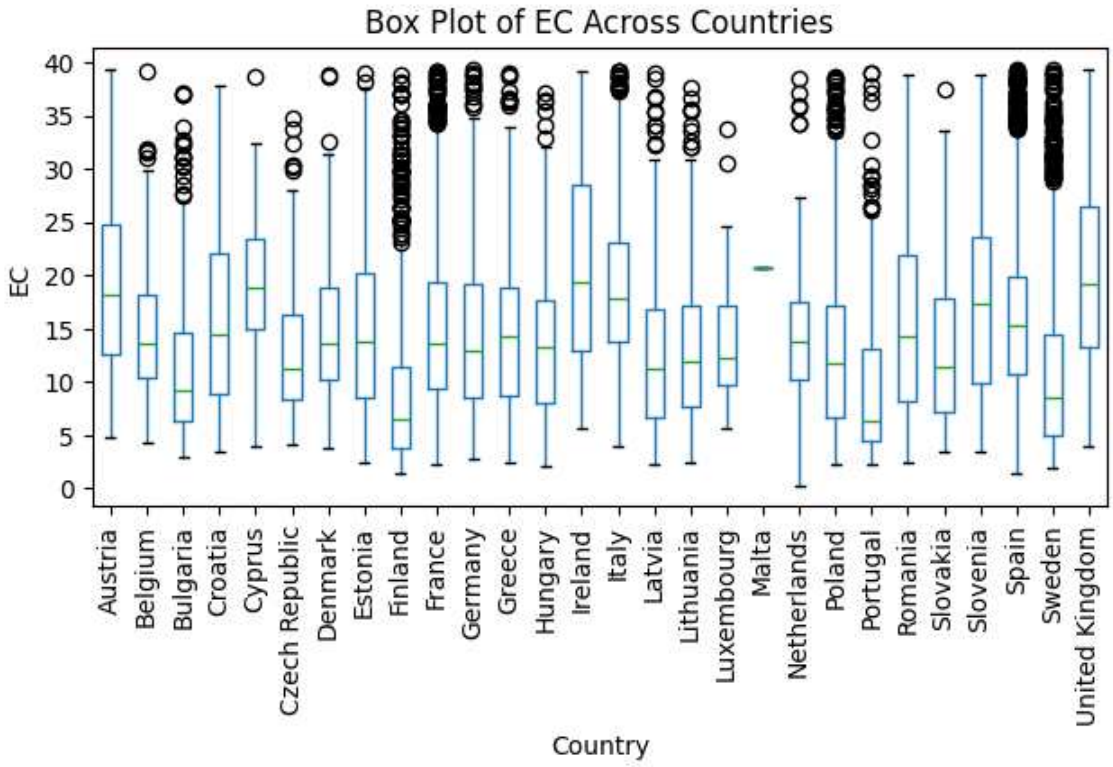

Figure 5.3 Box plots of electrical conductivity across countries

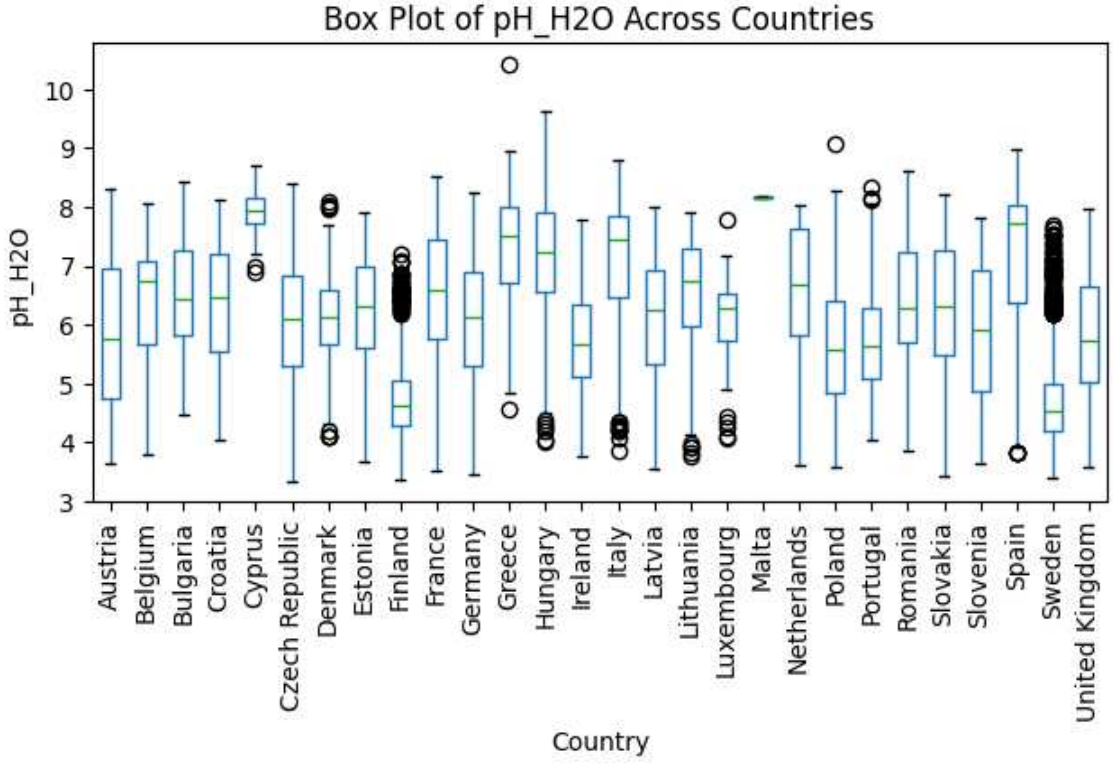

Figure 5.4 Box plots of pH measured