CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Czech University
of Life Sciences Prague

FACULTY OF AGROBIOLOGY, FOOD AND NATURAL RESOURCES

DEPARTMENT OF SOIL SCIENCE AND SOIL PROTECTION

Spatial prediction of potentially toxic element content in agricultural soil using digital soil mapping approaches, health risk and source distribution assessment

……..

Doctoral dissertation

Ph.D. Student: Prince Chapman Agyeman (Ing.)
Supervisor: Prof. Dr. Ing. Luboš Borůvka

Prague 2022

DECLARATION

I, Prince Chapman Agyeman, hereby declare that the dissertation thesis titled "Spatial prediction of potentially toxic element content in agricultural soil using digital soil mapping approaches, health risk and source distribution assessment" is the original work and, duly investigated using scientific protocols has not been submitted anywhere for any degree or professional qualifications.

Prague, November 2022

Ing. Agyeman Prince Chapman

ACKNOWLEDGEMENTS

Glory be to the almighty God for giving me the strength, wisdom, knowledge, and comprehensive insurance throughout this study. Despite the turbulence, He made sure that this academic voyage was successful.

I will also give thanks to my supervisor, who took the leap of faith to give me the opportunity and accommodate and tolerate me throughout these four years of study. Through his ingenuity, clinical submission, and mentorship, this study has come to a fruitful end.

Thanks be to the entire Department of Soil Science and Soil Protection at the Faculty of Agrobiology, Food, and Natural Resources, CZU, for providing the conducive atmosphere and haven for me to undertake this doctoral program.

Finally, I wish to give thanks to my immediate family who, through their love, prayers, and support, have seen me through these studies.

PREFACE

The thesis presented contains consolidated publications spanning from 2018 to 2022 that are partitioned into sections and intercorrelated based on the objectives and the proposed hypothesis that follow scientific protocols. The thesis uses a dataset from the Frydek Mistek district to apply various models to the intent prediction and mapping of potentially toxic elements (PTEs) in soil. Additional auxiliary datasets employed in the thesis are from Sentinel 2, Landsat 8, the digital elevation model, soil chemical properties, and visible near-infrared spectroscopy to improve the prediction of PTEs in agricultural soil. PTE namely Pb, Sb, Mn, Cr, Cd, Cr, Ni, and Mn were studied in conjunction with a series of learning algorithms and auxiliary datasets in the prediction and the mapping of PTE levels in the agricultural soils of the Frýdek-Místek district. In addition, a source distribution assessment and a health risk assessment of the study area were carried out. The Department of Soil Science and Soil Protection at the Czech University of Life Sciences (CZU), Prague, provided oversight for the entire thesis, and grant providers and co-authors are recognized in the relevant publications.

Nevertheless, the study also used different PTEs datasets, including legacy datasets and preferential sampling datasets, to predict PTEs in agricultural soil on a national scale (Czech Republic). To our knowledge, no other study has investigated this, possibly due to the uncertainty of the fusion of legacy data and preferential sampling datasets, as well as ensembles, in the prediction of PTE in agricultural soil on a national scale. The manuscript for this study is currently under review. The use of preferential sampling data and legacy data in conjunction with ensemble models improved PTE prediction on a national scale. The study will facilitate the use of preferential sampling data in known polluted areas to supplement legacy data and improve prediction on a national scale. However, the composition of papers based on scientific findings obtained from various papers in this thesis suggests that there is no single modeling approach that is deemed to be the best in all studies. Below are the papers that constitute the thesis.

Agyeman, P. C., *Khosravi, V., Kebonye, N. M., John, K., Borůvka, L., & Vašát, R. (2022). Using spectral indices and terrain attribute datasets and their combination in the prediction of cadmium content in agricultural soil.* ***Computers and Electronics in Agriculture****, 198, 107077.*

Agyeman, P. C., *Kingsley, J. O. H. N., Kebonye, N. M., Ofori, S., Borůvka, L., Vašát, R., & Kočárek, M. (2022). Ecological risk source distribution, uncertainty analysis, and application of geographically weighted regression cokriging for prediction of potentially toxic elements in agricultural soils.* ***Process Safety and Environmental Protection****, 164, 729-746.*

Agyeman, P. C., *John, K., Kebonye, N. M., Borůvka, L., Vašát, R., Drábek, O., & Němeček, K. (2021). Human health risk exposure and ecological risk assessment of potentially toxic element pollution in agricultural soils in the district of Frýdek-Místek, Czech Republic: a sample location approach.* ***Environmental Sciences Europe****, 33(1), 1-25.*

Agyeman, P. C., *Kebonye, N. M., Khosravi, V., John, K., Borůvka, L., & Vašát, R. (2023). Optimal zinc level and uncertainty quantification in agricultural soils via visible near-infrared reflectance and soil chemical properties****. Journal of Environmental Management****.*

Agyeman, P. C., *John, K., Kebonye, N. M., Khosravi, V., Borůvka, L., & Vašát, R. (2023).* Prediction of the *concentration of antimony in agricultural soil using data fusion, terrain attributes combined with regression kriging.* ***Environmental Pollution****.*

Agyeman, P. C., *Borůvka, L., Kebonye, N. M., Khosravi, V., John, K., Drabek O., Tejnecky V., Quantification of the optimal cadmium level in agricultural soil using legacy data, preferential sampling, Sentinel 2, Landsat 8 coupled with ensemble model.* ***Journal of Environmental Management*** *(Under review)*

TABLE OF CONTENTS

1.0 LITERATURE REVIEW

1.1 Soil pollution

The term soil pollution refers to the presence of an anomalous chemical or substance in a higher-than-normal concentration that has a negative impact on any non-targeted organism (FAO & ITPS, 2015). Potentially toxic elements (PTEs) have anthropogenic sources, can occur naturally in soils as mineral components, and can be dangerous in high quantities. Soil pollution is frequently not precisely assessed or visible, making it a hidden concern. Pollutants are becoming more diversified because of agrochemical and industrial advances. Soil surveys to differentiate pollutants are time-consuming and costly due to their diversification, leading to variation. The effects of soil pollution are further influenced by soil properties, which restrict contaminants' mobility, bioavailability, and residence time (FAO & ITPS, 2015). Industrialization, wars, mining, and intensification in farming have left a legacy of contaminated soils worldwide (Bundschuh et al., 2012; Luo et al., 2009). Soil has been used as a sink for dumping strong and liquid pollutants since urban expansion. It was carefully planned so once the toxins were buried and out of sight, they would pose no threat to human health or the environment and would eventually disappear (Swartjes & Siciliano, 2012). The primary sources of soil pollution are anthropogenic, resulting in a buildup of toxins in soils that may reach alarming levels (Cachada et al., 2017a). The attributes of soil allow it to naturally accommodate and support the mobility of inorganic chemicals or potentially toxic elements (PTEs) such as chromium (Cr), cadmium (Cd), arsenic (As), manganese (Mn), nickel (Ni), lead (Pb), zinc (Zn), copper (Cu), mercury (Hg), antimony (Sb), and cobalt (Co) (Sun & Chen, 2016). PTEs have been classified as the third most significant threat to soil processes in Europe and whole Eurasia, the fourth in North Africa, the sixth in Asia, the seventh in the Northwest Pacific, the eighth in North America, and the ninth in Sub-Saharan Africa and Latin America (FAO & ITPS, 2015).

1.2 Potentially toxic elements

Potentially toxic elements (PTEs) are abundant natural components of the earth's crust soils (Iñigo et al., 2011; Kabata-Pendias and Mukherjee, 2007). PTE is a generic terminology given to poisonous metal(loid)s that are detrimental to either human well-being or a sustainable

environment. Although part of PTEs has anthropogenic sources, some elements can happen naturally in soils as components of minerals. One of the many substantial detrimental effects of human practices on aquatic and terrestrial ecosystems is the widespread mobilization and distribution of pollutants from their natural reservoirs into the atmosphere, soil, and water (Hou et al., 2017; Zhao et al., 2014). Soil contamination cannot be regularly evaluated or outwardly seen, making it a concealed threat. The diverse variety of contaminants is continuously advancing due to agrochemical and industrial developments. Nevertheless, the impacts of soil contamination also depend upon soil properties since this control the mobility, bioavailability, and residence time of PTEs (FAO & ITPS, 2015). Attempting to address soil PTE contamination introduces some unique problems such as i) PTEs are non-destructible and frequently accumulate rather than degrade in soils (Maas et al., 2010); ii) they have a wide range of health effects, and the health vulnerability is complicated by their oxidation state and associated bioavailability disparities (Walker et al., 2003); and iii) there are numerous diffusional sources of PTE contamination (Qu et al., 2020). Excessive levels of PTEs in soils not only have an impact on soil health, but due to their persistence in the environment and long biological half-lives, they can accumulate in the food chain and potentially impair human health (Ackermann, 1980; N'guessan et al., 2009; Xie et al., 2012). Although the negative consequences of PTEs have long been known and that contemporaneous prominence and its impact on PTEs persists and is escalating in some locations, most formerly cultivated land, mining area, industrial area and metallurgical tailing dumpsites are now abandoned (Gholizadeh et al., 2015).

1.3 Sources of PTEs

PTEs can enter the soil through divergent pathways, namely geogenic sources and anthropogenic sources. The fundamental sources of soil contamination are anthropogenic, resulting in the sowing of contaminants in soils that may reach levels of concern (Cachada et al., 2017b). Source of pollutant occurs from different sources, namely natural enrichment, agricultural activities (land application of fertilizers, animal manures, composts, pesticides), industrial activities, transportation system, atmospheric deposition, waste management and treatment, and mining (Basta et al., 2005; Jiménez-Ballesta et al., 2017; Khan et al., 2008; Zhang et al., 2010). PTEs of anthropogenic sources are typically more mobile and bioavailable in soil than PTEs of lithogenous

or pedogenic origin (Kaasalainen & Yli-Halla, 2003; Keeperman, 2000). According to Seaward et al. (1990) Thevenon et al., (2011), natural processes such as weathering of rocks, erosion, rock formation and volcanic eruption play a significant role in the emission and exposure of enormous quantities of PTEs such as Al, Cu, Hg, Mn, Ni and Zn into the environment, particularly soil. Scragg, (2005) reported that agricultural production was the primary human influence exerted on the soil. The ever-growing human population is the fulcrum that pushes farmers to produce more and apply agrochemicals such as fertilizers and pesticides to enhance yield and productivity. Application of agrochemicals like foliar sprays rich in PTEs, for instance, Co, Cu, Fe, Mn, Mo, Ni and Zn, to soil essentially for plant growth (Lasat, 1999), successively during every crop season elevates the PTE concentration in the soil. However, recent publications by Liang et al. (2017); Luo et al., (2009); Nicholson et al., (2003) suggested that anthropogenic activities related to agronomic practices such as the use of fertilizers, fungicides and fossil fuel combustion have contributed to the high accumulation of Cu, Hg, Mn, Pb or Zn in soils. For example, lead arsenate and arsenate compounds used to control pests in fruit orchards in New Zealand and Australia are rich in Cu, Cr and As (Wuana and Okieimen, 2014); these elements are likely to increase the concentration of PTEs in soil beyond the tolerable limits. Basta et al. (2005) recounted that the application of biosolids like sewage sludge, industrial waste and compost to agricultural fields results in the increment of PTEs such as As, Cd, Cr, Cu, Pb, Hg, Ni, Se, Mo, Zn, Tl and Sb in the soil. Industrial activity coupled with mining as well as tailings discharges a large amount of PTEs that contaminates the soil. For instance, huge lead (Pb) and zinc (Zn) ore mining, as well as metal smelting, has the propensity in contaminating the soil and poses an ecological risk. According to the FAO and ITPS (2018) reports, the United Nations Environmental Assembly (UNEA-3) agreed on a resolution calling for expedited actions and collaboration to address and manage soil pollution worldwide.

1.4 Spatial distribution of PTEs

PTE's spatial distribution is primarily related to the source of pollution, which is primarily soil pollution, and is thus more pervasive (Borůvka et al., 2005) and may be detrimental to the environment and human health due to their degree of toxicity and tenacity in nature. However, the spatial distribution of confined pollution has a significant structure, with probable pollution

adjacent to the point source declining with distance away from the source (Borůvka et al., 2005). Soil physiochemical characteristics typically significantly impact PTE spatial distribution mechanisms (Mahmoudabadi et al., 2015). According to Zhao et al. (2010), the enrichment index of PTEs can sometimes be associated with various soil qualities and their spatial distribution of index enrichment for some PTEs like Cd, Ni, and Zn related to spatial structures of pH, OM, sand, and clay. The metal deposit can be enhanced by increasing soil pH, organic matter, cation exchange capacity, and the amount of iron and manganese oxides (Lake et al., 1984). PTE moieties have a significant impact on soil chemical mobility and bioavailability in soil (Zhang et al., 2018). The bioavailable proportion of a PTE in soils, on the other hand, is critical in its accumulation by organisms. PTEs emitted by anthropogenic activities are thought to have a high bioavailability (Bolan et al., 2014). As a result, recognizing the chemical form of PTEs in soils might be useful for analyzing their significant environmental concerns (Sun et al., 2019). The soil surface cation exchange model of PTEs binding reactions explains the bonding of metal ions to the surface of the mineral functional group to generate a more stable molecular unit (Christl & Kretzschmar, 1999). However, the major surface functional groups are inorganic hydroxyl groups which attach to surface Al, Fe, Mn, or Si on oxides or Al and Si exhibited on the margins of clay particles (Zachara & Westall, 2018). The surface complexation model has been predicated on observing that ion sorption occurs at specific surface sites (Christl & Kretzschmar, 1999).

1.5 Potentially toxic elements for study

1.5.1 Antimony

Antimony occurs naturally in the environment and enters the environment through a variety of human applications such as type-metal alloy (with lead to prevent corrosion), in electrical applications, pewter, in primers and tracer cells in munition manufacture, semiconductors, flameproof pigments and glass, medicines for parasitic diseases, as a nauseant, as an expectorant, combustion of fossil fuels (Bradl, 2005). However, it is mainly produced from the ores stibnite (Sb_2S_3) and valentinite (Sb_2O_3). Even though most Sb contamination appears to come from mining and industrial emission sources, such as smelting, it frequently co-occurs with arsenic (Telford et al., 2009).

1.5.2 Arsenic

Arsenic is chemically related to phosphorous since both belong to the same periodic table V-A (Bradl, 2005a). Arsenic is deposited into the soil anthropogenically through a variety of sources, including point sources, which are industrially orientated (mining, smelting, metal hardening, paints, textile, industrial dusts, medicinal, pharmaceutical, wastewater, pesticides, smelting of gold, production of iron and steel, industrial waste, combustion of fossil fuel, industrial waste). Furthermore, it enters the soil through diffusion pathways brought on by agricultural activities (application of arsenic in herbicides, cattle and sheep dips and insecticides). Pesticides account for about 80% of As production; however, because of their toxicity, pesticides no longer contain large amounts of As, although it is nevertheless a dominant PTEs in pesticides (Deschamps and Matschullat, 2011). Agricultural practices such as pesticides, fertilizers, sewage sludge, and manure are significant sources of As in agricultural soils (Kabata-Pendias & Szteke, 2015a). According to Falandysz and Rizal, (2016), As can exist in inorganic and organic forms, but the inorganic one is more prevalent in the soil. Due to its noxious nature, arsenic can be hyperaccumulating in plants such as mushrooms (Falandysz & Borovička, 2013).

1.5.3 Cadmium

Cadmium is regarded as one of the most environmentally hazardous metals, negatively impacting all biological functions (Bernard, 2008). Cadmium accumulation in the soil and the environment exhibits a very negative impact on the environment and food quality. The anthropogenic occurrences of Cd in the soil via point source are metallurgy, mining, phosphate fertilizer production, pigments and paints, electronics, industrial and incineration dust and fumes, wastewaters, pesticides, battery, PVC products, colour pigments. In contrast, some diffuse sources are accumulation from phosphatic fertilizers (containing 2-200 Cd mg/kg), domestic and sewage sludge, wear of automobile tires, lubricants and mining and metallurgical activities, pollutions from mining and smelting operations and atmospheric deposition from the combustion of fossil fuels (Smiljanić et al., 2019).

1.5.4 Chromium

Naturally, Cr is among the few PTEs that do not occur in elemental but exist in compound forms such as chromite (Wuana & Okieimen, 2011). Smith (1995) asserted that the mineral chromite, $FeCr_2O_4$, is mined as a significant ore product of chromium. It is applied as an electroplating process emission and the disposal of Cr-containing wastes primary sources of Cr-contamination. Jeřábková et al. (2018) and Smiljanić et al. (2019) reported that the anthropogenic sources of Cr can be put into two sources, namely point sources like mining and metallurgy, metal plating, rubber, photography, industrial dust and fumes, tanning, leather industry, chemical industry, fertilizers, textile industry, paints and pigments, and the diffuse ones that can be wastewater and sludge from dyeing and tanning industries. Mobility of Cr is affected by soil sorption properties such as clay content, iron oxide content, and the amount of organic matter present. Surface runoff can transport Cr in its soluble or precipitated form to surface waters; soluble and desorbed Cr complexes can leach from soil into groundwater; the leachability of Cr(VI) increases as soil pH increases; however, the majority of Cr released into natural waters is particle associated and is eventually deposited in the sediment or soil (Smith, 1995).

1.5.5 Copper

PTE copper is malleable, ductile, and a good conductor of heat and electricity. This is distinguished by a crystalline structure that absorbs frequencies in the visible range. Cu rapidly combines to organics in the soil, indicating that maybe a small fraction of copper will be detected in solution as ionic copper, Cu (II). Cu solubility is dramatically decreased at pH 5.5, close to the optimal farmland pH of 6.0–6.5 (Eriksson et al., 1997; Martínez & Motto, 2000). The sources of copper in the soil and the environment, as propounded by Smiljanić et al. (2019), are point sources such as mining and metallurgy, plating, rayon, electrical and electronic waste, pesticides, paints and pigments, textile industry, explosive, and diffuse sources like manures, fertilizers, pesticides, sewage sludge and atmospheric fall out resulting from the combustion of fossil fuels and industrial processes.

1.5.6 Manganese

Manganese is among the most prevalent metals in soils, occurring as oxides and hydroxides and cycling via its three oxidation states which are found mainly as pyrolusite (MnO_2) and, to a smaller extent, rhodochrosite ($MnCO_3$)(LENNTECH, 2008). The anthropogenic source of Mn that is introduced into the soil and the environment in a point and the diffuse source is through the production of ferromanganese steels, electrolytic manganese dioxide for use in batteries, alloys, catalysts, fungicides, antiknock agents, pigments, dryers, wood preservatives, coating welding rods (Bradl, 2005b). Manganese is a mineral nutrient that is required by all plant species (PlantProbs.net, 2019). On the other hand, Mn is accumulated by species such as diatoms, mollusks, and sponges. Manganese dioxide is employed as a catalyst, and when chemically linked to potassium to form potassium permanganate, it is a powerful oxidant and disinfectant. Manganese oxide (MnO) and manganese carbonate ($MnCO_3$) are two other compounds that have applications: the first is used in fertilizers and ceramics, while the second is the starting material to produce other manganese compounds (LENNTECH, 2008).

1.5.7 Nickel

Most of the nickel on earth is unavailable because it is trapped in the planet's iron-nickel molten core, which contains 10% nickel. It has been estimated that the entire amount of nickel dissolved in the sea is roughly 8 billion tons. Although the organic matter has a high capacity for metal absorption, coal and oil contain significant amounts. Nickel content in soil can range from 0.2 ppm to 450 ppm in some clay and loamy soils. The anthropogenic source from the point and the diffuse sources are fertilizers, manures, metal refining, smelting, burning of coal and industrial sewage sludge, emissions from mining and smelting operations, an atmospheric fallout from the combustion of fossil fuels, mining and metallurgy, electroplating, production of iron and steel, industrial dust, industrial aerosols, incineration of waste, fertilizers, combustion of coal, battery, chemical industries, food processing industries (Alloway, 2013). Khodadoust et al. (2004) reported that the most common application of Ni is as a constituent of steel and other metal products, and metal plating industries, fossil fuel burning, and nickel mining and electroplating are significant nickel contributors to pollution in the soil.

1.5.8 Lead

Lead (Pb) is a naturally occurring bluish-grey metal that is commonly found as a mineral in combination with other elements such as Sulphur (PbS, $PbSO_4$) or oxygen ($PbCO_3$), and its concentration in the earth's crust ranges from 10 to 30 mg/kg (USDHHS, 2007). The typical mean Pb content for surface soils worldwide is 32 mg/kg and ranges from 10 to 67 mg/kg (Kabata-Pendias, 2010a). The source of Pb introduced into the soil via point and diffuse sources are mining and metallurgy, industrial dust and fumes, application of lead in gasoline, combustion fossil fuel, solid waste, solid waste combustion and incineration, industrial waste, paints and pigments, explosives, ceramics and dishware, some types of PVC, pesticides, fertilizers, manufacturing of lead-acid batteries, urban runoff, exhaust gases of petrol engines, which account for nearly 80% of the total Pb in the air, pesticides, fertilizer impurities, emissions from mining and smelting operations, and atmospheric fallout from the combustion of fossil fuels. Soils near Pb mines may contain as high as 0.5% Pb content (Galušková et al., 2011; Alloway, 2013). In a global context, Pb in soils averages around 27 mg/kg, with level differences for individual soils ranging from 3-90 mg/kg. Pb concentrations in Cambisols and Histosols were significantly higher than in Arenosols (Kabata-Pendias and Szteke, 2015b).

1.5.9 Zinc

Anthropogenic activities increase Zn concentrations are rising abnormally. Due to the apparent accumulation of Zn in soils, plants frequently experience Zn uptake that their systems cannot handle. Greaney (2005) reported that Zn could disrupt soil activity by interfering with the activity of microbes and earthworms, which slows the degradation of organic matter. Zn concentrations in soils worldwide range from 30-100 mg/kg on average, while significantly more significant amounts can be found in calcareous and organic soils (Kabata-Pendias and Szteke, 2015). Similarly, anthropogenic Zn sourcing from diverse agricultural and mining operations may raise Zn levels in specific soils (Araújo et al., 2017). Point and diffuse sources of Zn are mining and metallurgy, galvanization, plating iron and steel, electroplating, fertilizers, metal waste, fertilizers, manures, pesticides, sewage sludge.

1.6 Pollution assessment-based receptor model (PAB-RM)

The validity and dependability of soils for crop production, particularly urban, peri-urban, and agricultural soils, should be investigated to assess the impact and toxicity of PTE pollution. Huang et al. (2018) and Sawut et al. (2018) asserted that indices could reliably measure the state of soil contamination and the extent to which human activity impacts the soil and the environment. The use of pollution indices allows for measuring environmental risk and the degree of soil deterioration, illustrating the systematic relevance of evaluating soil quality with indices (Adamu and Nganje, 2010). In addition, the index (enrichment factor, ecological risk) allows researchers to assess whether PTE accumulation in soil was generated by a human or natural source (Peter and Adeniyi, 2011). Nonetheless, computed pollution indices values notify researchers and other stakeholders about the extent of pollution in the environment, allowing them to take appropriate action when necessary. Moreover, pollution assessment indices are significant for monitoring soil quality and indicate long-term resilience, especially in urban, peri-urban, and agro-ecosystems (Norbaya et al., 2014). These indices are frequently used to quantify PTE pollution in agricultural soil, urban soil, peri-urban soil and the environment.

1.7 Digital soil mapping

Digital soil mapping (DSM) or predictive soil mapping is presently the most effective way to predict the spatial variation of soil/sediment over an area (McBratney et al., 2003a). According to Minasny and McBratney (2016), DSM or predictive soil mapping has become a successful subdiscipline of soil science. Iqbal et al. (2005) stated that spatial variability of soil physical properties within or between soils is, at most times, inherent due to geological and pedological soil formation factors. However, some of the variability may be caused by other management practices. The factors work together on a temporal and spatial scale, and the content is further adjusted by the spatial heterogeneity deposition of soil properties. Zhu et al. (2018) reported that environmental covariates and soil relationships in spatial predictions are fitted with a model and the learned nexus and are subsequently applied to spaces or locations where data (soil/sediment data) are unknown. Usually, DSM forms a quantitative soil environment relationship centered on the modelling points or sample observation points to characterize the nexus between soil and

environmental covariates such as climate variables, geological variables, slope and topographic wetness index (Penizek & Boruvka, 2008). DSM applies models to compute soil property values at unknown locations (Heung et al., 2016; McBratney et al., 2003a; Minasny and McBratney, 2016; Zhu et al., 2001). Globally, the soil science communities have adopted DSM for mapping soil properties and classes (Arrouays et al., 2014) and, to a significant extent, to predict the concentration of PTEs in the soil/sediments. Due to its high accuracy compared to conventional mapping, many stakeholders (e.g., FAO) have embraced DSM usage. DSM is consistent for sustainable land management (Padarian et al., 2019), and by extension, it is valuable and efficient in the spatial prediction of PTEs. Significant to the success and applicability of spatial predictions are the underlying assumptions employed in describing the relationships and how these relationships are characterized (Zhu et al., 2018b). Soil mapping techniques have improved by the progression of geographical information technology and computational technology (Zhang et al., 2017). Lagacherie and McBratney (2006) defined digital soil mapping as the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation and knowledge from related environmental variables. The accumulation of PTEs in the soils/sediments has been a worldwide concern (González-Macías et al., 2006; Liu et al., 2003), as it poses an utmost threat to human health (Chen et al., 2015). According to Chen et al. (2009), one of the feasible roles of studies is the inhibition of PTEs in the soil. On the other hand, spatial prediction of PTEs provides an avenue to delineate the distribution of potentially toxic elements, their concentration, occurrence and knowing their source of pollution.

DSM approaches include conventional statistical techniques, machine learning algorithms (MLA), geostatistical methods and hybrid approaches (Chen et al., 2019). The traditional approaches include the application of commonly used non-spatial statistical techniques such as multiple linear regression (Jiang et al., 2019), partial least square regression (Lago et al., 2021), generalized linear models and linear mixed models (Doetterl et al., 2013). The statistical methods were applicable in the modelling of soil organic carbon (SOC) (Gomes et al., 2019), PTEs (Ballabio et al., 2018) and soil properties (Shi et al., 2011). While such statistical methods are uncomplicated to implement, their requirements for unbiased and comparable distribution with huge datasets are

sometimes an obstacle (Chen et al., 2019). These approaches are also characterized by a relative paucity of spatial information, making them less dependable and inappropriate for identifying regional variations (Lian et al., 2009). According to Kempen et al. (2012), DSM applications are centered on scientific research and are regionally based specific. Human activities can pollute confined areas with well-defined borders from point sources or contaminate wider land surfaces diffusely. However, identifying the source of PTEs is frequently difficult if the point source cannot be determined at the place where large concentrations of the element are observed (Tóth et al., 2016). DSM of PTEs content of topsoil within specific regions of the earth, such as European coverage, aids in assessing spatial patterns and hotspots on the continent (Tóth et al., 2016). Whilst the spatial variability of PTEs is naturally associated with point sources of pollution, other parameters such as wind speed and direction are also relevant and should not be overlooked (Taghizadeh-Mehrjardi et al., 2021). Behrens et al. (2018) applied DSM in multi-scale terrain feature generation and their respective efficiency for a deep learning algorithm. Costa et al. (2018) reported that the digital elevation models (DEM) are extensively used in digital soil mapping (DSM) and are chosen based on metrics and indicators (quality criteria) that are supposed to reflect how effectively a particular DEM depicts the terrain surface.

1.7.1 Geostatistical approaches

Geostatistical approaches encapsulate simple/ordinary kriging, cokriging, universal kriging, and empirical Bayesian kriging. Geostatistical techniques are commonly used to interpolate geographical characteristics with considerable spatial autocorrelation, including climate factors, ecological soil properties, and geological elements, such as PTEs. The application of geostatistical modelling approaches cut across broader spectra such as application in soil SOC (Bangroo et al., 2020; Peng et al., 2013; Wang et al., 2015), soil properties (López-Granados et al., 2005; Vašát et al., 2013), PTEs (Ash et al., 2014; Linnik et al., 2020; Łyszczarz et al., 2020) and analyzing remote/proximal sensing images (Zawadzki et al., 2005; van der Meer, 2012). The advantages of geostatistical interpolation include the (i) capacity to account for directional factors, such as soil pollution (PTEs), soil erosion, siltation flow, lava flow, and wind movement, as well as (ii) the potential to surpass the lowest and highest point values. The limitations of geostatistical analysis

are the smoothing effects of kriging and the fact that spatial interpolation evaluates physical data in a continuous domain.

1.7.2 Sequential gaussian simulation (SGS)

The fundamental idea behind SGS is to reconstruct sequential grid points using the empirical distribution's temporary proportion (i.e., in this case the PTEs data). It generates an output that is relatively like the accurate spatial actuality of an interest parameter. Even though the data should be detectable, the interpolated points symbolize the variogram approach and the nugget effect's local noise (Goovaerts, 2001). Moreover, it is premised on the multi presumption of a random feature model (Goovaerts, 2001; Johari et al., 2020). The data set appears to provide the critical standard score change, ensuring the logic of the univariate data distribution at the very least. For more information on SGS, refer to Gholampour (2019).

1.7.3 Machine learning approaches

With low soil data and auxiliary environmental data, MLA techniques can tolerate nonlinearity and multicollinearity and counteract overfitting (Gautam et al., 2011). MLA models include an array of methods and are not limited to random forest, cubist, support vector machines, Bayesian regularized neural network, regularized random forests, conditional inference forest, extreme gradient boosting, gaussian process regression, multivariate adaptive regression splines, partial least square regression, Bayesian generalized linear model, M5 tree model and quantile regression forest.

Random forest (RF) is defined as the assemblage of diverse regression and/or classification trees. Breiman, (2001) created the algorithm and asserted that its higher accuracy level could be compared to adaptive boosting. Gislason et al. (2006) and Heung et al. (2014) articulated that the computational ability of RF is faster. The variable handling capacity of the RF is both categorical and continuous. According to Díaz-Uriarte and Alvarez de Andrés (2006), RF does not need variables preselection and it is capable of handling noise due to its robust nature. Cutler et al. (2007) documented that the algorithm begins with several tree samples (ntree) from the data sampled. The operation is modified employing which the predictors (mtry) are sampled

arbitrarily; afterwards, every ntree grows a regression tree and the RF algorithm chooses the utmost split between the variables sampled instead of all the variables (Nawar and Mouazen, 2017). The default value mtry is considered the square root of the totality of the variables (Abdel Rahman et al., 2014). Sega and Xiao (2011) proposed the formulae for RF regression which is given as

$$RF=\frac{1}{M}\sum_{m=1}^{M}\hat{f}_m^*(X_\theta)$$

In which M represents the m^{th} bagging repeatedly tree (m=1,….., M), X_θ represents the covariates and the $\hat{f}_m^*(X_\theta)$ also denotes the m^{th} tree of an individual test case.

Support vector machine (SVM) is an MLA that generates an optimum disengaging hyperplane to distinguish categories that have similarities and are not independent in a linear way. Vapnik (1995) developed the algorithm, which was meant for purposes of classification and in recent times, it has been adopted for solving regression-oriented problems. Li et al. (2014) communicated that SVM is one of the best classifier techniques and it has been applied in a diverse field. This study employs the regression aspect of SVM (support vector machine regression - SVMR). SVMR was initiated by Cherkassky and Mulier (2006) as a regression based on kernel and its computation functions with a linear regression model that possesses a multivariate space feature. John et al. (2020) indicated that the SVMR utilizes a hyperplane linear regression that establishes a nonlinear relationship, and it is possible for the space feature. Vohland et al. (2011) outlined that epsilon (ε)-SVMR utilizes a trained dataset to procure a represented model as an insensitive feature, which is used to map data independently with the optimum epsilon ε- deviation from dependent data training. The preset distance ε error inside is ignored from the actual value, and if the error is seen to be bigger than the epsilon (ε), it is compensated for by the soil property. The model also reduces the intricacy of training data to a broader subset of support vectors. The equation as proposed by Vapnik (1995) is given as

$$y(x) = \sum_{k=1}^{N}\alpha_k K(x, x_k) + b,$$

In which the b represents the scalar threshold, $K(x, x_k)$ representing the kernel function, α denoting the Lagrange multiplier, N symbolizing the number dataset, x_k representing the data

input and y is the data output. One of the critical kernels used is the SVMR operation with the Gaussian Radial Basis Function (RBF). The RBF kernel was applied to ascertain the optimum SVMR model that is essential to procure the finest penalty set factors C and the kernel parameters gamma (γ) for the PTEs training data.

Multivariate adaptive regression splines (MARS) were created by Friedman (1991) as a nuanced technique and a non-parametric regression method that generates multiple linear regression models across a wide range of predictor values. Its quantitative approach that divides training data into simple linear segments (splines) with varying gradients using a splitting approach (slope). MARS makes no assumptions about the primary relationships between the dependent variable and independent factors (Zhang et al., 2016). Splines are frequently connected smoothly with piecewise polynomials, also known as basic functions (BFs), resulting in a comprehensive framework that can compensate both linear and nonlinear behavior (Zhang et al., 2016). Friedman (1991) and Zhang et al. (2016) provide more information on the MARS algorithm.

M5 model tree is a numerical prediction algorithm, and its splitting criterion is based on the standard deviation of the values in the subset T of the training data that reaches a specific node (which is an analogue of entropy). M5 model tree however perfoms a binary decision tree with linear regression functions at the terminal (leaf) nodes that can forecast uninterrupted numerical attributes Quinlan (1992). The divide-and-conquer strategy is used to build tree-based models. Two steps are necessary for model tree generation. Making a decision tree using a splitting criterion is the first step. models.

Cubist implements a related approach to boosting but is called "committees," which make iterative decisions in sequence. This model employs instance and model-based coupling techniques to create a multivariate regression from training data. Quinlan (1992) and Kuhn and Johnson (2013) reported that the cubist primary value is to enhance the multiple trainings committees and also augment the weight to ensure it is well balanced. Similarly, the cubist model training committees (above one mostly) boosting method shares similarities with sequential series tree development with weight-adjusted. Kuhn et al. (2013) recounted that the cubist

model is typically used to apply amended depending on the number of neighbours, based on prediction rules. However, Kuhn et al. (2014) stated that the cubist regression model utilized in classification and regression is prevalent and extensively used in R as a package. The cubist model follows the same method as in random forests.

Partial least square regression (PLSR) technique has the leverage of eliminating the challenge of multidimensionality between many predictor variables (Mishra et al., 2020). The algorithm can be used to perform and analyze independently for each number of characteristics ranging from 1 to 10 (Agyeman et al., 2022). After superimposing the explanatory variables to an original space, the algorithm concurrently uncovers a linear regression method linking the predictor variables and the correlation between the explanatory variables in the new projected space (Gamon et al., 1992). Ehsani et al. (1999) provide more information on the PLSR algorithm.

The **Gaussian process regression (GPR)** is a technique of nonparametric modeling (Vasudevan et al., 2009; Y. Zhang & Xu, 2021). This is a supervised machine learning technique for general regression and probabilistic classification problems. Wang et al. (2020) attribute GPR's utility to its ease of use and high accuracy. GPR can also help to reduce dataset overfitting (Ballabio et al., 2019).

Quantile regression forest (QRF) is a variation on the RF method (Meinshausen, 2006). It keeps track of all observation samples in each decision tree node, their average values, and their variation. It also evaluates the provisional distribution of prediction results predicated on this information (Dharumarajan et al., 2019).

Bayesian techniques are the best way to solve learning problems, and any other approach that does not approximate them should also perform worse on average. They are extremely effective for data model comparative study because they automatically and quantitatively automatically embody (Jr et al., 2003). Under Bayes' Rule, complex approaches automatically self-punish. *Bayesian approaches complement neural networks (NNs)* by overcoming an excessively flexible network's proclivity to explore virtually nonexistent or excessively convoluted data models (Tchagang & Valdés, 2019). Traditional backpropagation NN training methods utilize a solitary set of variables (weights, biases, etc.), the Bayesian method to NN modelling techniques

considers all potential values of network parameters, weighted by the probability of each set of weights. Applying the Bayesian regularized neural network approach, Bayesian inference is utilized to derive the posterior probability distribution of weights and connected attributes from a prior probability distribution based on updates offered by the training set (Tchagang and Valdés, 2019).

Extreme gradient boosting (EGB) is a decision tree that uses a gradient-boosted method to enhance speed and precision (Climent et al., 2019). It is a regression and categorization problem that is solved sequentially by a set of limited prediction techniques, with each new design focusing to rectify the imperfections of the earlier models (Agyeman et al., 2022). The EGB is a pragmatic and modular application of Friedman's gradient boosting framework that is premised on Friedman's original gradient boosting technique (Climent et al., 2019).

Regularized random forest (RRF) is the latest change to random forest (RF), which applies a regularization structure to random forests and integrates it into the tree increasing algorithm (Deng, 2013). RRF, as displayed in, produces high feature subsets, and reduces the number of characteristics used in categorization and regression problems (Deng et al., 2012). To preclude overfitting, regularization typically involves adding a penalty to an error function.

Bayesian generalized linear model in a modeling approach that presumes generalized linear models (GLMs) with factors in an enclosed environment of prevalent preference (e.g., in monotonic or convex regression), but maintaining a genuine posterior variation backed by a system of linear restrictions and limitations can be challenging, particularly when some limitations are legitimate and implementable, culminating in a decreased feature subspace. Bayesian methods eliminate the need for a nonlinear remedy by repeatedly sampling from posterior probabilities. Another advantage of the Bayesian technique is its versatility in assessing ambiguity in calculated random impacts and hyperparameter functionalities. Bayesian inference is based on data obtained instead of the assumption of infinite data populations. These inferences advantage Bayesian techniques because all inferences are accurate and not approximated, and the results are understandable (Congdon, 2007; Ntzoufras, 2011).

Conditional inference forest (CIF) is a tree-growing technique popular in bioinformatics applications (Nicodemus et al., 2010). Theoretically, CIF varies from traditional random forests in that it separates the alternative of the isolating different factors from the agglomeration of the already selected partitioned variable's dividing point (Hothorn et al., 2012). The optimal control divide parameter is largely decided in the first step, and an associative test is performed between the potential split parameters and the response. CIF techniques are used in addition to a provisional grid for the potential composite relevance measure, enabling for superior appraisal of each parameter's independent commitment and discrimination of observables from erroneous relationships (Delerce et al., 2016). When sampling without substitution is used, the CIF two-step method yields a non-biased split variable alternative and a test statistic when the quadratic version is used.

1.7.4 Ensemble modeling

Putting various models together to leverage each other's strengths while avoiding each other's weaknesses to produce good modeling predictions gave way to the ensemble modeling approach. It is essentially a hybridized model that combines the output of multiple modeling approaches into a single modeling approach to produce a more efficient output. Stacking is an ensemble technique for achieving maximum generalization precision by combining the results of various machine learning methods into a single component technique (Breiman, 1996; Malone et al., 2014). The basic concept of a typical ensemble model is divided into two stages: the initial level (Level 0), which contains the sub model, and the final level, which contains all the sub model's predictions piped through a meta-learning algorithm to give a final prediction due to the sub model's departure from level 0. The ensemble model in the thesis is made up of four sub models and the stack tree or meta-learner is a standalone modeling approach that uses the weights generated by the sub models to produce the final predictions.

1.7.5 Application of proximal and remote sensing in DSM

Proximal and remote sensing technologies are changing into efficient means for obtaining vast amounts of geographical data (Brevik et al., 2016). The advent of new statistical techniques and, eventually, machine learning, enables new ways to interpret these data (Lu, 2010). These

techniques have significantly altered soil mapping today (McBratney et al., 2003b) and provided new approaches for the digital mapping of polluted soils with PTEs. The development of proximal and remote sensing has been acknowledged as a suitable and successful remote and contactless discovering method for detecting and analyzing soil contaminants (Choe et al., 2008a). Applying proximal and remote sensing methods can be a useful resource in the phases of pollution research and environmental concern assessment and can ultimately lead to a noticeable decrease in pollution issues in both natural and man-made environments (Asmaryan et al., 2014; Gholizadeh et al., 2018; Wu et al., 2007).Visible near infrared (Vis-Nir) spectroscopy has been successfully utilized by Bray et al. (2009) to detect quantified PTEs in polluted and unpolluted soil with varying concentrations of PTEs (Cu, Zn, Cd, and Pb). Ren et al. (2009), on the other hand, reported that applied reflectance spectroscopy was used to evaluate the content of PTEs such as As and Cu in mining areas. Soil samples were examined employing HyMap hyperspectral imagery and a varied multiple endmember spectral mixture modelling (VMESMA) computational technique to determine the distribution sequence and proportion of remnant tailing materials in the research area (Cocks et al., 1998). On the other hand, the application of remote sensing in the potentiality investigation of the spectral variability associated with PTEs to map the spread of PTEs in impacted areas of a mining region in southeast Spain using HyMap data (Choe et al., 2008).

1.8 Environmental covariates

The concept of vegetation/soil-environment relationships has frequently been presented in an equation with six key environmental factors. Jenny (1941) proposed that the nature and characteristics of soil at any location are the result of the interactions of five soil-forming factors, namely 'c' - climate; 'o' - vegetation and living organisms; 'r' - relief, topography, and landscape attributes; 'p' - parent material, lithology; and 'a' - time or age. McBratney et al. (2003) proposed the SCORPAN model, in which soil (as either soil classes, Sc, or soil attributes, Sa) is an empirical quantitative function of seven environmental covariates:

$S = f(s, c, o, r, p, a, n)$

where s: soil, other properties, or prior knowledge of the soil at a point; c: climate, climatic properties of the environment at a point; o: organisms, vegetation or fauna or human activity; r: topography, landscape attributes; p: parent material, lithology; a: age, the time factor; and n: space, relative spatial position. Multiple approaches may be used to incorporate auxiliary terrain information (Penížek & Borůvka, 2006). However, gathering terrain information utilizing digital techniques of the earth's surface is generally deposited in a repository that is guided by computers. The data classes typically obtained by researchers are presented in several categories employed in quantifying mapping PTEs, soil properties, and so on (Penížek et al., 2016). The technique for acquiring data for covariates generally considers the ground height, density, observation coordinates, and the GIS-based algorithm. In this thesis, the use of spectral indices was considered and explored to determine the feasibility of using a mathematical equation based on a remote sensing dataset.

1.9 Remote sensing

The "art and science of deriving information from measurements made at a distance" has been defined as remote sensing (Colwell, 1997). Measurements of electromagnetic radiation from the earth's surface are made in two ways: passive and active. Passive remote sensing collects electromagnetic data generated by the interaction of the sun's energy and surface materials, such as measurements collected by satellite sensors. Active remote sensing gathers data from the earth's surface because of an emitted signal, such as LiDAR (Light Detection and Ranging) or radar. Spectral data remote sensing provides direct information about the surface properties of soils, vegetation, or other materials. Remotely sensed spectral properties at the surface can be linked to environmental covariates that influence soil development. As a result, the spectral properties can potentially be used to infer other soil characteristics. Remote sensing data can be used to map variations in relief, climate, organisms, parent material, and even time (indirectly). The spatial detail, spectral wavelengths of imagery, and even the season of the year or other temporal aspects of the physical environment that influence data acquisition timing should all be considered. Satellite images can be obtained from free hub such as the European Space Agency's Copernicus Open Access Hub and the earthexplorer.usgs.gov. Some of the imageries that can be

extracted are Landsat and Sentinel. The popularly used imageries are the Sentinel 2 and the Landsat 8.

1.9.1 Image fusion

The goal of image fusion was to combine multiple input images into a more informative single composite image. Fusion usually blends low to medium spatial resolution hyper/multispectral images with high spatial resolution panchromatic images to produce an image that retains both the spectral and spatial resolution of the hyper/multispectral and panchromatic images. Depending on the fusion stage, image fusion is performed at three different levels (Pohl et al., 1998):

1- Pixel level 2- Feature level, 3-Decision level

The pixel level is the image composition's lowest processing level. The most popular and effective image fusion techniques at the pixel level are Hue, Intensity, Saturation (IHS), Gram Schmidt (GS), Principal Component Analysis (PCA), and wavelet.

1.10 Bivariate mapping

Bivariate mapping refers to the method of characterizing spatial objects including grid cells or area polygons based on the values of two variables (Speich et al., 2015). By visualizing the two variables as a single output employing a single-color legend, a bivariate color scheme is generated. A bivariate map depicts the spatial relationships between two raster layers (Tyner, 2010). Spatial relationship can then be examined as a single output map for use in a range of applications. When two variables have a spatial relationship, it suggests that they are interdependent. Beard and Mackaness (2006) express similar points of view in the scenario of uncertain spatial analysis, in which the feature and a method for analyzing its predictive ability are highly symbolically depicted in a bivariate map. Moreover, multiple studies have compared and demonstrated that the efficiency of bivariate maps varies, and the results vary depending on the map reader's knowledge and experience (Hope & Hunter, 2013; Roth, 2013). For more information on the bivariate mapping procedure, see Kebonye et al. (2022) and Trumbo (1981).

The map function would generate a bivariate map that incorporated spatially distinct features from both layers.

1.11 Knowledge deficit

Considering the above literature, spatial prediction of PTEs and their effects in agricultural soil has always been a challenge. The premium placed on agricultural soil due to its use for crop production has thus pushed a lot of research traffic in the direction of new ways to abate or reduce the influx of PTEs into agricultural soil to a tolerable minimum. DSM has bridged the gap in terms of improving the availability of updates and quantitative and reliable soil data and information to support decision-making in relation to sustainable soil management. Prediction of PTEs over the year using DSM incorporates a wide range of environmental covariates, including remote sensing, terrain attributes, a digital elevation model, and other allied auxiliary datasets that aid in the prediction of PTEs in the soil. Nonetheless, the potential for determining PTEs in agricultural soil by combining spectral indices estimated from remote sensing datasets with terrain attributes remains untapped. Spectral indices estimated from satellite imagery for the spatial prediction of PTEs in soil require a relationship with allied ancillary datasets with representation of the earth's surface, such as DEM, because they contain information about the elevation of geological (ground) features such as valleys, mountains, and landslides, to name a few. This provides details information needed in ensuring good predictions that are useful to decision makers and end user.

PTEs accretion in soil, as an important factor influencing soil structure and function, has a significant impact on cultivated land quality, causing soil compaction and nutrient loss, resulting in a decrease in agricultural product production and quality (Zhao et al., 2017). A high concentration of PTEs in agricultural soil reduces the productivity of agriculture, the microbial activity in the soil, makes the soil infertile, and it enters the food chain. (Toth et al. 2016, Vácha, 2021). More specifically, elevated PTEs in the soil have a higher tendency to be harmful to human health, soil flora and fauna. Different crop production methods can have varying effects on PTE uptake from soil to plant, posing various health risks to residents via the food chain (Antoniadis et al., 2017, Liu et al., 2013, Zhuang et al., 2009). The current concept of health risk assessment

in all areas of interest is largely based on the mean values of the study area, as well as the maximum and minimum values. These types of health risk assessments do not provide a complete picture of the health status of the study area, but rather a good idea of it. In contrast, using a sample location approach to assess health in an area of interest is a paradigm shift in health risk assessment that provides a comprehensive overview of the area of interest as well as detailed information on health risk status per each sampled location.

Understanding, analyzing, and controlling PTEs pollution requires knowledge of PTEs sources. Principal component analysis combined with multiple linear regression, UNMIX, chemical mass balance, and geostatistics combined with geographic information system (GIS) techniques have all been used to assign PTE sources (Fei et al., 2019). Positive matrix factorization (PMF), recommended by the United States Environmental Protection Agency (USEPA, 2014), is an ideal receptor model that can quantitatively calculate the contributions of potential sources to soil PTE contamination at each data point under nonnegative constraints and data uncertainty (Huang et al., 2018). Moreover, according to Paatero et al. (2014) and Brown et al. (2015), the application of the PMF receptor model has drawbacks, resulting in disparities between measured and predicted PTE content, which impacts factor contributions. One of the best ways to improve model performance in DSM, and thus reduce errors and improve prediction efficiency, is to hybridize the modeling approach with a different model. Based on this context, ecological risk was combined with PMF to produce ER-PMF, which provides less differences and has a higher likelihood of reducing errors between measured and predicted PTE content. This practice is novel and provides an appropriate orifice to reduce errors resulting from source distribution assessment to the minimal level.

Some metals, such as zinc (Zn) and iron (Fe), play important roles in crop production because they are essential nutrients for plant growth, development, and productivity. Elevated levels of metals in the soil, such as Zn, have a devastating effect on plant growth. It is common practice to use alternative auxiliary datasets to predict these PTEs in soil. However, the combination of pretreatment methods used in pretreating spectral datasets is uncommon. More importantly, the interaction of macronutrients and micronutrients in the soil is critical for maintaining soil balance for crop growth. Nonetheless, leveraging the stimulating and antagonistic effects of

macro- and micronutrients in soil, as well as a visible near-infrared spectroscopy dataset, for the prediction of PTEs in soil, particularly zinc, is unexplored terrain.

Monitoring agricultural soil on a regular basis to ensure that it is in good condition for agricultural production is critical. The collection and use of legacy data collected from national agencies and allied bodies that share pollution monitoring goals has gone a long way toward assisting in the monitoring of soil pollution levels and providing pragmatic and realistic solutions to the problem. Legacy data is frequently used in the prediction and mapping of PTEs in agricultural soil to monitor pollution levels and soil quality. The sampling regime appears to be wide at times, capturing few polluted areas during the sampling process, due to the vastness of the area sampled for periodic monitoring. The spatial process and sampling location are not thought to be stochastically independent in preferential sampling. When the area of interest is deemed polluted or has a peculiar problem, this sampling procedure is initiated. There have been studies that use either preferentially sampled or legacy data. What has not previously been done is the combination of preferential sampling datasets and legacy data with the goal of improving prediction performance. Preferential sampling in polluted areas within a larger area to supplement legacy data for soil quality and pollution monitoring improves prediction, identifies high polluted areas, and provides tailored solutions based on pollution outcomes.

Fusion is commonly characterized in remote sensing as the integration of two or more images with varying spectral and spatial features (Khosravi et al. 2022). As a result, the fusion product includes all the characteristics of both single images, making it more informative (Palsson et al., 2018). The fusion procedure must retain the resulting fused image's spectral and spatial resolutions while avoiding spectral and/or spatial distortion (Qu et al., 2018). Moreso, most study have applied data fusion of remote sensing imageries in prediction of PTEs or soil properties. Moreover, it is rare combining terrain attributes to data fusion of Sentinel 2 and Landsat 8 in the prediction of PTEs in agricultural (Agyeman et al. 2022). The geological terrain is an important and influential factor in predicting PTEs such as Sb in soil (Agyeman et al. 2022). Given the circumstances of pedogenesis and the evolution of development, environmental covariates have the greatest influence on the impactful categorization of the spatial variability of PTEs in soil (Zeraatpisheh et al., 2020). PTEs enrichment in agricultural soils results from a combination of

anthropogenic and natural processes, including parent material weathering and subsequent pedogenesis (Agyeman et al. 2022). Therefore, there is a dearth in combining the remote sensing dataset along with terrain attributes to harness the potential in prediction of PTEs in the soil. Nonetheless, the use of regression kriging with a combined auxiliary dataset from Sentinel 2 and Landsat 8 (data fusion), as well as terrain attributes, to improve PTE prediction efficiency has never been explored.

2.0 OBJECTIVES & HYPOTHESES

Paper 1: Using spectral indices and terrain attribute datasets and their combination in the prediction of cadmium content in agricultural soil

Hypothesis: Using prediction models with reasonable accuracy, soil pollution can be spatially predicted.

Objectives: The goals of this study are to (i) determine the variability of prediction Cd in agricultural soil using spectral indices or terrain attributes coupled with modeling algorithms and (ii) determine whether combining spectral indices and terrain attributes coupled with modeling algorithms can improve Cd prediction efficiency in agricultural soil.

Paper 2: Human health risk exposure and ecological risk assessment of potentially toxic element pollution in agricultural soils in the district of Frýdek-Místek, Czech Republic: a sample location approach

Hypothesis: Increased levels of PTEs in a study area beget carcinogenic and non-carcinogenic health risk exposures.

Objectives: The primary objective of this paper is to create a digitized soil map that highlights the human-related health risks posed by PTEs, as well as to estimate and map pollution indices outputs, the pattern of PTE spatial distribution, source apportionment, and determine carcinogenic and non-carcinogenic health exposures using a sample location approach.

Paper 3: Ecological risk source distribution, uncertainty analysis, and application of geographically weighted regression cokriging for prediction of potentially toxic elements in agricultural soils.

Hypothesis: The impact of agriculture and industry on soil health in the study area can be ascertained through the use of reliable pollution indices and multivariate statistics.

Objectives: The specific objectives of this paper are (i) to determine the environmental risk level of the study area, (ii) evaluate ER-PMF (ecological risk-positive matrix factorization) and PMF (positive matrix factorization) receptor models for estimating PTE source allotment, (iii) employ ecological risk-assessed PTE values to calculate PCA and a correlation matrix, (iv) estimate the

uncertainty based on the receptor models and assess the efficiency of the prediction of PTEs based on geographical weighted regression or a hybridized model.

Paper 4: Optimal zinc level and uncertainty quantification in agricultural soils via visible near-infrared reflectance and soil chemical properties

Hypothesis: The relationship between soil properties, pollutant contents, and auxiliary datasets allows for the application of reliable models to detect the concentration of these elements (PTEs).

Objectives: The study's objectives are to (i) quantify the concentration of Zn in cultivated soil based on a series of MLAs coupled with Vis-NIR spectral reflectance; (ii) determine whether combining Vis-NIR, soil chemical properties, and MLAs in the estimation of Zn content in agricultural soil will improve prediction accuracy; (iii) determine the level of uncertainty that will be propagated in both contexts, and (iv) evaluate the performance of a single pretreated method versus a combination of pretreatment methods.

Paper 5: Quantification of the optimal cadmium level in agricultural soil using legacy data, preferential sampling, Sentinel 2, Landsat 8 coupled with ensemble model.

Hypothesis: The combination of legacy data and data from preferential samples will improve soil pollution spatial prediction over a larger area.

Objectives: The study's specific objectives are to compare the prediction of Cd in agricultural soil using two distinct Cd datasets; (i) to apply the different spatial resolution of remote sensing datasets to the prediction of Cd in agricultural soil; (ii) to assess the propensity of ensemble models coupled with diverse Cd datasets and remote spatial resolution datasets; (iii) and (iv) finally, to assess uncertainty using ensemble-sequential gaussian simulation (EnSGS).

Paper 6: Prediction of the concentration of antimony in agricultural soil using data fusion, terrain attributes combined with regression kriging.

Hypothesis: The use of data fusion, terrain attributes, and a hybridized modeling approach improves prediction over using either data fusion or a terrain attribute coupled modeling algorithm for prediction.

Objectives: The specific goals of this study are as follows: (i) apply data fusion coupled with regression kriging approaches to the estimation of Sb concentration in agricultural soil (scenario 1); (ii) add terrain attributes to data fusion datasets combined with regression kriging techniques to estimate Sb content in agricultural soil (Scenario 2); (ii) compare scenario 1 and scenario 2, and (iv) map the uncertainties propagated by both scenarios.

3.0. METHODOLOGY

3.1. Study area

The study area is located in the Frýdek-Místek District within the lower reach of the Moravian-Silesian Region in the Czech Republic, Europe. The community is a combination of previous two independent towns, specifically Silesian Frýdek and the Moravian Místek. The area under study is positioned within the geographical coordinates ranging from Latitude 49° 41' 0" to 49° 50' 0" North and Longitude 18° 10' 0" to 18° 50' 0" East at an altitude ranging between 225 and 327 m above sea level, characterized by a cold temperate climate and a high amount of rainfall even in dry months. In Frýdek-Místek, the summers are hot and partly cloudy, and the winters are cold, dry, windy and mainly cloudy (Weather Spark, 2016). Temperatures range mostly from -5°C to 24°C throughout the year, rarely falling below -14°C or rising over 30°C, while the average annual precipitation ranges between 685 and 752 mm (Weather Spark, 2016). The area survey of the district is projected at 1208 km^2, with 39.38% of the land allocated for agricultural activities and 49.36% for forest land. The farmland within the study area is close to the Ostrava city in which the steel industry is located, and therefore it becomes a critical area for the evaluation of PTEs distribution and soil quality within and around surrounding communities. The PTEs pollution in the area is caused by atmospheric deposition emitted from the steel industry nearby, vehicular emission, abrasion from tires, and agricultural activities (e.g., pesticide and insecticide applications) (Agyeman et al., 2020). Major soil types are primarily Cambisols and Stagnosols (Kozák et al, 2010). These soils dominate the Czech Republic and are found at mean elevation ranges of 455.1 m for Stagnosols and 493.5 m for Cambisols. (Vacek et al., 2020). According to WRB, (2015), Cambisols cover about 1.5 billion hectares worldwide, and its reference soil group principally is well represented in the boreal and the temperate regions. The soils are primarily composed of colluvial, alluvial, or aeolian deposits. A cambic diagnostic horizon characterizes them with fine sandy loam texture, clay content of >4 % with less carbonate content by a lithic discontinuity (Kozák et al., 2010).

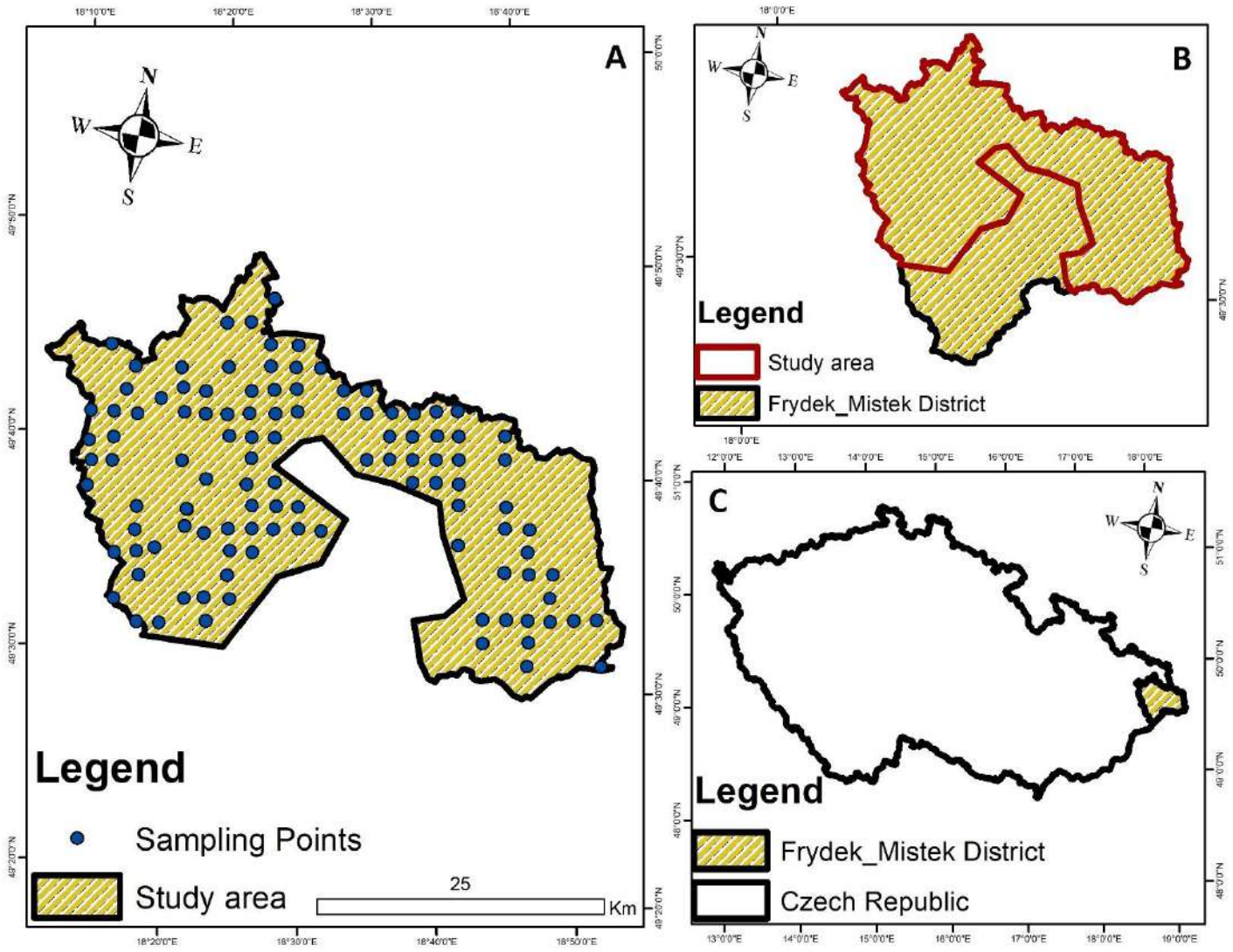


Figure 1. Study area and locations of the sampling points

3.1.2. Soil sampling and analysis

A total sample of 115 topsoil's was collected from agricultural land in the district of Frýdek-Mistek. A regular grid was the sample pattern adopted, and the soil sample intervals were 2 by 2 km using a handheld GPS unit (Leica Zeno 5 GPS). Samples were collected using a steel auger to the depth of 0 to 20 cm for topsoil. The samples obtained was packaged in Ziploc bags, correctly labelled, and transported to the laboratory. The samples were air-dried, crushed by a mechanical device (Fritsch disk mill pulverize) and then sieved (< 2 mm) to obtain a pulverized sample (Pavlů et al., 2018). A gram of the dried, homogenized, and sieved soil sample (sieve size <2 mm) was

inserted in a Teflon bottle and well labelled. 7 mL of 35% HCl and 3 mL of 65% HNO_3 (use automatic dispensers - a special dispenser for each acid) were dispensed in each Teflon bottle and gently closed the cap to enable the sample to remain overnight for reactions (aqua regia procedure). Then the mixture was placed on a hot metal plate for 2 hours to stimulate the process of digestion of the sample and left to cool. The mixture was transferred to a prepared 50mL volumetric flask and then diluted with deionized water to 50 mL. The diluted supernatant was then filtered into 50mL PVC tubes. Also, 1 mL of the diluted solution was further diluted with 9 mL of deionized water and filtered into a 12 mL test tube prepared for PTE pseudo-concentration of the PTEs in this sample. PTEs (As, Cd, Cr, Cu, Mn, Ni, Pb, Zn, Sb) concentration was measured by ICP-OES (inductively coupled plasma optical emission spectrometry) (Thermo Fisher Scientific company, USA) in compliance with standard procedures and protocols. The quality assurance and control (QA/QC) procedure was ensured by assessing the standard reference material for each sample (SRM NIST 2711a Montana II soil). PTEs with low or half detection limits were excluded from this study. The detection limits of the PTEs used in this investigation are 0.0002 mg/L (Cd), 0.0007 mg/L (Cr), 0.0060 mg/L (Cu), 0.0001 mg/L (Mn), 0.0004 mg/L (Ni), 0.0015 mg/L (Pb), 0.0067 mg/L (As), 0.0082 mg/L (Sb) and 0.0060 mg/L (Cu and Zn). Furthermore, the quality control and quality assurance process were ensured for each analysis by evaluating the reference criteria. Duplicate analysis was performed to guarantee that the error was minimized.

3.2. Modeling approaches

We applied the diverse modelling approach from the geostatistical, MLA and a hybrid model in the number of research undertaken. The varied techniques employed are empirical Bayesian kriging, ordinary kriging, inverse distance weighting, random forest, cubist, support vector machine regression, and self-organizing map. Furthermore, the following algorithms were applied, Bayesian regularized neural network, regularized random forests, conditional inference forest, extreme gradient boosting, gaussian process regression, multivariate adaptive regression splines, partial least square regression, Bayesian generalized linear model, M5 tree model, support vector machine regression, cubist M5 tree model and quantile regression forest. In addition, we applied a hybrid algorithm that hybridizes ordinary kriging with machine learning

models (random forest, cubist, conditional inference forest and extreme gradient boosting) to create a regression kriging approach.

3.3. Model performance

The performance of the models chosen for this investigation was assessed. The models were trained using 75% of the dataset (86 observation points), and then validated using the remaining 25% of the dataset (i.e., 29 observation points). The model's performance was assessed using mean absolute error (MAE), root-mean-square error (RMSE), median absolute error (MdAE), concordance correlation coefficient (CCC), ratio of performance to interquartile distance (RPIQ), and coefficient of determination (R^2)

3.4. Pollution assessment

3.4.1 Single pollution index (PI)

The Single Pollution Index (PI) is an index for determining which PTEs poses the greatest threat to a soil environment. Tomlinson et al. (1980) introduced the PI, and the equation is given as

$$\mathrm{PI} = \frac{\mathrm{C_n}}{\mathrm{B_n}}$$

Where B_n is the geochemical background value of the PTEs in the soil (mg/kg) and the C_n is the concentration of the PTE in the soil (mg/kg). PI is categorized into a low level ($PI \leq 1$), moderate level ($1 < PI \leq 3$), considerable level ($3 < PI \leq 6$), or high level ($PI > 6$).

3.4.2 Pollution load index (PLI)

The PLI is often used to measure the average amount of soil pollution indices. This index provides a direct way to display the soil deterioration resulting from the accumulation of PTEs. Tomlinson et al. (1980) introduced this equation, and the equation is given as

$$\mathrm{PLI} = \sqrt[\mathrm{n}]{\mathrm{PI_1} \times \mathrm{PI_2} \times \mathrm{PI_3} \times \ldots\ldots \times \mathrm{PI_n}}$$

Where n represents the number of analyzed PTEs, PLI is categorized into a low level ($PLI \leq 1$), moderate level ($1 < PLI \leq 2$), high level ($2 < PLI \leq 5$), or extremely high level ($PLI > 5$) based on the degree of pollution.

3.4.3 Ecological risk assessment (ER and RI)

Risk index (RI) is the index for determining the degree of ecological risk caused by soil concentrations of PTEs. The index (RI) was introduced and applied by Hakanson (1980) and the equation is given as

$$\mathrm{RI} = \sum_{i=1}^{n} E_r^i$$

In which n is the number of PTEs and E_r^I is the single index of the ecological risk index factor, which the equation is given by

$$E_r^I = T_r^i \times \mathrm{PI}$$

The T_r^i is the toxicity response coefficient of specific PTE (Hakanson, 1980) and the PI represents the single pollution index. The toxicity response coefficient of the PTEs used are 30 (Cd), 10 (As), 5 (Cu), 5 (Pb), 2 (Cr), 2 (Zn), 2 (Ni) and 1 (Mn) (Håkanson 1980). The Er has five classifications: low risk (Er ≤ 40), moderate risk (40 < Er ≤ 80), considerable risk (80 < Er ≤ 160), high risk (160 < Er ≤ 320), and very high risk (Er ≥ 320). The RI has four categories, namely, low risk (RI ≤ 150), moderate risk (150 < RI ≤ 300), considerable risk (300 < RI ≤ 600), or very high risk (RI > 600).

3.5. PMF model

Positive Matrix Factorization (PMF), EPA–PMF v5.0 (U.S. EPA, 2014) is a statistical method used to measure the contribution of the source of samples to the composition or fingerprints of the source. The U.S. Environmental Protection Agency uses this receptor model, developed by Paatero (1997) and Paatero &Tapper (1994). The model does not require any profile source and all the data is weighted by using uncertainty. According to Norris et al. (2008), PMF is used mainly in solving source contributions and source profile that is dataset composition based which is given by this equation

$$X_{ij} = \sum_{k=i}^{p} (g_{ik} f_{kj} + e_{ij})$$

In which X_{ij} is the ith elemental concentration measured in the jth sample p represents the factor number, f the source profile species, g the sample contribution, j and i are the number of samples

and chemical species, and e_{ij} denotes the species. Where K= 1 in the p source, i = 1 of the elements and j = 1 of the samples.

The determination of the contribution, as well as profiles factors, is given by this equation

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{X_{ij}}{U_{ij}} \right)^2$$

where *m* denotes the number of PTEs investigated, *n* signifies the number of soil samples, and U_{ij} means the uncertainty of PTEs *j* in soil sample *i*. U_{ij} is determined based on the PTEs content (C_{ij}), the relative standard deviation (σ) (that is standard deviation divided by the mean), and the method detection limit (C_{MDL}). Therefore, it implies that the PTEs content is above C_{MDL} value, U_{ij} is computed as:

$$U_{ij} = \sqrt{(\sigma \times C_{ij})^2 + C_{MDL}^2}$$

PMF model recommends that the data below the detection limit would be substituted with the value of $C_{MDL}/2$, i.e., data that does not occur in this study and the associated uncertainty is calculated as:

$$U_{ij} = {}^{5}/_{6}\, C_{MDL}$$

Moreover, the constraint of no significant negative contribution (Gik), the maximum optimal factors were derived using the multilinear engine algorithm in PMF. It is noteworthy to note that the minimum *Q* can be global or local. Consequently, multiple attempts using diverse starting points were carried out to reach the global minimum *Q* and reliable solution.

3.6. Health risk assessment

The ever-growing human population and human endeavour to ensure that the planet remains heaven for humanity are under constant constraint. Frequently, scientists, policymakers, and other stakeholders push the limits of research in several ways. However, no matter the initiative and the best course of utilizing research, the world is now and then polluted. Humans are exposed to PTEs in three different forms every day, including inhalation, ingestion, and dermal

contact (Wang et al., 2017). The following equations determine the exposure pathways to humans by PTEs (see Table 1).

$$\mathrm{CDI_{ing}} = \frac{\mathrm{C \times IR_{ing} \times EF \times ED}}{\mathrm{BW \times AT}}\ 10^{-6}$$

$$\mathrm{CDI_{inh}} = \frac{\mathrm{C \times\ IR_{inh} \times EF \times ED}}{\mathrm{PEF \times\ BW \times AT}}$$

$$\mathrm{CDI_{derm}} = \frac{\mathrm{C \times SA \times AF \times ABS \times EF \times ED}}{\mathrm{BW \times AT}} \times\ 10^{-6}$$

$$\mathrm{CDI_{total}} = \mathrm{CDI_{ing}} + \mathrm{CDI_{inh}} + \mathrm{CDI_{derm}}$$

The parameters CDI_{ing} (chronic dialy intake-ingestion), CDI_{inh} (chronic dialy intake-inhalation) and CDI_{derm} (chronic dialy intake-dermal) and reference values of the indices of the above equations are listed in Table 1.

Table 1: Exposure factors used in CDI estimation for non-carcinogenic and carcinogenic risk.

Variables	Description	Units	Values	Values	
			Adults	Children	
C	Concentration of PTEs of present study	mg/kg			
IRing	Ingestion rate	mg/d	100	200	US EPA, 2011
IRinh	Inhaling rate	m^3/d	20	7.65	USEPA 1991
EF	Exposure frequency	days/year	350	350	US EPA, 2011
ED	Exposure duration	year	24	6	US EPA, 2011
SA	Skin surface area	cm^2	1530	860	Eziz et al., 2018
AF	Soil adherence factor	$mg/cm^2/d$	0.07	0.2	US EPA, 2011
ABS	Dermal absorption factor		0.001	0.001	US EPA, 2011
PEF	Particle emssion factor	m^3/kg	1.36×10^9	1.36×10^9	US EPA, 2011
BW	Average body weight	kg	70	15	US EPA, 2013
AT N-C	Average time for non-Carcongenic risk	day	ED × 365	ED × 365	Wang et al., 2017; Eziz et al., 2018;
AT Ca	Average time for non-Carcongenic risk	day	70 × 365	70 × 365	Wu et al., 2019
CF	Units conversion factor	kg.mg-1	1×10^{-6}	1×10^{-6}	US EPA, 2002
Specific reference dose for ingestion	RfD_o	mg/kg/day	Cd (1×10-3), Cr (3×10-3), Cu (4.0×10-2), Ni (2×10-2), Pb (3.50×10-3), Zn (3×10-1), As (3×10-4)and Mn (0.14)		Li et al. 2015; USDOE 2011; Qing et al. 2015; De Miguel et al. 2007; Teng et al., 2015
Specific reference dose for dermal contact	RfD_{ABS}	mg/kg/day	Cd (5×10^{-5}), Cr (6×10^{-5}), Cu (1.2×10^{-2}), Ni (5.4×10-3), Pb (5.3×10-4), Zn (6×10-2) and Mn (0.05)		

Specific reference dose for inhalation	RfD_i	mg/m^3	Cd (1×10-3), Cr (2.86×10-5), Cu (4.02×10-2), Ni (2.06×10-2), Pb (3.52×10-3), Zn (3×10-1)and Mn (0.8)
Oral slope factor	SF_o	$((mg/kg/day)^{-1})$	Cd (15), Cr (0.5), Ni (0.84), Pb (0.28), and As (1.5)
Absorbed dose slope factor	SF_{ad}	$((mg/kg/day)^{-1})$	Cd (15), Cr (0.5), Ni (0.84), Pb (0.28), and As (3.66)
Inhalation slope factor	SF_i	$((mg/m^3)^{-1})$	Cd (15), Cr (0.5), Ni (0.84), Pb (0.28), and As (15.1)

3.6.1. Non – carcinogenic risk assessment

The equation of the potential non-carcinogenic risk for a single PTE was computed as the hazard quotient (H.Q), which is given by the equation:

$$\mathrm{HQ} = \frac{CDI_i}{\mathrm{RfD}}$$

Where RfD represents the reference dose (mg/kg/d), the estimated daily exposure to the human population and i is the exposure pathway (soil ingestion, dermal, or inhalation). The determination of comprehensive health risk of all the PTEs studied was done by computing HQ values. The values were summed up and expressed as the hazard index (HI), which is given by equation 14 (US EPA, 1989) :

$$\mathrm{HI} = \sum \mathrm{HQ} = \mathrm{HQ_{ing}} + \mathrm{HQ_{inh}} + \mathrm{HQ_{derm}}$$

Whereby HQ_{ing}, HQ_{inh} and HQ_{derm} represent the hazard quotient for ingestion, inhaling and dermal, respectively. A report from USEPA, (2002) explicitly outlined that when the HI < 1, then it presupposes that there is a potential to impact health if humans exposed to PTEs negatively. However, Eziz et al., (2018) mentioned that if HI > 1, there is also the propensity for non-carcinogenic health risks to emerge.

3.6.2 Carcinogenic risk assessment

The US EPA, (1989) report stated that the likelihood of cancer of any kind developing might be attributed to humans being exposed to carcinogenic risk (CR).

$$\mathrm{CR} = CDI_i \times \mathrm{SF}$$

$$\mathrm{TCR} = \sum \mathrm{CR} = \mathrm{CR_{ing}} + \mathrm{CR_{inh}} + \mathrm{CR_{derm}}$$

In which CR, TCR and SF values represent carcinogenic risk (no unit), total carcinogenic risk (no unit) and slope factor for carcinogenic PTEs (mg/kg/d), respectively. TCR values should range from 1×10^{-6} to 1×10^{-4}. That is the tolerable standard that proves no significant health threat to humans (Hu et al., 2012).

3.7. Spectral indices

A spectral index is a mathematical equation that is applied to an image's various spectral bands per pixel. The bands considered necessary for the computation of spectral indices such as clay mineral ratio (CLAYMR), ferrous mineral ratio (FMR), iron oxide ratio (IOR), carbonate normalized ratio (CNR), rock outcrop normalized ratio (RONR), and normalized difference built-up index (NDBI) were used to estimate the indices.

Clay Minerals Ratio (CLAYMR) is a geological index for identifying mineral features containing clay and other minerals, such as alunite, using two shortwave infrared (SWIR) bands (Drury. 1987, (Segal 1982, Kienast et al. 2017).

Ferrous minerals (FMR) are a geological index that uses the shortwave infrared (SWIR) and near-infrared (NIR) bands to identify rock features that contain some amount of iron-bearing minerals (Segal 1982, Kienast et al. 2017).

Iron oxide (IOR) index is a geological index that uses red and blue bands to identify rock features that have been oxidized by iron-bearing sulfides (Segal 1982, Kienast et al. 2017).

Carbonate normalized ratio (CNR) is a geological index that uses red and green bands to identify carbonate features that contains the calcium carbonate-bearing minerals (Segal 1982, Kienast et al. 2017).

Rock outcrop normalized ratio (RONR) is a geological index that uses green and short-wave infrared bands to identify sedimentary features that contains the sedimentary (bright pixels) verse igneous (dark pixels) parent material (Segal 1982, Kienast et al. 2017).

Normalized difference built-up index (NDBI) uses NIR, and SWIR bands are used in the built-up Index (NDBI) to highlight manufactured built-up areas. Urbanization is one of the most conspicuous soil pollutions because it includes anthropogenic causes, caused by human activities such as: urban fabric, industrial, commercial, and transportation units; mines, dumps, and construction sites; sports and leisure facilities, which endanger functional diversity; and a variety of environmental and spatial planning issues such as urban sprawl, food safety, community vulnerability to climate change, and pollution (soil, air, water, and noise). The intent of including this index is to determine the impact of built-up infrastructure on soil pollution (Zha et al. 2003). The spectral index formulas are given as

$$CLAYMR = \frac{SWIR\ 1}{SWIR\ 2}$$

$$\mathrm{FMR} = \frac{SWIR}{NIR}$$

$$\mathrm{IOR} = \frac{RED}{BLUE}$$

$$CNR = \frac{RED - GREEN}{RED + GREEN}$$

$$\mathrm{RONR} = \frac{SWIR\ 1 - GREEN}{SWIR\ 2 + GREEN}$$

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR}$$

(SWIR - short wave infrared, NIR - near infrared)

3.8. Data fusion

The Gram-Schmidt (GS) data fusion approach, which is based on an orthogonal vector algorithm (Khosravi et al. 2022), was used in this study. In this method, all images are converted to vector imagery while retaining the same pixel dimension at a transformed high spatial resolution scale. Thus, the GS data fusion transformation is carried out for the high spatial resolution bands (Laben et al., 2000). In this thesis the images employed Sentinel 2A and Landsat 8-OLI bands. The 20m spatial resolution Sentinel 2A bands 11 and 12 were downscaled to 10m using GS approach to obtain consistent spatial resolution with the band 2, band 3, band 4 and band 8. Similarly, the Landsat 8-OLI bands 2 to 7 were equally resampled from 30m spatial resolution to 10m spatial resolution using GS fusion approach. The Landsat 8-OLI bands 2 to 7 were fused to the 10m Sentinel Bands using GS fusion approach. These bands from Sentinel 2 and Landsat 8 were chosen because possess the same spectral similarities.

3.9. Methodology summary for each paper

3.9.1. Methodology 1

Using spectral indices and terrain attribute datasets and their combination in the prediction of cadmium content in agricultural soil.

The study explores the application of spectra indices estimated from sentinel 2 bands, the application of terrain attributes and the combination of both auxiliary dataset in the prediction of Cd in agricultural soil. The modeling approach was partitioned into three scenarios, comprised of the prediction using terrain attributes coupled with digital soil mapping (DSM) approaches (Scenario 1), prediction using spectral indices combined with DSMs approaches (Scenario 2), and prediction using a combination of terrain attributes, spectral indices, and DSMs approaches (Scenario 3). The study employed six modeling approaches including Gaussian process regression (GPR), partial least square regression (PLSR), extreme gradient boosting (EGB), multivariate adaptive regression splines (MARS), Bayesian regularized neural network (BRNN), regularized random forest (RRF), Bayesian generalized linear model (BGLM), and M5 tree models. The validation of the precision of the modeling approach was determined by the application of concordance correlation coefficient (CCC), root mean square error (RMSE), mean absolute error (MAE), median absolute error (MdAE), and the coefficient of determination (R^2).

3.9.2. Methodology 2

Human health risk exposure and ecological risk assessment of potentially toxic element pollution in agricultural soils in the district of Frýdek Místek, Czech Republic: a sample location approach.

The study performed comprehensive health exposure assessment applying the sample location approach in agricultural soil. The method applied in the study was unparallel to the normal approach where the estimated mean of each PTEs used are applied in the estimation of the carcinogenic and non-carcinogenic health risks. The sample location data from each sample point was used in the estimation of the health risk in other to decentralize the assessment based on the 2 by 2 km of the study area. The carcinogenic and non-carcinogenic effects of the PTEs (i.e., lead, arsenic, chromium, nickel, manganese, cadmium, copper, and zinc) were quantified using the health risk assessment equation. The output of the children's and the adults CDI_{total} (Chronic

Daily Intake total) values for non-carcinogenic risk and carcinogenic risk were mapped to determine the high, moderate and the low exposure degree in the study area. On the other hand, positive matrix factorization (PMF) was applied to determine the source distribution of the PTEs correlating it the potential pollutants within the study area.

3.9.3. Methodology 3

Ecological risk source distribution, uncertainty analysis, and application of geographically weighted regression cokriging for prediction of potentially toxic elements in agricultural soils.

In this study the application of hybridized models was applied to enhance the practicability and the efficiency of PMF and modeling algorithm such as cokriging and geographical weighted regression. A pollution index such as ecological risk was used to estimate the ecological risk of the study area, and the estimated output was applied in the PMF receptor models to compare the results with the parent model PMF. Validation criteria such as RMSE, MAE, and R^2 were used to determine the efficiency and ability to minimize error in the estimation of the source distribution. The study also investigated the feasibility of using geographical weighted regression (GWR) and the hybridization of GWR and cokriging (GWRCoK) in predicting the concentration of PTEs (i.e., lead, arsenic, chromium, nickel, manganese, cadmium, copper, and zinc) in the study area.

3.9.4. Methodology 4

Optimal zinc level and uncertainty quantification in agricultural soils via visible near-infrared reflectance and soil chemical properties.

This study determines the optimal level of Zn in the agricultural soil using two distinct approaches, namely: (1) employing visible near-infrared spectra reflectance along with machine learning algorithms (MLAs) (Context 1), and (2) applying visible near-infrared spectra reflectance, soil chemical properties (SCP), and MLAs (Context 2). As an auxiliary dataset, SCP such as magnesium (Mg), potassium (K), iron (Fe), copper (Cu), and phosphorus (P), which serve as micro and macro nutrients, were used as an auxiliary dataset combined with visible near-infrared spectra reflectance to predict Zn in the soil. The spectral ranges from 350 to 400, as well as 2401

to 2500 nm, were eliminated due to noise, and the spectral range from 400 to 2400 nm was pretreated using Savitzky-Golay filter (SG), logarithmic transformation (log(1/R), standard normal variate (SNV), correction maximum reflectance (CMR), discrete wavelet transformation (DWT), and multiplicative scatter correction (MSC). More so some pretreated techniques were combined to such as DWT-CMR (discrete wavelet transform-correction maximum reflectance), SG-LOG-MSC (savitzky-golay smoothing-logarithm1/R-multiplicative scatter correction). SG-LOG-SNV (savitzky-golay smoothing-logarithm1/R-standard normal variate), SG-SNV-MSC (savitzky-golay smoothing-standard normal variate - multiplicative scatter correction), DWT-SNV-MSC (discrete wavelet transform-standard normal variate-multiplicative scatter correction), DWT-LOG-MSC (discrete wavelet transform-logarithm1/R-multiplicative scatter correction) to determine their applicability and performance compared to the individual pretreated methods. The following MLAs were used: conditional inference forest (CIF), partial least square regression (PLSR), M5 tree model (M5), extreme gradient boosting (EGB), and support vector machine regression (SVMR). The uncertainty of the prediction was mapped based on each context using prediction intervals like mean, 95% and 5%. The figure represents the flowchart of the study (Figure2)

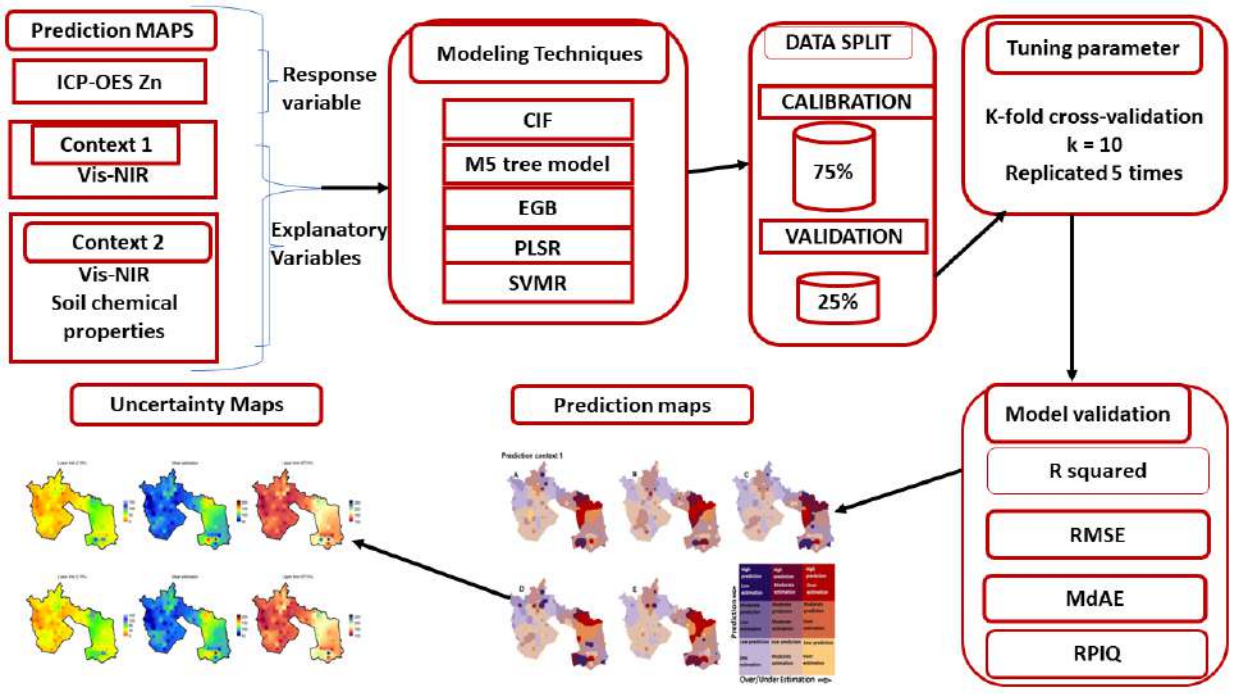


Figure 2. Schematic diagram for the workflow for this study

3.9.5. Methodology 5

Quantification of the optimal cadmium level in agricultural soil using legacy data, preferential sampling, Sentinel 2, Landsat 8 coupled with ensemble model.

The research was based on the application of legacy datasets (LD) and the usage of preferentially sampled datasets plus legacy data (PS-LD). Furthermore, the study also explored the usage of Sentinel 2 (S2) and Landsat 8 (L8) datasets from different spatial resolutions, that is, 10 m and 20 m. In S2, the 20m bands such as band 11 and 12 were downscaled to 10m bands to harmonize with bands 2, 3, 4, and 8. Alternatively, bands 2, 3, 4, and 8 were also upscaled to 20m bands to be in sync with bands 11 and 12. With this, therefore, we obtained two different auxiliary datasets from S2 with 20m and 10m spatial resolution. On the other hand, L8 bands from 2 to 7, which are of 30m spatial resolution, were resampled to 10m and 20m spatial resolution to obtain distinct auxiliary datasets. The resampling, downscaling, and upscaling were done using the bilinear approach in ArcGIS. The modeling approaches used were ensemble models with four sub-models and a meta learner (Figure 3). The modeling uncertainty was also determined using a hybridized model ensemble sequential gaussian simulation (EnSGS). The ensemble models used in this paper were composed of ten distinct algorithms that allow a model to appear twice in the ensemble models created. If a model appears as a meta learner, it will appear once as a sub-model, or altenatively a model appear twice as a sub-model in the four ensemble models created. This enables varying the predictive strength of modeling approaches in ensemble models. Despite this, the same data input was used in each ensemble model during the modeling process.

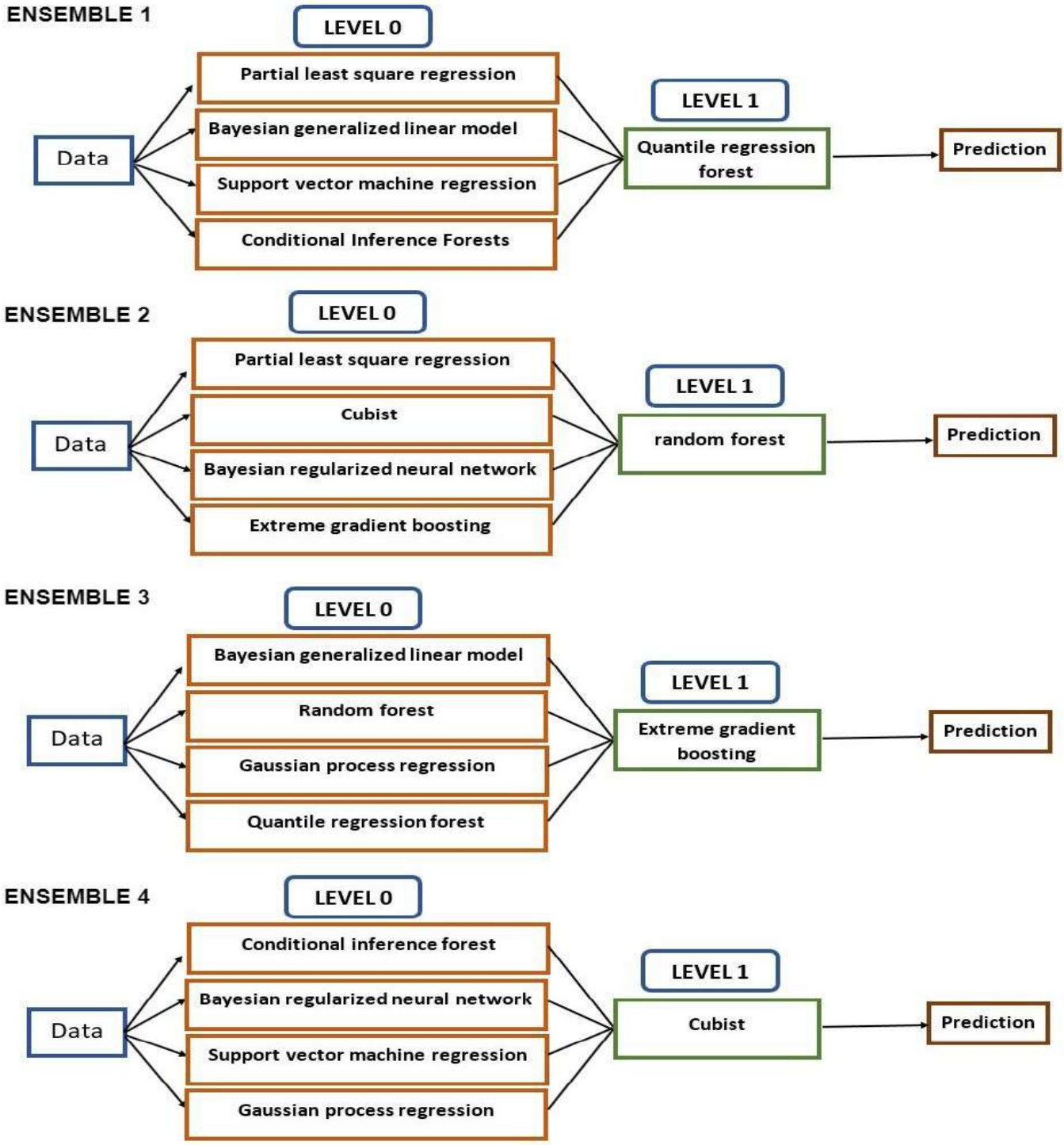


Figure 3. Schematic diagram for the workflow for this study

3.9.6. Methodology 6

Prediction of the concentration of antimony in agricultural soil using data fusion, terrain attributes combined with regression kriging.

This study harnesses the potential of combining remote sensing images such as Sentinel 2 and Landsat 8 into a data fusion. The research also applies terrain attributes in conjunction with the data fusion in the prediction of antimony (Sb) in the agricultural soil. Sb prediction was done using two approaches: prediction using data fusion coupled with regression kriging (scenario 1) and Sb prediction using data fusion, terrain attributes, and regression kriging (scenario 2). The modeling techniques used in the estimation of Sb concentration in agricultural soil included cubist regression kriging (cubist_RK), conditional inference forest regression kriging (CIF_RK), extreme gradient boosting regression kriging (EGB_RK), and random forest regression kriging (RF_RK). The model validation was performed using root mean square error (RMSE), mean absolute error (MAE), bias, and coefficient of determination (R^2). The figure represents the flowchart of the study (Figure 4).

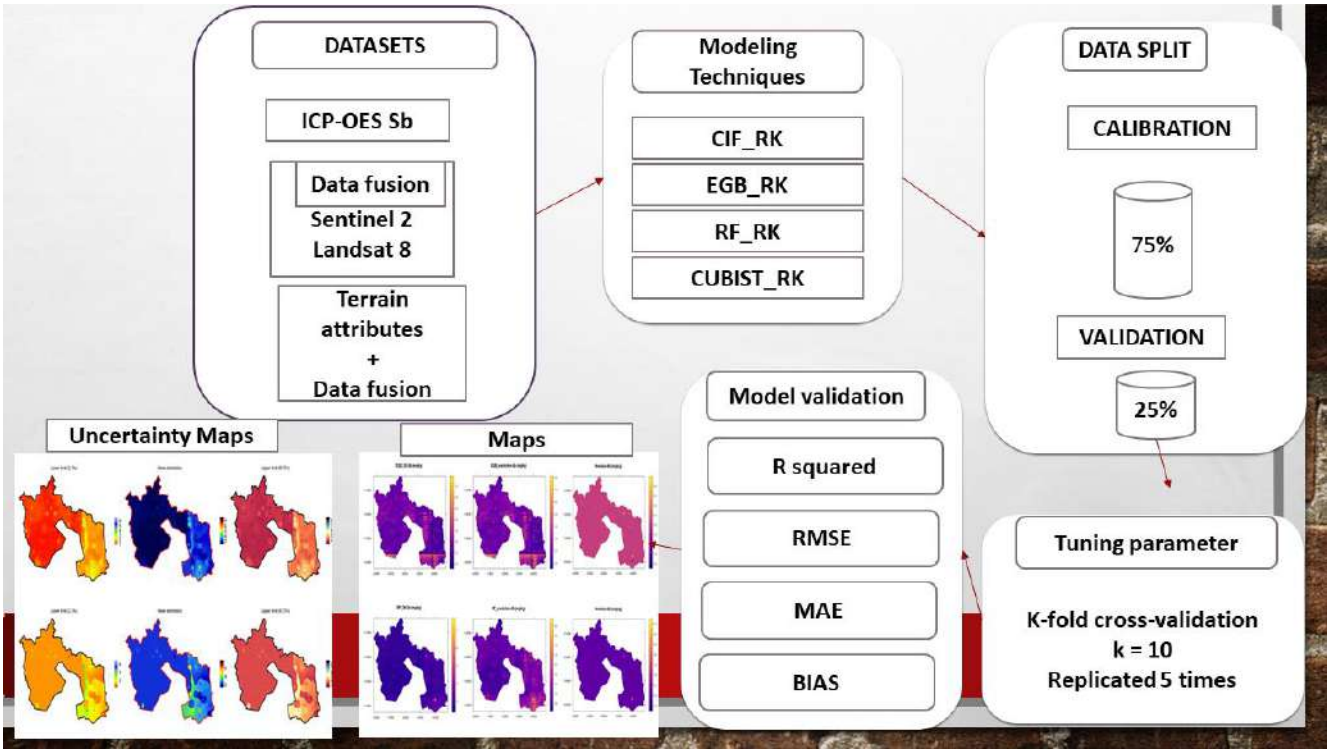


Figure 4. Represents a schematic diagram for the workflow of the study

4. SUMMARY AND CONCLUDING REMARKS

4.1 Summary of key findings and discussion

The general outlook of the study is to determine the spatial prediction of PTEs using digital soil mapping techniques in agricultural soil, with an extended focus on source distribution assessment as well as health risk exposure assessment. The auxiliary datasets applied in this study are Sentinel 2, terrain attributes, Landsat 8, visible near-infrared spectroscopy, and the application of soil properties to aid in the prediction of diverse PTEs in the soil. The study also makes use of a data fusion process that combines satellite images such as Sentinel 2 and Landsat 8 into a composite image to extract ancillary datasets to aid in enhancing as well as obtaining better prediction. The use of this diverse dataset aided in the exploration of various avenues for obtaining a better prediction of PTEs in the study area. In some cases, a combination of remote sensing datasets and terrain attributes were used solitary, combined, and compared to ascertain the best auxiliary dataset along with modeling algorithms in the prediction of PTEs in the soil. It was apparent that the combination of varied auxiliary datasets provided better results than using one auxiliary dataset. This combination increases prediction efficiency and decreases marginal errors.

Paper 1

The results indicate that in scenario 1, the eight modelling techniques employed used terrain attributes as auxiliary datasets in conjunction with Cd data quantified using ICP-OES. MARS had the lowest MdAE (0.21), accompanied by RRF (0.32), PLSR (032), BRNN (035), the M5 tree model (035), EGB (036), and GPR and BGLM (044 and 0.46, respectively). M5 tree modeling produced the lowest RSME (0.45), followed by BRNN (0.48), PLSR (0.51), GPR (0.53), BGLM (0.54), RRF (0.80), EGB (0.83), and MARS (1.29). According to the calculated MAE results, the M5 tree approach had the lowest MAE (0.37), accompanied by BRNN, which had the second lowest MAE (0.40), PLSR (0.42), GPR (0.45), BGLM (0.46), EGB (0.52), RRF (0.55), and MARS (0.66). According to the R^2 values, the M5 tree model produced the highest value of R^2 = 0.77 out of the eight modelling techniques used to predict the concentration of Cd in the soil. Other modeling approaches' R^2 values were within acceptable precision and accuracy ranges, namely 0.73 for

BRNN, 0.71 for PLSR, 0.70 for GPR, 0.69 for BLGM, 0.68 for MARS, and 0.61 for RRF. Only the EGB modeling method performed poorly, with R^2 = 0.47, which is unacceptable. Using the CCC assessment methods, the CCC prediction ranges between 0.41 and 0.73 for the modeling approaches. The M5 tree model had the highest CCC value, while MARS had the lowest. The consolidated results showed that the M5 tree modeling approach combined with terrain attributes and the measured Cd concentration was the best modeling approach for predicting Cd in the soil with the highest prediction efficiency and lowest error. Similarly, in scenario 2, where spectral indices were used as the auxiliary dataset with the same eight modeling approaches, EGB (extreme gradient boosting) was the best modeling approach for predicting Cd in agricultural soil, with the highest CCC and R^2 values of 0.76 and 0.83 and the lowest MAE of 0.33, respectively. In scenario 3 the prediction of Cd in agricultural soil was done using a combination spectral indices and terrain attributes along with similar 8 modeling approaches. The cumulative performance of the modeling approaches in predicting Cd in agricultural soil using spectral indices, terrain attributes, and modeling approaches indicated that the M5 tree (R^2 =0.84, RMSE = 0.39, MAE = 0.31, MdAE = 0.24 and CCC = 0.81) modeling approach is the best approach for predicting Cd with higher precision and a consistent minimal error margin.

When modeling Scenario 1 (prediction based on terrain attributes) and Scenario 3 (prediction based on terrain attributes and spectral indices) were compared, GPR, MARS, BRNN, and BGLM performed better using terrain attributes alone as auxiliary datasets than when combined. The PLSR, EGB, RRF, and M5 tree models, on the other hand, performed significantly better in Scenario 3 than in Scenario 1. When Scenarios 2 and 3 were compared, it was clear that GPR, EGB, MARS, RRF, and BGLM performed better in Scenario 2 than the respective modeling methods in Scenario 3. In Scenario 3, the PLSR, BRNN, and M5 tree models outperformed the respective modeling techniques in Scenario 2. It can be reported that the application of terrain attributes, spectral indices, and the integration of spectral indices as auxiliary datasets has demonstrated the ability of the PLSR and M5 tree model approaches to predict Cd consistently and optimize prediction performance in all scenarios with elevated efficiency and minimized error. According to Kalambukattu et al. (2018), a combination of terrain attributes and spectral indices has the potential to optimize results with high accuracy. Multiple studies, including

Goydaragh et al., (2021), and Xu et al., (2019), proposed and applied diverse auxiliary datasets like spectral datasets with environmental variables, such as terrain attributes, to boost modeling results, particularly in comparison to simply employing spectral datasets or environmental variables. In predicting PTEs such as Cd in soil, geological terrain is an important and influential factor. Soil parent composites are formed because of long-term relationships between bedrock, climatic conditions, and geomorphic mechanisms. Several parameters, including the number of data points, the type of model, the variability of soil properties, and the capabilities of environmental variables to explain soil variations, can all have an impact on model prediction performance (Taghizadeh-Mehrjardi et al., 2020). The M5 tree model has been used in a plethora of research and has proven to produce results with high accuracy and little error, as in the current study. Kumar and Deswal, 2020a, Heddam, 2021, and Sihag et al., 2019, investigated the efficiency of various modeling methods for the assessment of PTEs in soil and discovered that the M5 tree modeling approach was the best model for Cu and Zn assessments. Besides that, Biabani et al. (2016) and Rahimikhoob (2016) used the M5 tree model algorithm to assess the daily reference of evapotranspiration and predict the temporal evolution of clear water and discovered that the M5 tree model method produced satisfactory results when quantifiable metrics such as R^2, RMSE, and MAE were considered, with less deviation from arithmetic values. It can be concluded that using the M5 tree modeling technique with the integration of terrain attributes and spectral indices outperforms the use of spectral indices or the terrain attributes separately in predicting Cd in agricultural soil.The M5 tree model is composed of many tree structures constructed with subsets, and a tree configuration with the fewest errors must be constructed to avoid overfitting (Kumar and Deswal, 2020).

Paper 2

The results revealed that the pollution assessment of the soils in the study area using diverse pollution assessment indexes (pollution index, pollution load index, ecological risk and risk index), based on the application of the local background value and the European average value, displayed a range of pollution levels due to differences in the threshold limits from differing geochemical background levels. The principal components analysis and positive matrix factorization, respectively, identified the sources of pollution and the distribution of PTE sources.

The CDI_{total} (Chronic Daily Intake total) of the PTEs per sampled data implied that children were more exposed than adults regarding the non-carcinogenic risk. Even so, the quantified hazard quotient (HQ) for children based on non-carcinogenic risk appears to be higher than the HQ for adults. The estimated HQs values for PTEs at the minimum and maximum values (both children and adults) are as follows: As > Pb > Cr > Mn > Cd > Ni > Cu > Zn. The study proved that ingestion was the most likely route for PTEs exposure in the study area. The variability of the measured PTEs concentrations per sampled location revealed that the HI (for children) values estimated per 2 X 2 km indicated that 7 of the sampled locations were greater than 1. However, the calculated HI indicated that 6.1% (1.01E+00 to 2.05E+00, or 7 out of 115 sampled locations) of the overall study area posed a high non-carcinogenic risk to children. In terms of carcinogenic risk, the chronic daily intake of Cd, Cr, Pb, Ni, and As was calculated. For the carcinogenic risk, the CDI_{total} for adults and children is given in the following order: Pb > Cr > Ni > As > Cd. Children had higher CDI_{totals} than adults, regardless of PTE computed value.

PTEs can cause cardiovascular disease, poor respiratory function, cognitive deficits, reproductive toxicity, and bone damage in children (Madrigal et al. 2018). The Cr CDI_{totals} for the carcinogenic risk of adults and children were higher than those of the other PTEs. Moreover, the CDI_{total} of children was considerably higher than that of adults. The CR for all PTEs in adults was detected to be substantially lower than in children. According to Agyeman et al. (2021), children are more sensitive to the health effects of PTEs due to oral and finger practice and seem to be extremely susceptible to PTEs. Children's HI values were likwise revealed in the following studies: Agyeman et al. (2021), Han et al. (2020), Natasha et al. (2020), Wang et al. (2020), Bhandari et al. (202), and Zheng et al. (2020). The calculated HI for adults is not statistically significant because it is less than the reference value of 1; this implies that if exposed, a non-carcinogenic adverse impact on an adult is unlikely.

When the current TCR is compared to similar studies conducted in Ostrava, Czech Republic, by Weissmannova et al. (2019), it appears that Pb poses a significant carcinogenic risk to children, Cd poses a moderate risk, and Cr poses a very high risk. This supports the current study's findings that children are more vulnerable to the health risks associated with PTE than adults. In contrast, Kebonye et al. (2021) confirmed recent findings that children in riverine soils are more

susceptible to PTE exposure than adults (Czech Republic). PTEs accumulate in fat tissues and have a negative impact on the functions of the central nervous system, immune and endocrine systems, urogenital and cardiovascular systems, and normal cellular metabolism (Wang et al. 2015, Wang et al. 2013).

Paper 3

Substantial research has used these multivariate statistics in source assessment to evaluate the proportion of PTEs in soil, which include Chen et al. (2015), Tao et al. (2017), Agyeman et al. (2021), and Hossain Bhuiyan et al. (2021b). To enhance precision and reliability, the minimum Q value was reduced to regulate the residual matrix. The system was run 20 times, with run 3 being the crucial point for the factors loadings and discharged. For both receptor models (ER-PMF and PMF), three factors were released, indicating the various percentage contributions or PTE percentage contributions determined in the source distribution analyses. To be selected as a dominant element, PTEs must have a minimum 44.5% or higher percentage contribution in the factor loadings. Cu (64.6%) and Ni (71.1%) dominated factor 1 in the ER-PMF receptor model, while As (77.5%) and Cd (44.7%) dominated in the PMF receptor model. Cr (66.20%) and Mn (71.40%) influenced factor 2 in the ER-PMF receptor model, and Cu (72.6%) and Ni (76.6%) influenced factor 2 in the PMF receptor model. In the ER-PMF receptor model, As (75.7%), Cd (45.9%), and Pb (47.4%) dominated factor 3, whereas in the PMF receptor model, Cr (55%) and Mn (65.6%) dominated factor 3. The R^2, RMSE, and MAE values also indicated that the prediction of Cd and As in agricultural soil by GWRCoK (geographical weighted regression-Cokriging) range from 0.945 to 0.961, compared to 0.636 to 0.713 for GWR (geographical weighted regression). The error margins estimated using RMSE and MAE were 1.272 and 0.749 for GWRCoK, and 2.636 and 1.819 for MAE, respectively.

Arsenic and cadmium had similar spatial distribution patterns in both approaches. The PTE distribution pattern was observed in the southeastern part of the study area and moved anticlockwise to the northwestern area. The maps had moderate to high hotspots, but the GWRCoK map had more intense hotspots than the GWR. Cr and Cu exhibited a similar distribution pattern on the GWR spatial distribution map, principally in the southwestern to northwestern

quadrants. Chromium had more hotspots stretching from northeast to southeast than Cu. Cr and Cu were more concentrated in the maps southwestern to northwestern areas, but Cu also showed a patch of a hotspot in the map's northeastern area. John et al. (2021) achieved optimal results by combining cokriging with Gaussian process regression. Many papers have combined GWR and ordinary kriging, including Kumar et al. (2012), Wang et al. (2012), Ye et al. (2017), and Pereira et al (2018). Ye et al. (2017) compared the effectiveness of geographically weighted regression kriging (GWRK) with multiple linear regression kriging (MLRK) and ordinary kriging (OK) and discovered that combining GWR with geostatistical algorithms such as OK produced better results in predicting soil organic content than MLRK and OK. Besides this, Imran et al. (2015) used GWRK for growth and yield modeling in West Africa, leading to the conclusion that GWRK is superior to KEDLN (KED with a local kriging neighborhood) and regression kriging, with considerably lower prediction uncertainty.

Five of the eight PTEs evaluated (As, Cd, Ni, Pb, and Zn) produced a higher accuracy level (R2) in the ER-PMF approach than in the PMF approach. The closer the R^2 value is to one, the better the prediction accuracy, according to Li et al. (2016), John et al. (2020), and Kebonye et al. (2021). According to Molinaro et al. (2005), determining the error rate or generalizability of the chosen model is a critical process in presenting results. Based on the overall average of R^2, RMSE, and MAE values calculated for the receptor models, ER-PMF had a high R^2 average (0.93) with a low RMSE (2.63) and MAE average (1.55), whereas PMF had a high R^2 average (0.93) with a higher RMSE (13.11) and MAE average (8.20). This means that ER-PMF can identify sources with greater accuracy and less error than PMF in source apportionment. Guan et al. (2019) compared three receptor models (PMF, UNMIX, and grouped principal component analysis/absolute principal component scores (GPCA/APCS)) and concluded that the GPCA/APCS receptor model was optimal based on the estimated R^2 values. In a similar manner, Salim et al. (2019) likened PCA-MLR and PMF, and the authors used R^2 to evaluate which receptor model seemed to be more dependable with high model efficiency; PMF was found to be optimal. Furthermore, Salim et al. (2019) used the Nash-Sutcliffe efficiency and quantified the percentage error, which was previously used by Moriasi et al. (2007) and Yang et al. (2013b) to determine the receptor model with the lowest percentage error while optimizing efficiency.

Paper 4

The results indicate that integrating the CIF modeling method along with Vis-NIR datasets (RAW, CMR, MSC, DWTCMR, SGLOGMSC, SGLOGSNV, SGSNVMSC, DWTSNVMSC, DWTLOGMSC) produced satisfactory results in Zn prediction in agricultural soils. The integration of CIF and the MSC pretreated dataset produced the best overall Zn prediction results (R^2 = 0.70, RMSE = 21.42 mg/kg, MdAE = 9.89, RPIQ = 1.51). The use of PLSR in conjunction with the Vis-NIR spectral dataset produced similar results, apart from CMR and DWTCRM, which produced minimal results. Nonetheless, the DWTLOGMSC dataset, in conjunction with the PLSR (R^2 = 0.56, RMSE = 24.84 mg/kg, MdAE = 12.98, RPIQ = 1.03), proved to be the most effective method for Zn prediction in agricultural soils. Except for the DWTLOGMSC dataset, the results from the M5 modeling technique combined with the Vis-NIR spectral dataset are satisfactory. The M5 tree model was the best method for predicting Zn in agricultural soil when combined with the MSC dataset (R^2 = 0.72, RMSE = 21.08 mg/kg, MdAE = 13.69, RPIQ = 1.63). However, combining the DWTSNVMSC dataset and the EGB modeling approach produced the best Zn prediction results in agricultural soil (R^2 = 0.64, RMSE = 22.82 mg/kg, MdAE = 12.46, RPIQ = 1.08). The combination of SVM and VIS-NIR spectral reflectance produced satisfactory results for four of the nine VIS-NIR spectral datasets used as the auxiliary dataset for Zn prediction in agricultural soil (MSC, CMR, SGLOGMSC, SGSNVMSC). The fusion of the MSC dataset and the SVM yielded the best results (R^2 = 0.52, RMSE = 24.98 mg/kg, MdAE = 13.77, RPIQ = 0.77). The overall assessment of the modeling approaches in context 1 indicates that the combination of the M5 tree model and the MSC dataset (R^2 = 0.72, RMSE = 21.08, MdAE = 13.69, RPIQ = 1.63) was the best approach for predicting Zn concentration in agricultural soil in context 1 with higher accuracy and minimal errors. Similarly in context 2 the overall evaluation of the modeling approaches suggested that CIF-DWTLOGMSC + SCP was the best method that was able to predict Zn concerntration in agricultural soil with minimal errors and high accuracy.