in calcium chloride across countries

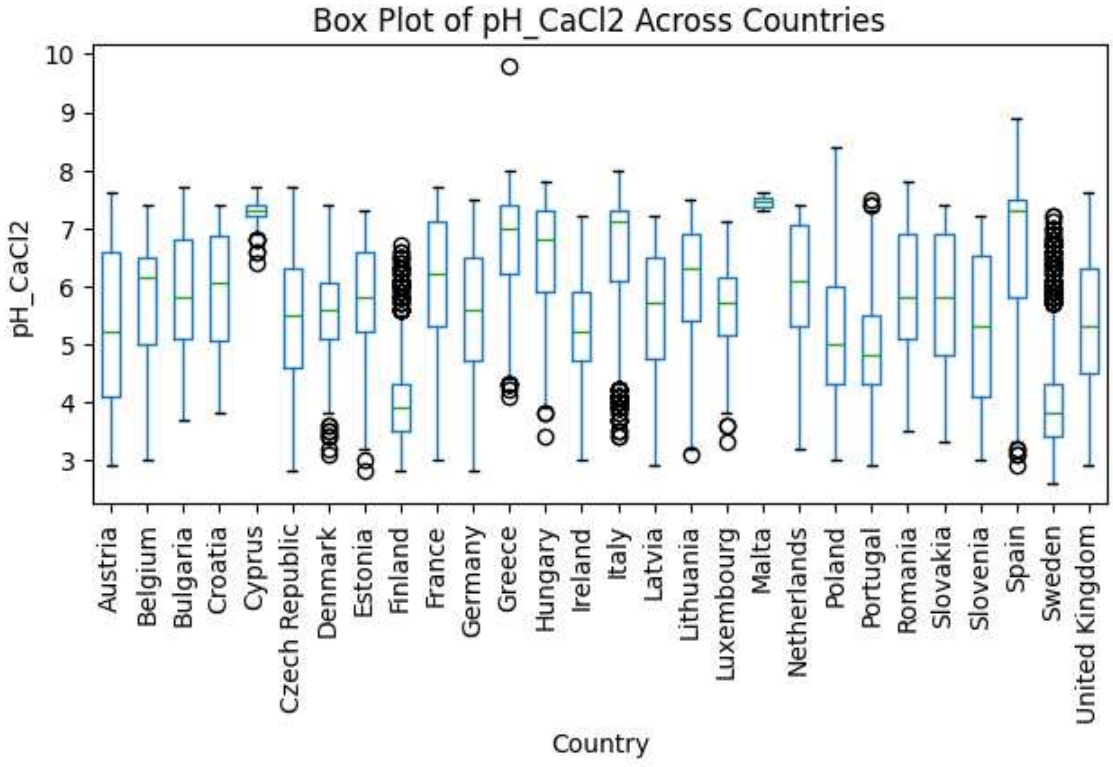

Figure 5.5 Box plots of pH measured in calcium chloride across countries

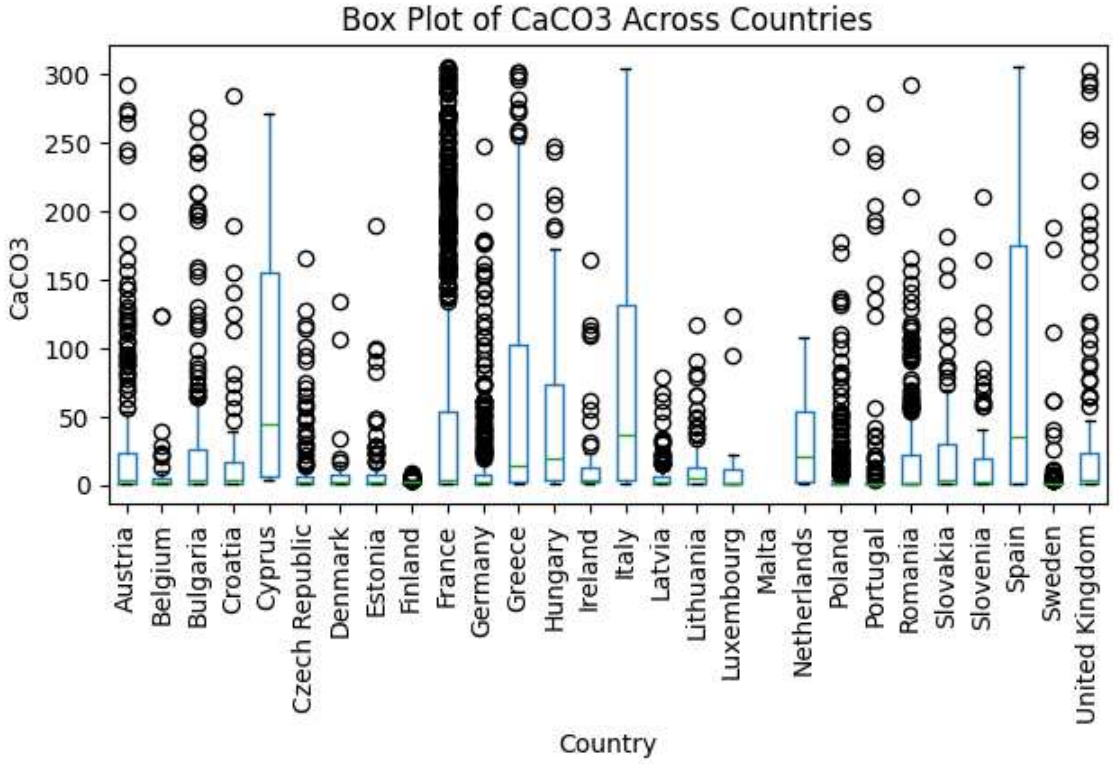