The overall evaluation of the best models in each of the five modeling approaches in both context 1 and 2 revealed that the combination of Vis-NIR spectral reflectance, soil chemical properties, and machine learning techniques produced the best prediction. Based on this, CIF-DWTLOGMSC + SCP was clearly the best overall technique for predicting Zn content in agricultural soil, with

significantly lower errors than the best models in the other modeling techniques and contexts.The use of Vis-NIR spectra reflectance in conjunction with the influence of micro and macro nutrients (soil chemical properties) on Zn prediction in agricultural soil has yielded remarkable results. The interaction of Zn as a micronutrient with the other micro and macronutrients may have had a significant impact on the best results in context 2. It is worth noting that the antagonistic and stimulating effects of soil macro and micronutrient interactions may have accounted for the best results in contexts 2 than in context 1. Geomorphological terrain has a significant impact on the quantification of PTEs such as Zn in soil, and interactions between bedrock, climatic conditions, and geomorphologic processes may result in the formation of soil parent composites. (Agyeman et al., 2022a, Agyeman et al., 2022b, Agyeman et al., 2022c). Kebonye et al. (2021) used soil chemical properties (Ca, Ti, Zn, Sr, Zr, Ba, Pb, and Th) in conjunction with MLAs to predict As concentrations in soil. Similarly, John et al., (2021) used MLAs in tandem with soil chemical properties such as potassium, calcium, sodium, magnesium, phosphorus, and vanadium to predict Sulphur in soil. John et al. (2020) used soil properties (i.e., Ca, Mg) in connection with terrain properties and a remote sensing dataset to predict soil organic carbon in alluvial soil. From another study, Hong et al., 2019a, Hong et al., 2019b used soil chemical properties in addition to Vis-NIR spectral reflectance, and the authors reported that the combination of Vis-NIR spectral reflectance, soil chemical properties, and an appropriate MLA model may improve prediction performance. The use of a pretreatment combined algorithms in conjunction with a single modeling and ensemble models to predict PTEs and soil organic carbon in a variety of soils and conditions has been tested and proven reliable (Biney et al., 2022; Biney et al., 2022c). The authors applied the hybridized pretreatment method in three different agricultural fields under three different measurement conditions (wet, dry, and field). However, no pretreatment is the best pre-processing method for predicting Zn concentration in soil, according to Kooistra et al. (2001). Even though raw spectra reflectance has a relatively high performance in the prediction of Zn in agricultural soil, the use of a combined predicted method and the inclusion of SCP has improved the prediction and reduced errors. Other pre-processing techniques must also be used to investigate the impact of various data treatment scenarios on the results of subsequent processing (Khosravi et al., 2018).

Paper 5

The result presents the Cd concentration prediction in agricultural soil using remote sensing datasets from Landsat 8 (L8) and Sentinel 2 (S2) (i.e., with a spatial resolution of 10 m) coupled with ensemble models as well as PS-LD and LD (Context 1). Four ensembling modeling approaches were used to predict Cd in agricultural soil using L8 and S2 as auxiliary datasets. The PS-LD results revealed that in ensemble 1, Cd prediction yielded R^2, RMSE, MAE, and MdAE values of 0.76, 0.66, 0.35, and 0.13 for L8 and 0.75, 0.67, 0.37, and 0.16 for S2. The L8 prediction of Cd in agricultural soil in ensemble 2 yielded R^2, RMSE, MAE, and MdAE values of 0.75, 0.65, 0.41, and 0.22, respectively, whereas in S8, Cd concentration was predicted with R^2, RMSE, MAE, and MdAE values of 0.58, 0.90, 0.48, and 0.19, respectively. L8 produced 0.64 (R^2), 0.82 (RMSE), 0.52 (MAE), and 0.22 (MdAE) in the prediction of Cd concentration in agricultural soil using ensemble 3, whereas S2 produced 0.71 (R^2), 0.69 (RSME), 0.42 (MAE), and 0.21 (MdAE) (MdAE). The ensemble 4 results indicated that using L8 Cd prediction yielded 0.74, 0.66, 0.38, and 0.17 for R^2, RMSE, MAE, and MdAE, respectively, whereas S2 yielded 0.69, 0.71, 0.44, and 0.21 for R^2, RMSE, MAE, and MdAE, respectively. Except for ensemble 3 of L8, which produced satisfactorily predicted Cd in agricultural soil with R^2, RMSE, MAE, and MdAE values of 0.58, 0.48, 0.37, and 0.14, the prediction results for LD of Cd using the four ensemble models produced abysmal results for both the S2 and L8 in 10m spatial resolution for both remote sensing datasets. However, ensembles 3 L8 and ensembles 1 L8, which produced the best prediction results in the prediction of Cd in agricultural soil, were the optimal modeling approaches based on the application of LD and PS-LD. Similarly, in context 2 (using 20m spatial resolution), the optimal prediction outputs from the LD and PS-LD coupled with ensemble models revealed that ensemble 1 of S2 was the overall best prediction approach for predicting Cd concentration in agricultural soil.

The cumulative best approach in the prediction of Cd in agricultural soil either using S2 and L8 from both 10m and 20m spatial resolution along with the ensemble models indicates unequivocally that the application of ensemble 1 of S2 of PS-LD with spatial resolution of 20m was the appropriate and best method for the prediction of Cd in agricultural soil with minimum errors and a higher R^2 value. This implies that using remote sensing datasets with higher spatial

resolution does not necessarily mean that prediction results will be improved; rather, it is dependent on the modeling techniques used as well as the spatial distribution of the dataset. Chen et al. (2004), who improved the accuracy of spectral unmixing by resampling the Ikonos image resolution from 4 to 30 m, observed the precision increase obtained by coarsening the image resolution. Obtaining better results from modeling an area is not solely dependent on the auxiliary dataset, but the ability to select the appropriate modeling approach in conjunction with the dataset may have a higher proclivity to produce good results. According to Zhou et al. (2021), predictions from modeling approaches created with coarse spatial resolution sensors can be comparable, if not superior, to models created with higher resolution sensors. The use of remote sensing images in the prediction of soil properties in a rural agricultural environment revealed that the low spatial resolution soil prediction approach demonstrated productive accuracy when compared to the higher spatial resolution approach (Xu et al., 2017). Kim et al. (2012) used a multi-scale modeling approach, soil series by remote sensing dataset application in a wetland ecosystem and discovered that datasets extracted from remote sensing images with lower or coarse spatial resolution performed better than datasets extracted from images with higher spatial resolution. Xia and Zhang (2022) conducted a comparative analysis of remote sensing images for the prediction of soil pH in the soil, and the authors discovered that using higher resolution remote sensing images in the prediction of soil properties in the soil does not necessarily increase prediction efficiency when compared to using medium resolution images.

Even though the current study is unique in that PS-LD and LD are evaluated using ensemble modeling, numerous studies have applied Sentinel 2 and Landsat 8 datasets and their combinations to a wide range of fields. The massive and prevalent data streams generated by satellite sensors, on the other hand, can ensure that soil surveillance and mapping procedures for large areas are created precisely, quickly, and successfully (Malenovsk et al., 2012). Furthermore, some satellite images are hampered by image quality factors. Because of its broad spatial coverage, quick revisit time, and ability to acquire data without regard for local air traffic limitations, satellite data can be valuable. Unfortunately, due to haziness or the need for parched and bald soil environmental conditions, these predefined reconsideration times may not be sufficient for adequate temporal coverage (Crucil et al., 2019). Other complexities for satellite

applications can include low image resolution and limited access to high-quality temporal and spatial images because of adverse atmospheric conditions and sensor requirements (Xiang et al., 2011). S2 has improved spatial and spectral capabilities for discriminating rangeland management practices (Sibanda et al., 2016), estimating forest canopy cover and leaf area index (LAI) (Korhonen et al., 2017), and increasing the categorization quality of built-up areas (Korhonen et al., 2017).

Resampling remote sensing datasets from coarse to fine spatial resolutions, or vice versa, does not always result in good prediction efficiency. Most of the time, these images loose quality during the resampling process, which can have an impact on the pixels that are extracted and used to predict PTEs or soil properties in the soil. The primary difference between down- and up-scaling synthetic and original images is that finer or coarser spatial details must be restored in the original down/up-scaling (Khosravi et al., 2022), and thus the inability to maintain spatial detail has an impact on image quality. In S2, some images have 20 and 10m spatial resolution, and not all bands were supposed to be resampled to either higher or lower spatial resolution, as in L8. For example, the use of some resampled bands in S2, such as Bands 2, 3, 4, and 8, from 10 m to 20m spatial resolution, in conjunction with unsampled bands 11 and 12, improved Cd prediction using PS-LD. The S2 and L8 prediction results were even closer when using bands that could be similar for both sensors, but the error outputs in the result for the spatial resolution of 20m of S2 were lower than for L8 in PS-LD. This implies that using original bands without resampling produces better results with less error. The combination of original bands and resampled bands produces better predictive modeling results than resampling all bands into different spatial resolutions. This means that the unsampled band retains the captured image details and qualities without distortion. The original bands contain useful data for predictive mapping. As a result, using original captured satellite images in prediction modeling is critical. Even though resampling can be useful for obtaining higher or coarser spatial resolution of bands for a specific goal, a combination of original bands in their original states and resampled bands has a better chance of producing good results.

Paper 6

The regression kriging (RK) approaches RF_RK, Cubist_ RK, EGB_ RK, and CIF_ RK produced R^2 values of 0.67, 0.49, 0.81, and 0.42 in scenario 1. The best results were obtained with the EGB_RK (R^2 = 0.81) approach, followed by the RF_RK approach (R^2 = 0.67). The regression kriging methods cubist_ RK (R^2 = 0.49) and CIF_RK (R^2 = 0.42) produced abysmal results with R^2 values below 0.5. EGB_RK had the lowest degree of error in the prediction of Sb in agricultural soil in terms of estimated error (RMSE and MAE). The modeling approach RF_RK had the least bias in the prediction of Sb in agricultural soil, with a bias of 0.31, followed by CIF_RK with a bias of 0.33, EGB_RK with a bias of 0.37, and cubist_RK with a bias of 0.40. According to the overall performance of the regression kriging modeling approaches, EGB_RK was the best modeling technique for predicting Sb in agricultural soil, with high prediction performance, low error margins, and detectable bias. In scenario 2, however, the cumulative prediction accuracy of the modeling techniques in predicting Sb concentration in agricultural soil revealed that the EGB_RK (R^2= 0.76, RMSE = 1.07, BIAS = 0.11, and MAE = 0.48) modeling approach was the best modeling method capable of predicting Sb concentration in agricultural soil with better efficiency, a lower error margin, and a satisfactory degree of bias. The cumulative assessment of the scenarios revealed that the three modeling approaches, EGB_RK, CIF_RK, and cubist_RK, significantly improved in scenario 2 compared to scenario 1. The overall modeling efficiency of the modeling techniques in predicting Sb in agricultural soil, on the other hand, indicated that the EGB_RK in the scenario 2 modeling approach is the best modeling method capable of predicting the concentration of Sb in agricultural soil with higher efficiency, minimal error margin, and a satisfactory degree of bias. According to Hengl et al. (2004), Umali et al. (2012), and Zhang et al. (2012), the use of RK in learning algorithms that incorporate spatial interpolation produces better spatial interpolation results in the prediction of soil properties and PTEs. When combined with an appropriate modeling algorithm, the spatial interpolation aspect of ordinary kriging produces good results. OK has the potential to produce good results when used in the prediction of PTEs and soil properties, according to Pham et al., 2019 and Pham et al., 2019. RK has consistently proven to be more precise when there is a strong correlation between predicted PTE and environmental covariates (Keskin and Grunwald, 2018). It is critical to emphasize that the

selection of ecologically consistent environmental covariates that correlate with the response variable with a robust autocorrelation with data makes RK more appropriate. More auxiliary datasets could be selected to improve the accuracy of the RK method (Pham et al., 2019). According to Kim et al. (2015), EGB tends to screen out the efficiency of modeling techniques by reducing the potential limitations of other modeling strategies, such as computational complexity. Furthermore, EGB can aid in modeling standardization (Jia et al., 2019), hyper-parameter tuning (Probst et al., 2019), local minima (Kawaguchi, 2019), elevated discrepancies (Li et al., 2020), and technology transfer (Kim et al., 2020).

4.2. Concluding remarks

The combination of various spectral pretreatment algorithms together with machine learning algorithms and appropriate auxiliary datasets improves Zn prediction outcomes.

The combination of data fusion, terrain attribute, and regression kriging modeling approaches produces optimal results with a high R^2 value, minimal errors, and bias.

For better prediction outcomes, proxies or additional data sets can be combined with soil characteristics that have a strong correlation with response variables.

Furthermore, combining terrain attributes with data fusion has the potential to reduce error, bias, and predict with high accuracy.

Combining preferential sampling with legacy datasets, as well as an appropriate modeling approach and a well-correlated remote sensing dataset, yields good results.

Using higher spatial resolution remote sensing datasets along with input data in the prediction of PTEs or soil properties in the soil does not necessarily mean good results will be produced.

The best results will be achieved through a combination of environmental covariates with a high correlation with the response variable, combined with appropriate modeling techniques predicting potentially toxic elements in agricultural soil.

The use of mean, maximum, and minimum values in health risk estimation does not provide a comprehensive picture of a research area's health state.

The continuous application of agriculturally related inputs such as phosphate fertilizers and other anthropogenic activities (e.g., the steel industry) can increase the level of PTEs in soils.

Using a pollution assessment-based receptor model (ER-PMF) has been shown to be reliable and practical in estimating distribution sources.

Geographical weighted regression cokriging proved to be more reliable and efficient in the mapping of PTEs in the agricultural soil than the application of geographical weighted regression.

Each study requires a different modeling approach that is appropriate for the type of dataset used because there is no one modeling approach that fits all datasets.

5.0 REFERENCES

Abdel Rahman, A. M., Pawling, J., Ryczko, M., Caudy, A. A., & Dennis, J. W. (2014). Targeted metabolomics in cultured cells and tissues by mass spectrometry: Method development and validation. *Analytica Chimica Acta*, *845*, 53–61. https://doi.org/10.1016/j.aca.2014.06.012

Ackermann, F. (1980). A procedure for correcting the grain size effect in heavy metal analyses of estuarine and coastal sediments. *Environmental Technology Letters*, *1*(11), 518–527. https://doi.org/10.1080/09593338009384008

Agyeman, P. C., Khosravi, V., Michael Kebonye, N., John, K., Borůvka, L., & Vašát, R. (2022). Using spectral indices and terrain attribute datasets and their combination in the prediction of cadmium content in agricultural soil. *Computers and Electronics in Agriculture*, *198*, 107077. https://doi.org/10.1016/J.COMPAG.2022.107077

Agyeman, P. C., Ahado, S. K., John, K., Kebonye, N. M., Vašát, R., Borůvka, L., Koracek M., Němeček, K. (2021). Health risk assessment and the application of CF-PMF: A pollution assessment–based receptor model in an urban soil. *Journal of Soils and Sediments*, *21*(9), 3117-3136.

Agyeman, P.C., Ahado, S.K., Kingsley, J., Kebonye, N.M., Biney, J.K.M., Borůvka, L., Vasat, R. and Kocarek, M., 2021. Source apportionment, contamination levels, and spatial prediction of potentially toxic elements in selected soils of the Czech Republic. Environmental geochemistry and health, 43(1), pp.601-620.

Agyeman, P.C., Kebonye, N.M., John, K., Borůvka, L., Vašát, R. and Fajemisim, O., 2022. Prediction of nickel concentration in peri-urban and urban soils using hybridized empirical bayesian kriging and support vector machine regression. *Scientific Reports*, *12*(1), pp.1-16.

Alloway, B. J. (2013). *Sources of Heavy Metals and Metalloids in Soils* (pp. 11–50). https://doi.org/10.1007/978-94-007-4470-7_2

Antoniadis, V., Shaheen, S. M., Boersch, J., Frohne, T., Du Laing, G., & Rinklebe, J. (2017). Bioavailability and risk assessment of potentially toxic elements in garden edible vegetables and soils around a highly contaminated former mining area in Germany. *Journal of environmental management*, *186*, 192-200.

Antonelli, J., Cefalu, M., Bornn, L., 2016. The positive effects of population-based preferential sampling in environmental epidemiology. Biostatistics 17, 764778. https://doi.org/10.1093/BIOSTATISTICS/KXW026

Araújo, D. F., Boaventura, G. R., Machado, W., Viers, J., Weiss, D., Patchineelam, S. R., Ruiz, I., Rodrigues, A. P. C., Babinski, M., & Dantas, E. (2017). Tracing of anthropogenic zinc sources in

coastal environments using stable isotope composition. *Chemical Geology*, *449*, 226–235. https://doi.org/10.1016/j.chemgeo.2016.12.004

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J. and dL Mendonca-Santos, M., 2014. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. Advances in agronomy, 125, pp.93-134.

Ash, C., Borůvka, L., Tejnecký, V., Nikodem, A., Šebek, O., & Drábek, O. (2014). Potentially toxic element distribution in soils from the Ag-smelting slag of Kutná Hora (Czech Republic): Descriptive and prediction analyses. *Journal of Geochemical Exploration*, *144*(PB), 328–336. https://doi.org/10.1016/j.gexplo.2013.11.010

Asmaryan, Sh G., V. S. Muradyan, L. V. Sahakyan, A. K. Saghatelyan, and Timothy Warner. "Development of remote sensing methods for assessing and mapping soil pollution with heavy metals." Arrouays D, Mc Kenzie N, Hempel J, de Forges AR, Mc Bratney A. GlobalSoilMap: basis of the global spatial soil information system. Leiden: CRC Press/Balkema (2014): 429-432.

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L. and Panagos, P., 2019. Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. Geoderma, 355, p.113912.

Ballabio, C., Panagos, P., Lugato, E., Huang, J.H., Orgiazzi, A., Jones, A., Fernández-Ugalde, O., Borrelli, P. and Montanarella, L., 2018. Copper distribution in European topsoils: An assessment based on LUCAS soil survey. Science of The Total Environment, 636, pp.282-298.

Bangroo, S. A., Najar, G. R., Achin, E., & Truong, P. N. (2020). Application of predictor variables in spatial quantification of soil organic carbon and total nitrogen using regression kriging in the North Kashmir Forest Himalayas. *Catena*, *193*, 104632. https://doi.org/10.1016/j.catena.2020.104632

Basta, N. T., Ryan, J. A., & Chaney, R. L. (2005). Trace Element Chemistry in Residual-Treated Soil: Key Concepts and Metal Bioavailability. *Journal of Environmental Quality*, *34*(1), 49–63. https://doi.org/10.2134/jeq2005.0049dup

Beard, K. and Mackaness, W., 1993. Visual access to data quality in geographic information systems. Cartographica: The International Journal for Geographic Information and Geovisualization, 30(2-3), pp.37-45.

Behrens, T., Schmidt, K., MacMillan, R. A., & Viscarra Rossel, R. A. (2018). Multi-scale digital soil mapping with deep learning. *Scientific Reports*, *8*(1), 1–9. https://doi.org/10.1038/s41598-018-33516-6

Bernard, A. (2008). Cadmium & its adverse effects on human health. *Indian Journal of Medical Research*, *128*(4), 557–564.

Bhandari, G., Atreya, K., Scheepers, P. T., & Geissen, V. (2020). Concentration and distribution of pesticide residues in soil: Non-dietary human health risk assessment. *Chemosphere*, *253*, 126594.

Bhuiyan, M. A. H., Karmaker, S. C., Bodrud-Doza, M., Rakib, M. A., & Saha, B. B. (2021). Enrichment, sources and ecological risk mapping of heavy metals in agricultural soils of dhaka district employing SOM, PMF and GIS methods. *Chemosphere*, *263*, 128339.

Bolan, N., Kunhikrishnan, A., Thangarajan, R., Kumpiene, J., Park, J., Makino, T., Kirkham, M.B. and Scheckel, K., 2014. Remediation of heavy metal (loid) s contaminated soils–to mobilize or to immobilize?. *Journal of hazardous materials*, *266*, pp.141-166.

Biney, J. K. M., Vašát, R., Bell, S. M., Kebonye, N. M., Klement, A., John, K., & Borůvka, L. (2022). Prediction of topsoil organic carbon content with Sentinel-2 imagery and spectroscopic measurements under different conditions using an ensemble model approach with multiple pre-treatment combinations. *Soil and Tillage Research*, *220*, 105379.

Biney, J.K.M., Vašát, R., Blöcher, J.R., Borůvka, L. and Němeček, K., 2022. Using an ensemble model coupled with portable X-ray fluorescence and visible near-infrared spectroscopy to explore the viability of mapping and estimating arsenic in an agricultural soil. *Science of The Total Environment*, *818*, p.151805.

Borůvka, L., Vacek, O., & Jehlička, J. (2005). Principal component analysis as a tool to indicate the origin of potentially toxic elements in soils. *Geoderma*, *128*(3-4 SPEC. ISS.), 289–300. https://doi.org/10.1016/j.geoderma.2005.04.010

Bradl, H.B., 2005. Sources and origins of heavy metals. In Interface science and technology (Vol. 6, pp. 1-27). Elsevier.

Bray, J. G. P., Rossel, R. V., & McBratney, A. B. (2009). Diagnostic screening of urban soil contaminants using diffuse reflectance spectroscopy. *Australian Journal of Soil Research*, *47*(4), 433–442. https://doi.org/10.1071/SR08068

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*(1), 49–64. https://doi.org/10.1007/BF00117832

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brevik, E. C., Calzolari, C., Miller, B. A., Pereira, P., Kabala, C., Baumgarten, A., & Jordán, A. (2016). Soil mapping, classification, and pedologic modeling: History and future directions. *Geoderma*, *264*, 256–274. https://doi.org/10.1016/j.geoderma.2015.05.017

Brown, S. G., Eberly, S., Paatero, P., & Norris, G. A. (2015). Methods for estimating uncertainty in PMF solutions: examples with ambient air and water quality data and guidance on reporting PMF results. Sci Total Environ, 518, 626-635.

Bundschuh, J., Litter, M. I., Parvez, F., Román-Ross, G., Nicolli, H. B., Jean, J. S., Liu, C. W., López, D., Armienta, M. A., Guilherme, L. R. G., Cuevas, A. G., Cornejo, L., Cumbal, L., & Toujaguez, R. (2012). One century of arsenic exposure in Latin America: A review of history and occurrence from 14 countries. *Science of the Total Environment*, *429*, 2–35. https://doi.org/10.1016/j.scitotenv.2011.06.024

Cachada, A., Rocha-Santos, T. and Duarte, A.C., 2018. Soil and pollution: an introduction to the main issues. In Soil pollution (pp. 1-28). Academic Press. https://doi.org/10.1016/B978-0-12-849873-6.00001-7

Chen, H., Teng, Y., Lu, S., Wang, Y., & Wang, J. (2015). Contamination features and health risk of soil heavy metals in China. *Science of the Total Environment*, *512–513*, 143–153. https://doi.org/10.1016/j.scitotenv.2015.01.025

Chen, L., Ren, C., Li, L., Wang, Y., Zhang, B., Wang, Z., & Li, L. (2019). A Comparative Assessment of Geostatistical, Machine Learning, and Hybrid Approaches for Mapping Topsoil Organic Carbon Content. *ISPRS International Journal of Geo-Information 2019, Vol. 8, Page 174*, *8*(4), 174. https://doi.org/10.3390/IJGI8040174

Chen, T., Liu, X., Li, X., Zhao, K., Zhang, J., Xu, J., Shi, J., & Dahlgren, R. A. (2009). Heavy metal sources identification and sampling uncertainty analysis in a field-scale vegetable soil of Hangzhou, China. *Environmental Pollution*, *157*(3), 1003–1010. https://doi.org/10.1016/j.envpol.2008.10.011

Cherkassky, V. and Mulier, F.M., 2007. *Learning from data: concepts, theory, and methods*. John Wiley & Sons. https://doi.org/10.1002/9780470140529

Choe, E., van der Meer, F., van Ruitenbeek, F., van der Werff, H., de Smeth, B., & Kim, K. W. (2008a). Mapping of heavy metal pollution in stream sediments using combined geochemistry, field spectroscopy, and hyperspectral remote sensing: A case study of the Rodalquilar mining area, SE Spain. *Remote Sensing of Environment*, *112*(7), 3222–3233. https://doi.org/10.1016/j.rse.2008.03.017

Christl, I., & Kretzschmar, R. (1999). Competitive sorption of copper and lead at the oxide-water interface: Implications for surface site density. *Geochimica et Cosmochimica Acta*, *63*(19–20), 2929–2938. https://doi.org/10.1016/S0016-7037(99)00266-5

Crucil, G., Castaldi, F., Aldana-Jague, E., van Wesemael, B., Macdonald, A., & Van Oost, K. (2019). Assessing the performance of UAS-compatible multispectral and hyperspectral sensors for soil organic carbon prediction. *Sustainability*, *11*(7), 1889.

Climent, F., Momparler, A. and Carmona, P., 2019. Anticipating bank distress in the Eurozone: An extreme gradient boosting approach. Journal of Business Research, 101, pp.885-896.https://www.sciencedirect.com/science/article/pii/S0148296318305678

Cocks, T., Jenssen, R., Stewart, A., Wilson, I. and Shields, T., 1998, October. The HyMapTM airborne hyperspectral sensor: The system, calibration and performance. In Proceedings of the 1st EARSeL workshop on Imaging Spectroscopy (pp. 37-42).

Congdon, P., 2007. *Bayesian statistical modelling*. John Wiley & Sons.

Costa, E. M., Samuel-Rosa, A., & dos Anjos, L. H. C. (2018). Digital elevation model quality on digital soil mapping prediction accuracy. *Ciencia e Agrotecnologia*, *42*(6), 608–622. https://doi.org/10.1590/1413-70542018426027418

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Davies, B.E. and Jones, L.H.P., 1988. Micronutrients and toxic elements. *Russell's soil conditions and plant growth. Eleventh edition*, pp.780-814.

Delerce, S., Dorado, H., Grillon, A., Rebolledo, M.C., Prager, S.D., Patiño, V.H., Garcés Varón, G. and Jiménez, D., 2016. Assessing weather-yield relationships in rice at local scale using data mining approaches. PloS one, 11(8), p.e0161620.https://doi.org/10.1371/JOURNAL.PONE.0161620

Deng, H., 2013. Guided random forest in the RRF package. *arXiv preprint arXiv:1306.0237*.

Deng, H. and Runger, G., 2012, June. Feature selection via regularized trees. In The 2012 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE. https://ieeexplore.ieee.org/abstract/document/6252640/

Deschamps, E., Matschullat, J. (2011). *Arsenic: Natural and Anthropogenic Arsenic in the Environment. Boca Raton, FL: CRC Press*.

Dharumarajan, S., Hegde, R., Janani, N. and Singh, S.K., 2019. The need for digital soil mapping in India. Geoderma Regional, 16, p.e00204.

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*. https://doi.org/10.1186/1471-2105-7-3

Doetterl, S., Stevens, A., van Oost, K., Quine, T. A., & van Wesemael, B. (2013). Spatially explicit regional-scale prediction of soil organic carbon stocks in cropland using environmental variables and mixed model approaches. *Geoderma*, *204–205*, 31–42. https://doi.org/10.1016/j.geoderma.2013.04.007

Doležalová Weissmannová, H., Mihočová, S., Chovanec, P., & Pavlovský, J. (2019). Potential ecological risk and human health risk assessment of heavy metal pollution in industrial affected soils by coal mining and metallurgy in Ostrava, Czech Republic. *International Journal of Environmental Research and Public Health*, *16*(22), 4495.

Dor, E. B., Ong, C., & Lau, I. C. (2015). Reflectance measurements of soils in the laboratory: Standards and protocols. *Geoderma*, *245*, 112-124.

Drury, S. A. (1987). Image Interpretation in Geology (London: Allen &Unwin).

Dinsdale, D., Salibian-Barrera, M., 2019. Modelling ocean temperatures from bio-probes under preferential sampling. https://doi.org/10.1214/18-AOAS1217 13, 713–745. https://doi.org/10.1214/18-AOAS1217

Ehsani, M.R., Upadhyaya, S.K., Slaughter, D., Shafii, S. and Pelletier, M., 1999. A NIR technique for rapid determination of soil mineral nitrogen. Precision agriculture, 1(2), pp.219-236.

EPA positive matrix factorization (PMF) 5.0 fundamentals and user guide https://www.epa.gov/sites/production/files/2015-02/documents/pmf_5.0_user_guide.pdf (2014)

Ergin, M., Saydam, C., Baştürk, Ö., Erdem, E., & Yörük, R. (1991). Heavy metal concentrations in surface sediments from the two coastal inlets (Golden Horn Estuary and İzmit Bay) of the northeastern Sea of Marmara. *Chemical Geology*, *91*(3), 269–285. https://doi.org/10.1016/0009-2541(91)90004-B

Eriksson, J., Andersson, A. and Andersson, R., 1997. The state of Swedish farmlands (Report 4778). Stockholm: Swedish Environmental Protection Agency.

Etemad-Shahidi, A., & Mahjoobi, J. (2009). Comparison between M5' model tree and neural networks for prediction of significant wave height in Lake Superior. *Ocean Engineering*, *36*(15–16), 1175–1181. https://doi.org/10.1016/J.OCEANENG.2009.08.008

Eziz, M., Mohammad, A., Mamut, A., & Hini, G. (2018). A human health risk assessment of heavy metals in agricultural soils of Yanqi Basin, Silk Road Economic Belt, China. *Human and Ecological Risk Assessment*, *24*(5), 1352–1366. https://doi.org/10.1080/10807039.2017.1412818

Falandysz, J., & Borovička, J. (2013). Macro and trace mineral constituents and radionuclides in mushrooms: Health benefits and risks. *Applied Microbiology and Biotechnology*, *97*(2), 477–501. https://doi.org/10.1007/S00253-012-4552-8

Falandysz, J., & Rizal, L. M. (2016). Arsenic and its compounds in mushrooms: A review. *Journal of Environmental Science and Health - Part C Environmental Carcinogenesis and Ecotoxicology Reviews*, *34*(4), 217–232. https://doi.org/10.1080/10590501.2016.1235935

FAO & ITPS. (2015). Intergovernmental Technical Panel on Soils. Status of the World's Soil Resources. *Intergovernmental Technical Panel on Soils*, 100–146.

Fei, X., Christakos, G., Xiao, R., Ren, Z., Liu, Y., & Lv, X. (2019). Improved heavy metal mapping and pollution source apportionment in Shanghai City soils using auxiliary information. *Science of the Total Environment*, *661*, 168-177.

Galušková, I., Boruvka, L., & Drábek, O. (2011). Urban soil contamination by potentially risk elements. *Soil and Water Research*, *6*(2), 55–60. https://doi.org/10.17221/55/2010-swr

Gamon, J.A., Penuelas, J. and Field, C.B., 1992. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. Remote Sensing of environment, 41(1), pp.35-44.https://www.sciencedirect.com/science/article/pii/003442579290059S

García, M., Saatchi, S., Ustin, S., & Balzter, H. (2018). Modelling forest canopy height by integrating airborne LiDAR samples with satellite Radar and multispectral imagery. *International Journal of Applied Earth Observation and Geoinformation*, *66*, 159–173. https://doi.org/10.1016/j.jag.2017.11.017

Gautam, R., Panigrahi, S., Franzen, D., & Sims, A. (2011). Residual soil nitrate prediction from imagery and non-imagery information using neural network technique. *Biosystems Engineering*, *110*(1), 20–28. https://doi.org/10.1016/j.biosystemseng.2011.06.002

Gholampour, A. and Johari, A., 2019. Reliability-based analysis of braced excavation in unsaturated soils considering conditional spatial variability. Computers and Geotechnics, 115, p.103163.https://www.sciencedirect.com/science/article/pii/S0266352X19302277

Gholizadeh, A., Borùvka, L., Vašát, R., Saberioon, M., Klement, A., Kratina, J., Tejnecký, V., & Drábek, O. (2015). Estimation of potentially toxic elements contamination in anthropogenic soils on a brown coal mining dumpsite by reflectance spectroscopy: A case study. *PLoS ONE*, *10*(2), e0117457. https://doi.org/10.1371/journal.pone.0117457

Gholizadeh, A., Saberioon, M., Ben-Dor, E., & Borůvka, L. (2018). Monitoring of selected soil contaminants using proximal and remote sensing techniques: Background, state-of-the-art and future perspectives. *Critical Reviews in Environmental Science and Technology*, *48*(3), 243–278. https://doi.org/10.1080/10643389.2018.1447717

Gia Pham, T., Kappas, M., Van Huynh, C. and Hoang Khanh Nguyen, L., 2019. Application of ordinary kriging and regression kriging method for soil properties mapping in hilly region of Central Vietnam. ISPRS International Journal of Geo-Information, 8(3), p.147.

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random forests for land cover classification. *Pattern Recognition Letters*, *27*(4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011

Gomes, L. C., Faria, R. M., de Souza, E., Veloso, G. V., Schaefer, C. E. G. R., & Filho, E. I. F. (2019). Modelling and mapping soil organic carbon stocks in Brazil. *Geoderma*, *340*, 337–350. https://doi.org/10.1016/j.geoderma.2019.01.007

González-Macías, C., Schifter, I., Lluch-Cota, D. B., Méndez-Rodríguez, L., & Hernández-Vázquez, S. (2006). Distribution, enrichment and accumulation of heavy metals in coastal sediments of Salina Cruz Bay, México. *Environmental Monitoring and Assessment*, *118*(1–3), 211–230. https://doi.org/10.1007/s10661-006-1492-8

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma, 103(1-2), pp.3-26. https://www.sciencedirect.com/science/article/pii/S0016706101000672

Goydaragh, M.G., Taghizadeh-Mehrjardi, R., Jafarzadeh, A.A., Triantafilis, J. and Lado, M., 2021. Using environmental variables and Fourier Transform Infrared Spectroscopy to predict soil organic carbon. *Catena*, *202*, p.105280.

Greaney, K.M., 2005. An assessment of heavy metal contamination in the marine sediments of Las Perlas Archipelago, Gulf of Panama. School of Life Sciences Heriot-Watt University, Edinburgh.

Guan, Q., Zhao, R., Pan, N., Wang, F., Yang, Y. and Luo, H., 2019. Source apportionment of heavy metals in farmland soil of Wuwei, China: Comparison of three receptor models. *Journal of Cleaner Production*, *237*, p.117792.

Han, Q., Wang, M., Cao, J., Gui, C., Liu, Y., He, X., He Y Liu, Y. (2020). Health risk assessment and bioaccessibilities of heavy metals for children in soil and dust from urban parks and schools of Jiaozuo, China. *Ecotoxicology and environmental safety*, *191*, 110157.

Hakanson, L. (1980). An ecological risk index for aquatic pollution control.a sedimentological approach. *Water Research*, *14*(8), 975–1001. https://doi.org/10.1016/0043-1354(80)90143-8

Heddam, S., 2021. New formulation for predicting soil moisture content using only soil temperature as predictor: multivariate adaptive regression splines versus random forest, multilayer perceptron neural network, M5Tree, and multiple linear regression. In *Water Engineering Modeling and Mathematic Tools* (pp. 45-62). Elsevier.

Hengl, T., Heuvelink, G. B., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, *120*(1-2), 75-93.

Heung, B., Bulmer, C. E., & Schmidt, M. G. (2014). Predictive soil parent material mapping at a regional scale: A Random Forest approach. *Geoderma*, *214–215*, 141–154. https://doi.org/10.1016/j.geoderma.2013.09.016

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma*, *265*, 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014

Hope, S. and Hunter, G.J., 2007. Testing the effects of thematic uncertainty on spatial decision-making. Cartography and Geographic Information Science, 34(3), pp.199-214. https://doi.org/10.1559/152304007781697884

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics, 15(3), 651-674. https://doi.org/10.1198/106186006X133933

Hong-Yan, R.E.N., Zhuang, D.F., Singh, A.N., Jian-Jun, P.A.N., Dong-Sheng, Q.I.U. and Run-He, S.H.I., 2009. Estimation of As and Cu contamination in agricultural soils around a mining area by reflectance spectroscopy: A case study. Pedosphere, 19(6), pp.719-726.

Hong, Y., Shen, R., Cheng, H., Chen, S., Chen, Y., Guo, L., He, J., Liu, Y., Yu, L. and Liu, Y., 2019. Cadmium concentration estimation in peri-urban agricultural soils: Using reflectance spectroscopy, soil auxiliary information, or a combination of both?. *Geoderma*, *354*, p.113875.

Hong, Y., Liu, Y., Chen, Y., Liu, Y., Yu, L., Liu, Y. and Cheng, H., 2019. Application of fractional-order derivative in the quantitative estimation of soil organic matter content through visible and near-infrared spectroscopy. *Geoderma*, *337*, pp.758-769.

Hou, D., O'Connor, D., Nathanail, P., Tian, L., & Ma, Y. (2017). Integrated GIS and multivariate statistical analysis for regional scale assessment of heavy metal soil contamination: A critical review. *Environmental Pollution*, *231*, 1188–1200. https://doi.org/10.1016/J.ENVPOL.2017.07.021

Huang, J., Guo, S., Zeng, G.M., Li, F., Gu, Y., Shi, Y., Shi, L., Liu, W. and Peng, S., 2018. A new exploration of health risk assessment quantification from sources of soil heavy metals under different land use. *Environmental pollution*, *243*, pp.49-58.

Hu, X., Zhang, Y., Ding, Z., Wang, T., Lian, H., Sun, Y., & Wu, J. (2012). Bioaccessibility and health risk of arsenic and heavy metals (Cd, Co, Cr, Cu, Ni, Pb, Zn and Mn) in TSP and PM2.5 in Nanjing, China. *Atmospheric Environment*, *57*, 146–152. https://doi.org/10.1016/j.atmosenv.2012.04.056

Iñigo, V., Andrades, M., Alonso-Martirena, J. I., Marín, A., & Jiménez-Ballesta, R. (2011). Multivariate statistical and GIS-based approach for the identification of Mn and Ni concentrations and spatial variability in soils of a humid mediterranean environment: La Rioja, Spain. *Water, Air, and Soil Pollution*, *222*(1–4), 271–284. https://doi.org/10.1007/s11270-011-0822-9

Iqbal, J., Thomasson, J. A., Jenkins, J. N., Owens, P. R., & Whisler, F. D. (2005). Spatial Variability Analysis of Soil Physical Properties of Alluvial Soils. *Soil Science Society of America Journal*, *69*(4), 1338–1350. https://doi.org/10.2136/sssaj2004.0154

Jeřábková, J., Tejnecký, V., Borůvka, L., & Drábek, O. (2018). Chromium in Anthropogenically Polluted and Naturally Enriched Soils: A Review. *Scientia Agriculturae Bohemica*, *49*(4), 297–312. https://doi.org/10.2478/sab-2018-0037

Friedman, J.H., 1991. Multivariate adaptive regression splines. The annals of statistics, 19(1), pp.1-67.

Jia, X., Hu, B., Marchant, B.P., Zhou, L., Shi, Z. and Zhu, Y., 2019. A methodological framework for identifying potential sources of soil heavy metal pollution based on machine learning: A case study in the Yangtze Delta, China. *Environmental Pollution*, *250*, pp.601-609.

Jiang, X., Zou, B., Feng, H., Tang, J., Tu, Y., & Zhao, X. (2019). Spatial distribution mapping of Hg contamination in subclass agricultural soils using GIS enhanced multiple linear regression. *Journal of Geochemical Exploration*, *196*, 1–7. https://doi.org/10.1016/j.gexplo.2018.10.002

Jiménez-Ballesta, R., García-Navarro, F. J., Bravo, S., Amorós, J. A., Pérez-de-los-Reyes, C., & Mejías, M. (2017). Environmental assessment of potential toxic trace element contents in the inundated floodplain area of Tablas de Daimiel wetland (Spain). *Environmental Geochemistry and Health*, *39*(5), 1159–1177. https://doi.org/10.1007/s10653-016-9884-3

Johari, A., Khani, M., Hadianfard, M.A. and JavidSharifi, B., 2020. System reliability analysis for seismic site classification based on sequential Gaussian co-simulation: A case study in Shiraz, Iran. Soil Dynamics and Earthquake Engineering, 137, p.106286.

John, K., Abraham Isong, I., Michael Kebonye, N., Okon Ayito, E., Chapman Agyeman, P. and Marcus Afu, S., 2020. Using machine learning algorithms to estimate soil organic carbon

variability with environmental variables and soil nutrient indicators in an alluvial soil. Land, 9(12), p.487.

John Lu, Z. Q. (2010). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(3), 693–694. https://doi.org/10.1111/j.1467-985x.2010.00646_6.x

John, K., Agyeman, P.C., Kebonye, N.M., Isong, I.A., Ayito, E.O., Ofem, K.I. and Qin, C.Z., 2021. Hybridization of cokriging and gaussian process regression modelling techniques in mapping soil sulphur. *Catena*, *206*, p.105534.

Gauch Jr, H.G. and Gauch, H.G., 2003. Scientific method in practice. Cambridge University Press.

Quinlan, J.R., 1992, November. Learning with continuous classes. In 5th Australian joint conference on artificial intelligence (Vol. 92, pp. 343-348).

Kaasalainen, M., & Yli-Halla, M. (2003). Use of sequential extraction to assess metal partitioning in soils. *Environmental Pollution*, *126*(2), 225–233. https://doi.org/10.1016/S0269-7491(03)00191-X

Kalambukattu, J.G., Kumar, S. and Arya Raj, R., 2018. Digital soil mapping in a Himalayan watershed using remote sensing and terrain parameters employing artificial neural network model. *Environmental earth sciences*, *77*(5), pp.1-14.

Kabata-Pendias, A. (2010a). Trace elements in soils and plants: Fourth edition. https://doi.org/10.1201/b10158

Kabata-Pendias, A. and Mukherjee, A.B., 2007. Humans (pp. 67-83). Springer Berlin Heidelberg.

Kabata-Pendias, A., & Szteke, B. (2015a). Trace Elements in Abiotic and Biotic Environments. In *Trace Elements in Abiotic and Biotic Environments*. https://doi.org/10.1201/b18198

Kawaguchi, K. and Bengio, Y., 2019. Depth with nonlinearity creates no bad local minima in ResNets. *Neural Networks*, *118*, pp.167-174.

Kebonye, N.M., Agyeman, P.C., Seletlo, Z. and Eze, P.N., 2022. On exploring bivariate and trivariate maps as visualization tools for spatial associations in digital soil mapping: A focus on soil properties. Precision Agriculture, pp.1-22.

Kebonye, N.M., Eze, P.N., John, K., Agyeman, P.C., Němeček, K. and Borůvka, L., 2022. An in-depth human health risk assessment of potentially toxic elements in highly polluted riverine soils, Příbram (Czech Republic). Environmental Geochemistry and Health, 44(2), pp.369-385.

Kebonye, N.M., John, K., Chakraborty, S., Agyeman, P.C., Ahado, S.K., Eze, P.N., Němeček, K., Drábek, O. and Borůvka, L., 2021. Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy. *Geoderma*, *384*, p.114792.

Kempen, B., Brus, D. J., Stoorvogel, J. J., Heuvelink, G. B. M., & de Vries, F. (2012). Efficiency Comparison of Conventional and Digital Soil Mapping for Updating Soil Maps. *Soil Science Society of America Journal*, *76*(6), 2097–2115. https://doi.org/10.2136/sssaj2011.0424

Keskin, H. and Grunwald, S., 2018. Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma, 326, pp.22-41.

Khan, S., Cao, Q., Zheng, Y. M., Huang, Y. Z., & Zhu, Y. G. (2008). Health risks of heavy metals in contaminated soils and food crops irrigated with wastewater in Beijing, China. *Environmental Pollution*, *152*(3), 686–692. https://doi.org/10.1016/j.envpol.2007.06.056

Khodadoust, A. P., Reddy, K. R., & Maturi, K. (2004). Removal of nickel and phenanthrene from kaolin soil using different extractants. *Environmental Engineering Science*, *21*(6), 691–704. https://doi.org/10.1089/EES.2004.21.691

Khosravi, V., Ardejani, F.D., Yousefi, S. and Aryafar, A., 2018. Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods. Geoderma, 318, pp.29-41.

Khosravi, V., Gholizadeh, A., & Saberioon, M. (2022). Soil toxic elements determination using integration of Sentinel-2 and Landsat-8 images: Effect of fusion techniques on model performance. *Environmental Pollution*, *310*, 119828.

Kim, J., Grunwald, S., Rivero, R.G. and Robbins, R., 2012. Multi-scale modeling of soil series using remote sensing in a wetland ecosystem. Soil Science Society of America Journal, 76(6), pp.2327-2341.

Kim, M. and Geum, Y., 2020. Predicting Patent Transactions Using Patent-Based Machine Learning Techniques. *IEEE Access*, *8*, pp.188833-188843.

Kienast-Brown, S., Libohova, Z., & Boettinger, J. (2017). Digital soil mapping. *Soil survey manual, USDA Handbook*, *18*, 295-354.

Kooistra, L., Wehrens, R., Leuven, R.S.E.W. and Buydens, L.M.C., 2001. Possibilities of visible–near-infrared spectroscopy for the assessment of soil contamination in river floodplains. Analytica chimica acta, 446(1-2), pp.97-105.

Korhonen, L., Packalen, P. and Rautiainen, M., 2017. Comparison of Sentinel-2 and Landsat 8 in the estimation of boreal forest canopy cover and leaf area index. Remote sensing of environment, 195, pp.259-274.

Kozák, J., Němeček, J., Borůvka, L., Lérová, Z., Němeček, K., Kodešová, R., Janků, J., Jacko, K., Hladík, J. and Zádorová, T., 2010. Atlas půd České republiky.[Soil Atlas of the Czech Republic.]. Prague, Czech University of Life Sciences Prague, 150.

Kuhn, M. and Johnson, K., 2013. Applied predictive modeling (Vol. 26, p. 13). New York: Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., Johnson, K., Kuhn, M., & Johnson, K. (2013). An Introduction to Feature Selection. In *Applied Predictive Modeling* (pp. 487–519). Springer New York. https://doi.org/10.1007/978-1-4614-6849-3_19

Kuhn, M., Weston, S., Keefer, C. and Coulter, N., 2014. C code for Cubist by Ross Quinlan. Cubist: rule-and instance-based regression modeling. R package version 0.0, 18.

Kumar, S., & Deswal, S. (2020). Phytoremediation capabilities of Salvinia molesta, water hyacinth, water lettuce, and duckweed to reduce phosphorus in rice mill wastewater. *International Journal of Phytoremediation*, *22*(11), 1097-1109.

Kumar, S., Lal, R., & Liu, D. (2012). A geographically weighted regression kriging approach for mapping soil organic carbon stock. *Geoderma*, *189*, 627-634.

Kuo, S.A., Heilman, P.E. and Baker, A.S., 1983. Distribution and forms of copper, zinc, cadmium, iron, and manganese in soils near a copper smelter1. *Soil science*, *135*(2), pp.101-109.

Smith, L.A. and Brauning, S.E., 1995. *Remedial options for metals-contaminated sites* (pp. 17-122). Boca Raton: CRC Press.

Laben, C.A. and Brower, B.V., Eastman Kodak Co, 2000. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. U.S. Patent 6,011,875.

Lagacherie, P., & McBratney, A. B. (2006). Chapter 1 Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping. *Developments in Soil Science*, *31*(C), 3–22. https://doi.org/10.1016/S0166-2481(06)31001-X

Lago, B. C., Silva, C. A., Melo, L. C. A., & Morais, E. G. de. (2021). Predicting biochar cation exchange capacity using Fourier transform infrared spectroscopy combined with partial least square regression. *Science of the Total Environment*, *794*, 148762. https://doi.org/10.1016/j.scitotenv.2021.148762

Lake, D. L., Kirk, P. W. W., & Lester, J. N. (1984). Fractionation, Characterization, and Speciation of Heavy Metals in Sewage Sludge and Sludge-Amended Soils: A Review. *Journal of Environmental Quality*, *13*(2), 175–183. https://doi.org/10.2134/jeq1984.00472425001300020001x

Lasat, M.M., 1999. Phytoextraction of metals from contaminated soil: a review of plant/soil/metal interaction and assessment of pertinent agronomic issues. Journal of Hazardous Substance Research, 2(1), p.5.

Lenntech. (2008). Manganese (Mn) - Chemical properties, Health and Environmental effects. http://www.lenntech.com/periodic/elements/mg.htm

Li, Z., Zhou, M., Xu, L. J., Lin, H., & Pu, H. (2014). Training sparse SVM on the core sets of fitting-planes. *Neurocomputing*, *130*, 20–27. https://doi.org/10.1016/j.neucom.2013.04.046

Li, L., Lu, J., Wang, S., Ma, Y., Wei, Q., Li, X., Cong, R. and Ren, T., 2016. Methods for estimating leaf nitrogen concentration of winter oilseed rape (Brassica napus L.) using in situ leaf spectroscopy. *Industrial Crops and Products*, *91*, pp.194-204.

Li, Y., Li, M., Li, C. and Liu, Z., 2020. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. *Scientific reports*, *10*(1), pp.1-12.

Lian, G., Guo, X., Fu, B., & Hu, C. (2009). Prediction of the spatial distribution of soil properties based on environmental correlation and geostatistics. *Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering*, *25*(7), 237–242. https://doi.org/10.3969/j.issn.1002-6819.2009.07.043

Liang, J., Feng, C., Zeng, G., Gao, X., Zhong, M., Li, X., Li, X., He, X., & Fang, Y. (2017). Spatial distribution and source identification of heavy metals in surface soils in a typical coal mine city, Lianyuan, China. *Environmental Pollution*, *225*, 681–690. https://doi.org/10.1016/j.envpol.2017.03.057

Linnik, V. G., Bauer, T. V., Minkina, T. M., Mandzhieva, S. S., & Mazarji, M. (2020). Spatial distribution of heavy metals in soils of the flood plain of the Seversky Donets River (Russia) based on geostatistical methods. *Environmental Geochemistry and Health*, 1–15. https://doi.org/10.1007/s10653-020-00688-y

Liu, W. X., Li, X. D., Shen, Z. G., Wang, D. C., Wai, O. W. H., & Li, Y. S. (2003). Multivariate statistical study of heavy metal enrichment in sediments of the Pearl River Estuary. *Environmental Pollution*, *121*(3), 377–388. https://doi.org/10.1016/S0269-7491(02)00234-8

Liu, X., Song, Q., Tang, Y., Li, W., Xu, J., Wu, J., Wang, F. and Brookes, P.C., 2013. Human health risk assessment of heavy metals in soil–vegetable system: a multi-medium analysis. *Science of the total environment*, *463*, pp.530-540.

López-Granados, F., Jurado-Expósito, M., Peña-Barragán, J. M., & García-Torres, L. (2005). Using geostatistical and remote sensing approaches for mapping soil properties. *European Journal of Agronomy*, *23*(3), 279–289. https://doi.org/10.1016/j.eja.2004.12.003

Loska, K., Cebula, J., Pelczar, J., Wiechuła, D., & Kwapuliński, J. (1997). Use of enrichment, and contamination factors together with geoaccumulation indexes to evaluate the content of Cd, Cu, and Ni in the Rybnik water reservoir in Poland. *Water, Air, & Soil Pollution*, *93*(1–4), 347–365. https://doi.org/10.1007/bf02404766

Luo, L., Ma, Y., Zhang, S., Wei, D., & Zhu, Y. G. (2009). An inventory of trace element inputs to agricultural soils in China. *Journal of Environmental Management*, *90*(8), 2524–2530. https://doi.org/10.1016/j.jenvman.2009.01.011

Luo, Y., Wu, L., Liu, L., Han, C., & Li, Z. (2009). Heavy metal contamination and remediation in Asian agricultural land. *National Institutes for Agro-Environmental Sciences*, *1(1)*, 1–9.

Iuss Working Group Wrb, 2015. World reference base for soil resources 2014, update 2015: International soil classification system for naming soils and creating legends for soil maps.

Łyszczarz, S., Błońska, E., & Lasota, J. (2020). The application of the geo-accumulation index and geostatistical methods to the assessment of forest soil contamination with heavy metals in the Babia Góra National Park (Poland). *Archives of Environmental Protection*, *46*(3), 69–79. https://doi.org/10.24425/aep.2020.134537

Maas, S., Scheifler, R., Benslama, M., Crini, N., Lucot, E., Brahmia, Z., Benyacoub, S. and Giraudoux, P., 2010. Spatial distribution of heavy metal concentrations in urban, suburban and agricultural soils in a Mediterranean city of Algeria. Environmental pollution, 158(6), pp.2294-2301.

Madrigal, J., Persky, V., Pappalardo, A. and Argos, M., 2018, September. Association of Heavy Metals with Measures of Pulmonary Function in Youth: Findings from the 2011-2012 National Health and Nutrition Examination Survey (NHANES). In *ISEE Conference Abstracts* (Vol. 2018, No. 1).

Mahmoudabadi, E., Sarmadian, F., & Nazary Moghaddam, R. (2015). Spatial distribution of soil heavy metals in different land uses of an industrial area of Tehran (Iran). *International Journal of Environmental Science and Technology*, *12*(10), 3283–3298. https://doi.org/10.1007/s13762-015-0808-z

Malone, B. P., Minasny, B., Odgers, N. P., & McBratney, A. B. (2014). Using model averaging to combine soil property rasters from legacy soil maps and from point data. *Geoderma*, *232–234*, 34–44. https://doi.org/10.1016/J.GEODERMA.2014.04.033

Martínez, C. E., & Motto, H. L. (2000). Solubility of lead, zinc and copper added to mineral soils. *Environmental Pollution*, *107*(1), 153–158. https://doi.org/10.1016/S0269-7491(99)00111-6

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003a). On digital soil mapping. *Geoderma*, *117*(1–2), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983–999. https://www.jmlr.org/papers/volume7/meinshausen06a/meinshausen06a.pdf

Molinaro, A.M., Simon, R. and Pfeiffer, R.M., 2005. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, *21*(15), pp.3301-3307.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D. and Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), pp.885-900.

Minasny, B., & McBratney, A. B. (2016a). Digital soil mapping: A brief history and some lessons. *Geoderma*, *264*, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Mishra, P. and Nikzad-Langerodi, R., 2020. Partial least square regression versus domain invariant partial least square regression with application to near-infrared spectroscopy of fresh fruit. Infrared Physics & Technology, 111, p.103547.

Müller, G. (1969). Index of geoaccumulation in sediments of the Rhine River. *GeoJournal, 2,* 108–118.

Nawar, S., & Mouazen, A. M. (2017). Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. *Sensors (Switzerland)*, *17*(10), 2428. https://doi.org/10.3390/s17102428

N'guessan, Y. M., Probst, J. L., Bur, T., & Probst, A. (2009). Trace elements in stream bed sediments from agricultural catchments (Gascogne region, S-W France): Where do they come from? *Science of the Total Environment*, *407*(8), 2939–2952. https://doi.org/10.1016/j.scitotenv.2008.12.047

Nicholson, F. A., Smith, S. R., Alloway, B. J., Carlton-Smith, C., & Chambers, B. J. (2003). An inventory of heavy metals inputs to agricultural soils in England and Wales. *Science of the Total Environment*, *311*(1–3), 205–219. https://doi.org/10.1016/S0048-9697(03)00139-6

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*(1), 1–13. https://doi.org/10.1186/1471-2105-11-110/FIGURES/6

Němeček, J., & Podlešáková, E. (1992). Retrospective experimental monitoring of heavy-metals containing in soils of the Czech Republic. *Rostlinna Vyroba*, *38*(6), 433-436.

Norris, G., 2008. Epa positive matrix factorization (pmf) 3.0 fundamentals & user guide, us. environmental protection agency. http://www. epa. gov/heasd/products/pmf/pmf. html.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. https://doi.org/10.1080/09332480.2012.685377

Nygård, T., Steinnes, E., & Røyset, O. (2012). Distribution of 32 elements in organic surface soils: Contributions from atmospheric transport of pollutants and natural sources. *Water, Air, and Soil Pollution*, *223*(2), 699–713. https://doi.org/10.1007/s11270-011-0895-5

Paatero, P. (1997). Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, *37*(1), 23–35. https://doi.org/10.1016/S0169-7439(96)00044-5

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*(2), 111–126. https://doi.org/10.1002/env.3170050203

Paatero, P., Eberly, S., Brown, S. G., & Norris, G. A. (2014). Methods for estimating uncertainty in factor analytic solutions. Atmos Meas Tech, 7(3), 781-797.

Padarian, J., Minasny, B., & McBratney, A. B. (2019). Using deep learning for digital soil mapping. *Soil*, *5*(1), 79–89. https://doi.org/10.5194/soil-5-79-2019

Palsson, F., Sveinsson, J. R., & Ulfarsson, M. O. (2018). Sentinel-2 image fusion using a deep residual network. *Remote Sensing*, *10*(8), 1290.

Pavlů, L., Drábek, O., Stejskalová, Š., Tejnecký, V., Hradilová, M., Nikodem, A., & Borůvka, L. (2018). Distribution of aluminium fractions in acid forest soils: Influence of vegetation changes. *IForest*, *11*(6), 721–727. https://doi.org/10.3832/ifor2498-011

Peng, G., Bing, W., Guangpo, G., & Guangcan, Z. (2013). Spatial distribution of soil organic carbon and total nitrogen based on GIS and geostatistics in a small watershed in a hilly area of northern China. *PLoS ONE*, *8*(12), e83592. https://doi.org/10.1371/journal.pone.0083592

Penížek, V., & Borůvka, L. (2006). Soil depth prediction supported by primary terrain attributes: A comparison of methods. *Plant, Soil and Environment*, *52*(9), 424–430. https://doi.org/10.17221/3461-pse

Penizek, V., & Boruvka, L. (2008). The digital terrain model as a tool for improved delineation of alluvial soils. *Digital Soil Mapping with Limited Data*, 319–326. https://doi.org/10.1007/978-1-4020-8592-5_28

Penížek, V., Zádorová, T., Kodešová, R., & Vaněk, A. (2016). Influence of elevation data resolution on spatial prediction of colluvial soils in a luvisol region. *PLoS ONE*, *11*(11). https://doi.org/10.1371/journal.pone.0165699

Pereira, O.J.R., Melfi, A.J., Montes, C.R. and Lucas, Y., 2018. Downscaling of ASTER thermal images based on geographically weighted regression kriging. *Remote Sensing*, *10*(4), p.633.

PlantProbs.net. (2019). *Manganese in plants and soil*. https://plantprobs.net/plant/nutrientImbalances/sodium.html

Pohl, C., & Van Genderen, J. L. (1998). Review article multisensor image fusion in remote sensing: concepts, methods and applications. International journal of remote sensing, 19(5), 823-854. https://doi.org/10.1080/014311698215748

Probst, P., Bischl, B. and Boulesteix, A.L., 2018. Tunability: Importance of hyperparameters of machine learning algorithms. *arXiv preprint arXiv:1802.09596*.

Qu, M., Chen, J., Huang, B., & Zhao, Y. (2020). Enhancing apportionment of the point and diffuse sources of soil heavy metals using robust geostatistics and robust spatial receptor model with categorical soil-type data. *Environmental Pollution*, *265*, 114964. https://doi.org/10.1016/J.ENVPOL.2020.114964

Qu, J., Lei, J., Li, Y., Dong, W., Zeng, Z., & Chen, D. (2018). Structure tensor-based algorithm for hyperspectral and panchromatic images fusion. *Remote Sensing*, *10*(3), 373.

Quinlan, J. R. (1992). Learning with continuous classes. In *Australian Joint Conference on Artificial Intelligence*. World Scientiic.

Ripin, S.N.M., Hasan, S., Kamal, M.L. and Hashim, N.M., 2014. Analysis and pollution assessment of heavy metal in soil, Perlis. *The Malaysian Journal of Analytical Sciences*, *18*(1), pp.155-161.

Roth, R.E., 2009. The impact of user expertise on geographic risk assessment under uncertain conditions. Cartography and Geographic Information Science, 36(1), pp.29-43.

Salim, I., Sajjad, R.U., Paule-Mercado, M.C., Memon, S.A., Lee, B.Y., Sukhbaatar, C. and Lee, C.H., 2019. Comparison of two receptor models PCA-MLR and PMF for source identification and apportionment of pollution carried by runoff from catchment and sub-watershed areas with mixed land cover in South Korea. *Science of the Total Environment*, *663*, pp.764-775.

Smith, L.A. and Brauning, S.E., 1995. *Remedial options for metals-contaminated sites* (pp. 17-122). Boca Raton: CRC Press.

Scragg, A. (2005). *Environmental biotechnology*.

Seaward, M. R. D. ; Richardson, D. H. S. (1990). Atmospheric sources of metal pollution and effects on vegetation. *Heavy Metal Tolerance in Plants Evolutionary Aspects*, 75–92.

Sega, M., & Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 80–87. https://doi.org/10.1002/widm.12

Segal, D. (1982). Theoretical basis for differentiation of ferric-iron bearing minerals, using Landsat MSS data. In *Proceedings of Symposium for Remote Sensing of Environment, 2nd Thematic Conference on Remote Sensing for Exploratory Geology, Fort Worth, TX* (pp. 949-951).

Shahid, M., Khalid, S., Bibi, I., Bundschuh, J., Niazi, N.K. and Dumat, C., 2020. A critical review of mercury speciation, bioavailability, toxicity and detoxification in soil-plant environment: Ecotoxicology and health risk assessment. Science of the total environment, 711, p.134749.

Shi, W., Liu, J., Du, Z., Stein, A., & Yue, T. (2011). Surface modelling of soil properties based on land use information. *Geoderma*, *162*(3–4), 347–357. https://doi.org/10.1016/j.geoderma.2011.03.007

Sibanda, M., Mutanga, O., & Rouget, M. (2016). Discriminating rangeland management practices using simulated hyspIRI, landsat 8 OLI, sentinel 2 MSI, and VENμs spectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(9), 3957-3969.

Sihag, P., Keshavarzi, A. and Kumar, V., 2019. Comparison of different approaches for modeling of heavy metal estimations. *SN Applied Sciences*, *1*(7), pp.1-11.

Smiljanić et al. (2019). The main sources of heavy metals in the soil and pathways intake. VI International Congress “Engineering, Environment and Materials in Processing Industry.

Speich, M. J. R., Bernhard, L., Teuling, A. J., & Zappa, M. (2015). Application of bivariate mapping for hydrological classification and analysis of temporal change and scale effects in Switzerland. *Journal of Hydrology*, *523*, 804–821. https://doi.org/10.1016/J.JHYDROL.2015.01.086

Sun, L., Guo, D., Liu, K., Meng, H., Zheng, Y., Yuan, F., & Zhu, G. (2019). Levels, sources, and spatial distribution of heavy metals in soils from a typical coal industrial city of Tangshan, China. *Catena*, *175*, 101–109. https://doi.org/10.1016/j.catena.2018.12.014

Sun, R. and Chen, L., 2016. Assessment of heavy metal pollution in topsoil around Beijing metropolis. PLoS One, 11(5), p.e0155350.

Swartjes, F. A., & Siciliano, S. (2012). Dealing with Contaminated Sites: From Theory towards Practical Application. *Soil Science Society of America Journal*, *76*(2), 748–748. https://doi.org/10.2136/sssaj2011.0004br

Taghizadeh-Mehrjardi, R., Fathizad, H., Hakimzadeh Ardakani, M. A., Sodaiezadeh, H., Kerry, R., Heung, B., & Scholten, T. (2021). Spatio-temporal analysis of heavy metals in arid soils at the catchment scale using digital soil assessment and a random forest model. *Remote Sensing*, *13*(9), 1698. https://doi.org/10.3390/rs13091698

Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., Valavi, R., Davatgar, N., Behrens, T. and Scholten, T., 2020. Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space. *Remote Sensing*, *12*(7), p.1095.

Takata, Y., Funakawa, S., Akshalov, K., Ishida, N. and Kosaki, T., 2007. Spatial prediction of soil organic matter in northern Kazakhstan based on topographic and vegetation information. Soil science and plant nutrition, 53(3), pp.289-299.

Tao, S.Y., Zhong, B.Q., Lin, Y., Ma, J., Zhou, Y., Hou, H., Zhao, L., Sun, Z., Qin, X. and Shi, H., 2017. Application of a self-organizing map and positive matrix factorization to investigate the spatial distributions and sources of polycyclic aromatic hydrocarbons in soils from Xiangfen County, northern China. *Ecotoxicology and environmental safety*, *141*, pp.98-106.

Tchagang, A. B., & Valdés, J. J. (2019). Prediction of the Atomization Energy of Molecules Using Coulomb Matrix and Atomic Composition in a Bayesian Regularized Neural Networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11731 LNCS*, 793–803. https://doi.org/10.1007/978-3-030-30493-5_75

Telford, K., Maher, W., Krikowa, F., Foster, S., Ellwood, M. J., Ashley, P. M., Lockwood, P. V., & Wilson, S. C. (2009). Bioaccumulation of antimony and arsenic in a highly contaminated stream adjacent to the Hillgrove Mine, NSW, Australia. *Environmental Chemistry*, *6*(2), 133–143. https://doi.org/10.1071/EN08097

Thevenon, F., Guédron, S., Chiaradia, M., Loizeau, J. L., & Poté, J. (2011). (Pre-) historic changes in natural and anthropogenic heavy metals deposition inferred from two contrasting Swiss Alpine lakes. *Quaternary Science Reviews*, *30*(1–2), 224–233. https://doi.org/10.1016/j.quascirev.2010.10.013

Tóth, G., Hermann, T., Szatmári, G., & Pásztor, L. (2016). Maps of heavy metals in the soils of the European Union and proposed priority areas for detailed assessment. *Science of the Total Environment*, *565*, 1054–1062. https://doi.org/10.1016/j.scitotenv.2016.05.115

Trumbo, B. E. (1981). A theory for coloring bivariate statistical maps. *American Statistician*, *35*(4), 220–226. https://doi.org/10.1080/00031305.1981.10479360

Tyner, J.A., 2010. Principles of map design. Guilford Publications.

Umali, B.P., Oliver, D.P., Forrester, S., Chittleborough, D.J., Hutson, J.L., Kookana, R.S. and Ostendorf, B., 2012. The effect of terrain and management on the spatial variability of soil properties in an apple orchard. Catena, 93, pp.38-48.

United States Environmental Protection Agency, 1989. Risk assessment guidance for superfund: Human health evaluation manual: Part A; Interim final. Environmental Protection Agency.

USDHHS. (2007). Toxicological Profile for Lead. *ATSDR's Toxicological Profiles*. https://doi.org/10.1201/9781420061888_ch106

USEPA. (2002). Supplemental Guidance for Developing Soil Screening Levels for Superfund Sites, Appendix D-dispersion Factors Calculations. USA. *United States Environmental Protection Agency, Washington, DC*, *pp. 4–24 .*

Vacek, O., Vašát, R., & Borůvka, L. (2020). Quantifying the pedodiversity-elevation relations. *Geoderma*, *373*, 114441. https://doi.org/10.1016/j.geoderma.2020.114441

Vácha, R. (2021). Heavy Metal Pollution and Its Effects on Agriculture. *Agronomy*, *11*(9), 1719.

Van der Meer, F. (2012). Remote-sensing image analysis and geostatistics. *International Journal of Remote Sensing*, *33*(18), 5644–5676. https://doi.org/10.1080/01431161.2012.666363

Vapnik, V. (1995). The Nature of Statistical Learning Theory. *Technometrics*, *38*(4), 409. https://doi.org/10.2307/1271324

Vašát, R., Pavlů, L., Borůvka, L., Drábek, O., & Nikodem, A. (2013). Mapping the topsoil pH and humus quality of forest soils in the north bohemian Jizerské hory mts. region with ordinary, universal, and regression kriging: Cross-validation comparison. *Soil and Water Research*, *8*(3), 97–104. https://doi.org/10.17221/62/2012-swr

Vasudevan, S., Ramos, F., Nettleton, E., & Durrant-Whyte, H. (2009). Gaussian process modeling of large-scale terrain. *Journal of Field Robotics*, *26*(10), 812-840.

Vohland, M., Besold, J., Hill, J., & Fründ, H. C. (2011). Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. *Geoderma*, *166*(1), 198–205. https://doi.org/10.1016/j.geoderma.2011.08.001

Walker, D.J., Clemente, R., Roig, A. and Bernal, M.P., 2003. The effects of soil amendments on heavy metal bioavailability in two contaminated Mediterranean soils. Environmental Pollution, 122(2), pp.303-312.

Wang, G., Liu, H.Q., Gong, Y., Wei, Y., Miao, A.J., Yang, L.Y. and Zhong, H., 2017. Risk assessment of metals in urban soils from a typical industrial city, Suzhou, Eastern China. International journal of environmental research and public health, 14(9), p.1025.

Wang, J., Yang, R., & Bai, Z. (2015). Spatial variability and sampling optimization of soil organic carbon and total nitrogen for Minesoils of the Loess Plateau using geostatistics. *Ecological Engineering*, *82*, 159–164. https://doi.org/10.1016/j.ecoleng.2015.04.103

Wang Q, Liu J, Cheng S (2015) Heavy metals in apple orchard soils and fruits and their health risks in Liaodong Peninsula, Northeast China. Environ Monit Assess 187. https://doi.org/10.1007/s10661-014-4178-7

Wang, S., Zhu, L., Fuh, J.Y.H., Zhang, H. and Yan, W., 2020. Multi-physics modeling and Gaussian process regression analysis of cladding track geometry for direct energy deposition. Optics and Lasers in Engineering, 127, p.105950.

Wang, Y. and Witten, I.H., 1996. Induction of model trees for predicting continuous classes. https://researchcommons.waikato.ac.nz/handle/10289/1183

Wang, F., Guan, Q., Tian, J., Lin, J., Yang, Y., Yang, L. and Pan, N., 2020. Contamination characteristics, source apportionment, and health risk assessment of heavy metals in agricultural soil in the Hexi Corridor. Catena, 191, p.104573.

Wang WX (2013) Dietary toxicity of metals in aquatic animals: recent studies and perspectives. Chinese Sci Bull 58:203–213. https://doi.org/10.1007/s11434-012-5413-7

Wang, K., Zhang, C. and Li, W., 2012. Comparison of geographically weighted regression and regression kriging for estimating the spatial distribution of soil organic matter. *Giscience & remote sensing*, *49*(6), pp.915-932.

Weather Spark. (2016). *Average Weather in Frýdek-Místek, Czechia, Year Round - Weather Spark*. https://weatherspark.com/y/83671/Average-Weather-in-Frýdek-Místek-Czechia-Year-Round

Weissmannová, H.D. and Pavlovský, J., 2017. Indices of soil contamination by heavy metals–methodology of calculation for pollution assessment (minireview). Environmental monitoring and assessment, 189(12), pp.1-25.

Wilson, N. J., Craw, D., & Hunter, K. (2004). Antimony distribution and environmental mobility at an historic antimony smelter site, New Zealand. *Environmental Pollution*, *129*(2), 257–266. https://doi.org/10.1016/j.envpol.2003.10.014

Wu, Y., Chen, J., Ji, J., Gong, P., Liao, Q., Tian, Q., & Ma, H. (2007). A Mechanism Study of Reflectance Spectroscopy for Investigating Heavy Metals in Soils. *Soil Science Society of America Journal*, *71*(3), 918–926. https://doi.org/10.2136/sssaj2006.0285

Wuana, R. A., & Okieimen, F. E. (2011). Heavy Metals in Contaminated Soils: A Review of Sources, Chemistry, Risks and Best Available Strategies for Remediation. *ISRN Ecology*, *2011*, 1–20. https://doi.org/10.5402/2011/402647

Wuana, R.A. and Okieimen, F.E., 2011. Heavy metals in contaminated soils: a review of sources, chemistry, risks and best available strategies for remediation. International Scholarly Research Notices, 2011.

Xia, C. and Zhang, Y., 2022. Comparison of the use of Landsat 8, Sentinel-2, and Gaofen-2 images for mapping soil pH in Dehui, northeastern China. Ecological Informatics, p.101705.

Xiang, H. and Tian, L., 2011. Development of a low-cost agricultural remote sensing system based on an autonomous unmanned aerial vehicle (UAV). Biosystems engineering, 108(2), pp.174-190.

Xie, X. L., Pan, X. Z., & Sun, B. (2012). Visible and Near-Infrared Diffuse Reflectance Spectroscopy for Prediction of Soil Properties near a Copper Smelter. *Pedosphere*, *22*(3), 351–366. https://doi.org/10.1016/S1002-0160(12)60022-8

Xu, Y., Smith, S. E., Grunwald, S., Abd-Elrahman, A., & Wani, S. P. (2017). Evaluating the effect of remote sensing image spatial resolution on soil exchangeable potassium prediction models in smallholder farm settings. *Journal of Environmental Management*, *200*, 423-433.

Xu, X., Du, C., Ma, F., Shen, Y., Wu, K., Liang, D. and Zhou, J., 2019. Detection of soil organic matter from laser-induced breakdown spectroscopy (LIBS) and mid-infrared spectroscopy (FTIR-ATR) coupled with multivariate techniques. *Geoderma*, *355*, p.113905.

Yang, B., Zhou, L., Xue, N., Li, F., Li, Y., Vogt, R.D., Cong, X., Yan, Y. and Liu, B., 2013. Source apportionment of polycyclic aromatic hydrocarbons in soils of Huanghuai Plain, China: comparison of three receptor models. *Science of the Total Environment*, *443*, pp.31-39.

Ye, H., Huang, W., Huang, S., Huang, Y., Zhang, S., Dong, Y. and Chen, P., 2017. Effects of different sampling densities on geographically weighted regression kriging for predicting soil organic carbon. *Spatial statistics*, *20*, pp.76-91.

Zachara, J.M. and Westall, J.C., 2018. Chemical modeling of ion adsorption in soils. In Soil physical chemistry (pp. 47-96). CRC Press.

Zawadzki, J., Cieszewski, C. J., Zasada, M., & Lowe, R. C. (2005). Applying geostatistics for investigations of forest ecosystems using remote sensing imagery. *Silva Fennica Monographs*, *39*(4), 599–617. https://doi.org/10.14214/sf.369

Zeraatpisheh, M., Jafari, A., Bodaghabadi, M.B., Ayoubi, S., Taghizadeh-Mehrjardi, R., Toomanian, N., Kerry, R. and Xu, M., 2020. Conventional and digital soil mapping in Iran: Past, present, and future. Catena, 188, p.104424.

Zha, Y., Gao, J., & Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International journal of remote sensing*, *24*(3), 583-594.

Zhang, S., Huang, Y., Shen, C., Ye, H. and Du, Y., 2012. Spatial prediction of soil organic matter using terrain indices and categorical variables as auxiliary information. Geoderma, 171, pp.35-43.

Zhang, G. lin, LIU, F., & SONG, X. dong. (2017). Recent progress and future prospect of digital soil mapping: A review. *Journal of Integrative Agriculture*, *16*(12), 2871–2885. https://doi.org/10.1016/S2095-3119(17)61762-3

Zhang, J., Li, H., Zhou, Y., Dou, L., Cai, L., Mo, L., & You, J. (2018). Bioavailability and soil-to-crop transfer of heavy metals in farmland soils: A case study in the Pearl River Delta, South China. *Environmental Pollution*, *235*, 710–719. https://doi.org/10.1016/j.envpol.2017.12.106

Zhang, M. K., Liu, Z. Y., & Wang, H. (2010). Use of single extraction methods to predict bioavailability of heavy metals in polluted soils to rice. *Communications in Soil Science and Plant Analysis*, *41*(7), 820–831. https://doi.org/10.1080/00103621003592341

Zhang, W. and Goh, A.T., 2016. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geoscience Frontiers, 7(1), pp.45-52.

Zhang, Y., & Xu, X. (2021). Fe-Based Superconducting Transition Temperature Modeling through Gaussian Process Regression. *Journal of Low Temperature Physics*, *202*(1–2), 205–218. https://doi.org/10.1007/S10909-020-02545-9

Zhao, K., Liu, X., Xu, J., & Selim, H. M. (2010). Heavy metal contaminations in a soil-rice system: Identification of spatial dependence in relation to soil properties of paddy fields. *Journal of Hazardous Materials*, *181*(1–3), 778–787. https://doi.org/10.1016/j.jhazmat.2010.05.081

Zhao, L., Xu, Y., Hou, H., Shangguan, Y. and Li, F., 2014. Source identification and health risk assessment of metals in urban soils around the Tanggu chemical industrial district, Tianjin, China. Science of the total environment, 468, pp.654-662.

Zhao, X., Ye, Y., Zhou, J., Liu, L., Dai, W., Wang, Q., & Hu, Y. (2017). Comprehensive evaluation of cultivated land quality and sensitivity analysis of index weight in hilly region of Pearl River Delta. *Transactions of the Chinese Society of Agricultural Engineering*, *33*(8), 226-235.

Zheng, S., Wang, Q., Yuan, Y. and Sun, W., 2020. Human health risk assessment of heavy metals in soil and food crops in the Pearl River Delta urban agglomeration of China. Food chemistry, 316, p.126213.

Zhou, T., Geng, Y., Ji, C., Xu, X., Wang, H., Pan, J., Bumberger, J., Haase, D. and Lausch, A., 2021. Prediction of soil organic carbon and the C: N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. Science of The Total Environment, 755, p.142661.

Zhuang, P., McBride, M. B., Xia, H., Li, N., & Li, Z. (2009). Health risk from heavy metals via consumption of food crops in the vicinity of Dabaoshan mine, South China. *Science of the total environment*, *407*(5), 1551-1561.

Zhu, A., Lu, G., Liu, J., Qin, C., & Zhou, C. (2018a). Spatial prediction based on Third Law of Geography. *Annals of GIS*, *24*(4), 225–240. https://doi.org/10.1080/19475683.2018.1534890

Zhu, A. X., Hudson, B., Burt, J., Lubich, K., & Simonson, D. (2001). Soil Mapping Using GIS, Expert Knowledge, and Fuzzy Logic. *Soil Science Society of America Journal*, *65*(5), 1463–1472. https://doi.org/10.2136/sssaj2001.6551463x

LIST OF PUBLICATIONS

WEB OF SCIENCE

1. **Agyeman, P. C.,** Kingsley, J., Kebonye, N. M., Khosravi, V., Borůvka, L., & Vašát, R. (2022). Prediction of the concentration of antimony in agricultural soil using data fusion, terrain attributes combined with regression kriging. ***Environmental Pollution*****,** 120697. **IF = 9.988**
2. **Agyeman, P. C.,** Kebonye, N. M., Khosravi, V., Kingsley, J., Borůvka, L., Vašát, R., & Boateng, C. M. (2023). Optimal zinc level and uncertainty quantification in agricultural soils via visible near-infrared reflectance and soil chemical properties. ***Journal of Environmental Management*****,** *326*, 116701. **IF = 8.91**
3. **Agyeman, P. C.,** Kingsley, J. O. H. N., Kebonye, N. M., Ofori, S., Borůvka, L., Vašát, R., & Kočárek, M. (2022). Ecological risk source distribution, uncertainty analysis, and application of geographically weighted regression cokriging for prediction of potentially toxic elements in agricultural soils. ***Process Safety and Environmental Protection*****,** *164*, 729-746. **IF = 7.926**
4. **Agyeman, P. C.,** Khosravi, V., Kebonye, N. M., John, K., Borůvka, L., & Vašát, R. (2022). Using spectral indices and terrain attribute datasets and their combination in the prediction of cadmium content in agricultural soil. **Computers and Electronics in Agriculture,** 198, 107077.**IF = 6.757**
5. **Agyeman, P. C.,** John, K., Kebonye, N. M., Borůvka, L., Vašát, R., Drábek, O., Němeček, K. 2021. Human health risk exposure and ecological risk assessment of potentially toxic element pollution in agricultural soils in the district of Frydek Mistek, Czech Republic: a sample location approach. **Environmental Sciences Europe**, 33(1), 1-25. **IF = 5.481**
6. **Agyeman, P. C.,** John, K., Kebonye, N. M., Borůvka, L., Vašát, R., Drábek, O. 2021. A geostatistical approach to estimating source apportionment in urban and peri-urban soils using the Czech Republic as an example**. Scientific reports**, 11(1), 1-15. **IF = 4.997**
7. **Agyeman, P. C.,** Kebonye, N. M., **John, K**., Borůvka, L., Vašát, R., Fajemisim, O. 2022. Prediction of nickel concentration in peri-urban and urban soils using hybridized empirical bayesian kriging and support vector machine regression. **Scientific Reports,** 12(1), 1-16. **IF = 4.997**
8. **Agyeman, P. C.,** John, K., Kebonye, N. M., Ahado, S. K., Borůvka, L., Němeček, K., Vašát, R., 2021. Multi-geochemical background comparison and the identification of the best normalizer for the estimation of PTE contamination in agricultural soil. **Environmental Geochemistry and Health** 1-17. **IF = 4.898**
9. **Agyeman, P. C.,** Ahado, S. K., Borůvka, L., Biney, J. K. M., Sarkodie, V. Y. O., Kebonye, N. M., & Kingsley, J. (2021). Trend analysis of global usage of digital soil mapping models in the prediction of potentially toxic elements in soil/sediments: a bibliometric review. *Environmental Geochemistry and Health*, *43*(5), 1715-1739. **IF = 4.898**
10. **Agyeman, P. C.,** Ahado, S. K., John, K., Kebonye, N. M., Biney, J. K. M., Borůvka, L., Vasat, R., Kocarek, M., 2021. Source apportionment, contamination levels, and spatial prediction of potentially toxic elements in selected soils of the Czech Republic.

Environmental geochemistry and health 43: 601-620. **IF = 4.898**

11. **Agyeman, P. C.,** John, K., Kebonye, N. M., Borůvka, L., & Vašát, R. (2022). Combination of enrichment factor and positive matrix factorization in the estimation of potentially toxic element source distribution in agricultural soil. ***Environmental Geochemistry and Health*****,** 1-27. **IF = 4.898**
12. **Agyeman, P. C.,** Ahado, S. K., John, K., Kebonye, N. M., Vašát, R., Borůvka, L., Kočárek, M., Němeček, K., 2021. Health risk assessment and the application of CF-PMF: a pollution assessment–based receptor model in an urban soil. **Journal of Soils and Sediments 21:** 3117-3136. **IF = 3.536**
13. Kebonye, N. M., **Agyeman, P. C**., Seletlo, Z., & Eze, P. N. (2022). On exploring bivariate and trivariate maps as visualization tools for spatial associations in digital soil mapping: A focus on soil properties. *Precision Agriculture*, 1-22. **IF = 5.875**
14. John, K., **Agyeman, P. C.,** Kebonye, N. M., Isong, I. A., Ayito, E. O., Ofem, K. I., Qin, C. Z., 2021. Hybridization of cokriging and gaussian process regression modelling techniques in mapping soil sulphur. **Catena 206:** 105534. **IF= 5.198**
15. Kebonye, N. M., Agyeman, P. C., & Biney, J. K. (2022). Using an innovative bivariate colour scheme to infer spatial links and patterns between prediction and uncertainty: an example based on an explainable soil CN ratio model. ***Modeling Earth Systems and Environment***, 1-8.
16. John, K., Bouslihim, Y., Ofem, K. I., Hssaini, L., Rezouk, R., **Agyeman, P. C.,** Kebonye, N. M., Penížek, V., Zádorová, T. 2022. Mapping of soil nutrients via different covariates combinations using kriging with external drift: theory and an example from Morocco. **Ecological Processes 11: 23. IF = 2.849**
17. John, K., Bouslihim Y., Ofem, K.I., Hssaini, L., Razouk, R., Okon, P. B., Isong A.I., **Agyeman, P. C., Kebonye,** N.M., Qin, C.Z., 2021. Do model choice and sample ratios separately or simultaneously influence soil organic matter prediction? **International Soil and Water Conservation Research.** https://doi.org/10.1016/j.iswcr.2021.11.003. **IF = 6.027**
18. Kebonye, N. M., Eze, P. N., **Agyeman, P. C.,** John, K., Kudjo, A. S. Efficiency of the t-distribution stochastic neighbor embedding technique for detailed visualization and modeling interactions between agricultural soil quality indicators. **Biosystems Engineering 210:** 282-298.
19. John, K., Afu, S. M., Isong, I. A., Aki, E. E., Kebonye, N. M., Ayito, E. O., **Chapman, P. A.,** Eyong, M. O., Penížek, V., 2021. Mapping soil properties with soil-environmental covariates using geostatistics and multivariate statistics. **International Journal of Environmental Science and Technology 18:** 3327–3342.
20. John, K., Afu, S. M., Isong, I. A., **Chapman, P. A.,** Kebonye, N. M., Ayito, E. O., 2021. Estimation of soil organic carbon distribution by geostatistical and deterministic interpolation methods: a case study of the Southeastern soils of Nigeria. **Environmental Engineering & Management Journal 20:** 1077-1085.
21. John, K., Isong, A. I., Kebonye, N.M., Ayito, E.O., **Agyeman, P.C.,** Afu, M. S., 2020. Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil**. Land 9:** 487. **IF= 3.398**
22. Kebonye, N. M., Eze, P. N., John, K., **Agyeman, P. C.,** Němeček, K., Borůvka, L., 2021. An

in-depth human health risk assessment of potentially toxic elements in highly polluted riverine soils, Příbram (Czech Republic). **Environmental Geochemistry and Health 44:** 369–385.

23. Kebonye, Ndiye M., John, K., Chakraborty, S., **Agyeman, P.C.,** Ahado, S. K., Eze, P.N., Němeček, K., Drábek, O., Borůvka, L., 2021. Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy. **Geoderma 384:** 114792.
24. Kebonye, N. M., **Agyeman, P. C.,** & Biney, J. K. (2022). Using an innovative bivariate colour scheme to infer spatial links and patterns between prediction and uncertainty: an example based on an explainable soil CN ratio model. ***Modeling Earth Systems and Environment*****,** 1-8.
25. Biney, JKM, Saberioon, M., Borůvka, L., Houška, J., Vašát, R., **Chapman Agyeman, P.,** ... & Klement, A. (2021). Exploring the Suitability of UAS-Based Multispectral Images for Estimating Soil Organic Carbon: Comparison with Proximal Soil Sensing and Spaceborne Imagery. ***Remote Sensing*** **,** *13* (2), 308.
26. John, K., Kebonye, N. M., **Agyeman, P. C.,** & Ahado, S. K. (2021). Comparison of Cubist models for soil organic carbon prediction via portable XRF measured data. ***Environmental Monitoring and Assessment*****,** *193*(4), 1-15.
27. Biney, J. K. M., Borůvka, L., **Chapman Agyeman, P.,** Němeček, K., & Klement, A. (2020). Comparison of field and laboratory wet soil spectra in the Vis-NIR range for soil organic carbon prediction in the absence of laboratory dry measurements. *Remote Sensing*, *12*(18), 3082.

SCOPUS

28. John, K., Abraham, I. I., Kebonye, N. M., **Agyeman, P. C.,** Ayito, E. O., Kudjo, A. S., 2021. Soil organic carbon prediction with terrain derivatives using geostatistics and sequential Gaussian simulation. **Journal of the Saudi Society of Agricultural Sciences 20:** 379-389.

OTHERS

29. Ofori, S., **Agyeman, P. C.,** Adotey, E. K., Růžičková, I., & Wanner, J. (2022). Assessing the influence of treated effluent on nutrient enrichment of surface waters using water quality indices and source apportionment. ***Water Practice & Technology***, *17*(7), 1523-1534.
30. Kebonye, N. M., **Agyeman, P. C.,** & Biney, J. K. (2023). Optimized modelling of countrywide soil organic carbon levels via an interpretable decision tree. Smart Agricultural Technology, 3, 100106.

IN REVIEW

Agyeman, P. C., Kebonye, N. M., Khosravi, V., Kingsley, J., Borůvka, L., Vašát, R., & Boateng, C. M. Using machine learning algorithms in combination with terrain attributes and visible near infrared to estimate and map lead variability in agricultural soil. CATENA

Agyeman, P. C., Borůvka, L., N., Kebonye, Khosravi, V., Kingsley, J. Quantification of the concentration of cadmium in agricultural soil using legacy data, preferential sampling, sentinel 2, Landsat 8, and an ensemble model. Journal of Environmental Management

Agyeman, P. C., N., Kebonye, Kingsley, J. Haghnazar H. Borůvka, L., Vašát, R., Compositional mapping, uncertainty assessment, and source apportionment via pollution assessment-based receptor models in urban and peri-urban agricultural soils. Soil and Sediments

Kebonye, N. M., **Agyeman, P. C.,** & Biney, J. K. On Exploring Umap for Heterogeneous Environmental Dataset Reduction and Visualization an Example Based on Soil Nutrient Levels. Computers and Electronics in Agriculture.

Khosravi, V., Gholizadeh, **A., Chapman Agyeman, P.,** Ardejani, F. D., Yousefi, S., and Saberioon, MFurther to quantification of content, can reflectance spectroscopy determine the speciation of toxic elements in a mine waste dump?

Computers and Electronics in Agriculture 198 (2022) 107077

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Using spectral indices and terrain attribute datasets and their combination in the prediction of cadmium content in agricultural soil

Prince Chapman Agyeman *, Vahid Khosravi, Ndiye Michael Kebonye, Kingsley John, Luboš Borůvka, Radim Vašát

Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague, Czech Republic

ARTICLE INFO

Keywords:
Agricultural soil
Cadmium
Digital soil mapping
Spectral indices
Terrain attributes
Environmental covariates

ABSTRACT

The continuous demand placed on farmland to yield optimal harvest is dependent on the continuous application of agrochemicals and fertilizers to increase soil fertility and manage diseases. Successive application of fertilizers and use of agrochemicals coupled with metal and steel industries introduce potentially toxic elements into the soil. Active agricultural activities and industrial emissions that result in atmospheric cadmium (Cd) injection and active deposition on agricultural soil (particularly from the primary metal industry, steel and iron industrial production). The concentration of cadmium in the study area exceeds the local background value. As a result, excessive cadmium soil concentration will contribute to increased toxic and carcinogenic effects, with negative implications for both environmental and human health. Therefore, determining the spatial distribution of Cd is critical for environmentally friendly agricultural production and reducing Cd emission into soils. The goals of this study are to (i) determine the variability of Cd prediction in agricultural soil using spectral indices or terrain attributes coupled with modeling algorithms, and (ii) determine whether combining spectral indices and terrain attributes coupled with modeling algorithms can improve Cd prediction efficiency in agricultural soil. The study applied three modelling scenarios, comprised prediction using terrain attributes coupled with digital soil mapping (DSM) approaches (Scenario 1), prediction using spectral indices combined with DSMs (Scenario 2), and prediction using a combination of terrain attributes, spectral indices, and DSMs (Scenario 3). Gaussian process regression (GPR), partial least square regression (PLSR), extreme gradient boosting (EGB), multivariate adaptive regression splines (MARS), Bayesian regularized neural network (BRNN), regularized random forest (RRF), Bayesian generalized linear model (BGLM), and M5 tree models were the DSMs used in the study. The M5 tree model and terrain attributes {Scenario 1 $R^2 = 0.77$, concordance correlation coefficient (CCC) = 0.73, root mean square error (RMSE) = 0.45, mean absolute error (MAE) = 0.37 and median absolute error (MdAE) = 0.35}, EGB and spectral indices {Scenario 2, $R^2 = 0.83$, CCC = 0.76, RMSE = 0.54, MAE = 0.33 and MdAE = 0.23} and the M5 tree model, spectral indices and terrain attributes {Scenario 3, $R^2 = 0.84$, CCC = 0.81, RMSE = 0.39, MAE = 0.31 and MdAE = 0.24} were the overall best combinations that predicted Cd in the agricultural soil. The overall evaluation of the approaches suggested that the combination of spectral indices, terrain attributes, and the M5 tree model in Scenario 3 was the optimal technique for predicting Cd in agricultural soil. Thus, a combination of environmental covariates with a high correlation with the response variable, combined with appropriate modeling techniques predicting potentially toxic elements in agricultural soil, will produce the best results.

1. Introduction

Soil pollution by potentially toxic elements (PTEs) is a worldwide problem due to the negative implications on the ecosystem and the prospective risk to human health. PTEs are an umbrella term for "heavy metals," "trace elements," and "toxic elements" with weight densities greater than or less than 5 g/cm^2 (Ali et al., 2013; Fang et al., 2016). The ever-increasing human population and efforts to meet their demands have culminated in the expansion of farmlands and the application of tons of agrochemicals and fertilizers to increase yield. According to Adimalla, (2020) and Song et al., (2018), as agricultural advancements have risen tremendously, PTEs pollution has been increased as a result of

* Corresponding author.
E-mail address: agyeman@af.czu.cz (P.C. Agyeman).

https://doi.org/10.1016/j.compag.2022.107077
Received 21 March 2022; Received in revised form 15 May 2022; Accepted 19 May 2022
Available online 26 May 2022

the application of fertilizers. Due to the high level of PTEs in agricultural soil, it has become challenging for soils to fulfill their functional role as a natural resource for the continued coexistence of plants, animals, and humanity, which is dependent on the balances of their structure and composition, as well as the chemical, biological, and physical properties (Gupta et al., 2020; Lehmann et al., 2020). The direct impact of PTEs, specifically cadmium (Cd), lead (Pb), mercury (Hg), and arsenic (As), has reduced the soil's ability to play its potential role as a habitat for macro- and microorganisms, resulting in soil deterioration, endangering food quality, reliability, and security, and exacerbating potential hazards to human health via the food chain (Jia et al., 2019; Shi et al., 2014b).

Over the years the continuous agricultural soil research and efforts to improve precision agriculture, thereby reducing the use of agrochemicals and the application of fertilizers that potentially pollute farmland, have piqued the interest of people all over the world. These investigations have resulted in a critical assessment of PTEs in agricultural soil to determine concentration levels as well as their effect, as the soil is used to grow food crops for human consumption. According to, Keshavarzi et al., (2018); Adimalla and Wang, (2018); Adimalla et al., (2019); and Agyeman et al. (2021), a premium has been placed on agricultural soil for two reasons: first, the agricultural food channel that is polluted is the primary source of consumption directly across different food products., such as vegetables, fruits, rice, and wheat, which can sometimes trigger health potential dangers; and second, densely gathered PTEs permeate through the microscopic pores and enter the soil and groundwater system, deteriorating soil/groundwater quality, which seems to have significant consequences on living beings.

Digital soil mopping (DSM) approaches can be classified in four categories namely, the conventional statistical techniques such as multiple linear regression (Agyeman et al., 2021), geostatistical approach like ordinary kriging (Agyeman et al., 2021), machine learning algorithm such as support vector machine (Asgari et al., 2020a,b; Sakizadeh et al., 2017; Tajik et al., 2019, 2020; Zeraatpisheh et al., 2020) and combination of to two or more modeling approaches to form a hybridized model such as ensemble (Agyeman et al., 2021; Chen et al., 2019). According to Minasny and McBratney (2016), DSM has effectively converted into an important subdiscipline of soil science. However, spatial variability of soil physiochemical properties within or between soils are generally intrinsic in nature due to geological and pedological soil formation factors, even though some variability might very well be induced by other management practices (Iqbal et al., 2005). The variables interfere on a spatially and temporally scale, and the distribution of soil properties modifies the content even further (Agyeman et al., 2021). Zhu et al., (2018) however indicated that, in spatial predictions environmental covariates (i.e., spectral indices, digital elevation model) and soil relationships are fitted with a model as well as the discovered relationship and then implemented to spaces or positions where the soil data or sediment data is available. DSM methods are technically quantitative soil-environment interactions based on observed points that distinguish the relationship between soil and environmental covariates like terrain attributes, spectral indices, and climatic datasets. DSM has been widely used in soil science around the world for mapping soil properties and classes, as well as, to a significant extent, predicting PTE concentrations in soil or sediments (Arrouays et al., 2014).

Over the years, the use of environmental covariates in DSM has proven to be useful, reliable, and effective in predicting and mapping PTEs (Azizi et al., 2022) or soil properties and farmers cannot apply DSM to render judgement due to spectral and spatial resolution issues (Zare et al., 2021). Multiplicity of research has applied either one or more terrain attributes (John et al., 2021), remote sensing dataset (Zhang et al., 2017) and climatic data (John et al., 2021b) to serve as an auxiliary dataset coupled with measured data to predict PTEs or other properties in the soil. Environmental covariates have enhanced multiple facets of soil surveying around the world, which include pre-mapping, designing efficient and effective field sampling strategies, and implementing spatial prediction approaches (Boettinger, 2010). Based on soil-forming factors and the SCORPAN technique, the DSM technique measure the connections between soil properties or PTEs and environmental factors (McBratney et al., 2003). There are several studies that have applied machine learning algorithm (MLA) in the prediction and mapping of PTE and soil properties including extreme gradient boosting (Goydaragh et al., 2021), cubist, (Biney et al., 2022), support vector machine regression, (wan et al., 2019), random forest, (Mao et al., 2021), partial least square regression (Vašát et al., 2017), extreme learning machine (Khosravi et al., 2018) and artificial neural network (Pyo et al., 2020). Numerous studies have used MLA in conjunction with environmental covariates such as terrain attributes and remote sensing datasets (e.g., spectral indices, Sentinel 2, Landsat 8) to predict PTEs in agricultural soil (Shi et al., 2014a; Wang et al., 2018, 2014; Zhao et al., 2012). However, finding research that uses spectral indices, terrain attributes, and their combinations to predict PTE content is challenging.

Active agriculture, metal and steel industries are among the industrial activities in the study area, so determining the chemical composition of agricultural soil is critical. The goals of this study are to (i) determine the variability of Cd prediction in agricultural soil using spectral indices or terrain attributes coupled with modeling algorithms, and (ii) determine whether combining spectral indices and terrain attributes coupled with modeling algorithms can improve Cd prediction efficiency in agricultural soil. Thus, the current study intends to use selected DSM approaches and environmental covariates such as terrain attributes and spectral indices estimated from Sentinel-2 satellite imagery for the prediction of Cd in agricultural soil by three distinct approaches, namely, assessing the impact of terrain attributes in the prediction of Cd (Scenario 1), assessing the impact of spectral indices in the prediction of Cd (Scenario 2) and assessing the impact of combined terrain attributes and spectral indices in the prediction of Cd (Scenario 3). We hypothesized that combining environmental covariates (spectral indices and terrain attributes) with an appropriate MLA has the potential to improve prediction efficiency.

2. Materials and methods

2.1. Study area

The study location is in the district of Frydek Mistek, Czech Republic. It is located at latitude of 49° 41′ 0′ north and longitude of 18° 20′ 0′ east, at elevation of 225 to 327 m above sea level (Agyeman et al., 2020). It has hilly topography and highlands from the exterior Carpathians. According to the Koppen classification, the study area has a Cfb = oceanic temperate climate with high rainfall even during dry months (John et al., 2021). Over the year, the temperature ranges from −5 to 24 °C, with average temperatures falling below −14 °C or increasing above 30 °C. The maximum annual precipitation is 83 mm, with a minimum overall accumulation of 17 mm (Weather Spark, 2016). Crops grown in the study area include oilseeds, corn, sunflower, and grapevines, as well as cereals like wheat, oats, barley, and rye. The district of Frydek Mistek has a total land area of 1208 km^2 (39.38 % for agricultural activities and 49.36% for forestland), and the land area utilized for this study is 889.8 km^2. The study area is identified by extensive agriculture activities as well as different metal works (such as fabrication, pneumatic cylinders, valves, regulators, and so on) and steel companies (such as the production of cold-rolled steel strips and sheets, anisotropic electrical steel strips and sheets). The color, structure, and carbonate composition of the soil are all easily distinguishable. Nonetheless, the soil's parent materials have a medium to fine texture. They are most commonly found in aeolian and colluvial deposits, which are also defined by top and subsurface mottles that can be observed in some soil regions and are primarily followed by concrete and whitening. They are differentiated by a cambic diagnostic horizon with fine sandy loam (e.g., cambic horizon and anthropedogenic horizons) texture, a clay concentration of more than 4%, and a lithic discontinuity with low carbonate

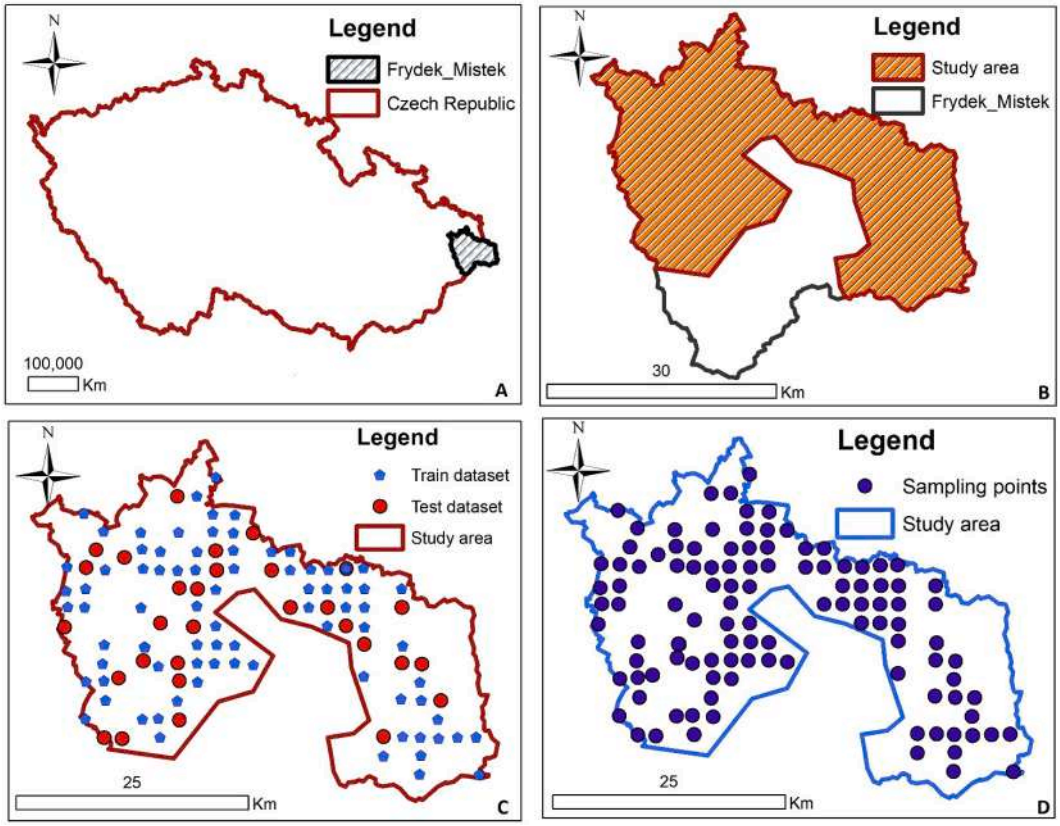


Fig. 1. Czech Republic (A), the district of Frydek Mistek (B), Study area with training and test dataset used (C) Study area with sampling locations (D).

content (Kozák et al. 2010). However, cambisols and stagnosols were the most prevalent soil types in the study area (Kozák et al. 2010). These soils are common throughout the Czech Republic and can be found at elevations ranging from 455.1 to 493.5 m (Vacek et al., 2020).

2.2. Soil analysis and sampling

A total of 115 topsoil samples were collected from agricultural land in the Frydek Mistek district (Fig. 1). The sample pattern was a standard grid, and the soil sample ranges were kept to 2 X 2 km using a handheld GPS (Leica-Zeno 5 GPS) device at depths ranging from 0 to 20 cm. Soil samples were placed in Ziploc bags, labelled, and delivered to the laboratory. The soil samples obtained were air-dried before being crushed by a machine (Fritsch disk mill pulverize) and sieved to obtain a pulverized soil sample (2 mm). In a Teflon container, one gram of the dried, 7 ml homogenized, and sieved soil sample (sieve size 2 mm) was placed and labelled. of 35% HCl and 3 ml of 65% HNO_3 were administered (utilizing automatic dispensers—one for each acid) into each teflon bottle, and the cap was gently closed to enable the sample to sit overnight for reactions to occur (aqua regia procedure) (Cools, 2016; Tejnecký et al., 2015). After the soil sample had been digested, the mixture was deposited on a heated metal plate for 2 h to aid digestion before being permitted to cool. The mixture was filtered to obtain supernatant. The supernatant was diluted to 50 ml with deionized water in a 50- ml volumetric flask. After that, the diluted supernatant was filtered into PVC tubes with a capacity of 50 ml. The concentration of PTEs was determined by ICP–OES (inductively coupled plasma–optical emission spectrometry) (Thermo Fisher Scientific Corporation, USA) in compliance with standard techniques and methods. Similarly, the quality control and quality assurance methods were ensured by examining the reference criteria for each study. To verify that the mistake was minimal, a duplicate analysis was performed.

2.3. Modelling techniques

The data was divided into a test dataset (with 25% for validation) and a training dataset using a random approach (75% for calibration). The training data was utilized to calibrate the regression models, and the test dataset was used to evaluate generalization abilities. All the modeling techniques were conditioned to a 10-fold cross-validation process that was repeated five times. The DSM approaches that were employed in this study are: Gaussian process regression (GPR), partial least square regression (PLSR), extreme gradient boosting (EGB), multivariate adaptive regression splines (MARS), Bayesian regularized neural network (BRNN), regularized random forest (RRF), Bayesian generalized linear model (BGLM) and the M5 tree model. Below is a brief description of the modeling approaches:

2.4. Gaussian process regression

The Gaussian process (GPR) is a nonparametric modelling method (Vasudevan et al., 2009; Zhang and Xu, 2021). This is a supervised machine learning approach for solving regression and probabilistic categorization issues in general. The present study examined the association between Cd levels and ancillary datasets, such as spectral indices and terrain attributes. Wang et al. (2020) credit the usefulness of GPR to its accessibility and high precision. Furthermore, GPR can aid in reducing dataset overfitting (Ballabio et al., 2019). The model was developed using Rstudio and applying the method "gaussprLinear," as well as libraries or packages (kernlab), and no tuning parameters were used.

2.5. Partial least square regression

The advantage of the PLSR algorithm is that it eliminates the problem of many features dimensionality among the predictor variables

(Mishra and Nikzad-Langerodi, 2020). The algorithm was applied in the present research to operate and evaluate independently for each number of features, ranging between 1 and 10. The algorithm simultaneously discovers a linear regression model connecting the predictor variables and the relationship between the explanatory variables in the new projected space after projecting the explanatory variables to an original space (Gamon et al., 1992). For more details on the PLSR algorithm, refer to Ehsani et al., (1999). The model was developed using Rstudio and applying the method "xgbTree," as well as libraries or packages that included the "xgboost" and "gamma" tuning parameters.

2.6. Extreme gradient boosting

Extreme gradient boosting (EGB) is a decision tree with a gradient-boosted approach that is improved for speed and precision (Climent et al., 2019). It is a sort of regression and characterization issue that is executed successively by an ensemble of restricted prediction algorithms, with each new design aiming to correct the flaws of the previous model. The EGB is based on Friedman's original gradient boosting approach (Climent et al., 2019), which is a practical and modular implementation of Friedman's gradient boosting framework. The model was developed using Rstudio and applying the method "xgbTree," as well as libraries or packages with the "xgboost" and "gamma" tuning parameters.

2.7. Multivariate adaptive regression splines

Friedman (1991) introduced MARS as a nuanced approach for organizing synergistic or interactive with minimal variable links among a collection of input factors and the target dependent. It is a nonparametric quantitative technique that uses a partitioning approach to split training datasets into simple linear segments (splines) with different gradients (slope). MARS assumes no hypotheses about the fundamental correlations of the dependent and independent factors (Zhang and Goh, 2016). The splines are commonly linked smoothly, and the piecewise polynomials, also referred to as basic functions (BFs), producing a comprehensive framework that can accommodate both linear and nonlinear behavior (Zhang and Goh, 2016). For more details on the MARS algorithm, refer to Friedman, (1991) and Zhang and Goh (2016). The model was constructed using Rstudio and applying the method "earth," as well as libraries or packages "earth" and tuning parameters "nprune".

2.8. Regularized random forest

Random forest which has been regularized (RRF), is the most recent variation of random forest (RF), which incorporates a regularization structure into the tree-increasing process (Deng and Runger, 212). RRF provides large feature subsets and decreases the number of features utilized in categorization and regression tasks, as shown in (Deng and Runger, 2012). Regularization usually entails applying a penalty to an error function to prevent overfitting. The model was built using Rstudio and applying the method " RRF," as well as "RRF" libraries and tuning parameters "mtry".

2.9. Bayesian regularized neural network

The best approach for dealing with learning problems is to use Bayesian approaches, and any other approach that does not resemble them should perform worse on average. They are especially useful for data model comparative studies because they embody automatically and quantitatively (Gauch et al., 2003). A Bayesian technique is a complex technique that is inherently self-punishing, according to Bayes' Rule. Tchagang and Valdés, (2019) proposed that Bayesian techniques complement neural networks (NNs) by overcoming an overly flexible network's inclination to explore almost nonexistent or excessively complicated data models. The Bayesian method for NN modelling techniques analyses all probable values of network parameters weighted by the likelihood of each set of weights. Traditional backpropagation NN training methods use a single set of variables (weights, biases, etc.). Using the Bayesian regularized neural network approach, Bayesian inference is used to construct the posterior probability distribution of weights, which are linked to the attributes of a prior probability distribution based on updates provided by the training set (Tchagang and Valdés, 2019). The model was generated using Rstudio and applying the "brnn" method, as well as libraries or packages "brnn" and "neurons" tuning parameters.

2.10. Bayesian generalized linear model (BGLM)

The Bayesian statistical approach infers generalized linear models (GLMs) with variables in a contained environment of common interest (e.g., in monotonic or convex regression), but establishing a legitimate posterior distribution supported by a system of linear constraints can be difficult, especially when some constraints are valid and enforceable, resulting in a reduced feature subspace. By sampling from posterior probabilities multiple times, Bayesian techniques obviate the necessity for a nonlinear solution. The versatility of the Bayesian technique for comprehensive evaluation of the ambiguity in the calculated random impacts and functionalities of hyperparameters is an additional feature. Bayesian inference is based on the collected data rather than on the presumption of limitless data populations. Bayesian approaches benefit from these inferences because all inferences are precise and not approximated, and the outcomes are comprehensible (Congdon, 2007; Ntzoufras, 2011). The model was developed using Rstudio and applying the "bayesglm" method, as well as libraries or packages "arm" and no tuning parameters.

2.11. M5 tree model

Quinlan (1992) created model trees as a type of regression tree that connects leaves to multivariate regression models. Model trees are a method of interacting with continuous class complications that provides conceptual recognition of the information as well as a nonlinear regression sit comfortably of the class (Etemad-Shahidi and Mahjoobi, 2009). They have a traditional decision tree structure, but instead of just discrete different classifiers, the leaves use linear functions. Quinlan (1992) developed M5 model trees, which were then recreated and enhanced in a framework named M5′ by Wang and Witten, (1996). An M5′ tree model, like regression trees, is an effective learning approach for determining real values that works well with large datasets. The M5′ tree model algorithm commences by iteratively partitioning the instance space to construct a regression tree, and the spliced principle is applied to reduce intrasubject variance in values bottom from the root via the subsidiary to the node (Etemad-Shahidi and Mahjoobi, 2009). The variance is determined by computing the standard deviation of the values that expanded from the root through the division to the node and evaluating the forecasted reduction in discrepancy as a direct consequence of analyzing each component at that node (Etemad-Shahidi and Mahjoobi, 2009). The model was created using Rstudio and applying the method "M5″, as well as the libraries or packages "RWeka: and tuning parameters "prune".

2.12. Environmental covariates

The use of geological spectral indices and terrain attributes as auxiliary datasets in the prediction of Cd in agricultural soil was chosen because of the impact and influence they have on the spatial distribution of PTEs in agricultural soil. For instance, according to Ding et al. (2017) terrain attributes such as slope and elevation have impact on the distribution of PTEs in the soil. Furthermore, PTEs are more effectively adsorbed in clay particles as the clay particle content increases, and the

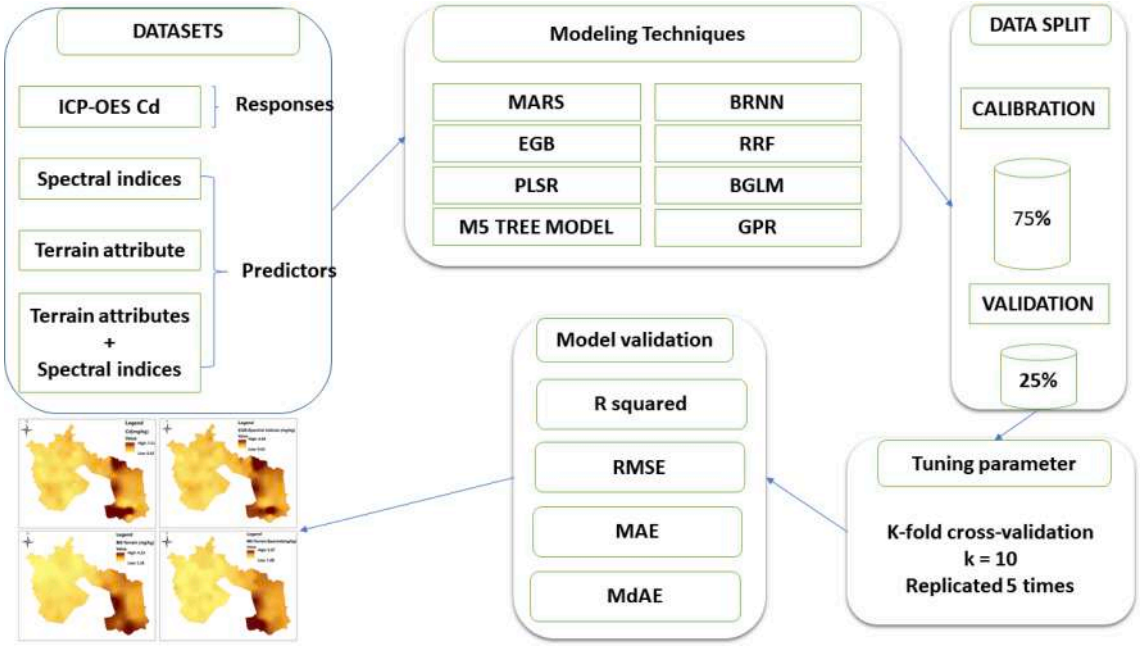


Fig. 2. The flow chart of the study.

elevated the clay mineral composition, the relatively easy it is to adsorb in soil.

In modelling Cd, terrain attributes were obtained with disparate sets of terrain derivatives that were sourced. The covariates were extracted from ASTER datasets utilizing a digital elevation model (DEM) with a spatial resolution of 30 m (https://doi.org/10.5067/ASTER/AST L1A.003) and treated with the SAGA-GIS terrain analysis toolbox. However, the 30 m spatial resolution processed DEM obtained was resampled using bilinear resampling method in ArcGIS to 10 m spatial resolution. The terrain attributes applied are elevation, slope, LS-factor, channel network base level, channel network drainage and relative slope position. The chosen terrain attributes are due to the relationship it has with Cd.

Sentinel 2 images of the study area were extracted from the European Satellite Agency (https://www.copernicus.eu/en/copernicus-services/emergency) in August 2020, and the bands were processed using SNAP software. The bands were at different resolutions of 10 m, 20 m, and 60 m. Bands were resampled into 10 m pixels using SNAP software to ensure that all the data were harmonized and have a uniform resolution. The spectral indices, such as clay mineral ratio (CLAYMR), ferrous mineral ratio (FMR), iron oxide ratio (IOR), carbonate normalized ratio (CNR), rock outcrop normalized ratio (RONR) and normalized difference built-up index (NDBI), were estimated using the bands required for its computation. The formulas of the spectral indices are given as:

$$CLAYMR = \frac{SWIR1}{SWIR2} \tag{1}$$

$$\mathrm{FMR} = \frac{SWIR}{NIR} \tag{2}$$

$$\mathrm{IOR} = \frac{RED}{BLUE} \tag{3}$$

$$CNR = \frac{RED - GREEN}{RED + GREEN} \tag{4}$$

$$\mathrm{RONR} = \frac{SWIR1 - GREEN}{SWIR2 + GREEN} \tag{5}$$

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR} \tag{6}$$

2.13. Model validation and accuracy assessment

Validation and accuracy assessment of the DSM modelling approaches employed in this study were performed using the coefficient of determination (R^2), root mean square error (RSME), mean absolute error (MAE) and median absolute error (MdAE). The regression model expresses R^2, which represents the variance of the proportion in the response. The RMSE determines the magnitude of the variations within the independent quantification categorize of the model predictive performance, whereas MdAE and MAE confirm the true quantifiable value. The study applied the Lin Concordance correlation coefficient (CCC) to determine the goodness of fit of the modeling approach measured (Lawrence and Lin, 1989). According to Viscarra Rossel et al., (2014), the CCC precise scale ranges from −1 to + 1, with 0.9 or greater denoting perfect agreement, 0.8 to 0.9 denoting substantial agreement, 0.65 to 0.8 denoting moderate agreement, and 0.65 denoting poor agreement. To determine the best model using the validation metrics, the R^2 and the CCC value should be close to 1, meaning the better the accuracy of the model. Similarly, the closer the RMSE, MdAE, and MAE to zero, the better the accuracy.

2.14. Data analysis

R studio was used to analyze the predictive performance of the modelling approaches. ArcGIS version 10.8 was used to create the optimal prediction as well as the spatial distribution maps of Cd. The spatial prediction intensity analysis was performed using the inverse distance weighting interpolation (IDW) technique. IDW estimates the values of the unknown situation within the sampling space, employing a linear confluence of values and assigns weights utilizing its inverse feature. IDW has a relatively lower marginal error than other interpolation techniques due to its capacity to allocate weights prior to prediction, making it more appropriate for generating precise spatial distribution maps. The flow chart of the study is presented in Fig. 2.

Table 1
Shows the statistical description of the PTE and environmental covariates.

	Median	Mean	SD	C V	Skewness	Kurtosis	Minimum	Maximum
Cd (N-115-mg/kg)	1.61	1.84	1.01	55.10	2.84	10.45	0.61	7.28
Dataset for modeling								
Training (N-85-mg/kg)	1.61	1.89	1.11	0.59	2.82	9.18	0.78	7.28
Test (N-30-mg/kg)	1.51	1.67	0.66	0.39	0.59	−0.26	0.61	3.21
Environmental covariates								
DEM	361.59	378.36	93.55	24.70	1.95	7.73	240.33	902.11
Slope	0.07	0.09	0.08	92.80	2.40	7.84	0.00	0.49
LS-Factor	0.79	1.26	1.65	130.60	4.10	24.00	0.01	13.08
CNBL	353.67	363.79	78.44	21.60	0.86	0.52	244.03	623.01
CND	7.89	15.06	29.17	193.70	6.83	59.49	0.00	279.10
RSP	0.04	0.07	0.11	163.70	4.89	34.39	0.00	0.92
clayMR	1.95	1.86	0.32	17.30	−0.82	−0.04	1.00	2.33
FMR	0.59	0.69	0.24	34.80	1.06	0.17	0.40	1.46
IOR	1.10	1.14	0.22	19.70	0.77	−0.21	0.82	1.78
RONR	0.78	0.73	0.18	24.20	−1.86	4.69	0.05	1.10
CNR	0.64	0.72	0.23	31.60	0.72	−0.68	0.41	1.29
NDBI	−0.27	−0.22	0.15	−65.00	0.75	−0.53	−0.45	0.17

3. Results and discussion

3.1. Data description

The summary descriptive statistics of the dataset used in this study are shown in Table 1. The maximum values, minimum values, median, mean, standard deviations (SD), coefficient of variations (CV), skewness, and kurtosis for the entire datasets, training dataset, test dataset, and environmental covariates are shown in table 1. The concentration of Cd varied from 0.61 (minimum value) to 7.28 mg/kg (maximum value). The coefficient of variation of Cd was estimated to be 55.10%. According to Wilding (1985) the criteria of the coefficient of variation can be categorized into high (CV greater than 35%), moderate (CV 15–35%) and lowest (CV less than 15%) variable classes. This result suggested that the CV of Cd in the study area is high, and the homogeneous distribution of Cd in the study area and its pollution source might be attributed to a local enrichment source (Agyeman et al., 2021c). The mean concentration of Cd (1.84 mg/kg) in the study area is relatively high compared to the local background value of Cd (0.2 mg/kg) reported by Nemecek (1992). It was 9.2 times higher than the local background values. In comparison, the mean concentration values of Cd to the national background value according to Czech decree No.152/2016 Coll. for agricultural soil (0.5 mg/kg) indicated that the national background value is 3.68 less than the reported mean concentration value of Cd. Similarly, the current study's mean concentration values of Cd in agricultural soil were found to be higher than the mean concentration values of Cd in agricultural soil in the Silesia region of Poland reported by, Piekut et al. (2018) {Bielski County (0.63), Czestochowa (1.06), Gliwice (0.50), Jastrzebie-Zdroj (0.50), Mikolow County (0.98), Myszkow County (0.99), Rybnik (0.50), Tychy (0.50), Wodzislaw County (0.50), Zabrze (1.63), Zory (0.50)}. The world average value (0.41 mg/kg), European average value (0.28 mg/kg) and upper continental crustal (0.10 mg/kg) level of Cd reported by Kabata -Pendias, (2011) were found to be lower than the current Cd mean concentration levels in the present study. The estimated skewness and kurtosis values were above 1 (see Table 1), which implies that the distribution of Cd is irregular, skewed in the right direction and leptokurtic based on Chandrasekaran et al., (2015) categorized description of the data distribution. The standard deviation (SD) of Cd was 1.01, implying a high level of heterogeneity due to the high concentration level of Cd. The statistical description of the environmental covariates mean values range from −0.22 to 378.36, median −0.27 to 3.61.59, SD 0.08 to 93.55, CV −0.27 to 193.70, skewness −0.82 to 6.83, kurtosis –0.04 to 59.49, maximum value of 0.19 to 902.11 and the minimum values −0.45 to 244.03.

Clay mineral ratio (CLAYMR), ferrous mineral ratio (FMR), iron oxide ratio (IOR), carbonate normalized ratio (CNR), rock outcrop normalized ratio (RONR) and normalized difference built-up index (NDBI), channel network base level (CNBL), channel network drainage (CND) and relative slope position (RSP), digital elevation model (DEM).

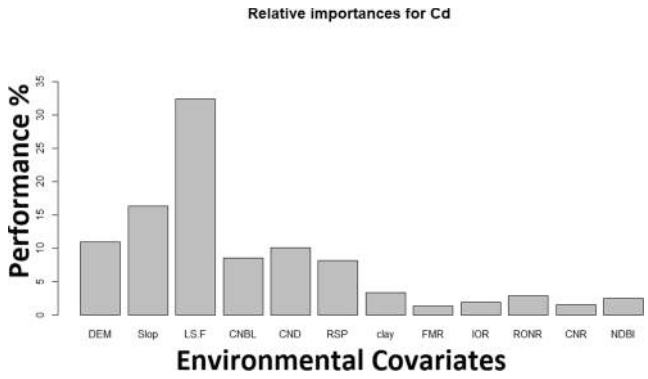


Fig. 3. Showing relative importance between Cd and environmental covariates.

3.2. Relative importance for Cd and environmental covariates

The chosen environmental covariates relative importance for the prediction of Cd concentration in the agricultural soil are presented in Fig. 3. The environmental covariates obtained for the prediction of Cd in the soil displayed a diverse degree of relative importance in the association with Cd based on the weights. However, based on the results, it was evident that the most important 6 environmental covariates that exhibited superior performance in the prediction of Cd in the agricultural soils are L.S. Factor, slope, DEM, CND, CNBL, and RSP. The weighted performances for these environmental covariates were 32.37%, 16.38%, 10.99%, 10.11%, 8.53%, and 8.17% for L.S. Factor, slope, DEM, CND, CNBL, and RSP, respectively. These 6 environmental covariates are all the covariates extracted from the terrain analysis. On the other hand, the environmental covariate spectral indices estimated from Sentinel 2 satellite imagery exhibited minimal performance with the weighted values 3.33%, 2.85%, 2.52%, 1.91%, 1.52% and 1.31% for clayMR, RONR, NDBI, IOR, CNR and FMR, respectively. The study area is distinguished by highlands and lowlands; the effect of slope length on erosion; and the steep slope factor, which most likely has an effect on slope steepness. This factor influenced the L.S factor, which was the most important covariate in the prediction of Cd in the agricultural soil. The terrain attributes were more relevant in the prediction of Cd in the agricultural soil, which is practically true for the terrain's attributes based on the geomorphology of the study area. Taghizadeh-Mehrjardi et al. (2020) applied terrain attributes and remote sensing datasets to the

Table 2
Assessment of modeling methods on Cd prediction using terrain attributes.

Modeling techniques	R^2	RMSE	MAE	MdAE	CCC
GPR	0.70	0.53	0.45	0.44	0.61
PLSR	0.71	0.51	0.42	0.32	0.64
EGB	0.47	0.83	0.52	0.36	0.47
MARS	0.68	1.29	0.66	0.21	0.41
RRF	0.61	0.80	0.55	0.32	0.46
BRNN	0.73	0.48	0.40	0.35	0.68
BGLM	0.69	0.54	0.46	0.46	0.60
M5	**0.77**	**0.45**	**0.37**	**0.35**	**0.73**

GPR-gaussian process regression, PLSR-partial least square regression, EGB-extreme gradient boosting, MARS- Multivariate adaptive regression splines, RRF-regularized random forest, BRNN- Bayesian regularized neural network, BGLM- Bayesian generalized linear model, M5-tree model.

prediction of soil organic carbon (SOC) in the soil, and terrain attributes such as slope and DEM were shown to be more relevant for the SOC content prediction in soil than remote sensing-based covariates. Besides, the impact of terrain attributes on predicting Cd content can be closely linked to variability in other soil physical properties as well as the presence of agricultural production in lower hilly terrain.

Geological terrain is an essential influential factor for the prediction of PTEs such as Cd in the soil. Over extended durations of time, relationships between bedrock, climatic conditions, and geomorphic mechanisms result in the development of soil parent composites. Environmental covariates have the greatest influence on the impactful categorization of the spatial variability of PTEs in soil, depending on the circumstance pedogenesis and the evolution development (Zeraatpisheh et al., 2020). The authors also suggested that the use of machine learning algorithms in the prediction of PTEs or soil properties in soil should consider the soil formation mechanism (Zeraatpisheh et al., 2020), which includes the mineral composition as well as the geological characteristics of the soil being studied. The enrichment of Cd in agricultural soils is due to a combination of anthropogenic and natural processes, such as parent material weathering and subsequent pedogenesis. As a result, the use of auxiliary datasets that are in tune with pedogenesis factors such as parent material (by employing a remote sensing - based spectral dataset to distinguish geochemical correlates of parent material, e.g., ferrous mineral ratio) in conjunction with a machine learning algorithm aid in predicting PTE (Cd) concentrations in soil. For example, Zhang et al. (2020) used machine learning algorithms (e.g., ANN, SVM) to spatially predict the concentration of PTEs (e.g., Cd) in urban soil by coupling soil parent materials such as Fe_2O_3, Al_2O_3 as auxiliary datasets to aid in the identification of the pollution source of the PTEs while increasing prediction efficiency. Wu et al. (2016), on the other hand, used machine learning algorithms such as ANN and GPR combined with soil formation factors (parent material, topography) and reported that machine-learning techniques adept of remedying non-linear problems, such as ANN and GPR, could also be used to develop soil background estimation models. Even though PTE pollution is largely determined by the concentration and transformation of PTEs in soils of various genesis, the determination of the spatial variability of PTEs in soil is dependent on the determination of the appropriate soil genesis factor that is linked to the PTEs under investigation.

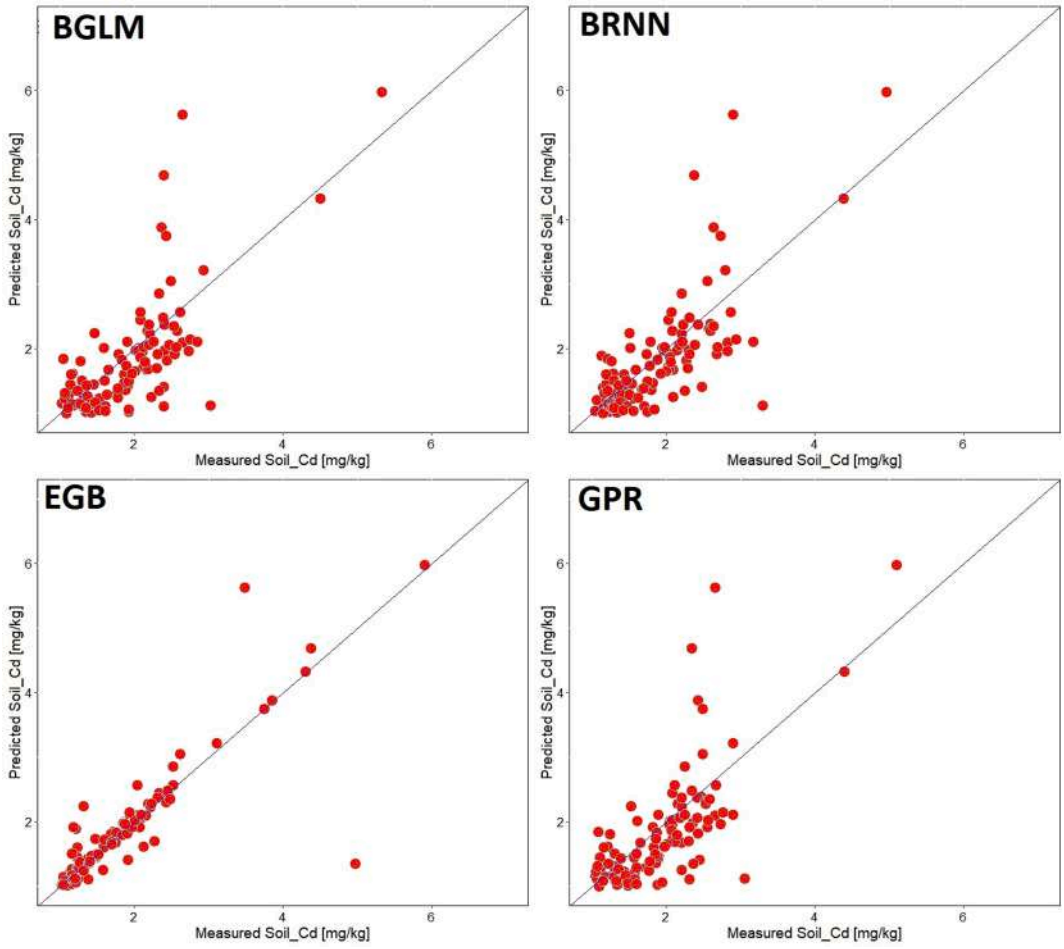


Fig. 4A. Cadmium content (mg/kg) values measured vs predicted in scenario 1 for each of the four model optimal fitness curves (BRNN- Bayesian regularized neural network, BGLM- Bayesian generalized linear model, EGB- extreme gradient boosting, GPR-gaussian process regression).

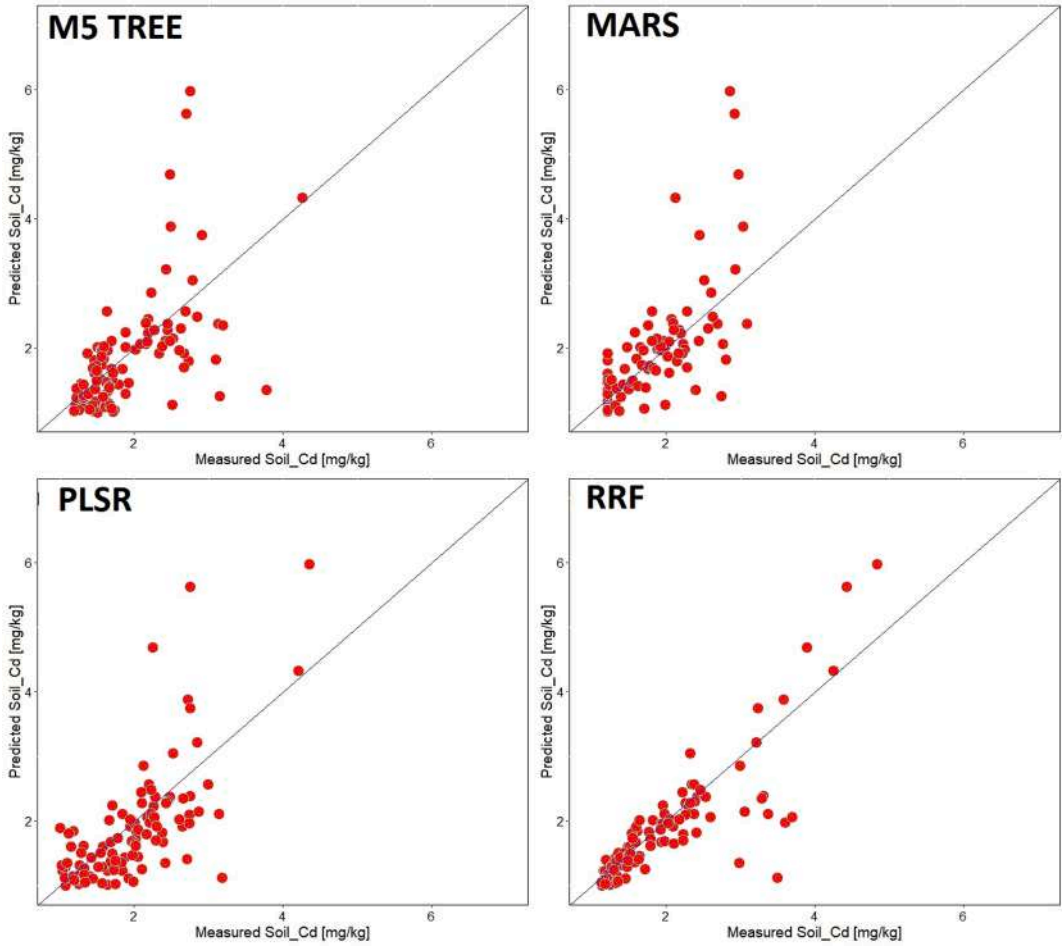


Fig. 4B. Cadmium content (mg/kg) values measured vs predicted in scenario 1 for each of the four model optimal fitness curves (M5 tree model, MARS- Multivariate adaptive regression splines, PLSR-partial least square regression, RRF-regularized random forest).

3.3. Prediction based on terrain attributes (Scenario 1)

Table 2 displays the performance of the various modelling approaches used for the prediction of Cd in agricultural soil, while the fitness curves for the predicted and measured Cd are presented in Figs. 4A and 4B. Eight modelling approaches were used, employing terrain attributes as the auxiliary datasets coupled with Cd data measured using ICP–OES. MARS had the least MdAE (0.21), followed by RRF (0.32), PLSR (0.32), BRNN (0.35), M5 tree model (0.35), EGB (0.36), and GPR and BGLM with 0.44 and 0.46, respectively. The M5 tree modelling approach yielded the lowest RSME value (0.45), followed by BRNN (0.48), PLSR (0.51), GPR (0.53), BGLM (0.54), RRF (0.80), EGB (0.83), and MARS (1.29). The estimated MAE results also revealed that M5 tree model had the lowest MAE (0.37), followed by BRNN having the second lowest MAE (0.40), PLSR (0.42), GPR (0.45), BGLM (0.46), EGB (0.52), RRF (0.55), and MARS (0.66). The closer the RMSE, MAE and MdAE values are to zero, the higher the precision and the more accurate the model is in predicting Cd in agricultural soil. The R^2 values revealed that, out of the 8 modelling approaches used to predict the concentration of Cd in the soil, M5 tree model obtained the highest value of $R^2 = 0.77$, which is a good prediction based on Li et al., (2016) precision and model assessment criteria. The R^2 values of the other modelling approaches were within the acceptable precision and accuracy range, i.e., 0.73 for BRNN, 0.71 for PLSR, 0.70 for GPR, 0.69 for BLGM, 0.68 for MARS and 0.61 for RRF. Only the EGB modelling approach performed poorly, exhibiting $R^2 = 0.47$, which is unacceptable. With the exception of the R^2 value obtained for EGB, the differences in the R^2 values of the modeling approaches in predicting Cd were relatively close and within the acceptable modeling assessment range. However, the differences in the computed modeling errors for each of the modelling approaches used were also relatively small, which did not exceed 1.3 for RMSE, 0.7 for MAE and 0.5 for MdAE. Based on the CCC assessment criteria, the Cd predictions from the modeling approaches ranged between 0.41 and 0.73. The M5 tree model, on the other hand, had the highest CCC for predicting Cd in agricultural soil, while MARS had the lowest. The R2 (0.77) of the M5 tree model, on the other hand, was the highest of all the techniques considered, with the CCC (0.73) providing the best 1:1 fit between measured and predicted Cd.

The cumulative results indicated that the combination of the M5 tree modeling approach coupled with terrain attributes and the measured Cd concentration was the optimal modeling approach that predicted Cd in the soil with higher prediction efficiency and minimal error. Neissi et al., (2020) applied M5 tree model and geographical information system in the spatial interpolation of the sodium absorption ratio and concluded that the performance of the M5 tree modelling combined with inverse distance weighting and kriging interpolation methods produced the best results. Kumar and Deswal (2020) assessed the capacity of the M5 tree model to predict phosphorus removal which yielded acceptable results ($R^2 = 0.987$, RMSE = 0.033 and MAE = 0.0258). The M5 tree model has been applied in various disciplines and proven to have the capacity to yield a high accuracy level with minimal error, as in the present study. Kumar and Deswal (2020), Sui et al., (2016), Heddam, (2021) and Sihag et al., (2019) assessed the performance of diverse modelling approaches for the evaluation of PTEs in soil and found that the M5 tree modelling approach was the optimal model for determinations of Cu and Zn. Similarly, Biabani et al., (2016) and Rahimikhoob, (2016) applied the

Table 3
Assessment of the performance of various modelling approaches using spectral indices.

Modeling techniques	R^2	RMSE	MAE	MdAE	CCC
GPR	0.71	0.52	0.46	0.51	0.65
PLSR	0.74	0.51	0.45	0.46	0.67
EGB	**0.83**	**0.54**	**0.33**	**0.23**	**0.76**
MARS	0.61	1.38	0.70	0.18	0.30
RRF	0.70	0.56	0.38	0.27	0.60
BRNN	0.75	0.50	0.44	0.47	0.68
BGLM	0.69	0.54	0.48	0.50	0.63
M5	0.76	0.49	0.42	0.44	0.68

GPR-gaussian process regression, PLSR-partial least square regression, EGB-extreme gradient boosting, MARS- Multivariate adaptive regression splines, RRF-regularized random forest, BRNN- Bayesian regularized neural network, BGLM- Bayesian generalized linear model, M5-tree model.

M5 tree model algorithm for assessment of the daily reference of evapotranspiration and the prediction temporal evolution of clear water and found that the M5 tree model method produced satisfactory results considering the performance indicators, such as R^2, RMSE and MAE, with less aberration from the arithmetical values.

3.4. Prediction based on spectral indices (Scenario 2)

Table 3 summarizes the performance of the modelling approaches used to predict Cd in agricultural soil using eight different models coupled with spectral indices as the auxiliary dataset, whereas the fitness curves between the measured dataset and the predicted dataset are shown in Figs. 5A and 5B. According to the estimated validation and accuracy results, M5 tree model had the lowest RSME value of 0.49. BRNN, was the second algorithm that obtained a minimal RSME (0.50) accompanied by PLSR (0.51), GPR (0.52), EGB (0.54), BGLM (0.54), RRF (0.56) and MARS (1.38). Except for MARS, which produced a high RMSE value in comparison to the others, the difference in the estimated RMSE was relatively close. The performance of algorithms in terms of MAE is in following order: EGB (0.33) > RRF (0.38) > M5 Tree (0.42) > BRNN (0.44) > PLSR (0.45) > GPR (0.46) > BGLM (0.48) > MARS (0.70). The MAE values of all modelling approaches were relatively close. Conversely, the obtained MdAE values ranged from 0.18 to 0.51. Only two of the modeling approaches' R^2 values were within the good performance zone, according to the results of their R^2 values (i.e., EGB $R^2 = 0.83$ and M5 tree model $R^2 = 0.76$). All the modelling approaches of the other 6 algorithms yielded acceptable results, and the R^2 values ranged from $R^2 = 0.61$ to 0.75. The CCC values of the modeling approaches ranged between 0.30 and 0.76, with MARS obtaining the lowest CCC value and EGB obtaining the highest CCC value. The R2 value of 0.83 obtained by the EGB modeling approach was nevertheless revealed to be the highest of all the techniques deemed, with a corresponding CCC value of 0.76 providing the optimal 1:1 fit between measured and predicted Cd. The precision and accuracy level of the modelling approaches in the prediction of Cd in the agricultural soil was relatively moderate to high based on the R^2, CCC, RMSE, MAE, and MdAE results obtained. Based on the cumulative performance of the modeling approaches in terms of the validation and accuracy assessment criteria, EGB was the best modeling approach for predicting Cd in

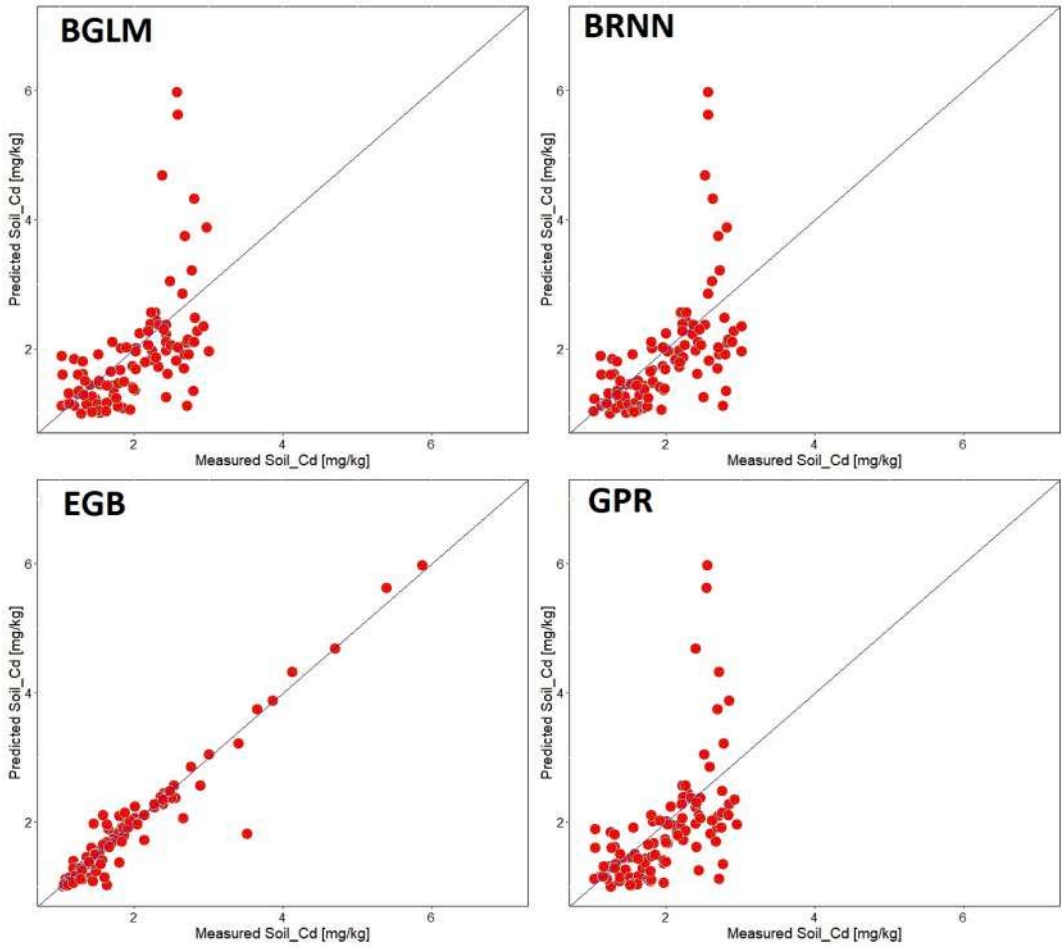


Fig. 5A. Cadmium content (mg/kg) values measured vs predicted in scenario 2 for each of the four model optimal fitness curves (BRNN- Bayesian regularized neural network, BGLM- Bayesian generalized linear model, EGB- extreme gradient boosting, GPR-gaussian process regression).

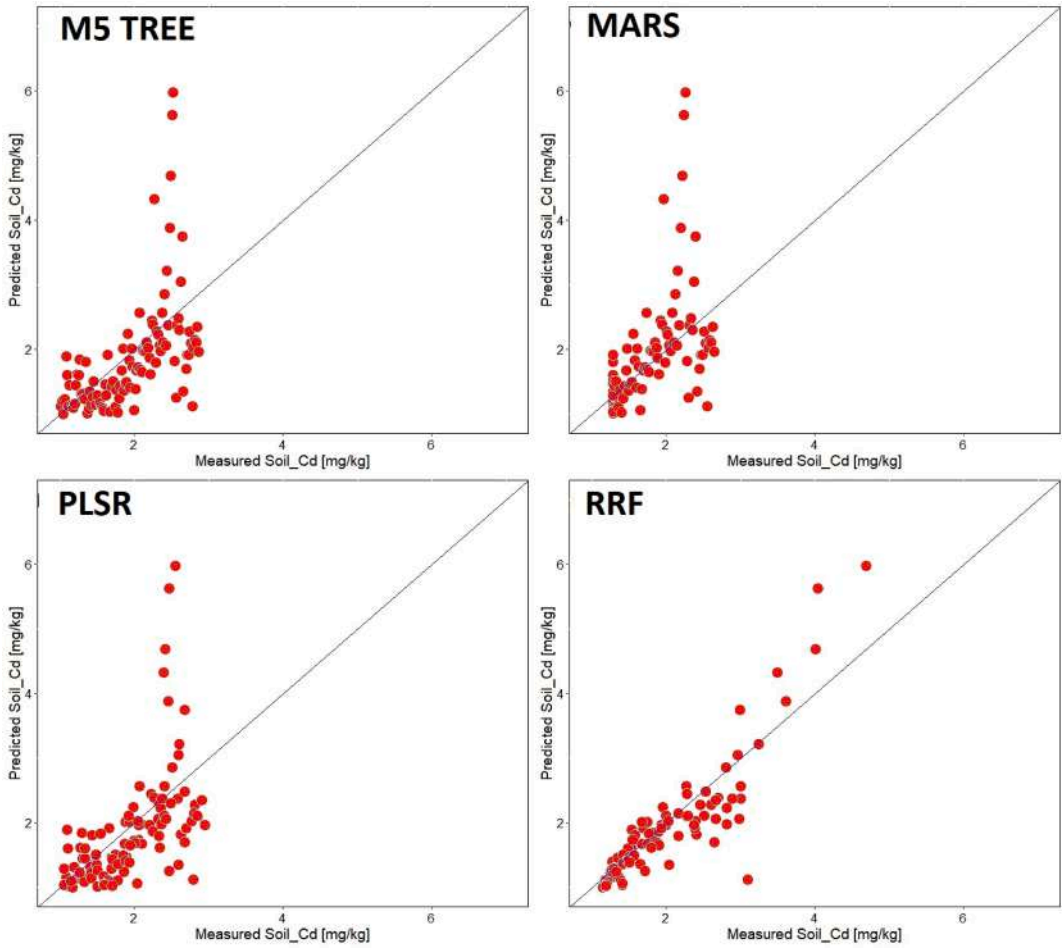


Fig. 5B. Cadmium content (mg/kg) values measured vs predicted in scenario 2 for each of the four model optimal fitness curves (M5 tree model, MARS- Multivariate adaptive regression splines, PLSR-partial least square regression, RRF-regularized random forest).

agricultural soil, with the highest CCC and R2 values of 0.76 and 0.83 and the lowest MAE of 0.33, respectively.

Goydaragh et al., (2021) applied EGB coupled with environmental variables as one of the modelling approaches in the prediction of soil organic carbon in the soil and yielded an R^2, RMSE, MAE, and MdAE values of 0.32, 0.72, 0.56, and 0.42 respectively. Comparing performance of EGB in this study to the results obtained in the above mentioned by, Goydaragh et al., (2021), it was evident that the EGB modeling technique in the current study performed extremely well in all validation and assessment criteria. EGB outperformed MLR, SVM, and RF models in predicting Mn removal in soil with the least error (RMSE = 1.4, MAE = 0.81) and the highest coefficient (R^2 = 0.88) (Bhagat et al., 2020). The R^2 result obtained by Bhagat et al. (2020) (R^2 = 0.88, RMSE = 1.4, MAE = 0.81) was higher than those R^2 value obtained in the current study (R^2 = 0.83, RMSE = 0.54, MAE = 0.33), but when the RMSE and MAE results from both studies are compared, it can be inferred that the estimated errors were 2.61 (RMSE) and 2.48 (MAE) times less in the current study. According to Ma et al., (2019), EGB has a significant advantage over other MLAs, such as ANN and SVM, in terms of selecting responsive features via relevance rankings and limiting model overfitting by defining the default orientation of splitting for missing datasets or values. Numerous research findings have shown that EGB can enhance predictive performance by incorporating key features in predicting metal ion concentrations, forest terrestrial biomass, and PTEs concentrations (Joharestani et al., 2019). Zhao et al., (2022) reported that comparing their studies to other studies, EGB has demonstrated reasonable model performance in estimating pollution indices using sensitive wavelengths. In the computation of PTEs concerntration in soil or sediments, EGB outperformed other MLA modelling approaches, such as RF, SVM, and ANN (Bhagat et al., 2021). However, EGB has the benefit of reducing underestimation and overestimation (Li et al., 2020). According to, Kim et al., (2015), EGB has a proclivity to filter out model performance, thereby minimizing potential limitations encountered in other modelling approaches, such as overfitting. Nonetheless, EGB can help to reduce modelling normalization issues, (Jia et al., 2019), the need for hyperparameter tuning, (Probst et al., 2019), local minima, (Kawaguchi and Bengio, 2019), higher irregularities, (Li et al., 2020), and strategies in technology transfer (Kim and Geum, 2020).

3.5. *Prediction based on terrain attributes and spectral indices (Scenario 3)*

The concentration of Cd was also predicted by combining the modelling approaches with terrain attributes and spectral indices, while the fitness curves for the measured dataset and predicted dataset are displayed in Figs. 6A and 6B. The performance of the modelling methods is presented in Table 4. Accordingly, R^2 values were within the acceptable prediction efficiency range (0.5 to 1), except for MARS (R^2 = 0.26), which exhibited abysmal performance. The M5 tree approach obtained the highest R^2 value of 0.84, accompanied by PLSR R^2 = 0.74, BRNN R^2 = 0.73, EGB R^2 = 0.67, GPR R^2 = 0.65, RRF R^2 = 0.62, and BGLM R^2 = 0.61. The computed RMSE values revealed that the M5 tree model yielded a minimal error value of 0.39, which is the preferred value because the smaller the error value is, the more efficient the modelling approach. Furthermore, BRNN was the next model that

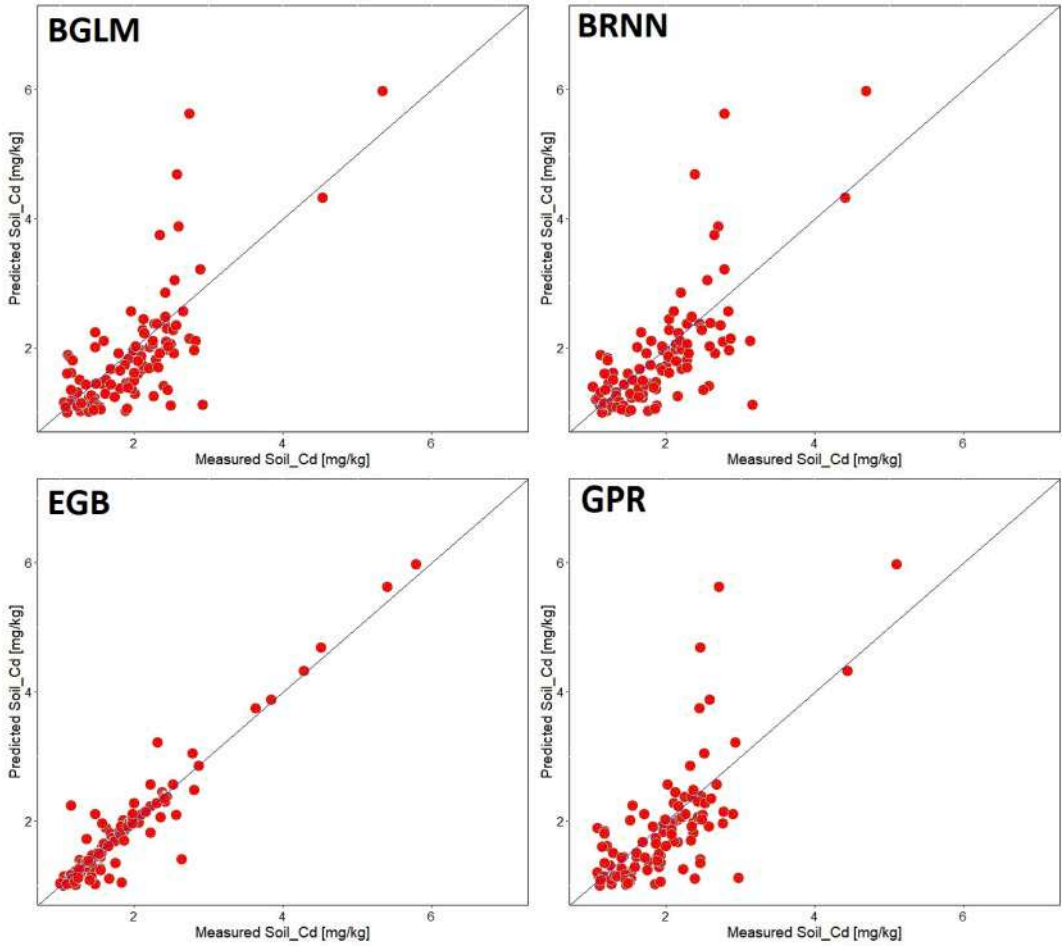


Fig. 6A. Cadmium content (mg/kg) values measured vs predicted in scenario 3 for each of the four model optimal fitness curves (BRNN- Bayesian regularized neural network, BGLM- Bayesian generalized linear model, EGB- extreme gradient boosting, GPR-gaussian process regression).

obtained the lowest RMSE value of 0.49, followed by PLSR (0.49), GPR (0.55), BGLM (0.58), EGB (0.68), RRF (0.76) and MARS (1.54). The performance of the modelling approaches with regards to MAE and MdAE ascends in this order: M5 tree > PLSR > BRNN > EGB > GPR > BGLM > RRF > MARS, for MAE and M5 tree > EGB > PLSR > RRF > MARS > BRNN > GPR > BGLM for MdAE. With the exception of MARS in RMSE and MAE, which had a slightly higher error margin, the margins between the estimated error criteria for the modeling approaches were relatively small. The estimated CCC values of the modeling approaches were within the range of 0.19 as the least for the MARS modeling approach and 0.81 as the highest for the M5 tree modeling approach. However, based on the R2 value of 0.84, the M5 tree model exhibited the highest of all the techniques deemed, with the CCC value of 0.81 providing the optimal 1:1 fit between measured and predicted Cd. The cumulative performance of the modelling approaches in the prediction of Cd in agricultural soil using spectral indices, terrain attributes, and MLA suggested that the M5 tree modelling approach is the optimal approach that predicts Cd with higher precision and a consistent minimal error margin.

Comparing modelling Scenario 1 (prediction based on terrain attributes) to modelling Scenario 3 (prediction based on terrain attributes and spectral indices), it is obvious that GPR, MARS, BRNN, and BGLM performed better using terrain attributes alone as the auxiliary datasets than their combination. However, the PLSR, EGB, RRF, and M5 tree models performed significantly better in Scenario 3 than in Scenario 1. On the other hand, comparing Scenarios 2 and 3, it was evident that GPR, EGB, MARS, RRF, and BGLM exhibited superior performance in Scenario 2 than in the respective modelling methods in Scenario 3. Contrariwise, the PLSR, BRNN, and M5 tree models performed better in Scenario 3 than the respective modelling techniques in Scenario 2. It can be inferred that the use of terrain attributes, spectral indices, and the combination of spectral indices and terrain attributes as auxiliary datasets has revealed the ability of the PLSR and M5 tree model approaches to consistently predict Cd and improve prediction efficiency in all scenarios with high efficiency and minimal error. Kalambukattu et al., (2018) reported that a combination of terrain attributes and spectral indices has the propensity to optimize results with good accuracy levels. In comparison to modelling Scenario 1, the R^2, RMSE, MAE, and MdAE values of the PLSR, EGB, RRF, and M5 tree models improved by a range of 0.50 to 17.60%, 1.98 to 9.46%, 2.05 to 10.55%, and 3.87 to 18.25%, respectively in Scenario 3, except for the MdAE of the RRF. Except for the R^2 of the BRNN modelling approach, the R^2, RMSE, MAE, and MdAE values of PLSR, BRNN, and the M5 tree model improved by 1.28 to 5.3%, 1.11 to 5.30%, 3.18 to 15.30%, and 12.12 to 28.12%, respectively in Scenario 3 than in Scenario 2. Several studies, such as, Goydaragh et al., (2021) and Xu et al., (2019), combined spectral datasets with environmental variables, such as terrain attributes, which improved the modelling results compared with using either spectral datasets or environmental variables.

3.6. Comparison of optimal models based on the modelling approaches

The three modelling Scenarios yielded great results, and the optimal modelling technique that predicted Cd in the agricultural soil for each modelling Scenarios was the M5 tree model for Scenario 1, the EGB for Scenario 2 and the M5 tree model for Scenario 3. Among the three

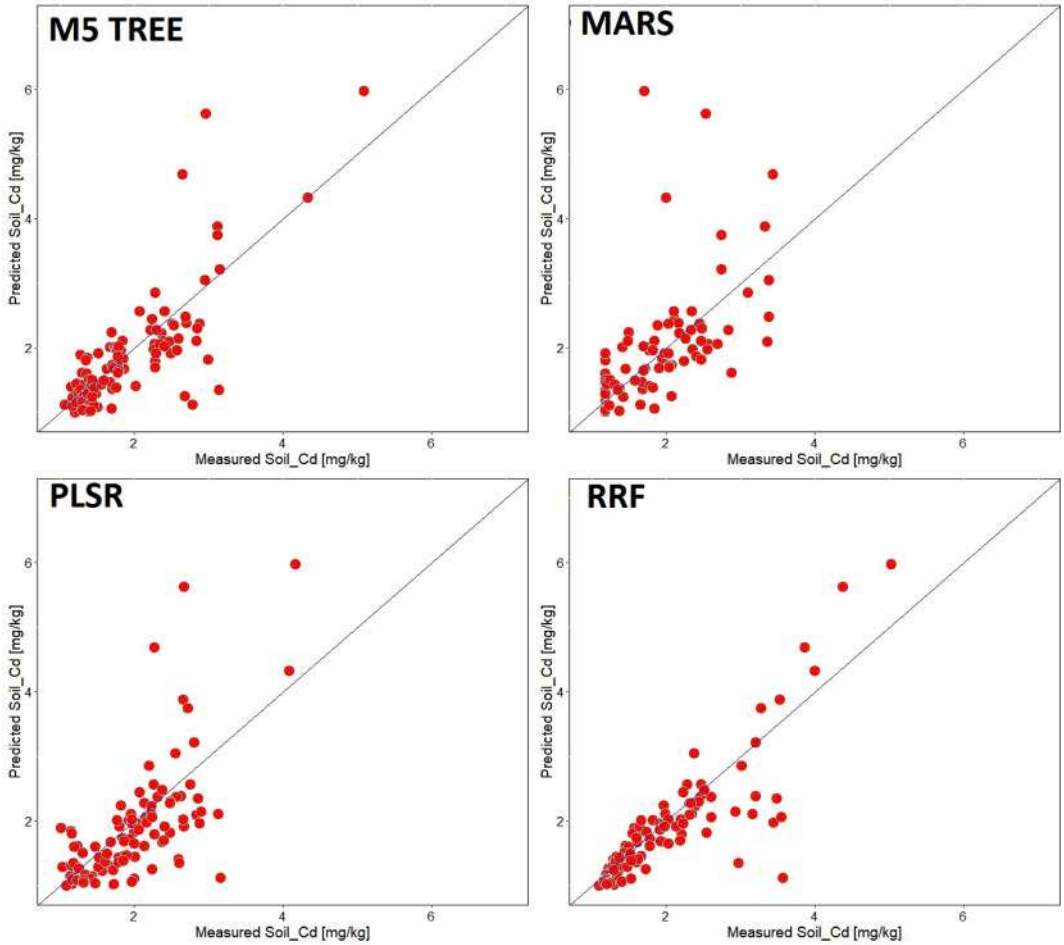


Fig. 6B. Cadmium content (mg/kg) values measured vs predicted in scenario 3 for each of the four model optimal fitness curves (M5 tree model, MARS- Multivariate adaptive regression splines, PLSR-partial least square regression, RRF-regularized random forest).

Table 4
Assessment of the performance of various modelling approaches using spectral indices and terrain attributes.

Modeling techniques	R^2	RMSE	MAE	MdAE	CCC
GPR	0.65	0.55	0.47	0.41	0.58
PLSR	0.74	0.49	0.41	0.30	0.68
EGB	0.67	0.68	0.42	0.26	0.64
MARS	0.26	1.54	0.85	0.35	0.19
RRF	0.62	0.76	0.52	0.32	0.48
BRNN	0.73	0.49	0.41	0.37	0.66
BGLM	0.61	0.58	0.50	0.46	0.55
M5	**0.84**	**0.39**	**0.31**	**0.24**	**0.81**

GPR-gaussian process regression, PLSR-partial least square regression, EGB-extreme gradient boosting, MARS- Multivariate adaptive regression splines, RRF-regularized random forest, BRNN- Bayesian regularized neural network, BGLM- Bayesian generalized linear model, M5-tree model.

techniques, the M5 tree model from Scenario 3 yielded the best results across all five accuracy and validation assessment approaches (R^2, RMSE, MAE, MdAE, CCC). Scenario 2's cumulative performance compared to Scenario 1 suggested that Scenario 2 produced better results than Scenario 1. When M5 tree model results for Scenario 3 are compared to M5 tree model results in Scenario 1, the M5 tree model in Scenario 3 improves the R^2 value by 4.6%. Similarly, the computed error in both scenarios (i.e., Scenario 1 and 3) shows that RMSE, MAE, and MdAE improved by 6.57%, 8.82%, and 18.25% in the M5 tree model in scenario 3 than in scenario 1. In comparison, the estimated CCC of the M5 tree model in scenario 1 (0.76) compared to the M5 tree model in scenario 3 (0.81) indicated an 2.87% improvement in performance, particularly in scenario 3. In contrast, comparing scenarios 3 and 2, it was clear that the R^2, RMSE and MAE improved by 0.95%, 15.64% and 2.51%, respectively, in favor of Scenario 3. Furthermore, when the computed CCC values for the optimal modeling approaches in the prediction of Cd in agricultural soil for Scenario 3 (0.81–M5 tree model) and 2 (0.71–EGB) were compared, the M5 tree model obtained CCC values improved by 5.24%. As a result, it can be extrapolated that the combination of terrain attributes and spectral indices coupled with the M5 tree modelling technique is an improvement over the application of terrain attributes or spectral indices in isolation with MLAs in the prediction of Cd in agricultural soil. The M5 tree model is made up of many tree structures built with subsets, and a tree configuration with the fewest error must be built so that can avoid overfitting (Kumar and Deswal, 2020). In comparison, the scatter plot between the measured and predicted for the optimal predictions for all approaches shows that the M5 tree model (Fig. 6B) had a better goodness of prediction than the other two optimal models (refer to Figs. 4B and 5A).

3.7. Spatial prediction intensity analysis

The spatial prediction intensity of the optimal modelling predictions per modelling approach and the spatial prediction of Cd in the agricultural soil are shown in Fig. 7A. The spatial prediction of Cd and the EGB-spectral index prediction map share the same similar hotspot pattern. The hotspot can be seen in the northwest to southwest in the clockwise direction. On the other hand, the spatial prediction of the M5 tree model-terrain and the M5 tree model-terrain spectral map distribution

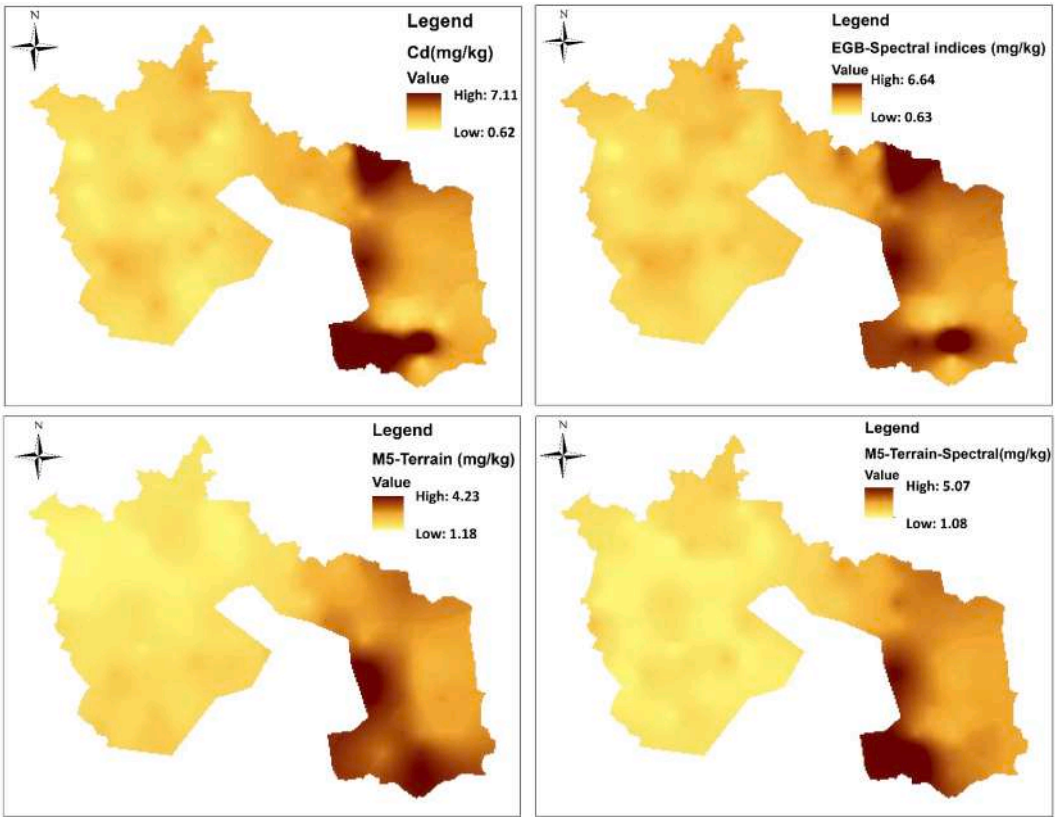


Fig. 7A. Cadmium spatial prediction and the best prediction techniques from the three modeling approaches [EGB- extreme gradient boosting, M5- M5 tree model].

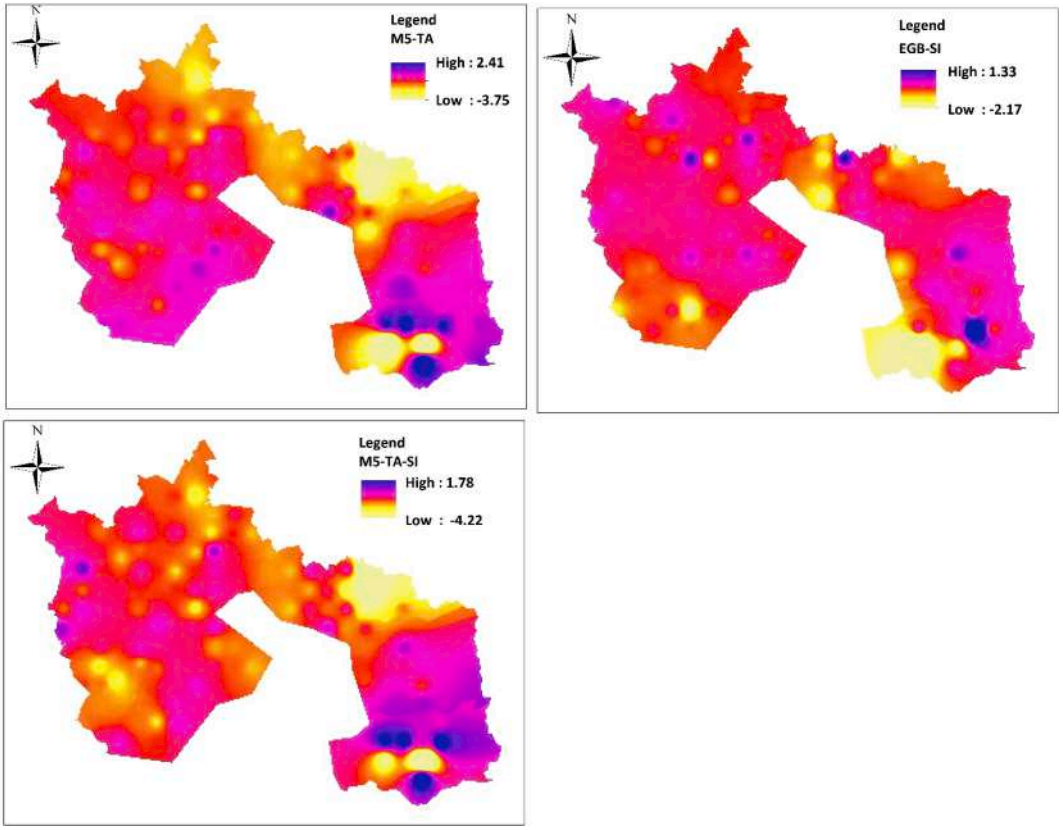


Fig. 7B. Spatial distribution variance between the observed Cd and predicted Cd for the optimal prediction in each scenario (M5-TA (M5 tree model-terrain attribute), EGB-SI (extreme gradient boosting-spectral indices) and M5-TA-SI (M5 tree model -terrain attributes-spectral indices)).

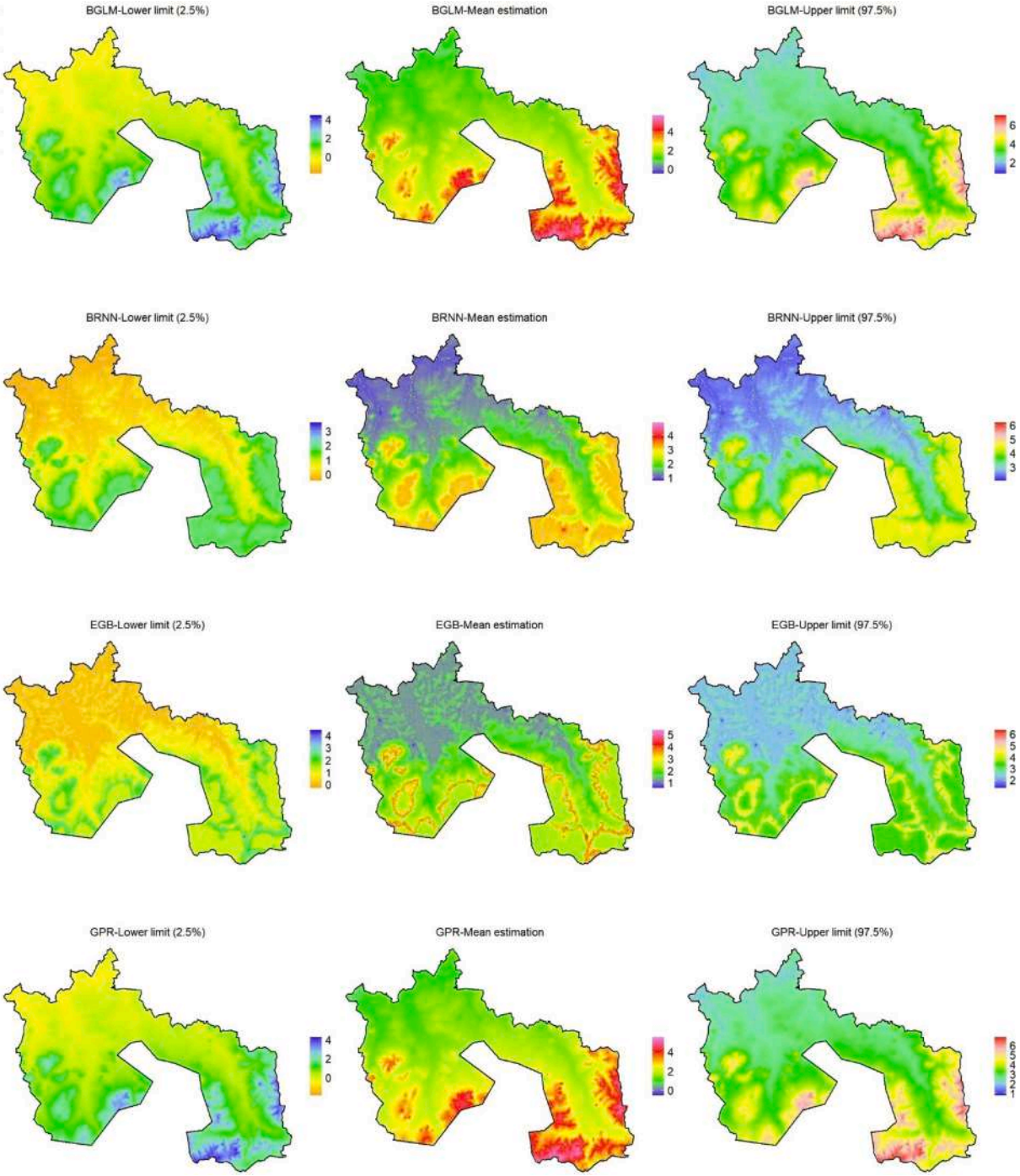


Fig. 8A. Maps showing the level of uncertainty at 2.5%, 97.5% and mean based on BGLM, BRNN, EGB and GPR modeling approaches for the prediction of Cd in the agricultural soil.

can be seen in the southwestern area of the map. The hotspot on the maps can be attributed to iron and steel production industries in the study area, phosphate fertilizer application on farmlands, particularly in the southeastern part, and atmospheric deposition of Cd on the soil. Hutton, (1983) suggested that iron and steel production and phosphate fertilizer application are the major anthropogenic sources of Cd in the soil. According to WHO (2010), the health effects of Cd include tubular renal dysfunction, acute pneumonitis with pulmonary oedema and lung cancer development. Cadmium poisoning can cause disruptions in the calcium metabolic rate and the creation of kidney stones, as well as softening of the bones and osteoporosis in those who are predisposed by working and living in cadmium-contaminated areas (WHO, 2010). The spatial prediction maps revealed that the M5-terrain and M5-terrain-spectral spatial distribution maps exhibited moderate prediction intensity patterns of 3.05 and 3.99 prediction intensities, respectively. Due to the active agricultural activities in that area, a high prediction intensity pattern was observed in the southeastern part. Cd and the EGB spatial prediction intensity map, on the other hand, revealed that their prediction intensity was very high, with prediction intensities of 6.49 and 6.01, respectively. The high prediction intensity in the Cd and the EGB spatial prediction intensity map is due to the metal and steel production and the agricultural activities (e.g., fertilizer application) in the northeastern and southeastern parts of the area. Men et al., (2019), on the other hand, argued that the smaller the extent, the more likely the

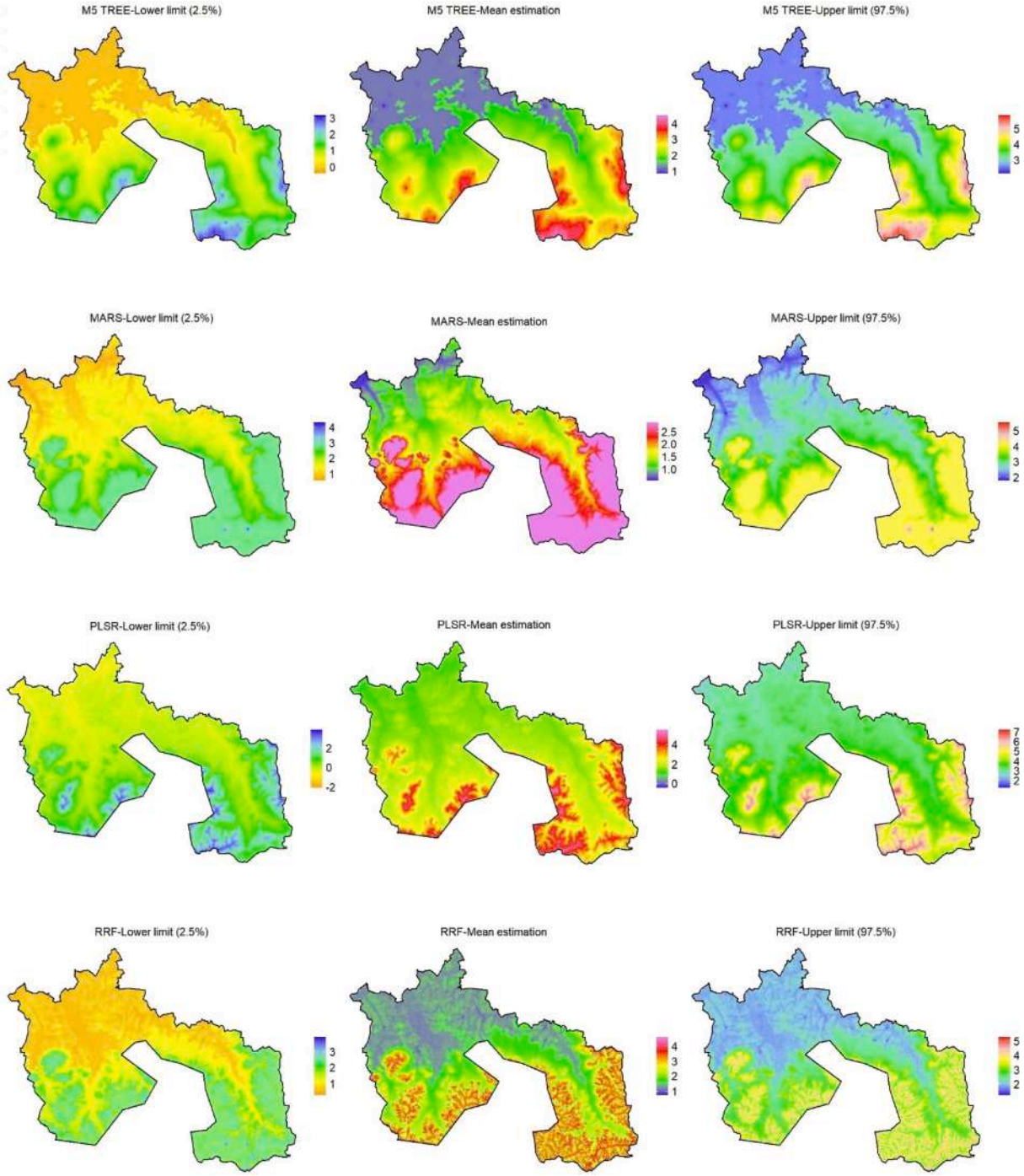


Fig. 8B. Maps showing the level of uncertainty at 2.5%, 97.5% and mean based on M5 tree, MARS, PLSR and RRF modeling approaches for the prediction of Cd in the agricultural soil.

sources are spatially homogeneous in nature. Similarly, the larger spatial intensity prediction implies that the impacts of myriad sources are primarily spatial.

Fig. 7B depicts the interpolation prediction variance between predicted and measured Cd values. The variance between the optimal Cd prediction and the observed Cd suggests that there is little over and under prediction, as shown by the spatial variance distribution maps for M5-TA, EGB-SI, and M5-TA-SI. The M5-TA and M5-TA-SI both displayed low spots prediction in the southeastern, northeastern, and northwest areas of the map. Similarly, EGB-SI and M5-TA-SI shared a low spot prediction in the southwestern area of the map, and EGB-SI exhibits a low spot prediction in the southwestern area of the map. The overprediction was most noticeable in the southeastern region of the spatial distribution variance map, which was more pronounced for the M5-TA and M5-TA-SI. The combination M5 tree model coupled with terrain attributes and spectral indices clearly revealed areas that were under predicted and were not revealed by using either covariate in the prediction of Cd in agricultural soil. It follows that both covariates complement each other, assisting in highlighting areas that were underpredicted by M5-TA and EGB-SI.

3.8. Uncertainty assessment

The efficiency of pollution assessment is dependent on effective and

precise mapping of PTEs in soil. However, in order to reduce the bias in pollution estimation induced by mapping techniques, it is critical to understand the uncertainty of soil PTE pollution evaluations initiated by mapping error as well as the differences in pollution analysis between multiple mapping techniques. The estimation of prediction uncertainty is critical in the tracking of ambiguity and out-of-time sample predictability. In this study, uncertainty was estimated for the 8 modeling algorithms used in the prediction of Cd in the agricultural soil. The uncertainty was predicted based on the 2.5% prediction interval, 97.5% prediction intervals, and the mean prediction. The uncertainty mapping patterns shared by the modeling techniques BGLM, PLSR M5 tree model, and GPR were similar, with low uncertainty dominating across the study area in the 2.5%, 97.5%, and mean uncertainty maps, respectively. However, patches of moderate to high uncertainty were exhibited in the southeastern and southwestern areas of the uncertainty map. The uncertainty for the BRNN and EGB maps displayed predominantly low uncertainty for the 2.5% uncertainty distribution map all over the study area, with a spot of moderate uncertainty in the southeastern and southwestern enclave of the map. However, the mean uncertainty and the 97.5% equally exhibited low to moderate uncertainty levels with slightly higher uncertainty for both modeling approaches in the southeastern and southwestern regions of the map. The RRF and MARS algorithms also displayed low to moderate uncertainty levels all over the study area for the 2.5% and 97.5% prediction intervals. Nevertheless, the mean predictions for the algorithms likewise exhibited a high level of uncertainty in the southeastern and southwestern areas of the map (see Figs. 8A and 8B).

4. Conclusion

This study applies a series of modeling algorithms to predict Cd in agricultural soil in an area in the southeastern part of the Czech Republic. Three different scenarios were applied by combining the terrain attributes coupled with modelling approaches (Scenario 1), spectral indices combined with modelling approaches (Scenario 2), and a combination of terrain attributes, spectral indices, and modelling approaches (Scenario 3). According to the obtained results, except for EGB, all the modelling approaches in Scenario 1 produced acceptable results. However, the overall assessment in Scenario 1 indicated that the M5 tree model was the best model capable of predicting Cd in agricultural soil using terrain attributes as an auxiliary dataset. The cumulative results in Scenario 2 suggested that EGB combined with spectral indices produced the best results. However, the performance of Scenario 2 compared to Scenario 1 suggested that using spectral indices as an auxiliary dataset combined with modelling approaches yielded better results than using terrain attributes. Except for MARS, all the models performed very well in Scenario 3. However, the cumulative analysis revealed that the M5 tree model combined with terrain attributes and spectral indices produced the best results in Scenario 3, with high R2 and CCC values and minimal error margins. The overall comparison of the optimal modeling scenarios revealed that the combination of spectral indices and terrain attributes to the M5 tree model produced the best results, yielding the highest Cd prediction results with higher R2 and CCC values and the lowest RMSE, MAE, and MdAE values. In conclusion, it was clear that not all modeling approaches produced optimal results when auxiliary datasets were combined. As a result, this study suggests that the best model for predicting Cd in agricultural soil is a combination of environmental covariates such as spectral indices and terrain attributes combined with the appropriate modeling approach like the M5 tree model.

CRediT authorship contribution statement

Prince Chapman Agyeman: Conceptualization, Methodology, Writing – original draft, Visualization. **Vahid Khosravi:** Data curation, Investigation. **Ndiye Michael Kebonye:** Software, Data curation. **Kingsley John:** Software, Visualization. **Luboš Borůvka:** Supervision. **Radim Vašát:** Data curation, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This study was supported by an internal PhD grant no. 21130/1312/3131 of the Faculty of Agrobiology, Food and Natural Resources of the Czech University of Life Sciences Prague (CZU). The support from the Ministry of Education, Youth and Sports of the Czech Republic (project No. CZ.02.1.01/0.0/0.0/16_019/0000845) is also acknowledged. Finally, The Centre of Excellence (Centre of the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially risk substances of anthropogenic origin: comprehensive assessment of the soil contamination risks for the quality of agricultural products, NutRisk Centre). that could have appeared to influence the scientific work in this manuscript.

References

Adimalla, N., 2020. Heavy metals contamination in urban surface soils of Medak province, India, and its risk assessment and spatial distribution. Environ. Geochem. Health 42 (1), 59–75. https://doi.org/10.1007/S10653-019-00270-1/FIGURES/6.

Adimalla, N., Qian, H., Wang, H., 2019. Assessment of heavy metal (HM) contamination in agricultural soil lands in northern Telangana, India: an approach of spatial distribution and multivariate statistical analysis. Environ. Monit. Assess. 191 (4), 1–15. https://doi.org/10.1007/S10661-019-7408-1/TABLES/6.

Adimalla, N., Wang, H., 2018. Distribution, contamination, and health risk assessment of heavy metals in surface soils from northern Telangana, India. Arab. J. Geosci. 11 (21), 1–15. https://doi.org/10.1007/S12517-018-4028-Y.

Agyeman, P.C., John, K., Kebonye, N.M., Borůvka, L., Vašát, R., Drábek, O., 2021a. A geostatistical approach to estimating source apportionment in urban and peri-urban soils using the Czech Republic as an example. Sci. Rep. 11 (1), 1–15.

Agyeman, P.C., Ahado, S.K., Borůvka, L., Biney, J.K.M., Sarkodie, V.Y.O., Kebonye, N.M., Kingsley, J., 2021. Trend analysis of global usage of digital soil mapping models in the prediction of potentially toxic elements in soil/sediments: a bibliometric review. Environ. Geochem. Health 43, 1715–1739. https://doi.org/10.1007/S10653-020-00742-9.

Agyeman, P.C., Ahado, S.K., John, K., Kebonye, N.M., Vašát, R., Borůvka, L., Kočárek, M., Němeček, K., 2021d. Health risk assessment and the application of CF-PMF: a pollution assessment–based receptor model in an urban soil. J. Soils Sediments 21, 3117–3136. https://doi.org/10.1007/S11368-021-02988-X/TABLES/6.

Agyeman, P.C., Ahado, S.K., Kingsley, J., Kebonye, N.M., Biney, J.K.M., Borůvka, L., Vasat, R., Kocarek, M., 2021e. Source apportionment, contamination levels, and spatial prediction of potentially toxic elements in selected soils of the Czech Republic. Environ. Geochem. Health 43 (1), 601–620.

Agyeman, P.C., John, K., Kebonye, N.M., Borůvka, L., Vašát, R., Drábek, O., Němeček, K., 2021f. Human health risk exposure and ecological risk assessment of potentially toxic element pollution in agricultural soils in the district of Frydek Mistek, Czech Republic: a sample location approach. Environ. Sci. Eur. 33 (1), 1–25. https://doi.org/10.1186/S12302-021-00577-W/TABLES/3.

Ali, H., Khan, E., Sajad, M.A., 2013. Phytoremediation of heavy metals—concepts and applications. Chemosphere 91 (7), 869–881.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., Zhang, G.L., 2014. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. Adv. Agron. 125, 93–134.

Asgari, N., Ayoubi, S., Demattê, J.A., Jafari, A., Safanelli, J.L., Da Silveira, A.F.D., 2020a. Digital mapping of soil drainage using remote sensing, DEM and soil color in a semiarid region of Central Iran. Geoderma Regional 22, e00302. https://doi.org/10.1016/J.GEODRS.2020.E00302.

Asgari, N., Ayoubi, S., Jafari, A., Demattê, J.A., 2020b. Incorporating environmental variables, remote and proximal sensing data for digital soil mapping of USDA soil great groups. Int. J. Remote Sens. 41 (19), 7624–7648.

Azizi, K., Ayoubi, S., Nabiollahi, K., Garosi, Y., Gislum, R., 2022. Predicting heavy metal contents by applying machine learning approaches and environmental covariates in west of Iran. J. Geochem. Explor. 233, 106921.

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L., Panagos, P., 2019. Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. Geoderma 355, 113912.

Bhagat, S.K., Tung, T.M., Yaseen, Z.M., 2021. Heavy metal contamination prediction using ensemble model: case study of Bay sedimentation, Australia. J. Hazardous Mater. 403, 123492.

Bhagat, S.K., Tiyasha, T., Tung, T.M., Mostafa, R.R., Yaseen, Z.M., 2020. Manganese (Mn) removal prediction using extreme gradient model. Ecotoxicol. Environ. Saf. 204, 111059. https://doi.org/10.1016/J.ECOENV.2020.111059.

Biabani, R., Halaghi, M.M., Ghorbani, K.H., 2016. M5 model tree to predict temporal evolution of clear-water abutment scour. Open J. Geol. 6 (9), 10451054. https://doi.org/10.4236/OJG.2016.69078.

Biney, J.K.M., Vašát, R., Blöcher, J.R., Borůvka, L., Němeček, K., 2021. Using an ensemble model coupled with portable X-ray fluorescence and visible near-infrared spectroscopy to explore the viability of mapping and estimating arsenic in an agricultural soil. Sci. Total Environ. https://doi.org/10.1016/J.SCITOTENV.2021.151805.

Boettinger, J.L., 2010. Environmental covariates for digital soil mapping in the western USA. In: Digital Soil Mapping. Springer, Dordrecht, pp. 17–27.

Chandrasekaran, A., Ravisankar, R., Harikrishnan, N., Satapathy, K.K., Prasad, M.V.R., Kanagasabapathy, K.V, 2015. Multivariate statistical analysis of heavy metal concentration in soils of Yelagiri Hills, Tamilnadu, India – Spectroscopical approach. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 137, 589–600. https://doi.org/10.1016/J.SAA.2014.08.093.

Chen, L., Ren, C., Li, L., Wang, Y., Zhang, B., Wang, Z., Li, L., 2019. A comparative assessment of geostatistical, machine learning, and hybrid approaches for mapping topsoil organic carbon content. ISPRS Int. J. Geo-Inf. 8 (4), 174. https://doi.org/10.3390/ijgi8040174.

Climent, F., Momparler, A., Carmona, P., 2019. Anticipating bank distress in the Eurozone: An extreme gradient boosting approach. J. Bus. Res. 101, 885–896.

Congdon, P., 2007. Bayesian Statistical Modelling. John Wiley & Sons.

Cools, N., B.D.V., 2016. Sampling and analysis of soil.“ Manual on methods and criteria for harmonized sampling, assessment, monitoring and analysis of the effects of air pollution on forests. [WWW Document] (accessed 4.5.22).

Decree No. 153/2016 Coll. Vyhláška č. 153/2016 Sb. ze dne 9. května 2016 o stanovení podrobností ochrany kvality zemědělské půdy a o změně vyhlášky č. 13/1994 Sb., kterou se upravují některé podrobnosti ochrany zemědělského půdního fondu (in Czech). In: Sbírka Zákonů České Republiky. 2016, částka vol. 59, pp. 2692–2704. ISSN 1211-1244.

Deng, H., Runger, G., 2012, June. Feature selection via regularized trees. In: The 2012 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.

Ding, Q., Cheng, G., Wang, Y., Zhuang, D., 2017. Effects of natural factors on the spatial distribution of heavy metals in soils surrounding mining regions. Sci. Total Environ. 578, 577–585.

Ehsani, M.R., Upadhyaya, S.K., Slaughter, D., Shafii, S., Pelletier, M., 1999. A NIR Technique for Rapid Determination of Soil Mineral Nitrogen. Precis. Agric. 1, 217–234. https://doi.org/10.1023/A:1009916108990.

Etemad-Shahidi, A., Mahjoobi, J., 2009. Comparison between M5′ model tree and neural networks for prediction of significant wave height in Lake Superior. Ocean Eng. 36, 1175–1181. https://doi.org/10.1016/J.OCEANENG.2009.08.008.

Fang, H., Huang, L., Wang, J., He, G., Reible, D., 2016. Environmental assessment of heavy metal transport and transformation in the Hangzhou Bay. J. Hazard. Mater. 302, 447–457.

Gamon, J.A., Penuelas, J., Field, C.B., 1992. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. Remote Sens. Environ. 41 (1), 35–44.

Gauch Jr., H.G., Gauch Jr., H.G., Gauch, H.G., 2003. Scientific method in practice. Cambridge University Press.

Goydaragh, M.G., Taghizadeh-Mehrjardi, R., Jafarzadeh, A.A., Triantafilis, J., Lado, M., 2021. Using environmental variables and Fourier Transform Infrared Spectroscopy to predict soil organic carbon. CATENA 202, 105280. https://doi.org/10.1016/J.CATENA.2021.105280.

Gupta, V.V.S.R., Roper, M.M., Thompson, J., Pratley, J. E., Kirkegaard, J., 2020. Harnessing the benefits of soil biology in conservation agriculture. Australian agriculture in, pp. 237–253.

Heddam, S., 2021. New formulation for predicting soil moisture content using only soil temperature as predictor: multivariate adaptive regression splines versus random forest, multilayer perceptron neural network, M5Tree, and multiple linear regression. In Water Engineering Modeling and Mathematic Tools, pp. 45–62. Elsevier. https://doi.org/10.1016/B978-0-12-820644-7.00027-X.

Hutton, M., 1983. Sources of cadmium in the environment. Ecotoxicol. Environ. Saf. 7 (1), 9–24.

Iqbal, J., Thomasson, J.A., Jenkins, J.N., Owens, P.R., Whisler, F.D., 2005. Spatial variability analysis of soil physical properties of alluvial soils. Soil Sci. Soc. Am. J. 69 (4), 1338–1350. https://doi.org/10.2136/sssaj2004.0154.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19 (1), 1–67.

Jia, Y., Jin, S., Savi, P., Gao, Y., Tang, J., Chen, Y., Li, W., 2019. GNSS-R soil moisture retrieval based on a XGboost machine learning aided method: Performance and validation. Remote Sens. 11 (14), 1655. https://doi.org/10.3390/rs11141655.

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., Talebiesfandarani, S., 2019. PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere 10 (7), 373. https://doi.org/10.3390/atmos10070373.

John, K., Afu, S.M., Isong, I.A., Aki, E.E., Kebonye, N.M., Ayito, E.O., Penížek, V., 2021a. Mapping soil properties with soil-environmental covariates using geostatistics and multivariate statistics. Int. J. Environ. Sci. Technol. 18 (11), 3327–3342. https://doi.org/10.1007/S13762-020-03089-X/TABLES/5.

John, K., Agyeman, P.C., Kebonye, N.M., Isong, I.A., Ayito, E.O., Ofem, K.I., Qin, C.Z., 2021b. Hybridization of cokriging and gaussian process regression modelling techniques in mapping soil sulphur. Catena (Amst) 206 Catena 206. https://doi.org/10.1016/j.catena.2021.105534.

John, K., Bouslihim, Y., Ofem, K.I., Hssaini, L., Razouk, R., Okon, P.B., Isong, I.A., Agyeman, P.C., Kebonye, N.M., Qin, C., 2021c. Do model choice and sample ratios separately or simultaneously influence soil organic matter prediction? Int. Soil Water Conserv. Res. https://doi.org/10.1016/J.ISWCR.2021.11.003.

Kalambukattu, J.G., Kumar, S., Arya Raj, R., 2018. Digital soil mapping in a Himalayan watershed using remote sensing and terrain parameters employing artificial neural network model. Environ. Earth Sci. 77 (5), 1–14. https://doi.org/10.1007/S12665-018-7367-9/FIGURES/13.

Kawaguchi, K., Bengio, Y., 2019. Depth with nonlinearity creates no bad local minima in ResNets. Neural Networks 118, 167–174.

Keshavarzi, B., Abbasi, S., Moore, F., Mehravar, S., Sorooshian, A., Soltani, N., Najmeddin, A., 2018. Contamination level, source identification and risk assessment of potentially toxic elements (PTEs) and polycyclic aromatic hydrocarbons (PAHs) in street dust of an important commercial center in Iran. Environ. Manage. 62 (4), 803–818. https://doi.org/10.1007/S00267-018-1079-5.

Khosravi, V., Doulati Ardejani, F., Yousefi, S., Aryafar, A., 2018. Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods. Geoderma 318, 29–41.

Kim, M., Geum, Y., 2020. Predicting Patent Transactions Using Patent-Based Machine Learning Techniques. IEEE Access 8, 188833–188843.

Kim, S., Choi, Y., Lee, M., 2015. Deep learning with support vector data description. Neurocomputing 165, 111–117.

Kozák, J., Němeček, J., Borůvka, L., Lérová, Z., Němeček, K., Kodešová, R., Zádorová, T., 2010. Atlas půd České republiky.[Soil Atlas of the Czech Republic.]. Czech University of Life Sciences Prague, Prague, p. 150.

Kumar, S., Deswal, S., 2020a. Phytoremediation capabilities of Salvinia molesta, water hyacinth, water lettuce, and duckweed to reduce phosphorus in rice mill wastewater. Int. J. Phytorem. 22 (11), 1097–1109.

Kumar, S., Deswal, S., 2020b. Phytoremediation capabilities of Salvinia molesta, water hyacinth, water lettuce, and duckweed to reduce phosphorus in rice mill wastewater. Int. J. Phytorem. 22 (11), 1097–1109.

Lawrence, I., Lin, K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 255–268.

Lehmann, J., Bossio, D.A., Kögel-Knabner, I., Rillig, M.C., 2020. The concept and future prospects of soil health. Nature Rev. Earth Environ. 1 (10), 544–553.

Li, L., Lu, J., Wang, S., Ma, Y.i., Wei, Q., Li, X., Cong, R., Ren, T., 2016. Methods for estimating leaf nitrogen concentration of winter oilseed rape (Brassica napus L.) using in situ leaf spectroscopy. Ind. Crops Prod. 91, 194–204.

Li, Y., Li, M., Li, C., Liu, Z., 2020. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. Sci. Rep. 10 (1), 1–12.

Wilding, L.P., 1985. Spatial variability: its documentation, accomodation and implication to soil surveys. In Soil spatial variability, Las Vegas NV, 30 November–1 December 1984, pp. 166–194.

Ma, J., Ding, Y., Cheng, J.C., Tan, Y., Gan, V.J., Zhang, J., 2019. Analyzing the leading causes of traffic fatalities using XGBoost and grid-based analysis: a city management perspective. IEEE Access 7, 148059–148072.

Mao, Y., Liu, J., Cao, W., Ding, R., Fu, Y., Zhao, Z., 2021. Research on the quantitative inversion model of heavy metals in soda saline land based on visible-near-infrared spectroscopy. Infrared Phys. Technol. 112 https://doi.org/10.1016/J.INFRARED.2020.103602.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.

Men, C., Liu, R., Wang, Q., Guo, L., Miao, Y., Shen, Z., 2019. Uncertainty analysis in source apportionment of heavy metals in road dust based on positive matrix factorization model and geographic information system. Sci. Total Environ. 652, 27–39.

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301–311.

Mishra, P., Nikzad-Langerodi, R., 2020. Partial least square regression versus domain invariant partial least square regression with application to near-infrared spectroscopy of fresh fruit. Infrared Phys. Technol. 111, 103547. https://doi.org/10.1016/J.INFRARED.2020.103547.

Neissi, L., Golabi, M., Gorman, J.M., 2020. Spatial interpolation of sodium absorption ratio: A study combining a decision tree model and GIS. Ecol. Indicators 117. https://doi.org/10.1016/J.ECOLIND.2020.106611.

Nemecek, J.P.E., 1992. Retrospective experimental monitoring of heavy-metals containing in soils of the Czech Republic [WWW Document].

Ntzoufras, I., 2011. Bayesian modeling using WinBUGS, vol. 698. John Wiley & Sons.

Piekut, A., Baranowska, R., Marchwińska-Wyrwał, E., Ćwieląg-Drabek, M., Hajok, I., Dziubanek, G., Grochowska-Niedworok, E., 2018. Is the soil quality monitoring an effective tool in consumers' protection of agricultural crops from cadmium soil contamination?—a case of the Silesia region (Poland). Environ. Monit. Assess. 190 (1), 1–9. https://doi.org/10.1007/S10661-017-6413-5.

Probst, P., Boulesteix, A.L., Bischl, B., 2019. Tunability: importance of hyperparameters of machine learning algorithms. J. Mach. Learn. Res. 20 (1), 1934–1965.

Pyo, J., Hong, S.M., Kwon, Y.S., Kim, M.S., Cho, K.H., 2020. Estimation of heavy metals using deep neural network with visible and infrared spectroscopy of soil. Sci. Total Environ. 741, 140162. https://doi.org/10.1016/J.SCITOTENV.2020.140162.

Rahimikhoob, A., 2016. Comparison of M5 model tree and artificial neural network's methodologies in modelling daily reference evapotranspiration from NOAA satellite images. Water Resour. Manage. 30 (9), 3063–3075. https://doi.org/10.1007/S11269-016-1331-9/TABLES/5.

Sakizadeh, M., Mirzaei, R., Ghorbani, H., 2017. Support vector machine and artificial neural network to model soil pollution: a case study in Semnan Province, Iran. Neural Comput. Appl. 28 (11), 3229–3238. https://doi.org/10.1007/S00521-016-2231-X/TABLES/8.

Shi, T., Chen, Y., Liu, Y., Wu, G., 2014a. Visible and near-infrared reflectance spectroscopy—An alternative for monitoring soil contamination by heavy metals. J. Hazard. Mater. 265, 166–176.
Shi, T., Liu, H., Wang, J., Chen, Y., Fei, T., Wu, G., 2014b. Monitoring arsenic contamination in agricultural soils with reflectance spectroscopy of rice plants. Environ. Sci. Technol. 48 (11), 6264–6272.
Sihag, P., Keshavarzi, A., Kumar, V., 2019. Comparison of different approaches for modeling of heavy metal estimations. SN Appl. Sci. 1 (7), 1–11. https://doi.org/10.1007/S42452-019-0816-6/FIGURES/11.
Song, H., Hu, K., An, Y., Chen, C., Li, G., 2018. Spatial distribution and source apportionment of the heavy metals in the agricultural soil in a regional scale. J. Soils Sediments 18 (3), 852–862. https://doi.org/10.1007/S11368-017-1795-0.
Sui, H., Li, L., Zhu, X., Chen, D., Wu, G., 2016. Modeling the adsorption of PAH mixture in silica nanopores by molecular dynamic simulation combined with machine learning. Chemosphere 144, 1950–1959.
Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., Valavi, R., Davatgar, N., Behrens, T., Scholten, T., 2020. Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space. Remote Sens. 12 (7), 1095.
Tajik, S., Ayoubi, S., Shirani, H., Zeraatpisheh, M., 2019. Digital mapping of soil invertebrates using environmental attributes in a deciduous forest ecosystem. Geoderma 353, 252–263. https://doi.org/10.1016/J.GEODERMA.2019.07.005.
Tajik, S., Ayoubi, S., Zeraatpisheh, M., 2020. Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. Geoderma Regional 20, e00256. https://doi.org/10.1016/J.GEODRS.2020.E00256.
Tchagang, A.B., Valdés, J.J., 2019, September. Prediction of the atomization energy of molecules using Coulomb matrix and atomic composition in a Bayesian regularized neural networks. In: International Conference on Artificial Neural Networks (pp. 793–803). Springer, Cham. https://doi.org/10.1007/978-3-030-30493-5_75.
Tejnecký, V., Šamonil, P., Matys Grygar, T., Vašát, R., Ash, C., Drahota, P., Šebek, O., Němeček, K., Drábek, O., 2015. Transformation of iron forms during pedogenesis after tree uprooting in a natural beech-dominated forest. Catena 132, 12–20.
Vacek, O., Vašát, R., Borůvka, L., 2020. Quantifying the pedodiversity-elevation relations. Geoderma 373, 114441. https://doi.org/10.1016/j.geoderma.2020.114441.
Vašát, R., Kodešová, R., Klement, A., Borůvka, L., 2017. Simple but efficient signal pre-processing in soil organic carbon spectroscopic estimation. Geoderma 298, 46–53. https://doi.org/10.1016/J.GEODERMA.2017.03.012.
Vasudevan, S., Ramos, F., Nettleton, E., Durrant-Whyte, H., 2009. Gaussian process modeling of large-scale terrain. J. Field Rob. 26 (10), 812–840. https://doi.org/10.1002/ROB.20309.
Viscarra Rossel, R.A., Webster, R., Bui, E.N., Baldock, J.A., 2014. Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Glob. Change Biol. 20, 2953–2970. https://doi.org/10.1111/GCB.12569.
Wan, M., Qu, M., Hu, W., Li, W., Zhang, C., Cheng, H., Huang, B., 2019. Estimation of soil pH using PXRF spectrometry and Vis-NIR spectroscopy for rapid environmental risk assessment of soil heavy metals. Process Saf. Environ. Prot. 132, 73–81. https://doi.org/10.1016/J.PSEP.2019.09.025.
Wang, F., Gao, J., Zha, Y., 2018. Hyperspectral sensing of heavy metals in soil and vegetation: Feasibility and challenges. ISPRS J. Photogramm. Remote Sens. 136, 73–84. https://doi.org/10.1016/J.ISPRSJPRS.2017.12.003.
Wang, J., Cui, L., Gao, W., Shi, T., Chen, Y., Gao, Y., 2014. Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy. Geoderma 216, 1–9. https://doi.org/10.1016/J.GEODERMA.2013.10.024.
Wang, S., Zhu, L., Fuh, J.Y.H., Zhang, H., Yan, W., 2020. Multi-physics modeling and Gaussian process regression analysis of cladding track geometry for direct energy deposition. Opt. Lasers Eng. 127, 105950.
Wang, Y., Witten, I.H., 1996. Induction of model trees for predicting continuous classes.
Weather Spark, 2016. Average Weather in Frýdek-Místek, Czechia, Year-Round - Weather Spark [WWW Document]. URL https://weatherspark.com/y/83671/Average-Weather-iFrýdek-Místek-Czechia-Year-Round.
World Health Organization (WHO), 2010. Preventing Disease through Healthy Environments. Exposure to Dioxins and Dioxin-like Substances: A Major Public Health Concern. Public Health and Environment. World Health Organization, Geneva, p. 27.
Wu, J., Teng, Y., Chen, H., Li, J., 2016. Machine-learning models for on-site estimation of background concentrations of arsenic in soils using soil formation factors. J. Soils Sediments 16 (6), 1787–1797.
Xu, X., Du, C., Ma, F., Shen, Y., Wu, K., Liang, D., Zhou, J., 2019. Detection of soil organic matter from laser-induced breakdown spectroscopy (LIBS) and mid-infrared spectroscopy (FTIR-ATR) coupled with multivariate techniques. Geoderma 355, 113905. https://doi.org/10.1016/J.GEODERMA.2019.113905.
Zare, E., Wang, J., Zhao, D., Arshad, M., Triantafilis, J., 2021. Scope to map available water content using proximal sensed electromagnetic induction and gamma-ray spectrometry data. Agric. Water Manag. 247, 106705 https://doi.org/10.1016/J.AGWAT.2020.106705.
Zeraatpisheh, M., Jafari, A., Bagheri Bodaghabadi, M., Ayoubi, S., Taghizadeh-Mehrjardi, R., Toomanian, N., Kerry, R., Xu, M., 2020. Conventional and digital soil mapping in Iran: Past, present, and future. Catena 188, 104424.
Zhang, G.-L., Liu, F., Song, X.-D., 2017. Recent progress and future prospect of digital soil mapping: A review. Journal of Integrative. Agriculture 16 (12), 2871–2885.
Zhang, W., Goh, A.T.C., 2016. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geosci. Front. 7, 45–52. https://doi.org/10.1016/J.GSF.2014.10.003.
Zhang, Y., Xu, X., 2021. Fe-Based Superconducting Transition Temperature Modeling through Gaussian Process Regression. J. Low Temp. Phys. 202, 205–218. https://doi.org/10.1007/S10909-020-02545-9.
Zhang, H., Yin, S., Chen, Y., Shao, S., Wu, J., Fan, M., Chen, F., Gao, C., 2020. Machine learning-based source identification and spatial prediction of heavy metals in soil in a rapid urbanization area, eastern China. J. Cleaner Prod. 273, 122858.
Zhao, D., Wang, J., Jiang, X., Zhen, J., Miao, J., Wang, J., Wu, G., 2022. Reflectance spectroscopy for assessing heavy metal pollution indices in mangrove sediments using XGBoost method and physicochemical properties. CATENA 211, 105967. https://doi.org/10.1016/J.CATENA.2021.105967.
Zhao, H., Xia, B., Fan, C., Zhao, P., Shen, S., 2012. Human health risk from soil heavy metal contamination under different land uses near Dabaoshan Mine, Southern China. Sci. Total Environ. 417–418, 45–54. https://doi.org/10.1016/J.SCITOTENV.2011.12.047.
Zhu, A., Lu, G., Liu, J., Qin, C., 2018. Spatial prediction based on Third Law of Geography. Taylor & Francis 24, 225–240. https://doi.org/10.1080/19475683.2018.1534890.

Process Safety and Environmental Protection 164 (2022) 729–746

Contents lists available at ScienceDirect

Process Safety and Environmental Protection

journal homepage: www.journals.elsevier.com/process-safety-and-environmental-protection

Ecological risk source distribution, uncertainty analysis, and application of geographically weighted regression cokriging for prediction of potentially toxic elements in agricultural soils

Prince Chapman Agyeman [a,*], Kingsley JOHN [a], Ndiye Michael Kebonye [a], Solomon Ofori [b], Luboš Borůvka [a], Radim Vašát [a], Martin Kočárek [a]

[a] Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague, Czech Republic

[b] Department of Water Technology and Environmental Engineering, Faculty of Environmental Technology. University of Chemistry and Technology, Technicka 5, 166 28 Praha 6- Devices, Prague, Czech Republic

ARTICLE INFO

Keywords:
Source distribution
Ecological risk-positive matrix factorization
Geographically weighted regression cokriging
Random forest
Uncertainty assessment

ABSTRACT

A resilient environment is essential for society's long-term viability. Receptor models have evolved into an excellent tool for detecting pollution sources and evaluating each source's empirical contributions based on ecological datasets. One hundred and fifteen soil sample were collected from the district of Frydek Mistek in the Czech Republic and the concentration of arsenic (As), cadmium (Cd), copper (Cu), chromium (Cr), manganese (Mn), nickel (Ni), lead (Pb)and zinc (Zn) measured inductively coupled plasma–optical emission spectrometry. The results suggested that the hybridized receptor models ER-PMF and PMF identified the following geogenic, steel industries, vehicular traffic, and agro-based activities such as pesticide and fertilizer applications as the primary sources in the source distribution. The ER-PMF source pollution identification efficiency ranged from R2 0.872–0.970, RMSE 0.128–17.344 and MAE 0.085–10.388, whereas the PMF R2 ranged from 0.883 to 0.960, RMSE 0.246–79.003 and MAE 0.145–49.925. The overall assessment of the efficiency of the receptor models suggests that the ER-PMF appears to yield more efficient results in pollution source identification compared to PMF. The PTEs mapping using geographical weighted regression (GWR) and a hybridized regression approach, geographical weighted regression cokriging (GWRCoK), revealed that GWRCoK had a higher goodness of fit in the spatial prediction maps than GWR. According to Hakanson's risk index classification, the ecological risk level in the study area was moderate to high (risk level = 51 observed locations out of 115, or 44.35%); however, Chen's risk index reclassification indicated that the toxicity level in the study area was moderate to extremely high (risk level = 113 observed locations out of 115, or 98.26%). However, the uncertainty assessment results indicated that the DISP interval ratio of the hybridized ER-PMF model was lower than that of the parent PMF model. However, it was clear that the random error that could occur in the DISP based on the DISP interval ratio was likely to be lower in the ER-PMF receptor model than in the parent model. The assessment of PTEs in soil has been widely published, but this study recommends using a pollution assessment-based receptor model (ER-PMF), which has been shown to be reliable and practical in estimating distribution sources.

1. Introduction

Due to the human population explosion, there has been a concerted effort over the years to till the land in order to increase yield in sequence to feed the ever-growing human population, which has seen rapid development with intensive industrial revolution to enhance agricultural activities and precision farming. Aside from the rapid development of industries and increased agricultural investment, it is well known that

Abbreviations: DISP -, Dsiplacement; ER-PMF, Ecological risk-positive matrix factorisation; PMF, Positive matrix factorisation; RMSE, Root means square error; MAE, Mean absolute error; PTEs, Potential toxic elements; GWR, Geographical weighted regression; GWR-CoK, Geographical weighted regression-Cokriging; KED, Kriging with extenal drift; R2, R squared.

* Corresponding author.
E-mail address: agyeman@af.czu.cz (P.C. Agyeman).

https://doi.org/10.1016/j.psep.2022.06.051
Received 17 March 2022; Received in revised form 6 June 2022; Accepted 23 June 2022
Available online 25 June 2022

agriculture production activities in third-world countries have plateaued, which has a devastating effect on traditional agriculture. Soil is an essential nonrenewable treasure that serves as the source and reservoir of various pollutants, such as potentially toxic elements (PTEs) (Hossain Bhuiyan et al., 2021a). Numerous anthropogenic practices, such as urban sprawl, large-scale farming, and rapid industrialization, have always been the source of soil nutrient toxicity with PTEs and, as a result, are of global concern. According to Kumar et al. (2019) and Keshavarzi and Kumar (2019), human activities such as urban expansion, intensive farming, and industrial growth, are the primary drivers that inject PTEs into urban soil and have generated concerns globally. PTEs can enter the environment through anthropogenic activities such as agricultural activities (i.e., of fertilizers, pesticides, livestock manures, sewage sludge) and natural source like parent materials (Wu et al., 2020; Bayraklı and Dengiz, 2020). As a result, the risk associated with the injection of PTEs into agricultural soil must be mitigated to protect the ecosystem, human health, and the environment. PTEs in agricultural soils showed high spatial heterogeneity due to natural sources and anthropogenic activities, making identifying specific risks a difficult task in most urban and peri-urban agricultural soils. It is well known that the advancement of urban expansion, as well as intensification of agricultural activities example has been the major orifice for potentially toxic element (PTE) pollution in agricultural soils (Kars and Dengiz, 2020), which has attracted worldwide attention due to their toxicity, endless availability, and tenacity (Tóth et al., 2016, Adimalla et al., 2019). PTE accumulation in agricultural soils might, however, result in deterioration of soil physiology and function (Beattie et al., 2018).

Pollution studies, notably soil pollution, are conducted worldwide to resolve various forms of soil type pollution, such as agricultural soil, urban soil, industrial soil, forest soil, and other environmental types of pollutants of interest to soil scientists, researchers, and other community stakeholders. Zhang et al. (2012) and Li et al. (2015) disclosed that PTEs are considered lethal since they appear to bioaccumulate and have significant implications for public health and environmental quality. Kelepertzis (2014) discovered that anthropogenic activities were primarily responsible for the elevation of the content of PTEs in urban and agricultural soils. Interestingly, in today's world, a variety of anthropogenic practices, such as the consistent application of fertilizers and pesticides to agricultural fields, reckless wastewater irrigation, vehicular traffic and atmospheric deposition (Zhang and Wang, 2020), play crucial roles in the accumulation of PTEs in the soil across various urban and peri-urban areas around the globe. However, PTEs introduced into the environment, especially the soil, directly and indirectly impact plants, animals, and humans. Nevertheless, according to Zhang and Wang (2020), the effect of human activities, inferred on the environment due to irrational land use planning and a poor understanding of environmental conservation and soil erosion worldwide, has increased tremendously.

The need to improve the spatially differing correlation between the dependent variable and the independent variables prompted the development of geographically weighted regression (GWR) from the multiple linear regression (MLR) model (Brunsdon et al., 1996, Wang et al., 2020). The creation of GWR has yielded better modelling results than some traditional models, such as MLR, in terms of improving the grasp of the spatially varied correlations between environmental variables and PTEs (Fu et al., 2021). In recent years, there has been a trend to combine ordinary kriging (OK) with GWR, which has produced excellent results in predicting PTEs, soil organic carbon, and other environmental variables in soil or water. Among such papers that hybridized GWR with a geostatistical algorithm are Kumar et al. (2012), Wang et al. (2012), Pereira et al. (2018) and Ye et al. (2017). In this study, we hybridize GWR with cokriging to explore other alternatives that can be used to replace OK to enhance prediction efficiency and reduce uncertainty. Geographically weighted regression cokriging (GWRCoK) is made up of two parts, namely, stochastic and deterministic, which are simulated individually. The deterministic part of GWR is simulated to predict the trend of prediction of a response or a targeted variable using environmental covariates. However, the stochastic part of cokriging predicts the targeted variables by adding the residuals to the predicted variables. Previous research has demonstrated that novel hybrid models correlate with individual models. A model such as GWRK performed better than the other models applied, such as the ordinary least squares (OLS) model (Sun et al., 2019), GWR, OK (Wang et al., 2020; Shen et al., 2019), and regression kriging (Shen et al., 2019) models in terms of prediction accuracy (Mitran et al., 2018). In terms of ambiguity and reliability, a two-step technique is preferable to accomplish good results other than geostatistics, which are highly heterogeneous ecological landscapes, be it urban or peri-urban agricultural soil (Chen et al., 2019). The hybridization of geostatistical models and GWR has proven to be effective and a resilient hybrid model that can predict PTEs in soil.

Several authors have carried out countless research in various journals reporting on environmental pollution, soil pollution, and urban pollution across the globe that are detrimental to ecology and humans. Many of such reports are performed in developed, developing and underdeveloped countries, such as Spain by Rodríguez et al. (2008), France by Escarré et al. (2011), China by Yang et al. (2018), Mexico by Morton-Bermea et al. (2009), Ghana by Kodom et al., (2012), Peru by Santos-Francés et al. (2017) and Sudan by Ashaiekh et al. (2019). In estimating pollution levels in a wide range of areas, such as towns, urban areas, peri-urban areas, hinterlands and forests, researchers and students use several pollution indices, such as the geoaccumulation index, enrichment factor, single pollution index, pollution load index, Nemerow pollution index, contamination level, potential ecological risk index and other host pollution indices. These indices have been described by USEPA (1998) as a versatile environmental pollution computing tool that assesses and coordinates data, providing a depth of information that exposes perceptions and uncertainties but also considers the likelihood of adverse ecological effects.

Over the years, the application of receptor models to determine source distribution in soil has become a popular approach in soil science. This multivariate analytical technique helps researchers quantify the source contribution of PTEs percentagewise in an area of study. Huang et al. (2018b) characterize the receptor model as a statistical tool that recognizes the source pathway of PTEs and quantifies the source distribution of PTEs understudied and aids in preventing soil pollution in the environment. Therefore, once the PTEs under investigation source contributions are quantified, it provides an opportunity to deliver mitigation measures to alleviate pollution. Various authors have utilized different receptor models to determine the proportion of PTE contributions in multiple regions. Among the most widely used receptor models are positive matrix factorization (PMF), UNMIX, chemical mass balance (CMB) and absolute principal component score-multiple linear regression (APCS-, MLR). In recent times, most authors have preferred to use PMF and APCS-MLR. The potential of PMF and APCS-MLR to produce a consistent result in source apportionment of a specified analysis is undeniable. However, in some cases, when both approaches are compared, one becomes superior to the other. In a comparative study, Guan et al. (2019) suggested that APCS-MLR produced better outcomes than PMF, while Yang et al. (2013a) likewise concluded that PMF outperformed APCS-MLR and UNMIX.