Figure 5.6 Box plots of extractable potassium across countries

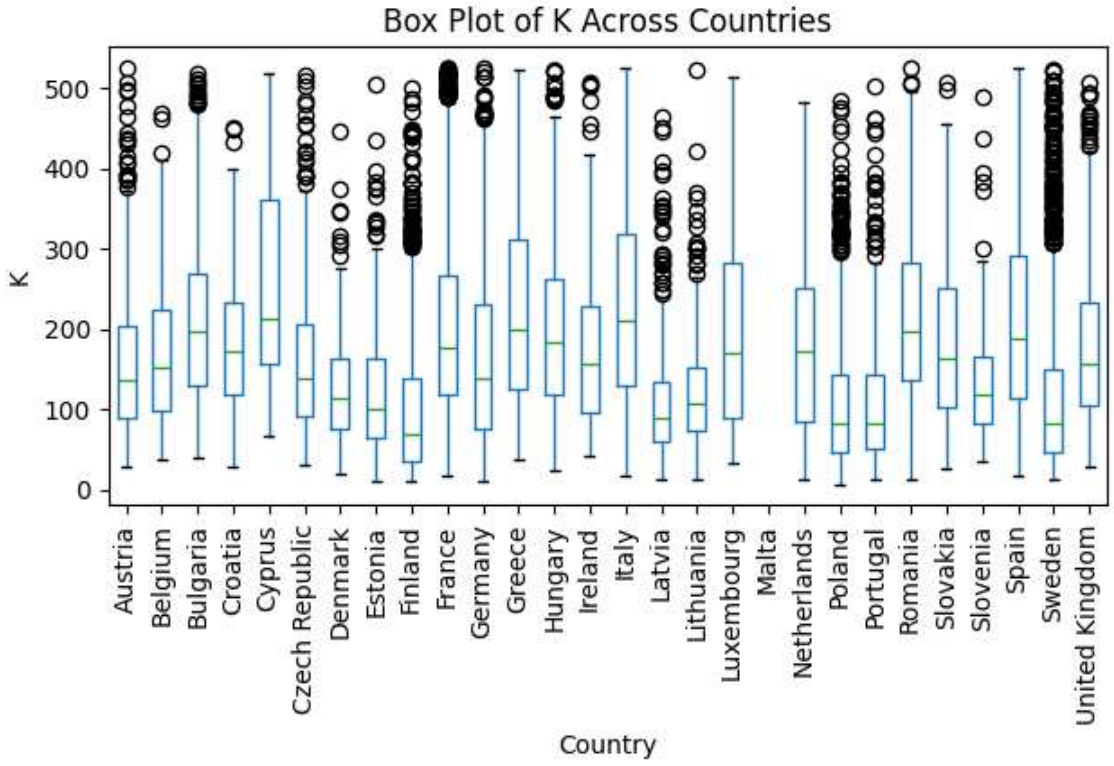

Figure 5.7 Box plots of extractable potassium across countries