Every effective model's efficiency and practicality stem from its capacity to predict or model with optimal efficiency and a low marginal error. Nonetheless, the current approach being introduced tends to combine PMF and a pollution assessment index to estimate source apportionment. Some of the drawbacks of PMF stated by some authors include inconsistencies in the predicted contribution for each probable pollution source (Haji Gholizadeh et al., 2016) and the inadequacy of PMF based on R^2 results (model efficiency) (Liu et al., 2020). The insidious nature of PTEs in the earth's crust, combined with the need for some of these metals or metalloids for human subsistence, has made it impossible to eliminate them from the environment. As a result, the

continuous pollution of the environment with these metals(loid)s is an eternal event. Numerous studies, including Liu et al. (2014), Sayadi et al. (2015a), Chen and Lu (2018), Wu et al. (2018), Zhu et al. (2018), Keshavarzi and Kumar (2020), Agyeman et al. (2021), Hossain Bhuiyan et al. (2021b) and many authors, have relied on either one or multiple of these pollution indices to determine the pollution levels of a piece of land, area, community, urban area, peri-urban areas and other meaningful places that are of interest to researchers. According to Sayadi et al. (2015a, 2015b), ecological risk is a technique for evaluating the potential risk of the environment impacted by exposure to one and sometimes more environmental factors through an environmental risk assessment. Gao et al. (2013) stated that the ecological risk index is commonly used to fully assess the possibility of toxicity and environmental dangers caused by PTEs. This index was developed by Håkanson (1980) to compute the level of pollution in soil and sediments. It is common and has been done frequently to use existing multivariate tools in research, such as PMF, ER, GWR, and Cok. To the best of our knowledge, no research has attempted to combine these multivariate statistical tools as has been done in the current paper. PMF will be hybridized with ER to produce a hybridized receptor model that will be compared to the parent receptor model (PMF). In contrast, GWR will be hybridized with Cok to create a new hybridized algorithm for predicting and mapping PTEs in soil. What is the environmental risk level of the productive soil in the study area? We hypothesized that determining the impact of agriculture and industries on the soil health of the study area would depend on the appropriate pollution indices used. What is the uncertainty based on the receptor models used? This research seeks to harness the potential of the ecological risk index to evaluate the possible PTE toxicity level and the ecological risk exposure of the study area. The specific objectives of this paper are to determine the environmental risk level of the study area, evaluate ER-PMF and PMF receptor models for estimating PTE source allotment, employ ecological risk-assessed PTE values to calculate PCA and a correlation matrix, estimate the uncertainty based on the receptor models and assess the efficiency of the prediction of PTEs based on geographical weighted regression or a hybridized model.

2. Materials and methods

2.1. Study area

The study area is situated within the district of Frýdek-Místek, with 57 peri-urban and urban areas (see Fig. 1). The research region is positioned at a latitude of 49°41^{0,,} north and a longitude of 18°20^{0,,} east at an elevation of 225–327 m above sea level, with a cold temperate temperature and a high amount of rainfall even in dry months. Frýdek-Místek has humid, partially wet summers and cold, dry, windy summers, and most winters are cloudy. Temperatures vary slightly between − 5°C and 24 °C throughout the year and are seldom below − 14 °C or above 30 °C, whereas the average annual precipitation is between 685 and 752 mm (Weather Spark, 2016). The geomorphology of the study area displays a rugged terrain that is considered part of the Moravian-Silesian Beskydy and the outer carpathian mountain and the highest mountains. The area survey of the district is estimated at 1208 km^{2,} with 39.38% of the land area for cultivation and 49.36% being forests. The measurement of the study area sampled from the Frýdek-Místek district is 889.83 km^2. A significant area for evaluating the distribution and related ecological impacts of PTEs is Trinec and Vitkovice. In and around the district, some parts of Ostrava, which form part of the area under investigation, are endowed with industries such as the steel industry and metal works (Agyeman et al., 2022).

Weather parent material, which is generally proportionately composed of sandy and silty soils, is used to characterize the physical and chemical properties of the soil (Kozák, 2010). The soil reaction, on

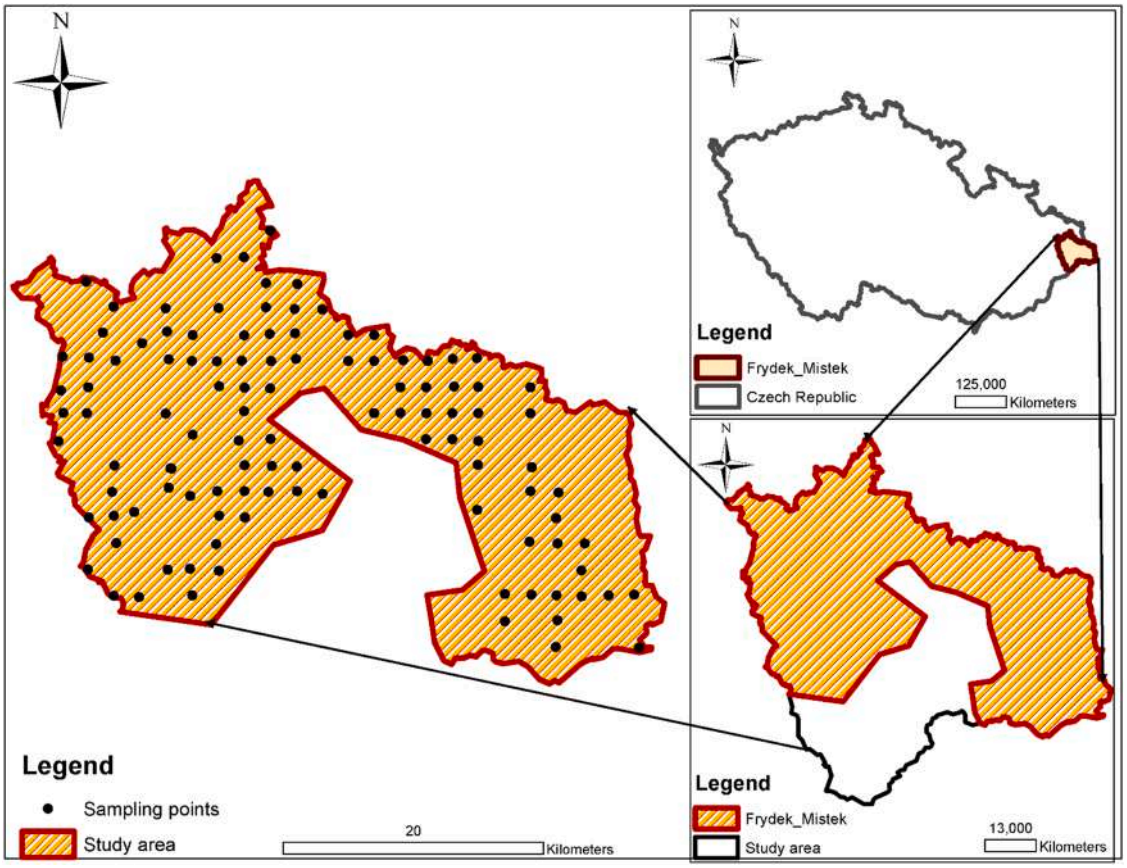


Fig. 1. Study area map and the sampling points.

the other hand, lies in the neutral to weakly acidic range, which is a favorable medium for aggressive microbial degradation and chemical fertility (Kozák, 2010). The soil texture has a moderate and satisfactory physical capability, with a high level of permeability and significant water retention, as well as interior draining (Vacek et al., 2020). The soil properties are clearly distinguished from the color, structure, and carbonate content of the soil. The soil contains medium and fine textures, which come from its parent materials. They are mainly colluvial, alluvial, or aeolian deposits. In certain sections of the field, there are mottles in the top and subsoil, followed mainly through concrete and blanching. However, cambisols and stagnosols are the predominant soil types (Kozák, 2010). In the Czech Republic, these soils are prevalent and range from 455.1 to 493.5 m (Vacek et al., 2020).

2.2. Soil sampling and soil analysis

A total sample of 115 topsoils was collected across 57 peri-urban and urban areas within the district of Frydek Mistek. A regular grid sampling technique was adopted, and the soil sampling gaps were 2 × 2 km using a handheld GPS unit (Leica Zeno 5 GPS) at a depth of 0–20 cm. The samples taken were packed in Ziploc polythene bags, properly labelled, and brought to the laboratory. The samples were air-dried, crushed by a mechanical device (Fritsch disk mill pulverize) and then sieved (< 2 mm) to obtain a pulverized sample. One gram of a dried but homogenized and sieved soil sample (sieve size< 2 mm) was deposited in a Teflon bottle and labelled. Seven milliliters of 35% HCl and three ml of 65% HNO_3 were dispensed into Teflon bottles, and the cup was closed halfway to allow a further reaction to take place overnight (aqua regia procedure). The mixture was placed on a hot metal plate for 2 h to stimulate the process of digestion of the sample and left to cool. The mixture was then filtered to obtain a supernatant. The supernatant was transferred into a 50 ml volumetric flask and then diluted with deionized water to 50 ml. The diluted supernatant was then filtered into 50 ml PVC tubes. In addition, 1 ml of the diluted solution was diluted with 9 ml of deionized water and filtered into a 12 ml test tube prepared for PTE pseudo concentration of PTEs in the samples. To identify metal concentrations, inductively coupled plasma–optical emission spectrometry (ICP–OES) (Thermo Fisher Scientific, USA) was utilized following standard procedures and protocols. The quality assurance and control (QA/QC) approach was ensured by analyzing the standard reference material for each sample (SRM NIST 2711a Montana II soil). PTEs with low or half detection limits were excluded from this study. The detection limits of the PTEs utilized in this study are 0.0002 (Cd), 0.0007 (Cr), 0.0060 (Cu), 0.0001 (Mn), 0.0004 (Ni), 0.0015 (Pb), 0.0067 (As), and 0.0060 (Zn). Moreover, for each analysis, the quality control and quality assurance processes were ensured by checking the reference criteria. Duplicate analysis was carried out to ensure that the error was minimized.

2.3. Pollution indices

To measure PTE pollution's influence and toxic effects, the homogeneity of peri-urban and urban areas must be analyzed. On this basis, pollution indicators, including the ecological risk and single pollution index (PI), were employed to quantify the pollution level in the research region. The use of pollution indices allows for a consistent assessment of the condition of environmental pollution and the amount of human influence affecting the landscape and ecosystem in particular (Sawut et al., 2018; Huang et al., 2018a). Therefore, these parameters are extensively utilized in monitoring PTE pollution in polluted soils or, from a larger perspective, the ecosystem.

2.4. Single pollution index (PI)

The proportion of PTE concentration to geochemical background values is defined as PI. PI was pioneered by Tomlinson et al. (1980) and the equation is given as

$$PI = \frac{C_n}{B_n}$$

where B_n is the geochemical background value of the PTE in the soil (mg/kg) and C_n is the concentration of the PTE in the soil (mg/kg). PI is categorized into an absent PI< 1, low level (1 < PI<2), moderate level (2 < PI<3), strong level (3 < PI<5), or high level (PI > 6).

2.5. Ecological risk

The potential ecological risk index (ER and RI) is utilized to quantify the magnitude of the ecological hazard associated with toxic element concentrations in the environment or in soil. Håkanson (1980) formulated the index, and the equation is given as

$$RI = \sum_{i=1}^{n} E_r^i$$

where n is the number of PTEs and E_r^I is the ecological risk index factor, which is given as

$$E_r^I = T_r^i \quad \times PI$$

T_r^i denotes the toxicity response coefficient of a specific PTE, (Hakanson, 1980)and PI represents the single pollution index. The toxicity response coefficients of the PTEs used were 30 (Cd), 10 (As), 5 (Cu), 5(Pb), 2(Cr), 2(Zn), 2(Ni) and 1(Mn). The ER has five classifications: low risk (EI ≤ 40), moderate risk (40 < EI ≤ 80), considerable risk (80 < EI ≤ 160), high risk (160 < EI ≤ 320), and very high risk (EI ≥ 320). The RI has four categories, namely, low risk (RI ≤ 150), moderate risk (150 < RI ≤ 300), considerable risk (300 < RI ≤ 600), or very high risk (RI > 600).

2.6. Positive matrix factorization (PMF)

The source distribution study was executed using a multivariate receptor model such as PMF (US-EPA PMF 5.0 software) (U.S. EPA, 2014). PMF is a receptor model developed to quantify chemical mass balance, and the data matrix (original) X is expressed in the following order m x n, which may be written as

$$X = GF + E$$

G (m × p) is a factor contribution matrix, F (p × n) is a factor profile matrix, and E (m × n) is a residual error matrix. *E* is given as

$$e_{ij} = \sum_{k=1}^{p} g_{ik} \quad f_{ki} - \quad x_{ij}$$

where i signifies elements 1 to *m*, *j* denotes elements 1 to *n*, and *k* represents the source from 1 to *p*.

The PMF model acquires the contributions and profiles of the released factor by decreasing the objective function Q underneath the limitation of nonnegative contributors, and the solution in the US-EPA PMF software is approximated by the Multilinear Engine-2 (ME-2) (Paatero, 1999).

$$Q = \quad \sum_{i-1}^{n} \sum_{j-i}^{m} \left(e_{ij}/u_{ij}\right)^2$$

where u_{ij} is the uncertainty in the *j*th chemical element for sample *i, and* the authors previously discussed the uncertainty and the parameters involved (Agyeman et al., 2021).

2.7. Ecological risk positive matrix factorization receptor model (ER-PMF)

The traditional approach in estimating source apportionment using the PMF receptor model uses raw data acquired after laboratory analysis

to compute the source apportionment. The ER-PMF receptor model approach uses the calculated ER values of each PTE from every sampled point instead of the raw data to calculate the source contribution of each PTE. The receptor model ER-PMF is given as

$$\left(ER_r^i\right)_{ij} = \sum T_r^i \times \frac{(C_n)_{ij}}{(B_n)i}$$

where $(ER_r^i)_{ij}$ is the computed ecological risk of each PTE from the jth source in the ith sampling location,$(C_n)_{ij}$ is the concentration of the single PTE in the soil in the jth source from the ith sampling site, $(B_n)i$ is the concentration of the respective PTE under investigation of its geochemical background level and T_r^irepresents the toxicity response.

2.8. Geographical weighted regression (GWR)

GWR is a deterministic model that is an extension of an ordinary least square regression. GWR in a modelling process always considers nonstationary but spatial predictor relationships and employs the varying spatial coefficient in a linear local model (Song et al., 2017). According to LI et al. (2020), the traditional regression approach cannot calculate the global variables with an elevated degree of precision because soil qualities fluctuate with environmental variability. To capture spatial variability, a GWR creates local variables as opposed to global ones, such as local R^2, local coefficient and local model residual (Costa et al., 2018). Please see the supplementary materials for more information on the equation and the definition of the GWR parameters.

2.9. Random forest

The assemblage of varied regression trees and categorization is known as a random forest (RF). Breiman (2001) developed the method and claimed that it is comparable to accuracy in adaptive boosting. The computing ability of RF is faster, according to Gislason et al. (2006) and Heung et al. (2014). The RF's variable handling capabilities are categorical and continuous. According to Díaz-Uriarte, Alvarez de Andrés (2006), RF does not require variable preselection, and due to its robust nature, it is capable of handling noise. Cutler et al. (2007) established that the algorithm starts with a number of tree samples (ntree) taken from the data sampled. The operation is changed so that the predictors (mtry) are randomly selected. Each ntree creates a regression tree, and the RF algorithm selects the utmost split between the variables sampled rather than all of them (Nawar and Mouazen, 2017). Please see the supplementary materials for more information on the application of the modeling approach random forest.

2.10. Data partitioning

A random data split approach was used to divide the data into a test dataset (with 25% for validation) and a training dataset (75% for calibration). The training dataset was used to calibrate the regression models, while the test dataset was utilized to assess generalization capabilities (Kooistra et al., 2003). This was done to determine the suitability of the various models used to estimate PTE source apportionment. All the models were subjected to a 10-fold cross-validation process that was repeated five times. Each receptor model's factor contributions or scores were employed as predictors or explanatory variables (PTEs) to predict the target variables. R was used to carry out the modelling procedure.

2.11. Validation and accuracy assessment

A selection of validation standards was utilized to determine the most reliable model suitable for predicting source apportionment such as pollution evaluation-based positive matrix factorization receptor models and PMF to examine the validity of the receptor model and its validation. The receptor models were evaluated using the mean absolute error (MAE), root mean square error (RMSE), and R square, also known as the coefficient dedication (R^2). R^2 expresses the proportionate interchange in the response using the regression model. The model prediction capability is defined by the RMSE and the fluctuation dimension inside the independent dimension, whereas the MAE determines the true quantitative value. Please see the supplementary materials for more information on the equations for the validation and accuracy criteria.

2.12. Data modelling techniques

This study utilized random forest (RF), which has been identified as the most successful and dependable technique for prediction and soil mapping in soil science in the current era. Furthermore, Kebonye et al. (2021) and John et al. (2020) validated the effectiveness, dependability, and practicality of the MLA technique in soil science prediction and mapping. MLA is essentially an automated methodology that allows for the definition of a learning process based on the amount of data, allowing for multicollinearity and nonlinearity. Multicollinearity and nonlinearity, according to Gautam et al. (2011), help to avoid overfitting in the case of constrained soil sample positions.

2.13. Data analysis

PMF EPA 5.0 were used to conduct quantitative models (for source distribution estimation). Both principal component analysis and Pearson correlation matrix assessment were performed using RStudio. Modelling and spatial distribution maps of the PTEs were analyzed using ArcGIS version 10.2.1, and the ordinary kriging (OK) interpolation technique was employed.

The OK interpolation technique allowed us to estimate the spatial distribution of PTEs in the location under investigation. Kriging is an interpolation that predicts values of variable at locations where data are not available based on the spatial pattern of the available data.

Cokriging (CoK) is a geostatistical interpolation technique that employs a variety of variable forms in the prediction of a specific variable (John et al., 2021). Tziachris et al. (2019) stated that the explanatory variable must have a sturdy correlation with the response variable. Please see the supplementary material for more information on the equation and the application of OK and CoK.

3. Results

3.1. Sample descriptive statistics

The statistical description of the PTE data is presented in Table 1, indicating the maximum and minimum PTE concentrations, median, mean, standard deviation, kurtosis, skewness, coefficient of variability and other average PTE values from a different country. The PTEs analyzed for this research are Cd, Cr, Cu, Mn, As, Ni, Pb, and Zn. The PTE mean values decrease in the order of Mn<Zn<Pb<Cr<Cu<Ni<As<Cd. The maximum and minimum values range between 7.28 mg/kg to 1691.76 mg/kg and 0.16 mg/kg to 186.02 mg/kg, respectively. The kurtosis and skewness of the PTE estimated values also range between 1.37 and 11.77 and 0.79–3.04, respectively. As evaluated by the standard deviation, the average value of variability in the dataset ranges from 1.01 to 259.35. The distribution of dataset points in series throughout the estimated means measured using the coefficient of variation (CV) also ranged from 33.01 to 92.96. The geochemical background from the world average value (WAV), upper continental crust (UCC), European average value (EAV) extracted from Kabata--Pendias (2011) are also captured in Table 1.

Table 1
The statistical description of the PTEs.

	Mn	Ni	Pb	Zn	As	Cd	Cr	Cu
Standard Deviation	259.35	6.78	18.51	34.35	4.95	1.01	9.38	9.98
Kurtosis	1.37	2.49	18.80	7.32	11.77	10.45	2.69	4.90
Skewness	0.79	1.63	3.67	2.11	3.04	2.84	1.33	2.04
Minimum (mg/kg)	186.02	4.86	9.56	37.48	1.85	0.61	10.90	7.88
Maximum (mg/kg)	1691.76	42.39	155.69	272.18	30.42	7.28	62.78	62.62
Median (mg/kg)	664.39	13.75	30.10	75.47	4.57	1.61	26.90	19.68
CV	37.10	41.97	54.68	40.31	92.96	55.16	33.01	44.27
Mean values (mg/kg)	699.03	16.15	33.86	85.22	5.32	1.84	28.43	22.54
[a] UCC(mg/kg)	900.00	20.00	15.00	70.00	1.80	0.10	100.00	17.30
[b] WAV (mg/kg)	488.00	29.00	27.00	70.00	6.83	0.41	59.50	38.90
[c] EAV (mg/kg)	524.00	37.00	32.00	68.10	11.60	0.28	94.80	17.30

Kabata-Pendias (2011) [abc], CV (Coefficient of variation)

4. Multivariate analysis

4.1. Ecological risk correlation matrix (ER-CM) and ecological risk principal component analysis (ER-PCA)

Table 2 indicates the metallic relationship of the following PTEs: ecological risk cadmium (ERCd), ecological risk arsenic (ERAs), ecological risk copper (ERCu), ecological risk chromium (ERCr), ecological risk nickel (ERNi), ecological risk manganese (ERMn), ecological risk lead (ERPb) and ecological risk zinc (ERZn). The estimated ER-CM of the PTEs showed a high correlation between ERCd and ERAs, with r = 0.90. Other high correlation values were observed between ERCd and ERPb, ERCd and ERAs, ERZn and ERPb, and ERAs and ERPb, with r values = 0.85, 0.78, 0.83 and 0.75, respectively. However, some of the PTEs exhibited a moderate relationship with one another, such as ERAs and ERZn, as well as ERCu and ERNi, which both recorded r = 0.62 and 0.68, respectively.

The whole dataset was analyzed using ecological risk-principal component analysis (ER-PCA). The ER-PCA discharged three principal components, accounting for 83% of the total cumulative explained variance captured in Table 3. The relationship between the ER-PTE values was very high, and therefore, the r-value was fixed at r = 0.75. The first component loadings (PC 1) were for the ER-PTE ERPb, ERZn, ERAs, and ERCd, with r values ranging from 0.81 to 0.92 (see Table 3). PC 2 loadings produced ERNi and ERCu as the dominant ER-PTEs, accounting for 23% of the explained variance with corresponding r values = 0.92 and 0.86, respectively. The final PC loadings (PC3) also explained 19% of the variance, accounting for r values r = 0.78 for ERMn and r = 0.87 for ERCr.

The computation of the source distribution in soil has gained popularity within the field of soil science and EPA. PMF is one of the useful multivariate statistical software programs used to estimate the source apportionment of most areas. Considerable research, including Chen et al. (2015), Tao et al. (2017), Agyeman et al. (2021) and Hossain Bhuiyan et al. (2021b), has applied this tool in source analysis to determine the contribution of PTEs in soil. To control the residual matrix, it was necessary to reduce the minimum Q value to enhance accuracy and assurance. The system ran 20 times and run 3 was the relevant factor selected for the factors loading discharged. Three factors were released for both receptor models (ER-PMF and PMF), indicating the various percentage contributions or the PTE percentage contributions ascertained in the source distribution analyses. For PTEs to prevail over a factor, a minimum of 44.5% or higher percentage contribution must be attained to be selected as a dominant element. Factor 1 was dominated by Cu (64.6%) and Ni (71.1%) for the ER-PMF receptor model and by As (77.5%) and Cd (44.7%) for the PMF receptor model (see Table 4). Factor 2 was highly influenced by Cr (66.20%) and Mn

Table 3
Correlation between the ER-PTE (ecological risk-potential toxic element) values.

ER for PTEs	PC 1	PC 2	PC 3
ERMn	0.29	0.17	0.78
ERNi	0.10	0.92	0.10
ERPb	0.90	0.16	0.17
ERZn	0.81	0.41	0.10
ERAs	0.91	-0.08	0.19
ERCd	0.92	0.15	0.22
ERCr	0.11	0.14	0.87
ERCu	0.16	0.86	0.24
Eigenvalues	3.29	1.85	1.54
% Variance explained	41.00	23.00	19.00
Cumulative % total			83.00

Source apportionment

Table 4
The source percentage contribution for each factor loading of the ER-PMF receptor model.

	F1%	F2%	F3%		F1%	F2%	F3%
	ER-PMF				PMF		
As	24.30		75.70		77.50	7.70	14.90
Cd	29.10	25.10	45.90		44.70	30.70	24.60
Cr	13.00	66.20	20.80		1.10	43.90	55.00
Cu	64.60	32.40	2.90		20.40	72.60	7.00
Mn	0.10	71.40	28.50		0.20	34.30	65.60
Ni	71.10	28.90	0.10		22.10	76.60	1.30
Pb	19.80	32.80	47.40		39.10	26.70	34.20
Zn	36.50	28.40	32.10		36.10	43.80	20.10

Table 2
Ecological risk correlation matrix.

	ERMn	ERNi	ERPb	ERZn	ERAs	ERCd	ERCr	ERCu
ERMn	1.00							
ERNi	0.24	1.00						
ERPb	0.42	0.21	1.00					
ERZn	0.38	0.45	0.83	1.00				
ERAs	0.38	0.07	0.75	0.62	1.00			
ERCd	0.43	0.30	0.85	0.78	0.90	1.00		
ERCr	0.49	0.27	0.28	0.27	0.25	0.34	1.00	
ERCu	0.40	0.69	0.35	0.44	0.16	0.31	0.29	1.00

(71.40%) in the ER-PMF receptor model, and Cu (72.6%) and Ni (76.6%) likewise dictated factor 2 in the PMF receptor model. The final factor (factor 3) in the ER-PMF receptor model was monopolized by As (75.7%), Cd (45.9%) and Pb (47.4%), and similarly, in the PMF receptor model, Cr (55%) and Mn (65.6%) predominated factor 3.

4.2. Pollution indices assessment

The estimated single pollution index showed varying pollution levels for each PTE (see Table S1). Unlike Cd, which exhibited pollution levels at all sample locations, Mn, Cr, Ni, Pb, Zn, Cu, and As showed some pollution levels at some locations. Of the (115) samples analyzed, 76 samples exhibited low concentrations for Mn, 9 for Ni, 63 for Pb and Zn, 24 for As, 3 for Cd, 2 for Cr and 7 for Cu. However, some of the samples reported moderate levels of pollution, such as Mn (16), Pb (4), Zn (7), As (2) and Cd (26). Cadmium displayed high and very high levels of contamination at 56 and 30 locations, respectively. Similarly, Mn, Pb, Zn and As also showed a high pollution level at a single sampled spot. Arsenic displayed a very high-level pollution level for 3 sampled sites, similar to Pb at a sampled location.

4.3. Potential ecological risk index

The estimated potential ecological risk index of PTEs investigated by ERs and RIs is shown in supplementary table S2. Most of the PTEs investigated, such as Mn, Ni, Pb, Zn, Cr, and Cu, exhibited a slight ecological risk entirely in the study area. Cadmium, on the other hand, posed a moderate to extremely high ecological risk, with 15 locations posing a moderate ecological risk, 77 posing a high ecological risk, 19 posing a relatively high ecological risk, and 4 posing an extremely high ecological risk. The risk level of the study area computed revealed that 64 of the locations had a slight risk level, while 44 of the locations had a moderate risk level. In addition, 6 of the total sampled locations also displayed high risk-prone areas, while a sampled location also fell within a quite strong risk location.

4.4. Spatial analysis

The spatial distribution of some of the PTEs showed the same distribution patterns. These were As and Cd, as well as Pb and Zn (see Fig. 2). The northeastern and southwestern parts of the spatial distribution map exhibited high levels of As and Cd, while the northwestern and southwestern parts indicated deficient levels of As and Cd. Similarly, Pb and Zn also displayed hotspots in the northeastern and southwestern parts of the map, with a moderate spatial pollution distribution level in the southwestern and northwestern parts of the map. The other PTEs, such as Cr, Cu, and Ni, showed similar hotspots in the northeastern part of the map and exhibited hotspots in the northwestern part of the map. Manganese, chromium and copper also displayed hotspots in the southwestern part of the map.

5. Discussion

5.1. Sample descriptive statistics

The concentration of PTEs under investigation compared to the same UCC PTEs in Table 1 suggested that Cu, Cd, As, Pb and Zn of the current studies were higher than the respective UCC PTEs. Similarly, the mean concentrations of Mn, Pb, Zn, and Cd from the WAV and EAV reported by Kabata-Pendias (2011) are found to be lower when compared to the mean concentrations of the respective PTEs in the current study. Comparing the mean values to the local background values of the following PTEs: Cd (0.2 mg/kg), ZN (80 mg/kg) and Mn (545 mg/kg) reported by Němeček, & Podlesakova (1992) suggests that their mean concentration is lower than the current respective PTEs in the current study. However, the following PTEs: As (10 mg/kg), Cr (70 mg/kg), Cu (25 mg/kg), Ni (30 mg/kg), and Pb (50 mg/kg) mean of concentrations of the background values reported Němeček, & Podlesakova, (1992) were found to be higher than the respective PTEs in the current study. Gholizadeh et al. (2015), reported the mean concentration of the PTEs such as Cu, Mn, Cd, Pb and Zn in six study areas including Pokrok [Cu 13.76 mg/kg, Mn 599.40 mg/kg, Cd 0.27 mg/kg, Pb 18.43 mg/kg and Zn 25.26 mg/kg], Radovesice [Cu 14.20 mg/kg, Mn 541.30 mg/kg, Cd 0.17 mg/kg, Pb 13.70 mg/kg and Zn 21.98 mg/kg], Březno [Cu 14.37 mg/kg, Mn 680.90 mg/kg, Cd 0.16 mg/kg, Pb 14.17 mg/kg and 41.50 Zn mg/kg], Merkur [Cu 12.22 mg/kg, 590.00 Mn mg/kg, Cd 0.16 mg/kg, Pb 17.53 mg/kg and Zn 13.56 mg/kg], Pruné řov [Cu 15.81 mg/kg, Mn 552.60 mg/kg, Cd 0.11 mg/kg, Pb 14.38 mg/kg and 26.83 Zn mg/kg] and Tumerity [Cu 15.03 mg/kg, Mn 753.10 mg/kg, Cd 0.12 mg/kg, Pb 12.25 mg/kg and Zn 25.61 mg/kg] proximate to mining industries compared to the mean concentration of the respective PTEs in the current study, which suggested that the PTEs' mean concentrations in the current study are higher, except for Mn (Tumerity). Similarly, Weissmannová et al. (2019) reported the total median concentration of Mn (1370.92 mg/kg), Pb (37.71 mg/kg), Zn (204.56 mg/kg), Cd (0.21 mg/kg), Cr (21.11 mg/kg), and Cu (17.46 mg/kg) in industrial affected soils by coal mining and metallurgy in Ostrava, located in the Moravia–Silesian Region of the Czech Republic closer to the current study area, and comparing their total median concentration to the current study area suggests that Cd, Cr, and Cu of the current study area are higher.

The distribution of the data skewness varied considerably due to the difference in the data values analyzed from each sampled location. The skewness of the data distribution revealed that, apart from Mn, all the calculated skewness values were greater than + 1. According to Chandrasekaran et al. (2015), if the skew value is greater than + 1, it can be concluded that the distribution is irregular. The data being analyzed is very anomalous, skewed in the right direction, and leptokurtic. The calculated standard deviation values indicated that the SD values were higher due to the high concentration of PTEs with high variable heterogeneity in the study area. The coefficient of variation (CV) showed the degree of heterogeneity in PTE concentrations in the soil. Therefore, if the CV is $\leq$ 20%, it is inferred to be low variability, 21%$<$CV $\leq$ 50% is said to be moderate variability, 50% $<$CV $\leq$ 100% indicates high variability, and if the CV is greater than 100%, it is thus considered to be exceptional variability. The CV of the PTEs in the study area indicated that Mn, Cr, Ni, Cu and Zn had accrued 37.10%, 33.01%, 41.97%, 44.27% and 40.31%, respectively. This suggests that Mn, Cr, Ni, Cu, and Zn have a moderately varied distribution, and hence a relatively homogeneous distribution. However, Pb, Cd and As showed high variability, with CV values of 54.68%, 55.16% and 92.96%, respectively. These results clarified that Pb, Cd and As are nonhomogeneous and that their pollution is motivated by anthropogenic activities in the study area.

5.2. Multivariate analysis

The relationship between the ER-PTEs suggested a nexus between the elements, which indicates that the elements are highly related and might share a close source. ER-PTEs with r = 0.75 or higher, such as ERPb and ERZn, share a more significant relationship and pose a higher metallic connection. On the other hand, ER-PTEs with r = 0.50–0.75 share a relatively moderate association, and their metallic connections are moderately bonded.

The source of the PTEs was established using the estimated ER values, and the results indicated that ERPb, ERZn, ERCd, and ERAs from PC1 were largely anthropogenic origins. The PC1 clustered PTEs may share the same or similar pattern of occurrence in the environment. Their elevation beyond the standard threshold may be due to vehicle traffic, industrial activities, atmospheric deposition, intensive farming, and urbanization. However, the results also explained that PC2 and PC3 ER-PTEs, such as ERMn, ERCu, ERNi and ERCr, are principally geogenic

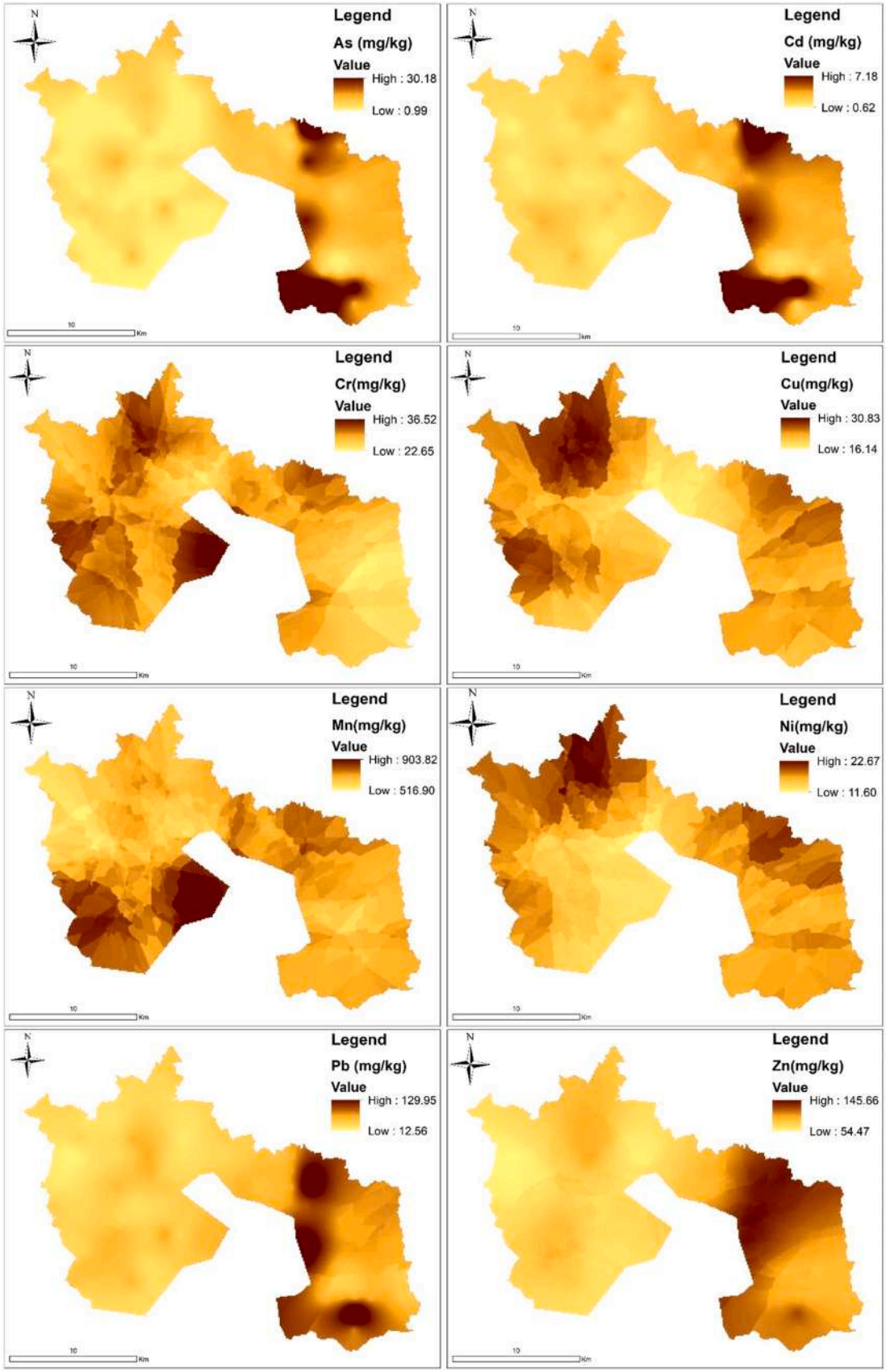


Fig. 2. Spatial distributions of PTEs using ordinary kriging interpolation.

elements. This is consistent with Borůvka et al. (2005) report that the positively associated PTEs within the same PC and their projected angle reveal their source of occurrence. The elevation in concentrations of PTEs from geogenic origin, compared to UCC, WAV, and EAV (see Table 1), might be due to a boost in anthropogenic activities. Forty percent of the land in the study area is engaged in intensive farming, and the district is home to the metal and steel industries. These human activities pose an ecological threat to the environment. The ensemble in PC2, whose primary source shows that the environment is primarily natural rather than anthropogenic, also implies that anthropogenic activities significantly support Ni and Cu. This is congruent with other studies confirming that agricultural soil has been improved by Ni-based fertilizers and Cu-based fertilizers (Harasim, Filipek, 2015a, 2015b; Agyeman et al., 2021). Other tenants, such as industrial activities, may also provide a boost (Harasim, Filipek, 2015a, 2015b). Clusters in PC 3 further suggested that the excess Mn and Cr in environmental soil have a geogenic origin and may be exacerbated by Mn and Cr buildup above the usual ecological threshold due to the steel industry and agriculture activities (Keshavarzi and Kumar, 2020).

5.3. Source distribution by receptor models

Factor 1 loadings in ER-PMF were dominated by Cu and Ni, controlling 51.89% of the total share of factor 1 loadings in ER-PMF, whereas in PMF it was predominated by As and Cd, controlling 50.66% of the total share of factor 1 loadings in PMF (See Figs. 3 and 6). The primary sources of Ni and Cu in the ER-PMF receptor model of factor 1 might be attributable to geological sources rather than anthropogenic sources, as supported by PCA analysis. The source of As and Cd in factor 1 of the PMF receptor model is attributable primarily to an anthropogenic source. Even though Ni and Cu of factor 1 loading of ER-PMF are more from geogenic sources, the spatial distribution map of Ni and Cu in Fig. 2 highlights hotspots in the southwestern and northwestern areas of the map that are active farmlands, implying that the anthropogenic source is augmenting the geogenic source. Furthermore, Huang et al. (2021) reported that the geographical empirical analysis performed can explicitly show local conditions during research, which is thought to be a highly efficient approach to determining pollution hotspots and investigating pollution sources. Literature has shown that Ni is primarily of geogenic origin (Antić-Mladenović et al., 2011; Hseu et al., 2017; Li, Öztürk and Dengiz et al., 2020, 2020). Their elevation in the study area suggests that anthropogenic activities such as nickel-based fertilizer for agriculture. Copper is regarded as one of the seven micronutrients required for plant growth, and 5–30 mg kg1 of Cu in crop tissues is generally considered adequate (Adriano, 2005). Cu has been widely

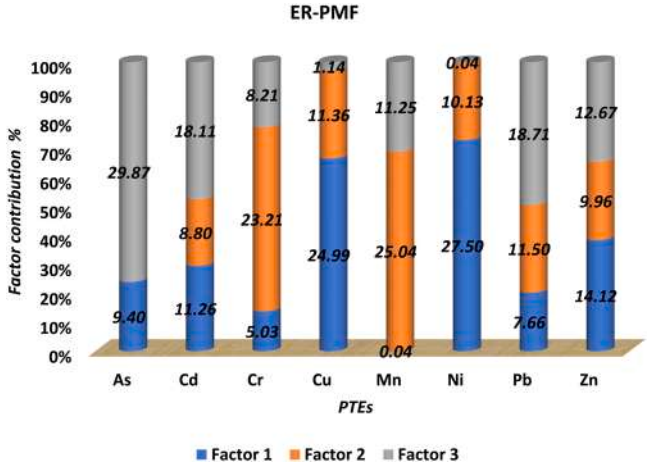


Fig. 3. Percentage proportion of each factor in each PTE of the ER-PMF receptor model.

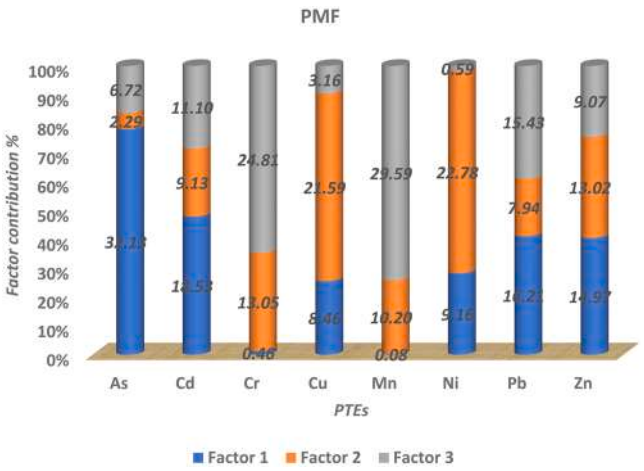


Fig. 4. Percentage proportion of each factor in each PTE of the PMF receptor model.

used as a fungicide, particularly in vineyards, for many decades (Ballabio et al., 2018). On the other hand, the sources of As and Cd in factor 1 of the PMF receptor model are attributable to anthropogenic sources, which is confirmed by the PCA analysis in this study. The spatial distribution of As and Cd in the study area exhibits hotspots in the northeastern and southeastern areas of the map where there are active agricultural activities and the presence of steel and metal industries. Arsenic's inclusion in animal feed and its application in agro-related products such as fertilizer, insecticides, herbicides, and fungicides account for its elevation in the soil study area. When combined with other PTEs, such as Zn, arsenic is an agronomically related and potent pair of elements that can be primarily found in agricultural products such as pesticides, farmyard manure, and fertilizers. Yang et al. (2017), Liang et al. (2017) and Hu and Cheng (2013) reported that PTEs such as As contribute significantly to soil pollution through the use of agro-based fertilizers, for instance, phosphate fertilizers and ammonium phosphate fertilizers, which lead to the elevation of As. Cadmium and arsenic are correlated due to their existence in chemical fertilizers such as phosphate fertilizers. This is consistent with reports made by Robertsa (2014), Corguinha et al. (2015), and Wang et al. (2016) confirming that phosphate fertilizers contain PTEs such as As and Cd. In addition, Mamut et al. (2018) and Shao et al. (2016) likewise posited that Cd enrichment in the soil might be primarily attributed to the use of fertilizers and pesticides to enhance crop productivity. The current results of this study are corroborated by Fei et al. (2019) report, emphasizing that excessive Cd concentrations in Shanghai agricultural soils should be attributed to Cd-rich agro-related practices.

Factor 2 of the ER-PMF receptor model was highly influenced by Cr and Mn, controlling 52.62% of the total share of the factor 2 loadings in ER-PMF as presented in Fig. 3, whereas in the PMF it was dictated by Cu and Ni, accruing 44.37% of the total share of factor 2 loadings in PMF. The source of Cr and Mn in the study area is primarily from geogenic sources, which is supported by the PCA analysis in this study. However, the spatial distribution of Mn and Cr on the map (Fig. 2) highlights hotspots in the southwestern and northwestern areas of the map. Human activities within the community foretell the involvement of associated anthropogenic activities that enhance the geogenic source of the PTEs to exceed the allowed level. Lv and Wang (2018), Gao and Wang (2018) and Cui et al. (2018) acknowledged that the Cr and Mn sources of occurrence are primarily natural sources. Similarly, studies by Men et al. (2018),Chen et al. (2018a), Lv (2019), and Ma et al. (2018) have also confirmed the same results but further added that the excesses of Cr and Mn in the soil beyond the tolerable limits may be due to human activities augmenting the geogenic source, especially in agro-based communities and industrial-based environs. Chromium elevation in some areas in the

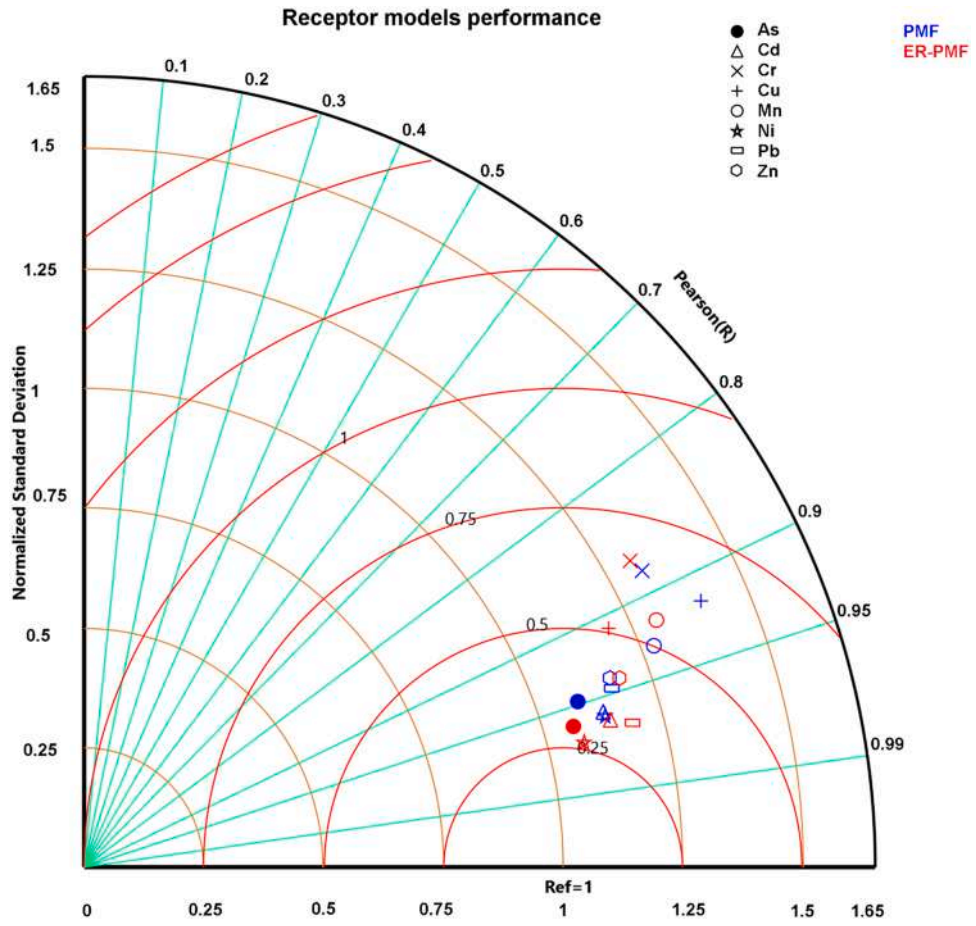


Fig. 5. Taylor's diagram comparing receptor model performance. The semicircles in red on axis x indicate the root mean square error values. The -90^0 curves moving from the y-axis to the x-axis denote the standard deviation values. The straight green line originating from the origin is the Pearson correlation value, and Ref= 1 is the reference value.

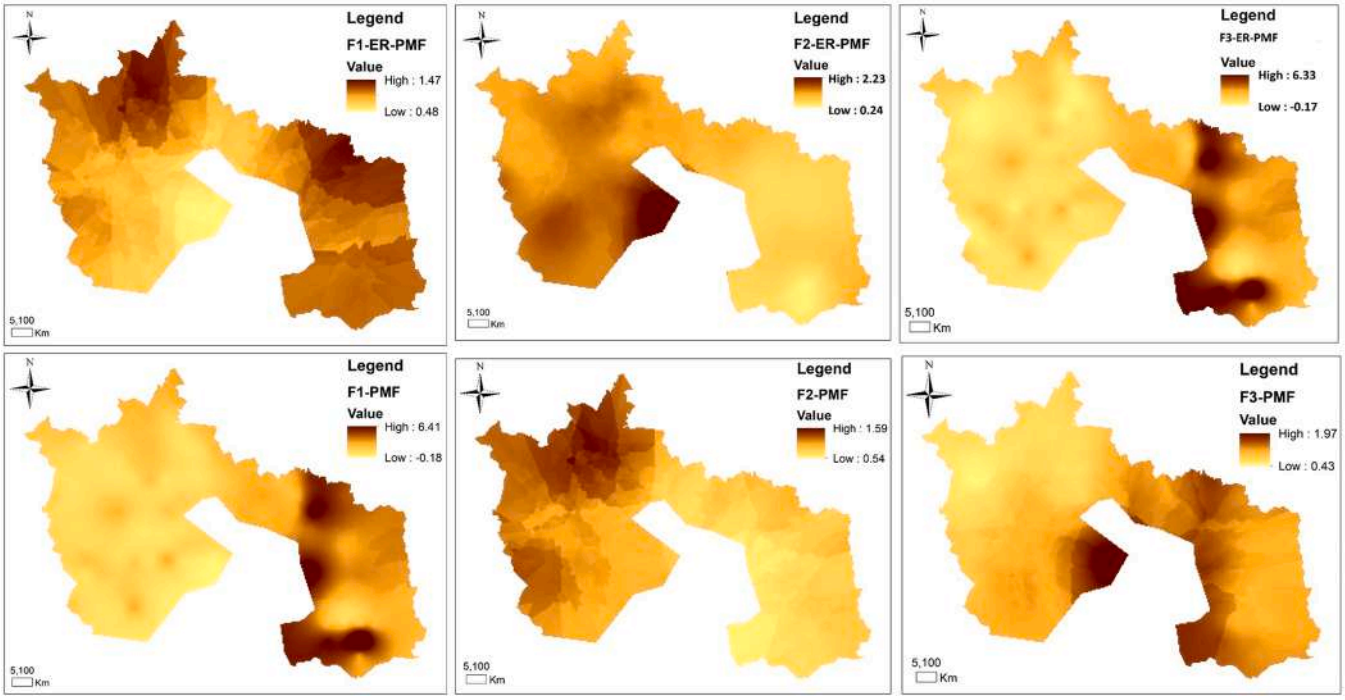


Fig. 6. Spatial prediction of receptor model factor contributions.

study areas (southwestern and the northwestern regions of the map) may result from a collaborative effort between a geogenic origin and anthropogenic sources such as active farming in the study area. Cr levels have risen in recent decades due to crop intensification, excessive use of pesticides and fertilizers, and agronomical reuse of treated wastes and byproducts of industrial activities (Gattullo et al., 2020). Zeremski-Škorić et al. (2010) reported that the occurrence of Cr has been detected in fertilizers used, particularly organic composts and phosphate fertilizers. Manganese, on the other hand, is a good metal utilized in the steel industry to manufacture ferromanganese steel. Mn elevation in the soil may be attributed to the steel industry in the study area.

Factor 3 of the ER-PMF receptor model was dominated by As, Pb, and Cd, accounting for 61.63% share of factor 3 loadings in ER-PMF, whereas Cu and Ni dominated in PMF, accounting for 44.37% share of factor 3 loadings in PMF. The sources of pollution of As, Pb, and Cd in factor 3 of ER-PMF are primarily anthropogenic, whereas the Cu and Ni factors 3 of PMF are more geogenic, as confirmed by the PCA analysis estimated and the spatial distribution map of the hotspots highlighted on the maps. The high level of Pd in the soil may be attributed to an anti-knock agent; vehicle tires, vehicle exhaust fumes, and chimneys account for a high concentration of Pb in some parts of the study area. Chalking, leaded paints, industrial smelting, alloying, and scraping are other factors contributing to Pb pollution in the soil. Gan et al. (2018) and Jin et al. (2019) reported that vehicular traffic, antiknock agents, car tires, break wear, and other vehicle-based processes are responsible for increasing the concentration of Pb in the soil in most communities. Furthermore, Li et al. (2013) iterated that pollution might originate from leaded gasoline usage. The sources of the other PTEs identified in factor 3 loading have already been discussed in the previous sections.

5.4. Model performance and spatial distribution of factors

The receptor models were subjected to accuracy and validation assessment via a machine learning algorithm (random forest). Both were cross validated using the coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE). The assessment was performed on the PTE understudy whose source apportionment was duly estimated using ER-PMF and PMF approaches. The computed R^2 for the PTEs from both models demonstrated that the ER-PMF model had a more significant percentage accuracy level for more PTEs than the PMF model (see Table 5). Of the 8 PTEs assessed, 5 (As, Cd, Ni, Pb and Zn) of the PTEs obtained a higher accuracy level (R^2) in the ER-PMF approach than in the PMF approach. According to Li et al. (2016), John et al. (2020) and Kebonye et al. (2021), the closer the R^2 value is to 1, the better the prediction accuracy. Molinaro et al. (2005) outlined that the critical process in presenting results is determining the error rate or generalizability of the chosen model. Therefore, the marginal error of the receptor model approaches was also computed alongside the accuracy assessment level. The RMSE of both receptor model approaches showed that the ER-PMF calculated errors were far lower than those of PMF. The error level for six of the eight PTEs (Cr, Cu, Mn, Ni, Pb, and Zn) out of the ER-PMF receptor model was significantly reduced (see Table 5). The overall average of R2, RMSE, and MAE values estimated for the receptor models revealed that ER-PMF had a high R2 average (0.938) with a low RMSE (2.634) and MAE average (1.559), whereas PMF had a high R2 average (0.936) with a higher RMSE (13.116) and MAE average (8.208). This means that in source apportionment, ER-PMF can identify sources with greater accuracy and less error than PMF.

Similarly, the computed MAE error level corroborated similar results as the RMSE and revealed that the closer the RMSE and MAE computed error was to zero, the better the model approach. As a result, the cumulative performance of the receptor models (R2, RMSE, and MAE) revealed that the ER-PMF approach outperforms the parent receptor model (PMF) in source apportionment estimation. Guan et al. (2019) compared three receptor models (PMF, UNMIX and grouped principal component analysis/absolute principal component scores (GPCA/APCS)) and based on the estimated R^2 values; the authors concluded that the GPCA/APCS receptor model was optimal. Similarly, Salim et al. (2019) compared PCA-MLR and PMF, and the authors used R^2 to determine the receptor model that is more reliable with high model efficiency; PMF was found to be optimal. Additionally, Salim et al. (2019) applied the Nash-Sutcliffe efficiency and quantified the percentage error, which has been previously used by Moriasi et al. (2007) and Yang et al. (2013b), to ascertain the receptor model with minimal percentage error while simultaneously optimizing efficiency. In this study, instead of the Nash-Sutcliffe efficiency, we applied R^2, RMSE and MAE, which are also widely utilized to determine model efficiency and error margin in modelling approaches or statistical evaluation. In a different case, Huang et al., (2018c) applied a modified receptor model to compute PTEs in the soil. The author outlined that principal component analysis-multiple linear regression with distance (PCA-MLRD) compared to PMF and APCS-MLR showed that PCA-MLRD was optimal. This suggests that the hybridization of the existing receptor model tends to yield better results, as similarly performed in this recent study. Furthermore, hybridizing PMF with pollution assessment indices (ecological risk) improves source apportionment efficiency while significantly reducing errors.

The receptor models were further subjected to additional performance assessment through the Taylor diagram (see Fig. 5) (Taylor, 2005). We have two variables in different colors on the Taylors diagram, red representing the ER-PMF receptor model and blue representing the PMF receptor model. Both receptor models have the same variables: the PTEs inscribed in the diagram with unique symbols but differentiated by color in the diagrammatic representation. The results presented in the Taylor diagram suggested that all the receptor models yielded equivalent values of normalized standard deviation spanning between 0.75 and 1. Furthermore, the results showed the ratio of the standard model deviation to the standard reference value deviation. Based on the Taylors diagram, it can be interpreted that out of the 8 PTE variables assessed for both models, the hybridized receptor model that is ER-PMF is optimal. Five (Cu, Pb, As, Cd and Ni) out of the 8 PTEs from the ER-PMF receptor model showed superior performance against similar PTEs from the PMF receptor model. However, Zn assessment was nuetral for both models. The cumulative performance of both receptor models from the Taylor diagram indicates the higher efficiency of the hybrid receptor model ER-PMF to the parent model PMF.

The spatial distribution map of the factor scores of both receptor

Table 5
Performance of the receptor models.

	As			Cd			Cr			Cu		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
ER_PMF	0.961	1.596	0.771	0.963	17.344	10.388	**0.872**	**0.128**	**0.085**	**0.910**	**0.447**	**0.279**
PMF	**0.948**	**1.27**	**0.594**	**0.958**	**0.246**	**0.145**	0.883	3.599	2.35	0.918	3.413	2.164
	Mn			**Ni**			**Pb**			**Zn**		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
ER_PMF	**0.918**	**0.174**	**0.115**	**0.970**	**0.22**	**0.152**	**0.967**	**0.872**	**0.502**	**0.943**	**0.292**	**0.181**
PMF	0.932	79.003	49.925	0.960	1.59	1.074	0.947	5.305	2.645	0.941	10.498	6.765

model approaches was mapped using ordinary kriging interpolation, as represented in Fig. 6. Factor 1 for the ER-PMF approach showed hotspots in the northwestern and northeastern spatial distribution maps. This result suggests that vigorous agricultural activities and industrial activities complementing the geogenic sources that pollute the soil in the study area. The PMF factor 1 spatial distribution map hotspot was envisaged at the northeastern and southeastern parts of the map. This will be attributed to the steel plant and agricultural activities in the southeastern part of the study area. The steel plant is located between the northeastern and eastern parts of the study area map. The factor 2 spatial distribution map of both receptor models was visualized within the northwestern and southwestern parts of the map. Both maps displayed hotspots with ER-PMF factors showing high hotspots in the southwestern area and moderate hotspots on the northwestern side. Nevertheless, PMF factor 2 showed a dotted hotspot in both the northwestern and southeastern parts of the map. The factor 3 spatial distribution map of the ER-PMF receptor model shares a similar hotspot with the factor 1 PMF spatial distribution due to the similar elements dominating the factor loadings (As and Cd). This implies that factor 1 of PMF and factor 3 of ER-PMF are anthropogenically oriented elements, validating that those continual human activities in the study area pollute the soil. Factor 3 of the PMF approach hotspots can be seen in small portions of the southwestern and southeastern spatial distribution maps. Hotpots can be attributed to multiple sources, such as the steel industry, atmospheric deposition, and agronomic practices (that is, the application of agrochemicals).

5.5. Potential ecological risk index

The potential ecological risk estimated based on Hakanson theory suggested that the study area's ecological risk level was minimal for all PTEs (Mn, Ni, Pb, Zn, Cr, Cu) except As and Cd. Even though the environment appears to be primarily a moderately risky area, a few sampled locations indicated a moderate level of environmental risk (As). Cadmium ecological risk assessment revealed that communities within the study area's enclave are at risk of Cd intake by ingestion, dermal ingestion, or inhalation. Much literature on the health risk of Cd to humans, flora, and fauna has indicated devastating repercussions on humans, whether ingested, dermal or inhaled. An array of literature on potential carcinogenic- and noncarcinogenic-related diseases, as well as ecological threats posed by Cd, are captured copiously by Åkesson et al. (2014), Wang et al. (2015), Wu et al. (2016), Yu et al. (2017), Satarug et al. (2017b), Satarug et al. (2017a) and Qasemi et al. (2019). The risk index (RI) assessment found that the toxicity level for such locations is high in the northeastern and southeastern parts of the RI kriging spatial distribution map (Fig. 7). The northeastern region of the map is home to a variety of industries, including the steel industry. In contrast, the southwestern hotspot may be attributed primarily to agricultural practices prevalent in the communities. Chen et al. (2018b) reclassified the RI grading standard, stating that if RI< 70, it denotes slight risk, 70 <RI ≤ 140 moderate risk, 140 <RI ≤ 280 strong risks, 280 <RI ≤ 560 quite strong and RI greater than 560 signifies extreme risk level. When the Chen et al. (2018) classification for RI is compared to the Hakanson (1980) classification, if Chen's classification is applied, the toxicity level of the study area is higher. Thus, the hotspots range from moderate to extremely risk-prone environments based on Chen's classification, i.e., risk level = 113 observed locations out of 115, or 98.26%, and based on Hankason's classification, i.e., risk level = 51 observed locations out of 115, or 44.35%.

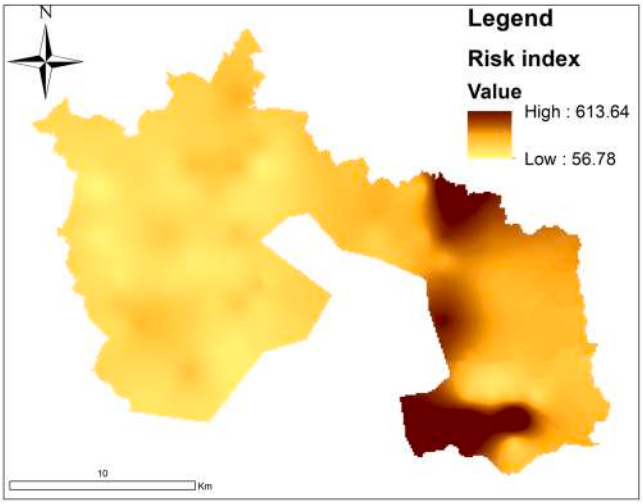


Fig. 7. Spatial distribution of risk index (RI) values using ordinary kriging.

5.6. Spatial analysis

The spatial distribution analysis of PTEs in urban and peri-urban areas of the study area identified some hotspots of PTEs in some areas, suggesting elevated PTEs in some parts of the spatial distribution map. Reimann (2005) and Reimann et al. (2008) proposed that creating a geochemical spatial distribution map is a very valuable tool that helps to extrapolate much information from the area. The spatial distribution maps (Fig. 2) indicated that As and Cd share the same distribution pattern with hotspots in the northeastern and southeastern maps. The steel and metal industries are located within the northeastern part of the study area, with moderate farming activity within that enclave. The land-use type visible in the southeastern part of the hotspot area is essentially agricultural land. Their hotspots in the northeastern and southeastern parts of the spatial distribution map might be due to the steel industry and agriculture. Cd's anticorrosive nature makes it easier to use as a coating agent for steel, brass, iron, and aluminum. According to Lambert et al. (2011), steel industries are significant sources of pollution to soils due to production activities such as scrap metal melting of metalloids in a furnace to determine which iron is recycled, and the pollution may linger in the soil for a while even if the steel industry is closed. Chromium, copper, manganese, and nickel shared relatively similar spatial distribution patterns at some sampled locations. Hotspots were exhibited by Cr, Cu and Ni in the northeastern and northwestern parts of the spatial distribution map (see Fig. 2). The steel industry and metal works are located in the northeastern part of the study area, which might be responsible for the elevation of the PTEs beyond the standard threshold. These PTEs (Cr, Cu and Ni) are very useful in the steel industry because of their alloy formative abilities and mechanical properties for steel, providing an anticorrosive property that keeps the metal impervious to corrosion (Satyendra, 2014a; Satyendra, 2014b; Blog, 2020).

On the other hand, the hotspot envisaged on the northeastern part of the PTEs (Cr, Cu and Ni) can be attributed to anthropogenic assistance such as atmospheric deposition and intensive agro-related activities in the area. This is consistent with a study carried out by Huang et al. (2019) reporting that anthropogenic sources such as fertilizers, atmospheric deposition, fungicides, the proximity of agricultural sites near industrial activities, sewage irrigation and plastic films can elevate the concentration of PTEs in soil. Zinc and lead also showed the same spatial distribution pattern and hotspots in the northeastern and southeastern map parts. This result might be ascribed to the vehicular traffic, steel industry, atmospheric deposition, and agricultural production in the southeastern area. Phosphatic fertilizers applied to the soil unintentionally potentially add Pb to the soil. However, manures and biosolids from livestock manure and compost applied to urban and peri-urban land or soil increase the concentration of PTEs such as Pb and Zn (Basta et al., 2005). Metal smelting in steel industries accounts for an extensive upsurge in Pb and Zn levels in the soil that is injurious to the health of the ecology (Wuana and Okieimen, 2014).

The selection of the best fitted variogram models for the spatial

prediction of the PTEs using ordinary kriging is presented in supplementary material Table S 5. The spatial dependence of the PTEs in the study was determined by assessing semivariance, which depicted the estimated index that was predicted with a spherical semivariance method with nugget effects. The nugget sill ratio was calculated using the Cambardella et al. (1994) criteria for spatial dependency. According to the criteria reported by Cambardella et al. (1994) a nugget sill ratio of 25% indicates that spatial prediction exhibited strong spatial dependence. However, if the ratio is between 25% and 75%, it implies a moderate spatial dependence, and when the ratio is above 75%, it suggests a weak or poor spatial dependence. The spatial dependency of the nugget to sill ratio $\mathbf{(C_0/C_0 + C)}$ was expressed in percentiles. Based on the estimated nugget to sill ratio, it was clear that the PTEs As (0.00), Cd (0.00), Cu (0.00), and Pb (2.29) had strong spatial dependence. Other PTEs, such as Cr (28.71), Mn (70.01), Ni (54.51), and Zn (36.37), on the other hand, exhibited moderate spatial dependence.

5.7. Uncertainty analysis

The effectiveness of the receptor models was subsequently evaluated by calculating the uncertainty of the receptor models in the source distribution and risk assessment. To validate the uncertainty evaluation of the receptor model results, Table S3 shows the base model displacement technique and the bootstrap base model approach. The dependability and resilience of the receptor models were emphasized in the precision of the results; however, displacement (Disp) was judged to be a crucial step for the screening solution. The rationale of the models was that the frequency of swaps factors depending on the model solution remained inversely proportional to the degree of swaps factors dependent on the model solution (Brown et al., 2015). The DISP swap results per receptor model in the present study were 0.05–0.07 for ER-PMF, while the PMF of the receptor models had modest swaps ranging from 0.03 to 0.06 (See Table S4). The DISP swap performance in the ER-PMF was well characterized with negligible data inaccuracies, while the other receptor model (PMF) efficiency was normal. Regardless, a swap appears to provide conditional random defects, well-defined receptor model solutions, and a prospective rotational uncertainty that might or may not show up in model performance (Paatero et al., 2014). The calculated DISP results were appropriate because of the lesser percentage change (i.e., less than 1) in the Q (DISP percent dQ, see Table S 4), which is compatible with Brown et al. (2015); Wu et al. (2020); and Wang et al. (2019, 2021) report.

The bootstrap technique used in the current work was set to 100 iterations, with factor loading BS mappings of 100% for all receptor models (See Table S4). This signifies that in this study, the receptor models' reliability in delivering ideal results per factor loading was 100% per receptor model. The heterogeneity of the source species contributions and the box plot of the factors at each receptor model primarily focused on the DISP and the BS assessment are shown in Figures S 1–3. The relevance of the source profile ambiguity retrieved from the model run indicates the receptor model's box plot's coherence. In certain aspects, the base run values for PTE concentrations from the receptor models were beyond the interquartile range. For instance, in the ER-PMF receptor model, Mn, Ni, As, and Cr were discovered outside the interquartile range in factor 1, Mn, Zn, Pb, and Cr were found outside the interquartile range in factor 2, and Mn, Ni, Zn, As Cr and Cu were found outside the interquartile range in factor 3 (see Figs. S4–6). However, the following PTEs were discovered outside the interquartile range in the PMF receptor model: Ni, Pb, Zn, As, Cu, Cr, and Cd for factor 1, Zn for factor 2, and Cr, Zn, and Pb for factor 3. Considering the BS's influence on the observation and the concentration error computation addressed in the base run, a biased BS run could be attributable. Most of the PTEs in the factor loading were discovered beyond the interquartile range, according to Wang et al. (2021) and Qiao et al. (2021).

Table S3 shows the uncertainty interval ranges (i.e., minimum 5th and maximum 95th) as well as the base values of the base error calculation for each receptor model. The receptor models' interval ratios are crucial in computing the interval ranges predicated on the split of the uncertainty, which corresponds to the midpoints (Brown et al., 2015; Paatero et al., 2014). The DISP and BS interval ratios, which showed performance comparable for each PTE in each receptor model, were used to maintain the coherence of uncertainty results for each receptor model, as shown in Fig S3. PTEs with large interval ratios showed greater uncertainty (Hu et al., 2020); for instance, in Fig. S3, the Cd interval ratio of the ER-PMF receptor model was shown to be closer to the apex, i.e., 10^1 to 10^2 for factors 1 and 3. It also had a high interval ratio, 10^2 for factor 3, as it approached the apex. Cd on the ER-PMF interval ratio chart, on the other hand, showed a high interval ratio for factors 1–3 that was closer to the apex 100. The BS interval ratios measured for each receptor model were significantly higher than the DISP interval ratio, premised on the computed uncertainty for each receptor model. According to Wu et al. (2020), this might be classified as random errors, which account for a significant fraction of the overall uncertainty. When the hybridized receptor model ER-PMF was compared to the parent model PMF, it was revealed that for factors 1–3 (ER-PMF), models exhibited decreased interval ratios in the BS uncertainty interval (PMF). When comparing the DISP uncertainty interval ratio, the ER-PMF showed a smaller interval ratio in more of the PTEs discharged by the factor loadings than the parent model PMF. The interval ratios decreased random error and boosted efficiency by combining the PMF with pollution assessment indicators such as ER.

5.8. Geographically weighted regression analysis

Mapping the PTE concentration of the agricultural soil was performed using geographically weighted regression and geographically weighted regression cokriging (GWRCoK) (see Figs. 8 and 9). The approaches used to predict the spatial distribution of PTEs resulted in a variety of spatial distribution patterns, as did the techniques used to investigate the spatial relationship between terrain attributes and PTEs. The model (GWR and GWRCoK) was evaluated based on the prediction accuracy of the PTE distribution in the soil using R^2, RMSE, and MAE. The spatial distribution of arsenic and cadmium exhibited similar spatial distribution patterns for both approaches (GWR and GWRCok; see Figs. 8 and 9). The PTE distribution pattern was seen in the southeastern part of the study area, moving to the northwestern area in the anticlockwise direction. The maps exhibited moderate to high hotspots, but the GWRCoK map showed more intense hotspots than the GWR. The R^2, RMSE, and MAE also suggested that the efficiency of the prediction of Cd and As in the agricultural soil by GWRCoK was 0.945–0.961 compared to 0.636–0.713 for GWR. The error margins estimated based on RMSE and MAE were 1.272 and 0.749 for GWRCoK and 2.636 and 1.819 for MAE, respectively (see Table 6).

For the GWR spatial distribution map, the distribution pattern for Cr and Cu showed a similar distribution pattern, primarily in the southwestern to northwestern quadrants. Chromium had more hotspots that extended from the northeast to the southeast than Cu. The GWRCoK map for Cr and Cu was more concentrated in the southwestern to northwestern areas of the map, but Cu also showed a patch of a hotspot in the map's northeastern area. The prediction accuracy of both models for Cr and Cu was significantly low, from 0.063 to 0.075 and 0.071–0.393 for R^2 for GWR and GWRCoK, respectively. The error margins range from 8.986 to 9.617, 6.422–6.457, 2.021–2.466, and 1.346–1.477 for GWR and GWRCoK, respectively. Manganese and nickel from the GWRCoK spatial distribution map shared a similar pattern, but Mn showed more hotspots in the northwest. In contrast, Ni also displayed hotspots in the southwestern and southeastern regions. Based on the GWR spatial distribution map for Mn and Ni, Mn showed more patches of hotspots in the southwestern and southeastern regions than Ni. The prediction efficiency based on both approaches for Mn and Ni was low for both models, but Ni prediction in GWRCok was good (0.794). The error margin for Mn for both models was too high, and GWR was 2.42 times higher than

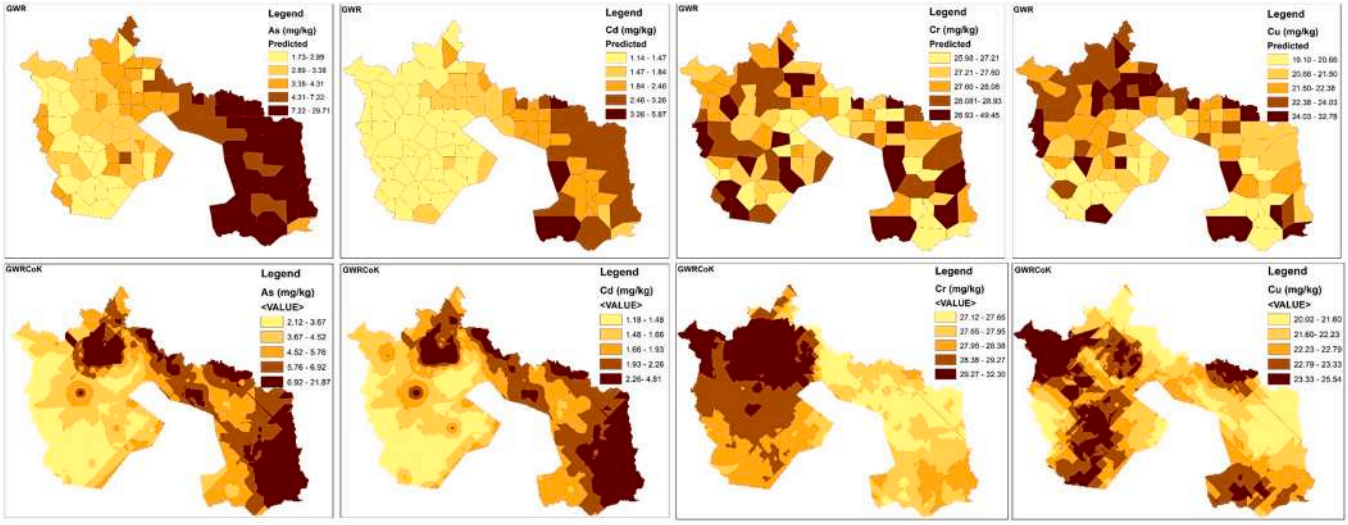


Fig. 8. shows the spatial distribution of PTEs using geographically weighted regression cokriging (GWRCoK) and GWR.

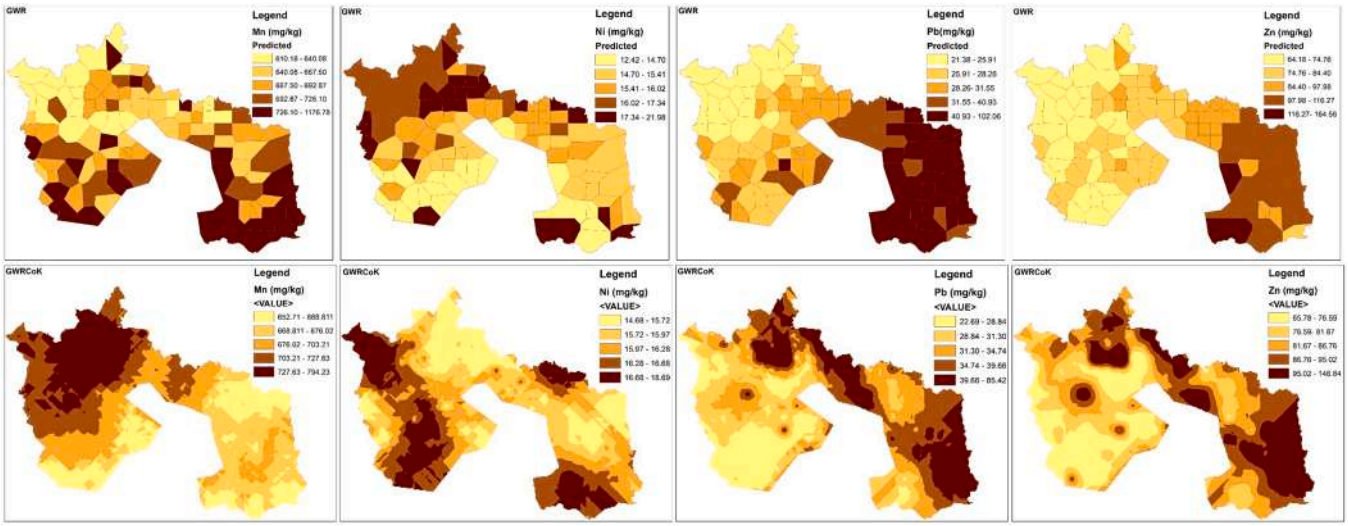


Fig. 9. shows the spatial distribution of PTEs using geographically weighted regression cokriging (GWRCoK) and GWR.

Table 6
shows the performance of the model used in the prediction of PTEs in soil.

	GWR			GWRCoK		
	R2	RMSE	MAE	R2	RMSE	MAE
As	0.713	2.636	1.819	**0.945**	**1.272**	**0.749**
Cd	0.636	0.608	0.359	**0.961**	**0.196**	**0.124**
Cr	0.075	8.986	6.457	**0.071**	**2.466**	**1.346**
Cu	0.063	9.617	6.422	**0.393**	**2.021**	**1.447**
Mn	0.107	243.943	186.396	**0.168**	**76.955**	**51.806**
Ni	0.041	6.462	4.761	**0.794**	**1.215**	**0.894**
Pb	0.425	13.981	7.824	**0.981**	**2.208**	**1.301**
Zn	0.293	28.753	19.613	**0.975**	**3.229**	**2.215**

GWRCoK. Zinc and lead likewise exhibited similar spatial distribution patterns with hotspots in the southeastern area of the map for both approaches, but Pb and Zn displayed hotspots in the northeastern region as well. GWRCok exhibited a goodness of fit for Pb and Zn, with R^2 values of 0.981 and 0.975, respectively, compared to 0.425 and 0.293 for GWR. The error margin for RMSE and MAE was relatively low, ranging between 1.301 and 2.208 for GWRCok and 19.613 and 28.753 for GWR. The magnitude of the error estimated by the GWR approach was 8.901 (RSME) and 8.94 (MAE) times greater than that of the GWRCok. This suggests that hybridizing cokriging with GWR yields better results than GWR. John et al. (2021) hybridized cokriging with Gaussian process regression and yielded optimal results. Numerous papers have hybridized GWR with ordinary kriging, such as Kumar et al. (2012), Wang et al. (2012), Ye et al. (2017) and Pereira et al. (2018). Ye et al. (2017) compared the performance of geographically weighted regression kriging (GWRK) with multiple linear regression kriging (MLRK) and ordinary kriging (OK), and the author concluded that hybridizing GWR with geostatistical algorithms such as OK yielded good results in predicting soil organic content is soil better than MLRK and OK. The author added that GWRK predicted SOC with less uncertainty and greater accuracy. This is consistent with the results obtained in this study. Similarly, Imran et al. (2015) used GWRK for growth and yield modelling in West Africa, concluding that GWRK is better than KEDLN (KED with a local kriging neighborhood) and regression kriging in general and that prediction uncertainty in GWRK was significantly reduced. This study demonstrated that the combination of cokriging and GWR can increase the efficiency of PTE prediction, thereby significantly reducing the error. This approach has undoubtedly optimized the channels of uncertainty provided by the sample selection design, empirical modelling techniques, and accessible covariate data sources,

as well as increased decision-makers' confidence by accounting for reliably calculated prediction ambiguity and enhanced efficiency level.

5.9. Advantages of ER-PMF over PMF

According to research findings such as Al-Anbari et al. (2015); Baran Jerzy Wieczorek Ryszard Mazurek Krzysztof Urban et al. (2017), pollution index preference is correlated to pollution threshold, source, and the ecological risk of PTEs. Assessing pollution thresholds utilizing varied pollution indices, including EF, Er, PI, and Igeo, gives the research scientist a rough inkling of the pollution degree as well as the corresponding source route of PTEs. Estimating ecological risk gives the research the fair idea of the pollution level of the research area and probable identify the source route based on the precise scale can outline a locality's environmental risk level. ER-PMF input source is the estimated ecological values that give the research a reasonable idea of the pollution source, whereas in PMF, UNMIX, and APCS-MLR receptor models' raw datasets are used, which means the source of the PTEs cannot be known until source distribution analysis is performed. ER-PMF analysis gives the researcher the validation based on the preliminary assessment made, whereas with PMF, the research is yet to determine the source distribution assessment made. The ER-PMF dataset obtained from ER estimation is focused on distinguishing between natural and anthropogenic pollution source processes, rendering it relatively easy to recognize pollution sources predicated on the precise scale utilized, whereas the PMF dataset needs to be analyzed to determine the natural and anthropogenic pollution source. Therefore, it easier and serves as a confirmatory analysis in ER-PMF than PMF. ER-PMF can apportion sources with high R2 and low error, whereas PMF can apportion sources with high R2 and a corresponding high error value. The use of ER allows for the easy recognition of impactful pollution sources, regardless of whether anthropogenic or geogenic (Gąsiorek et al., 2017; Z. Wang et al., 2015). ER-PMF have a relatively small DISP uncertainty interval than the parent model, indicating that random error is lowered and therefore performance is increased. In ER-PMF the dormant PTEs in every factor loading accrues higher percentage dominance per factor loading than in PMF. This increases the prediction efficiency in ER-PMF than in PMF. According to Paatero et al. (2014), the evaluated uncertainty in PMF assessment is to apply DISP intervals, and predicated on the result obtained, ER-PMF outperforms the parent model PMF. The estimation of ecological risk necessitates the use of a single pollution index or contamination factor, which necessitates the inclusion of geochemical background levels to provide a rough estimate of the amount of pollution caused by preindustrial sources. The significance of ER is that it provides an avenue for stakeholders to make decisions and manage natural resources while considering toxic levels, ecological sensitivity, and synergies between PTEs (Gąsiorek et al., 2017; Mazurek et al., 2017).

6. Conclusion

This study assessed source distribution using an ecological risk approach, estimated uncertainty assessment, and the application of geographically weighted regression cokriging for the prediction of the following PTEs: As, Cr, Cu, Cd, Pb, Mn, Ni, and Zn. The results indicated that the calculated risk index ranged from low to high according to the Hakanson categorization, but Chen's categorization increased the ecological risk from high to extremely high toxicity. The spatial assessment revealed varied hotspots for PTEs in peri-urban and urban areas, ranging from minimally spatially distributed to elevated hotspot areas. The steel industry, agrochemicals, fertilizer applications, vehicular traffic, and antiknock agents were all identified as potential polluters of the PTEs in the study by the receptor model. The evaluation of the receptor model efficiency and the magnitude of error computation revealed that when the ER-PMF and the PMF were compared, the ER-PMF outperformed the PMF in 5 (As, Cd, Ni, Pb, and Zn) of the 8 PTEs evaluated. Furthermore, the RMSE and MAE computed errors revealed that the hybrid receptor model ER-PMF margin of error was significantly reduced in six (Cr, Cu, Mn, Ni, Pb, and Zn) of the eight PTEs. The hybridized GWRCoK model outperformed GWR in predicting the distribution pattern of PTEs in the study area by improving the efficiency of PTE prediction, resulting in a significant reduction in error. The uncertainty assessment of the receptor models indicated that the DISP interval ratio of the hybridized ER-PMF model was less than that of the parent PMF model, and thus in the ER-PMF receptor model, random error that could arise in the DISP based on the DISP interval ratio was likely to be less than PMF. The cumulative performance of the receptor model indicated that ER-PMF was superior to PMF. Based on the comprehensive analysis, the ecological status of the study area revealed that the toxicity levels of some of the areas are potentially risky and have the tendency to pose a health-related risk to the people who live in the communities. As a result, practical measures should be implemented to mitigate the community's potentially risky and highlight risky areas.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This study was supported by an internal Ph.D. grant no. SV20–5-21130 of the Faculty of Agrobiology, Food and Natural Resources of the Czech University of Life Sciences Prague (CZU). Support from the Ministry of Education, Youth and Sports of the Czech Republic (project No. CZ.02.1.01/0.0/0.0/16_019/0000845) is also acknowledged. Finally, The Centre of Excellence (Centre of the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially risk substances of anthropogenic origin: comprehensive assessment of the soil contamination risks for the quality of agricultural products, NutRisk Centre).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.psep.2022.06.051.

References

Adimalla, N., Qian, H., Wang, H., 2019. Assessment of heavy metal (HM) contamination in agricultural soil lands in northern Telangana, India: an approach of spatial distribution and multivariate statistical analysis. Environ. Monit. Assess. 191 (4), 1–15.

Adriano, D.C., 2005. Trace Elements in Terrestrial Environments: Biogeochemistry, Bioavailability, and Risks of Metals, Vol. 860. Springer,, New York.

Agyeman, P.C., Kebonye, N.M., John, K., Borůvka, L., Vašát, R., Fajemisim, O., 2022. Prediction of nickel concentration in peri-urban and urban soils using hybridized empirical bayesian kriging and support vector machine regression. Sci. Rep. 12 (1), 1–16.

Agyeman, P.C., Ahado, S.K., Kingsley, J., Kebonye, N.M., Biney, J.K.M., Borůvka, L., Vasat, R., Kocarek, M., 2021. Source apportionment, contamination levels, and spatial prediction of potentially toxic elements in selected soils of the Czech Republic. Environ. Geochem. Health 43, 601–620. https://doi.org/10.1007/s10653-020-00743-8.

Åkesson, A., Barregard, L., Bergdahl, I.A., Nordberg, G.F., Nordberg, M., Skerfving, S., 2014. Non-renal effects and the risk assessment of environmental cadmium exposure. Environ. Health Perspect. 122 (5), 431–438. https://doi.org/10.1289/ehp.1307110.

Al-Anbari, R., A. O. A. H.M.J. , & A. F. H.A. (2015). Pollution loads and ecological risk assessment of heavy metals in the urban soil affected by various anthropogenic activities. ⟨https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Al-Anbari%2C+R.%2C+Abdul+Hameed%2C+M.+J.%2C+Obaidy%2C+Al%2C+%26+Fatima%2C+H.+A.+A.+%282015%29.+Pollution+loads+and+1036+ecological+risk+assessment+of+heavy+metals+in+the+urban+soil+affected+by+various+anthropogenic+1037+activities.+International+Journal+of+Advanced+Research%2C+2%2C+104%E2%80%93110.&btnG=⟩.

Antić-Mladenović, S., Rinklebe, J., Frohne, T., Stärk, H.J., Wennrich, R., Tomić, Z., Ličina, V., 2011. Impact of controlled redox conditions on nickel in a serpentine soil. J. Soils Sediment. 11 (3), 406–415. https://doi.org/10.1007/s11368-010-0325-0.

Ashaiekh, M.A., Eltayeb, M.A.H., Ali, A.H., Ebrahim, A.M., Salih, I., Idris, A.M., 2019. Spatial distribution of total and bioavailable heavy metal contents in soil from agricultural, residential, and industrial areas in Sudan. Toxin Rev. 38 (2), 93–105. https://doi.org/10.1080/15569543.2017.1419491.

Ballabio, C., Panagos, P., Lugato, E., Huang, J.H., Orgiazzi, A., Jones, A., Montanarella, L., 2018. Copper distribution in European topsoils: An assessment based on LUCAS soil survey. Sci. Total Environ. 636, 282–298.

Baran Jerzy Wieczorek Ryszard Mazurek Krzysztof Urban, A., Klimkowicz-Pawlas, Agnieszka, ski, Baran, A., Wieczorek, Á.J., Wieczorek, J., Mazurek, R., Urban, K., Klimkowicz-Pawlas, A., 2017. Potential ecological risk assessment and predicting zinc accumulation in soils. Environ. Geochem Health 40 (1), 435–450. https://doi.org/10.1007/s10653-017-9924-7.

Basta, N., Ryan, J., Chaney, R., 2005. Trace element chemistry in residual-treated soil: key concepts and metal bioavailability. J. Environ. Qual.

Bayraklı, B., Dengiz, O., 2020. An evaluation of heavy metal pollution risk in tea cultivation soils of micro-catchments using various pollution indexes under humid environmental condition. Rend. Lince-.-. 31 (2), 393–409. https://doi.org/10.1007/S12210-020-00901-1/FIGURES/7.

Blog, P. (2020). Importance of Nickel in Stainless Steel Industry? Pipingmart. ⟨https://www.pipingmart.com/blog/metals/importance-of-nickel-in-stainless-steel-industry/⟩.

Borůvka, L., Vacek, O., Jehlička, J., 2005. Principal component analysis as a tool to indicate the origin of potentially toxic elements in soils. Geoderma 128 (3–4 SPEC. ISS.), 289–300. https://doi.org/10.1016/j.geoderma.2005.04.010.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Brown, S.G., Eberly, S., Paatero, P., Norris, G.A., 2015. Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results. Sci. Total Environ. 518, 626–635.

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr. Anal. 28 (4), 281–298.

Cambardella, C.A., Moorman, T.B., Novak, J.M., Parkin, T.B., Karlen, D.L., Turco, R.F., Konopka, A.E., 1994. Field-scale variability of soil properties in central iowa soils. Soil Sci. Soc. Am. J. 58 (5), 1501–1511. https://doi.org/10.2136/SSSAJ1994.03615995005800050033X.

Chandrasekaran, A., Ravisankar, R., Harikrishnan, N., Satapathy, K.K., Prasad, M.V.R., Kanagasabapathy, K.V., 2015. Multivariate statistical analysis of heavy metal concentration in soils of Yelagiri Hills, Tamilnadu, India - Spectroscopical approach. Spectrochim. Acta - Part A: Mol. Biomol. Spectrosc. 137, 589–600. https://doi.org/10.1016/j.saa.2014.08.093.

Chen, H., Teng, Y., Lu, S., Wang, Y., Wang, J., 2015. Contamination features and health risk of soil heavy metals in China. Sci. Total Environ. 512–513, 143–153. https://doi.org/10.1016/j.scitotenv.2015.01.025.

Chen, X., Lu, X., 2018. Contamination characteristics and source apportionment of heavy metals in topsoil from an area in Xi'an city, China. Ecotoxicol. Environ. Saf. 151, 153–160. https://doi.org/10.1016/j.ecoenv.2018.01.010.

Chen, Y., Jiang, X., Wang, Y., Zhuang, D., 2018. Spatial characteristics of heavy metal pollution and the potential ecological risk of a typical mining area: A case study in China. Process Saf. Environ. Prot. 113, 204–219.

Chen, Y., Jiang, X., Wang, Y., Zhuang, D., 2018a. Spatial characteristics of heavy metal pollution and the potential ecological risk of a typical mining area: a case study in China. Process Saf. Environ. Prot. 113, 204–219. https://doi.org/10.1016/j.psep.2017.10.008.

Chen, Y., Jiang, X., Wang, Y., Zhuang, D., 2018b. Spatial characteristics of heavy metal pollution and the potential ecological risk of a typical mining area: a case study in China. Process Saf. Environ. Prot. 113, 204–219. https://doi.org/10.1016/j.psep.2017.10.008.

Corguinha, A.P.B., Souza, G.A. de, Gonçalves, V.C., Carvalho, C. de A., Lima, W.E.A. de, Martins, F.A.D., Yamanaka, C.H., Francisco, E.A.B., Guilherme, L.R.G., 2015. Assessing arsenic, cadmium, and lead contents in major crops in Brazil for food safety purposes. J. Food Compos. Anal. 37, 143–150. https://doi.org/10.1016/j.jfca.2014.08.004.

Cui, Z., Wang, Y., Zhao, N., Yu, R., Xu, G., Yu, Y., 2018. Spatial distribution and risk assessment of heavy metals in paddy soils of yongshuyu irrigation area from Songhua River Basin, Northeast China. Chin. Geogr. Sci. 28 (5), 797–809. https://doi.org/10.1007/s11769-018-0991-1.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792. https://doi.org/10.1890/07-0539.1.

Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinforma. 7 (1), 3. https://doi.org/10.1186/1471-2105-7-3.

Escarré, J., Lefèbvre, C., Raboyeau, S., Dossantos, A., Gruber, W., Cleyet Marel, J.C., Frérot, H., Noret, N., Mahieu, S., Collin, C., Van Oort, F., 2011. Heavy metal concentration survey in soils and plants of the Les Malines Mining District (southern France): Implications for soil restoration. Water Air Soil Pollut. 216 (1–4), 485–504. https://doi.org/10.1007/s11270-010-0547-1.

Fei, X., Xiao, R., Christakos, G., Langousis, A., Ren, Z., Tian, Y., Lv, X., 2019. Comprehensive assessment and source apportionment of heavy metals in Shanghai agricultural soils with different fertility levels. Ecol. Indic. 106, 105508 https://doi.org/10.1016/j.ecolind.2019.105508.

Gan, Y., Miao, Y., Wang, L., Yang, G., Li, Y.C., Wang, W., Dai, J., 2018. Source contribution analysis and collaborative assessment of heavy metals in vegetable-growing soils. J. Agric. Food Chem. 66 (42), 10943–10951. https://doi.org/10.1021/acs.jafc.8b04032.

Gao, H., Bai, J., Xiao, R., Liu, P., Jiang, W., Wang, J., 2013. Levels, sources and risk assessment of trace elements in wetland soils of a typical shallow freshwater lake, China. Stoch. Environ. Res. Risk Assess. 27 (1), 275–284. https://doi.org/10.1007/s00477-012-0587-8.

Gao, J., Wang, L., 2018. Ecological and human health risk assessments in the context of soil heavy metal pollution in a typical industrial area of Shanghai, China. Environ. Sci. Pollut. Res. 25 (27), 27090–27105. https://doi.org/10.1007/s11356-018-2705-8.

Gąsiorek, M., Kowalska, J., Mazurek, R., Chemosphere, M.P.-, 2017, undefined, 2017. Comprehensive assessment of heavy metal pollution in topsoil of historical urban park on an example of the Planty Park in Krakow (Poland). Elsevier,. ⟨https://www.sciencedirect.com/science/article/pii/S0045653517304794⟩.

Gattullo, C.E., Allegretta, I., Porfido, C., Rascio, I., Spagnuolo, M., Terzano, R., 2020. Assessing chromium pollution and natural stabilization processes in agricultural soils by bulk and micro X-ray analyses. Environ. Sci. Pollut. Res. 27 (18), 22967–22979.

Gautam, R., Panigrahi, S., Franzen, D., Sims, A., 2011. Residual soil nitrate prediction from imagery and non-imagery information using neural network technique. Biosyst. Eng. 110 (1), 20–28. https://doi.org/10.1016/j.biosystemseng.2011.06.002.

Gholizadeh, A., Borůvka, L., Mehdi Saberioon, M., Kozák, J., Vašát, R., Němeček, K., 2015. Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. Agric. Cz 10 (4), 218–227. https://doi.org/10.17221/113/2015-SWR.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognit. Lett. 27 (4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011.

Guan, Q., Zhao, R., Pan, N., Wang, F., Yang, Y., Luo, H., 2019. Source apportionment of heavy metals in farmland soil of Wuwei, China: Comparison of three receptor models. J. Clean. Prod. 237. https://doi.org/10.1016/j.jclepro.2019.117792.

Haji Gholizadeh, M., Melesse, A.M., Reddi, L., 2016. Water quality assessment and apportionment of pollution sources using APCS-MLR and PMF receptor modeling techniques in three major rivers of South Florida. Sci. Total Environ. 566–567, 1552–1567. https://doi.org/10.1016/j.scitotenv.2016.06.046.

Hakanson, L., 1980. An ecological risk index for aquatic pollution control.a sedimentological approach. Water Res. 14 (8), 975–1001. https://doi.org/10.1016/0043-1354(80)90143-8.

Håkanson, L., 1980. An ecological risk index for aquatic pollution control-a sedimentological approach. Water Res. 14, 975–1001.

Harasim, P., Filipek, T., 2015a. Nickel in the environment. J. Elem. 20 (2), 525–534. https://doi.org/10.5601/jelem.2014.19.3.651.

Harasim, P., Filipek, T., 2015b. Nickel in the environment. J. Elem. 20 (2), 525–534. https://doi.org/10.5601/jelem.2014.19.3.651.

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a Random Forest approach. Geoderma 214–215, 141–154. https://doi.org/10.1016/j.geoderma.2013.09.016.

Hossain Bhuiyan, M.A., Chandra Karmaker, S., Bodrud-Doza, M., Rakib, M.A., Saha, B.B., 2021a. Enrichment, sources and ecological risk mapping of heavy metals in agricultural soils of dhaka district employing SOM, PMF and GIS methods. Chemosphere *263*. https://doi.org/10.1016/j.chemosphere.2020.128339.

Hossain Bhuiyan, M.A., Chandra Karmaker, S., Bodrud-Doza, M., Rakib, M.A., Saha, B.B., 2021b. Enrichment, sources and ecological risk mapping of heavy metals in agricultural soils of dhaka district employing SOM, PMF and GIS methods. Chemosphere 263, 128339. https://doi.org/10.1016/j.chemosphere.2020.128339.

Hseu, Z.Y., Su, Y.C., Zehetner, F., Hsi, H.C., 2017. Leaching potential of geogenic nickel in serpentine soils from Taiwan and Austria. J. Environ. Manag. 186, 151–157. https://doi.org/10.1016/j.jenvman.2016.02.034.

Hu, Y., Cheng, H., 2013. Application of stochastic models in identification and apportionment of heavy metal pollution sources in the surface soils of a large-scale region. Environ. Sci. Technol. 47 (8), 3752–3760. https://doi.org/10.1021/es304310k.

Huang, J., Wu, Y., Sun, J., Li, X., Geng, X., Zhao, M., Fan, Z., 2021. Health risk assessment of heavy metal (loid) s in park soils of the largest megacity in China by using Monte Carlo simulation coupled with Positive matrix factorization model. J. Hazard. Mater. 415, 125629.

Huang, Y., Wang, L., Wang, W., Li, T., He, Z., Yang, X., 2019. Current status of agricultural soil pollution by heavy metals in China: a meta-analysis. Sci. Total Environ. 651, 3034–3042. https://doi.org/10.1016/j.scitotenv.2018.10.185.

Huang, Y., Chen, Q., Deng, M., Japenga, J., Li, T., Yang, X., He, Z., 2018. Heavy metal pollution and health risk assessment of agricultural soils in a typical peri-urban area in southeast China. J. Environ. Manag. 207, 159–168. https://doi.org/10.1016/j.jenvman.2017.10.072.

Huang, Y., Deng, M., Wu, S., Japenga, J., Li, T., Yang, X., He, Z., 2018a. A modified receptor model for source apportionment of heavy metal pollution in soil. J. Hazard. Mater. 354, 161–169. https://doi.org/10.1016/j.jhazmat.2018.05.006.

Huang, Y., Deng, M., Wu, S., Japenga, J., Li, T., Yang, X., He, Z., 2018b. A modified receptor model for source apportionment of heavy metal pollution in soil. J. Hazard. Mater. 354, 161–169. https://doi.org/10.1016/j.jhazmat.2018.05.006.

Imran, M., Stein, A., Zurita-Milla, R., 2015. Using geographically weighted regression kriging for crop yield mapping in West Africa. Int. J. Geogr. Inf. Syst. 29 (2), 234–257.

Jin, Y., O'Connor, D., Ok, Y.S., Tsang, D.C.W., Liu, A., Hou, D., 2019. Assessment of sources of heavy metals in soil and dust at children's playgrounds in Beijing using

GIS and multivariate statistical analysis. Environ. Int. 124, 320–328. https://doi.org/10.1016/j.envint.2019.01.024.
John, K., Isong, I.A., Kebonye, N.M., Ayito, E.O., Agyeman, P.C., Afu, S.M., 2020. Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. Land 9 (12), 1–20. https://doi.org/10.3390/land9120487.
John, K., Agyeman, P.C., Kebonye, N.M., Isong, I.A., Ayito, E.O., Ofem, K.I., Qin, C.Z., 2021. Hybridization of cokriging and gaussian process regression modelling techniques in mapping soil sulphur. Catena 206, 105534.
Kabata-Pendias, A., 2011. Trace Elements in Soils and Plants, fourth ed. CRC Press. Taylor and Francis Group. ISBN: 978-1-4200-9368-1.
Kars, N., Dengiz, O., 2020. Assessment of potential ecological risk index based on heavy metal elements for organic farming in micro catchments under humid ecological condition. Eurasia J. Soil Sci. 9 (3), 194–201. https://doi.org/10.18393/EJSS.719167/XML.
Kebonye, N.M., John, K., Chakraborty, S., Agyeman, P.C., Ahado, S.K., Eze, P.N., Němeček, K., Drábek, O., Borůvka, L., 2021. Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy. Geoderma *384*. https://doi.org/10.1016/j.geoderma.2020.114792.
Kelepertzis, E., 2014. Accumulation of heavy metals in agricultural soils of Mediterranean: Insights from Argolida basin, Peloponnese, Greece. Geoderma 221–222, 82–90. https://doi.org/10.1016/j.geoderma.2014.01.007.
Keshavarzi, A., Kumar, V., 2019. Ecological risk assessment and source apportionment of heavy metal contamination in agricultural soils of Northeastern Iran. Int. J. Environ. Health Res. 29 (5), 544–560. https://doi.org/10.1080/09603123.2018.1555638.
Keshavarzi, A., Kumar, V., 2020. Spatial distribution and potential ecological risk assessment of heavy metals in agricultural soils of Northeastern Iran. Geol., Ecol., Landsc. 4 (2), 87–103. https://doi.org/10.1080/24749508.2019.1587588.
Kodom, K., Preko, K., Boamah, D., 2012. X-ray fluorescence (XRF) analysis of soil heavy metal pollution from an industrial area in Kumasi, Ghana. Soil Sediment Contam.: An Int. J. 21 (8), 1006–1021.
Kooistra, L., Wanders, J., Epema, G.F., Leuven, R.S.E.W., Wehrens, R., Buydens, L.M.C., 2003. The potential of field spectroscopy for the assessment of sediment properties in river floodplains. Anal. Chim. Acta 484 (2), 189–200. https://doi.org/10.1016/S0003-2670(03)00331-3.
Kozák, J. (2010). Soil Atlas of the Czech Republic. 150.
Kumar, S., Lal, R., Liu, D., 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma 189, 627–634.
Kumar, V., Sharma, A., Kaur, P., Singh Sidhu, G.P., Bali, A.S., Bhardwaj, R., Thukral, A. K., Cerda, A., 2019. Pollution assessment of heavy metals in soils of India and ecological risk assessment: a state-of-the-art. Chemosphere 216, 449–462. https://doi.org/10.1016/j.chemosphere.2018.10.066.
Lambert, T.W., Boehmer, J., Feltham, J., Guyn, L., Shahid, R., 2011. Spatial mapping of lead, arsenic, iron, and polycyclic aromatic hydrocarbon soil contamination in Sydney, Nova Scotia: community impact from the coke ovens and steel plant. Arch. Environ. Occup. Health 66 (3), 128–145. https://doi.org/10.1080/19338244.2010.516780.
Li, H.B., Wang, J.Y., Chen, X.Q., Li, Y.P., Fan, J., Ren, J.H., Luo, X.S., Juhasz, A.L., Ma, L. Q., 2020. Geogenic nickel exposure from food consumption and soil ingestion: a bioavailability based assessment. Environ. Pollut. 265, 114873 https://doi.org/10.1016/j.envpol.2020.114873.
Li, L., Lu, J., Wang, S., Ma, Y., Wei, Q., Li, X., Cong, R., Ren, T., 2016. Methods for estimating leaf nitrogen concentration of winter oilseed rape (Brassica napus L.) using in situ leaf spectroscopy. Ind. Crops Prod. 91, 194–204. https://doi.org/10.1016/j.indcrop.2016.07.008.
Li, N., Kang, Y., Pan, W., Zeng, L., Zhang, Q., Luo, J., 2015. Concentration and transportation of heavy metals in vegetables and risk assessment of human exposure to bioaccessible heavy metals in soil near a waste-incinerator site, South China. Sci. Total Environ. 521–522, 144–151. https://doi.org/10.1016/j.scitotenv.2015.03.081.
Li, Z., Feng, X., Li, G., Bi, X., Zhu, J., Qin, H., Dai, Z., Liu, J., Li, Q., Sun, G., 2013. Distributions, sources and pollution status of 17 trace metal/metalloids in the street dust of a heavily industrialized city of central China. Environ. Pollut. 182, 408–416. https://doi.org/10.1016/j.envpol.2013.07.041.
Liang, J., Feng, C., Zeng, G., Gao, X., Zhong, M., Li, X., Li, X., He, X., Fang, Y., 2017. Spatial distribution and source identification of heavy metals in surface soils in a typical coal mine city, Lianyuan, China. Environ. Pollut. 225, 681–690. https://doi.org/10.1016/j.envpol.2017.03.057.
Liu, G., Yu, Y., Hou, J., Xue, W., Liu, X., Liu, Y., Wang, W., Alsaedi, A., Hayat, T., Liu, Z., 2014. An ecological risk assessment of heavy metal pollution of the agricultural ecosystem near a lead-acid battery factory. Ecol. Indic. 47, 210–218. https://doi.org/10.1016/j.ecolind.2014.04.040.
Liu, L., Dong, Y., Kong, M., Zhou, J., Zhao, H., Tang, Z., Zhang, M., Wang, Z., 2020. Insights into the long-term pollution trends and sources contributions in Lake Taihu, China using multi-statistic analyses models. Chemosphere 242, 125272. https://doi.org/10.1016/j.chemosphere.2019.125272.
Lv, J., 2019. Multivariate receptor models and robust geostatistics to estimate source apportionment of heavy metals in soils. Environ. Pollut. 244, 72–83. https://doi.org/10.1016/j.envpol.2018.09.147.
Lv, J., Wang, Y., 2018. Multi-scale analysis of heavy metals sources in soils of Jiangsu Coast, Eastern China. Chemosphere 212, 964–973. https://doi.org/10.1016/j.chemosphere.2018.08.155.
Ma, W., Tai, L., Qiao, Z., Zhong, L., Wang, Z., Fu, K., Chen, G., 2018. Contamination source apportionment and health risk assessment of heavy metals in soil around municipal solid waste incinerator: a case study in North China. Sci. Total Environ. 631–632, 348–357. https://doi.org/10.1016/j.scitotenv.2018.03.011.
Mamut, A., Eziz, M., Mohammad, A., 2018. Pollution and ecological risk assessment of heavy metals in farmland soils in Yanqi County, Xinjiang, Northwest China. Eurasia Soil Sci. 51 (8), 985–993. https://doi.org/10.1134/S1064229318080082.
Mazurek, R., Kowalska, J., Gąsiorek, M., Chemosphere, P.Z.-, 2017, undefined, 2017. Assessment of heavy metals contamination in surface layers of Roztocze National Park forest soils (SE Poland) by indices of pollution. Elsevier. ⟨https://www.sciencedirect.com/science/article/pii/S004565351631517X⟩.
Men, C., Liu, R., Wang, Q., Guo, L., Shen, Z., 2018. The impact of seasonal varied human activity on characteristics and sources of heavy metals in metropolitan road dusts. Sci. Total Environ. 637–638, 844–854. https://doi.org/10.1016/j.scitotenv.2018.05.059.
Mitran, T., Mishra, U., Lal, R., Ravisankar, T., Sreenivas, K., 2018. Spatial distribution of soil carbon stocks in a semi-arid region of India. Geoderma Reg. 15, e00192.
Molinaro, A.M., Simon, R., Pfeiffer, R.M., 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics 21 (15), 3301–3307. https://doi.org/10.1093/bioinformatics/bti499.
Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE Vol. 50 (Issue 3).
Morton-Bermea, O., Hernández-Álvarez, E., González-Hernández, G., Romero, F., Lozano, R., Beramendi-Orosco, L.E., 2009. Assessment of heavy metal pollution in urban topsoils from the metropolitan area of Mexico City. J. Geochem. Explor. 101 (3), 218–224. https://doi.org/10.1016/j.gexplo.2008.07.002.
Nawar, S., Mouazen, A.M., 2017. Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. Sens. (Switz.) 17 (10), 2428. https://doi.org/10.3390/s17102428.
Nemecek, J., & Podlesakova, E. (1992). RETROSPECTIVE EXPERIMENTAL MONITORING OF HEAVY-METALS. - Google Scholar. Rostlinna Vyroba.
Öztürk, E., Dengiz, O., 2020. Assessment and selection of suitable microbasins for organic agriculture under subhumid ecosystem conditions: a case study from Trabzon Province, Turkey. Arab. J. Geosci. 13 (22), 1–13. https://doi.org/10.1007/S12517-020-06200-1/TABLES/6.
Paatero, P., 1999. The multilinear engine—a table-driven, least squares program for solving multilinear problems, including the N-way parallel factor analysis model. J. Comput. Graph. Stat. 8 (4), 854–888. https://doi.org/10.1080/10618600.1999.10474853.
Paatero, P., Eberly, S., S, B.-A., 2014, undefined, 2014. Methods for estimating uncertainty in factor analytic solutions. AMT Copernic. Org. 7, 781–797. https://doi.org/10.5194/amt-7-781-2014.
Pereira, O.J.R., Melfi, A.J., Montes, C.R., Lucas, Y., 2018. Downscaling of ASTER thermal images based on geographically weighted regression kriging. Remote Sens 10 (4), 633.
Qasemi, M., Shams, M., Sajjadi, S.A., Farhang, M., Erfanpoor, S., Yousefi, M., Zarei, A., Afsharnia, M., 2019. Cadmium in groundwater consumed in the rural areas of gonabad and bajestan, iran: occurrence and health risk assessment. Biol. Trace Elem. Res. 192 (2), 106–115. https://doi.org/10.1007/s12011-019-1660-7.
Qiao, X., Shu, X., Tang, Y., Duan, L., Seyler, B.C., Guo, H., Zhang, H., 2021. Atmospheric deposition of sulfur and nitrogen in the West China rain zone: Fluxes, concentrations, ecological risks, and source apportionment. Atmos. Res. 256, 105569.
Reimann, C. (2005). Geochemical mapping: technique or art? Geochemistry: Exploration, Environment, Analysis.
Reimann, C., Filzmoser, P., Garrett, R.G., & Dutter, R. (2008). Statistical Data Analysis Explained: Applied Environmental Statistics with R. In Statistical Data Analysis Explained: Applied Environmental Statistics with R. https://doi.org/10.1002/9780470987605.
Robertsa, T.L., 2014. Cadmium and phosphorous fertilizers: the issues and the science. Procedia Eng. 83, 52–59. https://doi.org/10.1016/j.proeng.2014.09.012.
Rodríguez, J.A., Nanos, N., Grau, J.M., Gil, L., López-Arias, M., 2008. Multiscale analysis of heavy metal contents in Spanish agricultural topsoils. Chemosphere 70 (6), 1085–1096. https://doi.org/10.1016/j.chemosphere.2007.07.056.
Salim, I., Sajjad, R.U., Paule-Mercado, M.C., Memon, S.A., Lee, B.Y., Sukhbaatar, C., Lee, C.H., 2019. Comparison of two receptor models PCA-MLR and PMF for source identification and apportionment of pollution carried by runoff from catchment and sub-watershed areas with mixed land cover in South Korea. Sci. Total Environ. 663, 764–775. https://doi.org/10.1016/j.scitotenv.2019.01.377.
Santos-Francés, F., Martínez-Graña, A., Rojo, P.A., Sánchez, A.G., 2017. Geochemical background and baseline values determination and spatial distribution of heavy metal pollution in soils of the andes mountain range (Cajamarca-Huancavelica, Peru). Int. J. Environ. Res. Public Health 14 (8), 859. https://doi.org/10.3390/ijerph14080859.
Satarug, S., Vesey, D.A., Gobe, G.C., 2017a. Current health risk assessment practice for dietary cadmium: Data from different countries. Food Chem. Toxicol. 106, 430–445. https://doi.org/10.1016/j.fct.2017.06.013.
Satarug, S., Vesey, D.A., Gobe, G.C., 2017b. Health risk assessment of dietary cadmium intake: do current guidelines indicate how much is safe?. In: Environmental Health Perspectives, Vol. 125 Public Health Services, US Dept of Health and Human Services,, pp. 284–288. https://doi.org/10.1289/EHP108.
Satyendra. (2014a). Chromium in Steels. Ispatguru. ⟨https://www.ispatguru.com/chromium-in-steels/⟩.
Satyendra. (2014b). Copper in Steels. Ispatguru. ⟨https://www.ispatguru.com/copper-in-steels/⟩.
Sawut, R., Kasim, N., Maihemuti, B., Hu, L., Abliz, A., Abdujappar, A., Kurban, M., 2018. Pollution characteristics and health risk assessment of heavy metals in the vegetable bases of northwest China. Sci. Total Environ. 642, 864–878. https://doi.org/10.1016/j.scitotenv.2018.06.034.

Sayadi, M.H., Shabani, M., Ahmadpour, N., 2015a. Pollution index and ecological risk of heavy metals in the surface soils of amir-abad area in Birjand City, Iran. Health Scope 4 (1). https://doi.org/10.17795/jhealthscope-21137.

Sayadi, M.H., Shabani, M., Ahmadpour, N., 2015b. Pollution index and ecological risk of heavy metals in the surface soils of amir-abad area in Birjand City, Iran. Health Scope 4 (1). https://doi.org/10.17795/jhealthscope-21137.

Shao, D., Zhan, Y., Zhou, W., Zhu, L., 2016. Current status and temporal trend of heavy metals in farmland soil of the Yangtze River Delta Region: field survey and meta-analysis. Environ. Pollut. 219, 329–336. https://doi.org/10.1016/j.envpol.2016.10.023.

Shen, F., Mao, L., Sun, R., Du, J., Tan, Z., Ding, M., 2019. Contamination evaluation and source identification of heavy metals in the sediments from the Lishui River Watershed, Southern China. Int. J. Environ. Res. Public Health 16 (3), 336.

Tao, S., Yang, Zhong, Qing, B., Lin, Y., Ma, J., Zhou, Y., Hou, H., Zhao, L., Sun, Z., Qin, X., Shi, H., 2017. Application of a self-organizing map and positive matrix factorization to investigate the spatial distributions and sources of polycyclic aromatic hydrocarbons in soils from Xiangfen County, northern China. Ecotoxicol. Environ. Saf. 141, 98–106. https://doi.org/10.1016/j.ecoenv.2017.03.017.

Taylor, K.E. (2005). Taylor Diagram Primer. In Work. Pap (Issue January).

Tomlinson, D.L., Wilson, J.G., Harris, C.R., Jeffrey, D.W., 1980. Problems in the assessment of heavy-metal levels in estuaries and the formation of a pollution index. Helgoländer Meeresunters. 33 (1–4), 566–575. https://doi.org/10.1007/BF02414780.

Tóth, G., Hermann, T., Szatmári, G., Pásztor, L., 2016. Maps of heavy metals in the soils of the European Union and proposed priority areas for detailed assessment. Sci. Total Environ. 565, 1054–1062.

Tziachris, P., Aschonitis, V., Chatzistathis, T., Papadopoulou, M., 2019. Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters. Catena 174, 206–216.

U.S. EPA. (2014). Positive Matrix Factorization (PMF) 5.0-Fundamentals and User Guide.

USEPA, 1998. Guidelines for ecological risk assessment. Fed. Regist. Vol. 63.

Vacek, O., Vašát, R., Borůvka, L., 2020. Quantifying the pedodiversity-elevation relations. Geoderma 373, 114441. https://doi.org/10.1016/j.geoderma.2020.114441.

Wang, H.Y., Wen, S.L., Chen, P., Zhang, L., Cen, K., Sun, G.X., 2016. Mitigation of cadmium and arsenic in rice grain by applying different silicon fertilizers in contaminated fields. Environ. Sci. Pollut. Res. 23 (4), 3781–3788. https://doi.org/10.1007/s11356-015-5638-5.

Wang, K., Zhang, C., Li, W., 2012. Comparison of geographically weighted regression and regression kriging for estimating the spatial distribution of soil organic matter. GIsci. Remote Sens. 49 (6), 915–932.

Wang, L., Cui, X., Cheng, H., Chen, F., Wang, J., Zhao, X., Lin, C., Pu, X., 2015. A review of soil cadmium contamination in China including a health risk assessment. Environ. Sci. Pollut. Res. 22 (21), 16441–16452. https://doi.org/10.1007/s11356-015-5273-1.

Wang, S., Cai, L.M., Wen, H.H., Luo, J., Wang, Q.S., Liu, X., 2019. Spatial distribution and source apportionment of heavy metals in soil from a typical county-level city of Guangdong Province, China. Sci. Total Environ. 655, 92–101.

Wang, Y., Guo, G., Zhang, D., Lei, M., 2021. An integrated method for source apportionment of heavy metal (loid) s in agricultural soils and model uncertainty analysis. Environ. Pollut. 276, 116666.

Wang, Z., Wang, Y., Chen, L., Yan, C., Yan, Y., Bulletin, Q.C.-M.P., 2015, undefined, 2015. Assessment of metal contamination in coastal sediments of the Maluan Bay (China) using geochemical indices and multivariate statistical approaches. Elsevier,. ⟨https://www.sciencedirect.com/science/article/pii/S0025326X1500497X⟩.

Wang, Z., Xiao, J., Wang, L., Liang, T., Guo, Q., Guan, Y., Rinklebe, J., 2020. Elucidating the differentiation of soil heavy metals under different land uses with geographically weighted regression and self-organizing map. Environ. Pollut. 260, 114065.

Weather Spark. (2016). Average Weather in Frýdek-Místek, Czechia, Year Round - Weather Spark. ⟨https://weatherspark.com/y/83671/Average-Weather-in-Frýdek-Místek-Czechia-Year-Round⟩.

Weissmannová, H.D., Mihočová, S., Chovanec, P., Pavlovský, J., 2019. Potential ecological risk and human health risk assessment of heavy metal pollution in industrial affected soils by coal mining and metallurgy in ostrava, Czech Republic. Int. J. Environ. Res. Public Health 2019 16 (22), 4495. https://doi.org/10.3390/IJERPH16224495.

Wu, H., Liao, Q., Chillrud, S.N., Yang, Q., Huang, L., Bi, J., Yan, B., 2016. Environmental exposure to cadmium: health risk assessment and its associations with hypertension and impaired kidney function. Sci. Rep. 6. https://doi.org/10.1038/srep29989.

Wu, J., Lu, J., Li, L., Min, X., Luo, Y., 2018. Pollution, ecological-health risks, and sources of heavy metals in soil of the northeastern Qinghai-Tibet Plateau. Chemosphere 201, 234–242. https://doi.org/10.1016/j.chemosphere.2018.02.122.

Wu, J., Margenot, A.J., Wei, X., Fan, M., Zhang, H., Best, J.L., Gao, C., 2020. Source apportionment of soil heavy metals in fluvial islands, Anhui section of the lower Yangtze River: comparison of APCS–MLR and PMF. J. Soils Sediments 20 (9), 3380–3393.

Wuana, R.A., Okieimen, F.E., 2014. Heavy metals in contaminated soils: a review of sources, chemistry, risks, and best available strategies for remediation. Heavy Met. Contam. Water Soil.: Anal., Assess., Remediat. Strateg. 1–50. https://doi.org/10.1201/b16566.

Yang, B., Zhou, L., Xue, N., Li, F., Li, Y., Vogt, R.D., Cong, X., Yan, Y., Liu, B., 2013a. Source apportionment of polycyclic aromatic hydrocarbons in soils of Huanghuai Plain, China: Comparison of three receptor models. Sci. Total Environ. 443, 31–39. https://doi.org/10.1016/j.scitotenv.2012.10.094.

Yang, B., Zhou, L., Xue, N., Li, F., Li, Y., Vogt, R.D., Cong, X., Yan, Y., Liu, B., 2013b. Source apportionment of polycyclic aromatic hydrocarbons in soils of Huanghuai Plain, China: comparison of three receptor models. Sci. Total Environ. 443, 31–39. https://doi.org/10.1016/j.scitotenv.2012.10.094.

Yang, Q., Li, Z., Lu, X., Duan, Q., Huang, L., Bi, J., 2018. A review of soil heavy metal pollution from industrial and agricultural regions in China: pollution and risk assessment. Sci. Total Environ. Vol. 642, 690–700. https://doi.org/10.1016/j.scitotenv.2018.06.068.

Yang, Y., Christakos, G., Guo, M., Xiao, L., Huang, W., 2017. Space-time quantitative source apportionment of soil heavy metal concentration increments. Environ. Pollut. 223, 560–566. https://doi.org/10.1016/j.envpol.2017.01.058.

Ye, H., Huang, W., Huang, S., Huang, Y., Zhang, S., Dong, Y., Chen, P., 2017. Effects of different sampling densities on geographically weighted regression kriging for predicting soil organic carbon. Spat. Stat. 20, 76–91.

Yu, G., Dai, F., Wang, W., Zheng, W., Zhang, Z., Yuan, Y., Wang, Q., 2017. Health risk assessment of Chinese consumers to lead via diet. Hum. Ecol. Risk Assess. 23 (8), 1928–1940. https://doi.org/10.1080/10807039.2017.1338934.

Zeremski-Škorić, T., Ninkov, J., Sekulić, P., Milić, S., Vasin, J., Dozet, D., & Jakšić, S. (2010). Heavy metal content in some fertilizers used in Serbia. Ratarstvo i povrtarstvo/Field and Vegetable Crops Research, 47(1), 281–287.

Zhang et al. (2012). Development of the high-order decoupled direct method in three dimensions for particulate matter: enabling advanced sensitivity analysis in air quality models. Geoscientific Model Development 5. ⟨https://scholar.google.com/scholar_lookup?title=Development of the high-order decoupled direct method in three dimensions for particulate matter%3A enabling advanced sensitivity analysis in air quality models&journal=Geosci Model Dev&volume=5&pages=355–36⟩.

Zhang, Q., Wang, C., 2020. Natural and human factors affect the distribution of soil heavy metal pollution: a review. In: Water, Air, and Soil Pollution, Vol. 231. Springer. https://doi.org/10.1007/s11270-020-04728-2.

Zhu, D., Wei, Y., Zhao, Y., Wang, Q., Han, J., 2018. Heavy metal pollution and ecological risk assessment of the agriculture soil in Xunyang Mining Area, Shaanxi Province, Northwestern China. Bull. Environ. Contam. Toxicol. 101 (2), 178–184. https://doi.org/10.1007/s00128-018-2374-9.

Agyeman *et al.*
Environmental Sciences Europe (2021) 33:137
https://doi.org/10.1186/s12302-021-00577-w

Environmental Sciences Europe

RESEARCH Open Access

Human health risk exposure and ecological risk assessment of potentially toxic element pollution in agricultural soils in the district of Frydek Mistek, Czech Republic: a sample location approach

Prince Chapman Agyeman*, Kingsley John, Ndiye Michael Kebonye, Luboš Borůvka, Radim Vašát, Ondřej Drábek and Karel Němeček

Abstract

Background: Human activities considerably contribute to polluting potentially toxic element (PTEs) levels in soils, especially agricultural soils. The consistent introduction of PTEs in the environment and the soil pose health-related risks to humans, flora and fauna. One hundred and fifteen samples were collected in the district of Frydek Mistek (Czech Republic) in a regular grid form. The soil samples were air-dried, and the concentrations of PTEs (i.e. lead, arsenic, chromium, nickel, manganese, cadmium, copper, and zinc) were determined by ICP-OES (inductively coupled plasma optical emission spectrometry). The purpose of this study is to create digitized soil maps that expose the human-related health risks posed by PTEs, estimate pollution indices, ascertain the spatially distributed patterns of PTEs, source apportionment and quantify carcinogenic and non-carcinogenic health risks using the sample location approach.

Results: The results revealed that the pollution assessment of the soils in the study area using diverse pollution assessment indexes (pollution index, pollution load index, ecological risk and risk index), based on the application of the local background value and the European average value, displayed a range of pollution levels due to differences in the threshold limits from differing geochemical background levels. The principal components analysis and positive matrix factorization, respectively, identified the sources of pollution and the distribution of PTE sources. Mapping the health index and total carcinogenic risk highlighted hotspots of areas within the study area that require immediate remediation. The self-organizing map (SeOM) revealed a diversified colour pattern for the factor scores. A single neuron exhibited a high hotspot in all factor loadings on different blocks of neurons. Children's CDItotal (Chronic Daily Intake total) values for non-carcinogenic risk and carcinogenic risk were found to be greater than adults', as were their HQ (hazard quotients) and CR (carcinogenic risk) values. According to the health index of non-carcinogenic risk, 6.1% of the study area sampled posed a potential risk to children rather than adults. Corresponding to the sampled point-wise health risk assessment, 13.05% of the sampled locations are carcinogenic to children. The estimated health risk in

*Correspondence: agyeman@af.czu.cz
Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague, Czech Republic

the agricultural soil was high, with both carcinogenic and non-carcinogenic risks that could threaten persons living in the study area, particularly children.

Conclusion: In general, the continuous application of agriculturally related inputs such as phosphate fertilizers and other anthropogenic activities (e.g., steel industry) can increase the level of PTEs in soils. The use of mean, maximum, and minimum values in health risk estimation does not provide a comprehensive picture of a research area's health state. This study recommends using a sampled pointwise or location health risks assessment approach, which allows researchers to identify high-risk environments that exceeds the recommended threshold as well as areas on the verge of becoming high risk, allowing for rapid remedial action.

Keywords: Health risk, Source apportionment, Ecological risk, Spatial distribution, Principal component analysis, Self-organizing map

Background

Soil contamination suggests the presence of a chemical or foreign substance in concentrations above the normal threshold, which may be detrimental to an organism or humans [1]. This means of environmental pollution has become a primary ecological concern due to the timeless period of potentially toxic elements (PTEs) in nature coupled with the contamination of agricultural soils [2]. Although the more significant part of toxicity has anthropogenic origins, a few contaminants can typically happen in soils as components of weathering of rock deposits, and they can be toxic at high concentrations [3, 4]. Furthermore, contamination of the soil periodically cannot be precisely assessed or seen outwardly, rendering it a latent threat.

Human health depends on a sustainable agricultural sector with minimum human interference, which acts as a forerunner to a sustainable healthy livelihood. However, agricultural soil directly impacts human health, and it is crucial for food safety; PTEs are the most hazardous contaminants due to their build-up in crops [5]. There exists a considerable volume of literature indicating that the accumulation of PTEs in the soil is not exclusively the result of anthropogenic phenomena, but rather the result of a collaborative effort between geogenic and anthropogenic activities [6–8]. Due to the agrochemical and industrial developments, the numerous contaminants are continually progressing. These pollutant varieties tend to form complexes with certain organic compounds in the soil and produce various metabolites through their biological activity. All of these are combined with the soil system and extracted through laboratory analysis. PTEs such as aluminium, arsenic, beryllium, cadmium, lead, mercury, nickel, and radium may have the ability to exude toxic effects that are hazardous to humans, such as carcinogenic effects, teratogenic effects, and endothelial dysfunction [9–11]. According to FAO and ITPS [1], the adverse impact of contaminants from agricultural soils, as they regulate the mobility, bioavailability, and residence of PTEs, depends on their properties, respectively. These pollutants (PTE) have the potential to impact climate, soil, and water, as well as endangering organisms/animals, humans, food security, health, and life [12]. However, according to Zukowska and Biziuk [13], the presence of PTEs in the ecosystem (e.g., vegetable soil) causes them to change from a solid-state to either ionic ligands or, via biomethylation to metallic organic moieties, which can be potentially hazardous to the health of humans, animals, and the entire eco-environment via the food chain. PTEs exhibits potential danger to human health owing to environmental contamination and are classified into two risk categories: carcinogenic and non-carcinogenic risk. Crentsil and Anthony [14] argued that health risk assessment is a high-profile methodology recognized as a valuable, critical method for identifying anthropogenic tendencies that are detrimental to human health. Chen et al. [15] indicated that a detailed understanding of the potential health risks posed by soil PTEs is necessary for informed decision-making by stakeholders to reduce contamination, reduce human exposure and protect humans from risk.

There is no question of the natural source of PTEs in agricultural soil. Regrettably, their increase in agricultural soils is a direct consequence of over-fertilization, which pollutes the soil with PTEs such as Pb, Cd, Zn, Ni, Cu, as well as other polluting sources such as wastewater irrigation (As, Pb, Hg, Cd), compost (Pb, Co, Cd, Zn), pesticide application (Cd, Pb, Cu, Zn), sub-standard fertilizer, and industrial activities (Mn, Ni, As, Pb, Zn, Cr, Cu, Cd) [16, 17]. It has been suggested by Kim et al. [18] and Yang et al. [19] that soil-bound PTEs risk assessment is based on metallic soil content, which may lead to inaccuracy and the necessity for costly remediation of soil. Its important to note, however, that PTEs contamination is not limited to agricultural land. Nevertheless, it may also be detected in living tissues, where it is, for the most part, irreparable [20]. Eziz et al. [21] and Mamut et al. [2] disclosed that PTEs might potentially cause havoc to humans, flora and fauna in the environment. Extensive study has been undertaken in the contemporary era in

the disciplines of PTEs impact on human health, ecological risk, and highlighting environmental impacts [21–25]. Despite the abundance of literature on health concerns published worldwide, there is a dearth of documentation and research in the study area. However, according to Kampa and Castanas [26], health risk assessment is a practical and indispensable tool for recognizing and evaluating the dangers to human health caused by PTEs via various routes of exposure. The active agricultural production and number of industrial activities in the study area make monitoring human health exposure via PTEs critical. Indigenous health is a primary necessity in the study area. Hence a qualitative and comprehensive risk evaluation of agricultural soil health is necessary and appropriate. The primary objective of this paper is to create a digitized soil map that highlights the human-related health risks posed by PTEs, as well as to estimate and map pollution indices outputs, the pattern of PTE spatial distribution, source apportionment, and determine carcinogenic and non-carcinogenic health exposures using a sample location approach. This research will contribute significantly to the awareness of the dangers of PTE exposure in humans and livestock in the study area.

Materials and methods

Study area

The study site is located in the Czech Republic within the district of Frydek Mistek. Rugged terrain and mountains from the exterior Carpathians characterize the study area's geomorphology. The Carpathians, mountains and valleys are differentiated by natural rock and undulating relief. However, there are two mountain ranges in the northern section of the research region partitioned into highlands clusters by river valleys. The district's geological terrain is predominantly carbon-producing, making it an attractive shelter for Paskov and Staříč mining activities that are currently inactive [27].

The study area is characterized by extensive agricultural activity as well as various metal works (such as fabrication, pneumatic cylinders, valves, regulator, etc.) and steel industries (such as the production of cold-rolled steel strips and sheets, anisotropic electrical steel strips and sheets). It is geographically positioned at a latitude of 49°41′0′ North and a longitude of 18°20′0′ East at an altitude of 225–327 m above sea level [8]. Oilseeds, corn, sunflower, and grapevines are among the crops grown in the study area, as is the principal production of cereals such as wheat, oats, barley, and rye. Using the Koppen classification, the study area was classified as Cfb = oceanic temperate climate with high rainfall even during the dry months [28]. Throughout the year, the temperature ranges typically from − 5 °C to 24 °C, with temperatures rarely falling below − 14 °C or rising over 30 °C. The average highest rainfall for the year is 83 mm, with a minimum average total accumulation of 17 mm [29]. The estimated area for this study is 889.8 km^2 extruded from a total land area of 1208 km^2 (39.38% for agronomic activities and forest land 49.36%) for the district of Frydek Mistek. The soil's colour and its structure to its carbonate concentration of the soil's properties may be readily recognized from each other. The prevalent soils in the study area have bleached and paler coloration as well as dark colour in the topsoil. Nevertheless, the parent materials of the soil have a medium and fine texture. In most cases, they are found in aeolian and colluvial deposits, which are also characterized by mottles in the top and subsurface that can be seen in some soil regions, which are usually followed by concrete and whitening. A cambic diagnostic horizon distinguishes them with fine sandy loam texture, a clay concentration of more than 4%, and a lithic discontinuity with reduced carbonate content [30]. Nevertheless, the prevalent soil types in the study are cambisols and stagnosols [30]. These soils predominate the Czech Republic and can be found at elevations ranging from 455.1 to 493.5 m [31].

Soil sampling and analysis

A total of 115 topsoil samples were collected from agricultural land in the district of Frydek Mistek (Fig. 1). The sampling pattern was a standard grid, and the soil samples distances remained 2 × 2 km applying a hand-held GPS (Leica-Zeno 5 GPS) device at 0–20 cm deeper into the soil. The soil samples collected were deposited in Ziploc bags, categorized, and taken to the research laboratory. To obtain a pulverized soil sample, the obtained soil samples were air-dried before being crushed by a machine (Fritsch disk mill pulverize) and mesh sieved (2 mm). In the Teflon container, 1 g of the dried, homogenized, and sieved soil sample (sieve size 2 mm) was placed and labelled. 7 ml of 35% HCl and 3 ml of 65% HNO_3 (use automatic dispensers—a special dispenser for each acid) were dispensed in each bottle of Teflon and gently closed the cap to enable the sample to remain overnight for reactions to take place (aqua regia procedure). The supernatant was placed on a hot metal plate for 2 h to promote digestion of the sample and left to cool when the soil sample was dissolved.

The supernatant was transferred into a prepared 50-ml volumetric flask and then diluted with deionized water to 50 ml. The diluted supernatant was then filtered into 50 ml PVC tubes. In addition, 1 ml of diluted concentration was further diluted with 9 ml of deionized water and filtered into a 12 ml test tube prepared to evaluate the pseudo-total PTE content. The ICP-OES (inductively coupled plasma optical emission spectrometry) (Thermo Fisher Scientific Corporation, USA) was utilized

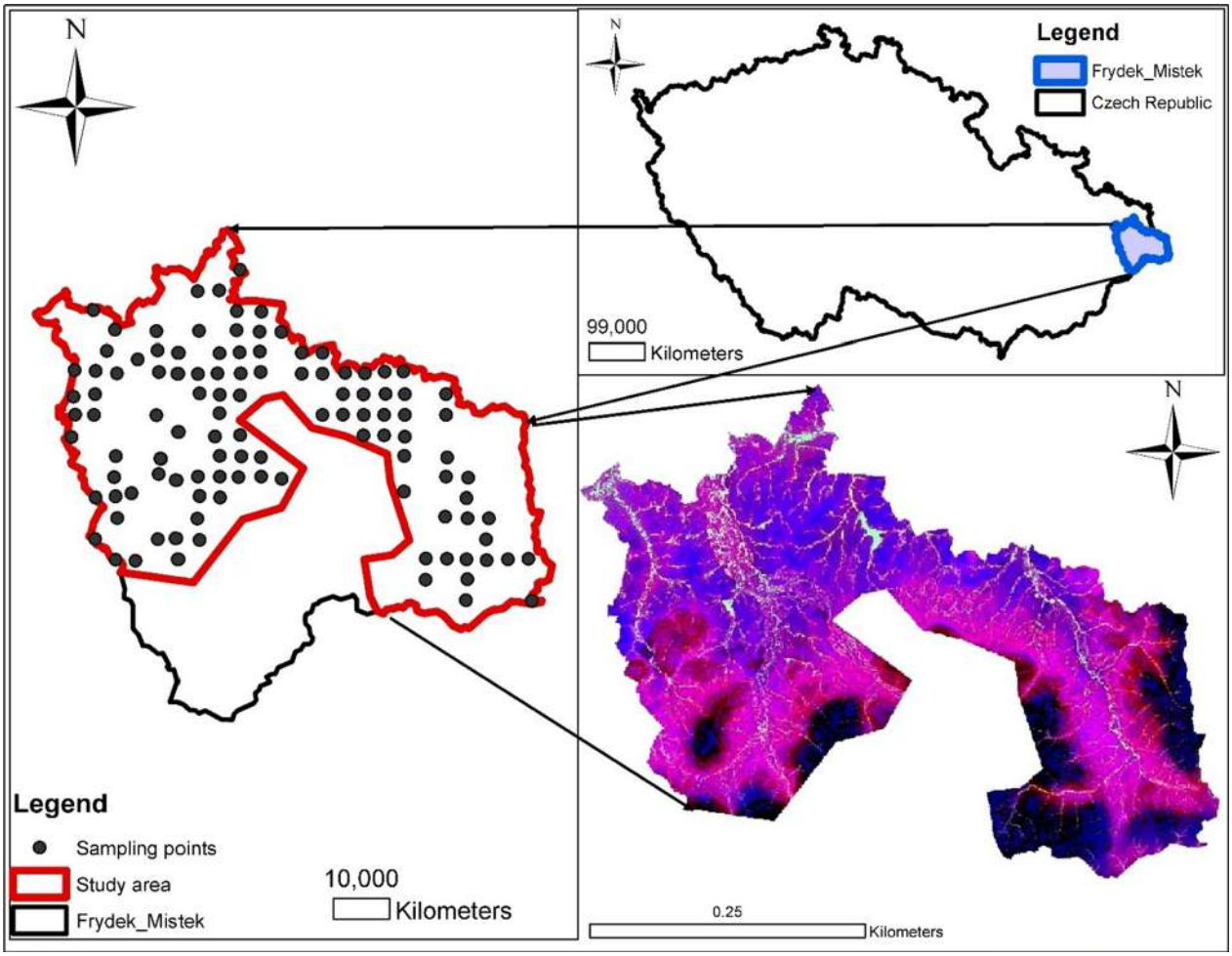


Fig. 1 Location map showing the sampled site with sampling points

to ascertain the concentration of PTEs (Mn, Ni, As, Pb, Zn, Cr, Cu, Cd) in compliance with standard procedures and protocols. Moreover, the quality control and quality assurance process were ensured by checking each sample's standards reference material (SRM NIST 2711a Montana II soil). The detection limits of the PTEs utilized in this study are 0.0002 (Cd), 0.0007 (Cr), 0.0060 (Cu), 0.0001 (Mn), 0.0004 (Ni), 0.0015 (Pb), 0.0067 (As), and 0.0060 (Zn). Duplicate analysis was carried out to ensure that the error was minimized. Pre-treatment analysis of soil samples was conducted at the Czech University of Life Science Prague.

Pollution indices assessment

The productive soil quality of agricultural land must be assessed to evaluate the effects and toxicity of PTE pollution. Based on this, various pollution indices such as the pollution index (PI), the pollution load index (PLI), the comprehensive ecological risk (ER) and the risk index (RI) were utilized to assess the pollution status of the study region. Huang et al., [32] and Sawut et al. [33] argue that indices can reliably measure the quality of soil contamination and the extent to which human activity impacts the soil environment. These indices are widely used in the assessment of PTE contamination in agricultural soil.

Single pollution index (PI)

The single pollution index (PI) is characterized as the concentration of PTE in a sample relative to its geochemical or geological background level. Tomlinson et al. [34] introduced the PI, and the equation is given as

$$\mathrm{PI} = \frac{C_n}{B_n}, \tag{1}$$

where B_n connotes the geochemical background values of the PTEs in the soil (mg/kg) and C_n symbolizes the PTE

concentrations in the soil (mg/kg). The PI precise scale is classified as $PI \leq 1$ (low level), $1 < PI \leq 3$(moderate level), $3 < PI \leq 6$ (considerable level) and $PI \geq 6$ (high level).

Pollution load index (PLI)

The PLI is often used to measure the average amount of soil pollution. This index provides a direct way to display the soil deterioration resulting from the accumulation of PTEs. Tomlinson et al. [34] introduced this equation, and the equation is given as

$$PLI = \sqrt[n]{PI_1 \times PI_2 \times PI_3 \times \ldots\cdots \times PI_n}, \quad (2)$$

where n represents the number of analysed PTEs, PLI is categorized into four classes such as $PLI \leq 1$ (low level), $1 < PLI \leq 2$ (moderate level), $2 < PLI \leq 5$ (high level), or $PLI > 5$ (extremely high level) centred on the intensity of pollution.

Ecological risk assessment (ER and RI)

Ecological risk (ER) is a measure employed to quantify the degree of ecological threat posed by PTE accumulation in soil. The index ER was pioneered and applied by Hakanson [35], and the equation is given as:

$$E_r^I = T_r^i \times PI. \quad (3)$$

The risk index (RI) is defined as the aggregate of each PTE's estimated ecological risk:

$$RI = \sum_{i=1}^{n} E_r^i. \quad (4)$$

The T_r^i is the toxicity response coefficient of specific PTE [34], and the PI represent the single pollution index. The toxicity response coefficient of the PTEs used are 30 (Cd),10 (As), 5(Cu), 5(Pb), 2(Cr), 2(Zn), 2(Ni) and 1(Mn). The ER has 5 classifications: $ER \leq 40$ (low risk), $40 < ER \leq 80$ (moderate risk), $80 < ER \leq 160$ (considerable risk), $160 < ER \leq 320$ (high risk), and $EI \geq 320$ (very high risk). The RI has 4 classes, namely $RI \leq 150$ representing the low risk, $150 < RI \leq 300$ indicating the moderate risk, $300 < RI \leq 600$ signifying the considerable risk and $RI > 600$ representing the very high risk.

Positive matrix factorization (PMF) model

The EPA–PMF v5.0 receptor model [36] is a multivariate receptor modelling approach used to estimate the contribution of the source of PTEs or hazardous substance samples to fingerprints or the composition of the source. The U.S. Environmental Protection Agency utilizes this receptor model, developed by Paatero [37]; Paatero and Tapper [38]. The model does not require any profile source, and all the data are weighted by using uncertainty. According to Norris et al. [39], PMF is used mainly in solving source contributions and source profile that is dataset composition based which is given by this equation:

$$X_{ij} = \sum_{k=i}^{p} \left(g_{ik} f_{kj} + e_{ij}\right), \quad (5)$$

in which p represents the factor number, f the source profile species, g the sample contribution, j and i signifies the quantity of samples and chemical species, and e_{ij} denotes the species.

This equation determines the contribution as well as profile factors:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{\varepsilon_{ij}}{u_{ij}}\right)^2, \quad (6)$$

in which m represents the quantity of analysed PTEs, n denotes the number of sampled soils, and U_{ij} refers to the uncertainty of PTE j in soil sample i. The parameters used to determine the uncertainty U_{ij} and the minimum Q were previously defined by the authors [8].

Health risk assessment

The ever-growing human population and human endeavour to ensure that the planet remains a haven for humanity are under constant constraint. Frequently, scientists, policymakers, and other stakeholders push the frontiers of research in many ways, and no matter the initiative and the best course of utilizing research, the world is now and then polluted. Humans are exposed to PTEs in three different forms every day, including inhalation, ingestion, and dermal contact. There are three procedures to assess the probability of human PTE exposure in peri-urban, urban, and rural settings, according to Wang et al. [40]. PTE exposure pathways to humans are calculated using the following equations:

$$CDI_{ing} = \frac{C \times IR_{ing} \times EF \times ED}{BW \times AT} 10^{-6}, \quad (7)$$

$$CDI_{inh} = \frac{C \times IR_{inh} \times EF \times ED}{PEF \times BW \times AT}, \quad (8)$$

$$CDI_{derm} = \frac{C \times SA \times AF \times ABS \times EF \times ED}{BW \times AT} \times 10^{-6}, \quad (9)$$

$$CDI_{total} = CDI_{ing} + CDI_{inh} + CDI_{derm}. \quad (10)$$

Additional file 1: Table S1 contains the definitions of the variables CDI_{ing}, CDI_{inh}, and CDI_{derm}, as well as reference values for the indices of the preceding Eqs. (7–10).

Non-carcinogenic risk assessment

The equation of the potential non-carcinogenic risk for a single PTEs was computed as the HQ (hazard quotient), which is given as Eq. (11):

$$HQ = \frac{CDI}{RfD}. \quad (11)$$

RfD (see Additional file 1: Table S1) represents the reference dose (mg/kg/d), and it is the estimated daily human population exposure. The determination of a particular health hazard of all the PTEs analysed was done by computing HQ values. The sum of the values was reported as the HI (hazard index), which is provided as Eq. (12) [41]:

$$HI = \sum HQ = HQ_{ing} + HQ_{inh} + HQ_{derm}, \quad (12)$$

whereby HQ_{ing}, HQ_{inh} and HQ_{derm} represent the hazardous quotient for inhaling, ingestion and dermal, respectively. A report from USEPA [42] explicitly outlined that when the HI < 1, it presupposes that there is a potential to negatively impact health if PTEs are exposed to humans. However, Eziz et al. [21] mentioned that if HI > 1, there is also the propensity for non-carcinogenic health risks to emerging.

Carcinogenic risk assessment

According to the USEPA's [41] findings, the possibility of developing cancer of any sort may be ascribed to humans being exposed to carcinogenic risk (CR). Equations (13 and 14) were employed to evaluate the carcinogenic risk of PTEs such as As, Ni, Pb, Cd, and Cr:

$$CR = CDI \times SF, \quad (13)$$

$$TCR = \sum CR = CR_{ing} + CR_{inh} + CR_{derm}, \quad (14)$$

in which the variables TCR, CR, and SF reflect total carcinogenic risk (no unit), carcinogenic risk (no unit), and slope factor for carcinogenic PTEs (mg/kg/d), respectively. TCR values should be in the range of $1 \times 10–6$ to $1 \times 10–4$. That is a reasonable standard that demonstrates no considerable risk to human health [43]. All the exposure factor values utilized in the health risk calculation are listed in Additional file 1: Table S1.

Analysis of data

The data were statistically analysed using kyplot for principal component analysis, RStudio for projected principal component loadings, EPA-PMF 5.0 to estimate source apportionment, and excel in quantifying the potential health risk as well as Pearson correlation matrix. PTE modelling, spatial distribution maps, and health risk assessment were interpolated using ordinary kriging in an R software environment. The factor scores of the PMF receptor model were likewise mapped using a self-organizing map (SeOM).

Kohonen [44] created SeOM by combining an artificial neural network with unsupervised learning techniques for organizing, evaluation, and predictions. SeOM was employed in this study to visualize factor score contribution as well as determine the number of clusters within the factor scores of the PMF receptor model in an agricultural urban and peri-urban soil. The SeOM assessment data act as an input dimensional vector variable [45, 46]. Melssen et al. [47] defined a neural network as having a single input layer that connects an input vector to a vector output with a unitary weight vector. SeOM generates a two-dimensional map composed of several neurons or nodes knitted together into a hexagonal, circular, or square topological layout based on their closeness [45]. Based on metrics, topographic error (TE) and quantization error (QE), map sizes were examined, and a SeOM model with 0.086 and 0.904 was chosen as a 55-map unit (5×11). The neuron structure was selected based on the empirical equation node number, which was given as:

$$m = 5 \times \sqrt{n}, \quad (15)$$

in which the m denotes the quantity of SeOM map neurons, n representing the input data quantity.

Results and discussion

PTEs concentration in soil

Statistical standards such as mean, median, skewness and kurtosis, standard deviation were employed to detect the PTEs concentration levels in the sampled soil (see Table 1). Table 1 includes PTEs estimated mean concentrations of the UCC (upper continental crust), WAV (world average values), and EAV (European average values) reported by Kabata-Pendias [48]. PTE concentrations (Zn, Pb, Mn, Cr, Cu, As, Ni, Cd) varied from 186.02 to 1691.76 mg/kg (Mn), 37.48 to 272.18 mg/kg (Zn), 9.56 to 155.69 mg/kg (Pb), 10.9 to 62.78 mg/kg (Cr), 7.88 to 62.62 mg/kg (Cu),4.86 to 42.39 mg/kg (Ni),1.85 to 30.42 mg/kg (As) and 0.61 io 7.28 (Cd) mg/kg. In the agricultural soil, the concentration of PTEs declined in the following order: Mn > Zn > Pb > Cr > C u > Ni > As > Cd > (see Table 1). The general PTEs mean

Table 1 PTE concentrations in the study site, basic data, toxic element, and geochemical background levels (number sample 115 per each PTE)

	PTEs (mg/kg)							
	Mn	Ni	Pb	Zn	As	Cd	Cr	Cu
Mean	699.03	16.15	33.86	85.22	5.32	1.84	28.43	22.54
Median	664.39	13.75	30.10	75.47	4.57	1.61	26.90	19.68
Local background value[A]	–	30.00	50.00	80.00	–	0.20	70.00	25.00
Finland[B]	–	60.00	60.00	150.00	10.00	0.50	100.00	100.00
Austria[C]	–	35.00	30.00	100.00	–	0.40	54.00	35.00
Spain[D]	–	25.50	26.50	57.00	14.00	–	57.00	17.50
Sweden	411.00	13.00	18.00	65.00	3.80	0.17	22.00	17.00
Japan	–	26.00	24.00	89.00	–	0.33	58.00	48.00
Brazil	535.00	25.00	22.00	73.00	–	0.18	86.00	109.00
USA	550.00	19.00	19.00	60.00	7.20	< 0.01–41	54.00	25.00
UCC[b]	900.00	20.00	15.00	70.00	1.80	0.10	100.00	17.30
WAV[a]	488.00	29.00	27.00	70.00	6.83	0.41	59.50	38.90
EAV[c]	524.00	37.00	32.00	68.10	11.60	0.28	94.80	17.30
Minimum	186.02	4.86	9.56	37.48	1.85	0.61	10.90	7.88
Maximum	1691.76	42.39	155.69	272.18	30.42	7.28	62.78	62.62
Range	1505.74	37.53	146.13	234.70	28.57	6.67	51.88	54.74
Standard Deviation	259.35	6.78	18.51	34.35	4.95	1.01	9.38	9.98
Kurtosis	1.37	2.49	18.80	7.32	11.77	10.45	2.69	4.90
Skewness	0.79	1.63	3.67	2.11	3.04	2.84	1.33	2.04
CV %	39.04	49.29	61.51	45.52	108.23	62.86	34.88	50.71

[a] World average value (WAV)

[b] Upper continental crust (UCC)

[c] European average value (WAV), [48] (page 41 and 42), coefficient of variability (CV) A [51], B [109], C [110], D [111]

concentration in the soil was relatively high than the EAV threshold, specifically Pb, Mn, Zn, Cd, and Cu. In the current study, the mean value of cadmium is 6.57 times greater than that of EAV (see Table 1), as are the concentrations of Mn (1.33), Pb (1.06), Cu (1.06), and Zn (1.06), (1.33). Although there may be a geogenic source, evidence suggests that anthropogenic activities significantly compensate for the elevation in PTE concentrations. Toth [49] reported some PTEs thresholds from the Ministry of Environment Finland (MEF), and As (5.0 mg/kg) and Cd (1.0 mg/kg) threshold limits were lower than the current study's corresponding PTEs. The mean concentration of Cr (100.0 mg/kg), Cu (100.0 mg/kg), Pb (60.0 mg/kg), Ni (50.0 mg/kg), and Zn (200.0 mg/kg) from the MEF threshold limits, on the other hand, was found to be greater than the respective PTEs mean concentration from the current study.

Conversely, the mean concentrations of the following PTEs, like Pb, Mn, Zn, and Cd, in our current study are similarly greater than the same PTEs from the world average value (WAV) threshold limit [48] (see Table 1). PTEs (Mn, Pb, Zn, and Cd) concentration levels in the current study are 1.48, 1.25, 1.21, and 4.49 times greater than WAV concentration threshold. Similarly, when the PTEs studied mean concentration values were compared to the PTEs of the UCC (Table 1), it was discovered that Zn, Pb, Cd and As are higher than the respective PTEs in the UCC. Comparatively, the mean concentration levels of Zn, Pb, Cd and As surpassed those of UCC by 2.26, 1.55, 3.57 and 18.4 times respectively.

The present study indicated that elevated values of some PTEs than those of UCCs indicate that anthropogenic sources play a vital role in pollution. The assertion is compatible with Jia et al. [50] point of view. The present mean concentration of PTEs in the current research relative to the mean concentration of PTEs in Sweden [48] suggests that the concentration of PTEs exceeded the threshold limits of the PTEs in Sweden (see Table 1). A comparable comparison to PTEs concentration levels reported in Brazil and the United States [48] shows that the following PTEs, Mn, Pb, Zn, and Cd, are lower than those reported in the present study (see Table 1). Comparing the concentration values of PTEs with those obtained from Japan [48] revealed that most of the PTEs under analysis were lower than those from Japan, except

for Pb and Cd (see Table 1). The current mean values of the PTEs (Ni, As, and Cr) in the study area were found to be lower when compared to the agricultural soil threshold limits in Finland, Austria, and Spain. On the other hand, Pb, Zn, and Cd levels in the current study were higher than the respective PTEs from Spain, but lower than the corresponding threshold limits from Finland and Austria. Cadmium mean concentrations in the present study were significantly higher than the threshold limit (see Table 1) from Finland and Austria. Predicated on the Nemecek and Podlesakova [51] report, the local background values (LBV) for Ni, Pb, Cr, and Cu from the district of Frydek Mistek revealed that the mean concentration of PTEs was within the permissible threshold limit (see Table 1). However, the mean concentrations of Cd and Zn in the study were higher than the local background values reported by Nemecek and Podlesakova [51] (see Table 1).

The estimated standard deviation values were high due to the concentration of PTEs with high variable heterogeneity in the study region. The computed skewness values were used to determine the normality and abnormalities of the distribution of PTEs values. According to Chandrasekaran et al. [52], if the PTE skew value ranges from 1 to − 1, it can be viewed as a regular distribution. Notwithstanding, if the PTE value is slightly skewed positively (> 1), the distribution is anomalous. The calculated kurtosis and skewness values were usually greater than 1; thus, the distribution of PTEs is believed to be irregular, skewed in the right direction and leptokurtic.

The CV (coefficient of variation) represents the extent of heterogeneity within PTE concentrations, pursuant to Karimi Nezhad et al. [53]. If the CV is between 0 and 20%, it is assumed that the PTEs are from a natural source, and if it is greater than 20%, it indicates the influence of anthropogenic activity. As a result, a CV of 20% shows low variability, a CV of 50% indicates moderate variability, a CV of 50% indicates significant variability, and a CV of 100% suggests extraordinarily high variability. The CV of the PTEs in the present agricultural soils declined in the following order As > Cd > Pb > Cu > Ni > Zn > Mn > Cr. The results evidenced that the PTEs Cr, Zn, Mn, and Cr are moderately variable and homogeneous. The high variability of Cd, Pb, and Cu inferred a non-homogeneous variability of PTEs, clearly indicating that the possible human-related influence. Arsenic (As) showed a very abnormal CV suggesting an exceptionally high variability. According to the distribution of Cd, Pb, and Cu non-spatial homogeneity, there is a likely local source of enrichment substance.

Table 2 Principal component illustrating the contributions of PTEs in the study area

PTEs	PC1	PC2
Mn	0.621	0.195
Ni	0.497	0.709
Pb	0.877	− 0.306
Zn	0.872	− 0.076
As	0.788	− 0.491
Cd	0.907	− 0.311
Cr	0.501	0.288
Cu	0.577	0.667
Eigenvalues	4.191	1.506
% variance explained	52.38	18.83
cumulative % total		71.21

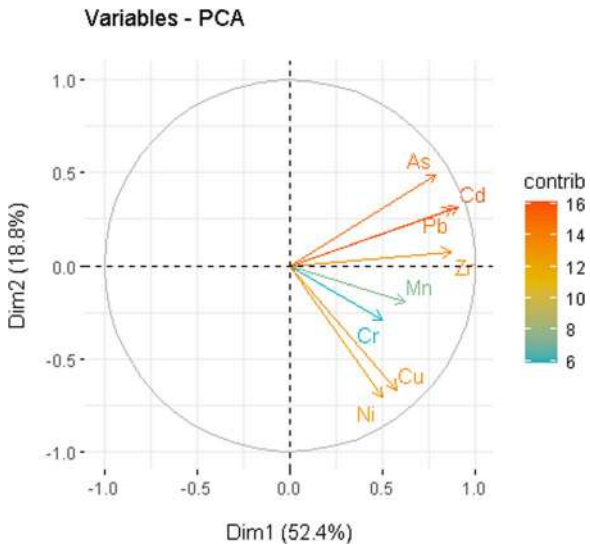


Fig. 2 Showing principal component contribution loadings projections

Chemometric approach

Multivariate analysis of PTEs

The primary source of pollution in the study area was detected utilizing principal component analysis (PCA). It is a supportive approach that can make valuable suggestions about PTE paths and primary sources [54]. The loadings of the principal components (PCs) extracted from the principal correlation values were fixed at or above 0.50 in this study (Table 2; Fig. 2). Following the criterion, PC 1 and PC 2 were statistically significant, accruing 71.21% of the data variance. PC1 explained 52.38% of the variance explained by the PTEs Pb, Zn, As, Mn, Cr, and Cd, in that order. According to the report in Table 2, some of the PTEs (Pd, Zn, As, and Cd) in PC1 had a

strong positive load ranging from 0.7 to 0.9, while other tenants, such as Mn and Cr, had a moderate positive load (0.5–0.7). This indicated that PC1 concentrations might be attributable to a variety of sources, including anthropogenic and parental material components. Agrochemicals such as lead arsenate herbicides or pesticides, which are essential sources of agricultural soil chemicals, are agronomically linked to As and Pb [55]. Existing studies by Nicholson et al. [56] and Luo et al. [57] established that livestock manure and fertilizer are important sources of As and Pb. The findings of the current study support this statement. The origins of Zn and Mn (*r* values = 0.872 and 0.621, respectively) may be traced back to the convergence of anthropogenic and geogenic sources (liming). According to Mantovi et al. [58], Zn concentration in soil surges may be linked to the application of waste resulting from animal husbandry and phosphate fertilizers. Cd and Cr accumulation in soils are related to the forging of metal, sewage and chemical fertilizers [59–62]. PC2 (18.83% of the overall variance) showed relatively high positive loading for PTEs such as Cu and Ni. As a result, Cu and Ni have a comparable source of pollution. Cu concentrations in topsoil are probably caused by fertilizers, other agricultural pollutants, and urban waste [63]. The presence of Nickel (Ni) in soil originates from both the parent material (lithosphere) and the anthropogenic deposition [64].

The correlation matrix (see Table 3) among the studied PTEs demonstrated the existence of a relationship between the PTEs. PTEs correlation revealed a strong relationship between the PTEs. Pb (lead) and Zn (zinc) demonstrated a strong positive connection with PbAs (*r*-value = 0.75), AsCd (*r*-value = 0.9), CdPb (*r*-value = 0.85), and CdZn (*r*-value = 0.78). As a result, it is critical to emphasize that they may have the same or nearly analogous origins. Other correlations between PTEs, such as As and Zn (*r*-value = 0.63) and Ni and Cu (*r*-value = 0.69), likewise exhibited a robust nexus, indicating that the pollution cause might be correlated or close together. Cd and As had the highest correlation value, while Ni and As had the least positive correlation (*r*-value = 0.07). All the PTEs had a positive relationship and no negative correlation.

Spatial distributions of PTEs in the study area

The PTEs spatial distributions in the study area are depicted in Fig. 3. As and Cd shared the same distribution map pattern, likewise Cr and Mn. The distribution pattern of As and Cd primarily was centred in the eastwards and the south-eastern area of the map. The map shows hotspots around the eastern (i.e. the steel industry) and the south-eastern part, but the As distribution map appears to be denser than Cd. Spatial variability of Cu and Ni showed hotspots across the northwestern, southwestern, and south-eastern parts of the map. The source distribution of Cu and Ni spatially in the map may be attributed to the steel industry and agrochemicals; this is coherent with the earlier study carried out by Krishna and Govil [65].

Moreover, Salonen and Korkka-Niemi [66] identified certain PTEs such as Ni and Cu as minute spatial and temporal distribution in world soils present in parent soil materials. Furthermore, Cr and Mn showed more undulated spatial distribution across the entire map except for the south-eastern part that looks relatively clean. Cr spatial variability appears to be denser than that of Mn. The abundance of Cr is caused by a variety of human-related activities such as electroplating. In addition, the industrial utilization of chromium in alloy creation, such as the steel industry and sewage discharge, are responsible for the Cr hotspots on the map. According to Goovaerts [67], the source of PTEs such as Cr, the geochemical/geological background of Cr is normal in generally. Nonetheless, its accumulation in agricultural soils may well be altered by anthropogenic sources related at times. Even though Mn is naturally occurring, the regular injection of manganese sulphate to farmland to boost yields in plants such as veggies and beans continuously raises

Table 3 Showing the correlation matric between PTEs

	Mn	Ni	Pb	Zn	As	Cd	Cr	Cu
Mn	1.00							
Ni	0.24	1.00						
Pb	0.42	0.21	1.00					
Zn	0.38	0.45	0.83	1.00				
As	0.38	0.07	0.75	0.62	1.00			
Cd	0.43	0.30	0.85	0.78	0.90	1.00		
Cr	0.49	0.27	0.28	0.27	0.25	0.34	1.00	
Cu	0.40	0.69	0.35	0.44	0.16	0.31	0.29	1.00

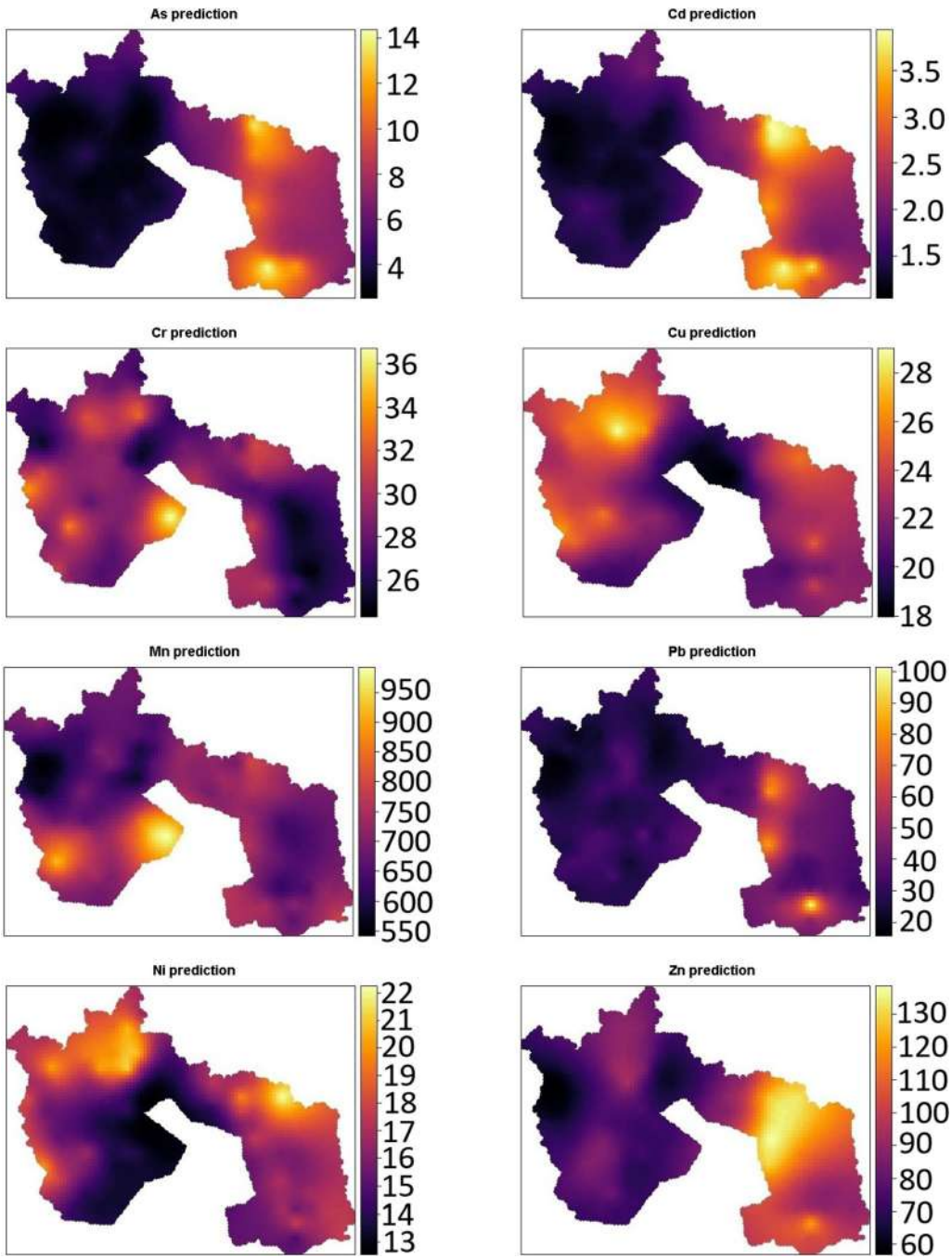


Fig. 3 The spatial distribution of PTEs in agricultural soil

the concentration of PTEs [68]. The eastern and southeastern areas of the map exhibited a sectorial spatial pattern distribution of Zn and Pb. The distribution of Pb and Zn spatially is linked directly to fertilizer application on farmlands, vehicular traffic, steel industry, and fuel knocking, which is consistent with the preceding study by Rodriguez et al., [69] stating that elevated levels of Zn and Pb in cultivated soil are due to anthropogenic factors composed by human-associated activities.

Source apportionment via PMF

The source apportionment of PTE contributions was performed applying the PMF receptor model, and the total number of samples included in the PMF analysis for each

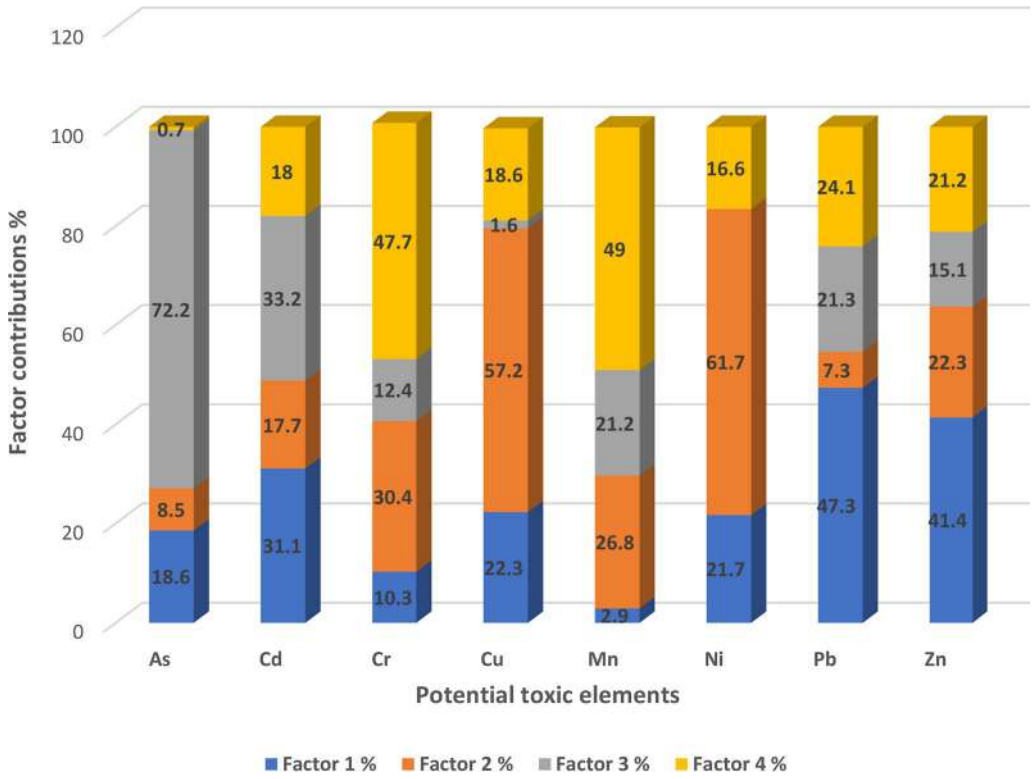


Fig. 4 Factor profile indicating the proportional contribution of PTEs from PMF receptor model

PTE was 115 (see Fig. 4). The minimum Q controls the residual matrix, which guarantees that an acceptable number of factors are generated. The PMF discharged factors loading that ran for twenty iterations, and all of the minimal Q converge in the current paper. Among the 20 iterations, run 14 was chosen to discharge the factor loadings and the proportional contributions of each PTE in the study. For a PTE to dominate a factor loading, the percentage dominant was fixed at 40% or more.

Factor 1 provided high factor loadings values comprising Pb and Zn (47.3% and 41.4%, respectively). The predominance of Pb and Zn in agricultural soil can be traced primarily to several sources. The dominant PTEs (Pb and Zn) in factor 1 are principally anthropogenic origin, evidenced by the projected principal component contribution loadings (Fig. 2) and have a strong correlation. They have elevated mean concentrations above the regulated thresholds, that is, WAV and EAV. Chakraborty et al. [70] and Khosravi et al. [71] reported that Zn and Pb are the principal PTEs pollutants in peri-urban and urban agricultural soil. The high level of Pb in the agricultural fields may be attributable to vehicular traffic, abrasion of tyres, knocking of fuel, and a limited geogenic source. Earlier reports from Tepanosyan et al. [72] and Li et al. [73] suggested that Pb accumulation in the soil may be attributable to automotive traffic, fuel knocking, and abrasion tyres. Similarly, Arditsoglou and Samara, [74] Hjortenkrans et al. [75] and Guan et al. [76] reported that Pb is deposited throughout agricultural fields via road networks used by automobile machines that connect vicinities, suburbs, and farmlands, where automobiles, agricultural-based machinery, and discharge equipment's which is Pb-containing exhaust, triggering soil pollution. Nevertheless, the source of Zn in the soil might be accredited to the steel industry within the study area and the wearing of vehicular tyres. Al-Khashman and Shawabkeh [77] and Wang et al. [78] recounted that the level of Zn in the soil might be attributable to the steel industry and tyre wearing. The metal and steel industries employ a high amount of Zn, which is generally used as an anti-corrosive agent in other metal products and exhibits galvanizing and alloy forming properties. Therefore, factor 1 source of pollution will be ascribed to the blend of the steel industry and vehicular traffic.

Factor 2 was controlled by Cu and Ni, accounting for 57.2% and 61.7% factor loading, respectively. The hotspots on Factor 2 spatial distribution map indicated that the Cu and Ni hotspots in the northwest enclave originated primarily from agricultural activities. The PCA projection in Fig. 2 suggests that the source of Ni and Cu

in the farmlands is mostly linked towards the geogenic origin. Nevertheless, Cu excess beyond the EAV also hints at a collaborative effort between geogenic sources and anthropogenic sources such as livestock manure. Copper accretion is correlated chiefly to cattle manure [79] because the confluence of Cu and Zn functions as a complement (anti-bacterial agent in the gut) [80], which boosts microbial activity and also modulates weaning patterns [81]. The application of Cu-rich manure (especially from livestock like pigs) and phosphate-based fertilizers, according to Cheng et al., [82] and Xiong et al. [79], may perhaps ultimately lead to Cu accumulation in agricultural soil. Even though Ni concentrations in agricultural area may be attributed to a geogenic source, the high levels recorded at specific sample locations are confirmed by the spatial distribution map (Fig. 3), demonstrating that the steel industry is the polluting catalyst. According to Al-Khashman [77] and Harasim and Filipek [83], the steel and metal industries, food processing, tyre wear, vehicular traffic, and corrosion appear to be the sources of Ni contamination. Numerous reports like Li et al. [84] and Chen et al. [85] have indicated that Ni arises through manufacturing activities such as steel manufacturing and metal processing. Ni plays a vital function in the creation of alloys such as nickel stain (a tin and nickel alloy), silver (a copper, nickel, zinc) and nickel bronze alloy (a tin and copper solution). Factor 2 pollution will primarily be attributable to geogenic sources, which will be bolstered by steel production industries and livestock manure.

Factor 3 was overshadowed by As, which had a source contribution of 72.2%. Most insecticides, herbicides, and pesticides, like sodium arsenate, calcium arsenate and lead arsenate, are high in arsenic and used in a variety of agricultural applications. Bhattacharya et al. [86] discovered that agrochemicals of this sort, such as sodium arsenate, calcium arsenate and lead arsenate, are high inorganic As. In previous research, Micó et al. [87] and Nicholson et al. [56], Jayasumana et al. [88, 89] suggested that the potential sources of As-enrichment in soil are agrochemicals. Furthermore, Liu et al. asserted that animal wastes containing organo-arsenic feed additives constitute a significant source of arsenic pollution in agricultural fields due to concentrated animal feeding activities. Factor 3 source of pollution will be ascribed to agrochemicals.

Cr and Mn controlled the final factor (factor 4) with a contribution load of 47% and 49%, respectively. Thus, the chromium concentration in the agricultural field might be attributed to a geogenic source. However, in some sampled locations, excesses based on maximum values also point to an anthropogenic source supplementing the geogenic source. In addition, the consistent application of phosphate fertilizer to the soil during each crop season introduces Cr into the soil, raising the concentration of Cr in farmlands. Liu et al. [90] recounted that the concentration of Cr per bag of phosphate fertilizer ranges from 30 to 3000 mg/kg. Nonetheless, current literature by Zhang et al. [91] indicated that high-level Cr concentrations in cultivated soils that exceed the permissible threshold limit are not limited to agro-related sources but rather a blend of parental material and anthropogenic sources. The mean concentration of Mn in the current agricultural soils is 1.43 and 1.33 times greater than the WAV and the EAV permissible threshold. This suggested that the high levels might be attributed to a diverse source such as the steel industry and fungicides. According to Bradl [92] Mn is used in the steel industry to produce ferromanganese steel. However, Shaw [93] reported that fungicides had been an integral component of plant disease management regimens for agronomic crops. Fungicides are applied to agricultural fields to prevent or limit the spread of fungus-caused disease. The successive application of manganese-based fungicides such as foliar fertilizers to increase yield elevates Mn concentration in agricultural soil. Factor 4 source pollution will be linked to a geogenic source that is actively augmented by the steel industry and fungicides.

The shown self-organizing map (SeOM) illustrates the concentration of PTEs in the PMF factor loadings as component planes composed of individual neurons. The component plane exhibited diverse colour patterns, as shown in Fig. 5. Based on the number of samples used in this study, the suitable neurons per mapped map was 55. The SeOMs were created with various colours, and the more similar the colour pattern, the more identical the sample attributes. Factor 1 and 3 components plane bore a striking resemblance in colour to the neighbour distance plot (U–Matrix). Factor 1 component plane was loaded with dominant PTEs such as Pb and Zn with a single high neuron on the left side of the map on the sixth block of neurons. Factor 2 component plane was loaded with the dominant PTEs Cu and Ni, exhibiting moderate to high neurons. The high neuron was envisaged on the fourth block of neurons of the map. Factor 3 was loaded with dominant PTE such as As, and the high neuron equally was seen on the Fourth block of neurons. Factor 4 was controlled by the PTEs like Mn and Cr, and the SeOM displayed a variety of colour shades from mild to moderate, moderate to a high neuron. The high neuron was seen on the tenth neurons block.

Generally, a redder colour neuron was displayed in the SeOMs. The component plane of the factor1 SeOM map showed a hotspot for the dominant PTEs. The proportion of Pb, Zn, Cd, and As is predominantly anthropogenic in origin, accounting for 70.76% of factor 1 loading, confirming that SeOM for the factor 1 component plane is

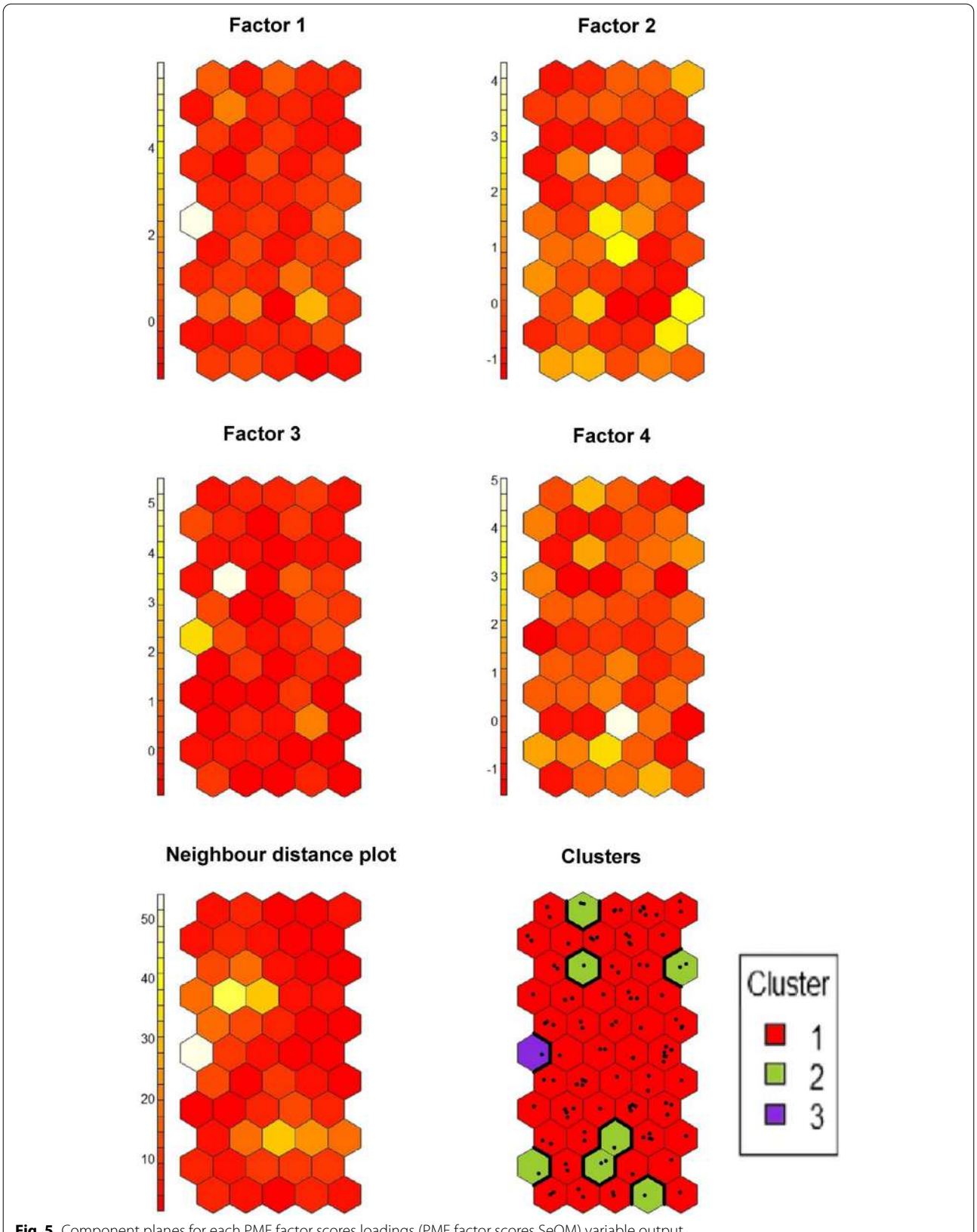


Fig. 5 Component planes for each PMF factor scores loadings (PMF factor scores SeOM) variable output

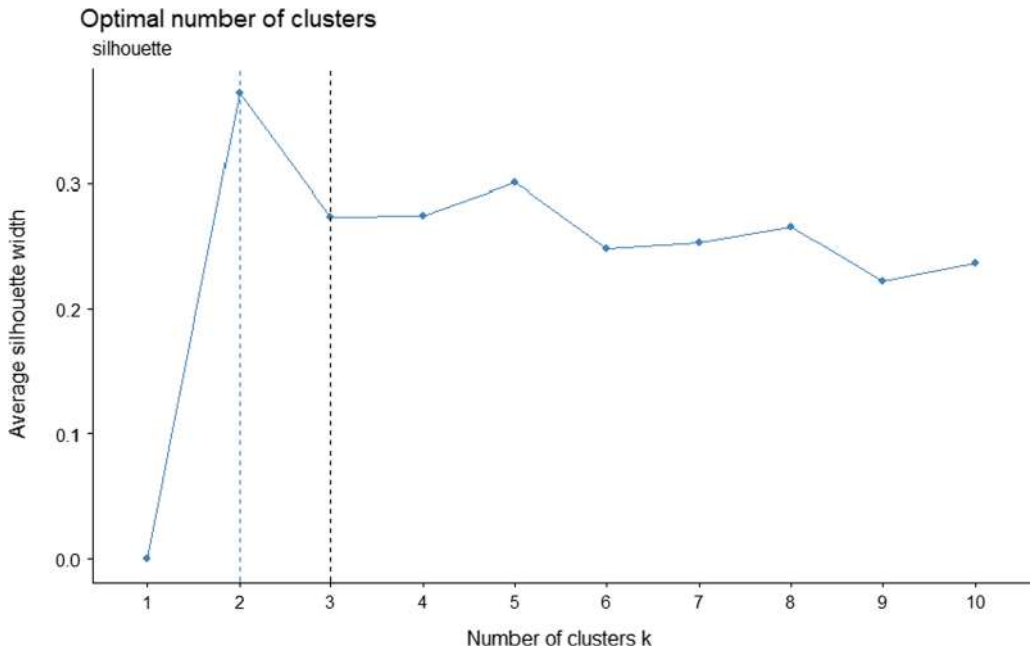


Fig. 6 Number clusters classification components

primarily anthropogenic. Factor 2 percentage proportion was dominated by geogenic PTEs (Ni, Cu, Mn, and Cr), which accounted for 75.94% of the cumulative variation. The factor 2 component plane was more geogenic due to the significant percentage proportion accrued by Ni, Cu, Mn, and Cr in factor 2 loading. Based on the PTEs (Pb, Zn, Cd, and As) percentage proportion (80.11%) accumulated, the factor 3 SeOM component plane is more anthropogenic. Based on the percentage proportion (67.33%) accrued by the PTEs (Ni, Cu, Mn and Cr), factor 4 component plane was ascribed to geogenic source.

K-means (silhouette) on the training map resulted in three distinct clusters (1–3). The partitioned three clusters developed using the K-means technique are displayed in distinct colours that correspond to the U-Matric component plane boundaries. Based on the silhouette technique (see Fig. 6), the cluster was ideal. The four-factor component planes represent the four-factor loadings in the PMF receptor model, which is simplified to allow for appropriate clustering interpretation [94]. The clustering of the 115 observation points allotted sampled points as follows; cluster 1 gathered the most soil samples, 102, out of a total of 115, cluster 2 received 12 samples, but Cluster 3 only obtained 1 (see Fig. 6). Due to the diverse anthropogenic and natural processes that influence soil formation, it is complicated to have appropriately differentiated cluster patterns in the distributed map [95].

Contamination assessment of PTEs based on local background (LBV) and European average values (EAV)

LBV and EAV were the geochemical background levels used in assessing pollution levels in the study area. The PTEs employed in the LBV, on the other hand, were Cd, Cu, Cr, Ni, Pb, and Zn, and the EAV As, Cd, Cu, Cr, Ni, Mn, Pb, and Zn. However, a comparison of pollution levels based on PI, PLI, ER, and RI was performed using the associated PTEs in both background levels.

Additional file 1: Table S2 shows the calculated single pollution index (PI) for the EAV, and the results suggested that the pollution level of the PTEs ranged from low to high. Mn pollution was observed in 22 of the 115 soil samples and As in 86, when PI was measured using EAV. Some of the locations sampled had a moderate level of pollution, and 92 of the areas sampled had a moderate level of Mn and As (67). (i.e. using the EAV). Manganese and arsenic pollution levels were high in a single observation location (sampled point 18 for Mn) and in 3 sampled areas for As. The PI for the following PTEs was estimated using both EAV and LBV as the geochemical background values: Cd, Cu, Cr, Zn, Ni, and Pb (see Additional file 1: Tables S2 and S5). Nickel, lead, zinc, chromium, and copper levels were low when EAV was used as the geochemical background level in 106, 44, 44, 113, and 108 sampled locations. Based on LBV as the geochemical background level, the number of sampled locations 106, 105, 61, 115 and 84 exhibited low pollution levels for the following PTEs Cu, Cr, Zn, Ni, and Pb. In EAV, Ni, Pb, Zn, Cd, Cr,

and Cu showed moderate pollution levels in 9, 67, 70, 29, 2, and 7 sampled locations, respectively, whereas, in LBV, moderate pollution levels were found in 9, 9, 27, 84 and 53 sampled locations, except for Cr, which showed none. In comparison, Pb, Zn, and Cd pollution levels estimated using EAV suggested that Pb, Zn, and Cd exhibited a considerable pollution level in 4, 1 and 73 sampled locations, respectively, whereas using LBV just a single sampled location showed a significant pollution level for Pb and Zn, and 27 for Cd. The number of sampled locations with high levels of Cd pollution in the study area was 13 based on EAV and 88 based on LBV.

The estimated pollution load index (PLI) exhibited a varied response for both background levels; however, all background levels revealed low pollution levels in 104 locations for EAV and 103 locations for LBV. Furthermore, for 7 locations, both background values showed moderate pollution. Only 3 locations had high pollution levels for EAV and 4 for LBV, but one (sample point 104 with PLI value 498: see Additional file 1: Table S4) had very high pollution levels for both background levels. The spatial distribution patterns of the pollution level based on both background levels were comparable (Fig. 7). The PLI maps revealed moderate pollution levels with patches of the hotspot and low spots in the south-eastern part of the map. These hotspots are consistent with the observed high PI values.

The ecological risk (ER) approach was utilized to examine the influence of various PTEs on cultivated soils. Except for Cd, the results of the ER assessment of cultivated soil samples indicated a low-risk analysis for all PTEs (Ni, Pb, Zn, Cr, and Cu) in both background levels used (see Additional file 1: Tables S3 and S6). Based on the background levels application, 15 of the 115 analysed locations revealed a moderate ecological risk level for the EAV, but none showed a moderate ecological risk level for LBV. On the other hand, 13 locations exhibited a considerable ecological risk level for LBV and the EAV 77 observed locations. Based on the background levels, the high ecological risk level was 78 for the LBV and 19 for the EAV. In contrast, both background levels exhibited very high ecological risk in 24 for LBV and 4 for EAV for the background levels in sampled locations, respectively.

The calculated risk index of the study region also indicated that 3 sampled locations had low ecological risk levels for LBV and 64 for EAV (see Additional file 1: Table S4). Relatively, the LBV and EAV exhibited moderate ecological risk levels in 70 and 44 sampled locations, respectively. Similarly, the risk index based on the application of the LBV as the geochemical background level revealed that 36 of the sampled locations were considerably risky, whereas the EAV, 6 sampled locations were considerably risky. Only 6 of the sampled locations had a very high ecological risk for the LBV, while EAV a sample location exhibited high ecological risk. The RI-OK (risk index ordinary kriging) spatial distribution map revealed that the majority of the risk-prone areas were in the northeastern and southwestern parts of the map for the potential ecological risk index based on EAV (PERI-EAV) and the northwestern to southwestern parts of the map for the potential ecological risk index based on LBV (PERI-LBV) (Fig. 8). According to the maps, the underlying cause of pollution in that region may be mostly traced to industrial and agricultural activities. The PLI and RI values for agricultural soils in Frydek Mistek's district indicated that pollution levels range from low PTE pollution to very high pollution risk. As a result, it's critical to identify PTE pollution sources on agricultural soil. LBV use is recommended, particularly

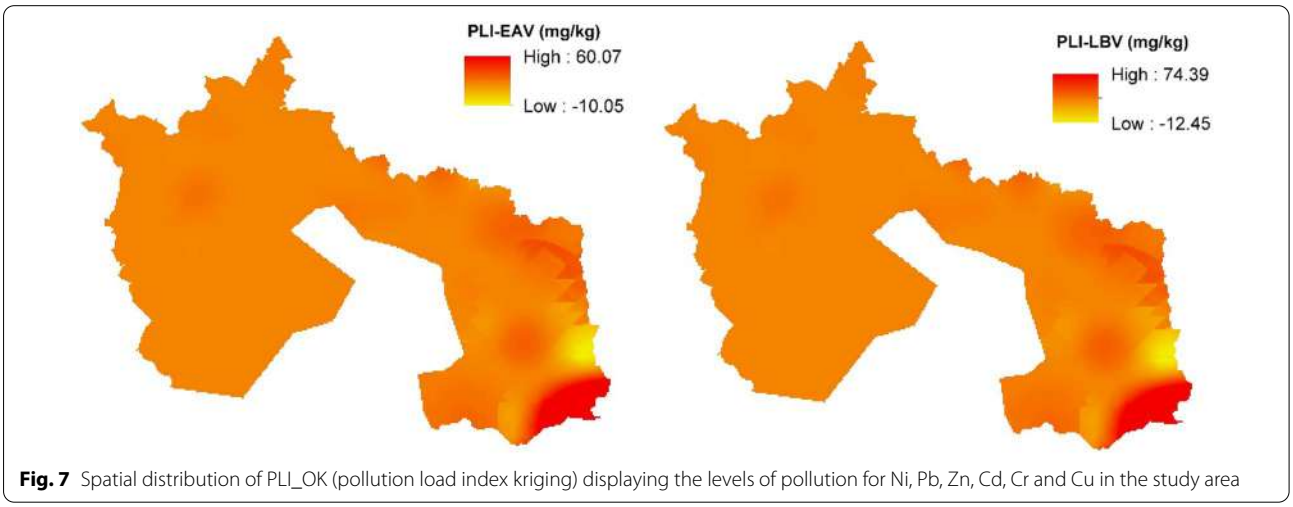


Fig. 7 Spatial distribution of PLI_OK (pollution load index kriging) displaying the levels of pollution for Ni, Pb, Zn, Cd, Cr and Cu in the study area

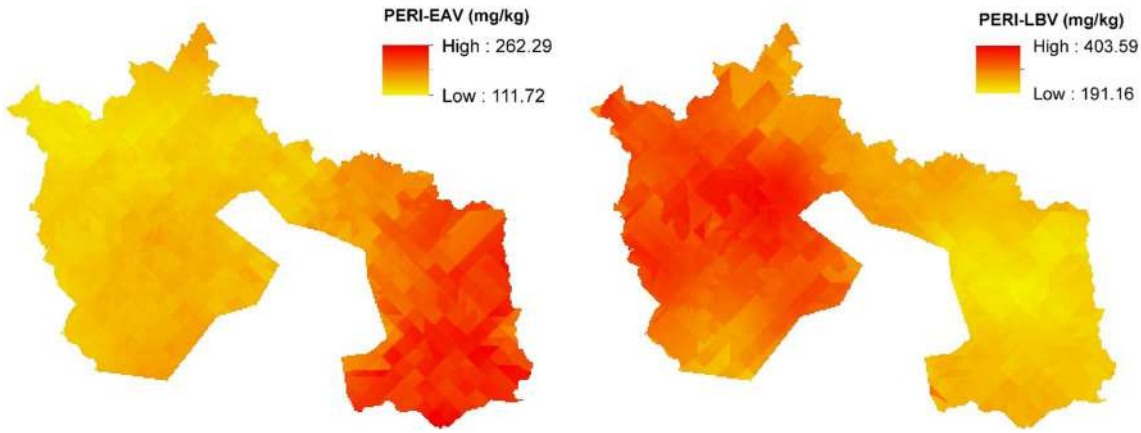


Fig. 8 Potential ecological risk index spatial distribution of (PERI-Kriging) showing pollution levels for Ni, Pb, Zn, Cd, Cr and Cu in the study area

when anthropogenic impacts and high levels of pollution are anticipated, because these levels might vary significantly among lithogenic contexts and should be examined in pedologically and geologically homogeneous areas [96]. However, using a reference geochemical background (e.g., EAV, UCC, and WAV) in the quantification of pollution level allows information about soil quality assessment to be analysed on a worldwide scale, enabling comparative studies beyond the local scale, and pollution indices that require reference geochemical background in their computation to be more multipurpose [96, 97].

Potential human health risk

Non-carcinogenic risk

The computed CDI_{totals}, HQ and HI values for non-Carcinogenic risk are displayed in Additional file 1: Tables S7–S10. The CDI_{total} distribution of PTEs in cultivated soils in the current research (children and adults) is presented in the following decreasing order: Mn > Zn > Pb > Cr > Cu > Cd > Ni > As (see Additional file 1: Tables S4 and S5). Additional file 1: Tables S4 and S5 illustrate the total non-carcinogenic intake (CDI_{total}) of adults and children. The CD_{total} values for children compared to adults indicate that children are slightly higher than adults. The CDI_{total} of the PTEs per sampled data (see Additional file 1: Table S7 and S8) suggested that the children exposure rate is higher than that of the adults. However, the children's computed hazard quotient (HQ) appears to be higher than the adults HQ (Additional file 1: Tables S9–S10). Based on the maximum and minimum range values of the HQs of children and adults per PTE, which fall between the following ranges such as 4.90E−02 to 2.82E−01 (Cr), 3.12E−03 to 2.72E−02 (Ni), 2.53E−03 to 2.01E−02 (Cu), 7.91E−02 to 1.30E+00 (As), 1.70E−02 to 1.55E−01 (Mn), 3.51E−02 to 5.72E−01 (Pb), 1.60E−03 to 1.16E−02 (Zn) and 8.42E−03 to 1.01E−01 (Cd) for children whereas the adults are 5.37E−03 to 3.10E−02 (Cr), 3.34E−04 to 2.92E−03 (Ni), 2.71E−04 to 2.15E−03 (Cu), 8.48E−03 to 1.39E−01 (As), 1.82E−03 to 1.66E−02 (Mn), 3.77E−03 to 6.14E−02 (Pb), 1.72E−04 to 1.25E−03 (Zn) and 9.19E−04 to 1.10E−02 (Cd). The calculated HQs values for PTEs of the minimum and maximum values (both children and adults) descend in this order As > Pb > Cr > Mn > Cd > Ni > Cu > Zn. The findings confirmed that ingestion was the most probable route for people in the study area to be exposed to PTEs. The variability of the measured PTEs concentration per sampled location revealed that the HI (for children) values estimated per 2×2 km suggested that 7 of the sampled location were higher than 1. Nonetheless, the HI estimated also suggested that 6.1% (1.01E+00 to 2.05E+00 that is 7 out of 115 sampled locations) of the total study area posed a high non-carcinogenic risk to children (see Additional file 1: Table S9). Similarly, 13.04% of the entire sampled area (i.e. 0.704–0.90, or 15 out of 115 for children) is on the verge of exceeding the allowable threshold if remedial action is not undertaken (see Additional file 1: Table S9). Children are more vulnerable to the health impacts of PTEs due to oral and finger practice, according to Agyeman et al. [98], and appear to be highly susceptible to PTEs. Numerous studies that employ PTE mean values to determine the risk to human health have consistently confirmed a High HI or lower risk level. Children's HI values in some studies that reported high or low HI values for children are as follows: Agyeman et al. [98] Han et al. [99], Natasha et al. [100], Wang et al. [101], Bhandari et al. [102] and Zheng et al. [103]. The computed HI for the adult is not statistically significant considering it is lower than the reference value 1; it thus implies that if

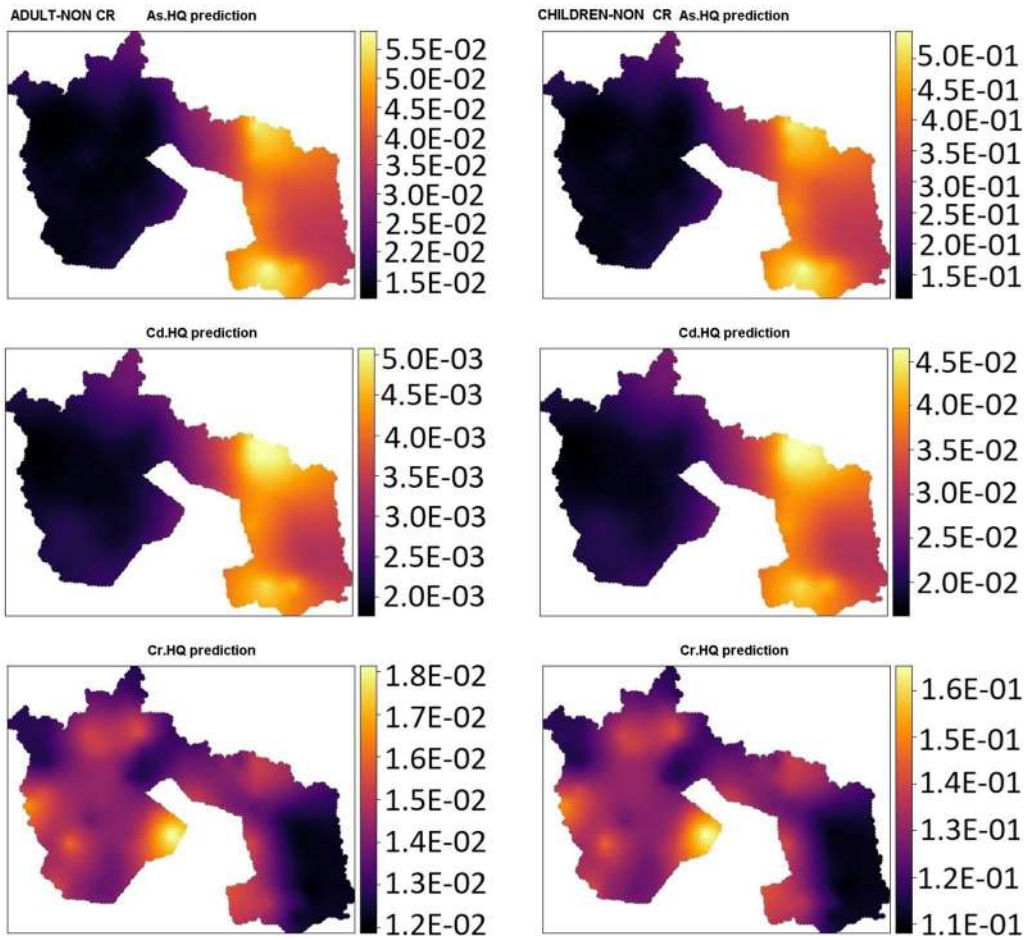


Fig. 9 Spatial distribution of hazard quotient (HQ) of PTEs (As, Cd and Cr) per sampled location

exposed, a non-carcinogenic adverse impact on an adult is not likely.

The spatial distribution of the hazard quotient of the PTEs per sample location suggested that the As, Pb, and Cd hazard quotient (AsHQ, PbHQs and CdHQs) for both children and adults showed similar colour patterns and hotspots in the northeastern and the south-eastern part of the map (see Figs. 9 and 11). The steel industry and agricultural activities in the suburbs are extremely probable to account for the hotspots, predicated on the commonality of the hazard quotient maps of As, Pb, and Cd. Chromium and manganese also share similar colour patterns of the hazard quotient spatial distribution map. Both (CrHQs and MnHQs) showed hotspots at the southwestern part of the map and moderate-to-low patches all over (see Figs. 9 and 10). This might be attributable to the usage of phosphate fertilizer and fungicides on agricultural fields to increase yield. This is supported by the estimated PMF, which revealed that Cr and Mn were the major PTEs in the factor 3 loadings.

Copper and nickel share similar hotspots pattern in the northwestern and the northeastern part of the hazard quotient spatial distribution map (see Fig. 10). Nevertheless, Ni showed more clearer or denser hotspots than copper. The PMF factor discharged confirms the hotspot pattern of Cu and Ni since Cu and Ni were the dominant PTEs in factor 4. Zinc showed a hotspot in the northwestern part of the map, which might be attributed to agriculture fertilizer and other tenants such as steel and metal industries that use zinc to coat iron and steel as a protective layer to inhibit corrosion.

The spatial distribution map of the adult and children hazard indexes has a similar colour pattern and hotspots. The children's degree of prediction based on the precise scale, on the other hand, suggested that the children residing within the enclave of the northeastern and south-eastern parts of the HI children spatial distribution map are exposed and vulnerable to PTEs (see Fig. 11). Therefore, premised on the children's HI distribution map scale, it can be inferred that HI values of

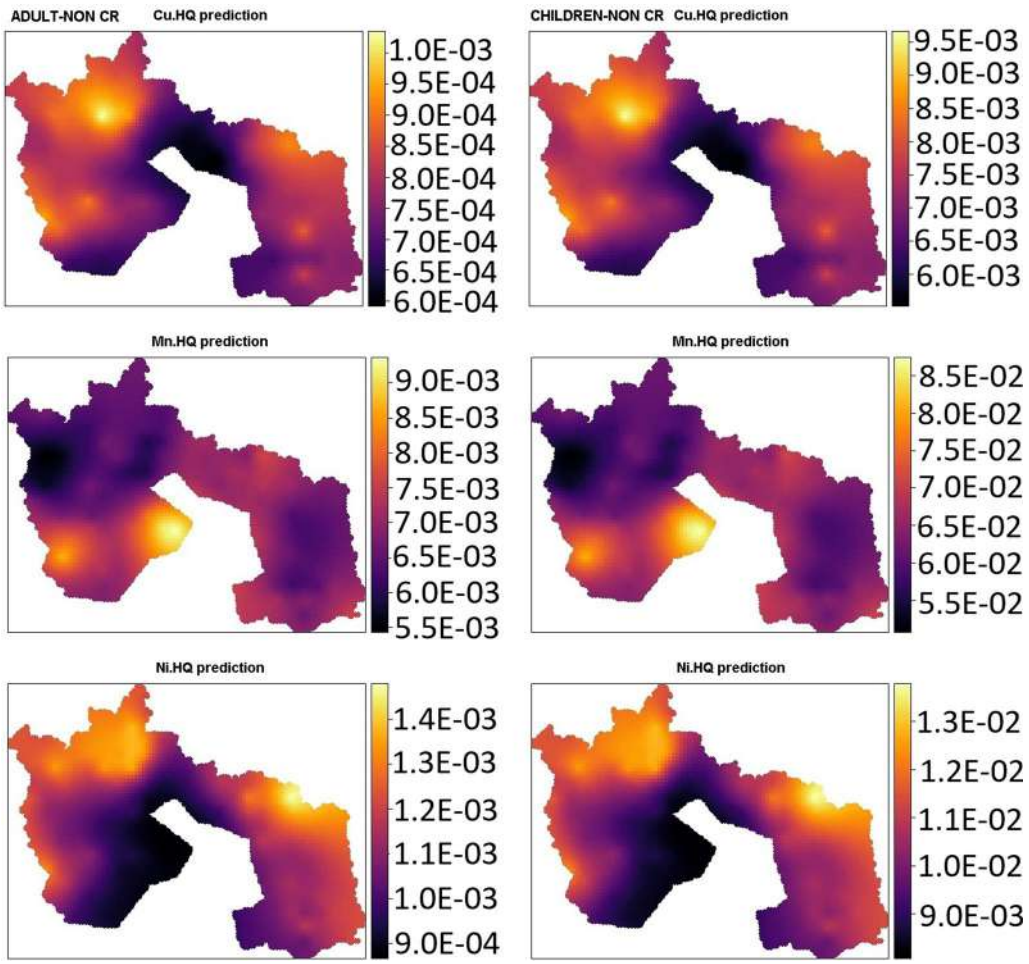


Fig. 10 Spatial distribution of hazard quotient (HQ) of PTEs (Cu, Mn and Ni) per sampled location

0.8 or higher are risk-prone areas, and thus corrective action should be made to mitigate the imminent threats to children.

Carcinogenic risk

CDI_{total}, TCR, and CR for both children and adults were computed, as shown in Additional file 1: Tables S11–S14. The chronic daily intake was calculated for Cd, Cr, Pb, Ni, and As. The CDI_{total} for adults and children are given in this descending order Pb > Cr > Ni > As > Cd. The CDI_{total} for children per sampled location for each PTE ranges between 1.20E−05 to 6.89E−05 (Cr), 5.33E−06 to 4.65E−05 (Ni), 2.03E−06 to 3.34E−05 (As) 1.05E−05 to 1.71E−04 (Pb) and 6.65E−07 to 7.99E−06 (Cd), whereas the adults Cr 5.13E−06 to 2.95E−05, Ni 2.29E−06 to 1.99E−05, As 8.71E−07 to 1.43E−05, Pb 4.50E−06 to 7.32E−05 and Cd 2.85E−07 to 3.42E−06. Regardless of the estimated value of the PTEs, children's CDI_{totals} were higher than adults. PTEs cause various health issues in children, including cardiovascular disease, poor respiratory function, cognitive deficits, reproductive toxicity, and bone damage [104]. Adults and children had higher Cr CDI_{totals} than the other PTEs. Furthermore, children's CDI_{total} was significantly higher than adults' (see Additional file 1: Tables S11 and S12). The CR for all PTEs in adults was found to be significantly lower than that of children.

The difference in measured values per sampled location exhibited different values for TCR. Based on the maximum and minimum values of the estimated TCR, it was apparent that the TCR of the children at all the observation points were found to be higher than that of the adult. The results revealed that 13.05% (i.e. 1×10^{-4} to $2.60E^{-04}$ 13.04%, that is 15 sampled points out of 115) of the sampled locations estimated TCR values for children were above 1×10^{-6} to 1×10^{-4}. However, the TCR estimated (for children) indicated that 45.22% (i.e. 7.02E−05 to 9.59E−05, that is 52 sampled points

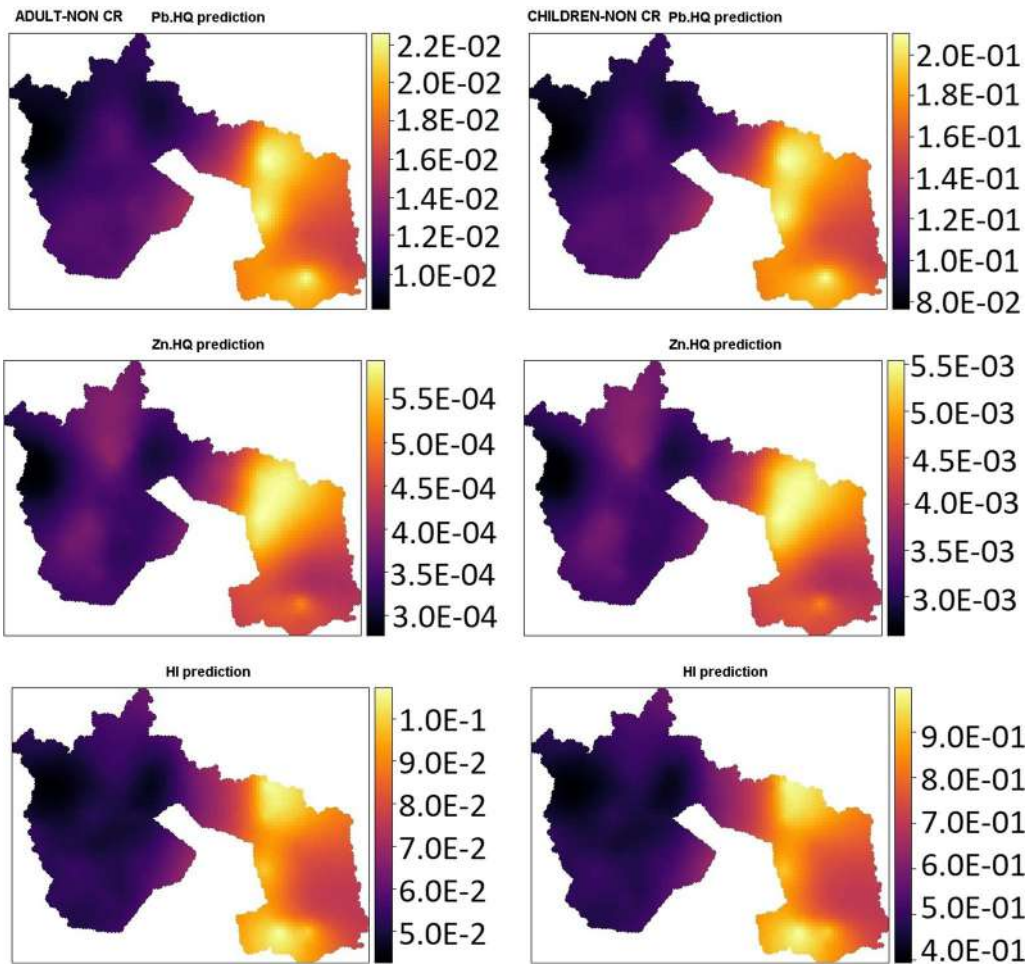


Fig. 11 Spatial distribution of hazard quotient (HQ) of PTEs (Pb and Zn) and health index (HI) per sampled location

out of 115) of the sampled locations are on the verge of exceeding the carcinogenic risk threshold if corrective measures are not enforced (see Additional file 1: Table S11 and S12). Nevertheless, several sampled locations' estimated TCR for adults exceeded the permissible threshold, while 2.16% (i.e. 7.26E−05 to 8.78E−05, or 3 out of 115 sampled locations) are on the cusp of exceeding the threshold. Due to the variability of measured PTEs values per sample location, the tendency of carcinogenic risk to befall a child is higher than that of an adult. Based on location-wise sampled data, the carcinogenic risk of the study area implies that some of the sample locations are carcinogenically risky to children compared to adults.

As a result, the likelihood of indigenous peoples, particularly children, being exposed to carcinogenic-related health risks is significant at some sample locations (13.04% or 15 sampling points out of 115) for children. Furthermore, the CR and HI of children were shown to be higher than that of adults, showing that children are nevertheless more likely to be exposed to PTEs because of their behavioural patterns, which increase the propensity for skin, particularly hand contact.

The spatial prediction of As and Cd carcinogenic risk for adults and children showed a similar hotspot pattern in the northeastern and south-eastern parts of the map (see Fig. 12). However, the hotspots anticipated that children with carcinogenic arsenic risk (CRAs) had a denser colour pattern, as evidenced by the predicted values. The spatial distribution of children's carcinogenic chromium risk (CRCr) revealed patches of hotspots, mainly in the northwestern and southwestern parts of the map (see Fig. 12). On the other hand, the adult displayed sporadic dotted moderate distribution with a broad scale of mild, moderate distribution across the study area. Nickel carcinogenic risk prediction, on the other hand, revealed moderate hotspots with a combination of high patches of hotspots in the northwest and the majority of the

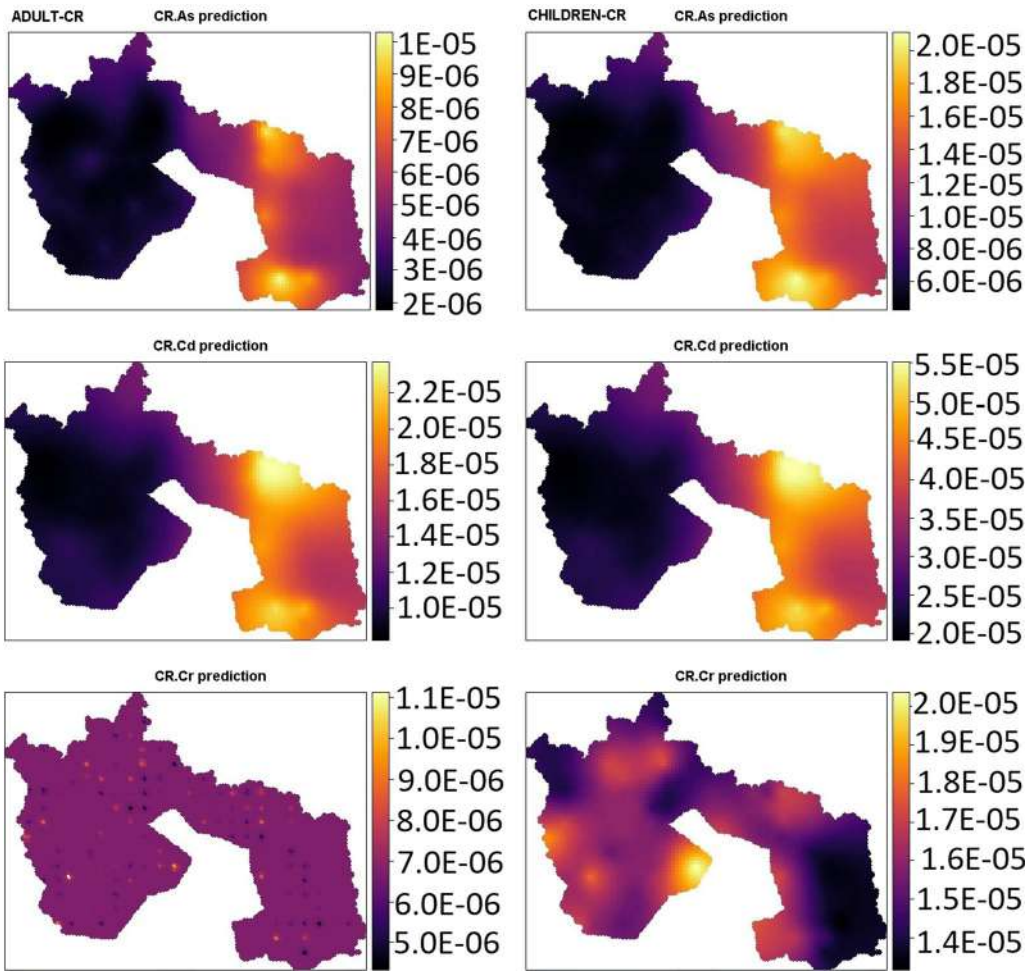


Fig. 12 Spatial distribution of hazard quotient (HQ) of PTEs (As, Cd and Cr) per sampled location

eastern enclave for both adults and children (see Fig. 13). Despite the similarities, the predicted values showed that the degree of exposure to children is greater than that of adults. The carcinogenic exposure rate of children to that of adults for Pb (CRPb) distribution map revealed moderate to high hotspots for children from the northeast to the southwest. Meanwhile, the adult CRPb map prediction revealed a continuous low hotspot with multiple dotted hotspots from the northeastern to the southwestern part of the map (see Fig. 13). When the current TCR is compared to similar studies conducted by Weissman nova et al. [105] in Ostrava, Czech Republic, it appears that Pb poses a significant health carcinogenic risk, Cd poses a moderate risk, and Cr poses a very high risk to children. This confirms the current study's findings that children are more vulnerable to PTE-related health risks than adults. In contrast, Kebonye et al. [106] affirmed the recent findings that children are more susceptible to PTEs exposure than adults in riverine soils, Příbram (Czech Republic).

The TCR maps for the children and adults have similar hotspot patterns from the northeast to the southwest (see Fig. 13). The TCR of that of children predicted values, on the other hand, revealed that the children residing within the enclave of the northeastern to the southern were in imminent danger. It can be inferred from the moderate to high hotspots patterns that begin at 1.0E-04 and higher (children TCR map) suggest that the risk associated with carcinogenic-related health issues in children, such as cancer of the skin, kidney, bladder, lung, prostate, and stomach, may occur earlier or later in their life journey. Numerous studies show that PTEs amass in fat tissues and subsequently negatively impact functions of the central nervous structure, immune and the endocrine systems, the urogenital and cardiovascular systems, and normal cellular metabolism [107, 108].

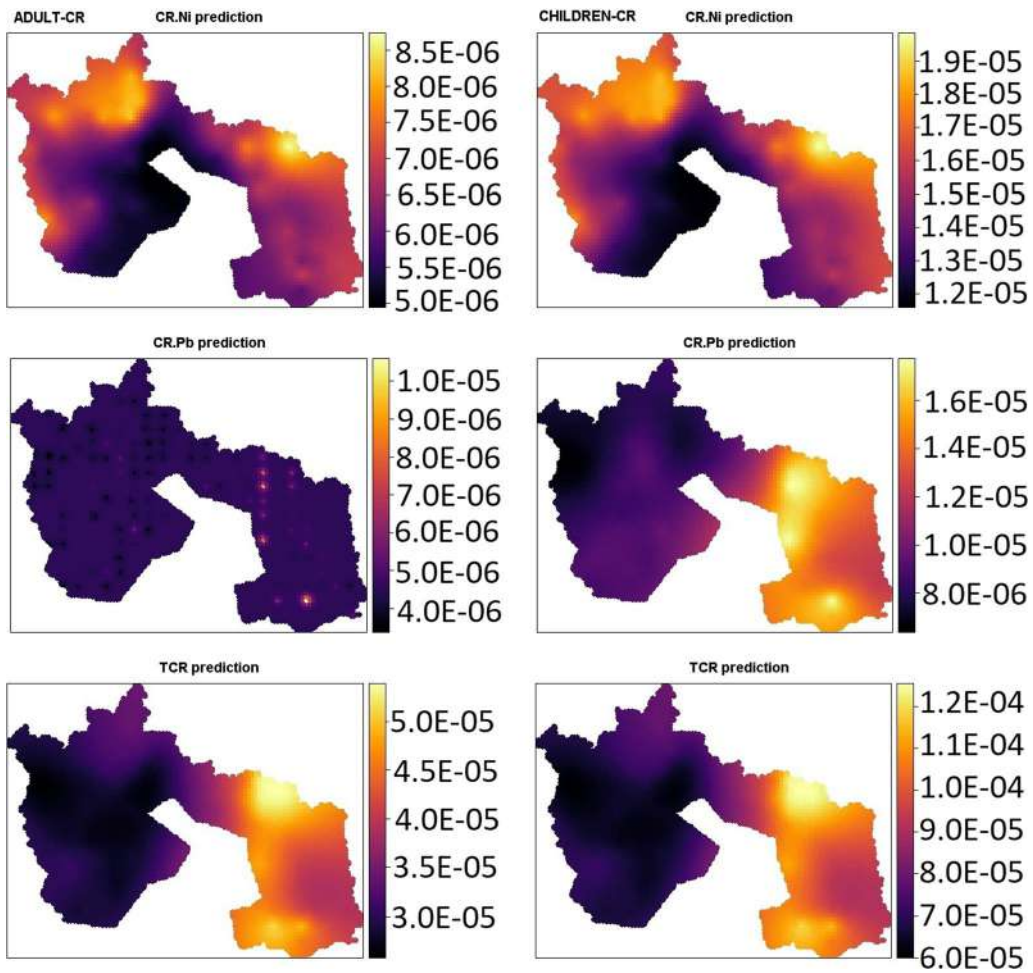


Fig. 13 Spatial distribution of hazard quotient (HQ) of PTEs (Ni and Pb) and total carcinogenic risk (TCR) per sampled location

Conclusion

In this study, a sample location technique was used to assess human health risk exposure and ecological risk of PTEs pollution in agricultural soils in the district of Frydek Mistek, Czech Republic. The utilization of the local background value and the European average value in the computation of pollution levels such as the single pollution index, pollution load index, and potential ecological risk revealed a variety of pollution levels based on dissimilarities in the threshold limits from disparate geochemical background levels. The PCA identified the primary pollution sources in the research area and confirmed the significant statistics of 71.21%. It suggested that the pollution source originated from a combination of sources, such as anthropogenic and geogenic sources. Pb and Zn (factor 1), Cu and Ni (factor 2), As (factor 3), and Mn and Cr (factor 4) predominate in various factor loadings, according to the source apportionment. The pollution assessment revealed that the pollution levels and ecological risk assessment ranged from low to high for pollution degrees and an exhibition of low to high pollution levels for pollution load index estimation. The health assessment risk for both carcinogenic and non-carcinogenic for adults and children indicated that the children are more exposed to adults. The sampled point-wise health risk assessment suggested that 13.05% of the totals sample locations are carcinogenically risky to children, and 6.04 of the sampled locations are likewise non-carcinogenically risky. The health risk spatial distribution map exposed the ecologically risky areas imminent to human health, especially children. PTEs in the soil can be increased by continually utilizing agricultural inputs and other anthropogenic activities such as the steel production industries. Due to the variability in observed PTEs concentration, the traditional approach of estimating health risk using mean concentration does not accurately

reflect the health condition of the area under study. We suggest that using the sampled location approach for future health risk assessment computations is essential. This enables the researcher to fully comprehend the study area and proffer remedial countermeasures at ecologically risky locations and on the verge of entering the high-risk zone. In general, the findings of this study are both informative and practical knowledge of the contamination of PTEs within the district of Frydek Mistek and the health-related risk status of individuals living in the neighbourhood.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12302-021-00577-w.

Additional file 1: Table S1. Exposure factors used in CDI estimation for non-carcinogenic and carcinogenic risk. **Table S2.** Shows the calculated single pollution index (PI) for the EAV, and the results suggested that the pollution level of the PTEs ranged from low to high. **Table S3**. Indicates the esctimated ecological risk index of the study area (geochemical background value used was European average value). **Table S4**. Indicates the esctimated risk index and pollution load index of the study area (local background value used was European average value). **Table S5**. Indicates the estimated single pollution index of the study area (geochemical background value used was local background value). **Table S6**. Indicates the estimated ecological risk of the study area (geochemical background value used was local background value) (geochemical background value used was local background value). **Table S7**. Children chronic daily intake total for PTEs (non carcinogenic). **Table S8**. Adult chronic daily intake total for PTEs (non carcinogenic). **Table S9**. Children Hazard quotient and health index. **Table S10**. Adult Hazard quotient and health index. **Table S11**. Children carcinogenic risk and total carcinogenic risk. **Table S12**. Adult carcinogenic risk and total carcinogenic risk. **Table S13**. Children chronic dailiy intake (carcinogenic risk). **Table S14**. Adult chronic daily intake (carcinogenic risk). **Table S15**. Raw data used for the study.