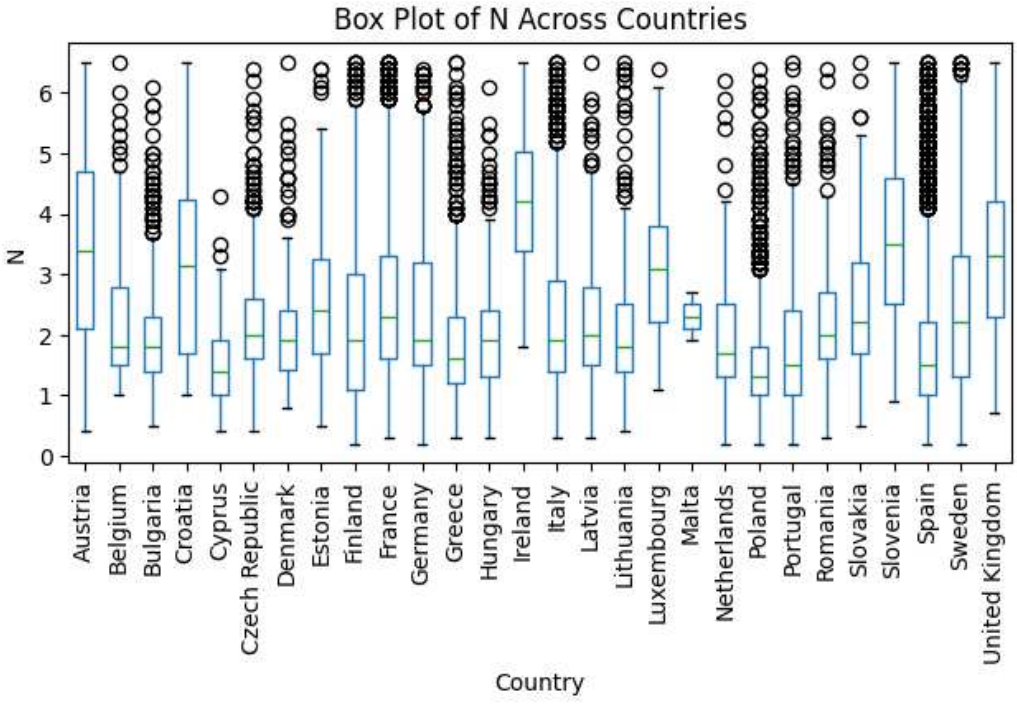

Figure 5.8 Box plots of total nitrogen across countries

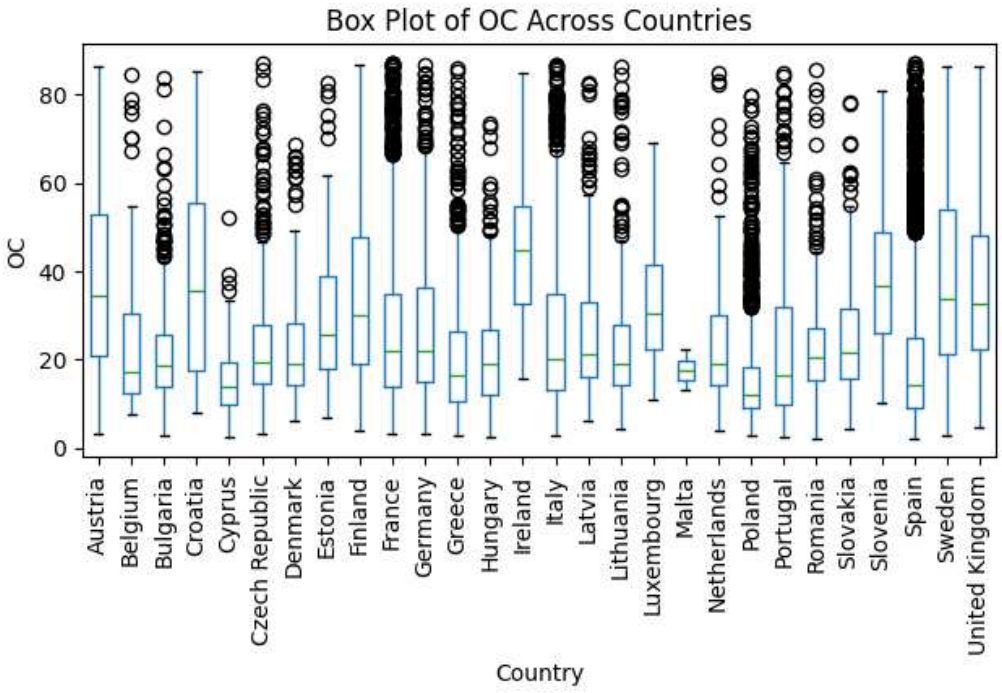

Figure 5.9 Box plots of organic carbon content across countries