Acknowledgements
The Czech University of Life Sciences Prague funded this research with an internal Ph.D. grant number. SV20-5-21130 from the Faculty of Agrobiology, Food, and Natural Resources (CZU). The Czech Ministry of Education, Youth, and Sports (Project No. CZ.02.1.01/0.0/0.0/16 019/0000845) also aided. Finally, there is the Centre of Excellence (Centre of the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially risk substances of anthropogenic origin: comprehensive assessment of the soil contamination risks for the quality of agricultural products, NutRisk Centre).

Authors' contributions
PCA: conceptualization, methodology, investigation, writing—original draft, writing—review and editing. JK: investigation, writing review and editing. NMK: investigation, writing review and editing. LB: investigation, supervision, review and editing. RV: data collection and proofreading. OD: data investigation and analysis. KM: data analysis. All authors read and approved the final manuscript.

Funding
Czech University of Life Sciences Prague (CZU) Agrobiology, Food and Natural Resources.

Availability of data and materials
The data are available upon request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Animal research
Not applicable.

Competing interests
The authors declare that they have no known competing personal interests or relationships that could have appeared to influence the scientific work in this manuscript.

Received: 13 August 2021 Accepted: 27 November 2021
Published online: 13 December 2021

References

1. FAO (2015) Status of the World's Soil Resources. Intergov Tech Panel Soils 123–126
2. Mamut A, Eziz M, Mohammad A, Anayit M (2017) Human and ecological risk assessment: An International Journal The spatial distribution, contamination, and ecological risk assessment of heavy metals of farmland soils in Karashahar-Baghrash Oasis, northwest China. Int J 23:1300–1314. https://doi.org/10.1080/10807039.2017.1305263
3. Khelfaoui M, Medjram MS, Kabir A et al (2020) Chemical and mineralogical characterization of weathering products in mine wastes, soil, and sediment from the abandoned Pb/Zn mine in Skikda, Algeria. Environ Earth Sci 79:293. https://doi.org/10.1007/s12665-020-09043-x
4. Massas I, Gasparatos D, Ioannou D, Kalivas D (2018) Signs for secondary buildup of heavy metals in soils at the periphery of Athens International Airport, Greece. Environ Sci Pollut Res 25(1):658–671
5. Zhang J, Li H, Zhou Y et al (2018) Bioavailability and soil-to-crop transfer of heavy metals in farmland soils: a case study in the Pearl River Delta, South China. Environ Pollut 235:710–719. https://doi.org/10.1016/j.envpol.2017.12.106
6. Dong B, Zhang R, Gan Y et al (2019) Multiple methods for the identification of heavy metal sources in cropland soils from a resource-based region. Sci Total Environ 651:3127–3138. https://doi.org/10.1016/j.scitotenv.2018.10.130
7. Wang S, Cai LM, Wen HH et al (2019) Spatial distribution and source apportionment of heavy metals in soil from a typical county-level city of Guangdong Province, China. Sci Total Environ 655:92–101. https://doi.org/10.1016/j.scitotenv.2018.11.244
8. Agyeman PC, Ahado SK, Kingsley J et al (2021) Source apportionment, contamination levels, and spatial prediction of potentially toxic elements in selected soils of the Czech Republic. Environ Geochem Health 43:601–620. https://doi.org/10.1007/s10653-020-00743-8
9. Douay F, Roussel H, Pruvot C, Waterlot C (2008) Impact of a smelter closedown on metal contents of wheat cultivated in the neighbourhood. Environ Sci Pollut Res 15:162–169. https://doi.org/10.1065/espr2006.12.366
10. Thomas LDK, Hodgson S, Nieuwenhuijsen M, Jarup L (2009) Early kidney damage in a population exposed to cadmium and other heavy metals. Environ Health Perspect 117:181–184. https://doi.org/10.1289/ehp.11641
11. Bilias F, Nikoli T, Kalderis D, Gasparatos D (2021) Towards a soil remediation strategy using biochar: effects on soil chemical properties and bioavailability of potentially toxic elements. Toxics 9(8):184
12. Bai J, Xiao R, Gong A et al (2011) Assessment of heavy metal contamination of surface soils from typical paddy terrace wetlands on the Yunnan Plateau of China. Phys Chem Earth 36:447–450. https://doi.org/10.1016/j.pce.2010.03.025
13. Zukowska J, Biziuk M (2008) Methodological evaluation of method for dietary heavy metal intake. J Food Sci 73:R21–R29

14. Bempah CK, Ewusi A (2016) Heavy metals contamination and human health risk assessment around Obuasi gold mine in Ghana. Environ Monit Assess 188. https://doi.org/10.1007/s10661-016-5241-3
15. Chen H, Teng Y, Lu S et al (2015) Contamination features and health risk of soil heavy metals in China. Sci Total Environ 512–513:143–153. https://doi.org/10.1016/j.scitotenv.2015.01.025
16. Keshavarzi A, Kumar V (2019) Ecological risk assessment and source apportionment of heavy metal contamination in agricultural soils of Northeastern Iran. Int J Environ 29:544–560
17. Dayani M, Mohammadi J (2010) Geostatistical assessment of Pb, Zn and Cd contamination in near-surface soils of the urban-mining transitional region of Isfahan, Iran. Pedosphere 20:568–577. https://doi.org/10.1016/S1002-0160(10)60046-X
18. Kim RY, Yoon JK, Kim TS et al (2015) Bioavailability of heavy metals in soils: definitions and practical implementation—a critical review. Environ Geochem Health 37:1041–1061. https://doi.org/10.1007/s10653-015-9695-y
19. Yang Q, Li Z, Lu X et al (2018) A review of soil heavy metal pollution from industrial and agricultural regions in China: pollution and risk assessment. Sci Total Environ 642:690–700. https://doi.org/10.1016/j.scitotenv.2018.06.068
20. Ma W, Tai L, Qiao Z et al (2018) Contamination source apportionment and health risk assessment of heavy metals in soil around municipal solid waste incinerator: A case study in North China. Sci Total Environ 631–632:348–357. https://doi.org/10.1016/j.scitotenv.2018.03.011
21. Eziz M, Mohammad A, Mamut A, Hini G (2018) A human health risk assessment of heavy metals in agricultural soils of Yanqi Basin, Silk Road Economic Belt, China. Hum Ecol Risk Assess 24:1352–1366. https://doi.org/10.1080/10807039.2017.1412818
22. Xu X, Zhao Y, Zhao X et al (2014) Sources of heavy metal pollution in agricultural soils of a rapidly industrializing area in the Yangtze Delta of China. Ecotoxicol Environ Saf 108:161–167. https://doi.org/10.1016/j.ecoenv.2014.07.001
23. Doabi SA, Karami M, Afyuni M, Yeganeh M (2018) Pollution and health risk assessment of heavy metals in agricultural soil, atmospheric dust and major food crops in Kermanshah Province. Iran Ecotoxicol Environ Saf 163:153–164. https://doi.org/10.1016/j.ecoenv.2018.07.057
24. Rinklebe J, Antoniadis V, Shaheen SM et al (2019) Health risk assessment of potentially toxic elements in soils along the Central Elbe River, Germany. Environ Int 126:76–88. https://doi.org/10.1016/j.envint.2019.02.011
25. Baltas H, Sirin M, Gökbayrak E, Ozcelik AE (2020) A case study on pollution and a human health risk assessment of heavy metals in agricultural soils around Sinop Province, Turkey. Chemosphere 241:125015. https://doi.org/10.1016/j.chemosphere.2019.125015
26. Kampa M, Castanas E (2008) Human health effects of air pollution. Environ Pollut 151:362–367. https://doi.org/10.1016/j.envpol.2007.06.012
27. czso, 2019. Characteristics of the Frýdek-Místek district CZSO in Ostrava [WWW Document]. URL https://www.czso.cz/csu/xt/charakteristika_okresu_frydek_mistek. Accessed 10.6.20
28. John K, Agyeman PC, Kebonye NM, Isong IA, Ayito EO, Ofem KI, Qin CZ (2021) Hybridization of cokriging and gaussian process regression modelling techniques in mapping soil sulphur. CATENA 206:105534
29. Weather Spark, 2016. Average Weather in Frýdek-Místek, Czechia, Year-Round - Weather Spark [WWW Document]. URL https://weatherspark.com/y/83671/Average-Weather-iFrýdek-Místek-Czechia-Year-Round
30. Kozák J (2010) Soil Atlas of the Czech Republic. 150
31. Vacek O, Vašát R, Borůvka L (2020) Quantifying the pedodiversity-elevation relations. Geoderma 373:114441. https://doi.org/10.1016/j.geoderma.2020.114441
32. Huang Y, Chen Q, Deng M et al (2018) Heavy metal pollution and health risk assessment of agricultural soils in a typical peri-urban area in southeast China. J Environ Manage 207:159–168. https://doi.org/10.1016/j.jenvman.2017.10.072
33. Sawut R, Kasim N, Maihemuti B et al (2018) Pollution characteristics and health risk assessment of heavy metals in the vegetable bases of northwest China. Sci Total Environ 642:864–878. https://doi.org/10.1016/j.scitotenv.2018.06.034
34. Tomlinson DL, Wilson JG, Harris CR, Jeffrey DW (1980) Problems in the assessment of heavy-metal levels in estuaries and the formation of a pollution index. Helgoländer Meeresuntersuchungen 33:566–575. https://doi.org/10.1007/BF02414780
35. Hakanson L (1980) An ecological risk index for aquatic pollution control. A sedimentological approach. Water Res 14:975–1001. https://doi.org/10.1016/0043-1354(80)90143-8
36. U.S. EPA (2014) Positive Matrix Factorization (PMF) 5.0-Fundamentals and User Guide
37. Paatero P, Tapper U (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5:111–126. https://doi.org/10.1002/env.3170050203
38. Paatero P (1997) Least squares formulation of robust non-negative factor analysis. In: Chemometrics and intelligent laboratory systems. Elsevier, pp 23–35
39. Norris G (2008) EPA Positive Matrix Factorization (PMF) 3.0 Fundamentals & User Guide, US. Environmental Protection Agency. http://www.epa.gov/heasd/products/pmf/pmf.html
40. Wang G, Liu HQ, Gong Y, et al (2017) Risk assessment of metals in urban soils from a typical Industrial city, Suzhou, Eastern China. Int J Environ Res Public Health 14. https://doi.org/10.3390/ijerph14091025
41. US EPA (1989) Risk Assessment Guidance for Superfund. Human Health Evaluation Manual Part A, Interim Final. United States Environ Prot Agency 1 part A:300
42. USEPA (2002) Supplemental Guidance for Developing Soil Screening Levels for Superfund Sites, Appendix D-dispersion Factors Calculations. USA. United States Environ Prot Agency, Washington, DC pp. 4–24
43. Hu X, Zhang Y, Ding Z et al (2012) Bioaccessibility and health risk of arsenic and heavy metals (Cd Co, Cr, Cu, Ni, Pb, Zn and Mn) in TSP and PM2.5 in Nanjing, China. Atmos Environ 57:146–152. https://doi.org/10.1016/j.atmosenv.2012.04.056
44. Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43:59–69. https://doi.org/10.1007/BF00337288
45. Fraser SJ, Dickson BL (2007) A New method for data integration and integrated data interpretation: self-organising maps
46. Li T, Sun G, Yang C et al (2018) Using self-organizing map for coastal water quality classification: towards a better understanding of patterns and processes. Sci Total Environ 628–629:1446–1459. https://doi.org/10.1016/j.scitotenv.2018.02.163
47. Melssen WJ, Smits JRM, Buydens LMC, Kateman G (1994) Using artificial neural networks for solving chemical problems. Part II. Kohonen self-organising feature maps and Hopfield networks. Chemom Intell Lab Syst. https://doi.org/10.1016/0169-7439(93)E0036-4
48. Kabata-Pendias A (2011) Trace Elements in Soils and Plants. CRC Taylor Fr Group, London New York
49. Tóth G, Hermann T, Da Silva MR, Montanarella L (2016) Heavy metals in agricultural soils of the European Union with implications for food safety. Environ Int 88:299–309
50. Jia Z, Li S, Wang L (2018) Assessment of soil heavy metals for eco-environment and human health in a rapidly urbanization area of the upper Yangtze Basin. Sci Rep 8 https://doi.org/10.1038/s41598-018-21569-6
51. Nemecek J, Podlesakova E (1992) Retrospective experimental monitoring of heavy-metals containing in soils of the Czech Republic. Rostlinna Vyroba 38(6):433–436
52. Chandrasekaran A, Ravisankar R, Harikrishnan N et al (2015) Multivariate statistical analysis of heavy metal concentration in soils of Yelagiri Hills, Tamilnadu, India—spectroscopical approach. Spectrochim Acta Part A Mol Biomol Spectrosc 137:589–600. https://doi.org/10.1016/j.saa.2014.08.093
53. Karimi Nezhad MT, Tabatabaii SM, Gholami A (2015) Geochemical assessment of steel smelter-impacted urban soils, Ahvaz, Iran. J Geochemical Explor 152:91–109. https://doi.org/10.1016/j.gexplo.2015.02.005
54. Hou D, He J, Lü C et al (2013) Distribution characteristics and potential ecological risk assessment of heavy metals (Cu, Pb, Zn, Cd) in water and sediments from Lake Dalinouer, China. Ecotoxicol Environ Saf 93:135–144. https://doi.org/10.1016/j.ecoenv.2013.03.012
55. Franco-Uría A, López-Mateo C, Roca E, Fernández-Marcos ML (2009) Source identification of heavy metals in pastureland by multivariate analysis in NW Spain. J Hazard Mater 165:1008–1015. https://doi.org/10.1016/j.jhazmat.2008.10.118

56. Nicholson FA, Smith SR, Alloway BJ et al (2003) An inventory of heavy metals inputs to agricultural soils in England and Wales. Sci Total Environ 311:205–219. https://doi.org/10.1016/S0048-9697(03)00139-6
57. Luo L, Ma Y, Zhang S et al (2009) An inventory of trace element inputs to agricultural soils in China. J Environ Manage 90:2524–2530. https://doi.org/10.1016/j.jenvman.2009.01.011
58. Mantovi P, Bonazzi G, Maestri E, Marmiroli N (2003) Accumulation of copper and zinc from liquid manure in agricultural soils and crop plants
59. Mishima S-I, Inoue T, Kimura R (2004) Estimation of cadmium load on Japanese farmland associated with the application of chemical fertilizers and livestock excreta. Soil Sci Plant Nutr 50:263–267. https://doi.org/10.1080/00380768.2004.10408476
60. Rezapour S, Atashpaz B, Moghaddam SS et al (2019) Cadmium accumulation, translocation factor, and health risk potential in a wastewater-irrigated soil-wheat (*Triticum aestivum* L.) system. Chemosphere 231:579–587. https://doi.org/10.1016/j.chemosphere.2019.05.095
61. Li B, Xiao R, Wang C et al (2017) Spatial distribution of soil cadmium and its influencing factors in peri-urban farmland: a case study in the Jingyang District, Sichuan, China. Environ Monit Assess 189:1–16. https://doi.org/10.1007/s10661-016-5744-y
62. Li X, Zhang J, Ma J et al (2020) Status of chromium accumulation in agricultural soils across China (1989–2016). Chemosphere 256:127036
63. Yaylali-Abanuz G (2011) Heavy metal contamination of surface soil around Gebze industrial area, Turkey. Microchem J 99:82–92. https://doi.org/10.1016/j.microc.2011.04.004
64. Echevarria G, Massoura ST, Sterckeman T, et al (2006) Assessment and control of the bioavailability of nickel in soils. In: Environmental toxicology and chemistry. pp 643–651
65. Krishna AK, Govil PK (2005) Heavy metal distribution and contamination in soils of Thane-Belapur industrial development area, Mumbai, Western India. Environ Geol 47(8):1054–1061
66. Salonen VP, Korkka-Niemi K (2007) Influence of parent sediments on the concentration of heavy metals in urban and suburban soils in Turku, Finland. Appl Geochem 22:906–918. https://doi.org/10.1016/j.apgeochem.2007.02.003
67. Goovaerts P (1997) Geostatistics for natural resources and evaluation
68. Al-Mughrabi KI, Vikram A, Poirier R et al (2016) Management of common scab of potato in the field using biopesticides, fungicides, soil additives, or soil fumigants. Biocontrol Sci Technol 26:125–135. https://doi.org/10.1080/09583157.2015.1079809
69. Rodríguez JA, Nanos N, Grau JM et al (2008) Multiscale analysis of heavy metal contents in Spanish agricultural topsoils. Chemosphere 70:1085–1096. https://doi.org/10.1016/j.chemosphere.2007.07.056
70. Chakraborty S, Weindorf DC, Paul S et al (2015) Diffuse reflectance spectroscopy for monitoring lead in landfill agricultural soils of India. Geoderma Reg 5:77–85. https://doi.org/10.1016/j.geodrs.2015.04.004
71. Khosravi V, Doulati Ardejani F, Yousefi S, Aryafar A (2018) Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods. Geoderma 318:29–41. https://doi.org/10.1016/j.geoderma.2017.12.025
72. Tepanosyan G, Sahakyan L, Belyaeva O, Saghatelyan A (2016) Origin identification and potential ecological risk assessment of potentially toxic inorganic elements in the topsoil of the city of Yerevan, Armenia. J Geochem Explor 167:1–11. https://doi.org/10.1016/j.gexplo.2016.04.006
73. Li X, Poon CS, Liu PS (2001) Heavy metal contamination of urban soils and street dusts in Hong Kong. Appl Geochem 16:1361–1368. https://doi.org/10.1016/S0883-2927(01)00045-2
74. Arditsoglou A, Samara C (2005) Levels of total suspended particulate matter and major trace elements in Kosovo: a source identification and apportionment study. Chemosphere 59:669–678. https://doi.org/10.1016/j.chemosphere.2004.10.056
75. Hjortenkrans D, Bergbäck B, Häggerud A (2006) New metal emission patterns in road traffic environments. Environ Monit Assess 117:85–98. https://doi.org/10.1007/s10661-006-7706-2
76. Guan Q, Wang F, Xu C et al (2018) Source apportionment of heavy metals in agricultural soil based on PMF: a case study in Hexi Corridor, Northwest China. Chemosphere 193:189–197. https://doi.org/10.1016/j.chemosphere.2017.10.151
77. Al-Khashman OA, Shawabkeh RA (2009) Metal distribution in urban soil around steel industry beside Queen Alia Airport, Jordan. Environ Geochem Health 31:717–726. https://doi.org/10.1007/s10653-009-9250-9
78. Wang J, Huang Y, Li T et al (2020) Contamination, morphological status and sources of atmospheric dust in different land-using areas of a steel-industry city, China. Atmos Pollut Res 11:283–289. https://doi.org/10.1016/j.apr.2019.10.014
79. Xiong X, Yanxia L, Wei L et al (2010) Copper content in animal manures and potential risk of soil copper pollution with animal manure use in agriculture. Resour Conserv Recycl 54:985–990. https://doi.org/10.1016/j.resconrec.2010.02.005
80. Sharif M, Rahman MA ur, Ahmed B, et al (2020) Copper nanoparticles as growth promoter, antioxidant and anti-bacterial agents in poultry nutrition: prospects and future implications. Biol Trace Elem Res 1–12. https://doi.org/10.1007/s12011-020-02485-1
81. Heo JM, Opapeju FO, Pluske JR et al (2013) Gastrointestinal health and function in weaned pigs: a review of feeding strategies to control post-weaning diarrhoea without using in-feed antimicrobial compounds. J Anim Physiol Anim Nutr (Berl) 97:207–237. https://doi.org/10.1111/j.1439-0396.2012.01284.x
82. Cheng Q, Guo Y, Wang W, Hao S (2014) Spatial variation of soil quality and pollution assessment of heavy metals in cultivated soils of Henan Province, China. Chem Speciat Bioavailab 26:184–190. https://doi.org/10.3184/095422914X14042081874564
83. Harasim P, Filipek T (2015) Nickel in the environment. J Elem 20:525–534. https://doi.org/10.5601/jelem.2014.19.3.651
84. Li XH, Tang ZL, Chu FY, Yang LY (2011) Characteristics of distribution and chemical speciation of heavy metals in environmental mediums around Jinchang mining city, Northwest China. Environ Earth Sci 64:1667–1674. https://doi.org/10.1007/S12665-009-0438-1
85. Chen T, Chang Q, Liu J et al (2016) Identification of soil heavy metal sources and improvement in spatial mapping based on soil spectral information: a case study in northwest China. Sci Total Environ 565:155–164
86. Bhattacharya P, Welch AH, Stollenwerk KG et al (2007) Arsenic in the environment: biology and chemistry. Sci Total Environ 379:109–120. https://doi.org/10.1016/j.scitotenv.2007.02.037
87. Micó C, Recatalá L, Peris M, Sánchez J (2006) Assessing heavy metal sources in agricultural soils of an European Mediterranean area by multivariate analysis. Chemosphere 65:863–872. https://doi.org/10.1016/j.chemosphere.2006.03.016
88. Jayasumana MACS, Paranagama PA, Amarasinghe, et al (2013) Possible link of Chronic arsenic toxicity with Chronic Kidney Disease of unknown etiology in Sri Lanka. J Nat Sci Res www.iiste.org ISSN 3
89. Jayasumana C, Fonseka S, Fernando A et al (2015) Phosphate fertilizer is a main source of arsenic in areas affected with chronic kidney disease of unknown etiology in Sri Lanka. Springerplus 4:1–8. https://doi.org/10.1186/s40064-015-0868-z
90. Liu L, An Y, Ma J et al (2020) Source apportionment of soil heavy metals in Beijing Urban Park based on the UNMIX model. Res Environ Sci 33:2856–2863. https://doi.org/10.13198/j.issn.1001-6929.2020.03.40
91. Zhang X, Zhong T, Liu L et al (2016) Chromium occurrences in arable soil and its influence on food production in China. Environ Earth Sci 75:1–8. https://doi.org/10.1007/s12665-015-5078-z
92. Bradl HB (2005) Chapter 1 Sources and origins of heavy metals. Interface Sci Technol 6:1–27. https://doi.org/10.1016/S1573-4285(05)80020-1
93. Shaw River (2017) Manganese Fact Sheet. 2
94. Alvarez-Guerra M, González-Piñuela C, Andrés A et al (2008) Assessment of Self-Organizing Map artificial neural networks for the classification of sediment quality. Environ Int 34:782–790. https://doi.org/10.1016/j.envint.2008.01.006
95. Wang Z, Xiao J, Wang L, Liang T, Guo Q, Guan Y, Rinklebe J (2020) Elucidating the differentiation of soil heavy metals under different land uses with geographically weighted regression and self-organizing map. Environ Pollut 260:114065
96. Kowalska JB, Mazurek R, Gąsiorek M, Zaleski T (2018) Pollution indices as useful tools for the comprehensive evaluation of the degree of soil contamination—a review. Environ Geochem Health 40:2395–2420
97. Agyeman PC et al (2021) Multi-geochemical background comparison and the identification of the best normalizer for the estimation of PTE contamination in agricultural soil. Environ Geochem Health. https://doi.org/10.1007/s10653-021-01109-4

98. Agyeman PC, Ahado SK, John K, et al (2021) Health risk assessment and the application of CF-PMF: a pollution assessment-based receptor model in an urban soil. J Soils Sediments 1–20. https://doi.org/10.1007/s11368-021-02988-x
99. Han Q, Wang M, Cao J et al (2020) Health risk assessment and bioaccessibilities of heavy metals for children in soil and dust from urban parks and schools of Jiaozuo, China. Ecotoxicol Environ Saf 191:110157. https://doi.org/10.1016/j.ecoenv.2019.110157
100. Natasha SM, Khalid S et al (2020) A critical review of mercury speciation, bioavailability, toxicity and detoxification in soil-plant environment: ecotoxicology and health risk assessment. Sci Total Environ 711:134749. https://doi.org/10.1016/j.scitotenv.2019.134749
101. Wang F, Guan Q, Tian J et al (2020) Contamination characteristics, source apportionment, and health risk assessment of heavy metals in agricultural soil in the Hexi Corridor. CATENA 191:104573. https://doi.org/10.1016/j.catena.2020.104573
102. Bhandari G, Atreya K, Scheepers PTJ, Geissen V (2020) Concentration and distribution of pesticide residues in soil: non-dietary human health risk assessment. Chemosphere 253:126594. https://doi.org/10.1016/j.chemosphere.2020.126594
103. Zheng S, Wang Q, Yuan Y, Sun W (2020) Human health risk assessment of heavy metals in soil and food crops in the Pearl River Delta urban agglomeration of China. Food Chem 316:126213. https://doi.org/10.1016/j.foodchem.2020.126213
104. Madrigal J, Persky V, Pappalardo A, Argos M (2018) Association of heavy metals with measures of pulmonary function in youth: findings from the 2011–2012 National Health and Nutrition Examination Survey (NHANES). ISEE Conf Abstr 2018. https://doi.org/10.1289/isesisee.2018.o03.03.26
105. Doležalová Weissmannová H, Mihočová S, Chovanec P, Pavlovský J (2019) Potential ecological risk and human health risk assessment of heavy metal pollution in industrial affected soils by coal mining and metallurgy in Ostrava, Czech Republic. Int J Environ Res Public Health 16(22):4495
106. Kebonye NM, Eze PN, John K, Agyeman PC, Němeček K, Borůvka L (2021) An in-depth human health risk assessment of potentially toxic elements in highly polluted riverine soils, Příbram (Czech Republic). Environ Geochem Health 1–17
107. Wang Q, Liu J, Cheng S (2015) Heavy metals in apple orchard soils and fruits and their health risks in Liaodong Peninsula, Northeast China. Environ Monit Assess 187. https://doi.org/10.1007/s10661-014-4178-7
108. Wang WX (2013) Dietary toxicity of metals in aquatic animals: recent studies and perspectives. Chinese Sci Bull 58:203–213. https://doi.org/10.1007/s11434-012-5413-7
109. IGME (2012) Geochemical atlas of Spain (Atlas Geoquímico de España). Instituto Geológico y Minero de España, Madrid In Spanish. Bravo S, García-Ordiales E, García-Navarro FJ, Amorós JÁ, Pérez-de-Los-Reyes C, Jiménez-Ballesta R, Higueras P (2019)
110. Erhart E, Hartl W, Putz B (2008) Total soil heavy-metal concentrations and mobile fractions after 10 years of biowaste-compost fertilization. J Plant Nutr Soil Sci 171(3):378–383
111. Tarvainen T, Kallio E (2002) Baselines of certain bioavailable and total heavy metal concentrations in Finland. Appl Geochem 17(8):975–980
112. Liu X, Zhang W, Hu Y et al (2015) Arsenic pollution of agricultural soils by concentrated animal feeding operations (CAFOs). Chemosphere 119:273–281. https://doi.org/10.1016/j.chemosphere.2014.06.067

Publisher's Note

Journal of Environmental Management 326 (2023) 116701

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

Optimal zinc level and uncertainty quantification in agricultural soils via visible near-infrared reflectance and soil chemical properties

Prince Chapman Agyeman [a,*], Ndiye Michael Kebonye [b,c], Vahid Khosravi [a], John Kingsley [a], Luboš Borůvka [a], Radim Vašát [a], Charles Mario Boateng [d]

[a] Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and NaturalResources, Czech University of Life Sciences Prague, 16500 Prague, Czech Republic
[b] Department of Geosciences, Chair of Soil Science and Geomorphology, University Of Tübingen, Rümelinstr. 19-23, Tübingen, Germany
[c] DFG Cluster of Excellence "Machine Learning", University of Tübingen, AI Research Building, Maria-von-Linden-Str. 6, 72076, Tübingen, Germany
[d] Department of Marine and Fisheries Sciences, University of Ghana, Legon, Ghana

ARTICLE INFO

Keywords:
Conditional inference forest
Zinc
Agricultural soil
Spectral reflectance
Uncertainty assessment

ABSTRACT

Zinc (Zn) is a vital element required by all living creatures for optimal health and ecosystem functioning. Therefore, several researchers have modeled and mapped its occurrence and distribution in soils. Nonetheless, leveraging model predictive performances while coupling information derived from visible near-infrared (Vis-NIR) and soils (i.e. chemical properties) to estimate potential toxic elements (PTEs) like Zn in agricultural soils is largely untapped. This study applies two methods to rapidly monitor Zn concentration in agricultural soil. Firstly, employing Vis-NIR and machine learning algorithms (MLAs) (Context 1) and secondly, applying Vis-NIR, soil chemical properties (SCP), and MLAs (Context 2). For the Vis-NIR information, single and combined pretreatment methods were applied. The following MLAs were used: conditional inference forest (CIF), partial least squares regression (PLSR), M5 tree model (M5), extreme gradient boosting (EGB), and support vector machine regression (SVMR) respectively. For context 1, the results indicated that M5-MSC (M5 tree model-multiplicative scatter correction) with coefficient of determination (R^2) = 0.72, root mean square error (RMSE) = 21.08 (mg/kg), median absolute error (MdAE) = 13.69 and ratio of performance to interquartile range (RPIQ) = 1.63 was promising. Regarding context 2, CIF with spectral pretreatment and soil properties [CIF-DWTLOGMSC + SCP (conditional inference forest-discrete wavelet transformation-logarithmic transformation-multiplicative scatter correction-soil chemical properties)] yielded the best performance of R^2 = 0.86, RMSE = 14.52 (mg/kg), MdAE = 6.25 and RPIQ = 1.78. Altogether, for contexts 1 and 2, the CIF-DWTLOGMSC + SCP approach (context 2) was the best Zn model outcome for the agricultural soil. The uncertainty map revealed a low to high error distribution in context 1, and a low to moderate distribution in context 2 for all models except CIF, which had some patches with high uncertainty. We conclude that a multiple optimization approach for modeling Zn levels in agricultural soils is invaluable and may provide fast and reliable information needed for area-specific decision-making.

1. Introduction

For many years, reasonably accurate and reliable conventional laboratory methods such as the Inductively Coupled Plasma Optical Emission Spectroscopy (ICP-OES) and Atomic Absorption Spectroscopy (AAS) have been solely used to quantify elemental levels in different environmental matrices including soils, vegetation and water (e.g. Gomez et al., 2007; Tighe et al., 2004; Nomngongo et al., 2013). Already, there are dozens of researches that confirm the importance of these methods for quantifying elemental levels specifically in soils. Nonetheless, the limits allied to the increase in costs of purchasing such methods plus the actual analytical procedures, time constraints involved in sample analysis as well as the non-environmentally friendly side of the methods remain a challenge. Since results from these analytical methods are important for different soil-related decisions and applications, alternative methods had to be developed to address this cause. Fortunately, proximal soil sensing has allowed for fast, rapid, cost-effective and environmentally friendly monitoring of soil elemental

* Corresponding author.
E-mail address: agyeman@af.czu.cz (P.C. Agyeman).

https://doi.org/10.1016/j.jenvman.2022.116701
Received 23 August 2022; Received in revised form 25 October 2022; Accepted 1 November 2022
Available online 14 November 2022

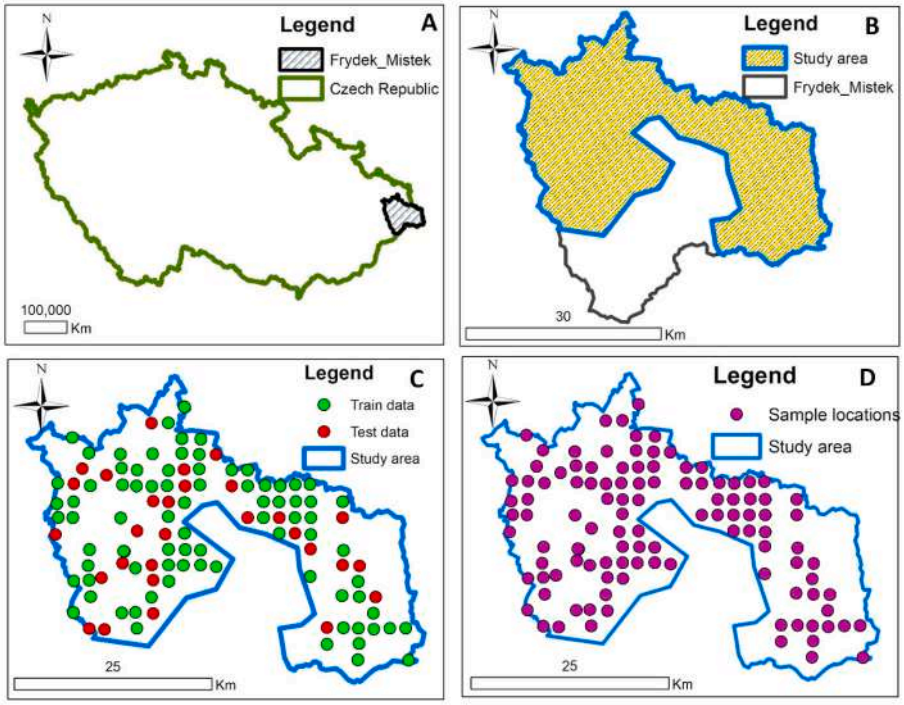


Fig. 1. The map Czech Republic (A), Frydek Mistek district (B), Research location coupled with the partitioned dataset employed (training and test) (C) Study area with sampling locations (D).

levels while leveraging machine learning algorithms (MLAs) to obtain reliable estimates (Bellon-Maurel et al., 2010). For these estimates, different proxies are normally applied alongside the MLAs to estimate variables of interest. Common proximal sensors used in soil-related studies include the portable visible near-infrared spectrometer (vis--NIR) (e.g. Khosravi et al., 2018), portable X-ray fluorescence spectrometer (pXRF) (Kebonye et al., 2021) and electromagnetic (EM) induction instruments such as the EM38 (e.g. Khongnawang et al., 2022).

Some of the many good examples of using MLAs coupled with proximal sensing data including vis-NIR (400–1200 nm) and shortwave infrared (1200–2500 nm) reflectance spectroscopy have been used to analyze the spectrally active attributes of soil and sediment samples, which include PTEs (Khosravi et al., 2018), soil organic carbon (Hutengs et al., 2019), soil organic matter (Hong et al., 2019a), sediments (Jiang et al., 2018), and soil attributes such as pH, soil organic matter and nitrogen content (Ahmadi et al., 2021). Certainly, vis-NIR combined with diverse soil-measured datasets such as those obtained via pXRF is gaining popularity in soil science because of the previously mentioned benefits. These benefits are also corroborated by Bellon-Maurel et al. (2010). Moreover, since the application of proximal sensors mostly involves minimal sample pre-processing (e.g. grinding and pulverization) to facilitate or improve measurements, therefore, vis-NIR may well be suitable for long and rapid surveillance of soil PTE contents. Even though soil PTEs are spectrally unresponsive (Stenberg et al., 2010), their interactions with spectrally active soil components such as clay, SOM, and Fe oxides might well permit for Vis-NIR prediction of these metals (Wang et al., 2014; Y. Wu et al., 2007). Low PTE concentrations lack spectral properties within the Vis-NIR region; nevertheless, the association between the content levels of these featureless elements and the responsive soil attributes can be used to estimate their enriched levels (Khosravi et al., 2018; Wu et al., 2007). The prediction of PTEs in soils and sediments using Vis-NIR spectral reflectance coupled with MLAs is increasing. In the past, Vis-NIR coupled with MLA was mostly used to predict soil properties including soil organic carbon and organic matter. The application of spectral reflectance coupled with MLAs for quantifying PTEs is widely reported by researchers including Biney et al. (2022); Cao et al. (2020); Devianti et al. (2019); Shi et al. (2014); Luce et al. (2017); Wang et al. (2014) and Xu et al. (2021). The combination of Vis-NIR and MLAs in the estimation of PTE content in soil has become an effective alternate tool for evaluating the concentration of PTEs in soil or sediments. Even though there is no one sure way of predicting the concentration of PTEs in the soil or sediments, the complementary relationship between Vis-NIR spectra reflectance coupled with MLAs has pushed the frontiers of predictive soil mapping. Despite the popularity of partial least squares, support vector machine, cubist, and random forest as complementary algorithms for spectra reflectance, it is worth emphasizing that there are no single MLAs that are best suited for spectral reflectance datasets for the estimation of PTE in sediment and soils. Researchers have applied different algorithms such as ensemble (Biney et al., 2022), extreme learning machine (Khosravi et al., 2018), Generalized Regression Neural Network (Xu et al., 2021), extreme gradient boosting (Zhao et al., 2022), stepwise multiple linear regression and multiple linear regression (Choe et al., 2009) coupled with spectra reflectance in the prediction of PTEs in sediments or soil. On account of the existing literature regarding the application of proximal sensors for monitoring PTEs in soils, generally few studies predict PTE levels in agricultural soils while leveraging information from Vis-NIR spectroscopy and soil chemical properties. The benefits of such an approach may facilitate rapid elemental monitoring in agricultural soils and also allow for conclusions to be drawn regarding the main drivers of PTEs within these specific soils.

The current study aims to capitalize on the potential for modeling Zn concentration levels in agricultural soil by combining a Vis-NIR spectral reflectance dataset with soil chemical properties and MLAs. The estimation of Zn concentration levels in cultivated soils is done in two ways: the prediction of Zn concentration in agricultural soil using machine learning algorithms and Vis-NIR spectra reflectance (Context 1) and the prediction of Zn concentration in agricultural soil using Vis-NIR spectra reflectance, soil chemical properties, and machine learning algorithms (Context 2). The study's objectives are to (i) quantify the concentration of Zn in cultivated soil based on a series of MLAs coupled with Vis-NIR

spectral reflectance; (ii) determine whether combining Vis-NIR, soil chemical properties, and MLAs in the estimation of Zn content in agricultural soil will improve prediction accuracy; (iii) determine the level of uncertainty that will be propagated in both contexts and (iv) to evaluate the performance of a single pretreated method versus a combination of pretreatment methods.

2. Materials and methods

2.1. Study location

The study setting is in the Czech Republic's Frydek Mistek local municipality (see Fig. 1). The study location's landscape is distinguished by steep terrain and mountainous from the outer Carpathians. The site location is differentiated by comprehensive commercial agriculture as well as a myriad of metal and steel industries, and it is geographically located at 49° 41′ 0′ north and 18° 20′ 0′ east, at an altitude of 225–327 m above sea level (Agyeman et al., 2020). However, using categorization by Koppen, the study location is categorized as having Cfb = temperate oceanic weather with elevated rainfall even throughout the dry months (John et al., 2021). The landmass used for this study is 889.8 km^2, which is fashioned out of a total landmass of 1208 km^2 for the entire Frydek Mistek district (39.38 percent for farmland activities and 49.36 percent used for forest cover). The soil's colour scheme, structure, and calcareous content can all be differentiated. Notwithstanding, the soil's parent materials have an intermediate to smooth texture. However, they are often commonly reported in aeolian and colluvial deposits, which are further best described by upper and subsoil mottles. These are visible in some soil geographic areas and are frequently accompanied by cementitious materials and bleaching. A cambic diagnostic horizon with a smooth sandy loam composition, a clay content greater than 4%, and a lithic disconnection with low carbonate content distinguishes them (Kozák et al., 2010). Notwithstanding, the most common soil types in the research setting were cambisols and stagnosols (Kozák et al., 2010). These soils can be found across the Czech Republic at elevations ranging between 160.6 m and 455.1 m for stagnosol and 59.6–493.5 m for cambisols Vacek et al. (2020).

2.2. Soil analysis and sampling

Some 115 topsoil samples were obtained from cultivated land in the Frydek Mistek local municipality (Fig. 1). The sample design was a regular grid, with soil sample intervals held to 2 × 2 km based on a handheld GPS (Leica-Zeno 5 GPS) device and maximum depths ranging from 0 to 20 cm. Before being transported to the research lab, the soils were pre-labeled and placed in polythene bags. The soil samples were dried with air before being crushed using a Fritsch disk mill pulverize machine and mesh sieved to achieve a finely ground and homogeneous soil sample (less than 200 mesh, 74 μm). Each 1 g processed sample (i.e., powdered, thoroughly mixed, mesh sieved) was inserted into a labeled Teflon bottle. For each Teflon bottle, 7 ml of 35% HCl and 3 ml of 65% HNO3 were added (via fully automated dispensers—one per acid), and the lid was delicately sealed to allow the sample to remain overnight for sample reactions to occur (aqua regia procedure) (Cools, 2016). After dissolving the soil sample, the mixed solution was placed for 120 min on a hot plate (metal) to facilitate digestion before being allowed to cool. The supernatant was obtained by filtering the mixture. The supernatant was poured into a 50-ml Pyrex beaker and watered down to the same volume with deionized water. The watered-down supernatant was then filtered further into 50 ml PVC tubes. Furthermore, 1 ml of the diluted concentration was diluted with 9 ml of de-ionized water and filtered into a 12 ml test tube to measure the pseudo-total PTE concentration in the solution. Following standard procedures and protocols, ICP-OES (inductively coupled plasma-optical emission spectrometry) (Thermo Fisher Scientific Corporation, USA) was used to determine the content levels of zinc, magnesium (Mg), potassium (K), iron (Fe), copper (Cu), and phosphorus (P) respectively. The detection limits of the elements were 0.0060 mg L-1 (Cu), 0.0184 mg L-1 (Fe), 0.0934 mg L-1 (k), 0.0029 mg L-1 (Mg), 0.0067 mg L-1 (P) and 0.0060 mg L-1 (Zn). Moreover, quality control and quality assurance (QC/QA) processes were guaranteed by going over the reference guidelines for each analysis. The duplicate analysis was conducted to ensure that errors were kept to a minimum.

Modeling using machine learning algorithms (MLAs).

Machine learning algorithms (MLAs) applied were, extreme gradient boosting, conditional inference forest, support vector machine, M5 tree model and partial least squares regression. Based on these MLAs, the datasets were each randomly split into two parts, the testing (25%) and training (75%) set. The training set generated the regression models showing the relationships between the response variable (i.e. zinc) and the predictor variables (i.e. Vis-NIR and soil chemical property data) while the testing data evaluated the performance of each model. Descriptions of the models are provided below together with the packages applied in the software R.

2.3. Conditional inference forest (CIF)

Conditional inference forest is a tree-growing method that is commonly used in bio-informatics applications (Nicodemus et al., 2010). Theoretically, CIF differs from the conventional random forest in that it has more robust splitting capabilities (Hothorn et al., 2006). These splitting capabilities result in less biased variable selections (Delerce et al., 2016). The software R package "party" was used to run this model.

2.4. Extreme gradient boosting (EGB)

The extreme gradient boosting (EGB) model is a form of decision tree algorithm although with slightly better performance (Chen and Guestrin, 2016). Most EGB applications are noted in mining (Chen and Guestrin, 2016). Furthermore, the EGB model can optimize as well as adjust its hyper-parameters based on the dataset being applied (Nguyen et al., 2022). For this study, the EGB model was applied and executed through the "XGBoost" package in software R.

Partial least squares regression (PLSR).

The PLSR algorithm has been widely applied for spectral data. This model is advantageous because it can eliminate the challenges associated with multidimensionality between different predictor variables (Mao et al., 2021). This model resembles a linear regression model by assuming a linear association between the response and predictor variables (Gamon et al., 1992). Ehsani et al. (1999) provide more details about this model. The PLSR model was implemented using the R package "PLS".

2.5. Support vector machine regression (SVMR)

Support vector machine is another famous model widely applied in different disciplines (Li et al., 2014). Initially proposed by Vapnik (1995) for supervised classification, this model now features in regression-based problems. In this study, the regression version of SVM (support vector machine regression-SVMR) is applied. This model creates an optimal disconnecting hyperplane to differentiate classifications that are comparable but not linearly autonomous (John et al., 2020). The SVMR model was applied in the software R based on the package "e1071".

2.6. M5 tree model (M5)

The M5 model is a kind of decision tree centered on regression than classification tasks (Etemad-Shahidi and Mahjoobi, 2009). Originally this model was developed by Quinlan (1992) but later Wang and Witten (1996) improve this model and call it M5 trees. The M5 initially builds

several decision trees via recursive splitting. At the end of each tree, a linear function is grafted following the pruning of the overgrown trees (Etemad-Shahidi and Mahjoobi, 2009). The M5 model was run using the "Rweka" package.

2.7. Spectral data preprocessing

Obtaining spectra reflectance measurements in the field and the laboratory is usually not easy because of the various unsuitable discrepancies. Thus, pretreatment and preprocessing of the spectra reflectance measurements are necessary to correct such discrepancies (Biney et al., 2022; Dor et al., 2015). One of the most obvious benefits of using data pre-processing is that it can help mitigate or significantly decrease the number of undesirable variations that can occur during sample collection and laboratory processing. Some of these undesirable inconsistencies include missing values, baseline variations, noises, and so on. According to Engel et al. (2013) these differences can sometimes mask the "true" chemically significant relationship of the internal mechanism, lowering the predictive outcome of the variable of interest. Before being subjected to the pre-processing methods described, the raw spectra were converted to reflectance. Multiple sample composite characteristics and spectroradiometer operational circumstances usually cause some nonlinear characteristics between the independent and dependent variables which result in random noise, numerous different scattering impacts, and threshold drift. Hence, some spectral preprocessing techniques are applied to mitigate these issues. The spectral data had a wavelength range of 350–2500 nm; some spectral ranges from 350 to 400, as well as 2401–2500 nm, were eliminated due to noise, and the spectral range from 400 to 2400 nm was pretreated. The following pretreatment techniques were applied in this study: Savitzky-Golay filter (SG), logarithmic transformation (log (1/R), standard normal variate (SNV), maximum reflectance correction (CMR), discrete wavelet transformation (DWT), multiplicative scatter correction (MSC), and raw spectra dataset. Moreover, several combinations such as DWT-CMR, SG-LOG-MSC, SG-LOG-SNV, SG-SNV-MSC, DWT-SNV-MSC and DWT-LOG-MSC were adopted to determine their applicability and performance in comparison to the individual pretreatment methods. The spectral set of data is influenced primarily by the device's processing, acquisition, and environmental factors (Dor et al., 2015; Mao et al., 2021) The pretreatment procedure was done in the software R, and the R packages employed were libraries signal, KernSmooth, pls, wavelets, and tripack.

3. Mapping procedures

3.1. Sequential Gaussian simulation (SGS) maps

The basic concept behind SGS is to simulate consecutive grid points using the provisional proportion of the empirical distribution (i.e., in this case, the PTEs data). It generates an output resembling the actual spatial reality of a variable of interest. Although it is expected that the datasets are identifiable, the interpolated points resemble the variogram approach and the local noise of the nugget effect (Goovaerts, 2001). Besides that, it is based on a random feature model's multi-Gaussianary assumption (Goovaerts, 2001; Johari et al., 2020). The dataset provides crucial regular score modification, ensuring at least the logic of the univariate data distribution. For more information on SGS refer to Gholampour and iranica (2019). In this study, the SGS method was mainly used to show the spatial distribution and characteristics of each soil chemical property. For each of the chemical properties, 1000 realizations of each property were generated and eventually, the means for each were computed. These are the individual maps observed in Fig. 3.

3.2. Bivariate maps

Bivariate mapping is the technique of classifying spatial objects like grid cells or area polygons based on the values of two parameters (Speich et al., 2015). A single colour legend is used to visualize the two variables as a single output called a bivariate colour scheme. A bivariate map displays the spatial interactions between two raster layers (Tyner, 2010). Spatial associations can then be explored as a single output map for different applications of interest. When two variables exhibit some spatial relationship, it is an indicator that there is some dependence between them. Beard & Mackaness (2006) express analogous points of view in the context of the ambiguity visual representation scenario, in which the characteristic and a method for determining its unpredictability are symbolically represented in a bivariate map. Moreover, the efficiency of bivariate maps is compared and shown to vary by multiple pieces of research, and in each case, the results are dependent on the map reader's knowledge and experience (Hope and Hunter, 2013; Roth, 2013). For details about the bivariate mapping procedure in digital soil mapping and regarding the theory we refer to research by Trumbo (1981). The Zn over/underestimation and Zn prediction raster layers were both used to generate a bivariate map showing the spatial associations of the two features as a single raster layer elucidating the cold and hot spots regions for contexts 1 and 2 respectively.

3.3. Zinc prediction and uncertainty maps

To map the Zn levels in the study area, a regression kriging approach was implemented whereby each of the models, extreme gradient boosting, conditional inference forest, support vector machine, M5 tree model and partial least squares regression were combined with ordinary kriging (i.e. hybridization). A covariate grid comprising the Vis-NIR and soil chemical spatial information was used to predict the Zn levels across the study area. Having obtained the final Zn predictions from each ordinary kriging hybridized model, the lower and upper prediction intervals would be generated by adding and subtracting 1.960 multiplied by the square root of the kriging variances for each model. The lower and the upper prediction intervals were used as uncertainty estimates for the Zn models. We refer to Kebonye et al. (2022) for the explanations regarding the hybridization procedure.

3.4. Assessment accuracy and validation of the models

The coefficient of determination (R2), the ratio of performance to interquartile range (RPIQ), bias, root mean square error (RSME), and median absolute error (MdAE) were used to assess the accuracy and validation of the modeling techniques used in this study. The R2, which represents the variance of the proportion in the response, is expressed by the regression model. The RPIQ is defined as the interquartile range (IQ = Q3-Q1) divided by the RMSE, and it represents the propagation of wider population residuals (Bellon-Maurel et al., 2010). The RMSE determines the size of the different variants within the individual measurement, which enables the approach prediction accuracy, whereas the MdAE confirms the true measurable value. The higher the R2 and the lower the RMSE, according to Li et al. (2016), the better the prediction and accuracy. According to Wang et al. (2014), for a model to be considered satisfactory/acceptable, it must have an R2 value between 0.5 and 0.75, and if the R2 value is 0.75 or higher, the model is considered good. To be considered satisfactory or good, a model must have R2 values ranging from 0.5 to 0.75 (satisfactory) and 0.75 or higher (good), with corresponding minimal error margins.

4. Results and discussion

4.1. Data description

The study elemental median concentration levels were 75.47 mg/kg (Zn), 18070.73 mg/kg (Fe), 1217.25 mg/kg (K), 608.83 mg/kg (P), and 19.68 mg/kg (Cu). The maximum and minimum concentration values of the elements ranged from 37.48 mg/kg to 272.18 mg/kg for Zn,

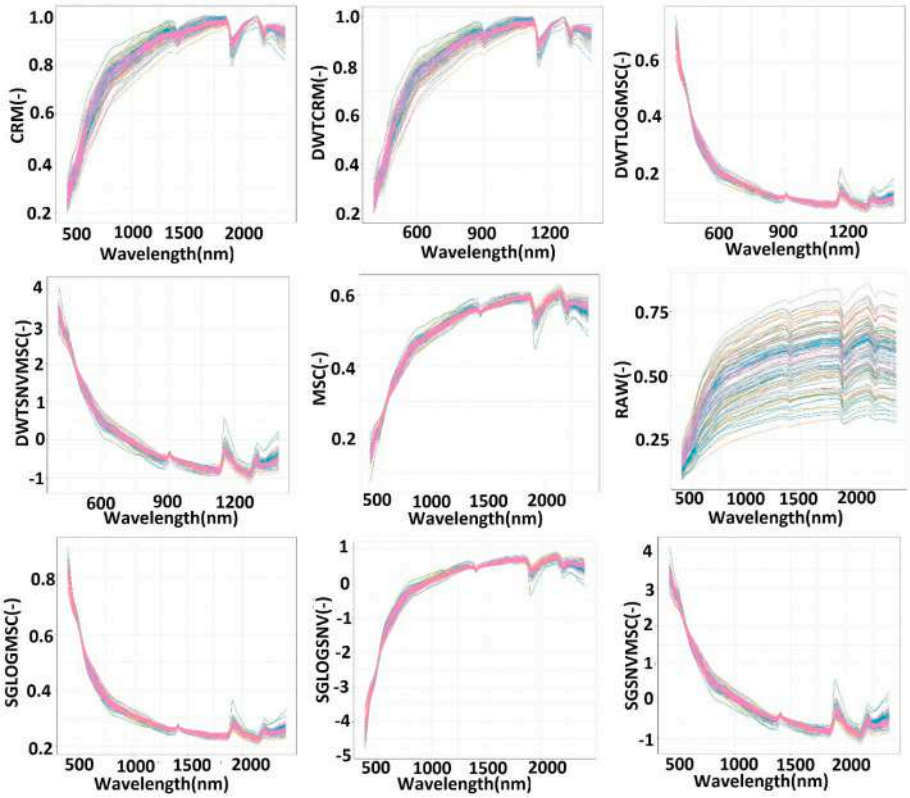


Fig. 2. Visible near infrared reflectance of various pretreated method both single and combined plotted employing the reflectance against the wavelength in RStudio using ggplot {RAW, CMR - correction maximum reflectance, MSC - multiplicative scatter correction, DWT - discrete wavelet transform, DWT-CMR (discrete wavelet transform-correction maximum reflectance), SG-LOG-MSC (savitzky-golay smoothing-logarithm1/R-multiplicative scatter correction). SG-LOG-SNV (savitzky-golay smoothing-logarithm1/R-standard normal variate), SG-SNV-MSC (savitzky-golay smoothing-standard normal variate - multiplicative scatter correction), DWT-SNV-MSC (discrete wavelet transform-standard normal variate-multiplicative scatter correction), DWT-LOG-MSC (discrete wavelet transform-logarithm1/R-multiplicative scatter correction)}.

8650.32 mg/kg to 79901.24 mg/kg for Fe, 497.51 mg/kg to 3535.68 mg/kg for K, 685.68 mg/kg to 5970.05 mg/kg for Mg, 294.55 mg/kg to 2903.09 mg/kg for P, and 7.88 mg/kg to 62.62 mg/kg for Cu. Based on the estimates of kurtosis and skewness the elemental datasets each showed non-normal distributions. Since the modeling approaches chosen for this study are nonparametric, there was no need to execute any form of data transformation before modeling. Based on the coefficient of variation (CV) criteria proposed by Wilding (1985), all the elements exhibited a high CV greater than 35% except for elements K and Mg, which displayed moderate CV estimates. The element levels decreased in the same order for all the percentiles 25th, 50th and 75th respectively. The decrease order was Fe > Mg > K > P > Zn>Cu. Conversely, Horák et al. (2018) reported the mean concentration of K (19045.20 mg/kg), P (467.23 mg/kg) and Cu (36.51 mg/kg) in arable land around Lovětín, Czech Republic. The mean concentration of P in the current study is relatively higher although estimates for K and Cu are lower. The

Table 1
Using visible near-infrared reflectance to predict Zn concentration (Context 1-model validation).

Auxiliary Dataset	R^2 CIF	RMSE	MdAE	RPIQ	R^2 PLSR	RMSE	MdAE	RPIQ	R^2 M5 TREE	RMSE	MdAE	RPIQ
RAW	0.70	21.81	13.78	1.48	0.60	24.65	16.26	0.99	0.66	22.56	14.18	1.43
CMR	0.71	21.44	11.82	1.53	0.18	31.28	23.55	0.11	0.65	22.95	13.51	1.36
MSC	0.70	21.42	9.89	1.50	0.55	25.28	13.82	0.99	0.72	21.08	13.69	1.63
DWT-CMR	0.71	21.46	12.59	1.52	0.17	31.31	23.63	0.11	0.66	22.79	11.88	1.32
SG-LOG-MSC	0.70	21.72	10.68	1.62	0.56	24.98	13.49	1.01	0.64	22.58	11.81	1.01
SG-LOG-SNV	0.65	22.51	11.32	1.44	0.26	29.20	21.78	0.12	0.61	23.59	14.88	1.30
SG-SNV-MSC	0.70	21.71	10.65	1.63	0.55	25.29	13.71	0.99	0.67	22.11	12.24	1.46
DWT-SNV-MSC	0.71	21.57	10.28	1.62	0.56	25.14	13.81	1.01	0.67	22.56	10.06	1.43
DWT-LOG-MSC	0.69	22.03	11.45	1.62	0.56	24.84	12.98	1.03	0.49	26.47	16.77	0.71
	R^2 EGB	RMSE	MdAE	RPIQ	R^2 SVMR	RMSE	MdAE	RPIQ				
RAW	0.59	23.82	12.50	1.18	0.30	28.37	16.59	0.31				
CMR	0.51	25.73	15.18	0.90	0.52	26.25	14.23	0.29				
MSC	0.64	22.87	13.00	1.11	0.52	24.98	13.77	0.77				
DWT-CMR	0.47	28.17	18.07	0.66	0.47	27.12	14.92	0.28				
SG-LOG-MSC	0.41	27.64	12.80	0.76	0.50	25.31	14.60	0.75				
SG-LOG-SNV	0.27	31.27	14.42	0.79	0.47	25.76	13.31	0.82				
SG-SNV-MSC	0.53	24.83	12.16	0.99	0.52	25.03	13.68	0.76				
DWT-SNV-MSC	0.64	22.82	12.46	1.08	0.47	26.12	14.31	0.62				
DWT-LOG-MSC	0.51	25.38	13.43	0.84	0.45	26.33	14.66	0.63				

Note: RAW, CMR - correction maximum reflectance, MSC - multiplicative scatter correction, DWT - discrete wavelet transform, DWT-CMR (discrete wavelet transform-correction maximum reflectance), SG-LOG-MSC (savitzky-golay smoothing-logarithm1/R-multiplicative scatter correction). SG-LOG-SNV (savitzky-golay smoothing-logarithm1/R-standard normal variate), SG-SNV-MSC (savitzky-golay smoothing-standard normal variate - multiplicative scatter correction), DWT-SNV-MSC (discrete wavelet transform-standard normal variate-multiplicative scatter correction), DWT-LOG-MSC (discrete wavelet transform-logarithm1/R-multiplicative scatter correction.

above-mentioned results may also suggest differences in agricultural activities across many parts of the country. Furthermore, it may be that various lithological factors uniquely contribute to the content levels of different abundant elements like K and Fe in agricultural soils. The mean Zn concentration in this study is relatively high when compared to the mean Zn concentration (80 mg/kg) of the local background value reported by Nemecek & Podlesakova (1992). According to the current study, the agricultural soil Zn levels may indicate some form of enrichment over time.

4.2. Spectral response of soil samples

Aside from the raw spectra reflectance, the selected spectra used in this study are shown in Fig. 2 after being pretreated using either a single pretreatment technique or a combined pretreatment technique. The diversity of the measured reflectance from the 115 soil samples is depicted on the raw plot in Fig. 2, which shows a weak peak of absorption from 450 nm to 650 nm that appears to overlap with the visible region. According to Song et al. (2012), some features in the visible region wavelengths including 430 nm, 500 nm, 530 nm, and 650 nm are caused by electronic transitions of the Fe3 + in oxy/hydroxides. Consequently, there is a possibility of Iron oxide absorption or bonding with other metal cations or hydroxyl groups, which has a visible spectral activity (Wu et al., 2005). CRM, RAW, DWTCRM, MSC, and SGLOGSNV pre-treated spectra reflectance plots are very similar. This was based on the soil samples' comparatively similar spectrally active properties, as well as differences between iron oxide minerals, which exhibit varying spectral responses at dissimilar wavelengths. The pretreated plotted DWTLOGMSC, DWTSNVMSC, SGLOGMSC, and SGSNVMSC, on the other hand, exhibit an inversely proportional characteristic to the pretreated plotted spectra reflectance CRM, DWTCRM, MSC and SGLOGSNV. The O–H clay minerals are impacted by the typical locations of the spectral reflectance peaks (Kooistra et al., 2003; Song et al., 2012). In contrast, the peak region 1410 nm–2210 nm of the spectra shown in Fig. 2 is commonly thought to be associated with the hydroxyl (O–H) bond (White, 1971). The peak is a composite material made up of clay minerals; otherwise, the hydroxyl includes kaolinite and smectite clay forms, with the former typically exhibiting peaks at 2210 nm (Nayak and Singh, 2007). Even so, some frail peaks around 2250 nm and 2450 nm were linked to the C–H bond in the organic matter, particularly lignin and humic acid (Ben-Dor et al., 1997), as well as carbonates (Ben-Dor et al., 1997; Ben-Dor and Banin, 1990; Gaffey, 1987).

4.2.1. Predicting Zn concentrations in agriculture soil using VIS-NIR (context 1)

Table 1 shows the validation of the models for predicting Zn concentration in agricultural soil using the Vis-NIR spectral dataset (Context 1). Figure SF1, on the other hand, shows a scatter plot of the measured and predicted values of the calibration model (the best model in each modeling approach), displaying the line of best fit and the relationship between the predicted and measured values. On the raw spectra reflectance, a single pretreatment technique and a combination of two or three pretreatment techniques were used, based on the five modeling techniques, to predict Zn in agricultural soil. The results showed that combining the CIF modeling approach with Vis-NIR datasets (RAW, CMR, MSC, DWTCMR, SGLOGMSC, SGLOGSNV, SGSNVMSC, DWTSNVMSC, DWTLOGMSC) produced satisfactory results in Zn prediction in agricultural soils. More specific, the combination of CIF and the MSC pretreated dataset yielded the overall best Zn prediction results (R2 = 0.70, RMSE = 21.42 mg/kg, MdAE = 9.89, RPIQ = 1.50) (See Table 1 model validation). The application of PLSR along with the Vis-NIR spectral dataset likewise produced satisfactory results, except for CMR and DWTCRM, which generated marginal results. Nevertheless, DWTLOGMSC dataset, along with the PLSR (R2 = 0.56, RMSE = 24.84 mg/kg, MdAE = 12.98, RPIQ = 1.03), was the best approach for Zn prediction in the agricultural soils. Based on the M5 modeling technique and the Vis-NIR spectral dataset, the results are satisfactory apart from the DWT-LOG-MSC dataset, which only yielded abysmal results. The M5 tree model, combined with the MSC dataset (R2 = 0.72, RMSE = 21.08 mg/kg, MdAE = 13.69, RPIQ = 1.63), was the best method to predict Zn in agricultural soil. The application of the EGB modeling approach coupled with the Vis-NIR spectral dataset revealed that out of the 9 spectral auxiliary datasets, 3 (DWTCRM, SGLOGSNV, SGLOGMSC) of the Vis-NIR spectral datasets applied along with EGB produced unsatisfactory results. However, combining the DWTSNVMSC dataset and the EGB modeling approach (R2 = 0.64, RMSE = 22.82 mg/kg, MdAE = 12.46, RPIQ = 1.08) produced the best Zn prediction results in agricultural soil. The combination of SVM and the VIS-NIR spectral reflectance yielded satisfactory results for 4 (MSC, CMR, SGLOGMSC, SGSNVMSC) of the 9 VIS-NIR spectral datasets used as the auxiliary dataset for Zn prediction in agricultural soil. The optimal approach was the combination of MSC dataset and the SVM (R2 = 0.52, RMSE = 24.98 mg/kg, MdAE = 13.77, RPIQ = 0.77). When comparing the optimal prediction performance per modeling technique in context 1, it was evident that M5-MSC produced the best R2 value of 0.72, followed by CIF-MSC R2 = 0.70, EGB-DWTSNVMSC R2 = 0.64, PLSR-DWTLOGMSC R2 = 0.56, and SVMR-MSC R2 = 0.52. The estimated RMSE of the optimal approaches, however, revealed that M5-MSC had the lowest RMSE value of 21.08 mg/kg, followed by CIF-MSC RMSE = 21.42 mg/kg, EGB-DWTSNVMSC RMSE = 22.82 mg/kg, PLSR-DWTLOGMSC RMSE = 24.84 mg/kg, and SVMR-MSC RMSE = 24.98 mg/kg. The MdAE estimated values of the optimal modeling technique, on the other hand, revealed that CIF-MSC had the lowest MdAE value of 9.89, followed by EGB-DWTSNVMSC MdAE = 12.46, PLSR-DWTLOGMSC MdAE = 12.98, M5-MSC MdAE = 13.69, and SVMR-MSC MdAE = 13.77. The optimal modeling approach's estimated RPIQ revealed that M5-MSC had the highest RPIQ of 1.63, followed by CIF-MSC RPIQ = 1.50, EGB-DWTSNVMSC RPIQ = 1.08, PLSR-DWTLOGMSC RPIQ = 1.03, and SVMR-MSC RPIQ = 0.77. The overall evaluation of the modeling approaches in context 1 suggests that the combination of the M5 tree model and the MSC dataset was the best approach in context 1 for predicting Zn concentration levels with higher accuracy and minimal errors.

The M5 tree model is a linear technique that estimates interdependences in the collection of data by exposing nonlinear data from each approach. The splitting requirements of an M5 model tree are derived from the computation of discrepancy at each node. The standard error of the class values that arrive at a node is used to examine the mistake in the M5 tree model. The model attribute with the highest projected results for error reduction is occasionally selected for partitioning there after each model attribute has been evaluated at that node (Sihag et al., 2019). Sihag et al. (2019), compared the performance of MLAs such as multilayer perceptron neural networks and the M5 model tree, and found that the M5 tree model predicted Zn content in the soil better than the multilayer perceptron neural network. M5 model Tree can present high correlation coefficient values and minimal MAE and RMSE errors and has been discovered to be more suitable for estimating bearing capacity (Khorrami et al., 2020). Sattari et al. (2018) applied the M5 tree model and support vector machine model in the prediction of ground water level, however, the author reported that the M5 tree model produces results that are simpler and clearer, easier to apply, and easier to decipher than the SVMR algorithm. The M5 model tree with wavelet regression can be more effective than artificial neural network modeling techniques in estimating sediment yield, as a comparison of the two reveals that the M5 model tree provides precise and understandable expressions for use by design engineers (Goyal, 2014). Agyeman, et al., (2022) used the M5 modeling technique and other machine learning algorithms, as well as spectral indices and terrain attribute data sets, to estimate cadmium concentration in cultivated soil. They discovered that the M5 modeling technique was the best approach for estimating Cd concentration in the soil with high performance and minimal errors.

Table 2
Using visible near-infrared reflectance and soil chemical properties to predict Zn concentration (Context 2-model validation).

Auxiliary Dataset	R^2 CIF	RMSE	MdAE	RPIQ	R^2 PLSR	RMSE	MdAE	RPIQ	R^2 M TREE	RMSE	MdAE	RPIQ
RAW + SCP	0.86	17.70	8.43	1.22	0.69	21.30	9.69	0.95	0.83	16.57	8.88	1.99
CMR + SCP	0.86	18.29	11.34	0.96	0.30	28.10	18.89	0.42	0.71	21.22	12.79	0.92
MSC + SCP	0.86	18.00	9.83	1.10	0.59	24.41	14.84	1.24	0.73	20.36	14.43	1.26
DWT-CMR + SCP	0.85	18.18	10.24	1.09	0.42	26.54	16.85	0.70	0.83	17.13	10.52	1.60
SG-LOG-MSC + SCP	0.86	17.69	10.34	1.10	0.59	24.07	13.29	1.27	0.79	18.46	10.12	1.21
SG-LOG-SNV + SCP	0.85	18.22	9.88	1.06	0.27	29.14	21.74	0.13	0.73	20.59	13.20	1.15
SG-SNV-MSC + SCP	0.86	17.88	10.24	1.18	0.58	24.41	14.73	1.24	0.73	20.28	14.40	1.54
DWT-SNV-MSC + SCP	0.87	17.61	9.76	1.08	0.62	23.36	15.62	1.34	0.69	21.35	15.21	1.25
DWT-LOG-MSC + SCP	0.86	14.52	6.25	1.78	0.63	23.05	13.96	1.34	0.77	19.11	10.38	1.36
	R^2 EGB	RMSE	MdAE	RPIQ	R^2 SVMR	RMSE	MdAE	RPIQ				
RAW + SCP	0.89	15.72	7.33	1.74	0.74	21.98	9.97	0.79				
CMR + SCP	0.72	21.03	10.12	1.05	0.67	24.59	13.35	0.43				
MSC + SCP	0.73	21.71	11.55	0.73	0.65	23.23	13.15	0.73				
DWT-CMR + SCP	0.76	19.88	10.02	1.19	0.65	25.75	14.31	0.33				
SG-LOG-MSC + SCP	0.74	20.66	10.37	0.97	0.61	23.99	13.86	0.72				
SG-LOG-SNV + SCP	0.65	22.58	8.15	0.80	0.62	23.78	13.19	0.69				
SG-SNV-MSC + SCP	0.77	20.35	10.74	0.94	0.65	23.26	13.08	0.73				
DWT-SNV-MSC + SCP	0.77	19.93	10.68	1.16	0.60	24.12	12.66	0.64				
DWT-LOG-MSC + SCP	0.72	21.11	11.52	1.41	0.60	24.13	12.50	0.62				

Note: RAW + SCP-(RAW-soil chemical properties), CMR-SCP – (correction maximum reflectance-soil chemical properties), MSC + SCP – (multiplicative scatter correction-soil chemical properties), DWT + SCP – (discrete wavelet transform-soil chemical properties), DWT-CMR + SCP (discrete wavelet transform-correction maximum reflectance-soil chemical properties), SG-LOG-MSC + SCP (savitzky-golay smoothing-logarithm1/R-multiplicative scatter correction-soil chemical properties), SG-LOG-SNV + SCP (savitzky-golay smoothing-logarithm1/R-standard normal variate-soil chemical properties), SG-SNV-MSC + SCP (savitzky-golay smoothing-standard normal variate - multiplicative scatter correction-soil chemical properties), DWT-SNV-MSC + SCP(discrete wavelet transform-standard normal variate-multiplicative scatter correction-soil chemical properties), DWT-LOG-MSC + SCP (discrete wavelet transform-logarithm1/R-multiplicative scatter correction-soil chemical properties).

4.3. Predicting the concentration of Zn using visible near-infrared reflectance and soil chemical properties (context 2)

Table 2 shows zinc predictions using VIS-NIR spectra reflectance and soil chemical properties (SCP) (Mg, K, Cu, P, and Fe) as ancillary datasets for the model's validation. Figure SF2(calibration model) depicts a scatter plot of the calibration model's measured and predicted values (the best model in each modeling approach), displaying the line of best fit and the relationship between the predicted and measured values. These soil chemical properties are micronutrients (Fe, Cu) and macronutrients (Mg, K, P) which interact differently with Zn. The results indicated that in the CIF modeling approach, combining the Vis-NIR + SCP with the CIF modeling techniques yielded good predictions with R2 values ranging between 0.85 and 0.87. Similarly, the RMSE, MdAE and RPIQ validation assessment criteria also ranged from 14.52 to 18.29 mg/kg (RMSE), 6.25 to 11.34 (MdAE) and 0.96 to 1.78 (RPIQ). Based on the results, it was clear that CIF-DWTLOGMSC + SCP (R2 = 0.86, RMSE = 14.52 mg/kg, MdAE = 6.25, RPIQ = 1.78) was the best predictive model approach for predicting Zn in soil with the minimum RMSE and MdAE values and the high RPIQ value. In the PLSR modeling approach, the results suggested that six of the 9 Vis-NIR + SCP combined with the PLSR modeling technique yielded satisfactory results with the other 3 producing unsatisfactory results. The R2, RMSE, MdAE and RPIQ values ranged between 0.27 and 0.69, 21.30–29.14 mg/kg, 9.69 to 21.14 and 0.13 to 1.34 respectively. The results indicated that the combination of PLSR and the auxiliary dataset RAW + SCP (R2 = 0.69, RMSE = 21.30 mg/kg, MdAE = 9.69, RPIQ = 0.95) was the optimal approach in the prediction of Zn in the soil with high R2 and minimal RMSE and MAE values. The M5 tree model approach combination with the auxiliary dataset yielded good and satisfactory results for all the 9-dataset used. However, the R2, RMSE, MdAE and the RPIQ results ranged from 0.71 to 0.83 (R2), 16.57–21.35 mg/kg (RMSE), 8.88 to 15.21 (MdAE) and 0.92 to 1.99 (RPIQ) correspondingly. The best combination pair in the PLSR modeling approach was the combination of PLSR and the RAW + SCP (R2 = 0.83, RMSE = 16.57 mg/kg, MdAE = 8.88, RPIQ = 1.99) auxiliary dataset. The EGB modeling approach coupled with the Vis-NIR + SCP datasets doled out good results with the R2, RMSE, MdAE and RPIQ values ranging between 0.65 and 0.89, 15.72–22.58 mg/kg, 7.33 to 11.55 and 0.94 to 1.74 correspondingly. Notwithstanding, the RAW + SCP dataset and PLSR were the most effective combinations for predicting the concentration of Zn in agricultural soil, with R2 = 0.89 and RPIQ = 1.74 values and low RMSE and MdAE values of 15.72 mg/kg and 7.33, respectively. Finally, the results proved that the results from the combination of SVMR, SCP, and Vis-NIR were satisfactory, with the R2, RMSE, MdAE, and RPIQ values ranging from 0.60 to 0.74 (R2), 21.98–25.75 mg/kg (RMSE), 9.97 to 14.31 (MdAE), and 0.62 to 0.79 (RPIQ). The outcome further revealed that the combination of SVMR and the auxiliary dataset RAW + SCP (R2 = 0.74, RMSE = 21.98 mg/kg, MdAE = 9.97, RPIQ = 0.79) produced the best results for Zn prediction in agricultural soil.

When the optimal modeling approaches for the five techniques were compared, it was definite that EGB-RAW + SCP had the best R2 value of 0.89, followed by CIF-DWTLOGMSC + SCP R2 = 0.86, M5-RAW + SCP R2 = 0.83, SVMR-RAW + SCP R2 = 0.74, and PLSR-RAW + SCP R2 = 0.69. The optimal approaches of the five modeling techniques were compared, and CIF-DWTLOGMSC + SCP had the lowest RMSE and MdAE obtained values, with RMSE and MdAE of 14.52 mg/kg and 6.52, respectively. The other optimal approaches for RMSE and MdAE performance per modeling techniques are as follows 15.72 mg/kg (EGB-RAW + SCP), 16.57 mg/kg (M5-RAW + SCP), 21.30 mg/kg (PLSR-RAW + SCP) and 21.98 mg/kg (SVMR-RAW + SCP) (RMSE) and 7.33 (EGB-RAW + SCP), 8.88 (M5-RAW + SCP), 9.69 (PLSR-RAW + SCP) and 9.97 (SVMR-RAW + SCP) (MdAE). The RPIQ results, on the other hand, indicated that the M5-RAW + SCP had the highest RPIQ (1.99), while the other predicting modeling approaches had 1.78 (CIF-DWTLOGMSC + SCP), 1.74 (EGB-RAW + SCP), 0.95 (PLSR-RAW + SCP), and 0.79 (SVMR-RAW + SCP) respectively. The cumulative assessment of the optimal prediction modeling approaches from the five modeling techniques indicated that CIF-DWTLOGMSC + SCP was the best method that was able to predict Zn in agricultural soil with minimal errors.

CIF which is not commonly applied in the field of soil science, is a type of decision tree technique for iterative binary partitioning. CIF incorporates the framework into a well-defined data analysis setting predicated on factorization tests, to distinguish between substantial and

unimportant advancements (Das et al., 2009). CIF has the proclivity to limit overfitting and model biases. However, CIF quantifies the independent variable significance of each parameter for every tree by initially splitting the relationship into combinations and then running tests on the tree with out-of-bag projections. The variable relevance in forests depends on the outcomes of numerous trees, attempting to avoid the uncertainty of different trees. The benefit of the new forest improvement algorithm over the conventional CART tree/forest is that it precludes ambiguous factors from being recognized as considerable simply because they have a larger number of categories or are iterative (Das et al., 2009). van Wesemael et al., (2019) applied CIF and found that it performed better in the selection of variables that have a strong relationship with a component that is consistent with the model's constraint. The CIF approach has been successfully used to relate aspects of soil heterogeneity to crop development heterogeneity (Goffart et al., 2022). Kapo et al. (2014) outlined that CIF can assess a comprehensive set of ecological variables that led to stressor-response presumptions at the statewide and eco-regional levels. Goydaragh et al. (2021)employed CIF in conjunction with ecological variables to estimate the concentration of soil organic carbon. While Cubist + Ba was the best model in that scenario, CIF + Ba and CIF outperformed random forest, EGB, CART, and conditional inference tree.

4.4. Comparison of the best models in contexts 1 and 2 predicated on modeling techniques

The optimal models from the five modeling techniques applied in Zn prediction in the agricultural soil are CIF-MSC, PLSR-DWTLOGMSC, M5-MSC, EGB-DWTSNVMSC, and SVMR-MSC respectively (Context 1) and CIF-DWTLOGMSC + SCP, PLSR-RAW + SCP, M5-RAW + SCP, EGB-RAW + SCP, and SVMR-RAW + SCP (Context 2). The assessment of the optimal model in each context using CIF modeling techniques based on R2, RMSE, MdAE, and RPIQ revealed that CIF-DWTLOGMSC + SCP performed better in context 2 than the optimal model CIF-MSC in context 1 (See Table ST1-model validation). The performance of the Vis-NIR spectra in conjunction with the SCP and CIF in context 2 resulted in an increase in the R2 value of 10.33%, an increase in the RPIQ value of 8.38%, and a decrease in the marginal errors of 19.21% for RMSE and 22.53% for MdAE. The combination of PLSR and the auxiliary datasets RAW + SCP (context 1) and DWT-LOG-MSC (context 2) and their output revealed that the fusion of PLSR with RAW + SCP predicted Zn content in the soil better than the application of PLSR and DWT-LOG-MSC. The prediction efficiency of the best models in PLSR revealed that R2 increased by 10.43% in context 2 over context 1, while marginal errors decreased by 7.67% and 14.53% for RMSE and MdAE in context 2 over context 1. The best prediction approach in the M5 tree model suggested that RAW + SCP (context 2) and MSC (context 1) were the best combinations for predicting the concentration of Zn in agricultural soil. Their performance output, however, suggested that M5-RAW + SCP predicted Zn in agricultural soil better than M5-MSC. The R2 and RPIQ values of M5-RAW + SCP are 7.54% and 9.89% higher than those of M5-MSC, respectively, while the RMSE and RPIQ values of M5-RAW + SCP are 11.97% and 21.33% lower than those of M5-MSC. Conversely, RAW + SCP and DWT-SNV-MSC were optimal models in EGB for both contexts. The output of the models EGB-RAW + SCP and EGB-DWTSNVMSC revealed that EGB-RAW + SCP model performed better than EGB-DWTSNVMSC with R2 and RPIQ 16.83% and 23.36% higher whiles the RMSE and MdAE values decreased by 18.45% and 25.89% in EGB-RAW + SCP than EGB-DWTSNVMSC. The RAW + SCP and MSC auxiliary dataset set combined with SVMR was the better auxiliary dataset for Zn prediction in the agricultural soil. Nonetheless, the combination of RAW + SCP and SVMR produced the best results when compared to MSC and SVMR. The study's results exhibited that while RMSE and MdAE error dropped by 6.39% and 15.98% in SVMR-RAW + SCP, respectively, the R2 and RPIQ values in SVMR-RAW + SCP were 17.13% and 1.23% higher than in SVMR-MSC. The cumulative assessment of the optimal models in the five modeling approaches revealed that the combination of Vis-NIR spectral reflectance, soil chemical properties, and machine learning techniques produced the best prediction. Based on this, it was blatantly evident that CIF-DWTLOGMSC + SCP was the best technique for predicting Zn content in agricultural soil overall, with much lower errors than all the best models in each modeling technique and context.

The combination of Vis-NIR spectra reflectance and the influence of micro and macro nutrients (soil chemical properties) on Zn prediction in agricultural soil has produced remarkable results. The interaction of Zn as a micronutrient with the other micro and macronutrient might have had a profound influence on the optimal results in context 2. It is important to note that the antagonistic and stimulation effects of the interaction of soil macro and micronutrients might have accounted for the optimal results in contexts 2. The influence of geomorphological terrain on the quantification of PTEs such as Zn in the soil is significant, and the interactions among bedrock, climatic conditions, and geomorphologic processes may lead to the formation of soil parent composites (Agyeman, et al., 2022). The utilization of MLAs in the estimation of PTEs in soil could consider the soil formation process, which incorporates the mineralogical composition coupled with the geomorphological properties of the soil being studied (Agyeman, et al., 2022; Zeraatpisheh et al., 2020). The right proportion of soil micronutrients and macronutrients provides optimal soil health. However, the act of applying Zn-based fertilizers such as ammoniated zinc, zinc sulfate, and chelated zinc soil to boost fertility may lead to the enrichment of some micronutrients such as Zn, rendering the soil potentially toxic. Kebonye et al. (2021) applied soil chemical properties (i.e., Ca, Ti, Zn, Sr, Zr, Ba, Pb and Th) combined with MLAs in the prediction of the concentration of As in the soil. John et al., (2021) similarly employed MLAs in conjunction with soil chemical properties such as potassium, calcium, sodium, magnesium, phosphorus, and vanadium in the prediction of sulphur in the soil. John et al. (2020) applied soil properties (i.e., Ca, Mg) coupled with terrain properties along with a remote sensing dataset for soil organic carbon prediction in alluvial soil. In the estimation of the content of cadmium and lead in polluted soil in Iran Bazoobandi et al. (2019), applied soil properties such as total nitrogen, phosphorus, and organic carbon. Due to the pedogenesis and evolutionary development of the area under investigation, the use of soil chemical properties as an auxiliary dataset in conjunction with modeling techniques in predicting the content of PTEs in soil has a greater influence on the response variable (e.g., Zn). Agyeman, et al., 2022 predicted nickel concentration in soil using soil chemical properties (Ca, Mg, k) and a hybridized MLA in peri-urban and urban soil. In another study,Hong et al. (2019) used soil chemical properties in addition to Vis-NIR spectral reflectance, and the authors reported that the combination of Vis-NIR spectral reflectance, soil chemical properties, and an appropriate MLA model may improve prediction performance.

Given the introduction of SCP in context 2, the results of the predictive performance of the modeling approaches in both contexts vary significantly. The differences in RMSE values of the modeling approach in contexts 1 and 2 are quite visible, which may be attributed to the significant variation between the predicted and actual measured values. The minimal variation, on the other hand, indicates the closeness of the variation between the response and the predicted values. This is consistent with Biney et al. (2022)'s report that high RMSE values indicate that the predicted and true responses differ significantly, whereas a small RMSE indicates that the predicted and true responses are very close. The application of a single pretreatment algorithm on raw data for the prediction of soil properties or PTEs in the soil is very common and reliable due to its environmental friendliness and cost-effectiveness. The act of hybridizing these known pretreatment algorithms on raw spectral reflectance for the prediction of PTEs or soil properties is uncharted. The use of a pretreatment combined algorithm in conjunction with a single modeling and ensemble models to predict PTEs and soil organic carbon in a variety of soils and conditions have

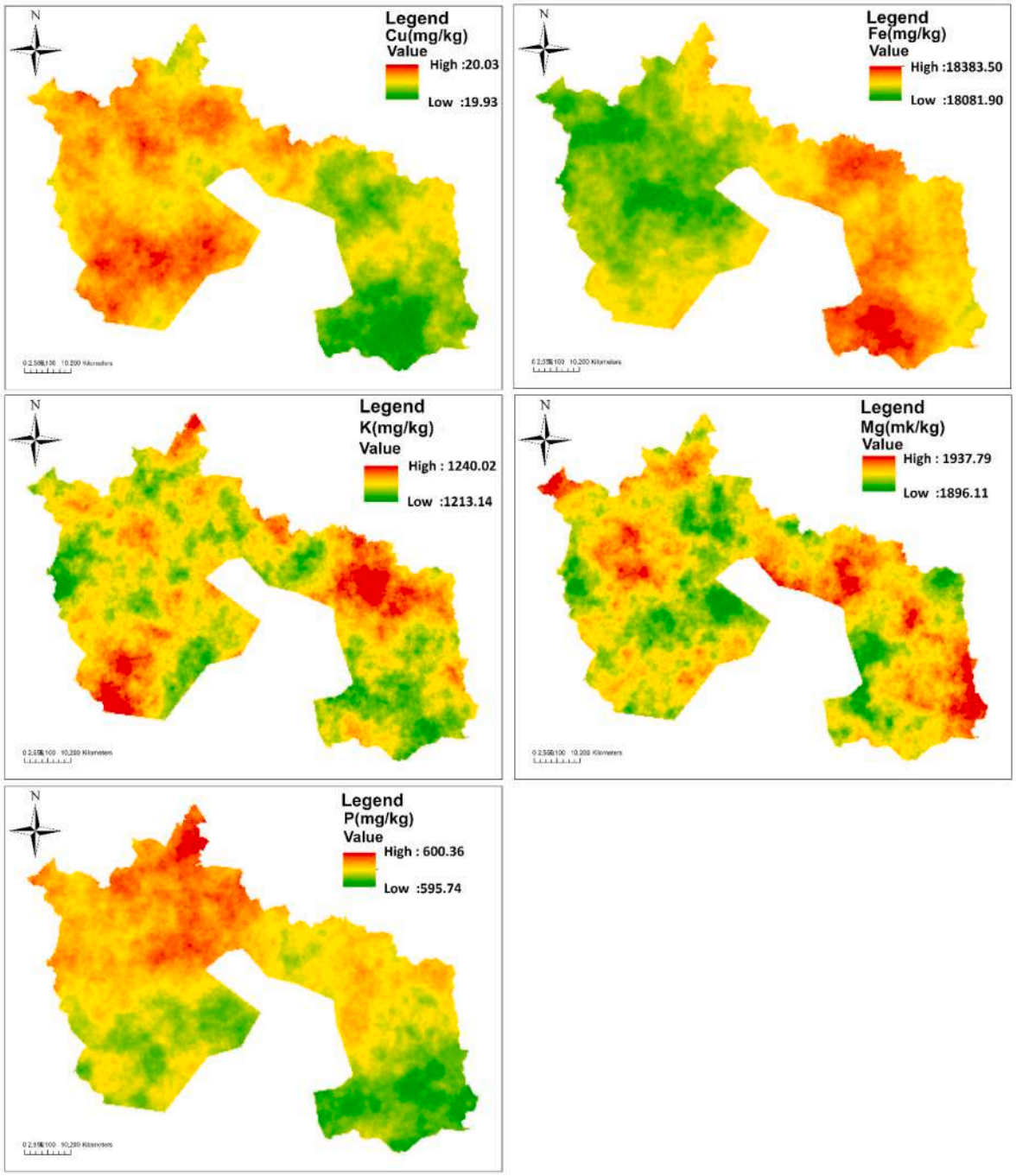


Fig. 3. Mean values (mg/kg) of 1000 SGS realizations of the soil chemical properties (Cu, Mg, K, P and Fe) in agricultural soils.

been tested and proven reliable (Biney et al., 2022; Biney et al., 2022c). The authors used the hybridized pretreatment method in three distinct agricultural fields under three distinct measurement conditions (wet, dry, and field). The combined pretreatment technique could be one option for eliminating or minimizing multiple artifacts at the same time. This implies that the application of combined pretreatment techniques along with an appropriate modeling approach is reliable and could be used in any soil type and under different conditions. The effectiveness and reliability of obtaining satisfactory and good prediction results are dependent on the conditions preceding the selection of appropriate modeling approaches as well as the best pretreatment combination used. Under certain circumstances, the spectral properties trend linked to a specific parameter during spectral quantification may intersect with the response pattern (e.g., PTEs) connected with some other factor. This might positively or negatively affect the prediction outcome due to the likelihood of masking out significant information. According to several authors, these factors may contribute to an increase or decrease in prediction accuracy (Dor et al., 2015; J. Biney et al., 2022). According to Kooistra et al. (2001) no pretreatment is the best pre-processing method for predicting Zn concentration in the soil. Even though the performance of raw spectra reflectance in the prediction of Zn in agricultural soil is relatively high, the use of a combined predicted method and the inclusion of SCP has improved the prediction and reduced errors. Other pre-processing techniques must also be used to explore the impact of numerous data treatment scenarios on subsequent processing results (Khosravi et al., 2018). The current application and combination of pretreatment algorithms, modeling approaches, and SCP in various contexts yielding satisfactory and good results conform to the

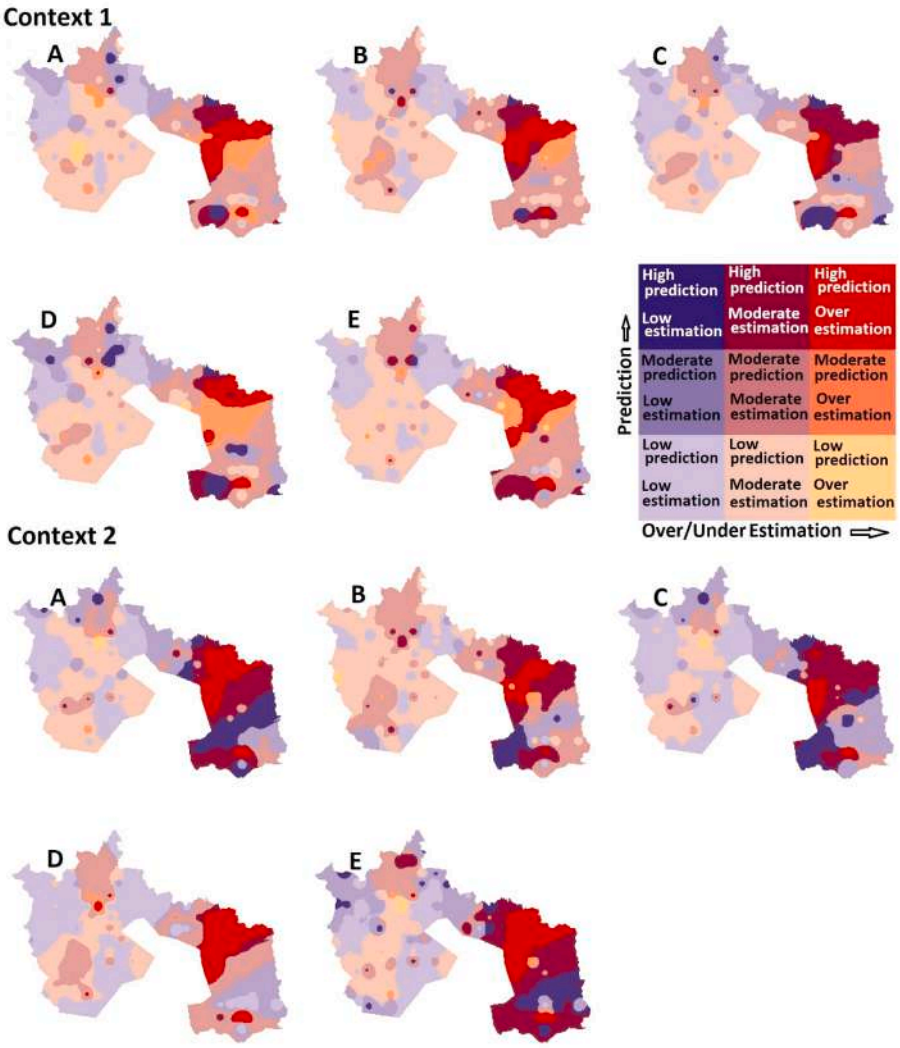


Fig. 4. Bivariate mapping showing the optimum modeling techniques' prediction of Zn concentration in agricultural soil and over or under prediction in prediction contexts 1 and 2 {optimal modeling approaches in prediction context 1 (A = CIF-MSC,B = EGB-DWTSNVMSC, C = M5-MSC, D = PLSR-DWTLOGMSC, E = SVM-MSC) optimal modeling approaches in prediction context 2 (A = CIF-DWTLOGMSC-SCP, B = EGB-RAW-SCP, C = M5-RAW-SCP, D = PLSR-RAW-SCP, SVM = RAW-SCP)}. (Note: For clarity, the bivariate map legend depicts the 9-colour range as well as the rating system used.). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

recommendations of Khosravi et al. (2018).

5. Mapping of the zinc, soil chemical properties and bivariate mapping of optimal modeling approaches and over/under predictions

The distribution of the soil chemical properties spatially in the agricultural soil is presented in Fig. 3. To model the soil chemical properties using SGS, semi-variograms were fitted. For all the semi-variogram plots (the soil chemical properties) the spherical approach was considered suitable. According to Heuvelink et al., (2001), the nugget sill ratio of a good spatial modeling approach should be less than 0.25, which implies that the model has higher spatial autocorrelation. A nugget sill ratio of 0.25–0.75 indicates moderate spatial autocorrelation, whereas a nugget sill ratio of 0.75 or greater indicates weak or poor spatial autocorrelation but greater spatial randomness. In this study, the nugget sill ratio of the elements mapped was above 0.75, except for Fe (0.56), which displayed moderate spatial autocorrelation (See Table ST2). This implies that the elements whose nugget sill ratio was above 0.75 exhibited stronger randomness, which thus further accentuates that the propensity of human activities to impact the concentration of those elements in the soil is somehow exceedingly high. The spatial variability of the elements, particularly Fe, Mg, and K, revealed a hotspot in the northeastern and southern areas of the research site. Mg and K, on the other hand, displayed hotspots in the northwest and southeast regions of the study area. Cu and P exhibited hotspots in the study area's southwest and northwest enclaves.

The bivariate map in Fig. 4 shows the distribution of the optimal prediction in each approach for each context. In Context 1, the CIF-MSC (A), EGB-DWTSNVMSC (B), and M5-MSC (C) approaches exhibited high prediction and overestimation in the southeast area of the map. Similarly, the high prediction and moderate estimation were displayed in the northeastern area of the map. The modeling approaches PLSR-DWTLOGMSC (D) and SVM-MSC (E) exhibited high prediction and overestimation in the northeastern area, and SVM-MSC (E) further showed overestimation in the southeastern area of the map. In the second context, high prediction and overestimation were displayed in the northeastern and southeast regions of the map for the following models: CIF-DWTLOGMSC-SCP (A), EGB-RAW-SCP (B), PLSR-RAW-SCP (D) and SVM = RAW-SCP (E). Equally, the modeling approaches CIF-DWTLOGMSC-SCP (A), EGB-RAW-SCP (B), PLSR-RAW-SCP (D) and SVM-RAW-SCP (E) exhibited high prediction and low estimation in the southeastern area, with patches of high prediction and low estimation in the northeastern region of the map. Again, the modeling approaches showed high prediction and moderate estimation were likewise spotted in the northeast and southeast regions of the map. The modeling approach PLSR-DWTLOGMSC (D) showed a high prediction and overestimation in the northeast and southeast areas of the map. Comparatively, the addition of soil chemical properties to visible near-infrared reflectance in the prediction of Zn in the agricultural soil has exposed

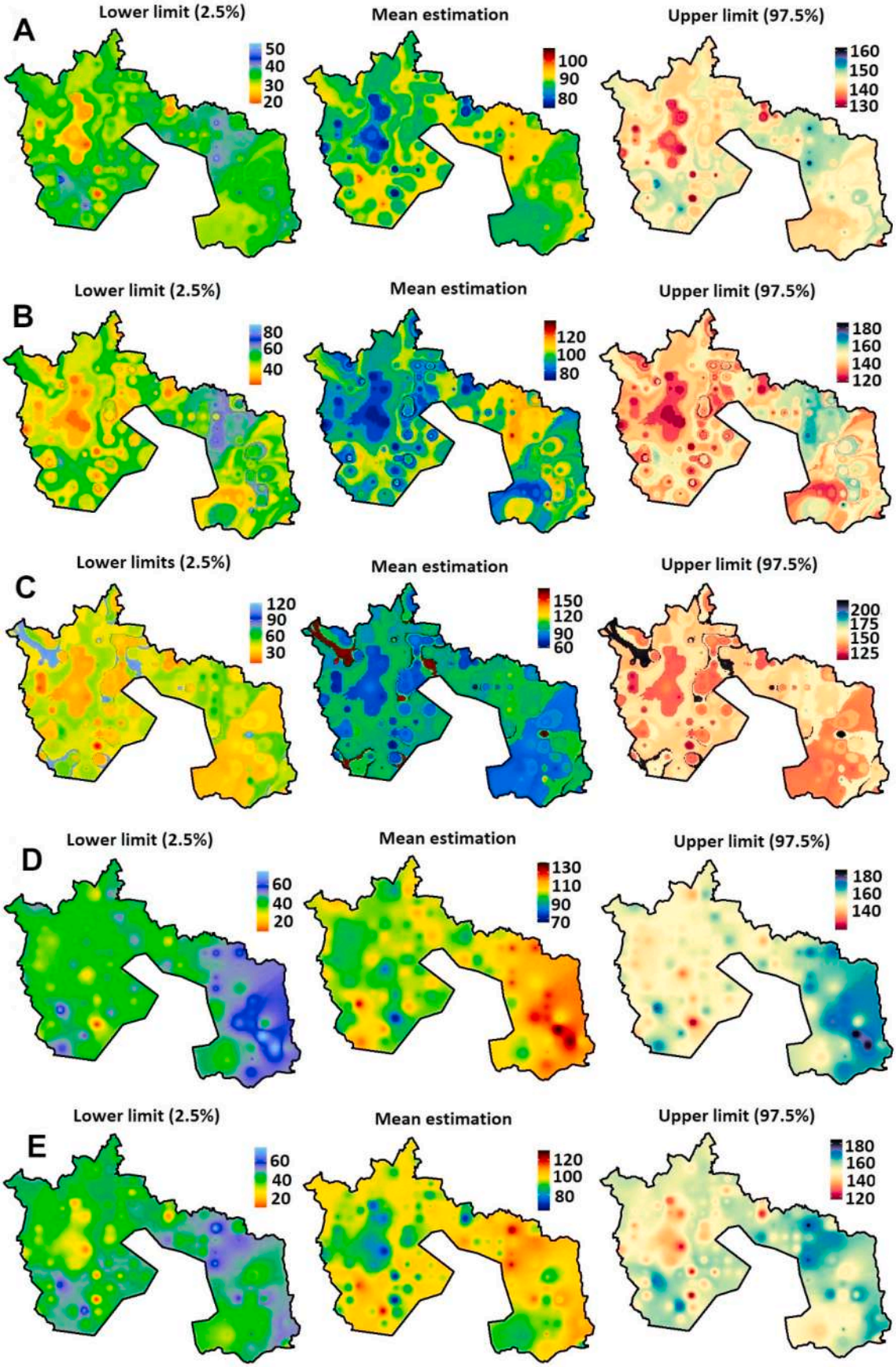


Fig. 5. Uncertainty assessment of Zn in the agricultural employing visible near infrared as the ancillary dataset (Context 1) {conditional inference forest (A), extreme gradient boosting (B), M5 tree model (C), Partial least squares regression (D), support vector machine (E).

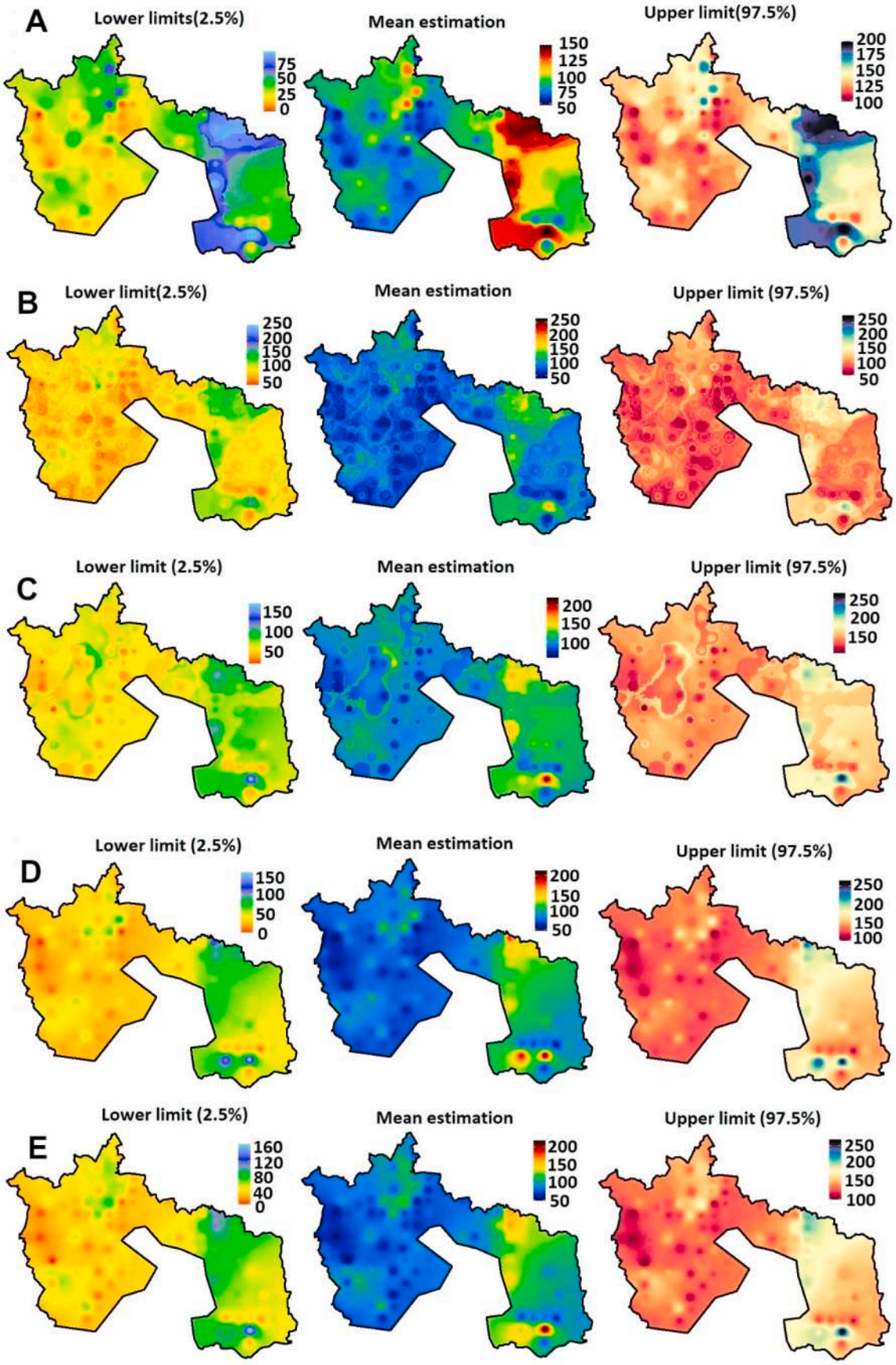


Fig. 6. Uncertainty assessment of Zn in the agricultural soil employing visible near infrared and soil cmeical properties as the ancillary dataset (Context 2) {conditional inference forest (A), extreme gradient boosting (B), M5 tree model (C), Partial least squares regression (D), support vector machine (E).

areas that are high, moderate, and low estimated as well as high prediction areas. In some regions with high predictions, anthropogenic impacts like the use of zinc in steel industries and the application of Zn-based fertilizers may be responsible for the variability of the concentration of zinc in the agricultural soil as shown by the predictions and the over/underestimated bivariate map. The content of PTEs in agricultural soil can be raised by anthropogenic causes such as fertilizers, air deposition, fungicides, the closeness of agricultural locations to industrial facilities, sewage irrigation, and plastic films (Huang et al., 2019). Conversely, the level of PTEs like Zn increases when composted manure, biosolids from animal manure, and compost are added to agricultural soil (Basta et al., 2005). A considerable increase in Zn levels in the soil caused by metal processing in the steel industry is harmful to the ecosystem (Wuana et al., 2011).

5.1. Uncertainty

Uncertainty assessment in mapping and modeling processes is an important part of evaluating the practical limits, potential ramifications, and effectiveness of risk performance analysis that is compatible with decision-making process steps. The uncertainty assessment in this study was performed in two distinct contexts, namely assessment of uncertainty based on the prediction of Zinc using Vis-NIR spectra reflectance (context 1) and uncertainty assessment premised on Zn prediction in the agricultural soil employing Vis-NIR spectra reflectance and soil chemical properties (context 2).

In context 1, the propagation of uncertainty using the lower limit (2.5) revealed that the CIF and EGB modeling approaches exhibited a low to moderate degree of uncertainty, with patches of high uncertainty in the southeast area of the map (Fig. 5). Similarly, the PLSR and SVM showed moderate levels of uncertainty with patches of high uncertainty in the northeast and southeast regions of the map. Nevertheless, the M5 tree model displayed mostly low to moderate levels of uncertainty in the whole study area, with patches of varying degrees of uncertainty in the northwest region of the map. Except for PLSR and SVM, which exhibited an elevated degree of uncertainty in the northeast and southeast areas of the study area, the mean uncertainty assessment for the modeling approaches was generally low to moderate. The upper limit (97.5) of the modeling approaches exhibited a low to moderate degree of uncertainty for all the modeling techniques, with patches of high uncertainty. In context 2, the uncertainties propagated in the lower limit were relatively low to moderate for all the modeling approaches except for the CIF modeling approach, which presented a high degree of uncertainty in the northeast and southeast regions of the study area (Fig. 6). Similarly, the mean uncertainty propagated by the modeling approaches displayed low to moderate uncertainties, with the CIF modeling approach displaying an elevated level of uncertainty in the northeast and southeast regions of the study area. The upper limit uncertainty propagated by the modeling approaches was likewise from low to moderate, with the CIF modeling approach exhibiting an elevated level of uncertainty in the northeastern and southeastern enclaves of the study area. Comparatively, adding the soil chemical properties to the Vis-NIR spectra reflectance has reduced the level of uncertainty propagated in all the modeling approaches in context 2, except for the CIF modeling approach, which exhibited a high level of uncertainty across the prediction intervals. The elevated level of uncertainty propagated in the northeastern and southeastern enclaves by CIF modeling approaches in the study area is expected due to the steel industry, metal works, and intensive agriculture in the southeastern part of the study region.

6. Conclusion

This study employs two distinct methods for Zn prediction in agricultural soil by combining visible near-infrared reflectance spectroscopy with machine learning algorithms (Context 1) and the application of visible near-infrared reflectance spectra, soil chemical properties, and machine learning algorithms (Context 2). The study's findings revealed that the combination of pretreatment techniques, as opposed to the application of a single technique, tends to produce optimal results. The study discovered that combining soil chemical properties (SCP) with visible near-infrared spectra reflectance can improve prediction performance. The study's findings suggested that SCP had a positive impact on RAW spectra reflectance in agricultural Zn prediction, which could be attributed to SCP's antagonistic and stimulating effect on Zn. Even though no pretreatment combined with SCP produced good results, the positive impact of pretreatment of RAW spectra reflectance before use reduced modeling errors, especially in context 2. Because there is no single pretreatment approach that is generically sufficient for soils, the application of combined pretreatment methods to spectral reflectance for prediction is promising. It can generate appropriate spectral reflectance capable of yielding good results, the combination of pretreatment techniques leverages its strengths on the weaknesses of other pretreatment techniques. The study demonstrates that using SCP in conjunction with spectral reflectance and an appropriate modeling approach yielded unrivaled results in uncertainty assessment, lowering the degree of uncertainty propagation in context 2 compared to context 1. The SCP, along with spectral reflectance and modeling approaches, is very appealing for environmental pollution surveillance, prediction assessments, and precision farming.

Credit authors statement

Prince Chapman Agyeman: Conceptualization, Methodology, Writing- Original draft preparation, Analysis, Visualization: Ndiye Michael Kebonye: software, Data curation. Kingsley JOHN: Software, Editing, Visualization. Vahid Khosravi: Data curation, Editing and Investigation. Luboš Borůvka: Supervision, Editing. Radim Vašát: Data Curation and Visualization. Charles Mario Boateng Editing, Analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

The Czech University of Life Sciences Prague supported this research with an internal Ph.D. grant no. SV20-5-21130 from the Faculty of Agrobiology, Food, and Natural Resources (CZU). The Ministry of Education, Youth, and Sports of the Czech Republic (project No. CZ.02.1.01/0.0/0.0/16 019/0000845) also assisted. Finally, there is the Centre of Excellence (Centre of the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially hazardous substances of anthropogenic origin: a comprehensive assessment of the soil contamination risks for the quality of agricultural products, NutRisk Centre).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2022.116701.

References

Agyeman, P.C., Kebonye, N.M., John, K., Borůvka, L., Vašát, R., Fajemisim, O., 2022a. 2022a. Prediction of nickel concentration in peri-urban and urban soils using hybridized empirical bayesian kriging and support vector machine regression. Sci. Rep. 12 (1 12), 1–16. https://doi.org/10.1038/s41598-022-06843-y.

Agyeman, P.C., Khosravi, V., Michael Kebonye, N., John, K., Borůvka, L., Vašát, R., 2022b. Using spectral indices and terrain attribute datasets and their combination in the prediction of cadmium content in agricultural soil. Comput. Electron. Agric. 198, 107077 https://doi.org/10.1016/J.COMPAG.2022.107077.

Agyeman, P.C., Khosravi, V., Michael Kebonye, N., John, K., Borůvka, L., Vašát, R., 2022c. Using spectral indices and terrain attribute datasets and their combination in the prediction of cadmium content in agricultural soil. Comput. Electron. Agric. 198, 107077 https://doi.org/10.1016/J.COMPAG.2022.107077.

Ahmadi, A., Emami, M., Daccache, A., He, L., 2021. Soil properties prediction for precision agriculture using visible and near-infrared spectroscopy: a systematic review and meta-analysis. Agronomy 11, 433. https://doi.org/10.3390/AGRONOMY11030433. Page 433 11.

Basta, N.T., Ryan, J.A., Chaney, R.L., 2005. Trace element chemistry in residual-treated soil: key concepts and metal bioavailability. J. Environ. Qual. 34, 49–63. https://doi.org/10.2134/JEQ2005.0049DUP.

Bazoobandi, A., Emamgholizadeh, S., Ghorbani, H., 2019. Estimating the Amount of Cadmium and Lead in the Polluted Soil Using Artificial Intelligence Models, pp. 933–951. https://doi.org/10.1080/19648189.2019.1686429 26, 10.1080/19648189.2019.1686429.

Beard, K., Mackaness, W., 2006. Visual Access to Data Quality in Geographic Information Systems, pp. 37–45. https://doi.org/10.3138/C205-5885-23M7-0664 30, 10.3138/C205-5885-23M7-0664.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M., McBratney, A., 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC, Trends Anal. Chem. 29 (9), 1073–1081.

Ben-Dor, E., Banin, A., 1990. Near-infrared reflectance analysis of carbonate concentration in soils. Appl. Spectrosc. 44, 1064–1069. https://doi.org/10.1366/0003702904086821.

Ben-Dor, E., Inbar, Y., Environment, Y.C.-R.S. of, 1997, Undefined, 1997. The Reflectance Spectra of Organic Matter in the Visible Near-Infrared and Short Wave Infrared Region (400–2500 Nm) during a Controlled Decomposition Process. Elsevier.