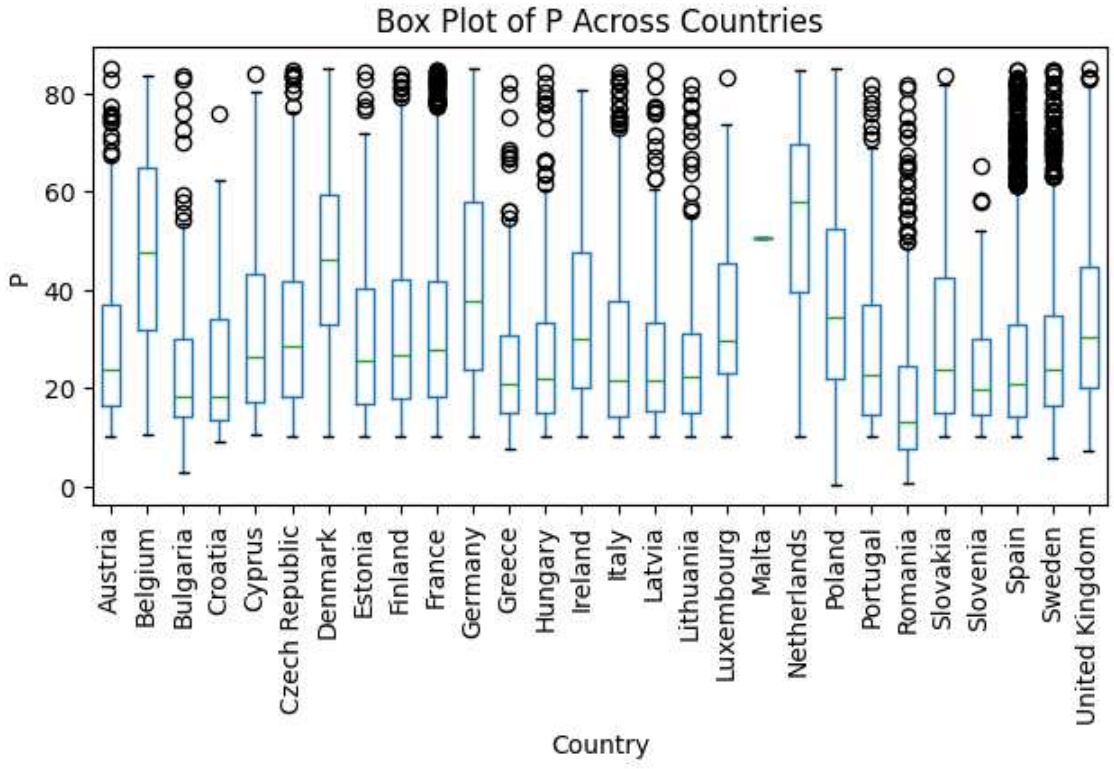

Figure 5.10 Box plots of total phosphorus across countries

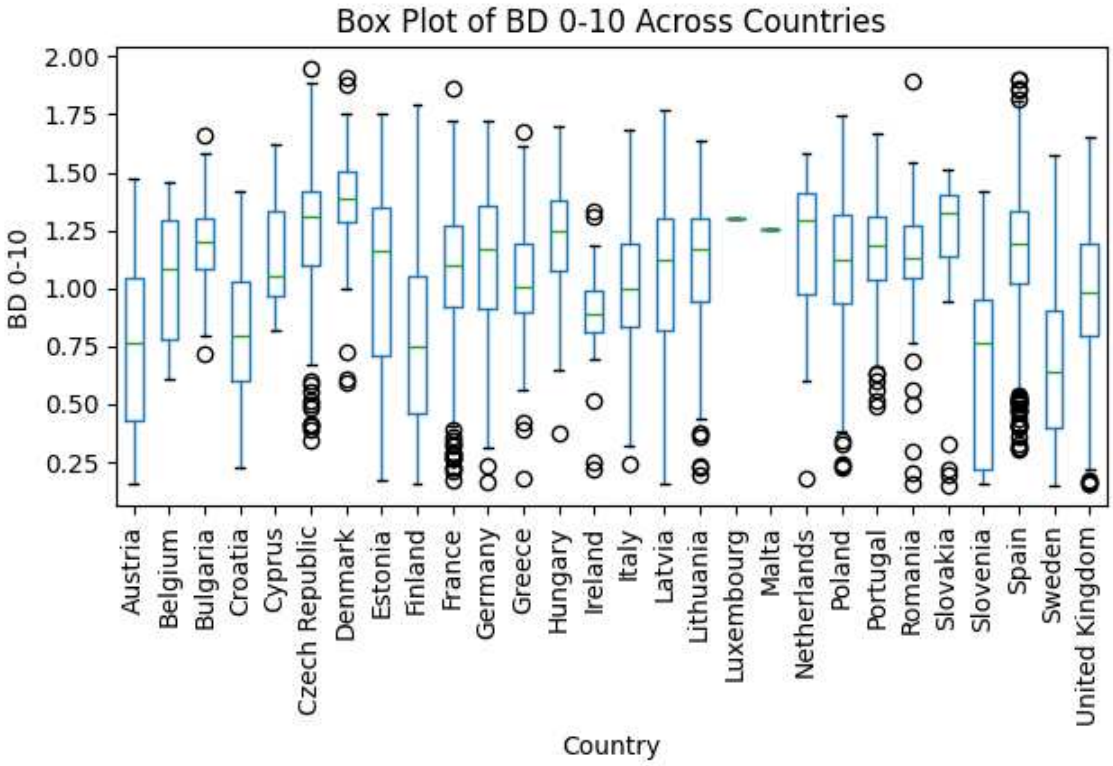

Figure 5.11 Box plots of bulk density countries for depth 0-10cm across countries

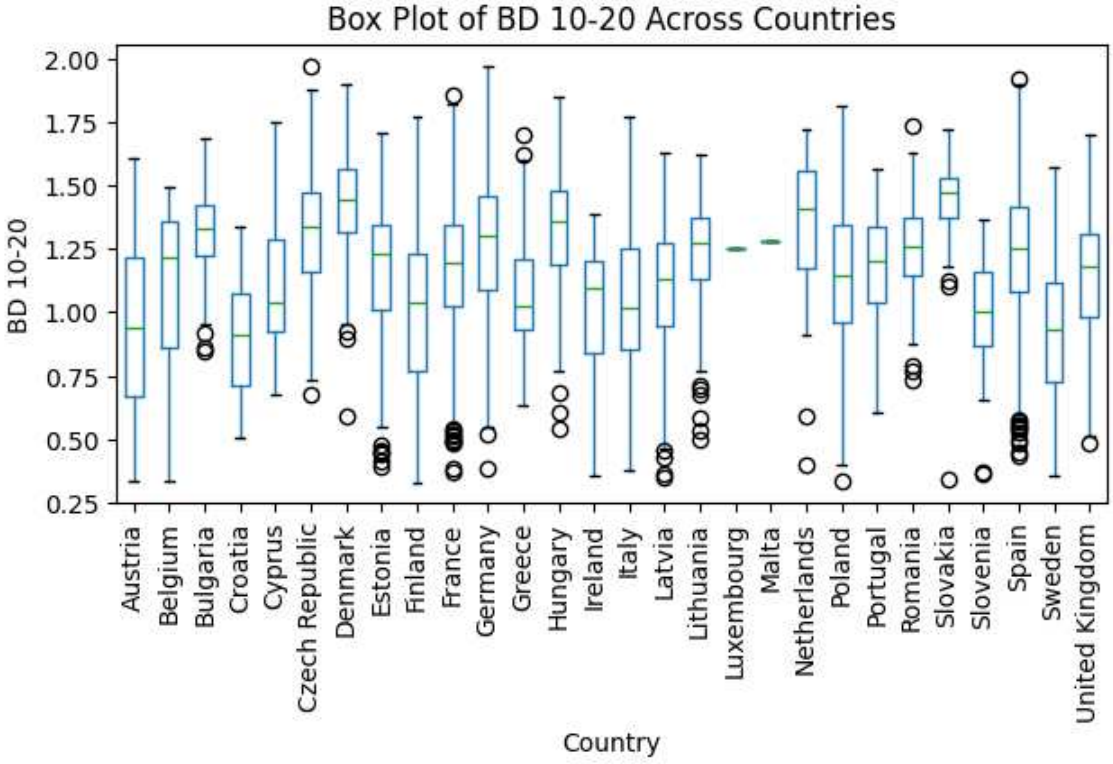

Figure 5.12 Box plots of bulk density countries for depth 10-20cm across countries