Biney, J., Vašát, R., Bell, S., N.K.-S. and T., 2022a. Undefined, 2022. Prediction of Topsoil Organic Carbon Content with Sentinel-2 Imagery and Spectroscopic Measurements under Different Conditions Using an Ensemble Model (Soil Tillage Res).

Biney, J.K.M., Vašát, R., Blöcher, J.R., Borůvka, L., Němeček, K., 2022b. Using an ensemble model coupled with portable X-ray fluorescence and visible near-infrared spectroscopy to explore the viability of mapping and estimating arsenic in an agricultural soil. Sci. Total Environ. 818, 151805 https://doi.org/10.1016/J.SCITOTENV.2021.151805.

Cao, J., Li, C., Wu, Q., Qiao, J., 2020. Improved mapping of soil heavy metals using a vis-NIR spectroscopy index in an agricultural area of eastern China. IEEE Access 8, 42584–42594. https://doi.org/10.1109/ACCESS.2020.2976902.

Chen, T., Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785. August-2016, 785–794.

Choe, E., Kim, K.W., Bang, S., Yoon, I.H., Lee, K.Y., 2009. Qualitative analysis and mapping of heavy metals in an abandoned Au-Ag mine area using NIR spectroscopy. Environ. Geol. 58, 477–482. https://doi.org/10.1007/S00254-008-1520-9/FIGURES/4.

Cools, N., B.D.V., 2016. Sampling and analysis of soil.“ Manual on methods and criteria for harmonized sampling, assessment, monitoring and analysis of the effects of air pollution on forests [WWW Document] (accessed 4.5.22).

Das, A., Abdel-Aty, M., Pande, A., 2009. Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. J. Saf. Res. 40, 317–327. https://doi.org/10.1016/J.JSR.2009.05.003.

Delerce, S., Dorado, H., Grillon, A., Rebolledo, M.C., Prager, S.D., Patiño, V.H., Varón, G. G., Jiménez, D., 2016. Assessing weather-yield relationships in rice at local scale using data mining approaches. PLoS One 11. https://doi.org/10.1371/JOURNAL.PONE.0161620.

Devianti, D., Sufardi, S., Zulfahrizal, Z., Munawar, A.A., 2019. Rapid and simultaneous detection of hazardous heavy metals contamination in agricultural soil using infrared reflectance spectroscopy. IOP Conf. Ser. Mater. Sci. Eng. 506, 012008 https://doi.org/10.1088/1757-899X/506/1/012008.

Dor, E.B., Ong, C., Lau, I.C., 2015a. Reflectance measurements of soils in the laboratory: Standards and protocols. Geoderma 245, 112–124.

Ehsani, M.R., Upadhyaya, S.K., Slaughter, D., Shafii, S., Pelletier, M., 1999. A NIR technique for rapid determination of soil mineral nitrogen. Precis. Agric. 1 (2), 219–236.

Engel, J., Gerretzen, J., Szymańska, E., J.J.-T.T., 2013. Undefined, 2013. Breaking with Trends in Pre-processing? TrAC Trends in Analytical Chemistry.

Etemad-Shahidi, A., Mahjoobi, J., 2009. Comparison between M5′ model tree and neural networks for prediction of significant wave height in Lake Superior. Ocean Eng. 36, 1175–1181. https://doi.org/10.1016/J.OCEANENG.2009.08.008.

Gaffey, S.J., 1987. Spectral reflectance of carbonate minerals in the visible and near infrared (0.35–2.55 um): anhydrous carbonate minerals. J. Geophys. Res. 92, 1429. https://doi.org/10.1029/JB092IB02P01429.

Gamon, J.A., Penuelas, J., Field, C.B., 1992. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. Rem. Sens. Environ. 41 (1), 35–44.

Gholampour, A., iranica, A.J.-S., 2019. undefined, 2019. Reliability analysis of a vertical cut in unsaturated soil using sequential Gaussian simulation. In: scientiairanica.sharif.edu. https://doi.org/10.24200/sci.2017.4571.

Goffart, D., Dvorakova, K., Crucil, G., Curnel, Y., Limbourg, Q., Oost, K. van, Castaldi, F., Planchon, V., Goffart, J.-P., Wesemael, B. van, 2022. UAV remote sensing for detecting within-field spatial variation of winter wheat growth and links to soil properties and historical management practices. A case study on Belgian loamy soil. Rem. Sens. 14, 2806. https://doi.org/10.3390/RS14122806. Page 2806 14.

Gomez, M.R., Cerutti, S., Sombra, L.L., Silva, M.F., Martínez, L.D., 2007. Determination of heavy metals for the quality control in argentinian herbal medicines by ETAAS and ICP-OES. Food Chem. Toxicol. 45 (6), 1060–1064.

Goovaerts, P., 2001. Geostatistical Modelling of Uncertainty in Soil Science. Geoderma.

Goyal, M.K., 2014. Modeling of sediment yield prediction using M5 model tree algorithm and wavelet regression. Water Resour. Manag. 28, 1991–2003. https://doi.org/10.1007/S11269-014-0590-6/FIGURES/7.

Goydaragh, M.G., Taghizadeh-Mehrjardi, R., Jafarzadeh, A.A., Triantafilis, J., Lado, M., 2021. Using environmental variables and Fourier Transform Infrared Spectroscopy to predict soil organic carbon. Catena 202, 105280. https://doi.org/10.1016/J.CATENA.2021.105280.

Heuvelink, G.B.M., Webster, R., 2001. Modelling soil variation: past, present, and future. Geoderma 100 (3–4), 269–301.

Hong, Y., Liu, Yaolin, Chen, Y., Liu, Yanfang, Yu, L., Liu, Yi, Cheng, H., 2019a. Application of fractional-order derivative in the quantitative estimation of soil organic matter content through visible and near-infrared spectroscopy. Geoderma 337, 758–769. https://doi.org/10.1016/J.GEODERMA.2018.10.025.

Hong, Y., Shen, R., Cheng, H., Chen, S., Chen, Y., Guo, L., He, J., Liu, Yaolin, Yu, L., Liu, Yi, 2019b. Cadmium concentration estimation in peri-urban agricultural soils: using reflectance spectroscopy, soil auxiliary information, or a combination of both? Geoderma 354, 113875. https://doi.org/10.1016/J.GEODERMA.2019.07.033.

Hope, S., Hunter, G.J., 2013. Testing the Effects of Thematic Uncertainty on Spatial Decision-Making, pp. 199–214.

Horák, J., Janovský, M., Hejcman, M., Šmejda, L., Klír, T., 2018. Soil geochemistry of medieval arable fields in Lovětín near Třešť, Czech Republic. Catena 162, 14–22. https://doi.org/10.1016/J.CATENA.2017.11.014.

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. J. Comput. Graph. Stat. 15 (3), 651–674.

Huang, Y., Wang, L., Wang, W., Li, T., 2019. Undefined, 2019. Current Status of Agricultural Soil Pollution by Heavy Metals in China: A Meta-Analysis (Science of the Total Environment).

Hutengs, C., Seidel, M., Oertel, F., Ludwig, B., Vohland, M., 2019. In situ and laboratory soil spectroscopy with portable visible-to-near-infrared and mid-infrared instruments for the assessment of organic carbon in soils. Geoderma 355, 113900. https://doi.org/10.1016/J.GEODERMA.2019.113900.

Jiang, Q., Liu, M., Wang, J., Liu, F., 2018. Feasibility of using visible and near-infrared reflectance spectroscopy to monitor heavy metal contaminants in urban lake sediment. Catena 162, 72–79. https://doi.org/10.1016/J.CATENA.2017.11.020.

Johari, A., Khani, M., M.H.-S.D., 2020. Undefined, 2020. System Reliability Analysis for Seismic Site Classification Based on Sequential Gaussian Co-simulation: A Case Study in Shiraz, Iran (Soil Dynamics and Earthquake Engineering).

John, K., Isong, I.A., Kebonye, N.M., Ayito, E.O., Agyeman, P.C., Afu, S.M., 2020. Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil. Nutrient Indicators in an Alluvial Soil. Land 9, 487. https://doi.org/10.3390/LAND9120487. Page 487 9.

John, K., Agyeman, P.C., Kebonye, N.M., Isong, I.A., Ayito, E.O., Ofem, K.I., Qin, C.Z., 2021. Hybridization of cokriging and Gaussian process regression modelling techniques in mapping soil sulphur. Catena 206, 105534. https://doi.org/10.1016/J.CATENA.2021.105534.

Kapo, K.E., Holmes, C.M., Dyer, S.D., de Zwart, D., Posthuma, L., 2014. Developing a foundation for eco-epidemiological assessment of aquatic ecological status over large geographic regions utilizing existing data resources and models. Environ. Toxicol. Chem. 33, 1665–1677. https://doi.org/10.1002/ETC.2557.

Kebonye, N.M., John, K., Chakraborty, S., Agyeman, P.C., Ahado, S.K., Eze, P.N., Němeček, K., Drábek, O., Borůvka, L., 2021. Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy. Geoderma 384, 114792. https://doi.org/10.1016/J.GEODERMA.2020.114792.

Kebonye, N.M., Agyeman, P.C., Seletlo, Z., Eze, P.N., 2022. On exploring bivariate and trivariate maps as visualization tools for spatial associations in digital soil mapping: a focus on soil properties. Precis. Agric. 1–22. https://doi.org/10.1007/s11119-022-09955-7.

Khongnawang, T., Zare, E., Srihabun, P., Khunthong, I., Triantafilis, J., 2022. Digital soil mapping of soil salinity using EM38 and quasi-3d modelling software (EM4Soil). Soil Use Manag. 38 (1), 277–291. https://doi.org/10.1111/sum.12778.

Khorrami, R., Derakhshani, A., Moayedi, H., 2020. New explicit formulation for ultimate bearing capacity of shallow foundations on granular soil using M5' model tree. Measurement 163, 108032. https://doi.org/10.1016/J.MEASUREMENT.2020.108032.

Khosravi, V., Doulati Ardejani, F., Yousefi, S., Aryafar, A., 2018. Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods. Geoderma 318, 29–41. https://doi.org/10.1016/J.GEODERMA.2017.12.025.

Kooistra, L., Wehrens, R., Leuven, R.S.E.W., Buydens, L.M.C., 2001. Possibilities of visible–near-infrared spectroscopy for the assessment of soil contamination in river floodplains. Anal. Chim. Acta 446, 97–105. https://doi.org/10.1016/S0003-2670(01)01265-X.

Kooistra, L., Wanders, J., Epema, G., R.L.-A.C., 2003. Undefined, 2003. The Potential of Field Spectroscopy for the Assessment of Sediment Properties in River Floodplains (Elsevier).

Kozák, J., Němeček, J, Borůvka, L., Lérová, Z., Němeček, K, Kodešová, R, Zádorová, T., 2010. Atlas půd České republiky.[Soil Atlas of the Czech Republic. Czech University of Life Sciences, Prague, Prague, p. 150.

Li, H., Leng, W., Zhou, Y., Chen, F., Xiu, Z., Yang, D., 2014. Evaluation models for soil nutrient based on support vector machine and artificial neural networks. Sci. World J. 288.

Li, L., Lu, J., Wang, S., Ma, Y., Wei, Q., Li, X., Ren, T., 2016. Methods for estimating leaf nitrogen concentration of winter oilseed rape (Brassica napus L.) using in situ leaf spectroscopy. Ind. Crop. Prod. 91, 194–204.

Luce, M.S., Ziadi, N., Gagnon, B., Karam, A., 2017. Visible near infrared reflectance spectroscopy prediction of soil heavy metal concentrations in paper mill biosolid-and liming by-product-amended agricultural soils. Geoderma 288, 23–36.

Mao, Y., Liu, J., Cao, W., Ding, R., Fu, Y., Zhao, Z., 2021. Research on the quantitative inversion model of heavy metals in soda saline land based on visible-near-infrared spectroscopy. Infrared Phys. Technol. 112, 103602 https://doi.org/10.1016/J.INFRARED.2020.103602.

Nayak, P.S., Singh, B.K., 2007. Instrumental characterization of clay by XRF, XRD and FTIR. Bull. Mater. Sci. 30, 235–238. https://doi.org/10.1007/S12034-007-0042-5.

Nemecek, J., Podlesakova, E., 1992. RETROSPECTIVE EXPERIMENTAL MONITORING OF HEAVY-METALS. Google Scholar [WWW Document]. Rostlinna Vyroba. URL.

Nguyen, T.T., Pham, T.D., Nguyen, C.T., Delfos, J., Archibald, R., Dang, K.B., Hoang, N. B., Guo, W., Ngo, H.H., 2022. A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. Sci. Total Environ. 804, 150187 https://doi.org/10.1016/J.SCITOTENV.2021.150187.

Nicodemus, K.K., Malley, J.D., Strobl, C., Ziegler, A., 2010. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinf. 11, 1–13. https://doi.org/10.1186/1471-2105-11-110/FIGURES/6.

Nomngongo, P.N., Ngila, J.C., Kamau, J.N., Msagati, T.A., Moodley, B., 2013. Preconcentration of molybdenum, antimony and vanadium in gasolsine samples using Dowex 1-x8 resin and their determination with inductively coupled plasma–optical emission spectrometry. Talanta 110, 153–159.

Quinlan, J.R., 1992 November. Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence, 92, pp. 343–348.

Roth, R.E., 2013. The Impact of User Expertise on Geographic Risk Assessment under Uncertain Conditions, pp. 29–43.

Sattari, M.T., Mirabbasi, R., Sushab, R.S., Abraham, J., 2018. Prediction of groundwater level in ardebil plain using support vector regression and M5 tree model. Ground Water 56, 636–646. https://doi.org/10.1111/GWAT.12620.

Shi, T., Chen, Y., Liu, Y., Wu, G., 2014. Visible and near-infrared reflectance spectroscopy—an alternative for monitoring soil contamination by heavy metals. J. Hazard Mater. 265, 166–176. https://doi.org/10.1016/J.JHAZMAT.2013.11.059.

Sihag, P., Keshavarzi, A., Kumar, V., 2019. Comparison of different approaches for modeling of heavy metal estimations. SN Appl. Sci. 1, 1–11. https://doi.org/10.1007/S42452-019-0816-6/FIGURES/11.

Song, Y., Li, F., Yang, Z., Ayoko, G., Frost, R., Science, J.J.-A.C., 2012. Undefined, 2012. Diffuse Reflectance Spectroscopy for Monitoring Potentially Toxic Elements in the Agricultural Soils of Changjiang River Delta, China.

Speich, M.J.R., Bernhard, L., Teuling, A.J., Zappa, M., 2015. Application of bivariate mapping for hydrological classification and analysis of temporal change and scale effects in Switzerland. J. Hydrol. (Amst.) 523, 804–821. https://doi.org/10.1016/J.JHYDROL.2015.01.086 st. Luce.

Stenberg, B., Rossel, R., agronomy, A.M.-A., 2010. Undefined, 2010. Visible and Near Infrared Spectroscopy in Soil Science (Advances in agronomy).

Tighe, M., Lockwood, P., Wilson, S., Lisle, L., 2004. Comparison of digestion methods for ICP-OES analysis of a wide range of analytes in heavy metal contaminated soil samples with specific reference to arsenic and antimony. Commun. Soil Sci. Plant Anal. 35 (9–10), 1369–1385.

Trumbo, B.E., 1981. A theory for coloring bivariate statistical maps. Am. Statistician 35, 220–226. https://doi.org/10.1080/00031305.1981.10479360.

Tyner, J.A., 2010. Principles of Map Design [WWW Document]. [Sl]. URL. https://scholar.google.co.uk/scholar?hl=en&as_sdt=0%2C5&q=Tyner%2C+J.+A.+%282010%29.+Principles+of+map+design.+New+York%3A+Guilford+Press.&btnG=.

van Wesemael, B., Chartin, C., Wiesmeier, M., von Lützow, M., Hobley, E., Carnol, M., Krüger, I., Campion, M., Roisin, C., Hennart, S., Kögel-Knabner, I., 2019. An indicator for organic matter dynamics in temperate agricultural soils. Agric. Ecosyst. Environ. 274, 62–75. https://doi.org/10.1016/J.AGEE.2019.01.005.

Vacek, O., Vašát, R., Borůvka, L., 2020. Quantifying the pedodiversity-elevation relations. Geoderma 373, 114441.

Vapnik, V., 1995. The nature of statistical learning theory. Technometrics 38, 409. https://doi.org/10.2307/1271324.

Wang, Junjie, Cui, L., Gao, W., Shi, T., Chen, Y., Gao, Y., 2014. Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy. Geoderma 216, 1–9. https://doi.org/10.1016/J.GEODERMA.2013.10.024.

Wang, J., Cui, L., Gao, W., Shi, T., Chen, Y.G.-, 2014. Undefined, 2014. Prediction of Low Heavy Metal Concentrations in Agricultural Soils Using Visible and Near-Infrared Reflectance Spectroscopy.

Wang, Y., Witten, I.H., 1996. Induction of model trees for predicting continuous classes.

White, W.B., 1971. Infrared Characterization of Water and Hydroxyl Ion in the Basic Magnesium Carbonate Minerals (geoscienceworld.org).

Wilding, L.P., 1985. Spatial variability: its documentation, accomodation and implication to soil surveys. In: Soil Spatial Variability 166–194.

Wu, Y., Chen, J., Ji, J., Gong, P., Liao, Q., Tian, Q., Ma, H., 2007. A mechanism study of reflectance spectroscopy for investigating heavy metals in soils. Soil Sci. Soc. Am. J. 71, 918–926. https://doi.org/10.2136/SSSAJ2006.0285.

Wu, Y., Chen, J., Wu, X., Tian, Q., Ji, J., Qin, Z., 2005. Possibilities of reflectance spectroscopy for the assessment of contaminant elements in suburban soils. Appl. Geochem. 20 (6), 1051–1059.

Wuana, R.A., Okieimen, F.E., Montuelle, B., Steinman, A.D., 2011. Heavy metals in contaminated soils: a review of sources, chemistry, risks and best available strategies for remediation. downloads.hindawi.com 20. https://doi.org/10.5402/2011/402647.

Xu, X., Chen, S., Ren, L., Han, C., Lv, D., Zhang, Y., Ai, F., 2021. Estimation of heavy metals in agricultural soils using vis-NIR spectroscopy with fractional-order derivative and generalized regression neural network. Rem. Sens. 13, 2718. https://doi.org/10.3390/RS13142718. Page 2718 13.

Zeraatpisheh, M., Jafari, A., Bodaghabadi, M., Catena, S.A.-, 2020. Undefined, 2020. Conventional and Digital Soil Mapping in Iran: Past, Present, and Future (Elsevier).

Zhao, D., Wang, Junjie, Jiang, X., Zhen, J., Miao, J., Wang, Jingzhe, Wu, G., 2022. Reflectance spectroscopy for assessing heavy metal pollution indices in mangrove sediments using XGBoost method and physicochemical properties. Catena 211, 105967. https://doi.org/10.1016/J.CATENA.2021.105967.

Environmental Pollution 316 (2023) 120697

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

Prediction of the concentration of antimony in agricultural soil using data fusion, terrain attributes combined with regression kriging☆

Prince Chapman Agyeman [a,*], John Kingsley [a], Ndiye Michael Kebonye [b,c], Vahid Khosravi [a], Luboš Borůvka [a], Radim Vašát [a]

[a] Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500, Prague, Czech Republic
[b] Department of Geosciences, Chair of Soil Science and Geomorphology, University of Tübingen, Rümelinstr. 19-23, Tübingen, Germany
[c] DFG Cluster of Excellence "Machine Learning", University of Tübingen, AI Research Building, Maria-von-Linden-Str. 6, Tübingen, 72076, Germany

ARTICLE INFO

Keywords:
Regression kriging
Terrain attributes
Data fusion
Uncertainty
Agricultural soil

ABSTRACT

Potentially toxic elements in agricultural soils are primarily derived from anthropogenic and geogenic sources. This study aims to predict and map antimony (Sb) concentration in soil using multiple regression kriging in two distinct modeling approaches, namely Sb prediction using data fusion coupled with regression kriging (scenario 1) and Sb prediction using data fusion, terrain attributes, and regression kriging (scenario 2). Cubist regression kriging (cubist_RK), conditional inference forest regression kriging (CIF_RK), extreme gradient boosting regression kriging (EGB_RK) and random forest regression kriging (RF_RK) were the modeling techniques used in the estimation of Sb concentration in agricultural soil. The validation results suggested that in scenario 1, EGB_RK was the optimal modeling approach for Sb prediction in agricultural soil with root mean square error (RMSE) = 1.31 and mean absolute error (MAE) = 0.61, bias = 0.37, and high coefficient of determination R^2 = 0.81. Similarly, the EGB_RK was also the optimal modeling approach in scenario 2, with the highest R^2 = 0.76, RMSE = 0.90, bias = 0.06, and MAE = 0.48 values than the other regression kriging modeling approaches. The cumulative assessment suggested that the EGB_RK in scenario 2 yielded optimal results compared to the respective modeling approach in scenario 1. The uncertainty propagated by the modeling approaches in both scenarios indicated that the degree of uncertainty during the modeling process was distributed across the study area from a low to a moderate uncertainty level. However, cubist_RK in scenario 2 exhibited some elevated spots of uncertainty levels. As a result, the combination of data fusion, terrain attributes, and regression kriging modeling approaches produces optimal results with a high R^2 value, minimal errors as well as bias. Furthermore, combining terrain attributes with data fusion is promising for reducing model error, bias and yielding high-accuracy predictions.

1. Introduction

Potentially toxic elements (PTE) pollution is caused by both natural and anthropogenic activities. Some of the main anthropogenic sources are attributed to the steel industry, mining and smelting, refining, and processing of iron ore. As consequence, PTEs tend to be released into the soil, freshwater, and air where they eventually pose potential health concerns to residents in both urban and peri-urban areas (Agyeman et al., 2021a,b; Mohammadi et al., 2018; Saleh et al., 2019). Despite efforts to limit pollution and the proliferation of PTEs in soils and the environment, it has emerged as a much more significant challenge for the environment and public welfare in recent decades, notably with the advent of industries, urbanization, and agricultural production. According to, Babst-Kostecka et al. (2018), whenever PTEs are released into the ecosystem, they could remain there for decades or even generations, dispersing to remote places and accumulating in biotic and abiotic ecological processes. Agricultural practices such as continuous fertilizer application to nourish soil nutrients and for increasing yield have significantly changed the chemical composition of the natural contents of elements including Pb, Cu, Cd, As Ni, Cr, Sb and Zn in a

☆ This paper has been recommended for acceptance by Dr Hefa Cheng.

* Corresponding author.
E-mail address: agyeman@af.czu.cz (P.C. Agyeman).

https://doi.org/10.1016/j.envpol.2022.120697
Received 13 June 2022; Received in revised form 10 November 2022; Accepted 16 November 2022
Available online 17 November 2022

variety of agricultural soils (Nanos and Martín, 2012; Agyeman et al., 2021). Even for agricultural soils, PTEs are primarily derived from anthropogenic and geogenic sources. The buildup of PTEs in agricultural soil over time depletes the soil quality and impedes its proper functioning as well as soil biology activities, for instance, microbial activity (Beattie et al., 2018).

The constant presence of these elements in the environment causes devastating ramifications for human beings, animals, and plants. Long-term human exposure to PTEs has health implications, which in some cases prove fatal (Adimalla et al., 2020). Inhalation of certain PTE such as Sb causes detrimental effects including lung inflammation, chronic bronchitis, heart muscle damage, liver fibrosis, inactive tuberculosis, altered lung functioning and gastrointestinal disorders (Cao et al., 2010; Podsiki and Committee, 2008). Urban and peri-urban soils overburdened with high Sb levels provide opportunities for Sb to be absorbed into the body via diverse routes such as ingestion, inhalation, and dermal absorption (Bagherifam et al., 2019; Wang et al., 2018). According to Bolan et al. (2022), the United States Public Health Services (i.e., US-PHS, 1992) reported that exposure to Sb causes systemic, neurological, immunological, genotoxic, reproductive and developmental effects as well as cancer. Moreover, there is a handful of evidence supporting the detrimental effect of Sb pollutants and their level of carcinogenicity on humans (Nishad and Chemosphere, 2021). Sb trioxide ($Sb2O_3$) is a potential carcinogen to humans, according to the US National Toxicology Program (US-NTP), it was predicated on mutagenicity tests on mice models and findings from preclinical development studies (Program, 2018). Current studies, on the other hand, discuss the effects of multiple soil applications in the restoration of Sb-influenced locations on decreasing bioavailability and toxic effects of the pollutant (Rinklebe et al., 2020). Regardless, there is a scarcity of knowledge on the impact Sb has on humans to establish a link between the Sb-related oxides.

Understanding the spatial distribution of PTEs in the modern era has become a primary objective with the advent of digital soil mapping (DSM), which provides a platform for using various algorithms as well as incorporating measures such as terrain attributes, remotes sensing imageries or data fusion techniques to improve on model predictions or maps. According to, Minasny and McBratney (2016), the goal of DSM is to anticipate the spatial variability of PTEs using a variety of approaches, dependent on soil measurements and ecological factors. Taghizadeh-Mehrjardi et al. (2020), reported that the underpinning of DSM is the use of a mathematical model in conjunction with environmental variables to measure soil properties or PTEs. The use of DSM approaches in conjunction with environmental variables such as remote sensing imageries, data fusion and terrain attributes can help to elucidate the PTE content in various soils or sediments. Several research studies, including Jiang et al. (2019), Wang et al. (2018), and Wu et al. (2020), have employed environmental factors and DSM techniques to estimate PTEs contents in soil. The geomorphological features of land and terrain attributes are important in modeling processes because they are integral soil-forming processes and have an impact on soil distribution (McKenzie and Ryan, 1999; Zeraatpisheh et al., 2020). Terrain attributes are extracted from digital elevation models (DEMs) and used as an auxiliary variable in DSM modeling approaches. Due to their low cost and open access for monitoring, predicting, and mapping PTEs and soil properties, the use of remote sensing datasets such as Sentinel 2 and Landsat 8 imageries as auxiliary data for DSM has drawn increased interest in research. Remote sensing imageries provide unrivaled benefits for monitoring the earth at different scales and resolutions (Hu et al., 2019), and they play a significant role in worldwide and territorial soil or landscape surveillance and mapping (Ivushkin et al., 2019).

Digitally predicting PTE concentrations in the soil has been of high interest to researchers because of the threats associated with these elements on humans, plants and organisms in communities, habitats and agricultural fields. A variety of predictions for PTEs in sediments, soil, floodplains, and other environments have been generated using machine learning algorithms, geostatistical-based models, or a hybridization of the two. Some literature applied generalized regression neural networks and artificial neural networks in the prediction of Cu, Mn, and Ni, (Sergeev et al., 2019), regression kriging in the prediction of various PTEs (As, Cd, Cr, Cu, Hg, Pb, Zn, Sb, Co and Ni) (Tóth, Hermann, Szatmári, et al., 2016), ordinary kriging and cokriging to predict Zn, As, Cd, Cr, Cu, Ni, Pb, Hg. Zeng et al. (2021) and Cao et al. (2017) applied geostatistical models, comprising ordinary kriging (OK) and regression kriging for the prediction of Cd. Based on the growth and academic synergy in the field of soil science, it is now possible to combine various algorithms to complement and increase the predictive performances of algorithms through hybridizations and ensembling to obtain high-quality outcomes. Hybridization of algorithms from geostatistics and machine learning has been applied successfully to achieve superior results. Such an example where hybridization has yielded good results includes the study by John et al. (2021) which applies cokriging and Gaussian process regression for mapping sulfur levels in soils. Hybridization of algorithms allows for the complementation of models to optimize efficiency and minimize apparent errors. Regression kriging is a hybridized kriging approach that combines the kriging of predictions and residuals with either linear regression models or machine learning algorithms. For instance, Pouladi et al. (2019) applied cubist and random forest to ordinary kriging to generate hybridized regression kriging models, cubist regression kriging and random forest kriging. Regression kriging is a hybridized geostatistical model that somehow resembles universal kriging with an exterior trend. The residual variogram is computed first, and then simple kriging is applied to the residuals to obtain spatial prediction residuals, which is known as regression kriging (RK) (Bourennane et al., 2000; Hengl et al., 2003, 2007).

The current study explores the use of satellite imageries such as Landsat 8 and Sentinel 2 in a data fusion process, to harness the potential of using these composite data in conjunction with RK approaches to determine the content of Sb in agricultural soil. Secondly, this study aims to use the composite dataset (data fusion) in conjunction with terrain attributes, as well as regression kriging approaches, to estimate Sb for the same soil. Finally, exploits the potential of RK, which combines geospatial analysis and an algorithm that incorporates the impact of soil physical and chemical characteristics to accommodate specific variations between the application of satellite images dataset and terrain attributes. The specific goals of this study are to (1) apply data fusion coupled with regression kriging approaches to estimate Sb concentrations in agricultural soil (scenario 1); (2) add terrain attributes to data fusion datasets combined with regression kriging techniques to estimate Sb content in agricultural soil (scenario 2); (3) compare scenario 1 and scenario 2 (4), and map the uncertainties propagated by both scenarios.

2. Materials and methods

2.1. Study area

The study location is in the Frydek Mistek district of the Czech Republic (See Fig. 1). The study area is characterized by hilly terrain and uplands from the external Carpathians. The study area is distinguished by substantial crop production as well as multiple metal and steel industries, and it undulates within a latitude of 49° 41′ 0′ north and longitude of 18° 20′ 0′ east, at altitudes between 225 and 327 m above sea level (Agyeman et al., 2020). The region has a Cfb = oceanic humid climate with heavy precipitation (Koppen categorization) even throughout the dry months (John et al., 2021b). The Study area has a total landmass of 1208 km^2 (39.38 percent for agronomic activities and 49.36 percent for forest cover), and the landmass used for this research is 889.8 km^2. The color, carbonate composition, and structure of the soil are all conceivably discernible. Even so, the physical properties of the soil have a medium to smooth texture. They are most often found in colluvial and aeolian deposits, where they are distinguished by the top

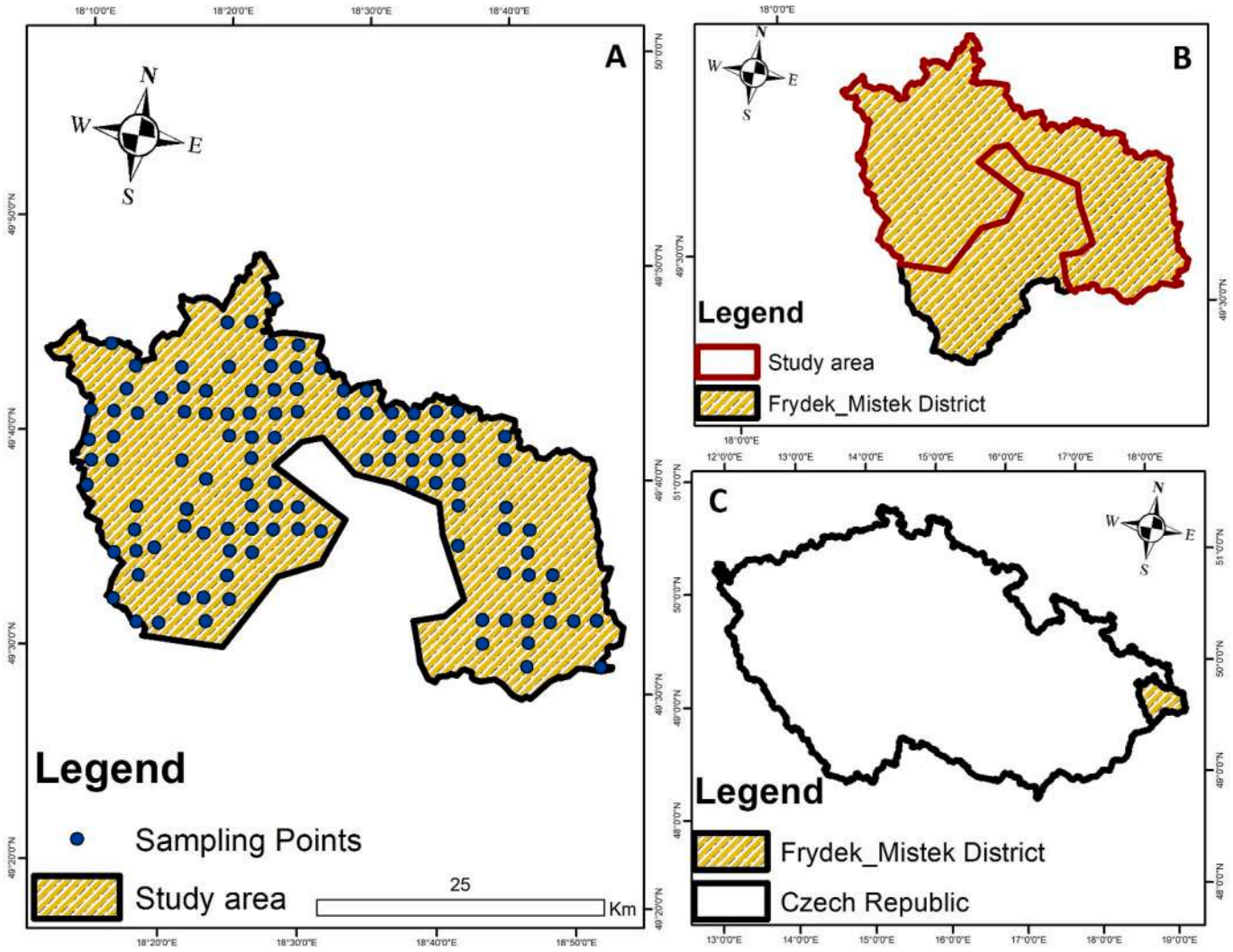


Fig. 1. Schematic showing the study area (A), District of Frydek Mistek (B) and Czech Republic (C).

surface and subsurface mottles that are visible in many soil areas and are usually accompanied by cementitious materials and bleaching. The soils are characterized by a cambic diagnostic horizon with a smooth sandy loam texture, a clay composition greater than 4%, and a lateritic disconnection with minimal calcareous composition (Kozák et al., 2010). Despite this, stagnosols and cambisols were the most common soil types in the study area (Kozák et al., 2010). These soil types are common in the Czech Republic, with altitudes ranging from 160.6 m to 455.1 m for stagnosol and 59.6–493.5 m for cambisols (Vacek et al., 2020).

2.2. Soil analysis and sampling

Topsoil samples (115 in total) were obtained from productive land in the Frydek Mistek district (Fig. 1). The sampling trend was a conventional grid, and the soil sampling distance was maintained at 2 × 2 km by employing a hand-held (GPS Leica-Zeno) gadget at depths spanning from 0 to 20 cm. Before transporting the sampled soil to the research laboratory, each sample was stored in plastic bags and pre-labeled. To obtain powdered soil samples, the soil samples were air-dried before being decimated with a mechanical device (Fritsch disk mill pulverize) and mesh sieved (2 mm). In a Teflon container, 1 g of dried, thoroughly mixed, and mesh-sieved soil sample (sieve size 2 mm) was placed and labeled. For each Teflon bottle, 7 ml of 35% HCl and 3 ml of 65% HNO_3 were discharged (employing fully automated dispensers—1 for every acid), and the lid was delicately sealed halfway to enable the sample to continue to stay overnight for reactions to occur (aqua regia procedure) (Cools, 2016). After dissolving the soil sample, the solution was placed on a hot plate for 2 h to speed up the digestion process before being allowed to cool. The supernatant was obtained by filtering the solution. The supernatant was transferred to a 50-ml (volumetric) flask and watered down with de-ionized water to 50 ml. The watered-down supernatant was then filtered into 50 ml PVC tubes. Besides that, 1 ml of the watered-down concentration was mixed with 9 ml of de-ionized water and filtrated into a pre-prepared 12 ml test tube to determine the pseudo total PTE concentration. Potential toxic element concentrations were determined following conventional guidelines employing the inductively coupled plasma–optical emission spectrometry (ICP–OES) (Thermo Fisher Scientific Corporation, USA). Moreover, the quality control and quality assurance procedures were guaranteed by checking the reference criteria for each study. The duplicate analysis was carried out to guarantee that errors were kept to a minimal level.

2.3. Geostatistics

Geostatistics is a statistical field that analyses and predicts the values related to spatial heterogeneity of physical processes. This includes the spatial and temporal coordinates of the datasets in the assessments. Numerous geostatistical techniques have been created as a pragmatic way to characterize spatial characteristics and linear interpolation

values for areas in which samples have not been collected. In this study, we employed ordinary kriging (OK) hybridizing it with machine learning algorithms for mapping and predicting Sb contents in soils. Ordinary kriging is an interpolation method that allows the user to quantify the spatial variability of soil properties at the investigational site (Agyeman, John, et al., 2021; Bishop and Geoderma, 2001). The equation is given as

$$Z^*(X_O) = \sum_{i-1}^{n} \lambda_i Z(X_I)$$

where $Z^*(X_O)$ represents the predicted value at the unquantified location (X_O), $Z(X_I)$ denotes the known or observed value at the location (X_I), λ_i is the coefficient weighting at the observed location (X_O) and n is the number of locations within the area under investigation.

Regression kriging (RK) is a type of interpolation approach where there is a combination of linear models of variables that are dependent and auxiliary variables, such as terrain attributes of variables in which the residuals are kriged alongside (Odeh et al., 1995). The RK approach will be used in the research to spatially interpolate the distribution of Sb in the following order:

- ➢ estimating the Sb prediction method approach by utilizing the regression technique in reciprocal directions,
- ➢ quantifying the Sb prediction modeling approach with residuals at every calibration position,
- ➢ modeling the covariance structure of the Sb residuals,
- ➢ spatially interpolating the Sb residuals using the variogram model parameters and
- ➢ obtaining the predicted map and combining the Sb prediction approach surface on the interpolated residual surface

2.4. Modeling using machine learning algorithms (MLAs)

The following machine learning algorithms (MLAs) were used in this study: extreme gradient boosting, random forest, conditional inference forest and cubist.

2.4.1. Conditional inference forest (CIF)

Conditional inference forest is a tree-developing model that is normally used in the application of bioinformatics (Nicodemus et al., 2010). CIF differs from the traditional random forest in theory because it distinguishes the selection of the splitting varying assortment of the splitting point of the already selected divided variable (Hothorn et al., 2006). The optimized divide variable is ascertained in the initial step and the associative test is run among the potential split parameters and the response. CIF approaches are used as a further tree-building technique with a conditional grid for the possible combination significance measure, allowing for superior assessment of each parameter's independent commitment and discrimination of observables from erroneous correlation (Delerce et al., 2016). If sampling is done without substitution and a test statistic, a quadratic version is utilized, and the CIF two-step methodology results in a non-biased split variable choice. The model is implemented in RStudio with the package "party".

2.4.2. Extreme gradient boosting (EGB)

EGB is classified as a form of decision tree algorithm that presents a boosted gradient method that enhances accuracy and speed (Climent et al., 2019). It is both a classification and regression algorithm. EGB was built on Friedman's previous gradient boosting method (Climent et al., 2019) which is a pragmatic and flexible function of Friedman's gradient boosting structure. The EGB modeling technique was applied in RStudio utilizing the R package "XGBoost".

2.4.3. Cubist

Quinlan (1992) developed the Cubist modeling approach as an augmentation of the M5 tree modeling method. The model's configuration is made up of preliminary factors that operate as various points in a decision tree and are merged with multivariate regression approaches. The trees are transformed into a collection of rules that are either exempted through clipping or integrated for easier assessment. The main benefit of the cubist approach is the addendum of multiple training committees and boosting to make the weights more affiliated (Quinlan, 1992; Kuhn et al., 2013). Cubist algorithms integrate boosting with training committees (generally more than one) and that is consistent with the approach of "boosting" by progressively constructing a sequence of trees mostly with modified weights. The cubist approach employs a large number of neighbors to alter the rule-based predictive model. The cubist modeling approach was applied in RStudio using the R package "cubist".

2.4.4. Random forest (RF)

A random forest (RF) model is a collective of multiple regression and classification trees. Breiman (2001) developed the technique and claimed that it outperformed adaptive boosting in terms of accuracy. According to Gislason et al. (2006) and Heung et al. (2014), RF's computational capacity is faster. The RF's parameter managing potential is categorical as well as uninterrupted. Due to its superior nature, RF does not require parameter preselection and can manage noise (Díaz-Uriarte & Alvarez de Andrés, 2006). The RF modeling approach was implemented in RStudio by applying the R package "randomForest".

2.4.5. Environmental covariates (EC)

The images of the Sentinel 2 satellite were obtained from a free satellite hub. Sentinel 2 was acquired by the European Space Agency Sentinel constellation wihin August 2020 (i.e., within the sampling period) (https://www.sentinel-hub.com/), and the bands were analyzed employing SNAP software. The Landsat 8-OLI satellite images were acquired (wihin August 2020) from the United States geological Earth-Explorer website. To obtain cloud-free Landsat 8-OLI and Sentinel-2 images of the study area, we selected images with very low cloud coverage and then mosaicked the most appropriate ones. Atmospheric correction was performed on the satellite imageries for sentinel 2 and Landsat 8.

Terrain attributes were derived using various sets of sourced terrain derivatives. The covariates were obtained from NASA EARTHDATA and processed with the SAGA-GIS terrain toolbox using a DEM with a spatial resolution of 30 m (https://earthexplorer.usgs.gov/). Moreover, the treated DEM procured at 30 m spatial resolution was rescaled to 10 m spatial resolution in ArcGIS utilizing the bilinear resampling method. Slope, elevation (DEM), LS-factor, CNBL (channel network base level), CND (channel network drainage), and RSP (relative slope position) are the terrain attributes used. The selected terrain attributes are largely attributable to the terrain's interaction with Sb.

2.5. Image fusion

Image fusion was accomplished by fusing multiple input images into a more informative single composite image. Fusion typically combines low to medium spatial hyper/multispectral images with a high spatial resolution panchromatic one to obtain an image preserving both spectral and spatial resolution of the hyper/multispectral and panchromatic images, respectively.

Depending on the fusion stage, image fusion is performed at three different levels (Pohl et al., 1998):

1 Pixel level,
2 Feature level,
3 Decision level

The pixel level is the lowest processing level of image composition. At the pixel level, the most popular and effective image fusion techniques are Hue, Intensity, Saturation (IHS), Gram Schmidt (GS),

Principal Component Analysis (PCA), and wavelet. GS approach was been used in this study because it has shown higher efficiency in the fusion of Sentinel 2 A and Landsat 8-OLI in a study conducted by Khosravi et al. (2022). It is based on an orthogonal vector algorithm, and all images are transformed to vector imagery maintaining equal dimensions of the pixels at a transformed high spatial resolution scale. The GS data fusion transformation process is thus carried out for the high spatial resolution bands (Laben, 2000).

This study employed Sentinel 2 A and Landsat 8-OLI bands. The 20 m spatial resolution Sentinel 2 A bands 11 and 12 were downscaled to 10 m using Gram-Schmidt (GS) approach to obtain a consistent spatial resolution with band 2, band 3, band 4 and band 8. Similarly, the Landsat 8-OLI bands 2 to 7 were equally resampled from 30 m spatial resolution to 10 m spatial resolution using the GS fusion approach. The Landsat 8-OLI bands 2 to 7 were fused to the 10 m Sentinel Bands using the GS fusion approach. These bands from Sentinel 2 and Landsat 8 were chosen because they possess the same spectral similarities.

2.6. Model assessment and approach

The set of data was divided into two parts: testing (25%) and training (75%), with the training data being used to generate the modeling regression and the testing data being employed to validate the performance of the designed models. All the modeling methods were exposed to a five-time replication of a ten-fold cross-validation procedure. The coefficient of determination (R^2), bias, root mean square error (RMSE), mean error and mean absolute error (MAE) and bias were calculated for the test dataset to assess the validity and precision of the DSM modeling techniques used in this study. The coefficient of determination, which represents the variability of the ratio in the response, is expressed by the regression model. The RMSE and degree of severity of the variations from within the independent quantification are used to classify the predictive model's performance, whereas MAE confirms the actual measurable value.

The modeling approach was done in two distinct approaches namely the application of data fusion of sentinel 2 and Landsat 8 coupled with regression kriging approaches to the estimation of Sb concentration in agricultural soil (scenario 1). The modeling approach allows the application of pixels extracted from the composite images' fusion (refer to supplementary Table ST1 for details of the bands used in the image fusion) from the 115 observation points in the study area as an auxiliary dataset coupled with modeling techniques to predict the concentration of Sb in agricultural soil (Scenario). In scenario 2 we employed the 115 pixels from the image fusion along with 115 samples each from terrain attributes (Slope, elevation or DEM, LS-factor, channel network base level, channel network drainage and relative slope position) combined with regression kriging techniques to estimate Sb content in agricultural soil (Scenario 2). Terrain attribute was chosen alongside data fusion of sentinel 2 and Landsat 8 because of the relationship it has with soils and also these features play a key role in the prediction of PTEs (Sb) in agricultural soil. Depending on the circumstances of pedogenesis and evolutionary development, environmental covariates have the strongest impact on the influential and effective categorization of the spatial variability of PTEs in soil (Zeraatpisheh et al., 2020). According to Ding et al. (2017), terrain attributes such as slope, and elevation have an impact on the distribution of PTEs in the soil. In any case, we are using data fusion of satellite images that are images of the Earth. Essentially environmental covariates such as slope are an important topographic factor that affects the migration and distribution of elements (Chu and Zhou, 2014).

3. Results and discussion

3.1. Data description and relative variable importance of environmental covariate to Sb

Table 1 presents the statistical description of Sb concentration in agricultural soil. Sb has a median, mean, maximum, and minimum value of 2.26 (mg/kg), 2.61 (mg/kg), 2.26 (mg/kg), and 9.72 (mg/kg), respectively. According to Wilding (1985), a coefficient of variation (CV) greater than 35% indicates high variability. The estimated CV of Sb is 41.30%, implying that the CV is high. This appears to suggest that the CV of Sb is homogeneous within agricultural soil and that the source of pollution may be anthropogenic. The estimated skewness of the Sb distribution in the agricultural soil suggests that the skewness value is greater than 1, which therefore suggests that the Sb distribution does not follow a normal distribution. Based on that, this study is inclined to use nonparametric regression models in the modeling of Sb in the agricultural soil, which means there is no need to log transform the dataset. Nevertheless, the estimated percentile distribution of Sb in the agricultural soil is 2.6 for the 25th, 50th, and 75th percentiles and 3.70 for the 90th percentile. Nakamaru et al. (2006) reported Sb concentrations in various agricultural soil groups with mean concentrations of 0.6 mg/kg (Andosol), 0.8 mg/kg (fluvisol), and 0.9 mg/kg (cambisol), which are 2.51–3.76 times lower than the current agricultural soil Sb concentrations. Zhong et al. (2020) also reported Sb concentrations in agricultural soil collected from different soil types: chestnut soil and red earth soil, with mean concentrations of 1.50 mg/kg and 2.73 mg/kg, respectively, in which the Sb concentration in the current study was found to be higher than the Sb concentration in the chestnut soil but lower than the Sb concentration in the red earth soil. The estimated crustal concentration of Sb reported by Cai et al. (2016) and Huang et al. (2012) lies between 0.2 and 0.3 mg/kg. However, according to the United States Environmental Protection Agency and the European Union Sb is a priority pollutant, that is analogous to arsenic (Cui et al., 2015). The median concentration of Sb in agricultural soil samples (0.30 mg/kg) in Italy and European agricultural soils (0.20 mg/kg) samples reported by Reimann et al. (2014) and Reimann et al. (2012) are lesser than the current Sb median concentration (2.26 mg/kg) in the study area. The agricultural soils in southern Poland reported Sb mean and median concentrations of 1.23 mg/kg and 0.95 mg/kg (Gruszecka-Kosowska et al., 2020) as compared to higher mean and median concentrations of 2.26 mg/kg and 2.61 mg/kg in the current study. Furthermore, in southwestern Poland, the mean and median concentration of Sb reported by Lewińska & Karczewska (2019) in the following towns Bardo (0.36 mg/kg, 0.18 mg/kg), Czarnów (1.3 mg/kg, 0.16 mg/kg), Głogów Cu smelter (2.05 mg/kg, 0.98 mg/kg) and Nowa Wieś Legnicka (1.55 mg/kg, 1.21 mg/kg) were found to be low compared to the current mean and median concentration in this study. Tóth et al. (2016a,b) detailed that the threshold concentration for Sb in agricultural soil in Europe is 2.00 mg/kg which is also lower than the current concentration of Sb in the agricultural soil of the current study area. Antimony, unlike lead,

Table 1
Statistical description of Sb.

Descriptive Statistics	Sb
Median (mg/kg)	2.26
Mean (mg/kg)	2.61
Minimum (mg/kg)	2.26
Maximum (mg/kg)	9.72
Standard deviation	1.08
Coefficient of variation	41.30
Skewness	4.22
Kurtosis	20.70
25th percentile	2.26
50th percentile	2.26
75th percentile	2.26
90th percentile	3.70

which has received more attention in the literature, is an emerging pollutant with metal and metalloid properties that are released into the environment either naturally or through anthropogenic sources. Sb is not widely discussed in terms of being a soil pollutant and a food toxic, as are Pb, Cd, Hg, and As.

The environmental covariates used in this study for Sb prediction in the cultivated soils are represented in Fig. 2. The corresponding importance of the relationship with Sb varied depending on the weights assigned to the environmental covariates used for Sb prediction in agricultural soil. Based on the relationship with the Sb, the relative importance of the environmental covariates obtained by RF, suggested that the most essential covariates decreases in this order B6 > LS. factor > slope > DEM > B3 > CNBL > CND > RSP > B4 > B2 > B5 > B1 with the corresponding percentile weights 35.34%, 27.90%, 13.23%, 3.93%, 4.06%, 4.00%, 3.24%, 2.82%, 2.44%, 1.93%, 1.78% and 1.56% respectively. Even though short-wave infrared (Band 6 - SWIR) was the most relevant covariate for Sb, within the most relevant 6 covariates, four of the terrain attribute covariates, including LS. factor, slope, DEM, as well as CNBL, were found to be within the most relevant covariates. This is true for the study area, which is characterized by relatively varying terrain (i.e. high and low terrain).

3.2. Prediction results using two different scenarios

Table 2 present the validation results obtained in the prediction of the concentration of Sb in agricultural soil using two distinct scenarios applied on varied auxiliary datasets. The regression kriging (RK) approaches RF_RK, Cubist_ RK, EGB_ RK, and CIF_ RK produced R^2 values of 0.67, 0.49, 0.81, and 0.42, correspondingly, in scenario 1. The EGB_RK ($R^2 = 0.81$) approach produced the best results, followed by the RF_RK approach (0.67). The regression kriging approaches cubist_ RK ($R^2 = 0.49$) and CIF_KK ($R^2 = 0.42$) produced abysmal results with R^2 values falling below 0.5, which according to Li et al., (2016) is unacceptable. The RMSE and MAE of the best-performing modeling approaches are listed in this order: EGB_RK, RF_RK, cubist_RK, and CIF_RK for RMSE, and EGB_RK, cubist _RK, RF_RK, and CIF_RK for MAE (refer to Table 2). In terms of estimated error (RMSE and MAE), EGB_RK had the lowest degree of error in the prediction of Sb in agricultural soil. The degree of bias in the prediction of Sb in agricultural soil based on the regression kriging approaches revealed that the modeling approach RF_RK had the least bias of 0.31, followed by CIF_RK with a bias of 0.33, EGB_RK with a bias of 0.37, and cubist_RK with a bias of 0.40. The overall performance of the regression kriging modeling approaches indicated that EGB_RK is the optimal modeling technique in the prediction of Sb in agricultural soil, with high prediction performance, low error margins, and appreciable bias.

In scenario 2, the regression kriging modeling approaches exhibited

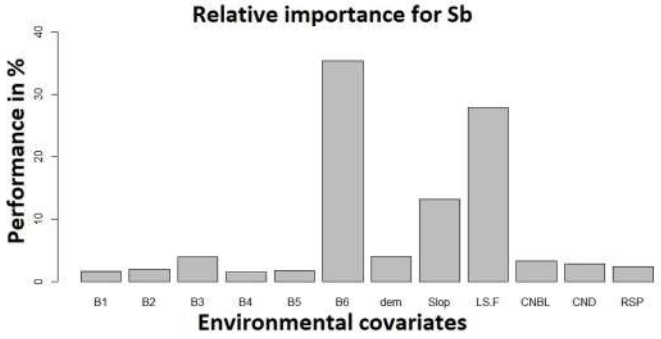


Fig. 2. Relative importance of environmental covariates to Sb (B1= Band 1, B2= Band 2, B3= Band 3, B4= Band 4, B5= Band 5, B6= Band 6, dem, slope, LS factor, CNBL = channel network base level, CND = channel network drainage, RSP = relative slope position).

Table 2
Depicts the performance of Sb prediction in agricultural soil employing data fusion as an ancillary dataset (scenario 1) and a combination of data fusion and terrain attributes as an ancillary dataset (scenario 2).

SCENARIO 1				
ALGORITHMS	R^2	RMSE	BIAS	MAE
RF_RK	0.67	1.40	0.31	0.70
CUBIST_RK	0.49	1.49	0.4	0.67
EGB_RK	0.81	1.31	0.37	0.61
CIF_RK	0.42	1.52	0.33	0.74
SCENARIO 2				
ALGORITHMS	R^2	RMSE	BIAS	MAE
RF_RK	0.48	1.15	0.05	0.6
CUBIST_RK	0.73	1.07	0.11	0.59
EGB_RK	0.76	0.90	0.06	0.48
CIF_RK	0.51	1.17	0.08	0.62

{EGB_RK (extreme gradient boosting-regression kriging), RF_RK (random forest_ regression kriging)m CUB_RK(cubist_ regression kriging) and CIF_RK(conditional inference forest_ regression kriging)}.

good performance, with the modeling approaches' R^2 values being 0.76, 0.73, 0.51, and 0.48 for EGB_RK, cubist_RK, CIF_RK, and RF_RK, respectively. The error values of the modeling approaches based on the least error obtained were 0.9 (EGB_RK), 1.07 (cubist_RK), 1.15 (CIF_RK) and 1.17 (RF_RK) for RMSE and the MAE of 0.48 (EGB_RK), 0.59 (cubist_RK), 0.60 (CIF_RK) and 0.62 (RF_RK). The prediction bias for the modeling approaches suggests that a 0.05 level of bias for the RF_RK modeling approach is the least prediction bias among the modeling approaches. The other modeling approaches such as EGB_RK, CIF_RK, and cubist_RK accrued degrees of biases of 0.06, 0.08, and 0.11 respectively. The cumulative prediction accuracy of the modeling techniques in predicting Sb concentration in agricultural soil revealed that the EGB_RK modeling approach is the best modeling method capable of predicting Sb concentration in agricultural soil with better efficiency, a lower error margin, and a satisfactory degree of bias.

When comparing scenarios 1 and 2, the obtained R^2 values of the RF_RK and EGB_RK in scenario 1 decreased by 16.04% for the RF_RK and 2.72% for the EGB_RK in scenario 2. Alternatively, the R^2 values for the cubist_RK and the CIF_RK increased by a margin of 19.79% for the cubist_RK and 8.75% for the CIF_RK in scenario 2 compared to scenario 1. Juxtaposing the estimated errors (RMSE and MAE) of the modeling approaches in both scenarios, it was evident that the error margin in scenario 2 decreased significantly by the marginal range of 9.97%–18.79% for RMSE and MAE, 6.23%–11.74% than in scenario 1. Conversely, the prediction biases of the modeling approaches in both scenarios also revealed that the biases accrued by the modeling approaches in scenario 2 decreased considerably by a marginal percentage range of 57.39%–72.04%. The cumulative assessment of the scenarios revealed that the three modeling approaches, EGB_RK, CIF_RK, and cubist_RK, significantly improved in scenario 2 compared to scenario 1. However, the overall modeling efficiency of the modeling techniques in predicting Sb in agricultural soil indicated that the EGB_RK in the scenario 2 modeling approach is the best modeling method capable of predicting the concentration of Sb in agricultural soil with higher efficiency, minimal error margin, and a satisfactory degree of bias.

The majority of the environmental factors used to build the connection with Sb were gathered from the ground and remote sensing satellites. The uncertainty of the environment, the spatial variation of Sb, and the predictive capabilities of modeling techniques all had a significant impact on the reliability of the prediction. The combination of data fusion and terrain attributes coupled with EGB_RK has proven to be effective in the prediction of PTE in agricultural soils with minimal error, acceptable bias, and a high coefficient of determination. It has been reported by Hengl et al. (2004); Umali et al. (2012) and Zhang et al. (2012) that the application of RK that introduces spatial interpolation into learning algorithms exhibits better spatial interpolation results in the prediction of soil properties and PTEs. The spatial interpolation

aspect of ordinary kriging when hybridized with an appropriate modeling algorithm yields good results. According to Pham et al. (2019a,b), OK has the tendency of yielding good results when applied in the prediction of PTEs and soil properties. The combination of data fusion datasets and terrain attributes, as well as the synergy established between covariates and modeling approaches such as EGB, has proven to produce acceptable results. Where there is a potent correlation between predicted PTE and environmental covariates, RK has consistently proven to be more precise (Keskin and Geoderma, 2018). In this study, it is evident that the relationship between the terrain attributes and Sb distribution has played a major role in the enhancement of the predicted results in scenario 2, especially in CIF_RK and cubist_RK. The low bias, RMSE and MAE in the data fusion-terrain attributes combination coupled with the modeling algorithms for the 3 RK models (EGB_RK, CIF_RK and cubist_RK) indicates that the terrain attributes combination with data fusion dataset information has served its purpose. It is imperative to highlight that the selection of environmental covariates that are ecologically consistent and correlate with the response variable with a robust autocorrelation with data makes RK more appropriate. More auxiliary datasets could be chosen to enhance the RK method's accuracy (Pham et al., 2019).

EGB possesses a substantial benefit above many MLAs in terms of selecting efficient attributes via a significance ranking system and limiting method prediction error by specifying the preset perspective of partitioning for omitted datasets or values (Ma et al., 2019). The findings suggest that by integrating major characteristics in evaluating metal ion prediction concentration levels, forest vegetation biomass, and PTE levels, EGB can improve prediction accuracy (Joharestani et al., 2019). In the estimation of PTEs in soils, EGB significantly outperforms other famous MLA prediction models including RF, ANN and SVM (Bhagat

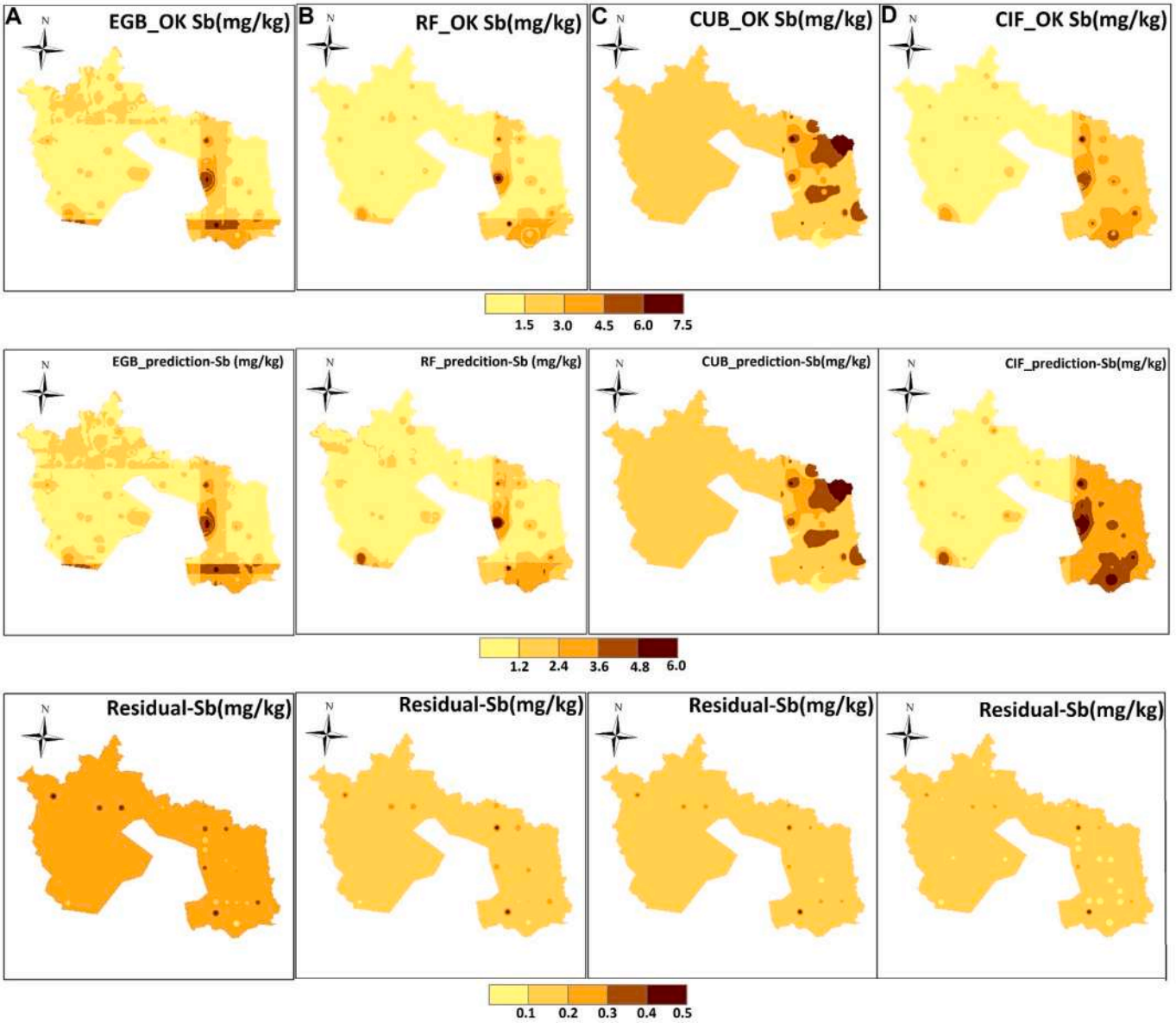


Fig. 3. Sb concentration maps based on residuals, algorithm predictions, and MLA and OK hybridization [using EGB_RK (A), RF_RK (B), CUB_RK (C), and CIF_RK (D)] (Scenario 1-using data of fusion Sentinel 2 and Landsat 8 as the predictors). Based on the residuals, algorithm, and combination of kriging and modeling approaches, the maps for each RK are arranged in columns for all the models, showing the prediction for the regression kriging models, the machine learning models, and the residuals. The legend beneath every row of the maps represents the precise scale for the concentration of Sb in the soil per model. {EGB_RK (extreme gradient boosting-regression kriging), RF_RK (random forest_ regression kriging)m CUB_RK (cubist_ regression kriging) and CIF_RK (conditional inference forest_ regression kriging)}.

et al., 2021). EGB, on the other hand, has the benefit of minimizing both overestimation as well as underestimation (Li et al., 2020). Based on Kim et al. (2015), EGB tends to screen out the modeling techniques efficiency by reducing the potential limitations that other modeling strategies have, including computational complexity. Moreover, EGB can help with modeling standardization issues (Jia et al., 2019), hyper-parameter tuning (Probst et al., 2019), local minima (Kawaguchi, 2019), elevated discrepancies (Li et al., 2020), and technology transfer (Kim et al., 2020), the demand for hyperparameter tuning (Probst et al., 2019).

3.3. Spatial prediction of Sb using regressing kriging approaches

The spatial distribution maps using both scenarios are presented in Figs. 3 and 4 for the modeling approaches RF_RK, cubist_RK, CIF_RK and EGB_RK. The spatial distribution maps for the modeling approaches in scenario 1 exhibited predominantly low to moderate spatial distribution of the mapped MLAs outputs and the hybridized modeling approaches RK with hotspots displayed in the southeastern area for the EGB_ (RK and prediction), cubist_ (RK and prediction) and CIF_RK spatial distribution map. However, the residual maps for the modeling approaches displayed low spatial distribution across the entire study area with exception of the residual for the EGB that displayed moderate spatial prediction. On the other hand, the spatial distribution maps for the modeling approaches in scenario 2 for the MLAs and the hybridized modeling approach displayed low to moderate spatial distribution for the modeling approaches. The cubist and the CIF modeling approach spatial distribution map showed hotspots in the northeastern and southeastern regions of the study area. The residuals map showed low distribution except for the EGB map which displayed moderate spatial distribution. The spatial distribution for the EGB_RK and the EGB_prediction map shared that same spatial distribution pattern. By comparing the RK maps for scenarios 1 and 2, the presence of terrain attributes

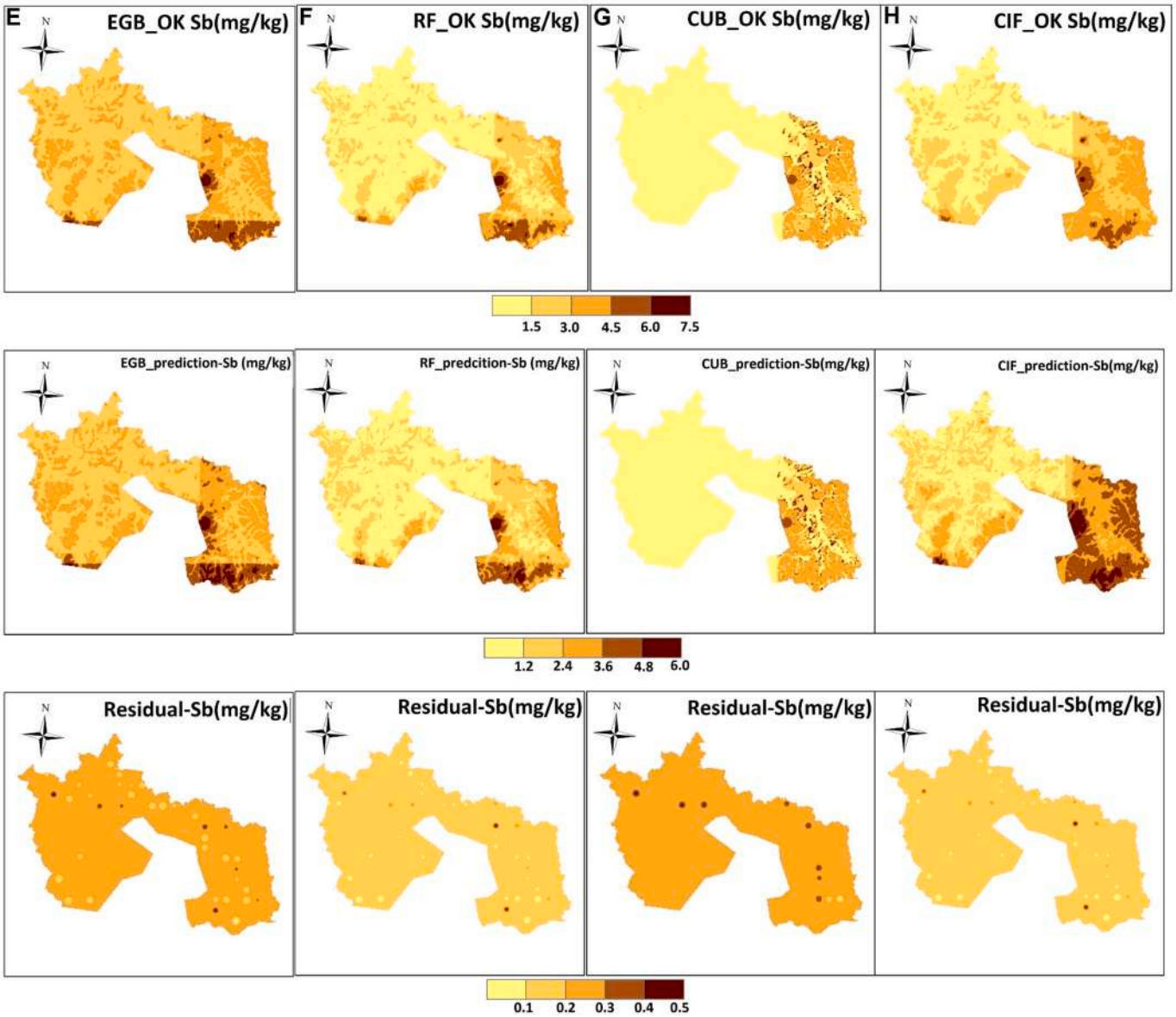


Fig. 4. Sb concentration maps based on residuals, algorithm predictions, and MLA and OK hybridization [using EGB_RK (E), RF_RK (F), CUB_RK (G), and CIF_RK (H)] (Scenario 1-using data fusion of sentinel 2 and Landsat 8 along with terrain attributes as the predictors). Based on the residuals, algorithm, and combination of kriging and modeling approaches, the maps for each RK are arranged in columns for all the models, showing the prediction for the regression kriging models, the machine learning models, and the residuals. The legend beneath every row of the maps represents the precise scale for the concentration of Sb in the soil per model. {EGB_RK (extreme gradient boosting-regression kriging), RF_RK (random forest_ regression kriging)m CUB_RK (cubist_ regression kriging) and CIF_RK (conditional inference forest_ regression kriging)}.

combined with the data fusion dataset coupled with the modeling approaches highlighted levels of Sb concentration that could not be shown when using data fusion alone as an auxiliary dataset for predicting Sb concentration in agricultural soil. The RK maps for the modeling approach show dissimilar spatial distribution of Sb in the study area. The combination of data fusion to terrain attributes used in the RKs spatial distribution maps has exhibited spatial differences in the maps which may be attributed to the terrain attributes that were added to the data fusion as predictor variables in the prediction of the concentration of Sb in the agricultural soil. Even though the application of auxiliary datasets to modeling approaches has the tendency of improving prediction accuracy and the maps produced with RKs display a comprehensive pattern for the Sb concentration in the agricultural soil. The RKs map for the modeling approaches in scenario 1 displays an erratic pattern while the RKs maps in scenario 2 exhibit consistent and natural spatial distribution patterns due to the introduction of terrain attributes as ancillary data in the mapping of Sb in the agricultural soil.

3.4. Uncertainty assessment based on scenarios 1 and 2

Model uncertainty can emerge from ignoring pertinent procedures like the linear nonlinear sorption process, using insufficient process characterizations like stable or state of balance, or specifying erroneous boundary constraints (Keller et al., 2002). The assessment of uncertainty propagation generates information-based and natural uncertainties that are dependent on the model's reliability, apart from variable calibration. Nonetheless, the prediction and uncertainty maps of the RK models for both scenarios are presented in Figs. 5 and 6, specifying the levels of uncertainty depicted by 2.5% and 97.5% prediction intervals while the mean prediction is denoted as such. In summary, these are presented in columns A to D and E to H for each modeling approach. Presented in Fig. 5 from columns A to D are CIF_RK (A), cubist_RK (B), EGB_RK (C) and RF_RK (D) for scenario 1 while in Fig. 6 the columns are such that there is CIF_RK (E), cubist_RK(F), EGB_RK(G) and RF_RK(H) for scenario 2. The evaluation of the prediction maps using the mean error (ME), mean absolute error (MAE) and root mean square error (RMSE) is presented in Table ST2 (Supplementary Table 2).

In scenario 1, the RKs approaches exhibited diverse levels of uncertainty propagation across the entire study area. However, the lower limit (2.5%) and the mean prediction propagation exhibited by the modeling approaches were generally low with patches of moderate degree of Sb seen in the south-eastern area for the cubist_RK modeling approach. On the other hand, the upper limit (97.5%) of uncertainty propagation by the modeling approaches is largely low for all the modeling approaches, with the south-eastern areas of the map for cubist_RK recording moderate degree of uncertainty propagation. The Sb mean estimate ME levels were 0.001 for EGB_RK, 0.004 for cubist_RK, 0.005 for CIF_RK and 0.003 for RF_RK. The estimated MAE and RMSE for the prediction maps in scenario 1 are 0.288 (EGB_RK), 0.306 (cubist_RK), 0.525 (CIF_RK) and 0.386 (RF_RK) for MAE and 0.579 (EGB_RK), 0.522 (cubist_RK), 0.971 (CIF_RK) and 0.723 (RF_RK). The error distribution estimated for the maps of each modeling approach based on RSME, ME, and MAE suggested that the degree of error propagated in the mean predictions is within 10%. The cumulative assessment of the error obtained by each modeling approach suggested that EGB_RK accrued the least error in the mapping output.

In scenario 2, the cubist_RK displayed moderate and spots of relatively high uncertainty propagation level in the north-eastern and the south-eastern enclaves of the map for their lower limit, upper limit and mean prediction maps. The estimated error margins of the mean prediction maps for each model based on ME ranged between 0.001 and 0.008, with EGB_RK obtaining the lowest (0.001) and the CIF_OK obtaining the highest (0.008) (See table ST2). The MAE and the RMSE estimated ranged between 0.260 and 0.464 for MAE and 0.513 to 0.903 for RMSE. EGB_RK obtained the least MAE (0.260) and RMSE (0.464) values whereas CIF_RK obtained the highest estimated MAE (0.513) and RMSE (0.903) values (refer to table ST2).

The cumulative assessment of the degree of uncertainty propagated by modeling approaches in the study area suggested that EGB_RK accrued the least error in the uncertainty maps based on the prediction

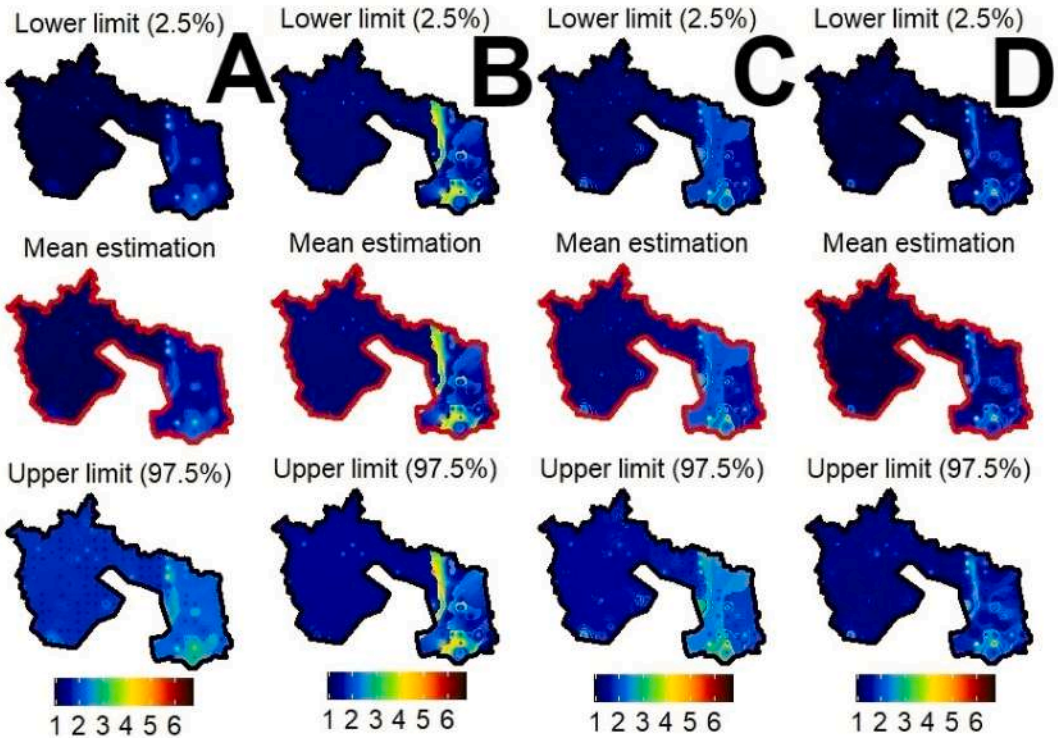


Fig. 5. Uncertainty propagation levels for Sb distribution in the study area based on the lower limit (2.5%), the upper limit (97.5%) and mean estimation using CIF_RK (A), cubist_RK (B), EGB_RK (C) and RF_RK (D) modeling approaches (Scenario 1).

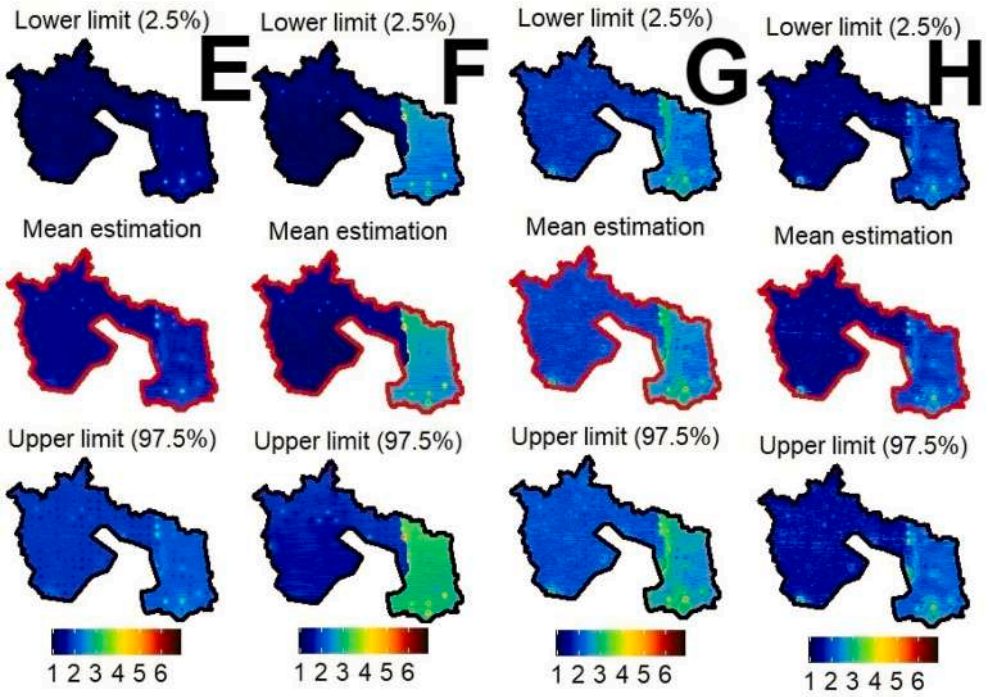


Fig. 6. Uncertainty propagation levels for Sb distribution in the study area based on the lower limit (2.5%), the upper limit (97.5%) and mean estimation using CIF_RK (E), cubist_RK (F), EGB_RK (G) and RF_RK (H) modeling approaches (Scenario 2).

intervals (2.5% and 97.5%). More so, the application of EGB_RK, data fusion of sentinel 2 and Landsat 8 along with terrain attributes accrued the least error when comparing all maps from both scenarios. Comparatively moderate Sb levels were propagated in the cubist_RK map in scenario 2 than in scenario 1. The areas that displayed the moderate level of uncertainty were the areas that exhibited the high level of Sb due to the steel factory and the metals works within that region. Fathololoumi et al. (2020) used data fusion and MLAs to predict soil properties, and the authors concluded that uncertainty propagation using the cubist model was high, which is somewhat consistent with the level of uncertainty propagated by cubist_RK in this study. The integration of terrain attributes and data fusion not only had a stronger impact on Sb prediction, particularly in cubist_RK, EGB_RK, and CIF_RK but also reduced prediction uncertainty in RF_RK and CIF_RK. Uncertainty quantification is required for spatial prediction and can corroborate the applicability of maps for managerial decision-making procedures (Fathololoumi et al., 2020). It is important to note, however, that the best model does not always result in the lowest uncertainty in the final map (Zeraatpisheh et al., 2022). The use of auxiliary datasets like data fusion and terrain attributes has proven to be reliable and capable of producing good results, reducing error margins in uncertainty mapping. Combining auxiliary datasets such as MCC and RST covariates can significantly reduce spatial prediction uncertainty (Zeraatpisheh et al., 2022). This is consistent with the results obtained in the current study.

4. Conclusion

The study assesses multiple regression kriging models for the prediction of Sb in the soil based on two scenarios, namely the prediction of Sb concentration in agricultural soil using data fusion and regression kriging approaches (scenario 1) and the prediction of the concentration of Sb in the soil using terrain attributes, data fusion and regression kriging approaches (scenario 2). The results revealed in scenario 1 that EGB_RK was the optimum modeling approach that predicted the concentration of Sb in the agricultural soil with minimal error, bias, and high R^2 value. In scenario 2, the results showed that EGB_RK was the optimal approach for the prediction of Sb concentrations in soil, with a high R^2 value and low RMSE and MAE values. Nevertheless, the cumulative performance of the models in both scenarios suggested that EGB_RK in scenario 2 performed better than the EGB_RK in scenario 1 based on the minimal error values as well as low bias compared to error and the bias values obtained in scenario 1. The uncertainty propagated by both scenarios was largely distributed from low to moderate uncertainty levels, but Cubist_RK exhibited spots of high uncertainty levels in the southeastern region of the map in scenario 2. The use of regression kriging approaches and auxiliary datasets such as data fusion from sentinel 2 and Landsat 8 datasets normally produces good results; however, combining it with terrain attributes reduces errors and bias and produces better results. This study recommends that the application of data fusion, terrain attributes coupled with an appropriate regression kriging approach can produce promising results while reducing marginal errors and bias.

Credit author statement

Prince Chapman Agyeman: Conceptualization, Methodology, Writing- Original draft preparation, Analysis, Visualization: Vahid Khosravi: Data curation, Editing and Investigation. Ndiye Michael Kebonye: software, Data curation. Kingsley JOHN: Software, Editing, Visualization. Luboš Borůvka: Supervision, Editing. Radim Vašát: Data Curation and Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

The Czech University of Life Sciences Prague supported this research with an internal Ph.D. grant no. SV20-5-21130 from the Faculty of Agrobiology, Food, and Natural Resources (CZU). The Ministry of Education, Youth, and Sports of the Czech Republic (project No. CZ.02.1.01/0.0/0.0/16 019/0000845) also assisted. Finally, there is the Centre of Excellence (Centre of the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially hazardous substances of anthropogenic origin: a comprehensive assessment of the soil contamination risks for the quality of agricultural products, NutRisk Centre).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2022.120697.

References

Adimalla, N., Chen, J., Qian, H., 2020. Spatial characteristics of heavy metal contamination and potential human health risk assessment of urban soils: a case study from an urban region of South India. Ecotoxicol. Environ. Saf. 194, 110406 https://doi.org/10.1016/J.ECOENV.2020.110406.

Agyeman, P.C., Ahado, S.K., Kingsley, J., Kebonye, N.M., Biney, J.K.M., Borůvka, L., Vasat, R., Kocarek, M., 2020. Source apportionment, contamination levels, and spatial prediction of potentially toxic elements in selected soils of the Czech Republic. Environ. Geochem. Health. https://doi.org/10.1007/s10653-020-00743-8.

Agyeman, P.C., Ahado, S.K., John, K., Kebonye, N.M., Vašát, R., Borůvka, L., Kočárek, M., Němeček, K., 2021a. Health risk assessment and the application of CF-PMF: a pollution assessment–based receptor model in an urban soil. J. Soils Sediments 21 (9), 3117–3136. https://doi.org/10.1007/s11368-021-02988-x.

Agyeman, P.C., John, K., Kebonye, N.M., Borůvka, L., Vašát, R., Drábek, O., 2021b. A geostatistical approach to estimating source apportionment in urban and peri-urban soils using the Czech Republic as an example. Sci. Rep. 11 (1), 1–15. https://doi.org/10.1038/s41598-021-02968-8.

Babst-Kostecka, A., Schat, H., Saumitou-Laprade, P., Grodzińska, K., Bourceaux, A., Pauwels, M., Frérot, H., 2018. Evolutionary dynamics of quantitative variation in an adaptive trait at the regional scale: the case of zinc hyperaccumulation in Arabidopsis halleri. Mol. Ecol. 27 (16), 3257–3273. https://doi.org/10.1111/mec.14800.

Bagherifam, S., Brown, T., 2019. Derivation Methods of Soils, Water and Sediments Toxicity Guidelines: a Brief Review with a Focus on Antimony. C. F.-J. of G.. Elsevier https://www.sciencedirect.com/science/article/pii/S0375674219300585.

Beattie, R., Henke, W., Campa, M., S B, T.H., 2018. Variation in Microbial Community Structure Correlates with Heavy-Metal Contamination in Soils Decades after Mining Ceased. Elsevier. https://www.sciencedirect.com/science/article/pii/S0038071718302645.

Bhagat, S., Tung, T., Materials, Z. Y.-J. of H., 2021. Heavy Metal Contamination Prediction Using Ensemble Model: Case Study of Bay Sedimentation, Australia. Elsevier. https://www.sciencedirect.com/science/article/pii/S0304389420314783.

Bishop, T., Geoderma, A.M.-, 2001. A Comparison of Prediction Methods for the Creation of Field-Extent Soil Property Maps. Elsevier. https://www.sciencedirect.com/science/article/pii/S001670610100074X.

Bolan, N., Kumar, M., Singh, E., Kumar, A., Singh, L., Kumar, S., Keerthanan, S., Hoang, S.A., El-Naggar, A., Vithanage, M., Sarkar, B., Wijesekara, H., Diyabalanage, S., Sooriyakumar, P., Vinu, A., Wang, H., Kirkham, M.B., Shaheen, S. M., Rinklebe, J., Siddique, K.H.M., 2022. Antimony contamination and its risk management in complex environmental settings: a review. Environ. Int. 158, 106908 https://doi.org/10.1016/J.ENVINT.2021.106908.

Bourennane, H., King, D., Couturier, A., 2000. Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. Geoderma 97 (3–4), 255–271. https://doi.org/10.1016/S0016-7061(00)00042-2.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Cai, F., Ren, J., Tao, S., Wang, X., 2016. Uptake, translocation and transformation of antimony in rice (Oryza sativa L.) seedlings. Environ. Pollut. 209, 169–176. https://doi.org/10.1016/j.envpol.2015.11.033.

Cao, H., Chen, J., Zhang, J., Zhang, H., Qiao, L., Men, Y., 2010. Heavy metals in rice and garden vegetables and their potential health risks to inhabitants in the vicinity of an industrial zone in Jiangsu, China. J. Environ. Sci. 22 (11), 1792–1799. https://doi.org/10.1016/S1001-0742(09)60321-1.

Cao, S., Lu, A., Wang, J., Huo, L., 2017. Modeling and mapping of cadmium in soils based on qualitative and quantitative auxiliary variables in a cadmium contaminated area. Sci. Total Environ. 580, 430–439. https://doi.org/10.1016/J.SCITOTENV.2016.10.088.

Chu, J., Zhou, J.F., 2014. Distribution and pollution of soil heavy metals in hilly upland around Pingdingshan coal mining area. Geogr. Res. 33 (7), 1383–1392.

Climent, F., Momparler, A., Carmona, P., 2019. Anticipating bank distress in the Eurozone: an extreme gradient boosting approach. J. Bus. Res. 101, 885–896. https://doi.org/10.1016/J.JBUSRES.2018.11.015.

Cools, N., V, B.D., 2016. Sampling and Analysis of soil." Manual on Methods and Criteria for Harmonized Sampling, Assessment, Monitoring and Analysis of the Effects of Air Pollution on Forests. Hünen Institute of Forest Ecosystems, Eberswalde.

Cui, X.D., Wang, Y.J., Hockmann, K., Zhou, D.M., 2015. Effect of iron plaque on antimony uptake by rice (Oryza sativa L.). Environ. Pollut. 204, 133–140. https://doi.org/10.1016/j.envpol.2015.04.019.

Delerce, S., Dorado, H., Grillon, A., Rebolledo, M.C., Prager, S.D., Patiño, V.H., Varón, G. G., Jiménez, D., 2016. Assessing weather-yield relationships in rice at local scale using data mining approaches. PLoS One 11 (8). https://doi.org/10.1371/JOURNAL.PONE.0161620.

Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinf. 7 https://doi.org/10.1186/1471-2105-7-3.

Ding, Q., Cheng, G., Wang, Y., Zhuang, D., 2017. Effects of natural factors on the spatial distribution of heavy metals in soils surrounding mining regions. Sci. Total Environ. 578, 577–585.

Fathololoumi, S., Vaezi, A.R., Alavipanah, S.K., Ghorbani, A., Saurette, D., Biswas, A., 2020. Improved digital soil mapping with multitemporal remotely sensed satellite data fusion: a case study in Iran. Sci. Total Environ. 721, 137703 https://doi.org/10.1016/J.SCITOTENV.2020.137703.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recogn. Lett. 27 (4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011.

Gruszecka-Kosowska, A., Baran, A., Wdowin, M., Mazur-Kajta, K., Czech, T., 2020. The contents of the potentially harmful elements in the arable soils of southern Poland, with the assessment of ecological and health risks: a case study. Environ. Geochem. Health 42 (2), 419–442. https://doi.org/10.1007/s10653-019-00372-w.

Hengl, T., Heuvelink, G., Stein, A., 2003. Comparison of kriging with external drift and regression-kriging. Technical Note, ITC 17. https://doi.org/10.1016/S0016-7061(00)00042-2.

Hengl, T., Heuvelink, G., Geoderma, A.S.-, 2004. A Generic Framework for Spatial Prediction of Soil Variables Based on Regression-Kriging. Elsevier. https://www.sciencedirect.com/science/article/pii/S0016706103002787.

Hengl, T., Heuvelink, G.B.M., Rossiter, D.G., 2007. About regression-kriging: from equations to case studies. Comput. Geosci. 33 (10), 1301–1315. https://doi.org/10.1016/j.cageo.2007.05.001.

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a Random Forest approach. Geoderma 214–215, 141–154. https://doi.org/10.1016/j.geoderma.2013.09.016.

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. J. Comput. Graph Stat. 15 (3), 651–674.

Hu, J., Peng, J., Zhou, Y., Xu, D., Zhao, R., Jiang, Q., Fu, T., Wang, F., Shi, Z., 2019. Quantitative estimation of soil salinity using UAV-borne hyperspectral and satellite multispectral images. Mdpi.Com 11, 736. https://doi.org/10.3390/rs11070736.

Huang, Y., Chen, Z., Liu, W., 2012. Influence of iron plaque and cultivars on antimony uptake by and translocation in rice (Oryza sativa L.) seedlings exposed to Sb(III) or Sb(V). Plant Soil 352 (1–2), 41–49. https://doi.org/10.1007/s11104-011-0973-x.

Ivushkin, K., Bartholomeus, H., Bregt, A., A. P.-R. sensing of, &amp, 2019. Global Mapping of Soil Salinity Change. Elsevier undefined. https://www.sciencedirect.com/science/article/pii/S0034425719302792.

Jia, X., Hu, B., Marchant, B., Zhou, L., Shi, Z., Pollution, Y.Z.-E., 2019. A Methodological Framework for Identifying Potential Sources of Soil Heavy Metal Pollution Based on Machine Learning: A Case Study in the Yangtze Delta. Elsevier. https://www.sciencedirect.com/science/article/pii/S0269749119302088.

Jiang, B., Adebayo, A., Jia, J., Xing, Y., Deng, S., Guo, L., Liang, Y., Zhang, D., 2019. Impacts of heavy metals and soil properties at a Nigerian e-waste site on soil microbial community. J. Hazard Mater. 362, 187–195. https://doi.org/10.1016/J.JHAZMAT.2018.08.060.

Joharestani, M.Z., Cao, C., Ni, X., Bashir, B., Talebiesfandarani, S., 2019. PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Mdpi.Com. https://doi.org/10.3390/atmos10070373.

John, K., Agyeman, P.C., Kebonye, N.M., Isong, I.A., Ayito, E.O., Ofem, K.I., Qin, C.Z., 2021a. Hybridization of cokriging and Gaussian process regression modelling techniques in mapping soil sulphur. Catena 206. https://doi.org/10.1016/j.catena.2021.105534.

John, K., Agyeman, P.C., Kebonye, N.M., Isong, I.A., Ayito, E.O., Ofem, K.I., Qin, C.Z., 2021b. Hybridization of cokriging and Gaussian process regression modelling techniques in mapping soil sulphur. Catena 206, 105534. https://doi.org/10.1016/j.catena.2021.105534.

Kawaguchi, K., Networks, Y. B.-N., 2019. Depth with Nonlinearity Creates No Bad Local Minima in ResNets. Elsevier. https://www.sciencedirect.com/science/article/pii/S0893608019301820.

Keller, A., Abbaspour, K.C., Schulin, R., 2002. Assessment of uncertainty and risk in modeling regional heavy-metal accumulation in agricultural soils. J. Environ. Qual. 31 (1), 175–187. https://doi.org/10.2134/jeq2002.1750.

Keskin, H., Geoderma, S.G.-, 2018. Regression Kriging as a Workhorse in the Digital Soil Mapper's Toolbox. Elsevier. https://www.sciencedirect.com/science/article/pii/S0016706117316567.

Khosravi, Vahid, Gholizadeh, Asa, Saberioon, Mohammadmehdi, 2022. Soil toxic elements determination using integration of sentinel-2 and Landsat-8 images: effect of fusion techniques on model performance. Environ. Pollut. 310, 119828.

Kim, M., Access, Y.G.-I., 2020. Predicting Patent Transactions Using Patent-Based Machine Learning Techniques. Ieeexplore.Ieee.Org. https://ieeexplore.ieee.org/abstract/document/9223646/.

Kim, S., Choi, Y., Neurocomputing, M.L.-, 2015. Deep Learning with Support Vector Data Description. Elsevier. https://www.sciencedirect.com/science/article/pii/S092523121500380X.

Kozák, J., Němeček, J., Borůvka, L., Lérová, Z., Němeček, K., Kodešová, R., Zádorová, T., 2010. Atlas půd České republiky.[Soil Atlas of the Czech Republic]. Czech University of Life Sciences, Prague, Prague, p. 150.

Kuhn, M., Johnson, K., Kuhn, M., Johnson, K., 2013. An introduction to feature selection. In: Applied Predictive Modeling. Springer, New York, pp. 487–519. https://doi.org/10.1007/978-1-4614-6849-3_19.

Laben, C., Brower, B.V., Eastman Kodak Company, 2000. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. U.S. Patent 6,011,875.

Lewińska, K., Karczewska, A., 2019. Antimony in soils of SW Poland—an overview of potentially enriched sites. Environ. Monit. Assess. 191 (2), 1–18. https://doi.org/10.1007/s10661-019-7214-9.

Li, Y., Li, M., Li, C., Liu, Z., 2020. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. Sci. Rep. 10 (1), 1–12.

Ma, J., Ding, Y., Cheng, J., Tan, Y., Access, V.G.-I., 2019. Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: a City Management Perspective. Ieeexplore.Ieee.Org. https://ieeexplore.ieee.org/abstract/document/8863366/.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89 (1–2), 67–94.

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: a brief history and some lessons. Geoderma 264, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017.

Mohammadi, A.A., Yousefi, M., Soltani, J., Ahangar, A.G., Javan, S., 2018. Using the combined model of gamma test and neuro-fuzzy system for modeling and estimating lead bonds in reservoir sediments. Environ. Sci. Pollut. Control Ser. 25 (30), 30315–30324. https://doi.org/10.1007/s11356-018-3026-7.

Nakamaru, Y., Tagami, K., Uchida, S., 2006. Antimony mobility in Japanese agricultural soils and the factors affecting antimony sorption behavior. Environ. Pollut. 141 (2), 321–326. https://doi.org/10.1016/j.envpol.2005.08.040.

Nanos, N., Martín, J.A.R., 2012. Multiscale analysis of heavy metal contents in soils: spatial variability in the Duero river basin (Spain). Geoderma 189, 554–562.

Nicodemus, K.K., Malley, J.D., Strobl, C., Ziegler, A., 2010. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinf. 11 (1), 1–13. https://doi.org/10.1186/1471-2105-11-110/FIGURES/6.

Nishad, P., Chemosphere, A.B.-, 2021. Antimony, a Pollutant of Emerging Concern: A Review on Industrial Sources and Remediation Technologies. Elsevier. https://www.sciencedirect.com/science/article/pii/S0045653521007219.

Odeh, I., McBratney, A., Geoderma, D.C.-, 1995. Further Results on Prediction of Soil Properties from Terrain Attributes: Heterotopic Cokriging and Regression-Kriging. Elsevier. https://www.sciencedirect.com/science/article/pii/001670619500007B.

Pham, T.G., Kappas, M., Huynh, C. van, Hoang, L., Nguyen, K., 2019a. Application of Ordinary Kriging and Regression Kriging Method for Soil Properties Mapping in Hilly Region of Central Vietnam. Mdpi.Com. https://doi.org/10.3390/ijgi8030147.

Pham, T.G., Kappas, M., Huynh, C. van, Nguyen, L.H.K., 2019b. Application of ordinary kriging and regression kriging method for soil properties mapping in hilly region of Central Vietnam. ISPRS Int. J. Geo-Inf. 8 (3), 147. https://doi.org/10.3390/IJGI8030147.

Podsiki, C., Committee, S., 2008. Chart of heavy metals, their salts and other compounds. November 24–29.

Pohl, C., 1998. Sensing, J. V. G.-I. journal of remote, & 1998, undefined. Review article multisensor image fusion in remote sensing: concepts, methods and applications. Taylor & Francis 19 (5), 823–854. https://doi.org/10.1080/014311698215748.

Pouladi, N., Møller, A.B., Tabatabai, S., Greve, M.H., 2019. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. Geoderma 342, 85–92. https://doi.org/10.1016/j.geoderma.2019.02.019.

Probst, P., Boulesteix, A.L., Bischl, B., 2019. Tunability: importance of hyperparameters of machine learning algorithms. J. Mach. Learn. Res. 20.

Program, N.T., 2018. Report on Carcinogens Monograph on Antimony Trioxide: RoC Monograph 13. https://pubmed.ncbi.nlm.nih.gov/34730921/.

Quinlan, J.R., 1992. Learning with Continuous Classes. World Scientific, pp. 343–348. https://www.worldscientific.com/doi/pdf/10.1142/9789814536271#page=356.

Reimann, C., Birke, M., Demetriades, A., Johnson, C.C., Team, G.P., 2012. Geochemical Atlas of European Agricultural and Grazing Land Soil. 34th International Geological Congress, 2011.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., 2014. Chemistry of Europe's Agricultural Soils. Part A — Schweizerbart science publishers.

Rinklebe, J., Shaheen, S., El-Naggar, A., W E, H., 2020. Redox-induced Mobilization of Ag, Sb, Sn, and Tl in the Dissolved, Colloidal and Solid Phase of a Biochar-Treated and Untreated Mining Soil. Elsevier. https://www.sciencedirect.com/science/article/pii/S0160412019345556.

Saleh, H.N., Panahande, M., Yousefi, M., Asghari, F.B., Oliveri Conti, G., Talaee, E., Mohammadi, A.A., 2019. Carcinogenic and non-carcinogenic risk assessment of heavy metals in Groundwater wells in Neyshabur plain, Iran. Biol. Trace Elem. Res. 190 (1), 251–261. https://doi.org/10.1007/s12011-018-1516-6.

Sergeev, A.P., Buevich, A.G., Baglaeva, E.M., Shichkin, A.V., 2019. Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals. Catena 174, 425–435. https://doi.org/10.1016/j.catena.2018.11.037.

Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., Valavi, R., Davatgar, N., Behrens, T., Scholten, T., 2020. Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate. Mdpi.Com 12, 1095. https://doi.org/10.3390/rs12071095.

Tóth, G., Hermann, T., da Silva, M.R., Montanarella, L., 2016a. Heavy metals in agricultural soils of the European Union with implications for food safety. Environ. Int. 88, 299–309. https://doi.org/10.1016/J.ENVINT.2015.12.017.

Tóth, G., Hermann, T., Szatmári, G., Pásztor, L., 2016b. Maps of heavy metals in the soils of the European Union and proposed priority areas for detailed assessment. Sci. Total Environ. 565, 1054–1062. https://doi.org/10.1016/j.scitotenv.2016.05.115.

Umali, B., Oliver, D., Forrester, S., Catena, D.C.-, 2012. The Effect of Terrain and Management on the Spatial Variability of Soil Properties in an Apple Orchard. Elsevier. https://www.sciencedirect.com/science/article/pii/S0341816212000136.

US-PHS, 1992. Toxicological profile of antimony and compounds Agency for Toxic Substances and Disease Registry. U.S. Public Health Service.

Vacek, O., Vašát, R., Borůvka, L., 2020. Quantifying the pedodiversity-elevation relations. Geoderma 373, 114441. https://doi.org/10.1016/j.geoderma.2020.114441.

Wang, N., Wang, A., Kong, L., 2018a. Calculation and Application of Sb Toxicity Coefficient for Potential Ecological Risk Assessment. Environment, M. H.-S. of the T., & 2018, undefined. Elsevier. https://www.sciencedirect.com/science/article/pii/S0048969717319897.

Wang, F., Gao, J., Zha, Y., 2018b. Hyperspectral sensing of heavy metals in soil and vegetation: Feasibility and challenges. ISPRS J. Photogrammetry Remote Sens. 136, 73–84. https://doi.org/10.1016/J.ISPRSJPRS.2017.12.003.

Wu, Z., Chen, Y., Han, Y., Ke, T., Liu, Y., 2020. Identifying the influencing factors controlling the spatial variation of heavy metals in suburban soil using spatial regression models. Sci. Total Environ. 717, 137212 https://doi.org/10.1016/J.SCITOTENV.2020.137212.

Zeng, L., Wang, Y., Jing, L., Cheng, Q., 2021. Quantitative determination of auxiliary information for mapping soil heavy metals and soil contamination risk assessment. Appl. Geochem. 130, 104964 https://doi.org/10.1016/j.apgeochem.2021.104964.

Zeraatpisheh, M., Jafari, A., Bodaghabadi, M., Catena, S.A.-, 2020. Conventional and Digital Soil Mapping in Iran: Past, Present, and Future. Elsevier. https://www.sciencedirect.com/science/article/pii/S0341816219305661.

Zeraatpisheh, M., Garosi, Y., Owliaie, H.R., Ayoubi, S., Taghizadeh-Mehrjardi, R., Scholten, T., Xu, M., 2022. Improving the spatial prediction of soil organic carbon using environmental covariates selection: a comparison of a group of environmental covariates. Catena 208, 105723.

Zhang, S., Huang, Y., Shen, C., Ye, H., Geoderma, Y.D.-, 2012. Spatial Prediction of Soil Organic Matter Using Terrain Indices and Categorical Variables as Auxiliary Information. Elsevier. https://www.sciencedirect.com/science/article/pii/S001670611100214X.

Zhong, Q., Ma, C., Chu, J., Wang, X., Liu, X., Ouyang, W., Lin, C., He, M., 2020. Toxicity and bioavailability of antimony in edible amaranth (Amaranthus tricolor Linn.) cultivated in two agricultural soil types. Environ. Pollut. 257, 113642 https://doi.org/10.1016/j.envpol.2019.113642.

Journal of Environmental Management

Quantification of the concentration of cadmium in agricultural soil using legacy data, preferential sampling, sentinel 2, Landsat 8, and an ensemble model.

--Manuscript Draft--

Manuscript Number:	JEMA-D-22-10526
Article Type:	Research Article
Keywords:	Preferential sampling; Ensemble models; Uncertainty assessment; Legacy data; Remote sensing.
Corresponding Author:	PRINCE CHAPMAN AGYEMAN, Msc Czech university of life science, Department of soil science and soil protection PRAGUE, CZECH REPUBLIC
First Author:	PRINCE CHAPMAN AGYEMAN, Msc
Order of Authors:	PRINCE CHAPMAN AGYEMAN, Msc
	Luboš Borůvka
	Ndiye Michael Kebonye
	Vahid Khosravi
	Kingsley JOHN
	Ondrej Drabek
	Vaclav Tejnecky
Abstract:	The current study assesses and predicts Cd concentration in agricultural soil using two cadmium (Cd) datasets, namely legacy data (LD) and preferential sampling-legacy dataset (PS-LD). The study predicts Cd in agricultural soil using four streams of auxiliary datasets extracted from Sentinel 2 (S2) and Landsat 8 (L8) bands. The study was divided into two contexts: Cd prediction in agricultural soil using a series of ensemble modeling approaches in conjunction with a LD (context 1) and Cd prediction in agricultural soil using PS-LD (context 2) coupled with a series of ensemble models. In context 1, ensemble 1, L8 with PS-LD was the cumulative optimal approach that predicted Cd in agricultural soil with a higher R2 value of 0.76, root mean square error (RMSE) of 0.66, mean absolute error (MAE) of 0.35, and median absolute error (MdAE) of 0.13. However, with R2 = 0.78, RMSE = 0.63, MAE = 0.34, and MdAE = 0.15, ensemble 1, S2 of PS-LD was the best prediction approach in predicting Cd concentration in agricultural soil in context 2. Overall, the predictions from both contexts indicated that ensemble 1 of S2 combined with PS-LD was the most appropriate and best model for Cd prediction in agricultural soil. The modeling approaches' uncertainty in both contexts was assessed using ensemble-sequential gaussian simulation (EnSGS), which revealed that the degree of uncertainty propagated in the study area was within 5% in both contexts. The combination of the PS dataset and the LD along with ensemble models and the remote sensing dataset, produced promising results. Nonetheless, the results demonstrated that the 20m spatial resolution band dataset used in the prediction of Cd in agricultural soil outperformed the 10m spatial resolution. When PS is combined with LD, an appropriate modeling approach, and a well-correlated remote sensing dataset are used, good results are obtained.
Suggested Reviewers:	Mojtaba Zeraatpisheh mojtaba.zeraatpisheh@henu.edu.cn
	Ruhollah Taghizadeh- Mehrjardi ruhollah.taghizadeh-mehrjardi@mnf.uni-tuebingen.de

Title page

Quantification of the concentration of cadmium in agricultural soil using legacy data, preferential sampling, sentinel 2, Landsat 8, and an ensemble model.

Prince Chapman Agyeman[1]*, Luboš Borůvka[1], Ndiye Michael Kebonye[2,3], Vahid Khosravi[1], Kingsley JOHN[1], Ondrej Drabek[1], Vaclav Tejnecky[1].

Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural

[1]Resources, Czech University of Life Sciences Prague, 16500 Prague, Czech Republic

*Correspondence E-mail: agyeman@af.czu.cz (P.C. Agyeman)

[2]Department of Geosciences, Chair of Soil Science and Geomorphology, University of Tübingen, Rümelinstr. 19-23, Tübingen, Germany

[3]DFG Cluster of Excellence “Machine Learning: New Perspectives for Science”, University of Tübingen, AI Research Building, Maria-von-Linden-Str. 6, 72076, Tübingen, Germany

Dear editors in chief

We wish to submit our research article entitled “Quantification of the concentration of cadmium in agricultural soil using legacy data, preferential sampling, sentinel 2, Landsat 8, and an ensemble model” for your consideration for publication in the journal of environmental management. In the manuscript, we evaluated diverse modeling approaches to determine their ability to combine these models with auxiliary datasets in the prediction of Cd in agricultural soil. It elucidates the ability to hybridize modeling approaches to an auxiliary dataset such as remotes sensing datasets as well as its accuracy in predicting PTEs in soil. We believe the manuscript is appropriate for journal of environmental management because it incorporates models that are not typically used in the combination process to improve the prediction and mapping of potentially toxic elements and soil properties in the environment.

Furthermore, it is important to note that this manuscript has not previously been published in any language; that it is not currently under consideration by another journal in any language; and that all authors have seen and agreed to the current version being submitted.

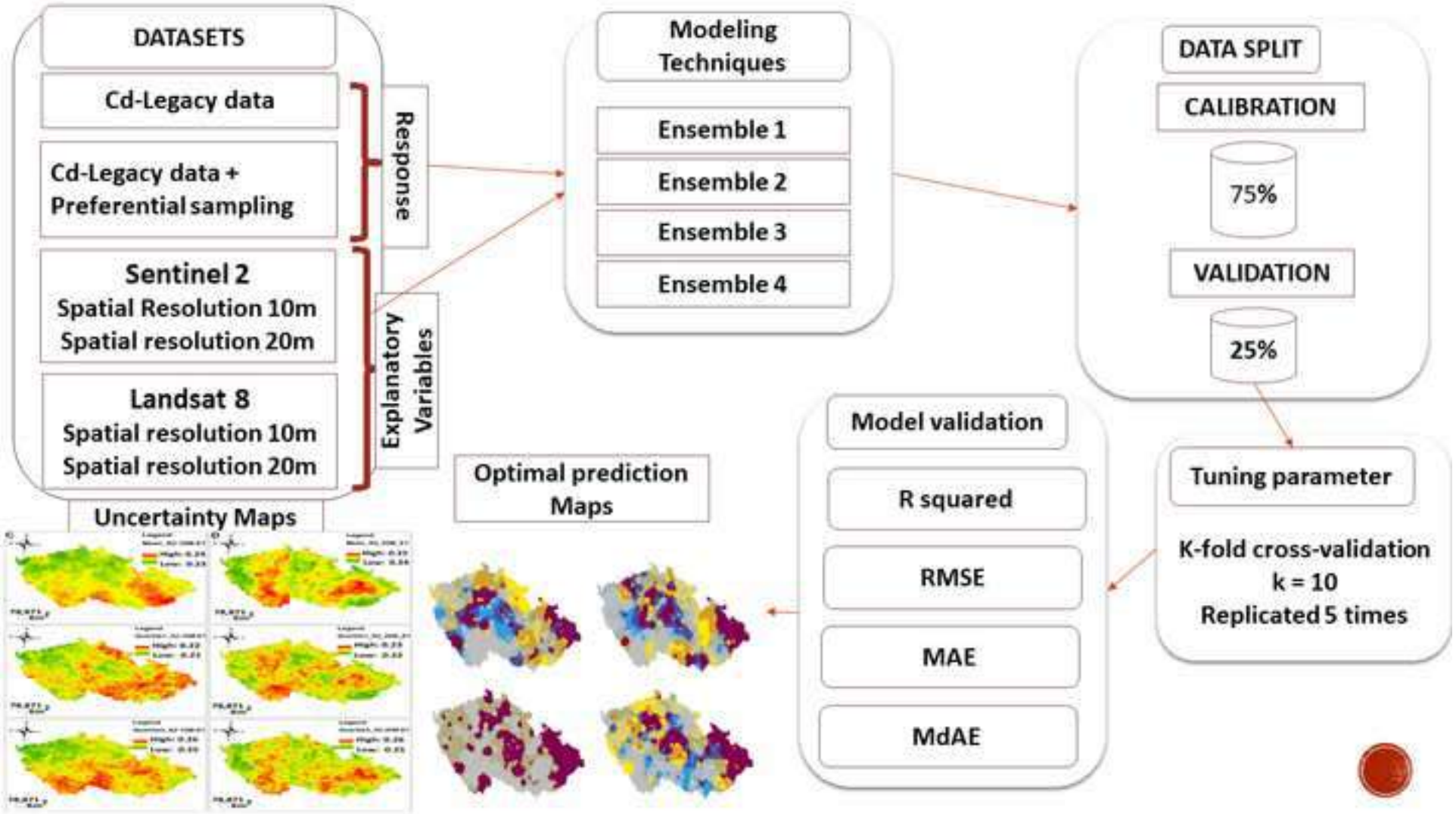
Thank you for your consideration and time.