Conclusion

All in all, our comprehensive study sheds light on the relationships between soil properties, geographical location, and land cover types. Our findings have been supported by application of statistical tests such as Kruskal Wallis test, along with regression techniques. Kruskal Wallis test was used to explore if there are differences in soil properties across different land cover types and different geographical regions.

Our findings are based on the application of statistical models and regression techniques. These methods, including K-Neighbours Regressor, Radius Neighbours Regressor, Ridge Regression, and Random Forest Regression models, have enabled us to generate predictions about these relationships. Each model provides us unique insight into understanding of the soil properties and their impact on land cover types and pH values.

Outcomes of our research have potential implications for various fields, including environmental science, land management, and agricultural practices. Insight gained could serve as a foundation for further research, contributing to the development of effective policies related to land management and environmental conversion. By understanding the nature of our soil and how it interacts with different geographical features and land cover types, we can work towards more suitable agricultural practices and better land use planning.

However, it is important to note that our study is not without its limits. While our models have been effective in predicting certain relationships, there is always a degree of uncertainty and room for improvement. Future research could focus on refining these models, including more diverse soil properties, or investigating different geographical regions. Further, it would be beneficial to integrate more contextual factors, such as climate data, to provide a more comprehensive understanding of the complex dynamics at play.

Overall, our research represents a stepping stone in understanding the intricate relationships between soil properties, geographical data, and land cover types. We hope that the insights gained from this study will spark further research, leading to more sustainable and informed decisions for our planet’s future.

Refences

1. Fernandez-Ugalde, O., Scarpa, S., Orgiazzi, A., Panagos P., Van Liedekerke, M., Marechal A. and Jones, A., 2022, *LUCAS 2018 Soil Module, Presentation of dataset and results.* European Commision
2. Ballabioa C., Lugatoa E., Fernández-Ugaldea O., Orgiazzia A., Jonesa A., Borrellib P., Montanarellaa P., Panagosa P., *Mapping LUCAS topsoil chemical properties at* European scale using Gaussian process regression, Geoderma, Vol. 355, (2019)
3. Ballabio C., Panagos P, Monatanarella L., *Mapping topsoil physical properties at European scare using the LUCAS database*, Geoderma, Vol. 261, (2015), pg. 110-123
4. Ballabio C., Panagos P., Lugato E., Huang b J., Orgiazzi A., Jones A., Fernández-Ugalde O., Borrelli P., Montanarella L., *Copper distribution in European topsoils: An assessment based on LUCAS soil survey,* Geoderma, Vol. 636, (2018), pg. 282-298
5. Panagos P., Meusburger K., Ballabio C., Borrelli P., Alewell C., *Soil erodibility in Europe: A high-resolution dataset based on LUCAS*, Geoderma, Vol. 479-480(2014), pg. 189-200
6. Gao Y., Liu L. Zhang X., Chen X., Mi J., Xie S., *Consistency Analysis and Accuracy Assessment of Three Global 30-m Land-Cover Products over the European Union Using the LUCAS Dataset,* Remote Sens., Vol. 12, (2020)
7. Karydas C. G., Gitas I. Z., Kuntz S., Minakou C., *Use of LUCAS LC Point Database for Validating Country-Scale Land Cover Maps*, Remote Sens., Vol. 7, (2015), pg. 5012-5041
8. Weigand M., Staab J., Wurm M., Taubenböck H., *Spatial and semantic effects of LUCAS samples on fully automated land use/land cover classification in high-resolution Sentinel-2 data,* Elsevier, Vol. 88, (2020), 102065
9. Martinez-Sanchez L., See L., Yordanov M., Verhegghen A., Elvekjaer N., Muraro D., d'Antrimont R., van der Velde M., *Autiomatic classification of land cover from LUCAS in-situ landscape photos using semantic segmentation and a Random Forest model*, Elsevier, Vol. 172, (2024), 105931

10. Kliment M., Kočica J., Kliment T., *Land Use Dataset Collection And Publiction Based on LUCAS and HILUCS*, Acta Horticulturae et Regiotecturae, Vol. 17, (2014), Issue 2, pg. 52-59

11. d'Andrimont R., Yordanov M., Martinez-Sanchez L., Eiselt B., Palmieri A., Dominici P., Gallego J., Reutner H. I., Joebges C., Lemoine G., van der Velde M., *Harmonised LUCAS in-situ land cover and use dataset for field surveys from 2006 and 2018 in the European Union,* Scientific Data, Vol. 7, (2020), article num. 352

12. Walpole R. E., Myers R. H., Myers S. L., Ye K., *Probability & Statistics for Engineers & Scentists*, Ninth Edition, (2012), pg. 668-669

13. Pohlert T., *The Pairwise Multiple Comparison of Mean Ranks Package(PMCMR)*, (2016), link: http://www.dpye.iimas.unam.mx/soriano/Estadistica2_2020_1/DOCUMENTOS/NOTAS/PMCMR.pdf, accessed: 1st of July

14. Walpole R. E., Myers R. H., Myers S. L., Ye K., *Probability & Statistics for Engineers & Scentists*, Ninth Edition, (2012), pg. 407

15. Frost J., *Mean Squared Error (MSE)*, (2024), link: https://statisticsbyjim.com/regression/mean-squared-error-mse/, accessed: 1st of July 2024

16. LaViale T., *Deep Dive on KNN: Understanding and Implementing the K-Nearest Neighbors Algorithm*, Arize, (2023, March), link: https://arize.com/blog-course/knn-algorithm-k-nearest-neighbor/#:~:text=KNN%20works%20in%20three%20main,value%20of%20the%20neighbors%2C%20respectively; accessed: 1st of July, 2024

17. Prakhar7, *Haversine formula to find distance between two points on a sphere*, GeeksForGeeks, (2022, September), link: https://www.geeksforgeeks.org/haversine-formula-to-find-distance-between-two-points-on-a-sphere/, accessed: 1st of July, 2024

18. Angela and Kezhan Shi, *Nearest Neigbours Regressors – A Visual Guide*, (2023, March), link: https://towardsdatascience.com/nearest-neighbors-regressors-a-visual-guide-78595b78072e, accessed: 1st of July 2024

19. Murel J., *What is ridge regression?*, (2023, November), link: https://www.ibm.com/topics/ridge-regression , accessed: 1st of July 2024

20. Sahai N., *Random Forest Regression – How it Helps in Predictive Analytics?*, (2023, September), link: https://www.analytixlabs.co.in/blog/random-forest-regression/#:~:text=A%20Random%20forest%20regression%20model,all%20the%20individual%20trees'%20predictions., accessed: 1st of July 2024

21. Garcia E., *Predictive Modeling – A.3 Multinominal logitistic regression*, (2024, April), link: https://bookdown.org/egarpor/PM-UC3M/app-ext-multinomialreg.html, accessed: 1st of July 2024

Sažetak

Ovo istraživanje ispituje odnos između svojstava, njihovog zemljopisnog položaja i tipova pokrova zemljišta u Europi koristeći podatke iz LUCAS baze podataka o tlu. Korištenjem Kruskal-Wallis testova i Conoverovih *post-hoc* uparenih testova, istraženo je postoje li razlike u svojstvima tla između tipova pokrova zemljišta i između zemljopisnih regija. Različiti statistički modeli i tehnike, uključujući *K-Neighbours* regresor, *Radius Neighbours* regresor, *Ridge* regresiju i *Random Forest* regresiju, korišteni su za predviđanje ovih odnosa. Rezultati pokazuju da svojstva tla značajno utječu na tipove pokrova, dok geoprostorni podaci poboljšavaju robusnost predikcija. Analiza ima implikacije za istraživanje okoliša, upravljanje zemljištima i poljoprivrednim praksama te bi moglo usmjeriti razvoj politika za iste.

Summary

This study examines the relationship between soil properties, geographical location and land cover types in Europe using the data from LUCAS Soil Database. Using Kruskal Wallis tests and post-hoc Conover's pairwise tests we have explored if there are differences in soil properties between land cover types and between geographical regions. Various statistical models and regression techniques, including K-Neighbours Regressor, Radius Neighbours Regression, Ridge Regression, and Random Forest Regression were used to predict these relationships. The findings indicate that soil properties significantly influence land cover types. Geospatial data improved the robustness of the predictions. The study has implications for environmental research, land management, and agricultural practices and could guide the development of related policies.