Sincerely

Prince Chapman Agyeman

(Corresponding Author)

Graphical Abstracts



Highlights

Combining preferential sampling with legacy data yielded good results.

An ensemble sequential gaussian simulation was used for the uncertainty assessment.

Application of 20m spatial resolution performed better than 10m spatial resolution.

Ensemble 1 was the optimal ensembling approach for predicting cadmium in the soil.

Quantification of the concentration of cadmium in agricultural soil using legacy data, preferential sampling, sentinel 2, Landsat 8, and an ensemble model.

Prince Chapman Agyeman[1]*, Luboš Borůvka[1], Ndiye Michael Kebonye[2,3], Vahid Khosravi[1], Kingsley JOHN[1], Ondrej Drabek[1], Vaclav Tejnecky[1].

[1]Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, 16500 Prague, Czech Republic

*Correspondence E-mail: agyeman@af.czu.cz (P.C. Agyeman)

[2]Department of Geosciences, Chair of Soil Science and Geomorphology, University of Tübingen, Rümelinstr. 19-23, Tübingen, Germany

[3]DFG Cluster of Excellence “Machine Learning: New Perspectives for Science”, University of Tübingen, AI Research Building, Maria-von-Linden-Str. 6, 72076, Tübingen, Germany

Abstract

The current study assesses and predicts Cd concentration in agricultural soil using two cadmium (Cd) datasets, namely legacy data (LD) and preferential sampling-legacy dataset (PS-LD). The study predicts Cd in agricultural soil using four streams of auxiliary datasets extracted from Sentinel 2 (S2) and Landsat 8 (L8) bands. The study was divided into two contexts: Cd prediction in agricultural soil using a series of ensemble modeling approaches in conjunction with a LD (context 1) and Cd prediction in agricultural soil using PS-LD (context 2) coupled with a series of ensemble models. In context 1, ensemble 1, L8 with PS-LD was the cumulative optimal approach that predicted Cd in agricultural soil with a higher R^2 value of 0.76, root mean square error (RMSE) of 0.66, mean absolute error (MAE) of 0.35, and median absolute error (MdAE) of 0.13. However, with R2 = 0.78, RMSE = 0.63, MAE = 0.34, and MdAE = 0.15, ensemble 1, S2 of PS-LD was the best prediction approach in predicting Cd concentration in agricultural soil in context 2. Overall, the predictions from both contexts indicated that ensemble 1 of S2 combined with PS-LD was the most appropriate and best model for Cd prediction in agricultural soil. The modeling approaches' uncertainty in both contexts was assessed using ensemble-sequential gaussian simulation (EnSGS), which revealed that the degree of uncertainty propagated in the study area was within 5% in both contexts. The combination of the PS dataset and the LD along with ensemble models and the remote sensing dataset, produced promising results. Nonetheless, the results demonstrated that the 20m spatial resolution band dataset used in the prediction of Cd in agricultural soil outperformed the 10m spatial resolution. When PS is combined with LD, an appropriate modeling approach, and a well-correlated remote sensing dataset are used, good results are obtained.

Keywords. Preferential sampling; Ensemble models; Uncertainty assessment; Legacy data; Remote sensing.

Quantification of the concentration of cadmium in agricultural soil using legacy data, preferential sampling, sentinel 2, Landsat 8, and an ensemble model.

Introduction

Soil pollution has caused a decline in soil quality around the world, destroying many soil habitats and rendering the soil unable to support levels of ecosystems for its intended purpose. Soil as a repository is regarded as the complexity of the ecological system that facilitates human food production processes and a diverse range of ecosystems. Soil is regarded as a record of human operations from the ancient history to today, as well as a reflection of natural phenomena that are part of evolution's heritage. For instance, soil alteration for agriculture and the burial of archaeological evidence are both examples of this (Mishra et al., 2016). The impact of pollution on soil, particularly agricultural soil, reduces soil efficiency and has a negative impact on the physicochemical and biological properties of the soil, resulting in a decrease in output. The unintentional application of fertilizers, pesticides, organic manure, and chemicals to improve soil quality in order to increase yield each crop season contributes significantly to soil toxicity due to the accumulation effect. Even so, soil is subjugated to anthropogenic impacts such as agricultural practices, which result in extreme contamination (Khosravi et al., 2022a) with potentially toxic elements (PTEs) such as cadmium, lead, arsenic, copper, zinc, chromium, and so on. PTEs, specifically cadmium (Cd), lead (Pb), mercury (Hg), and arsenic (As), have had a significant effect on the soil's capacity to fulfill its prospective role as a habitat for macro- and microorganisms, arising in soil degradation, negatively affecting quality of food, durability, and safety, and aggravating possible future dangers to human health via the food chain (Jia et al., 2019; Shi et al., 2014).

Sustained agricultural field exploratory studies and attempts to enhance smart farming, thus further minimizing the use of agricultural inputs like agrochemicals and the implementation of potentially polluting fertilizers, have sparked the curiosity of people all over the globe over the years (Agyeman et al., 2022b). These studies have contributed to a rigorous analysis of PTEs in agricultural soil to evaluate levels of concentration as well as their impacts as the soil is utilized to grow food crops for human and animal consumption(Agyeman et al., 2022a, 2022b).

Continuous agricultural soil research is essential for providing a method to regulate soil health issues and soil characteristics as the underpinnings of sustainable farming, attempting to address prospective health and nutrition balance problems, and tackling climatic, natural, and humanitarian emergencies.

Cadmium remains one of the extremely crucial metals in the soil contaminant because it is ingested by humans and there is a narrow margin between daily caloric intake and the consumption that could have a substantial impact on human health (Asami, 1984). Unlike most other PTEs, Cd accumulates quickly and easily in edible plant parts to levels that are detrimental to nutrition while having no negative effect on plant development(Wang et al., 2019). Long-term nutritional Cd ingestion at elevated amounts can cause health complications (Nordberg et al., 2002; Zhang et al., 2014). For instance, grains such as rice, as the basic food for roughly half of the world's population, have comparably low concentrations of bioavailable Zn and Fe, which may translate into higher Cd absorption efficiency in the human body when particularly in comparison to other foods (Chaney, 2015). Cd toxic effect in plants disrupts the antioxidant defense mechanism and increases the development of reactive oxygen species (ROS) (which including superoxide ions, hydrogen peroxide, and hydroxyl radicals), causing pigments, lipids, proteins, DNA, and some cellular molecules to be damaged (Srivastava et al., 2020; Unsal et al., 2020). Antioxidant enzymatic like guaiacol peroxidase, superoxide dismutase, glutathione reductase, catalase, and ascorbate peroxidase aid in the reduction of oxidative stress triggered by ROS exacerbated by Cd toxic effects(Latif et al., 2020; Shiyu et al., 2020). Furthermore, Cd pollution has a negative impact on soil enzymatic operations, biogeochemistry cycles, and microbiota(Aponte et al., 2020; Suhani et al., 2021). Extreme Cd exposure to humans can cause kidney disease, respiratory disease, hepatocellular, skeletal, childbearing issues, and carcinogenic effects (Nordberg et al., 2018). Preceding assessments, for example, by the World Health Organization, have recognized renal damage as the paramount effect of long-term Cd exposure. Vertebral implications such as significantly reduced bone mineral density and a steadily increasing incidence of fractures were observed in clusters of the general population residing in the near Cd-exposed area (Nordberg et al., 2003).

The application of remote sensing datasets such as Sentinel 2 and Landsat 8 as auxiliary data for the prediction of PTEs in soil has been at the forefront of predictive mapping and prediction analysis. Despite being a low-cost tool for predictive mapping and assessment of PTEs and soil properties, it provides unrivaled benefits in monitoring the ground at different spatial resolutions and scales. Previous research such as Agyeman, Khosravi, et al., (2022b); Gorji et al., (2020); Khosravi et al., (2022b); Taghizadeh-Mehrjardi et al., (2020a); J. Wang et al., (2020) has effectively predicted PTEs and soil properties in the soil using Sentinel 2 and Landsat 8 separately or in tandem. At spatial resolutions of 30 m, Landsat 8 has been extensively used for optimized global ecological and security surveillance, as well as particularly precision mapping of PTEs or soil properties in the soil(Wulder et al., 2019). Moreover, the precision of multispectral assessment is frequently limited, owing to the spatial resolution that is coarse and broader bandwidths, as well as the benefits of the broader wavelength spectrum and comparatively preferred imaging effects (Peng et al., 2019). On the other hand, Sentinel 2, which has finer spatial resolution (10 or 20 m), more obtainable spatial frequency bands, and relatively large swath widths (290 km), has piqued scientists' curiosity and demonstrates DSM's remarkable prediction accuracy (Zhou et al., 2021).

Ensemble learning incorporates various varieties of modeling techniques and merges projections via a meta-learning to achieve elevated efficiency than a single learner (Sagi & Rokach, 2018). Ensemble learning models, as opposed to bagging, boosting, and averaging techniques, are occasionally used in digital soil mapping(Opitz and Maclin, 1999) but have gained popularity as many organizations have implemented the computational capabilities and sophisticated analytics software required to run such models. The stacking technique in the ensemble modeling approach can hold the competitive edge of the features of various machine learning techniques, decrease the variability of the solitary machine learning algorithms, and facilitate improved and more reliable predictions(Li et al., 2020). Previous studies including Afrifa et al., (2021); Bhagat et al., (2021); Lin et al., (2022); Tan et al., (2021); Tian et al., (2017); Q. Wang et al., (2015) have reported the efficiency of ensemble model in the prediction of PTEs.

The use of legacy data in predictive mapping and prediction analysis in digital soil mapping is becoming more popular. Legacy data (LD) is data collected from a specific time or period, and it can span a year or more to monitor soil quality or pollution over a given period. A preferential sampling (PS) however, is one in which the spatial process and sampling location are not thought to be stochastically independent. This sampling procedure is initiated when the area of interest is deemed polluted or has a peculiar problem. In this study, we will use LD in the prediction of Cd in the agricultural soil and, likewise, use a combination of preferential sampling and legacy data (PS-LD) in the prediction of Cd in the soil. The combination of preferentially sampled datasets with legacy data from prediction mapping and analysis, on the other hand, is uncharted territory. The use of multi-source satellite-derived covariates via individual ML techniques (Level 0) and stacking them (Level 1) with streams of diverse resample auxiliary dataset has not been extensively explored (DAS et al. 2022). More so, the use of ensembling models in prediction is not new, nor is the use of remote sensing datasets such as Landsat 8 and Sentinel 2 in prediction. However, the use of ensemble models, Landsat 8, and Sentinel 2 at various spatial resolutions, along with preferentially sampled datasets and legacy data, has never been done before. The utilization of legacy data in tandem with data from preferential samples will influence soil pollution spatial prediction over a larger area. The current research focuses on predicting Cd in agricultural soil using a combination of preferential sampling and legacy datasets. We will analyze and compare the prediction of Cd in agricultural soil using a series of ensemble modeling approaches along with a legacy dataset (LD) (context 1) and the prediction of Cd in agricultural soil using preferential sampling-legacy dataset (PS-LD) (context 2) coupled with a series of ensemble models in this study. The study's specific objectives are to compare the prediction of Cd in agricultural soil using two distinct Cd datasets; to apply the different spatial resolution of remote sensing datasets to the prediction of Cd in agricultural soil; to assess the propensity of ensemble models coupled with diverse Cd datasets and remote spatial resolution datasets; and finally, to assess uncertainty using ensemble-sequential gaussian simulation (EnSGS).

Material and methods

Study area

Diverse agricultural lands such as arable land, orchards, hopfields, vineyards, agricultural land (greening) and grassland are the agricultural areas used for this study. The choice of agricultural land used in this study is scattered across the Czech Republic (Figure 1), which forms part of the 4.2 million hectares of agricultural land that is approximately 42% of the total landmass of the Czech Republic. The Czech Republic is a Central European country located at 48° 33’-51°03’ N, 12°05’-18°51’ E, with elevations ranging from 115 to 1603 meters above sea level. The Koppen classification of the climatic conditions of the Czech Republic is cool subarctic climate (Dfc), humid continental climate (Dfb), and temperate oceanic climate (Cfb). Based on the Czech meteorological institute, the average precipitation yearly ranges from 559 to 893mm with a corresponding yearly average temperature of 6.8 to 8.9 degree Celsius. A diversity of soil classes is found in the Czech Republic, but the predominant soil class is the Cambisol, with other minor soil types such as stagnosols and fluvisols (Kozák et al, 2010). Soil substrates are made up of a diverse scope of materials that reflect the area's multifaceted geological structure, with higher elevations influenced by gradient sediments of numerous solid rocks and reduced elevation sediments inferred primarily from aeolian, fluvial, and lacustrine sources (Chlupáč et al, 2002). The farmland has a spinning topographic feature with a mountainous region and bumpy characteristics, with plains restricted to the lowest height (Žížala et al., 2022).

Insert figure 1 close to this section

Soil sampling and analysis

Two hundred and twenty-one legacy datasets (LD) were collected from agricultural land (arable land, orchards, hopfields, vineyards, agricultural land (greening) and grassland and additional hundred and fifteen preferential sampled (PS) dataset collected from the district of Frydek Mistek. The legacy dataset was obtained from Basal Soil Monitoring through the Central Institute for Supervising and Testing in Agriculture (ÚKZÚZ). Thus, issues of different sampling homogeneity are addressed by the Basal Soil Monitoring methodology and sampling design. Regardless of the small number of samples collected, it is safe to say that these adequately

represent the overwhelming majority of the landscape under consideration. The method used to extract the Cd for the legacy dataset is the same as for the preferential sampled dataset. The legacy data used in this study was from 2017, as historical data might not be reflective of the present situation due to the dynamics and the recovery period.

The area that was preferentially sampled is a relatively polluted area with steel industries, metal works, and intensive agricultural. The obtained soil samples were air dried prior to crushing with a machine-driven device (Fritsch disk grinder) and sieved to ascertain powdery soil sample of less than 200 mesh, 74 microns. Each 1-gram milled sample (finely ground, mixed thoroughly, mesh sieved) was placed in a clearly labelled Teflon bottle. 7 ml of 35% HCl and 3 ml of 65% HNO_3 were incorporated to each Teflon bottle (via fully automated dispensers—one for each acid), and the lid was daintily encased to enable the sample to continue to stay overnight for sample reactions to occur (aqua regia procedure)(Cools, 2016). Upon dissolving the soil sample, the mixed solution was deposited on a hot plate (metal) for 120 minutes to aid digestion before being permitted to cool. The mixture was filtered to obtain the supernatant. The supernatant was injected into a 50-ml Pyrex beaker and diluted with deionized water to the same volume. The supernatant was then filtered again into 50 ml PVC tubes. Besides that, 1 ml of the diluted concentration was diluted with 9 ml of de-ionized water and filtered into a 12 ml test tube to determine the solution's pseudo-total PTE concentration. ICP-OES was used to detect cadmium content levels in conformance with classic protocols and methods.

Modeling procedures and ensemble modeling approach

The set of data was randomly divided into two parts: testing (25%) and training (75%), with the training data used to create the modeling regression and the testing data used to validate the efficiency of the created model. Putting diverse models together to leverage each other's strengths and weaknesses to produce good modeling predictions paved the way for the ensemble modeling approach. It is essentially a hybridized model that combines the output of various modeling approaches into a single modeling approach to produce a more efficient output. Stacking is an ensemble technique that consolidates the results of various machine learning methods into a solitary component technique to achieve maximum generalization

precision (Breiman, 1996; Malone et al., 2014). The conceptual structure for a typical ensemble model follows two stages, such as the initial level (Level 0), which contains the sub model, and the final level, which contains all the predictions of the sub model piped through a meta-learning algorithm to give a final prediction owing to the sub model's departure from (level 0). Every ensemble model in this study is made up of four sub models, as shown in Figure SF1, and the stack tree or meta learner is a standalone modeling approach that uses the weights generated by the sub models to produce the final predictions. Based on the initial simulation performed on the individual models to determine the best models to be used as a meta learner for the ensemble models, the meta learners were used because of their superior individual performance over the sub-models. The packages used in the model were caretEnsemble, caretStack, brnn, bayesglm, rf, qrf, pls, cforest, xgbTree, cubist, gaussprLinear, and svmRadial.

Environmental covariates (EC)

Images of the Sentinel 2 satellite were obtained from an unrestricted satellite hub, and Sentinel 2 was obtained from the European Space Agency in August 2020 within the sampling period, (https://www.sentinel-hub.com/) and the bands were obtained using SNAP software. The Landsat 8-OLI satellite images were downloaded from the United States geological Earth-Explorer website. Atmospheric correction was applied to satellite images. The Sentinel 2A and Landsat 8-OLI bands were used in this study at two different spatial resolutions of 10m and 20m for the selected bands listed in Figure SF2 to estimate the concentration of Cd in agricultural soil using the legacy dataset alone and in combination with the preferential sampling dataset and legacy dataset. The bands used in this study had spatial resolutions of 10 and 20 m. (refer to figure SF2 for in supplementary material for details). The bands exhibited in Figure SF2 were selected because they share wavelengths within a specific range and have similar spectral properties. The following Sentinel 2A bands 11 and 12 (that is, of 20m spatial resolution) were downscaled to 10m in ArcGIS using a bilinear approach to achieve consistent 10m spatial resolution with the bands 2, 3, 4, and 8. This was done to create the first auxiliary dataset for predicting Cd in agricultural soil. Similarly, bands 2, 3, 4, and 8 with 10m spatial resolution were upscaled to 20m spatial resolution to be in sync with bands 11 and 12 for the second auxiliary

dataset. This was accomplished in ArcGIS using a bilinear approach. Landsat 8-OLI bands 2–7 (refer to figure 3), on the other hand, were downscaled from 30m spatial resolution to 20m and 10m spatial resolution, respectively, using the bilinear approach in ArcGIS to obtain two different auxiliary datasets with two different spatial resolutions. The pixels from the 336 observed points of the study area's sample locations were extracted in ArcGIS from all two distinct spatial resolutions for Sentinel 2 and Landsat 8 to be used for Cd content prediction in agricultural soil. In all four streams of auxiliary dataset was obtained such as

I. 10m spatial resolution of Sentinel 2 of band 2, 3, 4, 8, 11 and 12,
II. 20m spatial resolution of Sentinel 2 of band 2, 3, 4, 8, 11 and 12,
III. 10m spatial resolution of Landsat 8 of band 2, 3, 4, 5,6 and 7 and
IV. 20m spatial resolution of Landsat 8 of band 2, 3, 4, 5,6 and 7.

Bivariate mapping

The approach of categorizing spatial objects such as grid cells or area polygons according to the values of two variables is known as bivariate mapping (Speich et al., 2015). A bivariate color scheme is created by visualizing the two parameters as a single output using a single-color legend. A bivariate map illustrates the spatial interrelations of two raster layers (Tyner, 2010). Spatial correlations can then be analyzed as a single output map for various applications. When two variables have a spatial connection, it indicates that they are dependent on each other. Beard and Beard & Mackaness, (2006) reflect similar viewpoints in the scenario of the uncertain spatial analysis scenario, in which the feature and a technique for evaluating its predictive ability are highly symbolic depicted in a bivariate map. Furthermore, numerous studies have contrasted and demonstrated that the effectiveness of bivariate maps varies, and the results in each case are contingent on the map reader's knowledge and experience (Roth, 2013; Hope & Hunter, 2013). We refer to Kebonye et al., (2022) and Trumbo, (1981) research for more information on the bivariate mapping procedure. The optimal models for the Cd preferential sampling-legacy dataset raster layer and the Cd prediction using the legacy data raster layer both produce bivariate maps based on a bivariate and a corresponding spatial extent. The raster layers of these

predictions were fused together in R using the map function, it would result in a bivariate map with spatially distinct features from both layers.

Sequential gaussian simulation (SGS)

SGS's basic idea is to replicate sequential grid points utilizing the empirical distribution's temporary proportion (i.e., in this case the PTEs data). It produces an output that is comparable to the precise spatial actuality of a parameter of interest. Even though the data are anticipated to be detectable, the interpolated points represent the variogram technique and the nugget effect's local noise (Goovaerts, 2001). Furthermore, it is predicated on the multi presumption of a random feature model (Goovaerts, 2001; Johari et al., 2020). The set of data seems to provide the critical standard score alteration, making sure the logic of the univariate data distribution at the very least. Refer to Gholampour et al., (2019) for more information on SGS.

In this study, SGS was combined with an ensemble model to form a model known as ensemble sequential gaussian simulation (EnSGS), which was used to generate an uncertainty map for predicting Cd content in agricultural soil. The uncertainty was estimated using prediction intervals such as the first and third quarters, as well as a mean prediction for each approach.

Assessment accuracy and validation of the models

To evaluate the precision and validation of the modelling methods employed in this study, the coefficient of determination (R2), root mean square error (RSME), mean average error (MAE), and median absolute error (MdAE) were used. The regression model expresses R^2, which reflects the variability of the proportion in the response. The RMSE and MAE determine the size of the various versions within the individual measurement, allowing the approach prediction accuracy to be determined, whereas the MdAE affirms the true measurable value.

RESULTS AND DISCUSSION

Data description

The descriptive summary statistics of the dataset used in this study are presented in Table 1. Presented in the table are the maximum, minimum, median, mean, geometric mean (GM), lower and upper quartile, the tenth and the ninetieth percentile, standard deviation (SD), coefficient of variation (CV), skewness, and kurtosis. The minimum and maximum Cd concentration values from the legacy dataset (LD) as well as preferential sampling-legacy dataset (PS-LD) are the same, such as 0.10 mg/kg and 8.84 mg/kg. The concentration of Cd from the LD is 0.46 mg/kg whereas the mean concentration of Cd from PS-LD is 0.93mg/kg. The national limits for cadmium concentration in agricultural soil based on the Czech decree No.152/2016 Coll are fixed at 0.5 mg/kg, which is higher than the LD used in this study but based on the addition of the preferential sampling to the legacy dataset, the estimated mean concentration (0.93 mg/kg) is higher than the national background values for agricultural soil. Frydek Mistek is a relatively polluted area owing to intensive agriculture as well as the steel industry and metal works in the sub region. The geometric mean of Cd for LD and PLSD is 0.28 mg/kg and 0.51 mg/kg, with corresponding lower and upper quartile values ranging between 0.17 mg/kg and 0.36 mg/kg (LD) and 0.20 mg/kg to 1.40 mg/kg (PS-LD), respectively. Nevertheless, the 10th and the 90th percentile of LD range between 0.13 mg/kg and 0.84 mg/kg for LD and for PS-LD from 0.14 mg/kg to 2.11 mg/kg. The SD for both LD and PS-LD are 0.87 and 1.13, respectively, with corresponding skewness and kurtosis values of 6.22 and 47.84 for LD and 2.86 and 12.27 for PS-LD. The CV of the dataset was above 100, indicating a high heterogeneity of Cd in the study due to the accretion of Cd concentration in the agricultural soil from diverse pollutants.

Insert Table 1 close to this section

Prediction Cd using remote sensing dataset at 10m and 20m spatial resolution via ensemble models

The prediction of the concentration of Cd in agricultural soil using remote sensing datasets from Landsat 8 (L8) and Sentinel 2 (S2) with a spatial resolution of 10m coupled with ensemble models as well as PS-LD and LD (Context 1) is presented in table 2. Four ensembling models were applied in the prediction of Cd in agricultural soil using L8 and S2 as the auxiliary datasets. The PS-LD results showed that in ensemble 1 Cd prediction yielded R2, RMSE, MAE, and MdAE values of 0.76, 0.66, 0.35, and 0.13 for L8 and 0.75, 0.67, 0.37, and 0.16 for S2. In ensemble 2, the L8 prediction of Cd in agricultural soil yielded R2, RMSE, MAE, and MdAE values of 0.75, 0.65, 0.41, and 0.22, however in S8, Cd concentration was predicted with R2, RMSE, MAE, and MdAE values of 0.58, 0.90, 0.48, and 0.19, respectively. The prediction of Cd concentration in agricultural soil using ensemble 3 revealed that L8 produced 0.64 (R2), 0.82 (RMSE), 0.52 (MAE), and 0.22 (MdAE), whereas S2 produced 0.71 (R2), 0.69 (RSME), 0.42 (MAE), and 0.21 (MdAE). The results of ensemble 4 suggested that using L8 Cd prediction yielded 0.74, 0.66, 0.38, and 0.17 for R2, RMSE, MAE, and MdAE, respectively, while the results of S2 yielded 0.69, 0.71, 0.44, and 0.21 for R2, RMSE, MAE, and MdAE, correspondingly. The prediction results for LD of Cd applying the four ensemble models generated abysmal results for both the S2 and L8 in 10m spatial resolution for both remote sensing datasets, except for ensemble 3 of L8, which produced satisfactorily predicted Cd in agricultural soil with R2, RMSE, MAE, and MdAE values of 0.58, 0.48, 0.37, and 0.14 correspondingly. Ensemble 1 provided the best prediction of Cd in agricultural soil in context 1 based on L8, and ensemble 1 also obtained the best model in S2 coupled with PS-LD prediction of Cd in agricultural soil. Ensemble 3 of LD was the optimal modeling approach in the prediction of Cd in agricultural soil. In the prediction of Cd in agricultural soil, Ensemble 3 of LD was the best modeling approach. However, the optimal modeling approaches based on the application of LD and PS-LD were ensembles 3 L8 and ensembles 1 L8, which produced the best prediction results in the prediction of Cd in agricultural soil.

Insert Table 2 close to this section

In context 2, prediction of Cd concentration in the agricultural soil was done using the 20m spatial resolution of the Sentinel 2 and Landsat 8 datasets along with PS-LD, LD, and ensemble models. In the PS-LD of S2 and L8, the prediction of Cd yielded satisfactory results for ensemble 1 with R2, RMSE, MAE and MdAE of 0.64, 0.88, 0.43, and 0.14 for L8 and 0.78, 0.63, 0.34, and 0.15 for S2. Ensemble 2 produced relatively satisfactory results for both remote sensing datasets with the following R2, RMSE, MAE, and MdAE results: 0.70, 0.78, 0.49, and 0.26 for L8 and 0.71, 0.72, 0.46, and 0.24 for S2. Conversely, in ensemble 3, the prediction of the concentration of Cd in the agricultural soil yielded the following results: 0.60(R2), 0.88 (RMSE), 0.55(MAE), 0.23(MdAE) for L8 and 0.69(R2), 0.72(RMSE), 0.46 (MAE) and 0.25(MdAE) for S2. Ensemble 4 also produced satisfactory results for Cd prediction in agricultural soil, yielding prediction values of 0.74 (R2), 0.74 (RMSE), 0.44 (MAE), and 0.18 (MdAE) for L8 and 0.71 (R2), 0.69 (RMSE), 0.44 (MAE), and 0.21 (MdAE) for S2. The application of LD along with the ensemble models in the prediction of Cd in the agricultural soil produced appalling results apart from ensemble 3 of L8 the generated satisfactory result with R2 value of 0.56, RMSE 0.50, MAE 0.29 and MdAE 0.15. The optimal approach in the prediction of Cd in the approach based on the application of PS-LD in L8 was ensemble 4 and for S2 ensemble 2. Nevertheless, with LD, the optimal approaches were ensemble 3 for L8 and ensemble 1 for S2.

Insert Table 3 close to this section

The performance of the ensemble models in the prediction of Cd in the results displayed in Table 2 and 3 showcases the ability of stacking models to predict the concentration of Cd in agricultural soil at a national scale. The modelling precision for PS-LD using 10m spatial resolution for L8 and S2 is ensemble 1> ensemble 2> ensemble 4> ensemble 3 for L8 and ensemble 1> ensemble 3>ensemble 4> ensemble 2 for S2. On the other hand, the modelling precision for the prediction of Cd using 10m spatial resolution-based LD for L8 and S2 are as follows ensemble 3> ensemble 2> ensemble 1> ensemble 4 for L8 and ensemble 1> ensemble 3> ensemble 2> ensemble 4 for S2. Similarly, the modeling accuracy for the prediction of Cd in the soil using remote sensing dataset of 20m spatial resolution suggested that the prediction accuracy based on the usage of Cd PS-LD is in this order ensemble 4> ensemble 2> ensemble 1> ensemble 3 for L8 and ensemble 1> ensemble 4> ensemble 2> ensemble 3 for S2. Likewise, the prediction precision Cd based on

LD follows in the order ensemble 3> ensemble 2> ensemble 1> ensemble 3 for L8 and ensemble 1> ensemble 4> ensemble 3> ensemble 2.

In context 1, the ensemble modeling approach prediction of Cd in agricultural soil using the 10m spatial resolution of the remote sensing dataset based on PS-LD from L8 and S2 revealed that Ensemble 1_L8 was the best modeling approach with a high R2 (0.76) value and minimal MAE (0.35) and MdAE (0.13) values. On the other hand, the prediction results from LD suggested that the application of L8 of 10m spatial resolution coupled with ensemble 3 was likewise the optimal technique in the prediction of the Cd concentration in agricultural soil. Based on the output of the prediction of Cd in the agricultural soil using S2 and L8 from LD, it was evident that using L8 of 10m spatial resolution as an auxiliary dataset coupled with ensemble 1 with predicted values of R2 (0.58), RMSE (0.48), MAE (0.27), MdAE (0.14) was the appropriate method in the prediction of Cd in agricultural soil. When the two optimal predictions using PS-LD and LD along with the ensemble models were compared, it was clear that ensemble 1 L8 (10m spatial resolution) along with PS-LD was the cumulative optimal approach in context 1 that predicted Cd in agricultural soil with a higher R2 value. Even though ensemble 3_L8 from LD obtained minimal RSME (0.48), the marginal increase for R2 (31.03%) in favor of ensemble 1_L8_PS-LD as against the marginal decrease of RMSE (27.27%-ensemble 3_L8_LD) indicates that ensemble 1_L8_PS-LD is the optimal approach in predicting the concentration of Cd in agricultural with minimal MdAE and MAE results and a corresponding high R2 value.

The predictions result in context 2 based on the usage of 20m spatial resolution of remote sensing dataset coupled with ensemble models and PS-LD (Cd) however suggested that ensemble 1 of S2 (R2 = 0.78, RMSE= 0.63, MAE= 0.34 and MdAE=0.15) was the utmost modeling method in the prediction of Cd in the agricultural soil. Even though there were other modeling approaches that predicted Cd with satisfactory results, ensemble 1 of S2 provided the best prediction with minimum errors. Nevertheless, the prediction outputs from LD revealed that the best prediction modeling method in the prediction of Cd in agricultural soil was ensemble 3 of L8 with prediction output 0.56 (R2), 0.5 (RMSE), 0.39 (MAE) and 0.15 (MdAE). Comparing the optimal predictions outputs from the LD and PS-LD coupled with ensembles models it was apparent that ensemble 1

of S2 was the overall utmost prediction approach in predicting the Cd concentration in agricultural soil.

The overall best prediction approach in the prediction of Cd in agricultural soil either applying S2 and L8 from both 10m and 20m spatial resolution along with the ensemble models indicates unequivocally that the application of ensemble 1 of S2 of PS-LD with spatial resolution of 20m was the appropriate and best method for the prediction of Cd in agricultural soil with minimum errors and a higher R2 value. This implies that using remote sensing datasets with higher spatial resolution does not necessarily mean that prediction results will be improved; rather, it is dependent on the modeling techniques used as well as the spatial distribution of the dataset. Chen et al. (2004) observed the precision increment obtained by coarsening the image resolution, who improved the accuracy of spectral unmixing by resampling the Ikonos image resolution from 4 to 30 m. Obtaining better results from modeling an area is not solely reliant on the auxiliary dataset, but the ability to select the appropriate modeling approach along with the dataset may have a higher propensity to obtain good output. According to Zhou et al., (2021) predictions from modelling approaches created with coarse spatial resolution sensors can be comparable, if not superior, to models created with higher resolution sensors. The use of remote sensing images in the prediction of soil properties in a rural agricultural environment revealed that the soil prediction approach with a low spatial resolution evidenced productive accuracy when particularly in comparison to the approach with a higher spatial resolution (Xu et al., 2017). Kim et al., (2012) reported that the application of a multi-scale modeling approach, soil series by remote sensing dataset application in a wetland ecosystem and discovered that datasets extracted from remote sensing images with lower or coarse spatial resolution performed better than datasets extracted from images with higher spatial resolution. Xia & Zhang, (2022), applied remote sensing images in a comparative analysis for the prediction of soil pH in the soil, and the authors found that using higher resolution remote sensing images in the prediction of soil properties in the soil does not necessarily increase prediction efficiency when compared to using medium resolution images.

Even though the current study is unique in that PS-LD and LD are evaluated using ensemble modeling, however, there have been numerous studies that have applied Sentinel 2 and Landsat

8 datasets and their combinations to a variety of fields. Satellite sensors' massive and prevalent data streams can, however, guarantee that soil surveillance and mapping procedures for large regions are created precisely, quickly, and successfully (Malenovsk et al. 2012). Furthermore, some satellite images are hampered by factors that affect image quality. Satellite data can be valuable because of its broad spatial coverage, quick revisit time, and potential to acquire data without regard to local air traffic limitations. Unfortunately, due to haziness or the requirement for parched and bald soil environmental conditions, these predefined reconsider times may not be sufficient for adequate temporal coverage (Crucil et al. 2019). Sometimes, other complexities for satellite applications include low image resolution and limited accessibility of high-quality temporal and spatial images, owing to adverse atmospheric conditions and sensor requirements (Xiang et al. 2011). In Finland, S2 was discovered to perform relatively better than L8 when both remote sensors were evaluated for assessing canopy cover and LAI (Korhonen, Packalen, and Rautiainen 2017). Studies comparing S2 to L8 and previous Landsat sensors discovered that S2 has improved spatial and spectral capabilities for discriminating rangeland management practices (Sibanda, Mutanga, and Rouget 2016), estimating forest canopy cover and leaf area index (LAI) (Korhonen, Packalen, and Rautiainen 2017), and increasing the categorization quality of built-up areas (Pesaresi et al. 2016).

Resampling remote sensing datasets from a coarse or lower spatial resolution to a higher or finer spatial resolution or vice versa does not always result in good prediction efficiency. Most of the time, during the resampling process, these images lose quality, which can have an impact on the pixels that are extracted and used to predict PTEs or soil properties in the soil. The primary distinction between down-scaling and up-scaling synthetic and original images is that finer or coarser spatial details must be restored in the original down/up-scaling (Khosravi et al. 2022), and thus inability to maintain spatial detail has an impact on image quality. Some images in S2 have 20 and 10m and not all the bands were supposed to be resampled to either higher or lower spatial resolution, unlike in L8. For instance, the use of some resampled bands in S2, such as Bands 2,3,4, and 8, from 10 m spatial resolution to 20 m spatial resolution, in conjunction with unsampled bands 11 and 12, improved the prediction of Cd in agricultural soil using PS-LD. When using bands that could be similar for both sensors, the S2 and L8 prediction results were even

closer, but the error outputs in the result for the spatial resolution of 20m of S2 were lower than for L8 in PS-LD. This implies that the ability to use original bands without resampling tends to produce better results with less error. The combination of original bands and resampled bands has a higher predictive modeling ability to produce good results than resampling all the bands into different spatial resolutions. This implies that the unsampled band preserves the captured image details and qualities of the original without distortion. The original bands contain valuable information for predictive mapping. As a result, the use of original captured satellite images is critical in prediction modeling. Although resampling may be a good way to obtain higher or coarser spatial resolution of bands for a specific objective, a combination of original bands in their original states and resampled bands has a higher chance of producing good results.

Ensemble model performance employing preferential sampling + legacy data (PS-LD) and legacy data (LD)

The performance of the ensemble models using the standard stacking approach for the four ensemble models for the prediction of Cd in agricultural soil based on PS-LD and LD presented in tables 2 and 3 showed a wide range of prediction accuracy results, ranging from good to poor predictions. Generally, the ensemble models from the use of PS-LD exhibited high prediction accuracy compared to the prediction results from the LD used. This might be attributed to the preferential sampling collected from relatively polluted areas in the Czech Republic. Comparing the optimal models from PS-LD and LD from context 1 and context 2, it was apparent that the R2 increased by a margin of 31.03 to 114.29 % in the prediction results based on PS-LD rather than LD. Even though some of the prediction results showed lower errors, their R2 values were below 0.5, which is not significant statistically to be considered. According to Li et al., (2016), a model R2 prediction accuracy output of less than 0.5 is unacceptable. Nevertheless, Willmott, (1981) emphasized the importance of RMSE over R2. This assertion can be valid if the prediction accuracy of a model is 50% (0.5) or greater. In general, the ensemble 1 produced the best results for predicting Cd in agricultural soil, as evidenced by the optimal result scatter plot (Figures 2 and 3). Out of the 8 optimal models from PS-LD and LD from spatial resolutions of 10m and 20m, five optimal modelling approaches were from ensemble 1. This suggests that using quantile regression forest as the meta-learning models (stack tree) for the sub models was a successful

approach that evaluated the conditional variability of the prediction results of the weighted average from the sub models. In fact, stacking techniques combine the capabilities of multiple learning algorithms to improve predictive efficiency and render the predictive algorithm quite rigorous (Taghizadeh-Mehrjardi et al., 2020b). However, we highlight that the stacking approach in ensemble 1 provided more precise predictions than the other ensembling models due to the individual models' performance forming a resilient ensemble approach, which resulted in the best results. The effectiveness of the stacking technique is typically determined by two factors: (1) the training dataset does not always provide enough information to identify a single precise method; and (2) the learning procedures of the individual method may be severely flawed (Tajik et al., 2020; Wang, 2018). Overfitting and individual modeling approaches' biases are known to be reduced by stacking. Stacking varying machine learning models, on the other hand, enhances prediction accuracy and is confirmed for improving digital soil mapping (Das et al. 2022). To improve predictive performance, the capabilities of stacking can be enjoyed in the combined effect of heterogeneous weak learners via a meta-learning model (Das et al. 2022). Stacking improved reliability by balancing assessment and prediction bias and variability throughout individual validation.

The success of ensemble models is largely dependent on aggregating the strengths of all sub models to compensate for the weaknesses of all models used to make it more robust. Numerous papers that used ensemble models reported that the combination of algorithms in an ensemble approach in predictive mapping in the prediction of potentially toxic elements as well as soil properties produced optimal results (Wang. 2018; Sagi & Rokach, 2018). Biney et al., (2022) used an ensemble model to predict arsenic in agricultural soil, and the authors concluded that the application of an ensemble outperformed the application of individual models. In estimating soil organic carbon in Mollisols Tajik et al., (2020) applied ensemble model and the authors reported that the application of ensemble yielded outstanding results. The utilization of ensembles is prescribed for estimating farmland commodity premiums one month ahead, as a quite assertive efficiency is detected, allowing the crafted model's precision to be increased while reducing decision-making risk (Ribeiro and dos Santos Coelho, 2020). Ensemble models emerges to be a

progressive solution that requires opportunity of all mapping initiatives while also harmonizing existing maps from multiple datasets (Caubet et al., 2019).

The scatterplots in figures SF3 and SF4 display measured versus predicted data for the optimal modeling approach of Cd for PS-LD and LD for both remote sensing datasets. Figure SF3 exhibits measured versus predicted of the optimal modeling approach of Cd for LD and figure SF4 for PS-LD. Extreme points can be seen in the plots, which can be classified as outliers, but a positive outlier that enhances the prediction results from the modeling. The composition of a dataset for prediction largely depends on the information that is embedded in the dataset during modeling. According to Frost, (2021) , the removal of outliers from a plot normally has the tendency to affect the prediction results because these points might possess vital information that can augment the prediction positively. However, the ability to obtain better predictions does not necessarily lie in the removal of outliers but in the ability to obtain the perfect random split or partition of a dataset for training and testing. That precise partitioned test data coupled with the appropriate modeling algorithm has the propensity to generate an optimal prediction. The scatter plot with PS-LD outliers (Figure SF4) shows more inference patterns and brings the outliers closer to the other observed points than the LD scatter plot. The act of combining legacy data to preferential sampled dataset has yielded positive results by increasing the prediction results and making the outliers less an outlier based on the scatter plots. Preferential sampling tends to be more dependable, and it compares very well to standard sampling approaches when there is no preferential sampling (Dinsdale and Salibian-Barrera, 2019). In terms of bias, preferential sampling outperforms complete spatial randomness, though it should be acknowledged that the extent of the bias is relative (Antonelli et al., 2016). Based on the impact of adding a preferential sampling dataset to the legacy dataset, it is necessary to report that the impact of preferential sampling is strengthened with the rising prevalence of observational data in the training dataset when appropriate model validation is used. With increasing sample size, there is a higher probability of including samples from preferentially obtained tangible and intangible test and training sets.

Spatial distribution of Cd based optimal modeling approaches in each scenario

The distribution pattern of the predicted Cd based on the application of diverse streams of auxiliary datasets such as 10m and 20m spatial resolution of S2 and L8 is presented in figure 2A and 2B, highlighting the mapping of the optimal predictions in each context. The optimal prediction using Cd LD spatial distribution maps revealed that the L8-10M-E3, L8-20M-E3, and S2-10m-E1 share the same spatial Cd distribution pattern with patches of hotspots in the northeastern, southeast, and central enclaves of the study area. Similarly, the modeling approach S2-10m-E1 showed hotspots in the study area's southeastern and northeastern regions. Regardless of the fact that the dominant hotspots were found in the southeast and central areas of the maps, pockets of hotspots were also found in the southwest for L8-10M-E3 and S2-10m-E1, and the northeast for S2-10m-E1 and L8-20M-E3. Irrespective of the auxiliary dataset used in the prediction of Cd in agricultural soil, it was clear that the central area and the southeast enclave of the study area have elevated levels of Cd in agricultural soil. This implies that anthropogenic effect churning of agronomic and industrial practices is quite common in those regions. The distribution pattern of maps generated by PS-LD of Cd shares a similar distribution pattern in all four optimal models in each context, with hotspots realized in the southeastern sector of the map. Furthermore, the optimal model L8-20M-E4 displayed pockets of hotspots in the study area's southern and northeastern regions.

The spatial distribution exhibits the association of the optimal prediction using diverse datasets in contexts 1 and 2. The visualization of this optimal prediction is mapped using the quantile breaks option in bivariate mapping. Quantile breaks bring out the comprehensive details of the spatial variability that dignifies the nexus between the respective variables (Kebonye et al., 2022). The high Cd levels seen in the eastern part of the maps displaying the red coloration are due to the preferential sampling showing the impact of intensive agricultural, metalworking, and steel industry within that enclave. Sporadic high level (red patches) of cadmium can be seen in the central regions of all the maps indicating a consistent elevation of Cd within the central part of the study area. The maps A, B and D exhibited all colors on the precise scale which portray the undulated levels of Cd within the Czech Republic. However, the prediction output for the optimal model in PS-LD and LD mapped together (map C) shares similar patterns and displays a spatial

relationship that elicits gray, red, and pale-yellow colors, indicating low, moderate, and high Cd content in the study area. The simplicity of overlaying maps based on sampling regimes has resulted in an understandable color classification scheme. For example, in the southern region, regardless of the datasets or ensemble models used, all the maps displayed a grey color that reflected the Cd level in the area. The various assessments of connection employed to compare categorization schemes all exhibited similar behavior (Speich et al., 2015). The bivariate maps depict the nations where composite preferences conform as a result of the application of various modeling techniques and the fusion of various sampling regimes. Based on agricultural output, connections, and the overall level of Cd in the nation, areas with strong ecological integrity were identified. The fusion of the preferentially sampled dataset with the legacy data revealed that hotspots of Cd distribution in the study area are primarily found in the southeastern and northeastern enclaves. Considering the fact that using the LD optimal prediction alone revealed other hotspots in other regions of the study area, it appears clear that the pervasiveness of Cd pollution in the study area is more pronounced in the southeastern and northeastern regions of the study area.

Insert figure 2A close to this section

Insert figure 2B close to this section

Uncertainty assessment of the optimal models in context 1 and context 2 using ensemble sequential gaussians simulation approach

The uncertainty assessment of the optimal models in context 1 and 2 was done using the fusion of ensemble and sequential gaussian simulation (EnSGS). The distribution of Cd in the agricultural uncertainty maps presented in figure 3 and 4 semi-variogram fit using the spherical approach in SGS presented diverse nugget sill ratio. Heuvelink et al., (2001) reported that a good modeling technique's nugget sill ratio should be less than 0.25 to indicate higher spatial autocorrelation; 0.25 to 0.75 to indicate moderate spatial autocorrelation; and 0.75 or greater to indicate poor autocorrelation but with higher spatial randomness. The nugget sill ratio of the map produced using the LD Cd datasets showed moderate spatial autocorrelations (0.25 to 0.75) for all the uncertainty maps except for map D (S2-20M-E1) which exhibited poor spatial autocorrelation

(0.87). Even though map D (S2-20M-E1) of LD showed low spatial autocorrelation it however displayed high spatial randomness. The estimated nugget sill ratio of PS-LD maps displayed high spatial autocorrelations for all the uncertainty maps with corresponding low spatial randomness.

The uncertainty map was categorized into two groups, namely: the uncertainty map based on LD and the uncertainty map based on PS-LD. The maps were produced using the first quartile, third quartile, and mean. Each letter from A to D represents a different optimal prediction uncertainty map for the mean, first quartile, and third quartile arranged in a column. The uncertainty maps A and B of LD share similar patches of spatial uncertainty distribution pattern over the study area and exhibit low to high uncertainty variation in the study area. The corresponding estimated uncertainty propagation degrees for A and B were 4.04% and 2.76%, respectively. The uncertainty propagation of the maps C and D showed a mainly concentrated degree of uncertainty being exhibited in the eastern, southern, and central sections of the study area. The estimated degree of uncertainty in the study area for C and D was 0.69% and 0.57%, correspondingly. Uncertainty map A exhibited the highest uncertainty degree, and the least level of uncertainty estimated was from map D. The uncertainty levels propagated in the PS-LD maps were generally centered in the following areas: the east and central sections of the study area for map A; the west, south, and central region for B; the central part of the study area for C; and the north, south, and eastern enclave for D. The estimated degree of uncertainty propagated in the study area for each map was as follows: 4.46% (D), 3.95% (A), 3.92% (C), and 4.65% (A). Map B exhibited the highest uncertainty level, and the least level of uncertainty propagated was from map C.

Insert figure 3 close to this section

Insert figure 4 close to this section

The uncertainty map produced in the study area based on the optimal prediction were generally low which did not exceed 5%. The propagation of uncertainty across the study area based on varying degrees from low to high establishes the heterogeneity of the distribution of uncertainty through the combination of diverse ensemble models and SGS (EnSGS). Liao et al., (2016) contrasted bootstrap and SGS approach for modeling soil water temporal stability and

discovered that uncertainties derived by SGS have greater heterogeneity than uncertainty maps acquired by bootstrapping. Similarly, Sharififar, (2022) applied SGS and machine learning algorithms to estimate uncertainty and concluded that data variability should be regarded as a source of uncertainty that has the tendency to decrease accuracy. Based on the Cd datasets from LD and PS-LD it was obvious that the uncertainty level in LD was smaller than that of PS-LD. According to Odeh et al. (2012), the use of legacy data in DSM may introduce high levels of prediction uncertainty. This is contrary to the results obtained in this study due to the age of the legacy data that was used was quite current. The differences in the degree of uncertainty could be attributed to the legacy data being more evenly distributed, thereby spreading the degree of uncertainty more broadly throughout the area. However, because preferential sampling data is more recent but too concentrated in one region, when combined with legacy data, the uncertainty level increases due to more samples concentrated in a specific area. The combination of ensemble and SGS in uncertainty assessment is a novel practice. However according to Szatmári et al., (2019) the application of SGS in the quantification of uncertainty yield good results than the application of quantile regressing forest. The minimal level of uncertainty degree obtained in this study might likely be attributed to SGS models leveraging it higher precision on the ensemble model to obtain the great results.

Insert table 4 close to this section

Conclusion

The study applies remote sensing datasets such as Sentinel 2, Landsat 8, a combination of preferential sampling and legacy datasets, as well as legacy datasets coupled with ensemble models in the prediction of Cd in agricultural soil. Two contexts were applied, that is, the prediction of the concentration of Cd in agricultural soil using remote sensing datasets from Landsat 8 (L8) and Sentinel 2 (S2) with a spatial resolution of 10m coupled with ensemble models as well as PS-LD and LD (Context 1) and the prediction of Cd concentration in the agricultural soil was done using the 20m spatial resolution of the S2 and L8 datasets along with PS-LD, LD, and ensemble models (Context 2). The results suggested that in context 1, the application of ensemble 1_L8 along with PS-LD 10m spatial resolution was the overall best approach in the prediction of Cd in soil. In context 2, the results suggested that the application of PS-LD (Cd), ensemble 1 of S2 and of 20 m spatial resolution was the overall optimal approach in the prediction of Cd in agricultural soil. However, the cumulative comparison of the optimal models from both contexts (1 and 2) revealed that PS-LD (Cd), ensemble 1 of S2 and of 20 m spatial resolution, was the overall best method for the prediction of Cd in agricultural soil in this study. The combination of ensemble and SGS (EnSGS) in uncertainty estimation was under 5%. Thus, it is obvious that hybridizing SGS with an ensemble model yielded great results. The study highlights that the application of high spatial resolution of a remote sensing dataset does not necessarily mean that the best prediction results will be obtained. Nevertheless, the combination of the auxiliary dataset with an appropriate algorithm has a higher tendency to produce good results. It is also worth mentioning that the combination of preferential sampling with legacy data can generate high prediction accuracy. Therefore, regions with a study area that shows a high level of pollution should be preferentially sampled and added to legacy data to increase the prediction efficiency. Hybridizing SGS with a machine learning algorithm in uncertainty assessment has the propensity to yield good results.

Acknowledgement

The Czech University of Life Sciences Prague supported this research with an internal PhD scholarship no. 21130/1312/3131 from the Faculty of Agrobiology, Food, and Natural Resources (CZU). We acknowledge the Czech Science Foundation, project No. 17-27726S and the Basal Soil Monitoring, which was kindly provided by the Central Institute for Supervising and Testing in Agriculture (ÚKZÚZ).

Declaration of Competing Interest

The authors declare that they have no known competing personal interests or relationships that could have appeared to influence the scientific work in this manuscript.

Refences

Afrifa, G.Y., Ansah-Narh, T., Loh, Y.S.A., Sakyi, P.A., Chegbeleh, L.P., Yidana, S.M., 2021. Estimation of groundwater heavy metal pollution indices via an amalgam of stack ensemble learning. International Conference on Electrical, Computer, and Energy Technologies, ICECET 2021. https://doi.org/10.1109/ICECET52533.2021.9698570

Agyeman, P.C., John, K., Kebonye, N.M., Borůvka, L., Vašát, R., 2022a. Combination of enrichment factor and positive matrix factorization in the estimation of potentially toxic element source distribution in agricultural soil. Environ Geochem Health 1–27. https://doi.org/10.1007/S10653-022-01348-Z/TABLES/4

Agyeman, P.C., Khosravi, V., Michael Kebonye, N., John, K., Borůvka, L., Vašát, R., 2022b. Using spectral indices and terrain attribute datasets and their combination in the prediction of cadmium content in agricultural soil. Comput Electron Agric 198, 107077. https://doi.org/10.1016/J.COMPAG.2022.107077

Antonelli, J., Cefalu, M., Bornn, L., 2016. The positive effects of population-based preferential sampling in environmental epidemiology. Biostatistics 17, 764–778. https://doi.org/10.1093/BIOSTATISTICS/KXW026

Aponte, H., Meli, P., Butler, B., Paolini, J., … F.M.-S. of the T., 2020, undefined, 2020. Meta-analysis of heavy metal effects on soil enzyme activities. Science of the Total Environment.

Asami, T., 1984. Pollution of Soils by Cadmium. Changing Metal Cycles and Human Health 95–111. https://doi.org/10.1007/978-3-642-69314-4_6

Beard, K., Mackaness, W., 2006. Visual Access to Data Quality in Geographic Information Systems. https://doi.org/10.3138/C205-5885-23M7-0664 30, 37–45. https://doi.org/10.3138/C205-5885-23M7-0664

Bhagat, S.K., Tung, T.M., Yaseen, Z.M., 2021. Heavy metal contamination prediction using ensemble model: Case study of Bay sedimentation, Australia. J Hazard Mater 403, 123492. https://doi.org/10.1016/J.JHAZMAT.2020.123492

Biney, J.K.M., Vašát, R., Blöcher, J.R., Borůvka, L., Němeček, K., 2022. Using an ensemble model coupled with portable X-ray fluorescence and visible near-infrared spectroscopy to explore the viability of mapping and estimating arsenic in an agricultural soil. Science of The Total Environment 818, 151805. https://doi.org/10.1016/J.SCITOTENV.2021.151805

Breiman, L., 1996. Stacked regressions. Mach Learn 24, 49–64. https://doi.org/10.1007/BF00117832

Caubet, M., Román Dobarco, M., Arrouays, D., Minasny, B., Saby, N.P.A., 2019. Merging country, continental and global predictions of soil texture: Lessons from ensemble

modelling in France. Geoderma 337, 99–110. https://doi.org/10.1016/J.GEODERMA.2018.09.007

Chaney, R.L., 2015. How Does Contamination of Rice Soils with Cd and Zn Cause High Incidence of Human Cd Disease in Subsistence Rice Farmers. Curr Pollut Rep 1, 13–22. https://doi.org/10.1007/S40726-015-0002-4

Chlupáč, I., B.R., K.J., S.Z., 2002. Chlupáč, I., Brzobohatý, R., Kovanda, J., Straník,... - Google Scholar

Cools, N., and B.D.V., 2016. Sampling and analysis of soil." Manual on methods and criteria for harmonized sampling, assessment, monitoring and analysis of the effects of air pollution on forests.

Crucil, G., Castaldi, F., Aldana-Jague, E., van Wesemael, B., Macdonald, A., & Van Oost, K. (2019). Assessing the performance of UAS-compatible multispectral and hyperspectral sensors for soil organic carbon prediction. *Sustainability*, *11*(7), 1889.

Dinsdale, D., Salibian-Barrera, M., 2019. Modelling ocean temperatures from bio-probes under preferential sampling. https://doi.org/10.1214/18-AOAS1217 13, 713–745. https://doi.org/10.1214/18-AOAS1217

Frost J, 2021. Guidelines for removing and handling outliers in data - Google Scholar [WWW Document]. Statistics . URL

Gholampour, A., iranica, A.J.-S., 2019, undefined, 2019. Reliability analysis of a vertical cut in unsaturated soil using sequential Gaussian simulation. scientiairanica.sharif.edu. https://doi.org/10.24200/sci.2017.4571

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma.

Gorji, T., Yildirim, A., Hamzehpour, N., Tanik, A., Sertel, E., 2020. Soil salinity analysis of Urmia Lake Basin using Landsat-8 OLI and Sentinel-2A based spectral indices and electrical conductivity measurements. Ecol Indic 112, 106173. https://doi.org/10.1016/J.ECOLIND.2020.106173

Heuvelink, G., Geoderma, R.W.-, 2001, undefined, 2001. Modelling soil variation: past, present, and future. Elsevier.

Hope, S., Hunter, G.J., 2013. Testing the Effects of Thematic Uncertainty on Spatial Decision-making. http://dx.doi.org/10.1559/152304007781697884 34, 199–214. https://doi.org/10.1559/152304007781697884

Jia, Y., Jin, S., Savi, P., Gao, Y., Tang, J., Chen, Y., sensing, W.L.-R., 2019, undefined, 2019. GNSS-R soil moisture retrieval based on a XGboost machine learning aided method: Performance and validation. mdpi.com. https://doi.org/10.3390/rs11141655

Johari, A., Khani, M., … M.H.-S.D. and, 2020, undefined, 2020. System reliability analysis for seismic site classification based on sequential Gaussian co-simulation: A case study in Shiraz, Iran. Soil Dynamics and Earthquake Engineering .

Kebonye, N.M., Prince, ·, Agyeman, C., Seletlo, Z., Peter, ·, Eze, N., 2022. On exploring bivariate and trivariate maps as visualization tools for spatial associations in digital soil mapping: A focus on soil properties. Precision Agriculture 2022 1–22. https://doi.org/10.1007/S11119-022-09955-7

Khosravi, V., Gholizadeh, A., Saberioon, M., 2022a. Soil toxic elements determination using integration of Sentinel-2 and Landsat-8 images: Effect of fusion techniques on model performance. Environmental Pollution 310, 119828. https://doi.org/10.1016/J.ENVPOL.2022.119828

Kim, J., Grunwald, S., Rivero, R.G., Robbins, R., 2012. Multi-scale Modeling of Soil Series Using Remote Sensing in a Wetland Ecosystem. Soil Science Society of America Journal 76, 2327–2341. https://doi.org/10.2136/SSSAJ2012.0043

Kozák, J., 2010. Soil Atlas of the Czech Republic 150.

Korhonen, L., Packalen, P., & Rautiainen, M. (2017). Comparison of Sentinel-2 and Landsat 8 in the estimation of boreal forest canopy cover and leaf area index. *Remote sensing of environment*, *195*, 259-274.

Latif, J., Akhtar, J., Ahmad, I., Mahmood-ur-Rehman, M., Shah, G.M., Zaman, Q., Javaid, T., Farooqi, Z.U.R., Shakar, M., Saleem, A., Rizwan, M., 2020. Unraveling the effects of cadmium on growth, physiology and associated health risks of leafy vegetables. Revista Brasileira de Botanica 43, 799–811. https://doi.org/10.1007/S40415-020-00653-0

Li, L., Lu, J., Wang, S., Ma, Y., Wei, Q., Li, X., Cong, R., Ren, T., 2016. Methods for estimating leaf nitrogen concentration of winter oilseed rape (Brassica napus L.) using in situ leaf spectroscopy. Ind Crops Prod 91, 194–204. https://doi.org/10.1016/J.INDCROP.2016.07.008

Li, X., Luo, J., Jin, X., He, Q., Niu, Y., 2020. Improving Soil Thickness Estimations Based on Multiple Environmental Variables with Stacking Ensemble Methods. Remote Sensing 2020, Vol. 12, Page 3609 12, 3609. https://doi.org/10.3390/RS12213609

Liao, K., Lai, X., Lv, L., Research, Q.Z.-S., 2016, undefined, 2016. Uncertainty in predicting the spatial pattern of soil water temporal stability at the hillslope scale. Soil Research,.

Lin, N., Jiang, R., Li, G., Yang, Q., Li, D., Yang, X., 2022. Estimating the heavy metal contents in farmland soil from hyperspectral images based on Stacked AdaBoost ensemble learning. Ecol Indic 143, 109330. https://doi.org/10.1016/J.ECOLIND.2022.109330

Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma 232–234, 34–44. https://doi.org/10.1016/J.GEODERMA.2014.04.033

Malenovský, Z., Rott, H., Cihlar, J., Schaepman, M. E., García-Santos, G., Fernandes, R., & Berger, M. (2012). Sentinels for science: Potential of Sentinel-1,-2, and-3 missions for scientific observations of ocean, cryosphere, and land. *Remote Sensing of environment*, *120*, 91-101.

Mishra, R., Mohammad, N., Sangyan, N.R.- van, 2016, undefined, 2016. Soil pollution: Causes, effects and control. researchgate.net.

Nordberg, G., Jin, T., Bernard, A., … S.F.-A. a journal of the, 2002, undefined, 2002. Low bone density and renal dysfunction following environmental cadmium exposure in China. BioOne. https://doi.org/10.1579/0044-7447-31.6.478

Nordberg, G., Nordberg, G.F., Lundstrom, N.G., Gunnarsson, D., Svensson, M., Bernard, A., Buchet, J., Fierens, S., Dumont, X., Jin, T., Zeng, X., Lu, J., Wu, X., Jiang, X., Ye, T., Kong, Q., Frech, W., Nordberg, M., 2003. Cadmium and human health: A perspective based on recent studies in China. The Journal of Trace Elements in Experimental Medicine 16, 307–319. https://doi.org/10.1002/JTRA.10039

Nordberg, G.F., Bernard, A., Diamond, G.L., Duffus, J.H., Illing, P., Nordberg, M., Bergdahl, I.A., Jin, T., Skerfving, S., 2018. Risk assessment of effects of cadmium on human health (IUPAC Technical Report). Pure and Applied Chemistry 90, 755–808. https://doi.org/10.1515/PAC-2016-0910/ASSET/GRAPHIC/J_PAC-2016-0910_FIG_002.JPG

Opitz, D., Maclin, R., 1999. Popular Ensemble Methods: An Empirical Study. Journal of Artificial Intelligence Research 11, 169–198. https://doi.org/10.1613/JAIR.614

Peng, J., Biswas, A., Jiang, Q., Zhao, R., Hu, J., Hu, B., Geoderma, Z.S.-, 2019, undefined, 2019. Estimating soil salinity from remote sensing and terrain data in southern Xinjiang Province, China. Geoderma.

Ribeiro, M.H.D.M., dos Santos Coelho, L., 2020. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Appl Soft Comput 86, 105837. https://doi.org/10.1016/J.ASOC.2019.105837

Roth, R.E., 2013. The Impact of User Expertise on Geographic Risk Assessment under Uncertain Conditions. http://dx.doi.org/10.1559/152304009787340160 36, 29–43. https://doi.org/10.1559/152304009787340160

Sagi, O., Rokach, L., 2018a. Ensemble learning: A survey. Wiley Interdiscip Rev Data Min Knowl Discov 8. https://doi.org/10.1002/WIDM.1249

Sagi, O., Rokach, L., 2018b. Ensemble learning: A survey. Wiley Interdiscip Rev Data Min Knowl Discov 8. https://doi.org/10.1002/WIDM.1249

Sharififar, A., 2022. Accuracy and uncertainty of geostatistical models versus machine learning for digital mapping of soil calcium and potassium. Environmental Monitoring and Assessment 2022 194:10 194, 1–16. https://doi.org/10.1007/S10661-022-10434-9

Shi, T., Liu, H., Wang, J., Chen, Y., Fei, T., Wu, G., 2014. Monitoring arsenic contamination in agricultural soils with reflectance spectroscopy of rice plants. Environ Sci Technol 48, 6264–6272. https://doi.org/10.1021/ES405361N

Shiyu, Q., Hongen, L., Zhaojun, N., Pedosphere, Z.R.-, 2020, undefined, 2020. Toxicity of cadmium and its competition with mineral nutrients for uptake by plants: A review. Pedosphere .

Sibanda, M., Mutanga, O., & Rouget, M. (2016). Discriminating rangeland management practices using simulated hyspIRI, landsat 8 OLI, sentinel 2 MSI, and VENμs spectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(9), 3957-3969.

Speich, M.J.R., Bernhard, L., Teuling, A.J., Zappa, M., 2015. Application of bivariate mapping for hydrological classification and analysis of temporal change and scale effects in Switzerland. J Hydrol (Amst) 523, 804–821. https://doi.org/10.1016/J.JHYDROL.2015.01.086

Srivastava, V., Vaish, B., Singh, R.P., Singh, P., 2020. An insight to municipal solid waste management of Varanasi city, India, and appraisal of vermicomposting as its efficient management approach. Environ Monit Assess 192. https://doi.org/10.1007/S10661-020-8135-3

Suhani, I., Sahab, S., Srivastava, V., Singh, R.P., 2021. Impact of cadmium pollution on food safety and human health. Curr Opin Toxicol 27, 1–7. https://doi.org/10.1016/J.COTOX.2021.04.004

Szatmári, G., Geoderma, L.P.-, 2019, undefined, 2019. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma.

Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., Valavi, R., Davatgar, N., Behrens, T., Scholten, T., 2020a. Improving the Spatial Prediction of Soil Organic Carbon Content in Two Contrasting Climatic Regions by Stacking Machine Learning Models and Rescanning Covariate Space. Remote Sensing 2020, Vol. 12, Page 1095 12, 1095. https://doi.org/10.3390/RS12071095

Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., Valavi, R., Davatgar, N., Behrens, T., Scholten, T., 2020b. Improving the Spatial Prediction of Soil Organic Carbon Content in Two Contrasting Climatic Regions by

Stacking Machine Learning Models and Rescanning Covariate Space. Remote Sensing 2020, Vol. 12, Page 1095 12, 1095. https://doi.org/10.3390/RS12071095

Tajik, S., Ayoubi, S., Zeraatpisheh, M., 2020a. Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. Geoderma Regional 20. https://doi.org/10.1016/J.GEODRS.2020.E00256

Tajik, S., Ayoubi, S., Zeraatpisheh, M., 2020b. Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. Geoderma Regional 20, e00256. https://doi.org/10.1016/J.GEODRS.2020.E00256

Tan, K., Ma, W., Chen, L., Wang, H., Du, Q., Du, P., Yan, B., Liu, R., Li, H., 2021. Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning. J Hazard Mater 401, 123288. https://doi.org/10.1016/J.JHAZMAT.2020.123288

Tian, L., Liu, X., Zhang, B., Liu, M., Wu, L., 2017. Extraction of Rice Heavy Metal Stress Signal Features Based on Long Time Series Leaf Area Index Data Using Ensemble Empirical Mode Decomposition. International Journal of Environmental Research and Public Health 2017, Vol. 14, Page 1018 14, 1018. https://doi.org/10.3390/IJERPH14091018

Trumbo, B.E., 1981. A theory for coloring bivariate statistical maps. American Statistician 35, 220–226. https://doi.org/10.1080/00031305.1981.10479360

Tyner, J.A., 2010. Principles of map design [WWW Document]. [Sl]. URL https://scholar.google.co.uk/scholar?hl=en&as_sdt=0%2C5&q=Tyner%2C+J.+A.+%282010%29.+Principles+of+map+design.+New+York%3A+Guilford+Press.&btnG= (accessed 7.9.22).

Unsal, V., Dalkıran, T., … M.Ç.-A. pharmaceutical, 2020, undefined, 2020. The role of natural antioxidants against reactive oxygen species produced by cadmium toxicity: a review. Advanced pharmaceutical bulletin .

Wang, C.R.-T.J. of P.C., 2018, undefined, 2018. Significantly improving the prediction of molecular atomization energies by an ensemble of machine learning algorithms and rescanning input space: A stacked. ACS Publications 122, 8868–8873. https://doi.org/10.1021/acs.jpcc.8b03405

Wang, C.-T.J. of P.C., 2018, undefined, 2018. Significantly improving the prediction of molecular atomization energies by an ensemble of machine learning algorithms and rescanning input space: A stacked. ACS Publications 122, 8868–8873. https://doi.org/10.1021/acs.jpcc.8b03405

Wang, J., Ding, J., Yu, D., Teng, D., He, B., Chen, X., Ge, X., Zhang, Z., Wang, Y., Yang, X., Shi, T., Su, F., 2020. Machine learning-based detection of soil salinity in an arid desert region,

Northwest China: A comparison between Landsat-8 OLI and Sentinel-2 MSI. Science of The Total Environment 707, 136092. https://doi.org/10.1016/J.SCITOTENV.2019.136092

Wang, P., Chen, H., Kopittke, P.M., Zhao, F.J., 2019. Cadmium contamination in agricultural soils of China and the impact on food safety. Environmental Pollution 249, 1038–1048. https://doi.org/10.1016/J.ENVPOL.2019.03.063

Wang, Q., Xie, Z., Li, F., 2015. Using ensemble models to identify and apportion heavy metal pollution sources in agricultural soils on a local scale. Environmental Pollution 206, 227–235. https://doi.org/10.1016/J.ENVPOL.2015.06.040

Willmott, C.J., 1981. ON THE VALIDATION OF MODELS. Phys Geogr 2, 184–194. https://doi.org/10.1080/02723646.1981.10642213

Wulder, M., Loveland, T., Roy, D., … C.C.-R. sensing of, 2019, undefined, 2019. Current status of Landsat program, science, and applications. Remote Sens Environ.

Xia, C., Zhang, Y., 2022. Comparison of the use of Landsat 8, Sentinel-2, and Gaofen-2 images for mapping soil pH in Dehui, northeastern China. Ecol Inform 70, 101705. https://doi.org/10.1016/J.ECOINF.2022.101705

Xiang, H., & Tian, L. (2011). Development of a low-cost agricultural remote sensing system based on an autonomous unmanned aerial vehicle (UAV). *Biosystems engineering*, *108*(2), 174-190.

Xu, Y., Smith, S., Grunwald, S., … A.A.-E.-J. of environmental, 2017, undefined, 2017. Evaluating the effect of remote sensing image spatial resolution on soil exchangeable potassium prediction models in smallholder farm settings. J Environ Manage.

Zhang, W., Du, Y., Zhai, M., environment, Q.S.-S. of the total, 2014, undefined, 2014. Cadmium exposure and its health effects: a 19-year follow-up study of a polluted area in China. Science of the total environment .

Zhou, T., Geng, Y., Ji, C., Xu, X., Wang, H., Pan, J., Bumberger, J., Haase, D., Lausch, A., 2021a. Prediction of soil organic carbon and the C:N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. Science of The Total Environment 755, 142661. https://doi.org/10.1016/J.SCITOTENV.2020.142661

Zhou, T., Geng, Y., Ji, C., Xu, X., Wang, H., Pan, J., Bumberger, J., Haase, D., Lausch, A., 2021b. Prediction of soil organic carbon and the C:N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. Science of The Total Environment 755, 142661. https://doi.org/10.1016/J.SCITOTENV.2020.142661

Žížala, D., Minařík, R., Skála, J., Beitlerová, H., Juřicová, A., Reyes Rojas, J., Penížek, V., Zádorová, T., 2022. High-resolution agriculture soil property maps from digital soil mapping methods, Czech Republic. Catena (Amst) 212, 106024. https://doi.org/10.1016/J.CATENA.2022.106024

Figure

Click here to access/download;Figure;figures.docx

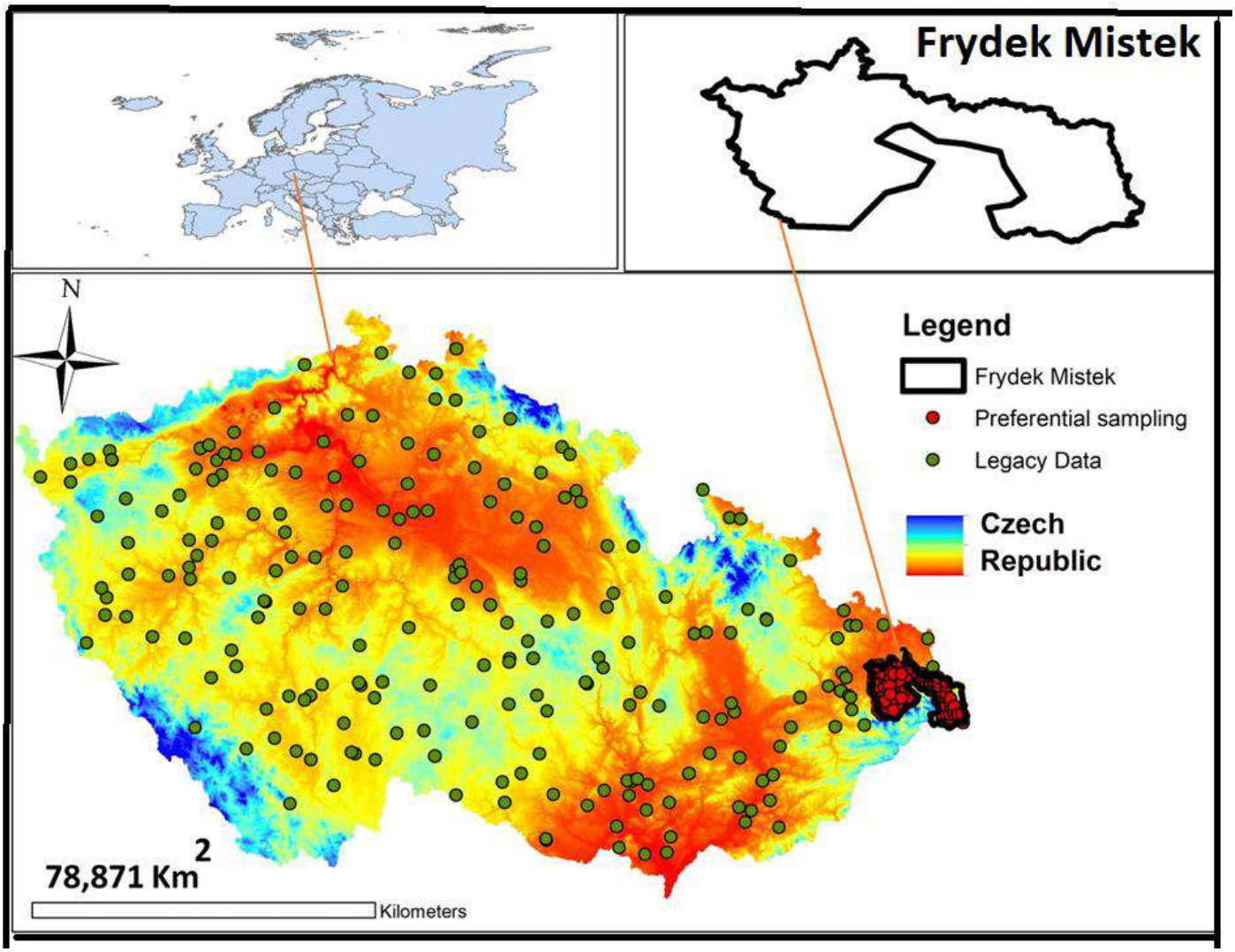


Figure 1 displaying the study area with sampled dataset from the Czech Republic

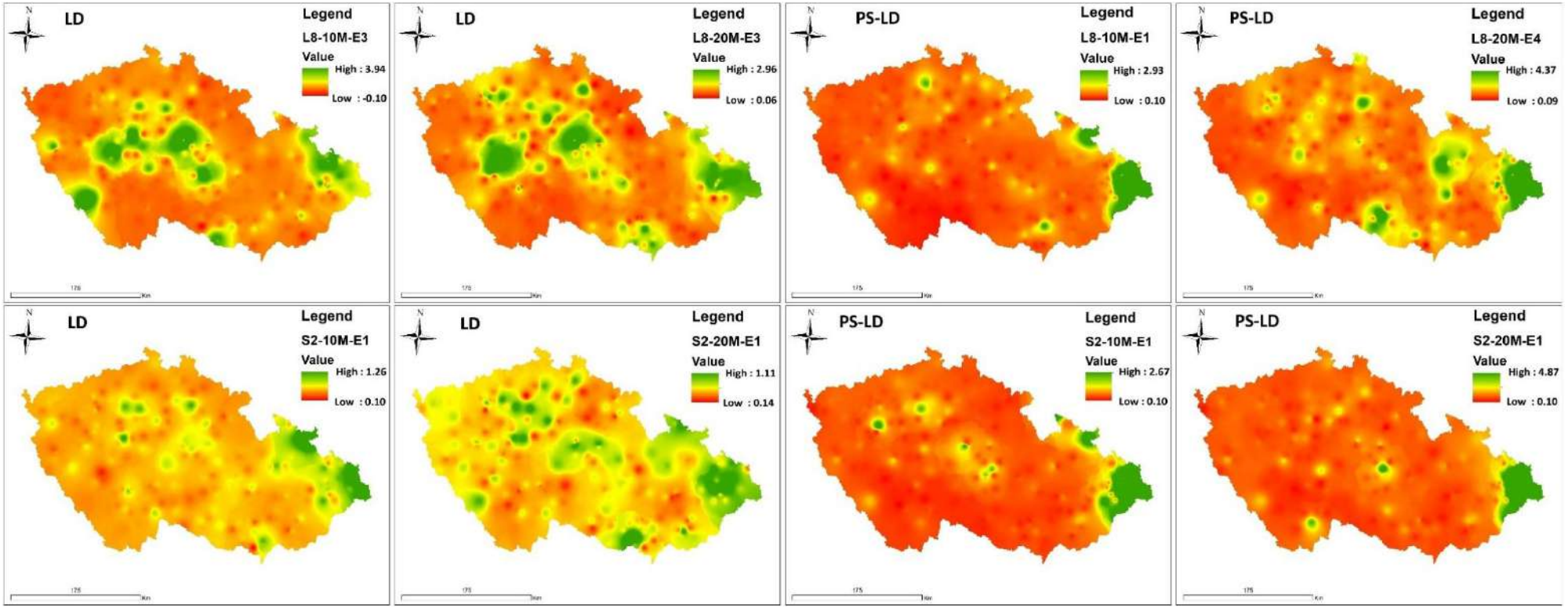


Figure 2A shows a spatial distribution of the optimal approaches in each scenario using PS-LD (preferential sampling-legacy data) and LD (legacy dataset) auxiliary datasets such as 10m and 20m spatial resolution of Sentinel 2 and Landsat 8 A (L8-10m-E4-PS-LD, L8-20m-E4-PS-LD, S2-10m-E1-PS-LD, L8-10m-E3-LD, S2-20m-E1-PS-LD, L8-20m-E3-LD, S2-10M-E1-LD, and D, S2-20m-E1-LD).

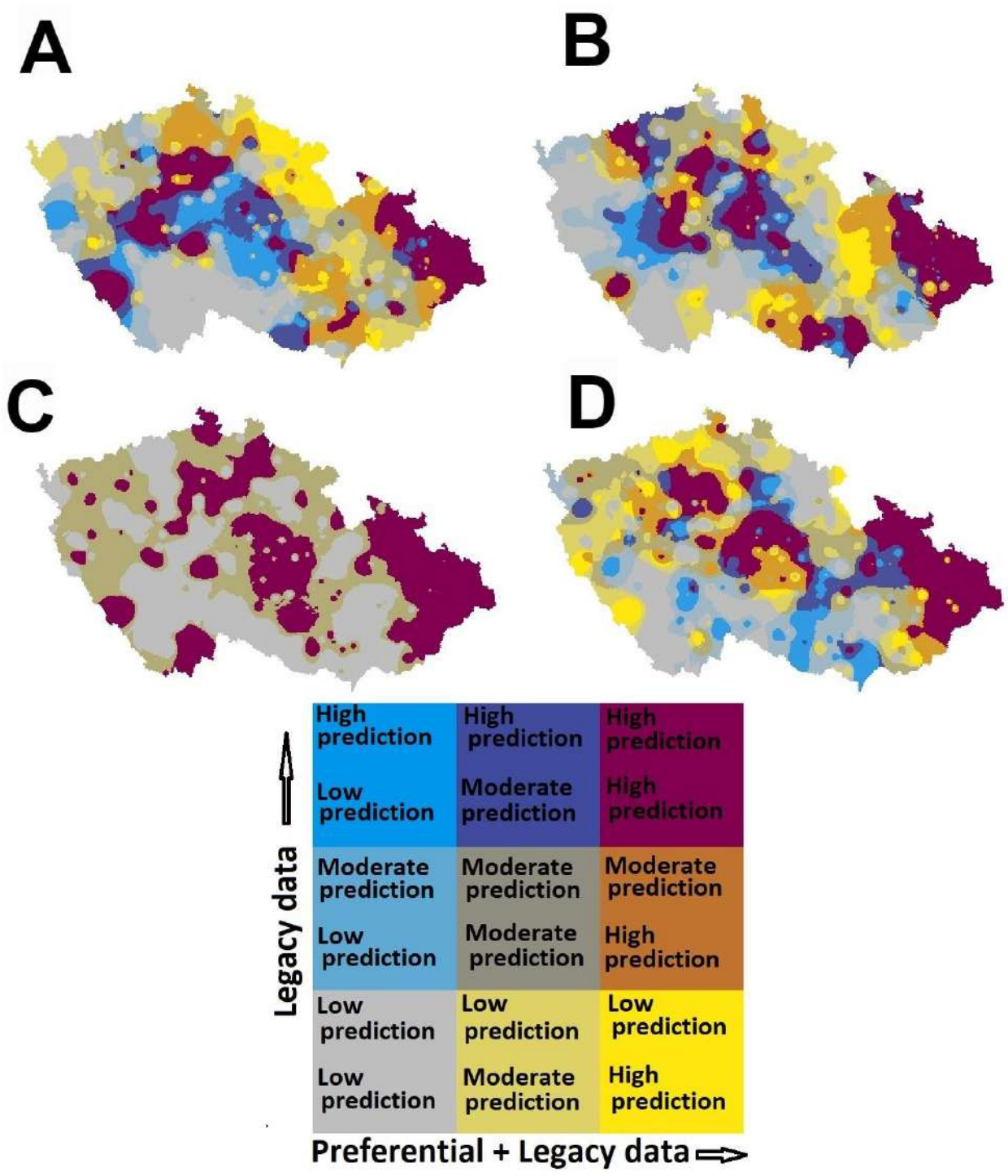


Figure 2B displays bivariate maps of the spatial distribution of cadmium using preferential sampling-legacy data(PS-LD) and legacy data (LD) across the Czech Republic based on quantile breaks A (L8-10m-E4-PS-LD and L8-10m-E3-LD), B(L8-20m-E4-PS-LD and L8-20m-E3-LD), C (S2-10m-E1-PS-LD and S2-10M-E1-LD) and D (S2-20m-E1-PS-LD and S2-20m-E1-LD).

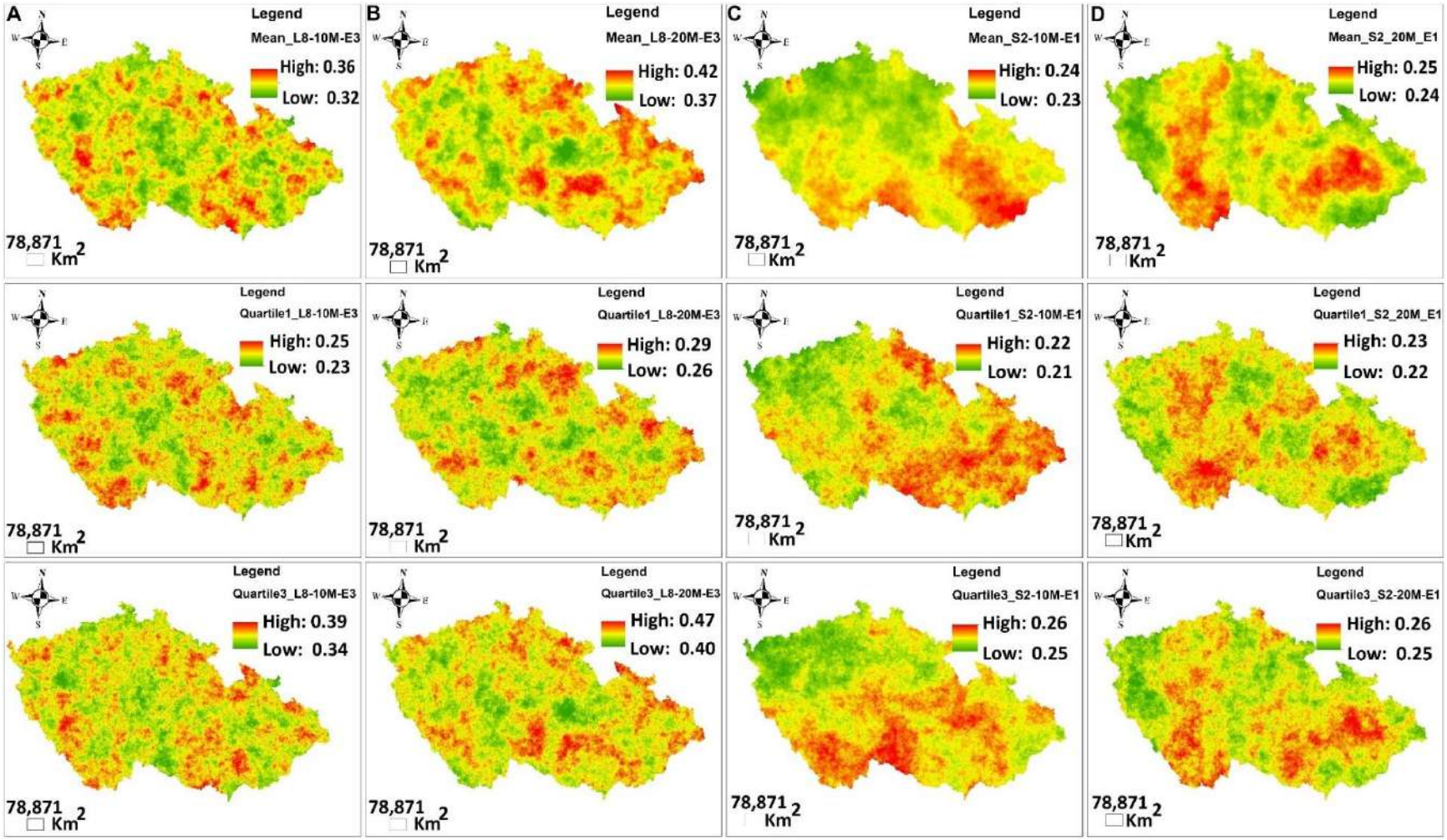


Figure 3 Using legacy dataset, Sentinel 2 and Landsat 8 datasets at 10 and 20 m spatial resolutions, respectively, in the uncertainty mapping (mean, quartile 1 and quartile 3) for cadmium concerntration in Czech agricultural soils.

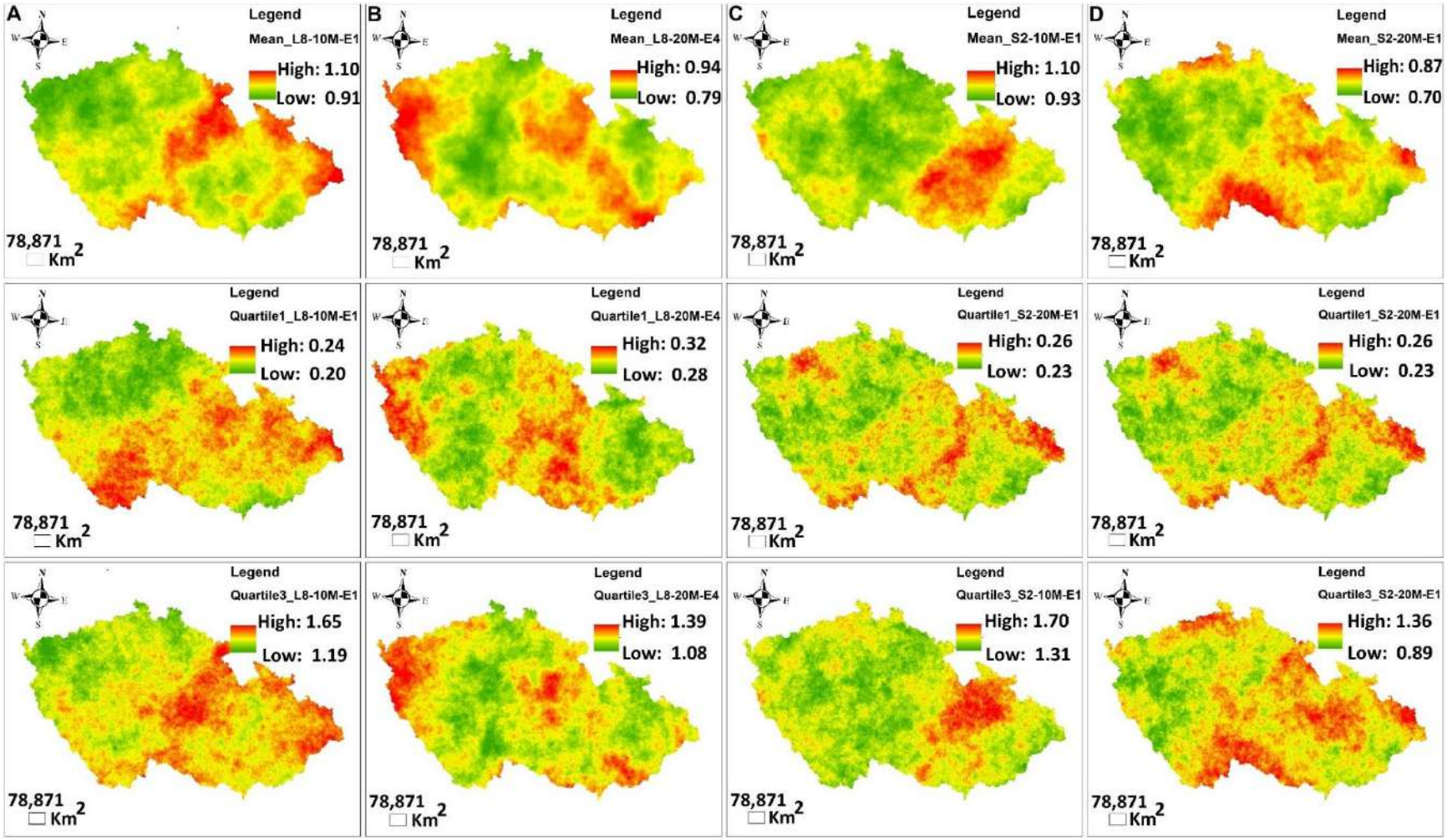


Figure 4 Using preferential sampling-legacy dataset, Sentinel 2 and Landsat 8 datasets at 10 and 20 m spatial resolutions, respectively, in the uncertainty mapping (mean, quartile 1 and quartile 3) for cadmium concerntration in Czech agricultural soils.

Table 1 summary of the statistical description of cadmium.

Description	Cd (LD)	Cd (PSLD)
Mean mg/kg	0.46	0.93
Geometric (Mean) mg/kg	0.28	0.51
Median mg/kg	0.24	0.37
Minimum mg/kg	0.10	0.10
Maximum mg/kg	8.84	8.84
Lower (Quartile) mg/kg	0.17	0.20
Upper (Quartile) mg/kg	0.36	1.40
Percentile (10th) mg/kg	0.13	0.14
Percentile (90th) mg/kg	0.84	2.11
Standard deviation	0.87	1.13
Coefficient Variation	190.29	121.54
Skewness	6.22	2.86
Kurtosis	47.84	12.27

(Cd (LD) represent cadmium from legacy dataset and Cd (PS-LD) represent cadmium from preferential sampling and legacy dataset)

Table 2 showing the prediction of Cd using remote sensing datasets, legacy datasets, and preferential sampling plus legacy dataset (Context 1)

LANDSAT 8 10M SPATIAL RESOLUTION					**SENTINEL 2 10M SPATIAL RESOLUTION**			
	Legacy and preferential sampling dataset							
	R2	RMSE	MAE	MdAE	R2	RMSE	MAE	MdAE
Ensemble 1	0.76	0.66	0.35	0.13	0.75	0.67	0.37	0.16
Ensemble 2	0.75	0.65	0.41	0.22	0.58	0.90	0.48	0.19
Ensemble 3	0.64	0.82	0.52	0.22	0.71	0.69	0.42	0.21
Ensemble 4	0.74	0.66	0.38	0.17	0.69	0.71	0.44	0.21
	Legacy dataset							
	R2	RMSE	MAE	MdAE	R2	RMSE	MAE	MdAE
Ensemble 1	0.23	0.58	0.28	0.09	0.35	0.54	0.27	0.12
Ensemble 2	0.37	0.51	0.33	0.16	0.26	0.54	0.30	0.17
Ensemble 3	0.58	0.48	0.37	0.14	0.37	0.51	0.34	0.21
Ensemble 4	0.30	0.63	0.39	0.20	0.39	0.73	0.34	0.20

Table 3 showing the prediction of Cd using remotes sensing datasets, legacy datasets, and preferential sampling plus legacy dataset (Context 2)

	LANDSAT 8 20M SPATIAL RESOLUTION				SENTINEL 2 20M SPATIAL RESOLUTION			
	Legacy and preferential sampling dataset							
	R2	RMSE	MAE	MdAE	R2	RMSE	MAE	MdAE
Ensemble 1	0.64	0.88	0.43	0.14	0.78	0.63	0.34	0.15
Ensemble 2	0.70	0.78	0.49	0.26	0.71	0.72	0.46	0.24
Ensemble 3	0.60	0.88	0.55	0.23	0.69	0.72	0.46	0.25
Ensemble 4	0.74	0.74	0.44	0.18	0.71	0.69	0.44	0.21
	Legacy dataset							
	R2	RMSE	MAE	MdAE	R2	RMSE	MAE	MdAE
Ensemble 1	0.29	0.57	0.27	0.09	0.44	0.58	0.32	0.09
Ensemble 2	0.49	0.48	0.32	0.24	0.17	0.59	0.37	0.20
Ensemble 3	0.56	0.50	0.39	0.15	0.27	0.60	0.38	0.22
Ensemble 4	0.37	0.66	0.44	0.29	0.35	0.72	0.34	0.14

Table 4 showing the semi variogram model fitted for Cd using the spherical technique and uncertainty assessment.

	LD-EnSGS				
	RANGE	NUGGET (C0)	SILL (C0 + C)	NUGGET /SILL RATIO (C0/C0 + C)	UNCERTAINTY%
S2-20M-E1 (D)	83709.64	0.89	0.13	0.87	0.57
L8-10M-E3(A)	37709.12	0.73	0.25	0.74	4.04
S2-10M-E1(C)	338782	0.74	0.4	0.65	0.69
L8-20M-E3(B)	55907	0.74	0.31	0.70	2.76
	PRES-LD- EnSGS				
S2-20M-E1(D)	148929.03	0.03	0.66	0.04	4.46
L8-10M-E1(A)	347410.47	0.05	1.27	0.04	3.95
S2-10M-E1 (C)	208852.42	0.01	0.85	0.01	3.92
L8-20M-E4(B)	198495	0.19	0.88	0.18	4.65

(L8 represent Landsat 8, S2 -sentinel 2. E1 to 4 referring to ensemble 1 to 4, LD-legacy dataset, SGS-sequential gaussian simulation, PREs-preferential sampling

Quantification of the concentration of cadmium in agricultural soil using legacy data, preferential sampling, sentinel 2, Landsat 8, and an ensemble model.

Prince Chapman Agyeman[1]*, Luboš Borůvka[1], Ndiye Michael Kebonye[2,3], Vahid Khosravi[1], Kingsley JOHN[1], Ondrej Drabek[1], Vaclav Tejnecky[1].

Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural

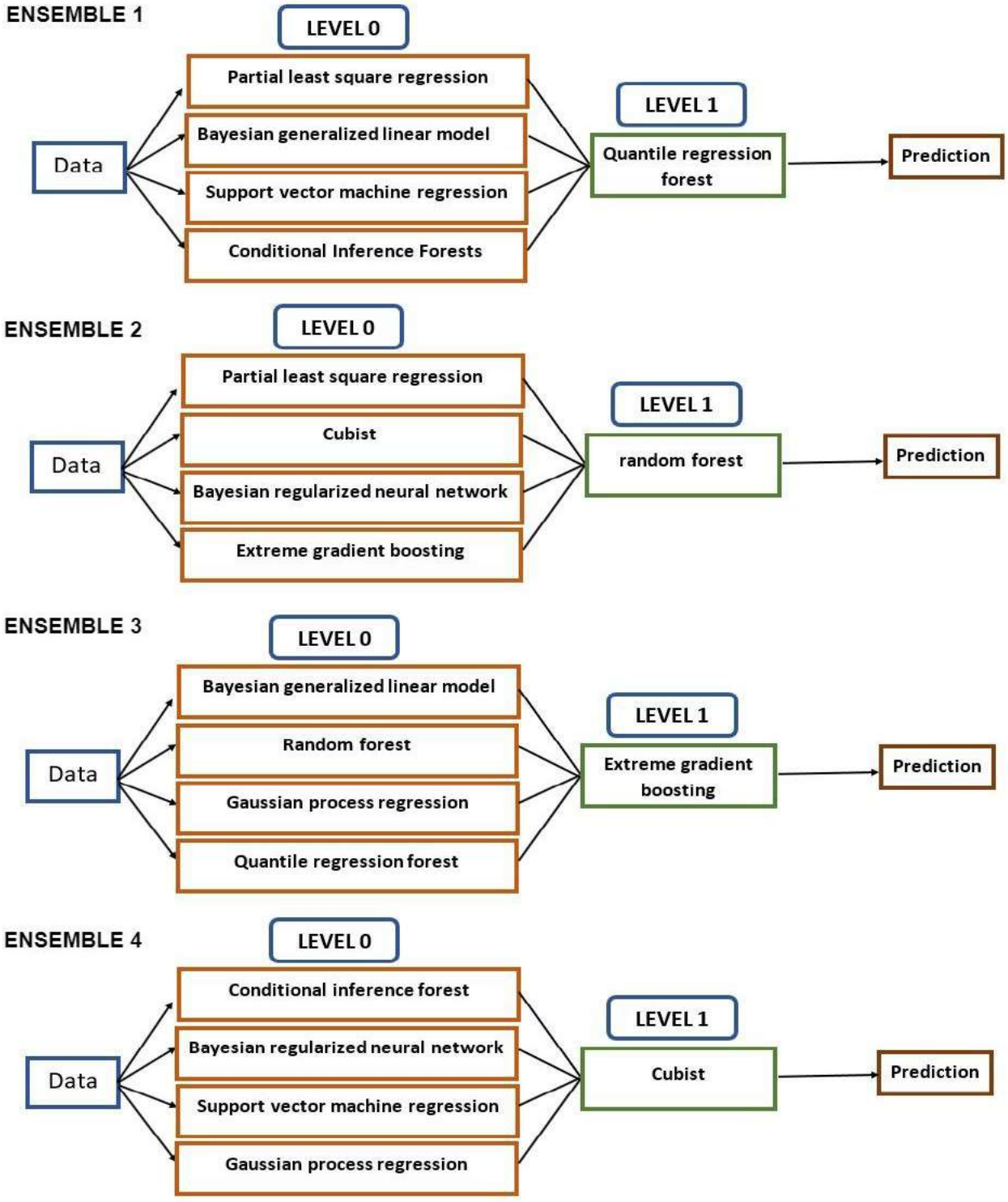
[1]Resources, Czech University of Life Sciences Prague, 16500 Prague, Czech Republic

*Correspondence E-mail: agyeman@af.czu.cz (P.C. Agyeman)

[2]Department of Geosciences, Chair of Soil Science and Geomorphology, University of Tübingen, Rümelinstr. 19-23, Tübingen, Germany

[3]DFG Cluster of Excellence “Machine Learning: New Perspectives for Science”, University of Tübingen, AI Research Building, Maria-von-Linden-Str. 6, 72076, Tübingen, Germany

Figures SF1 represent the general concept of the ensemble techniques applied in this study.



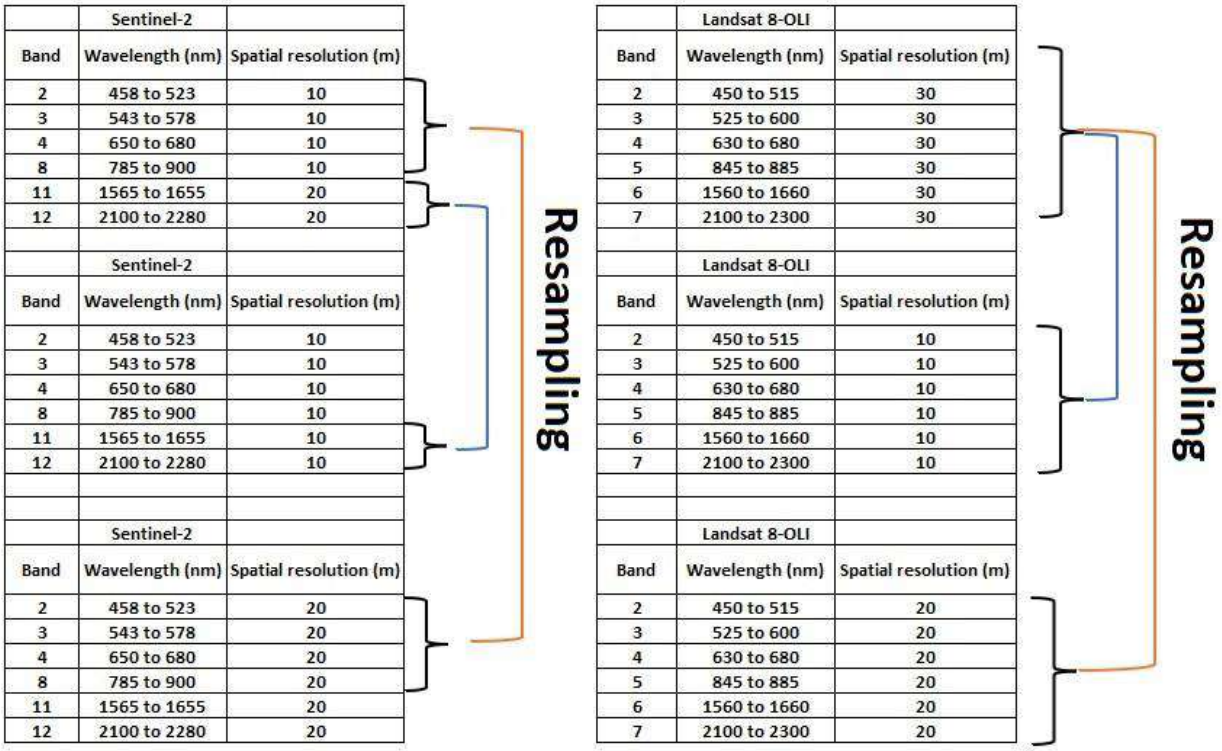


Figure SF2 illustrates the resampling of Sentinel 2 and Landsat 8 bands into spatial resolutions of 10 m and 20 m into four streams of auxiliary datasets used in the prediction of Cd in agricultural soil using legacy datasets and preferential sampling.

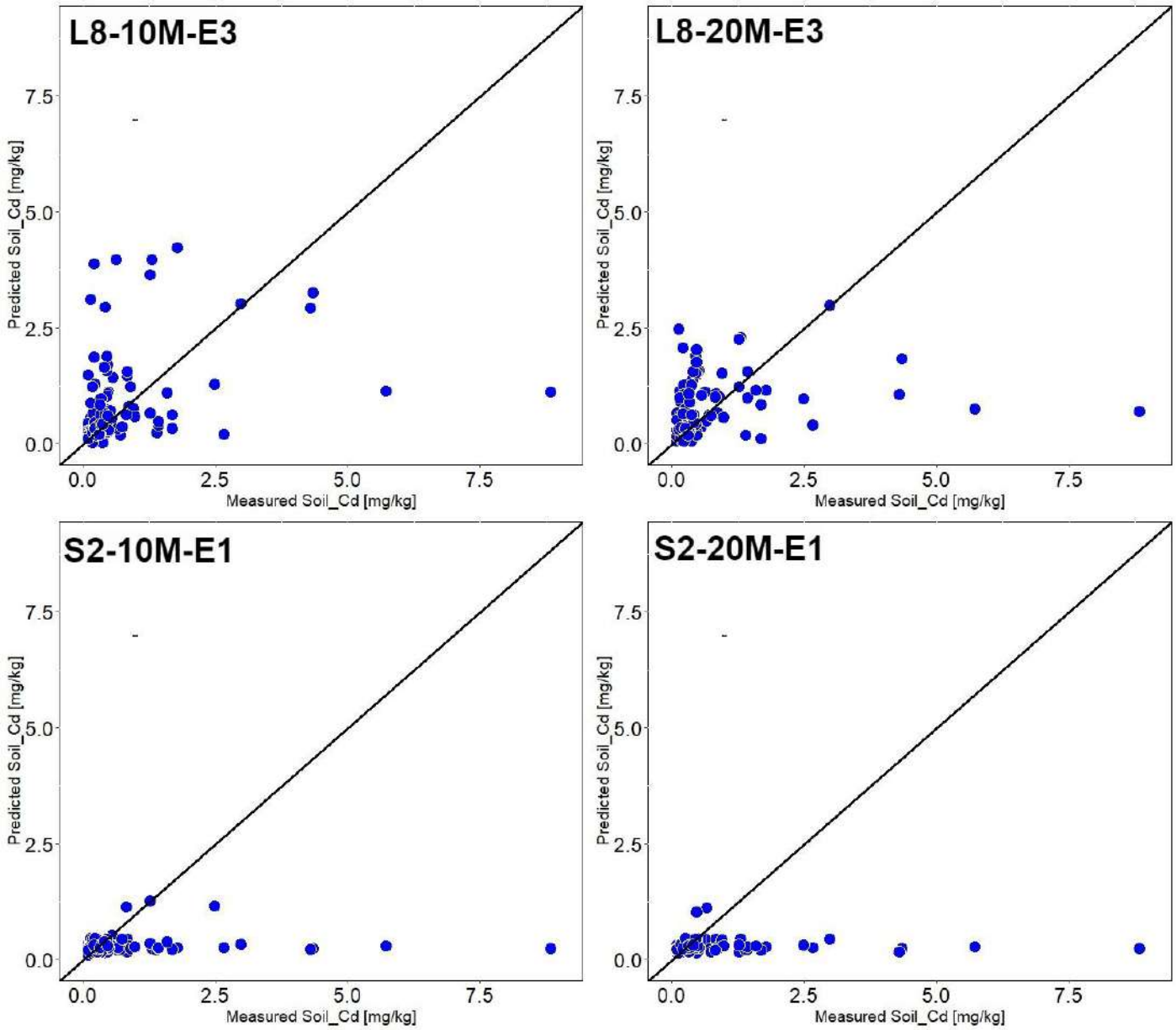


Figure SF3 depicts a scatter plot of measured and predicted cadmium concentrations for the legacy dataset using remote sensing datasets with 10m and 20m spatial resolution as auxiliary data (S2 represents sentinel 2 and L8 represent Landsat 8).

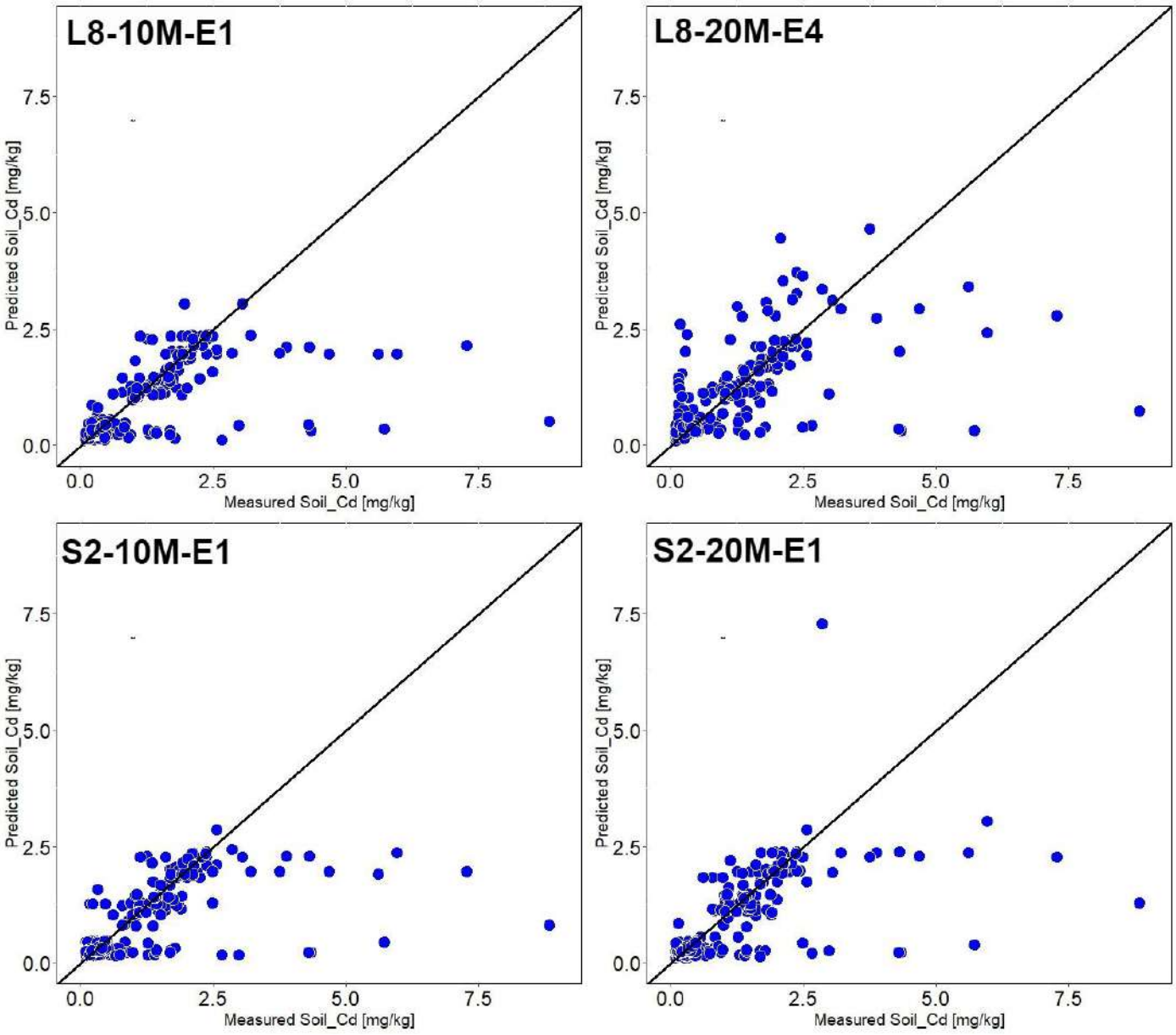


Figure SF4 illustrates a scatter plot of measured and predicted cadmium concentrations for the preferentially sampled-legacy dataset using remote sensing datasets with 10m and 20m spatial resolution as auxiliary data (S2 represents sentinel 2 and L8 represent Landsat 8).

Credit Author Statement

Prince Chapman Agyeman: Conceptualization, Methodology, Writing- Original draft preparation, Analysis**,** Visualization. **Luboš Borůvka**: Supervision, Editing.

Ndiye Michael Kebonye: software, Data curation. **Vahid Khosravi:** Data curation, Editing and Investigation. **Kingsley JOHN:** Software, Editing, Visualization. **Ondrej Drabek**: Data Curation and Visualization. **Vaclav Tejnecky** Editing, Analysis.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: