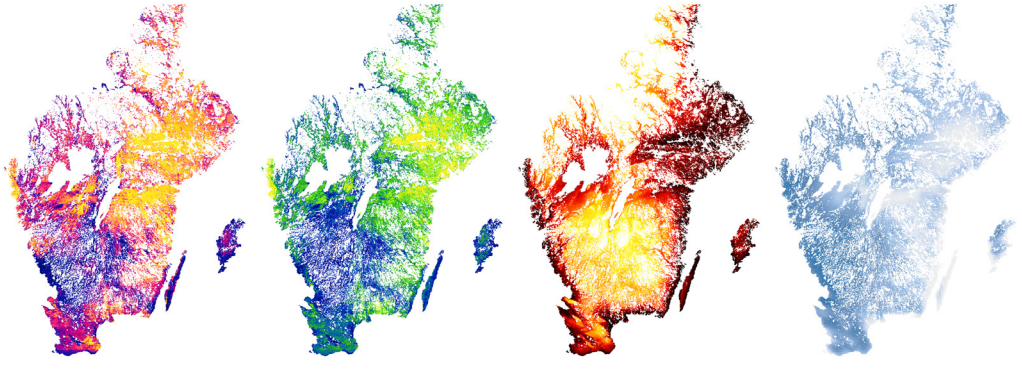
Acta Universitatis Agriculturae Sueciae

Doctoral Thesis No. 2022:33
Faculty of Natural Resources and Agricultural Sciences

Digital soil mapping and portable X-ray fluorescence prediction of cadmium, copper and zinc concentrations as decision support for crop production

Karl Adler



Digital soil mapping and portable X-ray fluorescence prediction of cadmium, copper and zinc concentrations as decision support for crop production

Karl Adler
Faculty of Natural Resources and Agricultural Sciences
Department of Soil and Environment
Skara

SLU
SWEDISH UNIVERSITY OF AGRICULTURAL SCIENCES

DOCTORAL THESIS
Skara 2022

Acta Universitatis Agriculturae Sueciae
2022:33

Cover: A selection of four spatially extensive environmental covariates used in this thesis for digital soil mapping of copper concentration in agricultural topsoil. These cover southern Sweden. From left to right: Measured soil uranium (mg kg^{-1}), measured soil thorium (mg kg^{-1}), elevation above sea level (m) and annual precipitation (mm). Different colours were used for artistic purposes and to showcase how environmental properties vary in geographical space. Changes in colour refer to changes in values.

Creator: K. Adler, using the Python programming language and the package Datashader.

ISSN 1652-6880
ISBN (print version) 978-91-7760-941-4
ISBN (electronic version) 978-91-7760-942-1

Skara
Print: SLU Service/Repro, Uppsala 2022

Digital soil mapping and portable X-ray fluorescence prediction of cadmium, copper and zinc concentrations as decision support for crop production

Abstract

Trace element concentrations in agricultural soil are important for crop production. Certain trace elements, e.g. copper (Cu) and zinc (Zn), are essential for crops to complete their life cycle. Other trace elements, e.g. cadmium (Cd), can be harmful to crops or the end-consumer. Hence, it is important to have maps of soil concentrations of trace elements or methods for determining concentrations in soil samples. This thesis investigated the possibility of predicting trace element concentrations (Zn, Cu, Cd) in soil samples using portable X-ray fluorescence (PXRF) measurements. It also examined usefulness of digital soil mapping (DSM) to create maps of Cu and Cd concentrations in agricultural topsoil in Sweden.

Portable X-ray fluorescence models were validated at national and farm level. Predicted Zn concentrations were found to be comparable to those obtained in conventional laboratory analysis, while predicted Cd and Cu concentrations were less accurate. The most accurate PXRF models were created using non-linear machine learning algorithms, e.g. random forest.

Digital soil mapping of Cd concentrations in Skåne County, combined with data from grain sampling, revealed that low Cd concentrations in winter wheat grain were associated with predicted low concentrations in soil. The map could thus be used to identify arable soils suitable for producing winter wheat for products with strict quality criteria, e.g. baby food. Digital soil mapping of Cu concentrations at national level revealed that 47% of arable soils are highly likely not at risk of Cu deficiency. Covariate importance analysis indicated importance of airborne gamma radiation measurement data in DSM of Cu and Cd concentrations.

Keywords: Crop production, machine learning, trace elements, winter wheat, PXRF.

Author’s address: Karl Adler, Swedish University of Agricultural Sciences, Department of Soil and Environment, Skara, Sweden.

Digital markkartering och prediktion med portabel röntgenfluorescens av kadmium-, koppar- och zinkkoncentrationer som beslutsunderlag för produktion av grödor

Sammanfattning

Vissa spårelement är nödvändiga för att en gröda ska kunna växa och fungera normalt, t.ex. koppar (Cu) och zink (Zn). Andra kan vara skadliga för grödan eller konsumenten, t.ex. kadmium (Cd). Därför är det viktigt att ha kartor över dessa spårelements halter i åkermark samt metoder för att bestämma halter i jordprover. Denna avhandling utforskade om mätningar med portabel röntgenflourescens (PXRF)-teknik kan användas för att prediktera halter av Zn, Cu och Cd i jordprover. Vidare genomfördes digital markkartering (digital soil mapping; DSM), för att skapa kartor över halter av Cu och Cd i matjord.

Portabel röntgenfluorescens-modeller validerades på nationell nivå och gårdsnivå. Prediktioner av Zn-halter var jämförbara med konventionell laboratorieanalys, medan prediktioner av Cd- och Cu-halter var mindre träffsäkra. Resultat visade även att PXRF-modeller baserade på icke-linjära maskininlärningsalgoritmer, t.ex. random forest presterade bäst.

Digital markkartering av Cd-halter i Skåne län tillsammans med data från grödprover visade att låga halter av Cd i höstvetekärna var associerade med låga predikterade halter av Cd i jord. Kartan kan därmed användas för att hitta områden som är särskilt lämpliga för produktion av höstvete med särskilda kvalitetskrav för t.ex. barnmat. Digital markkartering av Cu-halter på nationell nivå visade att åtminstone 47% av svensk åkermark sannolikt inte har risk för kopparbrist. Gammastrålningsdata från flygmätningar var mycket viktiga hjälpvariabler vid digital markkartering av Cu och Cd-halter.

Nyckelord: Växtproduktion, maskininlärning, spårelement, höstvete, PXRF.

Författarens adress: Karl Adler, Sveriges lantbruksuniversitet, inst. mark och miljö, Skara, Sverige.

Dedication

Till Ebba, mamma och pappa

Contents

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I. Adler, K., Piikki, K., Söderström, M., Eriksson, J., & Alshihabi, O. (2020). Predictions of Cu, Zn, and Cd concentrations in soil using portable X-ray fluorescence measurements. Sensors, 20 (2), pp. 474

II. Adler, K., Piikki, K., Söderström, M., Pettersson, C.G., & Eriksson, J. (2022). Digital soil mapping of cadmium: Identifying arable land for producing winter wheat with low concentrations of cadmium. (Submitted)

III. Adler, K., Piikki, K., Söderström, M., & Eriksson, J. (2022). Digital soil mapping of copper in Sweden: Using the prediction and uncertainty as decision support in crop micronutrient management. (Submitted)

The contribution of Karl Adler to the papers included in this thesis was as follows:

I. Planned the study together with K. Piikki and M. Söderström. Analysed the soil samples together with O. Alshihabi, using the PXRF device. Wrote the code, analysed the results and wrote the original manuscript draft. Edited and revised the manuscript based on feedback and with support from the co-authors.

II. Planned the study together with K. Piikki and M. Söderström. Analysed the soil samples together with O. Alshihabi, using the PXRF device. Wrote the code, analysed the results and wrote the original manuscript draft. Edited and revised the manuscript based on feedback and with support from the co-authors.

III. Planned the study together with K. Piikki and M. Söderström. Analysed the soil samples together with O. Alshihabi, using the PXRF device. Wrote the code, analysed the results and wrote the original manuscript draft. Edited and revised the manuscript based on feedback and with support from the co-authors.

List of tables

List of figures

Abbreviations

DEM	Digital elevation model
DSM	Digital soil mapping
E	Nash-Sutcliffe model efficiency coefficient
GBR	Gradient boosting regression
LOD	Limit of detection
MAE	Mean absolute error
MARS	Multivariate adaptive regression splines
MLR	Multiple linear regression
NaN	Not a number
PICP	Prediction interval coverage probability
PXRF	Portable X-ray fluorescence
RF	Random forest regression

1. Introduction

Soil, a vital resource, is a product of geology, time, topography/relief, climate, geographical location and biology (Jenny, 1941). Soil as a commodity is finite and its composition varies in geographical space. This variability provides different conditions for agricultural production. Acknowledging, managing and adapting to this variability is important in modern agriculture in order to maximise yields and minimise environmental impacts. One promising management framework is precision agriculture, which is, according to the International Society of Precision Agriculture (ISPA), defined as:

Precision Agriculture is a management strategy that gathers, processes and analyzes temporal, spatial and individual data and combines it with other information to support management decisions according to estimated variability for improved resource use efficiency, productivity, quality, profitability and sustainability of agricultural production. (ISPA, 2020).

In order to implement this management strategy, spatially explicit information about the soil is obviously needed. Conventionally, this information is obtained by soil sampling and subsequent chemical analyses performed in the laboratory to provide data on e.g. soil texture or organic matter content (Viscarra Rossel et al., 2011). However, this conventional method can be time-consuming and expensive, making it difficult to apply in extensive mapping of agricultural fields (Viscarra Rossel & McBratney, 1998; Gholizadeh & Kopačová, 2019). When the intention is to perform soil mapping, it is preferable to utilise data from sources that are spatially extensive or less time-consuming and expensive than classical laboratory measurements.

Remote sensing and proximal sensing are becoming increasingly popular complements to conventional laboratory analysis (Mulder et al., 2011). Remote sensing involves measurements by various sensors mounted on e.g. satellites, airplanes or unmanned aerial vehicles. Proximal sensing involves measuring in close proximity (<2 m) to the object in question, using either a hand-held device or on-the-go using a field vehicle (Adamchuk & Viscarra Rossel, 2010). Both methods make it possible to gather spatially extensive information about soil, albeit at different geographical levels.

Hand-held portable X-ray fluorescence (PXRF) devices are gaining increasing popularity as a proximal soil sensing method because they measure fast and non-destructively and are cheap to use (Lemière, 2018). These devices have been proven to be flexible, as they can be used to measure soil *in situ* or in a more controlled laboratory setting *ex situ* (Hu et al., 2014; Weindorf et al., 2014).

The concentration of a trace element is one of the soil properties of interest in crop production. Certain trace elements, such as boron (B), zinc (Zn), manganese (Mn), copper (Cu) and molybdenum (Mo), are essential for plant growth and functioning (Fageria et al., 2002). Other trace elements, such as cadmium (Cd) and mercury (Hg), have no known positive effects and can be harmful for the plant or the consumer of the plant (Smolders & Mertens, 2013; Steinnes, 2013). Hence, knowledge about trace element concentrations in soil can be important in terms of sustainable crop production.

Future crop production using precision agriculture will most likely require decision support in map format. Digital soil mapping (DSM), or predictive soil mapping, is a method that combines measurements on soil samples with environmental covariates gathered from proximal or remote sensing to make maps (McBratney et al., 2003; Scull et al., 2003). In the past, DSM was conducted by interpolating the space between soil samples using geostatistical methods, e.g. various kriging methods (Burgess & Webster, 1980; Minasny & McBratney, 2016). Nowadays, DSM together with machine learning has become a more frequently used method, due to advances in machine learning algorithms and in the processing power of computers and greater availability of data (Arrouays et al., 2020a; Piikki et al., 2021; Wadoux et al., 2021a).

2. Aim and objectives

2.1 Overall aim

The overall aim of this thesis was (i) to use and assess the applicability of PXRF measurements in predicting Zn, Cu and Cd concentrations in agricultural soil and (ii) to create, evaluate and interpret Cu and Cd maps at different geographical levels.

2.2 Specific objectives

Specific objectives were:

1. To evaluate if and how PXRF measurements can be used to predict trace element concentrations (Papers I-III).
2. To use PXRF measurements to predict trace element concentrations in soil samples not analysed by wet chemistry in the laboratory, to create a large DSM calibration dataset (Papers II & III).
3. To create and evaluate DSM models of trace element concentrations in Swedish agricultural soil at national, regional and farm level, using various environmental covariates (Papers II & III).
4. To identify environmental covariates that are important for DSM of trace elements (Papers II & III).
5. To assess and discuss the applicability of DSM models and maps of trace element concentrations as decision support in crop production (Papers II & III).

3. Background

3.1 Trace elements in Swedish agriculture

3.1.1 Trace elements and agriculture

Trace elements are found in low concentrations in nature, e.g. in soil. Trace elements can be essential or non-essential to crops and humans, and can therefore be beneficial, non-essential or even harmful for crop development and for the end-consumer of the crop (Oorts, 2013; Smolders & Mertens, 2013). Some essential trace elements in crop nutrition, e.g. Cu and Zn, are referred to as micronutrients. Non-essential trace elements in crop nutrition, e.g. Cd, are potentially toxic to both crops and humans (Smolders & Mertens, 2013). Further, if the concentrations present in soil are sufficiently high, trace elements regarded as essential can also be toxic to plants, leading to malformations in roots, stem and leaves (Broadley et al., 2007; Alloway, 2013; Adrees et al., 2015). For example, soil Cu concentrations above 100 mg kg^{-1} can be regarded as potentially toxic to crops (Ballabio et al., 2018).

Mapping and assessing every trace element in agricultural soil that may be important in crop nutrition would be a massive task. Countries often have a set of trace elements that need special attention in crop production, depending on crops grown, management practices and environmental factors. For example, in Mediterranean countries Cu can occur in toxic concentrations, mainly due to fungicide application in vineyards (Ballabio et al., 2018), while countries south of the Sahel can have Zn deficiency due to highly weathered soils (Alloway, 2009). Hence, the trace elements deemed important or not are site-dependent.

In Sweden, critical trace elements for crop production are outlined in the annual report "*Recommendations for fertilizing and liming*" issued by the

Swedish Board of Agriculture. The trace elements covered are B, Cu, Mn and Zn, which at some sites in Sweden occur in sufficiently low soil concentrations to cause deficiency in crops.

The plant-available concentration of a particular trace element is the most important parameter, but information on the total or pseudo-total concentration in soil is commonly more available and it is the next best parameter. In Sweden, soil is regarded as being at risk of Cu deficiency if the pseudo-total Cu concentration is below 7 mg kg^{-1} (Swedish Board of Agriculture, 2020). Pseudo-total concentrations refer to near-total concentrations.

While this risk limit has been established for Cu, there is no Swedish risk limit for Zn. In addition, the fact that some essential trace elements can be toxic at high concentrations means that there are other limits to consider. For example, sewage sludge may not be applied to agricultural soil in Sweden if the topsoil has pseudo-total concentrations above 40 mg Cu kg^{-1}, 100-150 mg Zn kg^{-1} and 0.4 mg Cd kg^{-1} (Swedish Environmental Protection Agency, 1998). Risk limits also change over time, based on research, observations or discussions. For example, the risk limit for Cu was set at 6-8 mg kg^{-1} for a long time, but then changed to 6-7 mg kg^{-1} in the 2022 rendition of the annual report (Swedish Board of Agriculture, 2021).

Limits for Cd concentrations mainly relate to the concentrations in grain or the resulting food product. At European level, maximum permissible levels of Cd have been established for different food products, e.g. cereal-based baby food may not contain more than 40 µg Cd kg^{-1} wet weight (European Commission, 2021). The Swedish Food Agency monitors Cd concentrations in foodstuffs available in Sweden and organises a yearly forum with participants from universities, industry and government agencies with the aim of reducing concentrations in foodstuffs.

3.1.2 Soil concentrations in Sweden and in Europe

Since the mid-1990s, Sweden has had an ongoing survey of agricultural topsoils organised by the Swedish University of Agricultural Sciences and funded by the Swedish Environmental Protection Agency (Eriksson, 2021). Results from this monitoring programme show that concentrations of the trace elements of interest in this thesis, i.e. Cu, Zn and Cd, have not changed significantly over the survey period. Hence, the concentrations of these trace

elements in Swedish agricultural soils can be regarded as quite stable at present.

Descriptive statistics on concentrations in Swedish and European agricultural soil indicate that soil Cu concentrations are slightly lower in Sweden (Table 1). According to Eriksson et al. (2017), 22% of Swedish agricultural land is below the deficiency risk limit for Cu (7 mg kg^{-1}). Hence, risk of Cu deficiency is not uncommon in Sweden. Concentrations of Zn are generally higher in Swedish agricultural soils (Table 1). Cadmium concentrations in agricultural soils are at more or less the same level in Sweden and Europe (Table 1).

Table 1: Minimum, 25th percentile, median, 75th percentile and maximum concentration (mg kg^{-1}) of copper (Cu), zinc (Zn) and cadmium (Cd) in agricultural soils in Europe and Sweden. European values obtained by Reimann et al. (2014a, 2014b) using aqua regia digestion (n = 2108). Swedish values obtained by Eriksson (2021) using nitric acid (HNO_3) digestion (n = 2029 for Cu and Cd, = 2028 for Zn).

Trace element	Statistic	Sweden	Europe
Copper	Minimum	2.0	0.3
	25th	7.8	8.3
	Median	12	15
	75th	20	24
	Maximum	190	395
Zinc	Minimum	5.0	2.8
	25th	37	27
	Median	55	45
	75th	77	65
	Maximum	560	1396
Cadmium	Minimum	0.04	<0.01
	25th	0.13	0.11
	Median	0.18	0.18
	75th	0.25	0.28
	Maximum	4.1	7.5

3.2 Portable X-ray fluorescence

3.2.1 History and fundamentals

The PXRF device can trace its roots back to the stationary X-ray fluorescence (XRF) device first introduced in the 1900s (Glanzman & Closs, 2007). Stationary XRF devices still exist today, and the major difference between the two is that PXRF devices have less available power due to their portability. The introduction of PXRF devices in the late 1970s was made possible by battery, microprocessor and software innovations (Glanzman & Closs, 2007; Weindorf et al., 2014). The X-ray fluorescence method involves exciting an atom and its electrons with X-rays, often emitted from an X-ray tube in the case of modern PXRF devices (Kalnicky & Singhvi, 2001). This source produces specific wavelengths from the X-ray part of the electromagnetic spectrum (Figure 1) that eject an inner-shell electron from an atom (Figure 2). The vacant spot left by the ejected electron is then filled by an outer-shell electron, and this change releases fluorescent energy in the X-ray region of the electromagnetic spectrum in an amount corresponding to the energy difference of the electron shells (Kalnicky & Singhvi, 2001) (Figure 2).

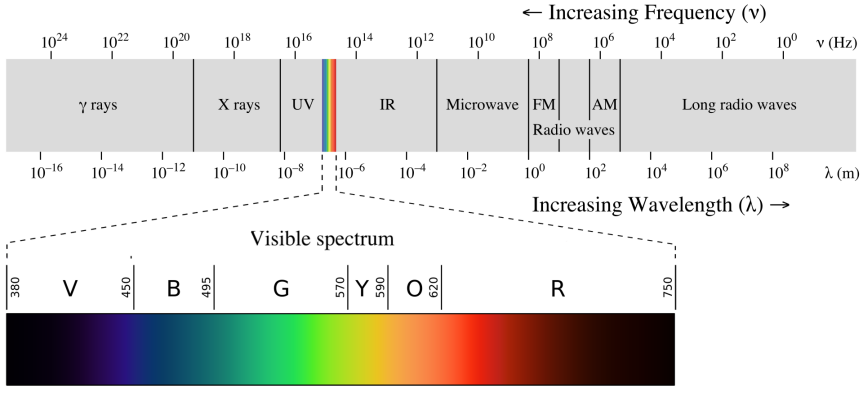


Figure 1: The electromagnetic spectrum and visible part (380 nm to 750nm). The X-ray part of the spectrum is of most interest for portable X-ray fluorescence (PXRF) devices. Creator: Philip Ronan, Gringer. Used under the creative commons licence (https://creativecommons.org/licenses/by-sa/3.0/).

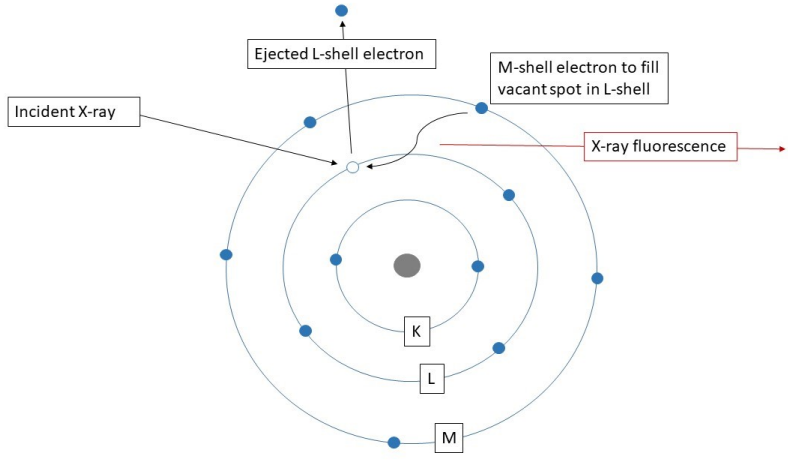


Figure 2: Illustration on how X-ray fluorescence works, from initial excitation of the electron to the end-product of X-ray fluorescence (left to right). Note that this example atom is solely for illustration purposes. K, L and M refer to different electron shells.

Hence, it is problematic to measure elements with few electron shells, e.g. elements lighter than magnesium (Mg) (Kalnicky & Singhvi, 2001; Lemière, 2018). Fluorescence is characterised by having lower outgoing energy, i.e. longer wavelength, than the incoming energy from the radiation source (Weindorf et al., 2014). The range of elements that can be measured depends on the PXRF device in question and how it is used, e.g. using vacuum or helium to minimise the attenuation effect of air on X-ray energy (Lemière, 2018).

The fluorescence wavelength is dependent on (i) the X-ray source, (ii) the element in question and (iii) the shell from which the electron was ejected and the shell from which the replacing electron originated. For instance, if the ejected electron originated from the K-shell and was replaced by an electron from the M-shell, the fluorescence is termed Kβ, with an associated wavelength. If the replacement electron originated from the L-shell then the fluorescence is termed Kα, with its specific fluorescence wavelength. Hence, an element can have multiple fluorescence peaks across the X-ray region in the electromagnetic spectrum that need to be accounted for (Kalnicky & Singhvi, 2001; Glanzman & Closs, 2007). The fluorescence wavelengths and their intensity provide information about the element present and its concentration in the sample (Weindorf et al., 2014).

When using a PXRF device, a model is used to convert these wavelength peaks into concentrations of trace elements. What is returned by the PXRF device is a spectrum with counts in each wavelength, dependent on the spectral resolution of the device. The model can either be supplied by the manufacturer of the PXRF device or created by the user using the raw PXRF spectra, together with chemometric methods (O'Rourke et al., 2016; Shresta et al., 2022).

3.2.2 Accuracy and limit of detection

According to United States Environmental Protection Agency (US EPA) method 6200, PXRF measurements should be confirmed against measurements of trace element concentrations deriving from conventional wet chemistry (US EPA, 2007). Comparisons should be made against reference samples that have accredited and accurate data on total or pseudo-total element concentrations from laboratory analyses, e.g. extraction with nitric acid (HNO_3) and analysis using inductively coupled plasma mass spectrometry (ICP-MS). Hence, during measurement with a PXRF device, reference samples are used to determine the accuracy of the measurements. It is common practice to report the recovery rates for measured elements (Weindorf & Chakraborty, 2020). The recovery rate shows how close the measured PXRF concentration is to the known quantity in a reference sample. For example, if the reference sample concentration is 100 mg Cu kg^{-1} and the PXRF measurement is 110 mg Cu kg^{-1}, this means that the corresponding recovery rate is 110% and that the PXRF concentration is an overestimate of the known concentration.

The accuracy of PXRF measurements is also very dependent on soil composition, i.e. soil moisture, particle size and organic matter content (Rouillon & Taylor, 2016; Ravansari & Lemke, 2018; Padilla et al., 2019). Sample preparation steps such as sieving, homogenising and drying often produce more accurate PXRF measurements (Goff et al., 2020).

The limit of detection (LOD) is an important concept in PXRF methodology. When using a PXRF device, measurements are commonly taken each second, e.g. for a total of 180 seconds. This results in a mean value for the 180 measurements and an associated standard deviation of the element concentration in question. This final reported concentration needs to be greater than three times the reported standard deviation to be above the LOD (Weindorf et al., 2012; Rouillon & Taylor, 2016). If the reported

measured concentration does not meet this criterion, then that measurement is below the LOD for the element and deemed too uncertain to be used.

3.3 Digital soil mapping and machine learning

3.3.1 Digital soil mapping

The underlying concept in DSM is to build mathematical models to predict soil observations using spatially extensive environmental covariates with the aid of computers (Minasny & McBratney, 2016). These covariates should be represented in the *scorpan* conceptual model (McBratney et al., 2003):

$$S = f(s, c, o, r, p, a, n) \quad eq. 1$$

where S is soil property or class, which in turn is a function of soil (s), climate (c), organisms (o), topography/relief (r), parent material (p), age (a) and spatial position (n).

The conceptual model is a variant of the existing *clorpt* model developed by Jenny (1941), but used more as a framework for prediction rather than an explanation of the soil environment (Malone et al., 2018). Examples of variable designation of spatially extensive covariates in the *scorpan* model are presented in Table 2.

Some types of covariates have a clear designation within the *scorpan* model, such as a digital elevation model (DEM) and the *scorpan* topographic/relief variable. It is difficult to find suitable covariates for some *scorpan* variables, such as age (Chen et al., 2022). However, a digital elevation model can provide information about the age of the soil, since the geomorphology of the environment is often a product of time (Grunwald, 2010; Chen et al., 2022). Hence, covariates can contain information that can be allocated to more than one *scorpan* variable.

Table 2: Common types of environmental covariates used in digital soil mapping (DSM) and the variable in the *scorpan* model that they represent.

Covariate type	***Scorpan* designation**
Digital elevation model	r, a
Remote sensing data	o, s, p
Proximal sensing data	o, s, p
Climate data	c, o
Coordinates	n

Digital soil mapping can be conducted at various geographical levels, depending on the geographical coverage of the covariates. The resulting map may cover a field, landscape, country, continent or the world. The spatial resolution of the covariates dictates the area of the pixels (raster) or the spacing between the grid points (vector) in the map. Some examples of geographical levels and their corresponding spatial resolution are presented in Table 3. As the examples in Table 3 show, DSM can be conducted on several geographical levels, with varying spatial resolutions. Generally, coarser spatial resolution is more common at larger geographical scales, with small geographical scales having finer spatial resolution (Minasny & McBratney, 2016; Piikki et al., 2021).

Table 3: Examples from the literature of geographical level, spatial resolution and predicted soil properties used in digital soil mapping (DSM).

Article	**Geographical level**	**Spatial resolution**	**Mapped property**
Ellili et al. (2019)	Landscape	10 m × 10 m	Soil organic carbon
Piikki & Söderström (2019)	Country	50 m × 50 m	i.a. clay content
Hengl et al. (2017)	Continent	250 m × 250 m	i.a. Cu, Mg and B
Guevara et al. (2018)	Continent	5 km × 5 km	Soil organic carbon
Poggio et al. (2021)	Global	250 m × 250 m	i.a. pH and nitrogen
Stockmann et al. (2015)	Global	1 km × 1 km	Soil organic carbon

3.3.2 Machine learning in digital soil mapping

In DSM, it is common to use machine learning techniques to map a variety of different soil properties at varying spatial resolution (Khaledian & Miller, 2020; Padarian et al., 2020). Machine learning is an umbrella term for computer algorithms that utilise data to build, i.e. calibrate or train, a prediction model, rather than creating a mechanistic model or having a model "hard coded" by a programmer (El Naqa, 2015). Machine learning thus encompasses many different algorithms, from e.g. simple linear regression to the more complex support vector machine regression. A popular algorithm is random forest and its different forms, which have been used to map e.g. soil pH globally (Poggio et al., 2021), soil organic carbon in the USA (Kim & Grunwald, 2016) and soil particle size fractions in Nigeria (Akpa et al., 2016).

There are no specific algorithms that are more suited than others for predictions of certain soil properties (Khaledian & Miller, 2020). Different machine learning algorithms excel at different tasks, depending on the covariates used and what is to be predicted. For example, random forest can handle linear and non-linear relationships in the data, while linear regression struggles with non-linear relationships (Hastie et al., 2009).

Recent advances in machine learning, increasing computing power and available covariates have had a positive effect in DSM, with e.g. more accurate maps (Minasny & McBratney, 2016; Wadoux et al., 2020). However, there has also been criticism about the use of machine learning in DSM. For example, Wadoux et al. (2021b) and Arrouays et al. (2020b) argue that pedological knowledge may be disregarded or lost with a machine learning framework. This can happen if a large number of covariates are used or can result from the low interpretability ('black box') of some machine learning models (Arrouays et al., 2020b; Khaledian & Miller, 2020). For instance, it can be problematic to identify factors that might influence soil pH if the resulting DSM model is a 'black box' calibrated with a large number of covariates. However, covariate importance methods and model interpretation can be used to hypothesise and unravel the 'black box' (Arrouays et al., 2020b; Wadoux & McBratney, 2021).

3.3.3 Digital soil mapping as decision support

Decision support exists in many forms. It can take the form of a complex computer system for farmers or policy-makers aimed at improving the quality of decisions (Zhai et al., 2020). It can also be a DSM product, such as a soil property map. When DSM is performed with the aim of producing a decision-making aid, it can be categorised as operational or practical DSM (Kidd et al., 2020). An example of an operational DSM product is the clay content map for Sweden created by Söderström et al. (2016) and refined by Piikki and Söderström (2019), which can be used to guide variable seed rate within an agricultural field. Another example is the Soil and Landscape Grid of Australia (SLGA) developed by Grundy et al. (2015) with information on 11 soil properties, such as bulk density at multiple depths. The Soil and Landscape Grid of Australia has been used by various stakeholders, and also by decision support system developers (Grundy et al., 2020; Kidd et al., 2020). Hence, DSM can be aimed towards some predefined problem or act as a foundation for decision support development.

3.3.4 Uncertainty in digital soil mapping

Quantification of uncertainty is often necessary to determine whether a map is suitable for its intended use (Heuvelink, 2014). Without quantification of uncertainty at every point, the end-user might believe that the prediction is the truth (Arrouays et al., 2020b). However, communicating, understanding and putting uncertainty to practical use can be complicated (Arrouays et al., 2017; Richer-de-Forges, 2019; Wadoux et al. 2021b).

The prediction interval is commonly estimated along with the predicted value (Heuvelink & Webster, 2022). A prediction interval is a range of values that will encompass all future observations, given a certain probability. Access to a predicted value and an accompanying prediction interval gives the end-user the possibility to assess the predicted value and its uncertainty. For example, assume that a predicted Cu concentration in soil at a point location is 7 mg kg^{-1}. At this location there is also an estimated 90% prediction interval, the lower bound of which is 4 mg Cu kg^{-1} and the upper bound of which is 10 mg Cu kg^{-1}. This information tells the end-user that the actual value at the point location lies with 90% probability between 4 mg Cu kg^{-1} and 10 mg Cu kg^{-1} (7 mg ± 3 mg Cu kg^{-1}).

The prediction interval can be estimated in two different ways. With the geostatistical method of kriging, the 90% prediction interval can be

determined by adding and subtracting 1.64 times the kriging standard deviation to the prediction, assuming normal distribution of values at the point (Heuvelink, 2014). Otherwise, e.g. with machine learning, the prediction interval is often related to percentiles of the empirical distribution function. For example, prediction of the 5^{th} and 95^{th} percentiles results in the 90% prediction interval, while calculating the 80% prediction interval would mean that the 10^{th} and 90^{th} percentiles have been predicted. The prediction interval needs to be validated (Szatmári & Pásztor, 2019). The prediction interval coverage probability (PICP) is often used to assess the validity of the prediction intervals in a map (Piikki et al., 2021). A PICP score shows how often true values are within the prediction interval (Shresta & Solomatine, 2006). The PICP score should be the same percentage as the prediction interval probability. A larger PICP score implies that the prediction interval is too wide, and a lower score implies that the interval is too narrow.

4. Materials and Methods

4.1 Soil samples and grain samples (Papers I-III)

4.1.1 National soil samples

Topsoil samples and laboratory analysis data from two national soil surveys were used in this thesis (Table 4). The first set of material, denoted NV, consisted of soil samples and data from the Swedish monitoring programme for arable soils funded by the Swedish Environmental Protection Agency (Eriksson et al., 2021). The sampling density in that programme is roughly one sample per 1300 ha of agricultural land and the dataset contains laboratory-analysed pseudo-total concentrations of Zn, Cd and Cu. These concentrations are determined using ICP-MS or inductively coupled plasma atomic emission spectroscopy (ICP-AES), after extraction using 7M HNO_3 in an autoclave at 120 °C for 30 minutes, according to Swedish standard SS 28311 (Swedish Institute for Standards, 2017). The NV data were used in Papers I-III for calibration and validation of the PXRF and DSM models.

The second set of material, denoted JV, consisted of soil samples from a sampling campaign of Swedish agricultural topsoil, funded by the Swedish Board of Agriculture (Swedish Board of Agriculture, 2015). Geographically, the JV set complement the NV set, with one sample taken roughly every 100 ha of agricultural land (Figure 3). The JV soil samples are not analysed for pseudo-total concentrations of Zn, Cd and Cu. The JV data were used in Papers II and III for calibration of the DSM models.

Table 4: Number of soil samples/data from the Swedish monitoring programme for arable soils (NV survey) and Swedish agricultural topsoil sampling (JV survey) used in each paper in this thesis.

Paper	NV	JV
I	1520	-
II	304	2097
III	1434	11 093

All soil samples in this thesis were composite samples consisting of nine subsamples taken at 0-20 cm depth within a 3 m radius from the selected sampling point. All soil samples were air-dried, homogenised and sieved (<2 mm).

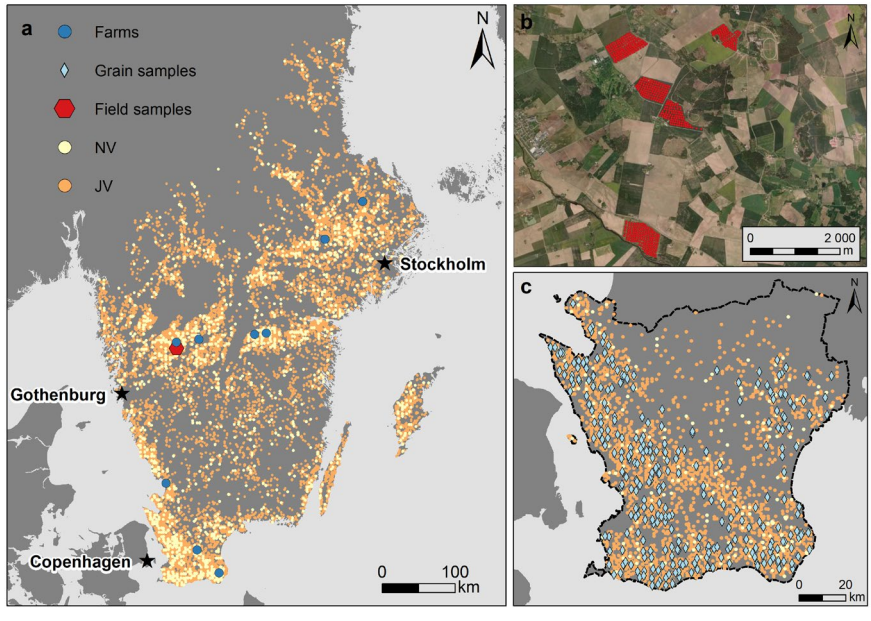


Figure 3: Overview map of (a) soil sampling locations in the Swedish monitoring programme for arable soils (NV) and Swedish agricultural topsoil sampling (JV), and location of the field soil samples Paper III, as well locations of the nine farms from Paper I. (b) Zoomed in map of the field soil samples from Bjertorp Farm used in Paper III, and (c) zoomed in map of Skåne County showing the NV and JV soil sample locations, and grain samples locations in Paper II. Legend applicable to panels (a-c). Large cities are marked as stars for spatial reference. Basemap in (b) courtesy of ESRI, Redlands, CA, USA.

The number of soil samples used differed between the papers, depending on the geographical extent of the investigation (Table 4). For example, the geographical spread of the NV soil sample sites extended across the whole of Sweden in Paper I, but only over the southern part in Paper III (Figure 3).

4.1.2 Field samples

Data on 179 soil samples from nine farms were used for validation of the PXRF models at farm scale (≈20 per farm) in Paper I (Figure 3a). Laboratory analysis was performed in the same way as for the NV dataset.

Data on soil samples from Bjertorp Farm, located north-east of Gothenburg (Figure 3b), were used for within-field validation of the Cu map produced (Paper III). These soil samples were collected from five different fields (25-47 ha), with roughly four samples per ha. All these soil samples were measured for pseudo-total Cu concentration, determined using ICP-AES after extraction with 2M HCl under heating in boiling water for two hours.

4.1.3 Grain samples

Data on grain samples were used to assess the relationship between soil Cd concentrations in the map produced and concentrations in winter wheat (*Triticum aestivum* L.) grain in the area covered by the map (Paper II). Data on 307 grain samples were used (Figure 3c). This set consisted of 196 samples collected in 1992 and 111 samples collected 2001-2007. Both sets had data on pseudo-total Cd concentrations, analysed after digestion using HNO_3. The samples from 1992 were digested at 135 °C for four hours and an additional hour at 100 °C. The samples from 2001-2007 were digested in a microwave oven at 120 °C for one hour. Determination was done with ICP-MS. There was a negligible difference in extracted Cd between the digestion methods. The grain data can be regarded as representative of the current situation (2022), since the Cd concentration in winter wheat grain has been stable over the past 29 years (Eriksson, 2021).

4.2 PXRF measurements (Papers I-III)

All soil samples in the JV and NV sets were subjected to PXRF measurements in the period 2018-2020. The measurements were made *ex situ*, using a Niton XL3t GOLDD+ PXRF device with a geometrically

optimised large area drift detector and a silver (Ag) anode of 50 kV and 200 µA (Thermo Scientific, Billerica, MA, USA). The PXRF device was mounted on a purpose-built static frame (Thermo Scientific, Billerica, MA, USA), and connected to a computer (Figure 4).

Each soil sample was pre-dried, homogenised and sieved (<2 mm), in compliance with recommendations for *ex situ* PXRF measurements (US EPA, 2007; Weindorf & Chakraborty, 2020). Each soil sample was measured for 180 s. Double-ended XRF sample cups, 32 mm diameter, with a 4 µm thick transparent polypropylene XRF film were used to contain the soil samples during measurement, following US EPA standards (U.S. EPA, 2007). The factory calibration "soil mode" was used to obtain concentrations in mg kg^{-1} from the fluorescence peaks.

The LOD was set at three times the standard deviation of the measurement, and trace element concentrations below this were set as Not a Number (NaN). Measured trace elements with <10% NaN values were used in Paper I as covariates in the PXRF models. This ensured that the trace elements used were commonly present above the LOD in Swedish soil. The resulting 13 trace elements were also used as covariates in the PXRF models in Papers II and III (Table 5).

Figure 4: Image of the portable X-ray fluorescence (PXRF) device used, shown mounted on its purpose-built frame with the lid closed.

During the measurement period, the reference standard 2709a from the National Institute of Standards and Technology (NIST) was subjected to PXRF measurements four times, to obtain the mean recovery rate and the standard deviation of the recovery rate, in order to check the measurement stability of the PXRF device (Table 5). As Table 5 shows, the concentrations measured by the PXRF device were relatively accurate, although some trace elements, such as caesium (Cs) and lead (Pb), were not accurately measured. However, the intention was not to measure concentrations of these trace elements with great accuracy, but rather to use the trace elements as covariates and their measured concentrations to predict concentrations of Cu, Zn and Cd. For this, the stability of the measurements was more important. For example, measured concentrations of Cs were suitable for modelling as they were consistently overestimated.

Table 5: The mean recovery rate from four measurements, and the standard deviation in recovery rate for each element used in the thesis.

Element	Recovery rate (%)	Recovery rate standard deviation (%)
Lead, Pb	63	10.8
Zirconium, Zr	65	0.9
Rubidium, Rb	83	0.8
Iron, Fe	84	0.6
Barium, Ba	87	2.4
Zinc, Zn	92	2.1
Strontium, Sr	92	0.8
Potassium, K	96	1.3
Manganese, Mn	97	2.7
Calcium, Ca	105	1.5
Titanium, Ti	114	1.8
Vanadium, V	123	18.6
Caesium, Cs	970	33.1

4.3 Machine learning algorithms (Papers I-III)

There are many machine learning algorithms to choose from, in several programming languages. In this thesis, the machine learning library Scikit-learn in the Python programming language was used (Pedregosa et al., 2018).

In Paper I, three different machine learning algorithms were chosen to create PXRF models. These were random forest regression (RF), multiple linear regression (MLR) and multivariate adaptive regression splines (MARS). These were selected because the aim was to use distinctly different models and MLR is linear, RF is non-linear and discrete, and MARS is non-linear and continuous. A discrete model cannot be used to extrapolate beyond the calibration data, while continuous model can be used extrapolate. The prediction made by an RF model is the mean of an ensemble of regression trees with bagging/bootstrapping, i.e. each regression tree is calibrated on a subset of the data (Breiman, 2001). A MARS model consists of several basis functions (piecewise linear regression models), that are created in a forward pass and later pruned to minimise overfitting in a backward pass (Hastie et al., 2009).

In Papers II and III, an ensemble of machine learning models was used, with the aim of combining their strengths. Therefore, a PXRF model consisted of several sub-models in an ensemble and the output of the PXRF model was the ensemble average, i.e. model averaging was applied. In model averaging, the output of several models is used e.g. to obtain a mean prediction (Hastie et al., 2009). In Paper II, the PXRF model consisted of a MARS model and an RF model. In Paper III, the PXRF model consisted of three extremely randomised tree (ERT) models and three gradient boosting regression (GBR) models, each with different hyperparameter values. An ERT model is similar to an RF model, but introduces a randomly chosen splitting threshold for each covariate and chooses the best one, while also omitting bagging/bootstrapping (Geurts et al., 2006). Gradient boosting regression works by fitting several shallow regression trees (stumps) in sequence in order to minimise a given loss function, such as least squares or quantile loss (Friedman, 2001). Providing full explanations and detailed descriptions of the algorithms was beyond of the scope of this thesis. I refer to Hastie et al. (2009) for more in-depth information.

4.4 Digital soil mapping (Papers II & III)

4.4.1 Calibration, algorithm, hyperparameter settings and application

The DSM models in Paper II and Paper III were calibrated using the NV dataset with laboratory measured concentrations and the JV dataset with concentrations predicted by a PXRF model. The JV and NV data also contained the environmental covariates used in each respective paper. The data on concentrations of Cu and Cd in the JV and NV datasets were thus the response variable, while the environmental covariates were the explanatory variables. The calibrated DSM models were then used to create maps of concentrations of Cu and Cd, as well as of the prediction interval.

Digital soil mapping was performed using the GBR algorithm for two reasons. First, using GBR makes it possible to predict specific percentiles, to create prediction intervals. Second, GBR can handle non-normally distributed data and non-linear dependencies as it is tree-based, just like RF.

Hyperparameters are the settings of a machine learning algorithm that are predefined by the user. In Paper II, testing showed improved performance with hyperparameter tuning. The hyperparameter values for tuning were chosen based on recommendations in Prettenhofer and Louppe (2014) and Elith et al. (2008). The hyperparameters and chosen values are presented in Table 6.

The prediction interval was created using GBR models calibrated based on quantile loss, where one GBR model was used to predict the 5^{th} percentile and another GBR model to predict the 95^{th} percentile. The 5^{th} percentile was the lower bound of the prediction interval, while the 95^{th} was the upper bound. The actual prediction was computed using a GBR model calibrated based on least squares loss.

Table 6: Description of hyperparameters frequently tuned for the Gradient Boosting Regression (GBR) algorithm in Paper II and values used after tuning.

Hyperparameter	Description	Settings used
Max depth	Maximum depth allowed in each tree	6
Max features	Fraction of covariates considered at each split	1.0
Learning rate	Shrinkage factor of each tree	0.011
Minimum samples	Samples needed for creating a leaf node	3
Subsampling	Fraction of samples used to construct each tree	0.6
Estimators	Number of trees	1000

In Paper III, hyperparameter tuning was omitted due to it being too computationally demanding, because of the large geographical extent of the study area. Instead, model averaging was done using three GBR models, but with different values of the most important hyperparameter for each sub-model. Initial testing revealed *subsampling* to be the most suitable hyperparameter to use, as the *max features* hyperparameter would tamper with the covariate importance. A total of three GBR models were calibrated, each with different settings of the hyperparameter subsampling (0.3, 0.6, 1.0). The mean of the output from the three GBR models was used as the prediction. This was done for the 5th percentile, mean and 95th percentile. A total of nine GBR models were thus calibrated. Lastly, the number of trees used in each GBR model was set to 500 instead of 1000, as testing revealed low performance improvement with increased number of trees.

4.4.2 Covariates

The foundation used for the DSM in Papers II and III was the covariate grid created by Piikki and Söderström (2019). The spatial resolution of this grid is 50 m × 50 m, it covers around 90% of Swedish agricultural land and it is restricted to non-organic soils (<20% organic matter), i.e. it has the same geographical coverage as the JV soil samples (see Figure 3). Covariates present in this grid are listed in Table 7.

Unique covariates used in Paper II were cokriged concentrations of Cd from biogeochemical data on Cd and Zn concentrations in the roots of sedges (*Carex* L.), meadowsweet (*Filipendula ulmaria* L.) and water moss (*Fontinalis* Hedw.) from small streams (Lax, 2009). Convergence index and topographic wetness index were computed from the DEM to obtain more topographic derivatives.

In Paper III, some covariates, such as cokriged biogeochemical data, convergence index and topographic wetness index, were omitted based on findings in Paper II or due to poor data availability. A 50 m × 50 m raster of soil moisture based on data originally reported by Ågren et al. (2021), resampled from 2 m × 2 m, was used. This raster provides nationwide soil moisture predictions, although mainly calibrated and validated on forest soils. A 4 km × 4 km grid of annual and seasonal climate data was obtained from the Swedish Meteorological and Hydrological Institute (SMHI, 2015).

The means for two reference periods provided in the dataset (1961-1990 and 1991-2013) were used.

Table 7: Environmental covariates, their type and source used for digital soil mapping in Papers II and III. Suffixes [II] and [III] indicate the paper in which each covariate was used. Original data source: * Geological Survey of Sweden, ** Lantmäteriet (Swedish Land Survery). DEM = digital elevation model, TPI = topographic position index.

Covariate	**Type**	**Source**
Uranium (^{238}U) (mg kg^{-1}) [II, III]	Gamma	Piikki & Söderström (2019)*
Thorium (^{232}Th) (mg kg^{-1}) [II, III]	Gamma	Piikki & Söderström (2019)*
Potassium (^{40}K) (%) [II, III]	Gamma	Piikki & Söderström (2019)*
Dose rate (nGy hr^{-1}) [III]	Gamma	Computed from U, Th and K
Topographic wetness index [II]	DEM	Computed from elevation
Convergence index [II]	DEM	Computed from elevation
TPI 5, 50 and 500 ha [II, III]	DEM	Piikki & Söderström (2019)*
Soil moisture [III]	DEM	Ågren et al. (2021)
Elevation (10 m × 10 m) (m) [II, III]	DEM	Piikki & Söderström (2019)**
Precipitation, annual (mm) [III]	Climate	SMHI (2015)
Precipitation, seasonal (MAM, JJA, SOM and DJF) (mm) [III]	Climate	SMHI (2015)
Temperature, annual (°C) [III]	Climate	SMHI (2015)
Temperature, seasonal (MAM, JJA, SOM and DJF) (°C) [III]	Climate	SMHI (2015)
Soil texture classes (Clay, Clay till, Till, Silt, Sand and Other) [II, III]	Soil texture	Piikki & Söderström (2019)*
Cokriged biogeochemical data [II]	Biogeochemical	Lax (2009)

Seasonal climate data were divided into spring (MAM; March, April and May), summer (JJA; June, July and August), autumn (SON; September, October and November) and winter (DJF; December, January and February). Lastly, dose rate was computed from airborne gamma radiation measurements, using the equation (Duval et al., 2005):

$$Dose\ rate = 13.2K + 5.48U + 2.72Th \quad eq.2$$

where *Dose rate* is in nGy hr^{-1}, *K* is measured potassium (^{40}K)concentration in %, *U* is measured uranium (^{238}U) concentration in mg kg^{-1} and *Th* is measured thorium (^{232}Th) concentration in mg kg^{-1}.

4.5 Validation

The validation metrics used were Nash-Sutcliffe model efficiency coefficient (Nash & Sutcliffe, 1970), denoted R^2 in Paper I and E in Paper II and III, and mean absolute error (MAE). An E value of 1 means that the model predicts perfectly, while an E value smaller than 0 means that using the prediction is less correct than using the mean value of the observations. An E value below 0 is possible. These metrics were mainly used to assess the performance in cross-validation and independent validation. Mean absolute error was chosen since it is less sensitive to outliers than the frequently used root mean squared error (RMSE). Squared Pearson correlation coefficient (r^2) from linear regression between predicted and measured concentrations was used in Paper III to assess whether the map could explain the variation within fields.

Precision and recall were used in Paper III to assess how accurately the DSM model predicted Cu concentrations in soil samples at or below the risk limit for Cu deficiency. Recall and precision was also calculated for other limits (1-60 mg kg^{-1}) to see the change in classifying performance with increasing concentration. A sample with a concentration at or below a set limit, e.g. 7 mg kg^{-1}, was a positive, and a sample with a concentration above the set limit was a negative. Recall shows the fraction of actual positives predicted as positives, precision shows the fraction of predicted positives that are actual positives.

In Paper I the classifying performance of the PXRF models were assessed using accuracy. Accuracy is the number of correctly predicted positives and negatives divided by the total number of predictions. This was done to see

how well the PXRF models could be used to correctly classify concentrations in soil samples below and above limits for sewage sludge application and Cu deficiency. In Paper I, the risk limit for Cu deficiency was set at 8 mg kg^{-1}, which was the upper bound of the risk limit range (6-8 mg kg^{-1}).

Cross-validation of the DSM models were done using five folds, and they were always validated against laboratory-analysed concentrations of Cu and Cd. Thus the DSM models were never validated against the PXRF-predicted concentrations in the JV dataset. For example, the calibration dataset of a cross-validation fold consisted of 80% of the NV dataset and the whole JV dataset. The remaining 20% of the NV dataset was then used for validation.

Cross-validation of the PXRF models were done using five folds in Paper II and III. Cross-validation of the PXRF models were done using leave-one-out in Paper I.

4.6 Covariate importance (Papers II & III)

Mean decrease in impurity (MDI) and permutation importance (PI) were used to assess the importance of covariates in the DSM models, since using two different methods minimises the risk of obtaining misleading results. Using covariate importance makes it possible to interpret the machine learning models and hypothesize, but it does not prove any causal relationship between covariates and a particular soil property.

Mean decrease in impurity is specific for tree-based algorithms and is included in the GBR algorithm provided by Scikit-learn. Mean decrease in impurity measures how many times a covariate is used for a split in the nodes of the regression trees and its hierarchy within the regression tree (Breiman, 2001). For example, a covariate used early for a split in the regression tree hierarchy is likely to be more important, as it impacts all subsequent splits.

Permutation importance is 'model agnostic' and thus applicable on any model type, provided that the model has been calibrated. It works by establishing a reference performance score for the calibrated model against the calibration or independent validation data (the performance score used in this thesis was E). Next, a covariate is permuted, i.e. randomly shuffled, in the dataset, this dataset with a permuted covariate is used for prediction and the deviation between the reference score and the new score is noted (Breiman, 2001; Strobl et al., 2008). This is done a number of times for each covariate, in order to obtain a mean deviation from the reference score. An

important covariate will result in large deviation from the reference score, resulting in a high PI score. In this thesis, the number of times each covariate was permuted was 10. The permutation importance was calculated based on the calibration data.

Since the DSM model in Paper III was an ensemble of three GBR models, the results obtained for MDI and PI were thus a mean of three values. The scores obtained in Paper II were not means, since that DSM model consisted of one GBR model.

5. Results and Discussion

5.1 PXRF modelling

The cross-validation results in Paper I confirmed that PXRF measurements can be used for accurate prediction of Zn concentrations, producing values very similar to those obtained in laboratory analysis. However, Cu and Cd concentrations were predicted less accurately (Figure 5). The performance of the PXRF models in predicting Cu and Cd concentrations in Papers II and III was similar to that in Paper I (Table 8; Table 9). The non-linear models performed better than MLR in predicting Cu and Cd concentrations, especially at farm level. However, MLR was as accurate as the non-linear models in predicting Zn concentrations. Predictions of Zn were considered accurate enough to be used when assessing whether soils can receive sewage sludge. The predictions from the PXRF model were more accurate than the PXRF-measured Zn concentrations (Paper I).

Table 8: Nash-Sutcliffe model efficiency coefficient (E) from cross-validation of each portable X-ray fluorescence (PXRF) model and trace element reported in Papers I-III. E values in brackets are from validation at the farm level. Model types: MLR = multiple linear regression, RF = random forest, MARS = multivariate regression splines, Paper II ensemble = MARS and RF, Paper III ensemble = extremely randomized trees and gradient boosting regression.

Paper and model	Copper, Cu	Cadmium, Cd	Zinc, Zn
Paper I, MLR	0.58 (0.90)	0.49 (0.74)	0.92 (0.96)
Paper I, RF	0.63 (0.84)	0.48 (0.74)	0.86 (0.94)
Paper I, MARS	0.59 (0.94)	0.70 (0.80)	0.92 (0.97)
Paper II, Ensemble	-	0.82	-
Paper III, Ensemble	0.66	-	-

Accuracy scores when classifying samples with concentrations below and above limits for sewage sludge application and Cu deficiency were high (>80%). However, accuracy can give misleading results with imbalanced datasets. This was the case as there were, e.g., were very few samples with concentrations above 40 mg Cu kg^{-1}. This made the PXRF models appear overly optimistic of its overall classifying performance. For example, the PXRF model was very good at classifying samples with concentrations below the sewage sludge limit for Zn, and most of the samples had concentrations below this limit. However, the few samples with concentrations above the limit were not accurately classified. Accuracy scores less impacted by imbalanced datasets would probably have been more suitable for communicating the overall classifying performance of the PXRF models, e.g. using balanced accuracy.

Conditional bias was especially present when predicting Cu and Cd concentrations in Paper II and Paper III, as it was problematic to predict very low and very high concentrations. In Paper III, this bias permeated into the subsequent DSM model, which had problems predicting concentrations lower than 5 mg Cu kg^{-1}. The tree-based PXRF models had problems predicting concentrations below 5 mg Cu kg^{-1} (Papers I & III). However, predicting Cu concentrations using PXRF measurements made it possible to predict below the LOD of Cu of the device, which was ≈ 20 mg kg^{-1} (Paper I).

Table 9: Mean absolute error (MAE) from cross-validation of each portable X-ray fluorescence (PXRF) model and trace element reported in Papers I-III (mg kg^{-1}). MAE values in brackets are from validation at the farm level. Model types: MLR = multiple linear regression, RF = random forest, MARS = multivariate regression splines, Paper II ensemble = MARS and RF, Paper III ensemble = extremely randomized trees and gradient boosting regression

Paper and model	Copper, Cu	Cadmium, Cd	Zinc, Zn
Paper I, MLR	3.9 (4.4)	0.07 (0.12)	5.6 (4.4)
Paper I, RF	3.5 (4.5)	0.05 (0.11)	5.9 (5.4)
Paper I, MARS	3.7 (3.2)	0.05 (0.09)	5.6 (4.0)
Paper II, Ensemble	-	0.08	-
Paper III, Ensemble	3.3	-	-

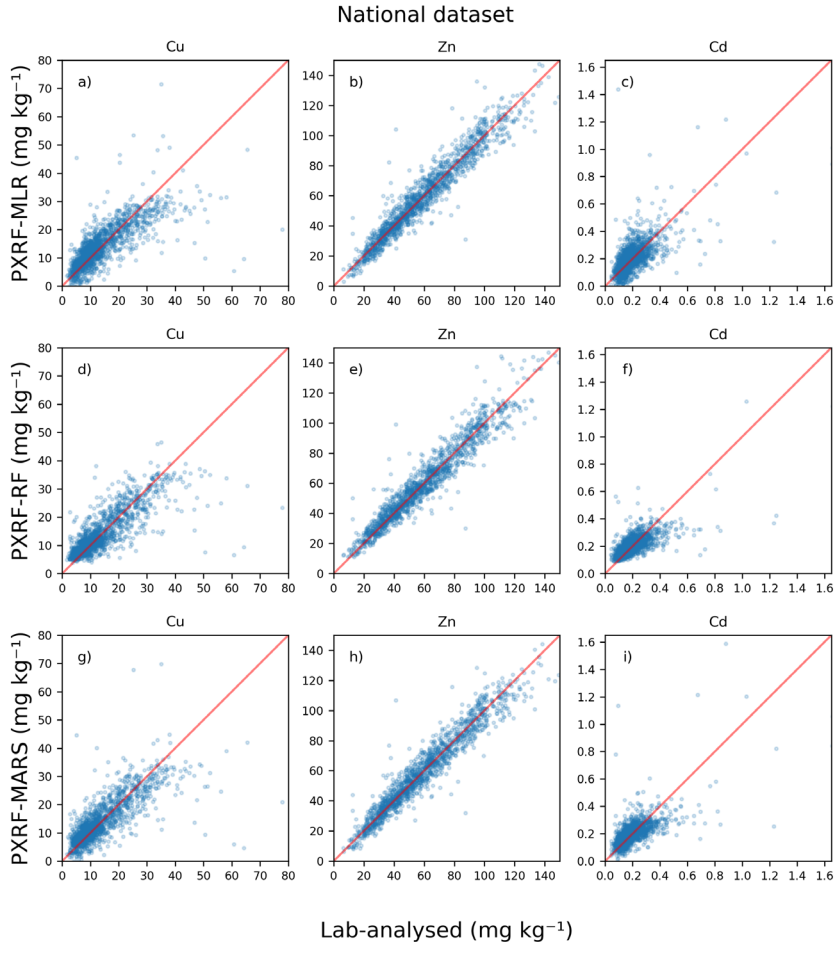


Figure 5: Cross-validation of the portable X-ray fluorescence (PXRF) models used in Paper I (n = 1520).

The PXRF model was used to predict Cu and Cd concentrations in a large proportion of the DSM calibration samples (Papers II & III). Thus, the biases present in the predictions of the PXRF models manifested to a certain degree in the DSM model predictions. This makes it tempting to create more range-specific models in the future or perhaps to implement bias correction. Accurately predicting the tails of the distribution, low and high

concentrations, is often more important in environmental applications (Belitz & Stackelberg, 2021). This was the case in this thesis, where models and maps were intended for identifying the risk of deficiency or the risk of exceeding limits. Future PXRF models and DSM models should perhaps be specifically aimed towards accurately predicting lower concentrations. This could perhaps be achieved by duplicating calibration samples with low concentrations, adding extra samples with low concentrations, i.e. spiking, only using calibration samples with low concentrations or using bias correction techniques as proposed by Song (2015), Sylvain et al. (2021) and Belitz and Stackelberg (2021). Another option for potentially improving performance of the PXRF model could be to utilise and calibrate it using the 'raw' spectrum, instead of the factory calibration measurements of trace elements. For example, an approach by Shresta et al. (2022) using PXRF spectra together with partial least squares support vector machine (PLS-SVM) produced more accurate predictions of Cd and Cu concentrations than was achieved in this thesis.

The conditional bias and overall weaker performance of PXRF models for Cu and Cd restricts their applicability as decision support for crop production. For example, the MAE reported in Paper III for predicted Cu concentrations was 3.3 mg kg^{-1}, a value that can be quite large when trying to determine whether a soil has a value below the risk limit of 7 mg kg^{-1}. Therefore, PXRF model predictions cannot be regarded as a direct replacement for conventional wet chemistry analysis, especially when certain limits are of interest. However, the speed and ease of PXRF measurements coupled with machine learning makes this an interesting complement to wet chemistry analysis.

5.2 PXRF predictions in DSM

One of the objectives in this thesis was to use PXRF models to predict concentrations of Cu and Cd in the JV soil samples, with the intention of using JV and NV data together as a calibration dataset. This substantially increased the size of the calibration dataset used for the DSM model in Papers II and III (see Table 4). The assumption was that a geographically denser calibration dataset with less accurate measurements was more important than a geographically sparser one with more accurate measurements. Some studies, such as those by Somarathna et al. (2017) and

Lai et al. (2021), have shown that increasing the number of samples for calibration can increase the accuracy and reduce the uncertainty of DSM models. However, increasing the number of samples will provide diminishing returns in terms of accuracy gains after a certain point. It should be noted that the two studies cited were on soil organic carbon and it cannot be assumed that this is also the case for DSM of trace elements.

In Paper III, it was revealed that the DSM model using both NV and JV as calibration data was more accurate (E = 0.58, MAE = 4.1 mg kg^{-1}) than the DSM model only using NV as calibration data (E = 0.40, MAE = 4.6 mg kg^{-1}). The reason for the better accuracy could be that more geographical variation was captured with the geographically denser calibration dataset. The use of predictions from PXRF measurements as calibration data for DSM is promising. The method is fast and provides measurements that can be used in a variety of ways to predict different soil properties. However, further investigation is needed to confirm whether a geographically denser calibration dataset is better than a more accurate, but sparser, dataset.

5.3 DSM modelling

Cross-validation revealed that the DSM models were less accurate than the corresponding PXRF models (Papers II & III) (Table 10, see also Table 8 and Table 9). This is probably because PXRF measurements were made directly on the soil, while the covariates used in the DSM model are indirect measurements. The accuracy of the DSM models, when compared against results from laboratory analyses, indicated that the maps produced should be regarded more as exploratory tools. For example, in Paper III the recall score for correctly predicting if a soil concentration was at or below the risk limit for Cu deficiency was around 0.08, i.e. 8% correctly classified.

Table 10: Cross-validation results of the digital soil mapping (DSM) models used in Papers II and III to predict cadmium (Cd) and copper (Cu) concentrations, respectively. E = Nash-Sutcliffe model efficiency coefficient, MAE = mean absolute error.

Paper and element	E	MAE (mg kg^{-1})
Paper II, Cd	0.69	0.11
Paper III, Cu	0.58	4.1

This means that the resulting map of Cu concentrations cannot be used alone to delineate agricultural land at risk. However, the precision was much higher, around 0.6. This precision score indicates that when the DSM model actually predicts a soil concentration to be at or below the risk limit, it is correct with 60% probability (Paper III). Increasing the risk limit to 8 mg kg^{-1} or higher made the recall and precision increase rapidly (Figure 6). For example, if the risk limit were to be reformulated to 10 mg kg^{-1}, the recall and precision score would be around 0.6 and 0.8, respectively (Figure 6). This means that the DSM model would correctly predict soil concentrations of Cu to be at or below 10 mg kg^{-1} in 60% of cases, and that a predicted concentration at or below 10 mg kg^{-1} would be correct with 80% probability (Figure 6). However, correctly predicting the concentration below the actual risk limit, 7 mg kg^{-1}, was problematic in this case as shown earlier.

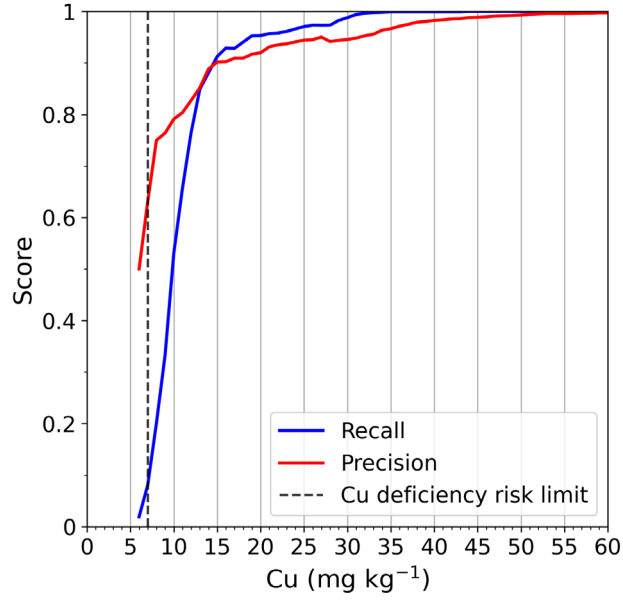


Figure 6: Recall and precision scores when predicting at or below concentrations in soil samples during cross-validation of the digital soil mapping (DSM) model for copper (Cu) (Paper III).

Figure 7 shows predicted Cu concentrations in agricultural soil, together with the prediction interval width. The spatial pattern of Cu concentrations generally followed the spatial variation in soil texture reported in Piikki and Söderström (2019) and in earlier Cu mapping by Eriksson et al. (2017). High predicted concentrations of Cu were mostly found on clayey soils, while lower concentrations were found on sandy soils. As indicated in Figure 7b, a wide prediction interval was often found on soils with a high predicted concentration of Cu. Only 3% of Swedish agricultural soil was predicted to have values at or below the risk limit of 7 mg kg^{-1}. This was very different from the value of 22% established by Eriksson et al. (2017). The low recall score of the DSM model at 7 mg kg^{-1} reflected the underestimation of soils below the risk limit in the Cu concentration map produced (see Figure 6).

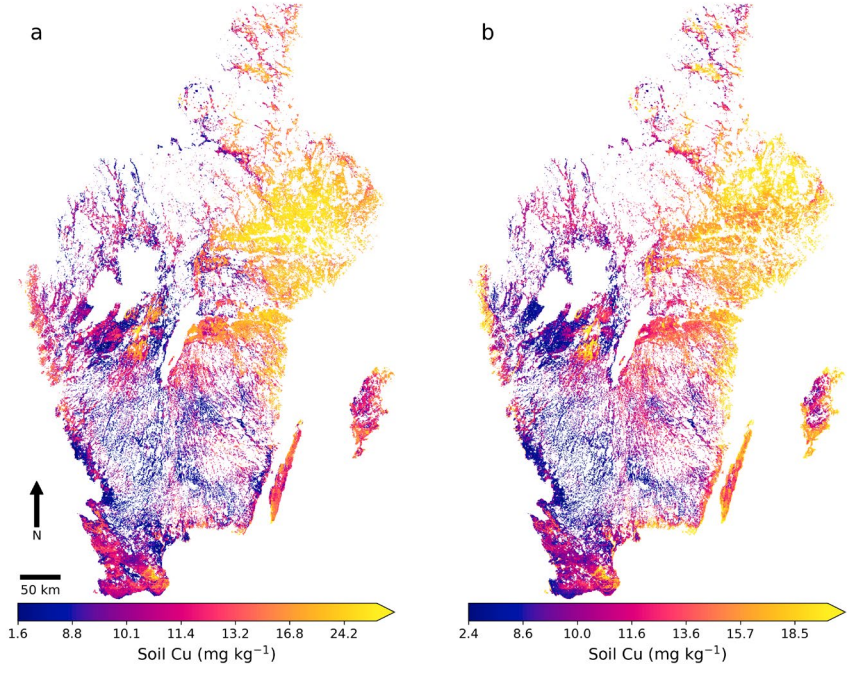


Figure 7: Maps of (a) predicted copper (Cu) concentrations in Swedish agricultural soil and (b) width of the 90% prediction intervals. The highest concentration presented in the colour bar of (a) and (b) corresponds to the 90th percentile (Paper III).

The underestimation of agricultural land at risk reveal that future work should perhaps focus on implementing procedures that will increase the accuracy of predictions at or below the risk limit. This could perhaps be achieved by using methods previously mentioned (see section 5.1, page 47-48), such as bias correction or calibration of DSM models specifically focused on predicting lower concentrations.

Zooming in on the validation farm of Bjertorp revealed that the prediction error was generally small (Figure 8). However, validation on this farm gave E and MAE of 0.56 and 2.0 mg kg^{-1}, respectively (Figure A1). This was a similar E value to that obtained for cross-validation, but almost half the MAE value. The PICP was 86% across all fields. This means that the prediction interval created was slightly too narrow for this farm. Some laboratory-analysed values were outside the lower prediction interval bound, and some were outside the upper prediction interval bound. This shows that the prediction interval is probably usable, although validated here only on one farm. A nationally derived PICP would be more representative of the validity of the prediction interval, and should be computed in the future. For example, the result from this farm might give an optimistic or pessimistic impression of the validity of the national prediction interval map.

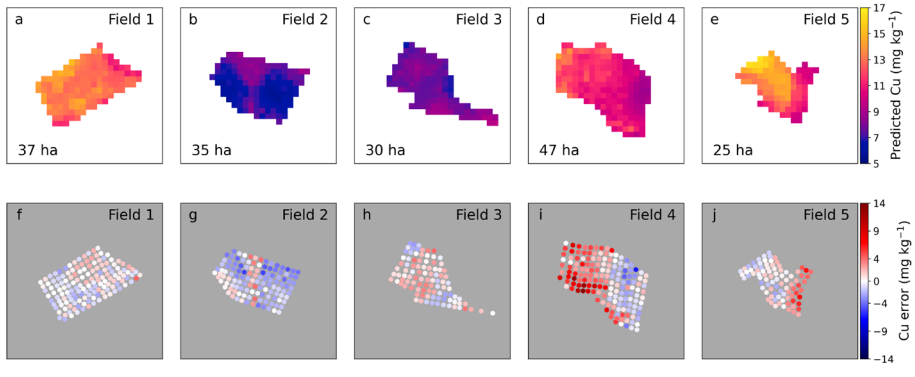


Figure 8: Maps of predicted copper (Cu) concentrations (a-e) zoomed in on the fields in Bjertorp and (f-j) the prediction error (Paper III).

As shown in Table 11, the field means were rather accurately predicted. The spatial resolution of the Cu map made it possible to map within-field, but the predicted within-field variation was not accurate (Table 11). A potential explanation for the problems with replicating within-field variation is that many of the covariates used have coarser spatial resolution than 50 m × 50 m, e.g. airborne gamma radiation measurements can have a footprint substantially larger than 50 m × 50 m.

Table 11: Field means of measured and predicted concentrations of copper (Cu) in the five fields at Bjertorp. Squared Pearson coefficient of a linear regression model (r^2) was used to assess how well the within-field variation was reproduced by the model (Paper III).

Field	**Measured field mean (mg Cu kg^{-1})**	**Predicted field mean (mg Cu kg^{-1})**	**r^2**
1	12.9	13.0	0.1
2	5.5	6.9	0.19
3	8.6	7.8	0.03
4	13.0	11.0	0.37
5	14.3	13.2	< 0.01

Figure 9 shows the map of Cd concentrations and the 90% prediction interval produced for the study area in Paper II. The spatial patterns obtained supported findings from previous work of high concentrations in Cd-rich parent material such as Cambrian sandstone or alum shale bedrock in the south-east part of the study area and in tills and glaciofluvial deposits influenced by this Cd-rich parent material (Söderström & Eriksson, 2013).

The prediction interval was generally wider on agricultural soils with high predicted concentrations of Cd, as was also the case in the Cu modelling in Paper III. A study by Poggio et al. (2021) found that wider prediction intervals often occurred in more sparsely sampled areas and in areas with high predicted values. In this thesis, areas with fewer samples, i.e. areas with less agricultural land, were not necessarily linked to a wide prediction interval. It would be interesting to compare the uncertainty produced by other methods, such as quantile regression forest, with that of GBR. This could reveal whether the prediction interval produced by the GBR model is subject to specific biases, as different mapping methods can yield different uncertainties (Heuvelink & Webster, 2022).

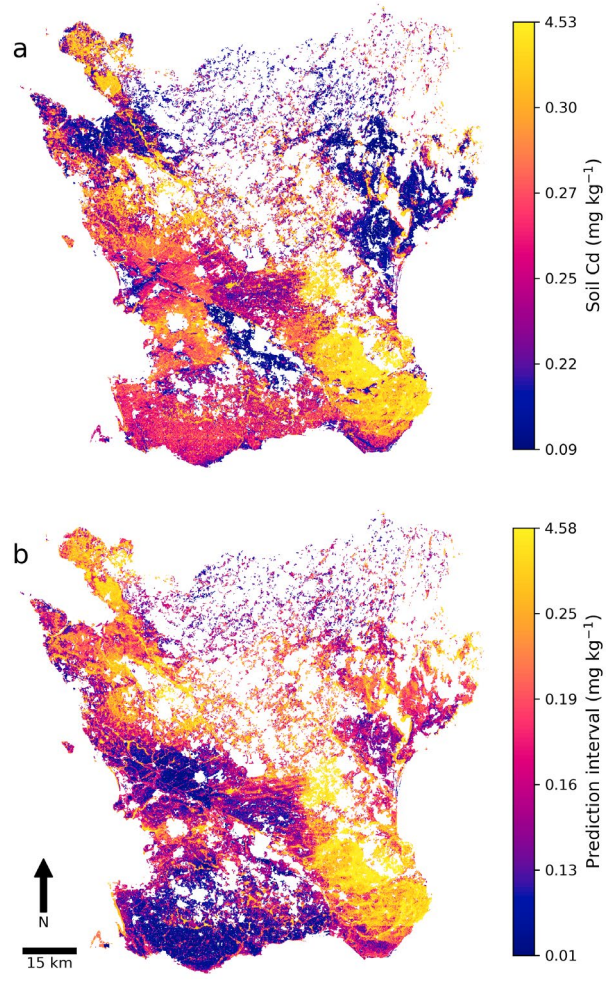


Figure 9: Maps of (a) predicted cadmium (Cd) concentrations in agricultural soil in Skåne County and (b) width of the 90% prediction intervals (Paper II).

5.4 DSM and its role as decision support

An attempt was made to explore if the Cd concentration map could be used as decision support in wheat production. First, areas that had Cd values at or below 0.196, 0.215 and 0.240 mg kg^{-1}, corresponding to the 30th, 40th and 50th percentile, respectively, of the analysed soil Cd concentrations in the NV

dataset, were selected. The overall aim was to assess whether Cd concentrations in grain could be kept low by selecting grains from areas with low predicted soil Cd concentrations. The results showed that the lower the predicted soil Cd concentration, the lower the median concentration in winter wheat grain (Figure 10).

Wheat grain grown on soils with Cd concentrations below the lowest limit, 0.196 mg kg^{-1}, generally had concentrations below 50 µg kg^{-1} (Figure 10). However, it is difficult to assess how these concentrations in grain would impact the final concentration in the potential end-product, e.g. baby food. Nevertheless, these soils could be suitable for production of winter wheat when Cd concentration in grain is needed to be kept as low as possible, to be used e.g. for baby food. The map could perhaps also be used as a planning tool, e.g. in sourcing winter wheat when specific requirements on Cd content in grain need to be met. Note that soils with low concentrations of Cd are rather uncommon in Scania County and there may be better scope for finding suitable areas in other regions of Sweden. Analysis of data on the Swedish winter wheat production area in 2020 revealed that only 4.4% of the area was within the lowest limit for soil Cd (Paper II).

The prediction interval is useful information, but not necessarily to the end user in its "raw" form (Wadoux et al., 2021b). Lark et al. (2022) argue that uncertainty values are best used in relation to some set value or limit, e.g. the probability of exceeding a certain concentration or risk limit. In Sweden, there are no guidelines or recommendations on Cd concentrations in agricultural soil, other than for sewage sludge application. Therefore, following on from Lark et al. (2022), the prediction uncertainties in some DSM products, including the Cd map in its current form in Paper II, can be purely decorative and not serve a clear purpose as decision support. Hence, it is not clear from Paper II how the prediction interval or its bounds could be used effectively as decision support, as no usable limit exists to relate the results to. One option could be to locate soils with low predicted Cd concentrations that also had a narrow prediction interval. However, this raises questions concerning how narrow the prediction interval should be and how low the predicted Cd concentration should be – which is difficult to answer.

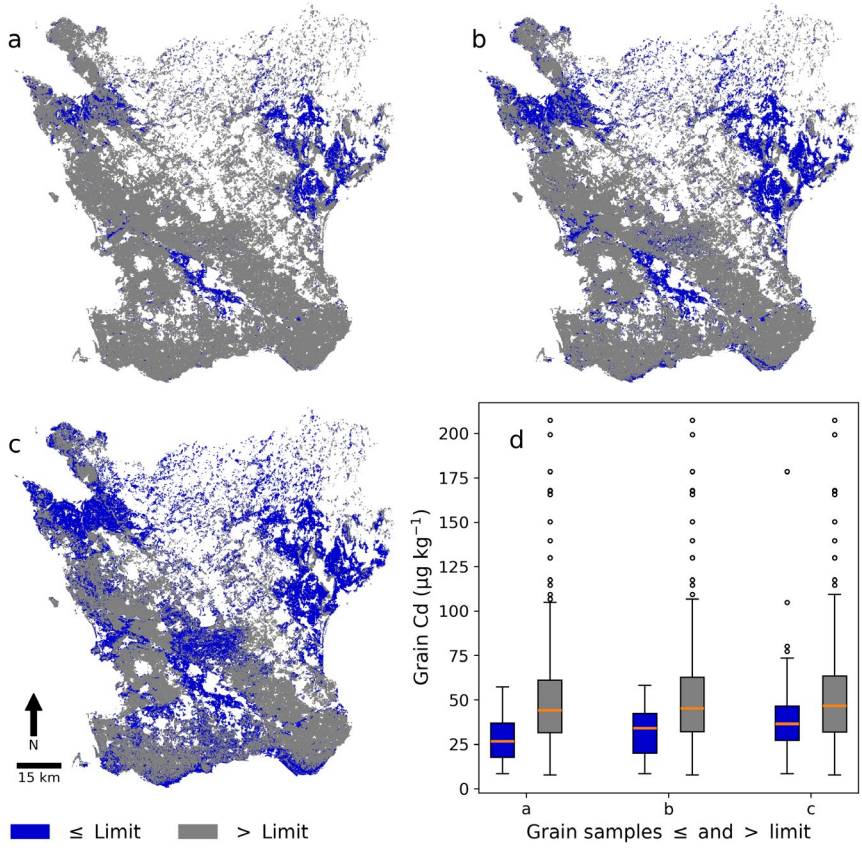


Figure 10: Maps showing (a-c) predicted soil cadmium (Cd) concentrations at or below defined limits and (d) the corresponding analysed grain Cd concentrations in winter wheat within and outside the delineated areas presented as boxplots (the orange line shows the median) (Paper II). The limits were 0.196, 0.215 and 0.240 mg kg^{-1}, corresponding to the 30th, 40th and 50th percentile of analysed Cd concentrations in the NV dataset, respectively.

The prediction interval bounds produced in Paper III could be used to answer questions about the risk limit of Cu deficiency in soil (7 mg kg^{-1}). Figure 11 shows agricultural land in Sweden where the lower bound of the prediction interval (5th percentile) was above the risk limit for Cu deficiency, i.e. where Cu concentrations in agricultural soils are highly likely not to be below the risk limit (95% probability). This equated to 47% of Swedish agricultural land (Figure 11).

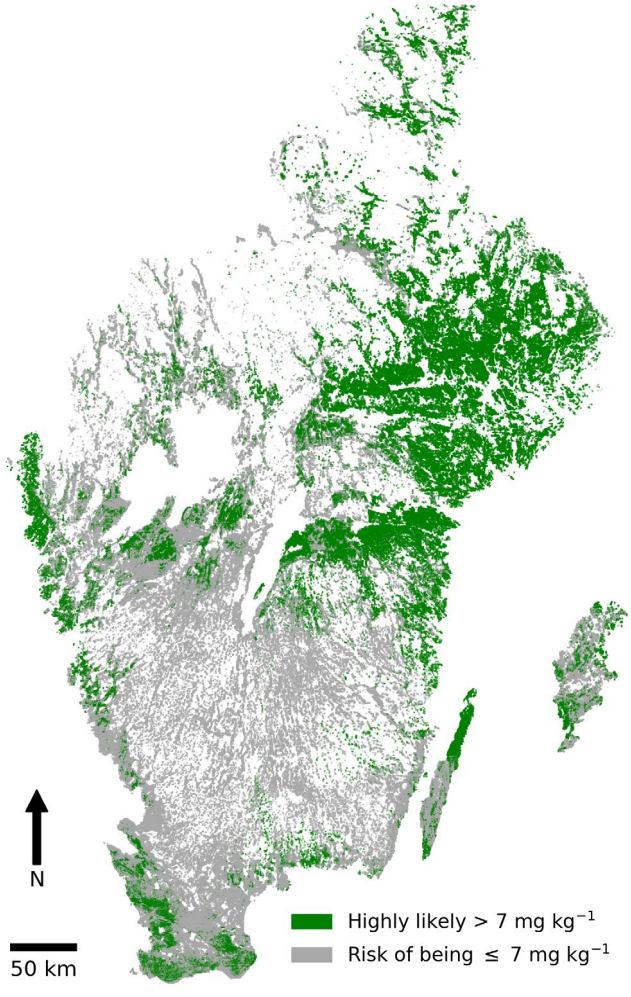


Figure 11: Map showing Swedish agricultural soils with copper (Cu) concentrations highly likely (95% probability) to be above the deficiency risk limit, and soils potentially below the risk limit. Based on mapping the lower bound of the prediction interval (5th percentile) (Paper III).

The remaining 53% of land cannot be excluded from a risk of being Cu-deficient, and more information, e.g. from laboratory analyses of soil samples, is needed. The map produced makes it possible to divert and redirect focus to these soils. It can also act as a guide for farmers and advisors

in discussions and decisions on whether laboratory analysis of Cu concentrations are needed and if so, where to take the soil samples.

The width of the prediction interval made it unsuitable for locating soils with concentrations highly likely to be below the risk limit. The upper bound was only below 7 mg kg^{-1} in two cells in the map. Therefore, it is obvious that more information is needed on the prediction uncertainty. It would probably be more correct to predict many percentiles (1, 2, 3 … 99), and not just the 95^{th} and 5^{th} as done in this thesis. This would have been particularly useful when assessing the varying probabilities of risk in the grey areas in Figure 11. For example, doing this would have made it possible to query a point in the map and see the predicted percentile at which the risk limit of Cu deficiency was reached. A predicted 30^{th} percentile of e.g. 7 mg kg^{-1} could imply a 30% probability of being at or below the risk limit at that point in the map. However, this would be more computationally demanding, as every percentile from the 1^{st} to the 99^{th} would have to be predicted when using GBR. It would also be important to assess the validity of these percentiles and the resulting prediction intervals, e.g. using PICP. The 0 and 100^{th} percentiles should not be predicted, as they are the minimum and maximum, and predicting these would mean that it would be impossible to find future observations outside the prediction interval.

5.5 Covariate importance in DSM

In Paper III, airborne gamma radiation measurement data ranked high for both methods used for assessing covariate importance (Table 12, Figure A2). Airborne gamma radiation measurements can indicate the degree of weathering and mineralogy (International Atomic Energy Agency, 2003). They are thus a good predictor of soil texture, which in turn is a good predictor of Cu concentrations. This probably explains why the soil texture classes ranked low in the assessment in Paper III, as the airborne gamma radiation data already contained sufficient information. Therefore, the covariates of soil texture classes may be redundant.

Table 12: Covariate importance ranking of the digital soil mapping (DSM) model for cadmium (Cd) (Paper II) and the DSM model for copper (Cu) (Paper III). MDI = mean decrease in impurity, PI = permutation importance, BioGeo = cokriged biogeochemical data, TWI = topographic wetness index, ConvInd = convergance index, TPI = topographic position index, U = Uranium (^{238}U), Th = Thorium (^{232}Th), K = Potassium (^{40}K). Each covariate type is colour-coded: Green = airborne gamma radiation measurement data, red = topographic data, blue = biogeochemical data, yellow = soil texture classes, grey = climate data.

Rank	Cd, MDI	Cd, PI	Cu, MDI	Cu, PI
1	U	U	Th	Th
2	BioGeo	Th	Dose rate	Prec MAM
3	Th	BioGeo	U	U
4	Elevation	K	Prec MAM	Dose rate
5	K	TPI 500	Elevation	Temp SON
6	TPI 5	Elevation	Temp DJF	Temp DJF
7	TPI 500	TPI 5	Prec DJF	Temp annual
8	TPI 50	TPI 50	Prec SON	Temp MAM
9	ConvInd	ConvInd	TPI 5	Elevation
10	TWI	TWI	Prec JJA	Prec DJF
11	Sand	Silt	K	K
12	Till	Till	Temp SON	Temp JJA
13	Silt	Sand	Temp JJA	Prec JJA
14	Clay till	Clay till	Temp annual	TPI 5
15	Clay	Other	TPI 50	Prec annual
16	Other	Clay	Temp MAM	Prec SON
17	-	-	TPI 500	TPI 50
18	-	-	Clay	TPI 500
19	-	-	Soil moisture	Clay
20	-	-	Prec annual	Soil moisture
21	-	-	Clay till	Clay till
22	-	-	Sand	Sand
23	-	-	Silt	Silt
24	-	-	Other	Other
25	-	-	Till	Till

However, this highlights one of the problems with MDI, whereby covariates with many unique values, e.g. dose rate, provide more opportunities for finding splitting thresholds in the regression tree nodes than the binary

classes of soil texture (Strobl et al., 2007). It could be the case that airborne gamma radiation measurement data contain better information about soil texture, or that they simply provide more opportunities for splitting thresholds.

In Paper III, climate covariates, and especially seasonally subdivided climate covariates, were identified as being important in the DSM model. Hengl et al. (2017) also found that climate covariates were important when mapping soil Cu in sub-Saharan Africa. Concentrations of Cu in soil are a product of environmental factors, including climate (Oorts, 2013), and there may be causal relationships between climate variables and soil properties. However, it is also possible that climate covariates simply act as spatial partitioning covariates, which was probably the case in Paper III. For example, higher amounts of precipitation fall in western Sweden and this region also has more coarse-textured soils, which are often linked to low Cu concentrations. Lastly, elevation ranked highest of all the DEM covariates in Paper III. Chen et al. (2012) found that topography could explain the spatial variation structure in soil Cu concentrations, with higher concentrations in lower parts of watersheds, and such relationships may also occur in Sweden. The DEM derivatives, such as topographic position index (TPI) and soil moisture, perhaps contained redundant information compared with elevation above sea level and precipitation data. However, it should be mentioned that the soil moisture map is not calibrated for agricultural soils.

In Paper II, the results were fairly similar to those in Paper III, with airborne gamma radiation covariates ranking highest (Table 12, Figure A3). Uranium was important in the DSM model for Cd, which was expected considering the strong correlations between gamma radiation-measured U and soil Cd concentrations identified by Söderström and Eriksson (2013), with soils influenced by Cd-rich alum shale often having elevated U concentrations. Cokriged biogeochemical data, i.e. a rough delineation of Cd concentrations in the landscape, were also highlighted as important in the DSM model in Paper II. Elevation was identified as the most important DEM covariate. Qiu et al. (2020) showed that Cd concentrations in soil are related to elevation gradients, and this may be the case in Skåne County as well. Hence, DEM covariates might be an important factor when conducting DSM of Cd. Nevertheless, based on the results from Papers II and III, it could be argued that elevation above sea level provides most of the necessary topographical information needed.

The performance of the DSM models for Cd and Cu in Papers II and III, respectively, indicated that more covariates are probably needed or that some critical covariate may be missing. Some previous DSM studies, e.g. those by Hengl et al. (2017) and Poggio et al. (2021), concluded that more covariates were needed, as this could improve model performance. In contrast, the results in this thesis indicate that many covariates may be redundant and could perhaps be omitted to increase model parsimony. However, it is important that the natural system is well represented by the covariates. Covariates on management practices, or proxies of these, could be beneficial, especially as management practices can influence Cu concentrations in agricultural soil (Vavoulidou et al., 2011; Oorts, 2013). How management practices can be represented as a covariate remains unclear, as to my knowledge there is no national dataset of this kind. An option could be to use indirect information on management practices for each agricultural field, e.g. the most commonly grown crop in each field. It would also be interesting to implement spatial covariates such as the oblique geographical coordinates (OGCs) proposed by Møller et al. (2020) in future DSM. Oblique geographical coordinates could perhaps be more suitable as spatial partitioning covariates than the climate covariates used in this thesis.

6. Conclusions and future prospects

In this thesis, DSM was conducted to create maps of Cu and Cd in agricultural soils Sweden. Prediction uncertainty, as the 90% prediction interval, was also mapped. The overall aim was to create maps that could serve as decision support for crop production. The results demonstrated that the Cd map created could be used to source winter wheat grain with low concentrations of Cd. This mapping framework can be scaled up to cover the whole of Sweden, providing opportunities to identify other suitable areas for production of winter wheat with low Cd concentrations in grain. Digital soil mapping of Cu revealed that the prediction interval bounds could be used to exclude agricultural soils from the risk of being Cu deficient.

Prediction interval mapping proved to be an important part of DSM, but needs to be improved in order to answer probabilistic questions with regard to crop production, e.g. the probability of the risk limit being exceeded at a certain point. The lower concentration range of Cu and Cd is of most interest for applications in crop production, so future DSM modelling should perhaps focus more on specific ranges of concentrations or on using classification models rather than regression models. In the future, it would also be interesting to improve the maps at local level, either by recalibrating the DSM model using local soil samples, i.e. spiking, or by using residual or regression kriging, as done by Söderström et al. (2016) and Nijbroek et al. (2018).

Predicting trace elements using PXRF measurements is promising, and can produce very accurate estimates of e.g. Zn concentration. However, prediction biases at lower concentrations of Cu and Cd reveal that there is still room for improvement. The future of PXRF models may lie in using bias correction or by using the raw spectra and data from other sensors, as tested by Shresta et al. (2022). In this thesis, using the concentrations of Cu

predicted by a PXRF model, and thus increasing the calibration dataset size, made the subsequent DSM model more accurate. Hence, the speed and ease of PXRF measurements makes it a powerful tool in DSM when more calibration data are needed.

The soil samples in the JV and NV datasets make it possible to conduct future DSM on many more trace elements or soil properties. For example, the NV dataset also contains measured pH and Mn concentrations. All these soil samples have been analysed with the PXRF device. Hence, Swedish soil mapping is promising and can be expanded to include many more variables in the future.

Covariate importance analysis revealed high importance of airborne gamma radiation measurement data in DSM of Cu and Cd concentrations. It would be interesting to improve the spatial resolution of airborne gamma radiation data so that within-field variation can potentially be predicted more accurately. Covariate redundancy of e.g. soil texture classes and DEM derivatives should be avoided in future Swedish DSM of Cu and Cd. Ideally, new covariates should contain novel information from new sources.

Lastly, I argue that future DSM in Sweden should be performed in close collaboration with stakeholders, so that the products developed are useful and desirable to the individual farmer, advisor, authority or industry. Significant efforts should be made to produce educational maps that suit the intended purpose. These maps should perhaps also be interactive, e.g. making it possible to zoom in at individual fields, obtain meaningful statistics or query probabilities of exceeding user-set limits. However, issues relating to handling sensitive information that these DSM products might contain, such as Cd concentrations in agricultural soils, need to be further assessed in future Swedish DSM.

References

Adamchuk, V., & Viscarra Rossel, R. A. (2010). Development of On-the-Go Proximal Soil Sensor Systems. In R.A. Viscarra Rossel, A. McBratney & B. Minsasny (Eds.), *Proximal Soil Sensing* (1st ed., p. 15-28). Springer. https://doi.org/10.1007/978-90-481-8859-8_2

Adrees, M., Ali, S., Rizwan, M., Ibrahim, M., Abbas, F., Farid, M., Zia-ur-Rehman, M., Irshad, M. K. I., & Bharwana, S. A. (2015). The effect of excess copper on growth and physiology of important food crops: a review. *Environmental Science and Pollution Research*, 22, 8148-8162. https://doi.org/10.1007/s11356-015-4496-5

Akpa, S. I., Odeh, I. O., Bishop, T. F., Hartemink, A. E., & Amapu, I. Y. (2016). Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma*, 271, 202-215. https://doi.org/10.1016/j.geoderma.2016.02.021

Alloway, B. J. (2009). Soil factors associated with zinc deficiency in crops and humans. *Environmental Geochemistry and Health*, 31, 537-548. https://doi.org/10.1007/s10653-009-9255-4

Alloway, B. J. (2013). Heavy Metals and Metalloids as Micronutrients for Plants and Animals. In B. J. Alloway (Ed.), *Heavy Metals in Soils* (3rd ed., p. 195-209). Springer. https://doi.org/10.1007/978-94-007-4470-7_7

Arrouays, D., Lagacherie, P., & Hartemink, A. E. (2017). Digital soil mapping across the globe. *Geoderma Regional*, 9, 1-4. https://doi.org/10.1016/j.geodrs.2017.03.002

Arrouays, D., McBratney, A., Bouma, J., Libohova, Z., Richer-de-Forges, A. C., Morgan, L. S. M., Roudier, P., Poggio, L., & Mulder, V. L. (2020b). Impressions of digital soil maps: The good, the not so good, and making them ever better. *Geoderma Regional*, 20, e00255. https://doi.org/10.1016/j.geodrs.2020.e00255

Arrouays, D., Poggio, L., Salazar Guerrero, O. A., & Mulder, V. L. (2020a). Digital soil mapping and *GlobalSoilMap*. Main advances and ways forward. *Geoderma Regional*, 21, Article e00265. https://doi.org/10.1016/j.geodrs.2020.e00265

Ballabio, C., Panagos, P., Lugato, E., Huang, J-H., Orgiazzi, A., Jones, A., Fernándes-Ugalde, O., Borrelli, P., & Montaranella, L. (2018). Copper distribution in European topsoils: An assessment based on LUCAS soil survery. *Science of the Total Environment*, 636, 282-298. https://doi.org/10.1016/j.scitotenv.2018.04.268

Belitz, K., & Stackelberg, P. E. (2021). Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models. *Environmental Modelling & Software*, 139, 105006. https://doi.org/10.1016/j.envsoft.2021.105006

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

Broadley, M. R., White, P. J., Hammond, J. P., Zelko, I., & Lux, A. (2007). Zinc in plants. *New Phytologist*, 173, 677-702. https://doi.org/10.1111/j.1469-8137.2007.01996.x

Burgess, T. M., & Webster, R. (1980). Optimal interpolation and isarithmic mapping of soil properties: I. The semi-variogram and punctual kriging. *Journal of Soil Science*, 31, 315-331. https://doi.org/10.1111/j.1365-2389.1980.tb02084.x

Chen, S., Arrouays, D., Mulder, V. L., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de-Forges, A. C., & Walter, C. (2022). Digital mapping of *GlobalSoilMap* soil properties at a broad scale: A review. *Geoderma*, 409, 115567. https://doi.org/10.1016/j.geoderma.2021.115567

Chen, Y., Liu, Y., Liu, Y., Lin, A., Kong, X., Liu, D., Li, X., Zhang, Y., Gao, Y., & Wang, D. (2012). Mapping of Cu and Pb Contaminations in Soil Using Combined Geochemistry, Topography, and Remote Sensing: A Case Study in the Le'an River Floodplain, China. *International Journal of Environmental Research and Public Health*. 9, 1874-1886. https://doi.org/10.3390/ijerph9051874

Duval, J.S., Carson, J.M., Holman, P.B., Darnley, A.G. (2005). *Terrestrial radioactivity and gamma-ray exposure in the United States and Canada, USGS Open-File Report 2005-1413*. http://pubs.usgs.gov/of/2005/1413/

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77, 802-813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Ellili, Y., Walter, C., Michot, D., Pichelin, P., & Lemercier, B. (2019). Mapping soil organic carbon stock change by soil monitoring and digital soil mapping at the landscape scale. *Geoderma*, 351, 1-8. https://doi.org/10.1016/j.geoderma.2019.03.005

El Naqa, I., & Murphy, M. J. (2015). What is Machine Learning? In I. El Naqa, R. Li, M. Murphy (Eds.), *Machine Learning in Radiation Oncology* (1st ed., p. 3-11). Springer. https://doi.org/10.1007/978-3-319-18305-3_1

Eriksson, J. (2021). *Current status of Swedish arable soils and cereal crops. Data from the period 2011 – 2017*. https://pub.epsilon.slu.se/23486/1/eriksson_j_210514.pdf

Eriksson, J., Dahlin, S. A., Sohlenius, G., Söderström, M., & Öborn, I. (2017). Spatial patterns of essential trace element concentrations in Swedish soils

and crops. *Geoderma Regional*, 10, 163-174. http://dx.doi.org/10.1016/j.geodrs.2017.07.001

European commission. (2021). *Commission Regulation (EC) No 1881/2006 (2021) Setting maximum levels for certain contaminants in foodstuffs*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02006R1881-20210831

Fageria, N. K., Baligar, V. C., & Clark, R. B. (2002). Micronutrients in Crop Production. *Advances in Agronomy*. 77, 185-268. https://doi.org/10.1016/S0065-2113(02)77015-6

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals Of Statistics*, 29, 1189-1232. https://doi.org/10.1214/aos/1013203451

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3-42. https://doi.org/10.1007/s10994-006-6226-1

Gholizadeh, A., & Kopačková, V. (2019). Detecting vegetation stress as a soil contamination proxy: a review of optical proximal and remote sensing techniques. *International Journal of Environmental Sciences*. 16, 2511-2524. https://doi.org/10.1007/s13762-019-02310-w

Glanzman, R. K., & Closs, L. G. (2007). Field Portable X-Ray Fluorescence Geochemical Analysis – Its Contribution to Onsite real-time Project Evaluation. In B. Milkereit (Ed.), *Proceedings of Exploration 07: Fifth Decennial International Conference on Mineral Exploration* (p. 291-301). Decennial Mineral Exploration Conferences.

Goff, K., Schnaetzl, R. J., Chakraborty, S., Weindorf, D. C., Kasmerchak, C., & Bettis III, E. A. (2020). Impact of sample preparation methods for characterizing the geochemistry of soils and sediments by portable X-ray fluorescence. *Soil Science Society of America Journal*, 84, 131-143. https://doi.org/10.1002/saj2.20004

Grundy, M. J., Searle, R., Meier, E. A., Ringrose-Voase, A. J., Kidd, D., Orton, T. G., Triantafilis, J., Philip, S., Liddicoat, C., Malone, B., Thomas, M., Gray, J., & McLearn Bennet, J. (2020). Digital soil assessment delivers impact across scales in Australia and the Philippines. *Geoderma Regional*, 22, e00314. https://doi.org/10.1016/j.geodrs.2020.e00314

Grundy, M. J., Viscarra Rossel, R. A., Searle, R. D., Wilson, P. L., Chen, C., & Gregory, L. J. (2015). Soil and Landscape Grid of Australia. *Soil Research*, 53, 835-844. https://doi.org/10.1071/SR15191

Grunwald S. (2010). Current State of Digital Soil Mapping and What Is Next. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink & S. Kienast-Brown (Eds.), *Digital Soil Mapping. Progress in Soil Science, vol 2* (1st ed., p. 3-12). Springer. https://doi.org/10.1007/978-90-481-8863-5_1

Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., Arévalo, G. E., Arroyo-Cruz, C. E., Bolivar, A., Bunning,

S., Bustamante Cañas, N., Cruz-Gaistardo, C. O., Davila, F., Dell Acqua, M., Encina, A., Figueredo Tacona, H., Fontes, F., Hernández Herrera, J. A., Ibelles Navarro, A. R., … Vargas, R. (2018). No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. *SOIL*, 4, 173-193. https://doi.org/10.5194/soil-4-173-2018

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer-Verlag: New York, NY, USA, 2009. https://doi.org/10.1007/978-0-387-84858-7

Hengl, T., Leenaars, J. G. B., Shepherd, K. D., Walsh, M. G., Heuvelink, G. B. M., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E., Wheeler, I., & Kwabena, N. A. (2017). Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutrient Cycling in Agroecosystems*, 109, 77-102. https://doi.org/10.1007/s10705-017-9870-x

Heuvelink, G. B. M. (2014). Uncertainty quantification of GlobalSoilMap products. In D. Arrouays, N. McKenzie, J. Hempel, A. Richer de Forges & A. B. McBratney (Eds.), *GlobalSoilMap. Basis of the Global Spatial Soil Information System* (p. 335-340). CRC Press.

Heuvelink, G. B. M., & Webster, R. (2022). Spatial statistics and soil mapping: A blossoming partnership under pressure. *Spatial Statistics*, 100639. https://doi.org/10.1016/j.spasta.2022.100639

Hu, W., Huang, B., Weindorf, D. C., & Chen, Y. (2014). Metals Analysis of Agricultural Soils via Portable X-ray Fluorescence Spectrometry. *Bulletin of Environmental Contamination and Toxicology*, 92, 420-426. https://doi.org/10.1007/s00128-014-1236-3

International Atomic Energy Agency. (2003). *Guidelines for radioelement mapping using gamma ray spectrometry data*. https://www-pub.iaea.org/MTCD/Publications/PDF/te_1363_web.pdf

Jenny, H. (1941). *Factors of Soil Formation, A System of Quantitative Pedology*. McGraw-Hill.

Kalnicky, D. J., & Singhvi, R. (2001). Field portable XRF analysis of environmental samples. *Journal of Hazardous Materials*, 83, 93-122. https://doi.org/10.1016/S0304-3894(00)00330-7

Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling*, 81, 401-418. https://doi.org/10.1016/j.apm.2019.12.016

Kidd, D., Searle, R., Grundy, M., McBratney, A., Robinson, N., O'Brien, L., Zund, P., Arrouays, D., Thomas, M., Padarian, J., Jones, E., McLean Bennett, J., Minasny, B., Holmes, K., Malone, B. P., Liddicoat, C., Meier, E. A., Stockmann, U., Wilson, P., … Triantafilis, J. (2020). Operationalising Digital Soil Mapping – Lessons from Australia. *Geoderma*, 23, e00335. https://doi.org/10.1016/j.geodrs.2020.e00335

Kim, J., & Grunwald, S. (2016). Assessment of carbon stocks in the topsoil using random forest and remote sensing images. *Journal of Environmental Quality*, 45, 1910-1918. https://doi.org/10.2134/jeq2016.03.0076

Lai, Y. Q., Wang, H. L., & Sun, X. L. (2021). A comparison of importance of modelling method and sample size for mapping soil organic matter in Guangdong, China. *Ecological Indicators*, 126, 107618. https://doi.org/10.1016/j.ecolind.2021.107618

Lark, R. M., Chaguumaira, C., & Milne, A. E. (2022). Decisions, uncertainty and spatial information. *Spatial Statistics*, 100619. https://doi.org/10.1016/j.spasta.2022.100619

Lax, K. (2009). Biogeochemical Data from SGU: Properties and Applications [Doctoral dissertation, Luleå University of Technology, Luleå, Sweden]. DiVa. http://ltu.diva-portal.org/smash/get/diva2:990669/FULLTEXT01.pdf

Lemière, B. (2018). A review of pXRF (field portable X-ray fluorescence) applications for applied geochemistry. *Journal of Geochemical Exploration*, 188, 350-363. https://doi.org/10.1016/j.gexplo.2018.02.006

Malone, B.P., Odgers, N. P., Stockmann, U., Minasny, B., & McBratney, A. B. (2018). Digital soil mapping of Soil classes and Continuous Soil Properties. In A. B. McBratney & U. Stockmann (Eds.), *Pedometrics* (1st ed., p. 373-413). Springer. https://doi.org/10.1007/978-3-319-63439-5_12

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

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

Mulder, V. L., de Bruin, S., Schaepman, M. E., & Mayr, T. R. (2011). The use of remote sensing in soil and terrain mapping – A review. *Geoderma*. 162, 1-19. https://doi.org/10.1016/j.geoderma.2010.12.018

Møller, A. B., Beucher, A. M., Pouladi, N., & Greve, M. H. (2020). Oblique geographic coordinates as covariates for digital soil mapping. *SOIL*, 6, 269-289. https://doi.org/10.5194/soil-6-269-2020

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I – A discussion of principles. *Journal of Hydrology*, 10, 282-290. https://doi.org/10.1016/0022-1694(70)90255-6

Nijbroek, R., Piikki, K., Söderström, M., Kempen, B., Turner, K. G., Hengari, S., & Mutua, J. (2018). Soil Organic Carbon Baselines for Land Degradation Neutrality: Map Accuracy and Cost Tradeoffs with Respect to Complexity in Otjozondjupa, Namibia. *Sustainability*, 10, 1610. https://doi.org/10.3390/su10051610

Oorts, K. (2013). Copper. In B. Alloway (Ed.), *Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability* (3rd ed., p. 367-394). Springer. https://doi.org/10.1007/978-94-007-4470-7_13

O'Rourke, S. M., Minasny, B., Holden, N. M., & McBratney, A. B. (2016). Synergistic use of Vis-NIR, MIR, and XRF spectroscopy for the determination of soil geochemistry. *Soil Science Society of America Journal*, 80, 888-899. https://doi.org/10.2136/sssaj2015.10.0361

Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: A review aided by machine learning tools. *SOIL*, 6, 35-52. https://doi.org/10.5194/soil-6-35-2020

Padilla, J. T., Hormes, J. F., & Selim, H. M. (2019). Use of portable XRF: Effect of thickness and antecedent moisture of soil on measured concentrations of trace elements. *Geoderma*, 337, 143-149. https://doi.org/10.1016/j.geoderma.2018.09.022

Pedregosa, F., Varaquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnayet, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Piikki, K., & Söderström, M. (2019). Digital soil mapping of arable land in Sweden – Validation of performance at multiple scales. *Geoderma*, 352, 342-350. https://doi.org/10.1016/j.geoderma.2017.10.049

Piikki, K., Wetterlind, J., Söderström, M., & Stenberg, B. (2021). Perspectives on validation in digital soil mapping of continuous attributes – A review. *Soil Use and Management*, 37, 7-21. https://doi.org/10.1111/sum.12694

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7, 217-240. https://doi.org/10.5194/soil-7-217-2021

Prettenhofer, P., & Louppe, G. (2014). Gradient Boosting Regression Trees in Scikit-learn. In Proceedings of the PyData 2014, London, UK, 21–23 February 2014. https://orbi.uliege.be/bitstream/2268/163521/1/slides.pdf. Accessed 27 February 2022

Qiu, M., Yuan, C., & Yin, G. (2020). Effect of terrain gradient of cadmium accumulation in soils. *Geoderma*, 375, 114501. https://doi.org/10.1016/j.geoderma.2020.114501

Ravansari, R., & Lemke, L. D. (2018). Portable X-ray fluorescence trace metal measruement in organic rich soils: pXRF response as a function of organic matter fraction. *Geoderma*, 319, 175-184. https://doi.org/10.1016/j.geoderma.2018.01.011

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., & O'Connor, P. (2014a). *Chemistry of Europé's Agricultural Soils, Part A: Methodology and*

interpretation of the GEMAS Data Set, Geologisches Jahrbuch Reihe B, Band B 102. Schweizebart Science Publishers, Stuttgart.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., & O'Connor, P. (2014b). *Chemistry of Europé's Agricultural Soils, Part B: General Background Informations and Further Analysis of the GEMAS Data Set, Geologisches Jahrbuch Reihe B, Band B 103*. Schweizebart Science Publishers, Stuttgart.

Richer-de-Forges, A. C., Arrouays, D., Bardy, M., Bispo, A., Lagacherie, P., Laroche, B., Lemercier, B., Sauter, J., & Voltz, M. (2019). Mapping of Soil and Land-Related Environmental Attributes in France: Analysis of End-Users´ Needs. *Sustainability*, 11, 2940. https://doi.org/10.3390/su11102940

Rouillon, M., & Taylor, M. P. (2016). Can field portable X-ray fluorescence (pXRF) produce high quality data for application in environmental contamination research? *Environmental Pollution,* 214, 255-264. https://doi.org/10.1016/j.envpol.2016.03.055

Scull, P., Franklin, J., Chadwick, O. A., & McArthur, D. (2003). Predictive soil mapping: a review. *Progress in Physical Geography*, 27, 171-197. https://doi.org/10.1191/0309133303pp366ra

Shresta, G., Calvelo-Pereira, R., Roudier, P., Martin, A. P., Turnbull, R. E., Kereszturi, G., Jeyakumar, P., & Anderson, C. W. N. (2022). Quantification of multiple soil trace elements by combining portable X-ray fluorescence and reflectance spectroscopy. *Geoderma*, 409, 115649. https://doi.org/10.1016/j.geoderma.2021.115649

Shresta, D. L., & Solomatine, D. P. (2006). Machine learning approaches for estimation of prediction interval for the model output. *Neural Networks*, 19(2), 225-235. https://doi.org/10.1016/j.neunet.2006.01.012

SMHI. (2015). *Klimatscenarier för Sverige – Bearbetning av RCP-scenarier för meteorologiska och hydrologiska effektstudier (in English: Climate scenarios for Sweden – Processing of RCP-scenarios for meteorological and hydrological effect studies)*. http://www.smhi.se/polopoly_fs/1.165049!/Klimatologi_15%20Klimatscenarier%20f%C3%B6r%20Sverige%20-%20Bearbetning%20av%20RCP-scenarier%20f%C3%B6r%20meteorologiska%20och%20hydrologiska%20effektstudier.pdf

Smolders E., & Mertens, J. (2013). Cadmium. In: B. J. Alloway (Ed.), *Heavy Metals in Soils* (3rd ed., p. 283-311). Springer. https://doi.org/10.1007/978-94-007-4470-7_10

Somarathna, P. D. S. N., Minasny, B., & Malone, B. P. (2017). More Data or a Better Model? Figuring Out What Matters Most for the Spatial Prediction of Soil Carbon. *Soil Science Society of America Journal*, 81, 1413-1426. https://doi.org/10.2136/sssaj2016.11.0376

Song, J. (2015). Bias corrections for Random Forest in regression using residual rotation. *Journal of the Korean Statistical Society*, 44, 321-326. https://doi.org/10.1016/j.jkss.2015.01.003

Steinnes, E. (2013). Mercury. In B. J. Alloway (Ed.) *Heavy Metals in Soils* (3rd ed., p. 411-428). Springer. https://doi.org/10.1007/978-94-007-4470-7_15

Strobl, C., Boulesteix, A. L., & Augustin, T. (2007). Unbiased Split Selection for Classification Trees Based on the Gini Index. *Computational Statistics & Data Analysis*, 52, 483–501. https://doi.org/10.1016/j.csda.2006.12.030

Strobl, C., Boulestix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9, 307. https://doi.org/10.1186/1471-2105-9-307

Stockmann, U., Padarian, J., McBratney, A., Minasny, B., de Brogniez, D., Montanarella, L., Hong, S. Y., Rawlins, B. G., & Field, D. J. (2015). Global soil organic carbon assessment. *Global Food Security*, 6, 9-16. https://doi.org/10.1016/j.gfs.2015.07.001

Swedish Board of Agriculture. (2015). *Nationell jordartkartering: Matjordens egenskaper i åkermarken (National soil mapping of topsoil properties)*. https://www2.jordbruksverket.se/download/18.4288f19214fb7ec78849af18/1441973777932/ra15_19.pdf

Swedish Board of Agriculture. (2020). *Rekommendationer för gödsling och kalkning 2021 (In English: Recommendations for fertilizing and liming 2021)*. https://www2.jordbruksverket.se/download/18.dc97d8e176cea4b0ec29b80/1609846154443/jo20_12.pdf

Swedish Board of Agriculture. (2021). *Rekommendationer för gödsling och kalkning 2022 (In English: Recommendations for fertilizing and liming 2022)*. https://www2.jordbruksverket.se/download/18.2e41a0a017fbf0833c41c938/1648218102903/jo21_9v2.pdf

Swedish Environmental Protection Agency. (1998). *Statens Naturvårdsverks Föreskrifter om Ändring i Kungörelsen (SNFS 1994:2) med Föreskrifter om Skydd för Miljön, Särskilt Marken, när Avloppslam Används i Jorbruket (The Swedish Environmental Protection Agencies Changes on When Sewage Sludge Can Be Used in Agriculture). Statens Naturvårdsverks Författningsssamling; Stockholm, Sweden: 1998*. https://www.naturvardsverket.se/globalassets/nfs/1998/nfs1998-04.pdf

Swedish Institute for Standards. (2017). *Soil analysis – Determination of trace elements in soil by extraction with nitric acid*. https://www.sis.se/en/produkter/environment-health-protection-safety/soil-quality-pedology/chemical-characteristics-of-soils/ss283112017/

Sylvain, JD., Anctil, F., & Thiffault, E. (2021). Using bias correction and ensemble modelling for predictive mapping and related uncertainty: A case study in digital soil mapping. *Geoderma*, 403, 115153. https://doi.org/10.1016/j.geoderma.2021.115153

Szatmári, G., & Pásztor, L. (2019). Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. *Geoderma*, 337, 1329-1340. https://doi.org/10.1016/j.geoderma.2018.09.008

Söderström, M., & Eriksson, J. (2013). Gamma-ray spectrometry and geological maps as tools for cadmium risk assessment in arable soils. *Geoderma*, 192, 323-334. http://dx.doi.org/10.1016/j.geoderma.2012.07.014

Söderström, M., Sohlenius, G., Rodhe, L., & Piikki, K. (2016). Adaptation of regional digital soil mapping for precision agriculture. *Precision Agriculture*, 17, 588-607. https://doi.org/10.1007/s11119-016-9439-8

United States Environmental Protection Agency (US EPA). (2007). *Method 6200 – Field Portable X-Ray Fluorescence Spectrometry for the Determination of Elemental Concentrations in Soil and Sediment*. https://www.epa.gov/sites/default/files/2015-12/documents/6200.pdf

Vavoulidou, E., Avramides, E. J., Papadopoulos, P., Dimirkou, A., Charoulis, A., & Konstantinidou-Doltsinis, S. (2011). Copper Content in Agricultural oil Related to Cropping Systems in Different Regions of Greece. *Communications in Soil Science and Plant Analysis*, 36, 759-773. https://doi.org/10.1081/CSS-200043367

Viscarra, R. A., Adamchuk, V. I., Sudduth, K. A., McKenzie, N. J., & Lobsey, C. (2011). Chapter Five – Proximal Soil Sensing: An effective Approach for Soil measurements in Space and Time. *Advances in Agronomy*, 113, 243-291. https://doi.org/10.1016/B978-0-12-386473-4.00005-1

Viscarra Rossel, R. A., & McBratney, A. B. (1998). Soil chemical analytical accuracy and costs: implications from precision agriculture. *Australian Journal of Experimental Agriculture*. 38, 765-775. https://doi.org/10.1071/EA97158

Wadoux, A. M. J.-C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, 210, 103359. https://doi.org/10.1016/j.earscirev.2020.103359

Wadoux, A. J. –C., & McBratney, A. B. (2021). Hypotheses, machine learning and soil mapping. *Geoderma*, 383, 114725. https://doi.org/10.1016/j.geoderma.2020.114725

Wadoux, A. M. J.-C., Román-Dobarco, M., & McBratney, A. B. (2021a). Perspectives on data-driven soil research. *European Journal of Soil Science*, 72, 1675-1689. https://doi.org/10.1111/ejss.13071

Wadoux, M. J. –C., Heuvelink, G. B. M., Lark, R. M., Lagacherie, P., Bouma, J., Mulder, V. L., Libohova, Z., Yang, L., & McBratney, A. B. (2021b). Ten challenges for the future of pedometrics. *Geoderma*, 401, 115155. https://doi.org/10.1016/j.geoderma.2021.115155

Weindorf, D. C., Zhu, Y., Chakraborrty, S., & Huang, B. (2012). Use of portable X-ray fluorescence spectrometry for environmental quality assessment of peri-urban agriculture. *Environmental Monitoring and Assessment*, 184, 217-227. https://doi.org/10.1007/s10661-011-1961-6

Weindorf, D. C., Bakr, N., & Zhu, Y. (2014). Chapter one – Advances in Portable X-ray Fluorescence (PXRF) for Environmental, Pedological, and Agronomic Applications. *Advances in Agronomy*. 128, 1-45. https://doi.org/10.1016/B978-0-12-802139-2.00001-9

Weindorf, D. C., & Chakraborty, S. (2020). Portable X-ray fluorescence spectrometry analysis of soils. *Soil Science Society of America Journal*, 84, 1384-1392. https://doi.org/10.1002/saj2.20151

Zhai, Z., Martínez, J. F., Beltran, V., & Marínes, N. L. (2020). Decision support systems for agriculture 4.0: Survery and challenges. *Computers and Electronics in Agriculture*, 170, 105256. https://doi.org/10.1016/j.compag.2020.105256

Ågren, A. M., Larson, J., Paul, S. S., Laudon, H., & Lidberg, W. (2021). Use of multiple LIDAR-derived digital terrain indices and machine learning for high-resolution national-scale soil moisture mapping of the Swedish forest landscape. *Geoderma*, 404, 115280. https://doi.org/10.1016/j.geoderma.2021.115280

Populärvetenskaplig sammanfattning

Marken har många egenskaper som är viktiga för grödors tillväxt, och i vissa fall även för vår egen hälsa när vi konsumerar skördeprodukterna. En av dessa egenskaper är innehållet av spårelement. Spårelement är grundämnen som förekommer i relativt låg halt i naturen. Grödor tar upp spårelement ur jorden. Vissa spårelement är mikronäringsämnen, till exempel koppar och zink. Mikronäringsämnen är nödvändiga för att växten ska kunna växa och utvecklas normalt. Underskott av koppar eller zink i växt kan leda till symptom såsom dåligt utvecklade organ eller hämmad rottillväxt. I ett jordbrukssammanhang kan detta leda till minskad skörd av grödor. Vissa spårelement, t.ex. kadmium, är dock inte nyttiga för oss slutkonsumenter. Ökat intag av kadmium via maten har kopplats till benskörhet och cancer hos människor. Detta innebär att det är bra om halter av dessa spårelement i matjord kan kartläggas, för att få ett bättre beslutsstöd i växtproduktionen. Till exempel, för att veta var det kan behövas koppargödsling om halten i matjorden bedöms vara för låg. Information om kadmium i matjord skulle kunna hjälpa oss att hitta områden där låga halter i gröda är sannolika.

När intentionen är att skapa bra kartor behövs oftast många jordprovsanalyser, vilket kan vara dyrt och tidskrävande. Ett möjligt alternativ är att använda handburen röntgenflourescens (PXRF)-teknik, vilket är en snabb, enkel och relativt billig metod att mäta total halt av spårelement i jord. Metoden fungerar genom att "skjuta" röntgenstrålning på ett prov, för att sedan registrera våglängderna på energin som kommer tillbaka. Dessa våglängder ger information om vilka spårelement som finns i provet och i vilka halter.

Digital markkartering, *digital soil mapping* (DSM) på engelska, är en metod att för kartera jordegenskaper, t.ex. halter av spårelement. Det man gör är att etablera statistiska sammanband mellan geografiskt täckande

hjälpvariabler och uppmätta halter av spårelement i jordprover för att skapa en modell. Hjälpvariabler kan vara t.ex. information om höjd över havet, årlig nederbörd eller marktexturklasser. Modellen kan användas för att kartera spårelement.

Den här avhandlingen undersökte om PXRF-mätningar kan användas för att bestämma halter av zink, koppar och kadmium i jordprover. Avhandlingen handlade även om digital markkartering av koppar och kadmium, och hur dessa kartor kan användas.

Resultat från PXRF-modellering visade att det var möjligt att bestämma halter av zink, koppar och kadmium med modeller som kombinerar maskininlärning och PXRF-mätningar. Zinkhalter framtagna på detta sätt var jämförbara med dem från konventionell laboratorieanalys, medan skattningar av halter av kadmium och koppar var mindre träffsäkra men fortfarande användbara. Denna metod är ett intressant alternativ eller komplement till konventionell laboratorieanalys. Metoden användes för att utöka mängden jordprover med kadmium- och kopparhalt för den digitala markkarteringen.

Digital markkartering av kadmium resulterade i en karta med skattad halt i olika delar av Skåne och beräkningar av hur säkra de framtagna värdena var. Jämförelse med data på halter i höstvetekärna visade ett samband mellan låg halt i kärna och låg halt i matjord. Det innebär att kadmiumkartan kan användas för att lokalisera områden lämpliga för produktion av höstvete med särskilt låg halt kadmium i kärna, till exempel för produktion av barnmat. Digital markkartering av koppar resulterade i en nationell karta med skattade halter och hur säkra de var. Utifrån detta kunde risk för kopparbrist med stor sannolikhet uteslutas i 47% av svensk matjord. De resterande 53% av svensk matjord har däremot medelstor till stor risk för kopparbrist och laboratorieanalys av jordprover krävs för säker bedömning.

Popular science summary

The chemical properties of soils are important for crop growth and to ensure food safety of crop-based products to end-consumers. One such chemical property is the concentration of trace elements in the soil. Trace elements are elements that occur in relatively small amounts in nature and are taken up by plants from the soil. Some trace elements, such as copper and zinc, are useful micronutrients in crop nutrition, and are needed by plants in order to function properly. Deficiency of copper or zinc in plants can lead to symptoms such as malformation of organs or limited root growth. In an agricultural context, this can lead to reduced crop yield. Some other trace elements in soil, such as cadmium, are especially undesirable for the end-consumer of plant products. Accumulation of these harmful trace elements can be toxic to the end-consumer. For example, elevated human intake of cadmium with food has been linked to bone brittleness and cancer. Therefore, the concentrations present in grain sold for food uses are strictly regulated by legislation. At both farm level and national level, it would be beneficial to have information about the concentrations of copper, zinc and cadmium in agricultural soils. For example, if copper concentrations in certain fields were known to be too low for crop health, copper could be added in trace amounts in fertiliser. Alternatively, spatially explicit information on cadmium concentrations in soil could help identify suitable soils for production of grains with particularly high quality requirements, e.g. for baby food production.

For the purposes of accurate and detailed soil mapping, many soil samples would need to be collected and analysed, which can be time-consuming and expensive. An alternative is to use portable X-ray fluorescence (PXRF) sensing, which is a quick, easy and relatively inexpensive method for direct measurement of total concentrations of trace elements in soil. This method works by ‘shooting’ X-rays at soil samples and recording the wavelengths of

the energy returned from different elements. The returning energy provide information on elements present in the sample and their concentrations.

Digital soil mapping (DSM) is a popular method for spatial estimation of soil properties, such as concentrations of trace elements. Digital soil mapping involves establishing statistical relationships between spatially extensive covariates (environmental information) and known concentrations of trace elements at point locations, to create a model. Examples of covariates are elevation, soil texture classes and annual rainfall. The model can then be used to predict concentrations of trace elements spatially, in order to produce a map.

This thesis explored the use of PXRF measurements to predict concentrations of zinc, copper and cadmium in soil samples. The thesis also explored the application and use of digital soil mapping of copper and cadmium in Swedish agricultural soil.

Results of PXRF modelling showed that it was possible to predict concentrations of zinc, cadmium and copper using PXRF measurements with machine learning models. Predicted concentrations of zinc were comparable to the values obtained in conventional laboratory analysis, while predictions of cadmium and copper were less accurate, but still usable. This methodology could be an interesting alternative or complement to conventional laboratory analysis. It was used in this thesis to predict concentrations of cadmium and copper in soil samples, in order to increase the number of soil samples for digital soil mapping.

Digital soil mapping of cadmium in Skåne County resulted in a map of Cd concentrations and of the uncertainty associated with those predicted concentrations. Comparisons with data on grain samples revealed that low cadmium concentrations in winter wheat grain were associated with low predicted concentrations in soil. Hence, the map could be used to identify agricultural soils especially suitable for growing winter wheat with low cadmium concentrations.

Digital soil mapping of copper resulted in a national map of concentrations and of the uncertainty associated with those predicted concentrations. This national map indicated that for 47% of agricultural soil in Sweden the risk of copper deficiency is very small, while for the remaining 53% of agricultural soil the risk of copper deficiency is medium to high. In these areas laboratory analysis of soil samples would be needed to more in detail determine the risk.

Acknowledgements

First and foremost, I would like to thank my principal supervisor, Kristin Piikki, for impeccable guidance, aid and patience throughout my studies. She was never too brash or too passionate, and was always cool, calm and collected. Kristin, it was really educational being your student, I learned a lot and hope many more students can benefit from your supervision in the future.

I would also like to thank my second-in-command supervisor, Jan Eriksson, who is really a treasure-trove of knowledge about soil. Getting feedback from Jan was always scary, but he taught me a lot and really helped me with writing scientific papers.

I must also thank Mats Söderström, who was not an official supervisor but extremely vital for realisation of my science shenanigans. Without his help and knowledge, the whole project would have fallen flat in a heartbeat.

Many thanks to Omran Alshihabi for the help with the PXRF measurements!

I am very grateful to the Region Västra Götaland together with SLU (contracts: RUN 2018-00141 and RUN 2021-00020) for making this project a reality.

I would also like to thank Sandra Wolters, Johanna Wetterlind, Bo Stenberg, Lena Engström, Sofia Delin, Karin Andersson, Henrik Stadig, C.G. Pettersson and others I met during my time at SLU. Even though a large part of my stay was during a pandemic and I am a socially awkward primate, you all made me feel at home and part of something bigger. Many thanks to Mary McAfee for the very quick and nice proofreading.

In my private sphere, I would like to thank my long-time partner and girlfriend, Ebba. Without you I would be lost. Oh how you have helped me in trying times, and you have made good times even better.

Many thanks also to my mum and dad. Your help and support has been invaluable - a luxury beyond measure.

I would also like to thank my good friend Joakim for many fun hours with gaming and merriment that helped me relax during this period. We sure do suck at Frisbee golf, but we have fun playing it. I would also like to mention Erik and Mikael for good companionship. Dota 2 with the boys was always a highlight.

Lastly, I would like to honour my late friend Niclas Ogeryd, who always said mathematics doesn't have to be so arduous. You were right, my friend, even though I struggle with it at times. You left too early. I did not do enough, I'm sorry.

Appendix

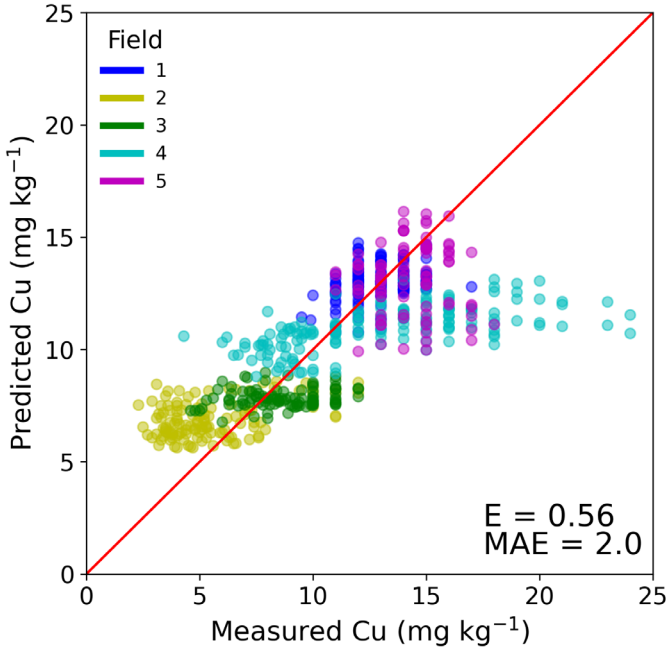


Figure A1: Scatter plot of validation results for the five fields on Bjertorp Farm. 'Predicted Cu' is the copper (Cu) concentration map produced in Paper III.

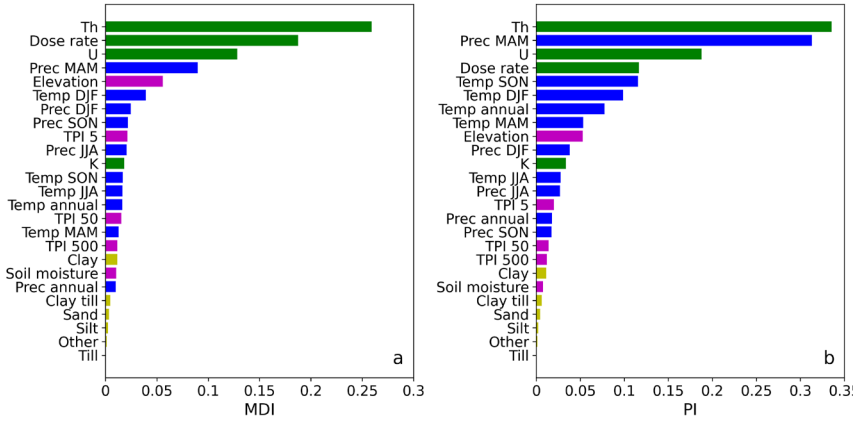


Figure A2: Covariate importance score and ranking using mean decrease in impurity (MDI) and permutation importance (PI) for the digital soil mapping (DSM) model for copper (Paper III). Each covariate type is colour-coded.

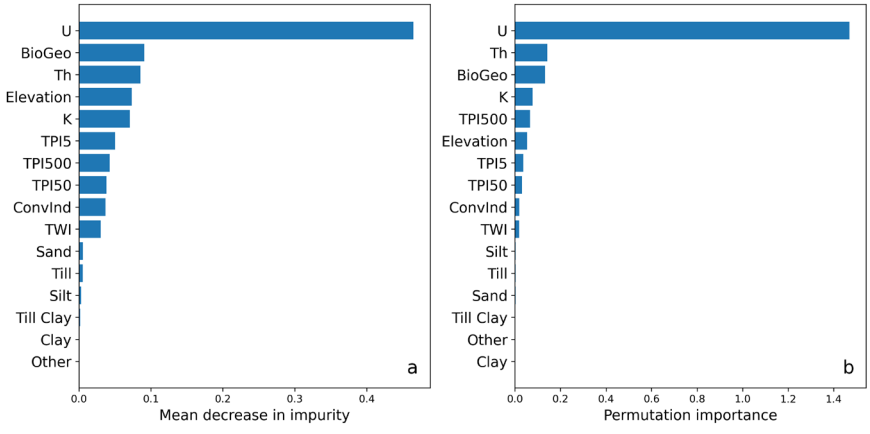


Figure A3: Covariate importance score and ranking using mean decrease in impurity and permutation importance for the digital soil mapping (DSM) model for cadmium (Paper II).

I

sensors

MDPI

Article

Predictions of Cu, Zn, and Cd Concentrations in Soil Using Portable X-Ray Fluorescence Measurements

Karl Adler *, Kristin Piikki, Mats Söderström, Jan Eriksson and Omran Alshihabi

Department of Soil and Environment, Swedish University of Agricultural Sciences, SE-75007 Uppsala/SE-53223 Skara, Sweden; kristin.piikki@slu.se (K.P.); mats.soderstrom@slu.se (M.S.); jan.o.eriksson@slu.se (J.E.); omran.alshihabi@slu.se (O.A.)

* Correspondence: Karl.Adler@slu.se

Received: 2 December 2019; Accepted: 12 January 2020; Published: 14 January 2020

Abstract: Portable X-ray fluorescence (PXRF) measurements on 1520 soil samples were used to create national prediction models for copper (Cu), zinc (Zn), and cadmium (Cd) concentrations in agricultural soil. The models were validated at both national and farm scales. Multiple linear regression (MLR), random forest (RF), and multivariate adaptive regression spline (MARS) models were created and compared. National scale cross-validation of the models gave the following R^2 values for predictions of Cu ($R^2 = 0.63$), Zn ($R^2 = 0.92$), and Cd ($R^2 = 0.70$) concentrations. Independent validation at the farm scale revealed that Zn predictions were relatively successful regardless of the model used ($R^2 > 0.90$), showing that a simple MLR model can be sufficient for certain predictions. However, predictions at the farm scale revealed that the non-linear models, especially MARS, were more accurate than MLR for Cu ($R^2 = 0.94$) and Cd ($R^2 = 0.80$). These results show that multivariate modelling can compensate for some of the shortcomings of the PXRF device (e.g., high limits of detection for certain elements and some elements not being directly measurable), making PXRF sensors capable of predicting elemental concentrations in soil at comparable levels of accuracy to conventional laboratory analyses.

Keywords: PXRF; soil; copper; zinc; cadmium; machine learning; precision agriculture

1. Introduction

Mapping concentrations of micronutrients or toxic elements in agricultural soil is important but is not commonly done. This kind of information could be useful in precision agriculture, where the goal is optimal management in space and time [1]. For instance, zinc (Zn) and copper (Cu) are important elements in crop production due to their roles in photosynthesis, respiration, and other plant functions [2,3]. However, excessively high concentrations can be toxic for crops (e.g., an excessive concentration of Cu can lead to malformation of root systems) [3]. Hence, there is a need to detect both low and high concentrations. Cadmium (Cd) is also toxic to consumers of crop products above certain threshold concentrations [2]. Thus, it can be useful to map Zn, Cu, and Cd at the field scale in order to rectify deficiencies and toxicities, and to safeguard crop quality and food safety. At present, there are no public field-scale maps of these elements in Sweden.

In Sweden, deficiency of Cu in crops is known to occur in sandy and organic soils [4], whereas availability of Zn is regarded as less of a problem. However, Zn deficiency in agricultural soil is a common problem in many other parts of the world [5]. Very high concentrations of Cd are typically related to the soil's parent material, which can vary substantially within an agricultural field [6]. In Sweden, a soil is deemed to be at risk of Cu deficiency at concentrations below 6–8 mg kg^{-1} [7]. There are no regulations governing Cd concentration in agricultural soil, but there are national laws

that prohibit application of sewage sludge when soil concentrations are above the stated limits for Cu (40 mg kg^{-1}), Zn (100–150 mg kg^{-1}), and Cd (0.4 mg kg^{-1}) [7,8].

To derive accurate maps of elemental concentrations in soil, many soil samples need to be analyzed. The conventional method involves element extraction with acids followed by analysis using the inductively coupled plasma (ICP) technique [9,10]. However, wet chemistry laboratory analyses can be expensive, time-consuming, and destructive to the sample [9,10]. The portable X-ray fluorescence (PXRF) technology is becoming an interesting option as it is a cheap, fast, and non-destructive method for analyzing element concentrations in soil samples [11]. This makes it very suitable for tasks where high sampling density is needed (e.g., mapping and geostatistics) [12]. The method works by exciting atoms with an energy source from the PXRF device, often an X-ray tube [13]. The atoms then emit X-ray fluorescence at specific wavelengths depending on the element in question, which is then measured by a sensor in the PXRF device [13]. The method can be accurate when combined with a simple preparation of the soil sample, and can provide high-quality data comparable to those obtained with conventional methods for quantification of certain elements in soil samples [14]. The PXRF technology is recognized as an official method for analyzing trace elements in soil by the United States Environmental Protection Agency (U.S. EPA) [15].

The aims of the present study were to:

- Use PXRF measurements to create national models for prediction of soil Cu, Zn, and Cd concentrations in agricultural soils;
- Validate these models at the national scale using cross-validation, and at the farm scale using an independent dataset;
- Compare the performance of three model types: multiple linear regression (MLR), multivariate adaptive regression splines (MARS), and random forest regression (RF);
- Test whether the best model for Cu can accurately predict whether a sample has concentrations above or below recommended levels;
- Test whether the best model for each element can accurately predict whether a soil sample has Cu, Zn, and Cd concentrations above or below the permissible level for sewage sludge application to agricultural soil.

2. Materials and Methods

2.1. Soil Sampling

The study area included all agricultural land in Sweden. Swedish crop production (mostly small-grain crops, oilseeds, pastures, and meadows covering about 2.5 Mha) is mainly concentrated in young, marine, and lacustrine post-glacial sediments from the time after the Weichselian glaciation [16]. More than 90% of the agricultural area is located in the southern area of the country (the sample distribution in Figure 1 accurately depicts the occurrence of arable land). Eutric and dystric cambisols are the dominant cropland soil types [16]. Cropland soil texture ranges from heavy clays in the eastern parts, to loam and sandy loam generally dominating in the south and southwestern agricultural areas [16–18]. For a general soil and texture map of Sweden, see Figures 2 and 4 by Eriksson et al. [19]. For an overview of topsoil properties of arable land in Sweden, see maps by Eriksson et al. (pp. 75–90) [7]. Descriptive statistics of the soil properties in the calibration samples are presented in Table 1.

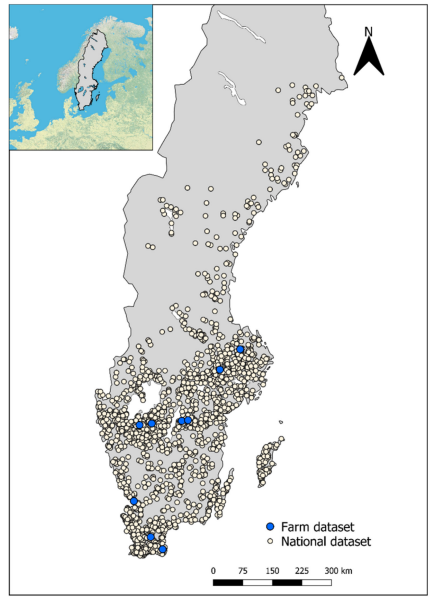


Figure 1. Map of Sweden showing soil sampling locations used in the present study. Farm dataset refers to the nine farms that were used for independent validation of the models. National dataset refers to the calibration samples. Base map courtesy of Environmental Systems Research Institute (ESRI) (Redlands, CA, USA).

Table 1. The minimum, maximum, mean, median, and standard deviation (SD) of cation exchange capacity (CEC) at pH 7 ($cmol_c$ kg^{-1}) for base saturation (%), soil organic matter (SOM) (%), clay content (%), and pH in the topsoil samples of arable land in Sweden used in the analyses (n = 1520).

	Minimum	Maximum	Mean	Median	SD
CEC	3	70	17	15	8
Base saturation	8	100	69	72	21
SOM	0.8	16.6	4.5	4.2	1.8
Clay content	2	80	23	19	15
pH	4.5	8.4	6.2	6.2	0.6

The total number of topsoil samples available from the national monitoring program for arable soils in Sweden was 1833 [7]. Sampling locations in the monitoring program were selected using a random stratified sampling design covering all arable land [7]. Soil samples from nine farms (n = 179, ~ 20 from several fields per farm) were used for validation at the farm scale (Figure 1). The nine farms were originally selected for a previous study in order to represent a wide range of Cd concentrations and different geologies, based on maps presented in Eriksson et al. [7]. Each soil sample consisted of nine subsamples collected with an auger at a depth of 0–20 cm within a 3–5 m radius of the sample coordinates. The soil samples were air-dried, homogenized, and sieved (< 2 mm) prior to analysis.

2.2. PXRF Measurements

The soil samples were analyzed ex situ using a Niton XL3t GOLDD+PXRF device with a geometrically optimized large area drift detector and an Ag anode that operates at 50 kV and 200 µA (Thermo Scientific, Billerica, MA, USA), which were connected to a computer and mounted on a static frame specially designed for the PXRF device (Thermo Scientific, Billerica, MA, USA).

The PXRF device was set in "soil mode", an instrument-specific measurement configuration optimized for soil materials, and measurement time was set to 180 s per sample. Each soil sample was dried, homogenized, and sieved (< 2 mm) according to recommendations for ex situ PXRF analysis [13,15]. Each soil sample was placed in a 32-mm double-ended XRF sample cup (filled to three-quarters volume) with a 4-μm thick transparent polypropylene XRF film in line with U.S. EPA standards [15] and placed on the PXRF aperture. The reference standard 2709a from the National Institute of Standards and Technology (NIST) was measured four times during the project to check the measurement stability of the PXRF device (see Supplementary Materials). Measurements were found to be stable over the course of the project.

The limit of detection (LOD) was set at three times the standard deviation of the measurement. The PXRF device measured each second for the duration of measurement (180 s). Hence, a final concentration and standard deviation were provided for the element in question when the measurement was completed. As each measurement has its own individual standard deviation for an element, there is no common LOD for an element. Measured values below this limit were denoted "not a number" (NaN). Only elements with < 10% NaN values in the national dataset were included in the modelling to ensure that the measured concentrations of elements used as explanatory variables were generally above the LOD of the PXRF device. Hence, future measurements with a similar PXRF device can be used with high probability. All samples that exhibited NaN values for any of the included elements were excluded from the modelling. The total number of samples used for calibration was 1520 (Soil properties of these samples can be seen in Table 1).

2.3. Laboratory Analyses

Pseudototal concentrations of Cu, Zn, and Cd in the soil samples were determined by extraction with 7M HNO_3 in an autoclave at 120 °C for 30 min, as stated by Swedish standard SS 28 31 11 [20]. Measurement was performed using inductively coupled plasma atomic emission spectroscopy (ICP-AES) for Zn and Cu, and inductively coupled plasma mass spectrometry (ICP-MS) for Cd. Hereafter, "lab-analyzed" refers to results obtained with this extraction and analysis method.

2.4. Modelling

2.4.1. Model Selection

Three different machine learning algorithms were chosen for modelling Zn, Cu, and Cd concentrations, namely MLR, RF, and MARS. The intention was to have a simple linear model (MLR) and two distinct non-linear models (RF and MARS). The RF and MARS algorithms produce non-linear models with discrete and continuous predictions respectively. RF consists of an ensemble of decision trees with bagging, where each decision tree is made from a partitioning algorithm based on conditional statements. The term bagging means creating several decision trees from different subsets of the data, making the final predicted value the mean value of several tree models [21]. MARS is based on building several piecewise linear regression models (basis functions) that are valid within certain intervals of the explanatory variables and defined by hinge functions [21]. The MARS algorithm first creates basis functions in a forward pass, later to be pruned in a backwards pass to reduce model complexity and risk of overfitting [21]. For a more detailed description of MLR, RF, and MARS, see Hastie et al. [21].

2.4.2. Model Implementation

The MLR and RF algorithms were implemented using the Scikit-learn machine learning package (version 0.19.1) for Python [22]. MARS was implemented using the Py-earth package (version 0.1.0) for Python, originally made for the R programming language [23]. Both RF and MARS were used in their default setting. For example, MARS was set as default to be additive. This was done to reduce overfitting and make the models more robust. The only hyperparameter set was with the RF models,

as the number of bagged trees needed to be specified (number of trees was set to 100). For a complete description of the default settings, see the respective model descriptions in the Py-earth and Scikit-learn packages. Predictions of negative concentrations were set to 0 mg kg^{-1}.

2.4.3. Validation

The performance metrics used were the mean absolute error (MAE) and the coefficient of determination (R^2), often named the Nash–Sutcliffe model efficiency coefficient [24], which is defined as in Equation (1):

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2} \quad (1)$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\overline{y}$ is the mean of the actual values of the response variable.

Cross-validation was performed on the national dataset using the leave-one-out method for each MLR, RF, and MARS model. Validation metrics are also presented based on how well the Cu and Cd models performed at lower concentrations (arbitrarily chosen range of interest (ROI)) of 0–20 mg kg^{-1} and 0–0.5 mg kg^{-1}, respectively. This was done to generate validation statistics that give a better understanding of how well the predictions perform around concentrations of practical interest.

In addition, confusion matrices were created to assess whether the models could be used to determine if a soil element concentration is above or below a given threshold in the cross-validation for Cu deficiency and sewage sludge application with regard to Cu, Zn, and Cd concentrations. The upper boundary of the Cu deficiency threshold was used (8 mg kg^{-1}). The models chosen for this task were those that performed best in terms of R^2 in the cross-validation for each element. Agreement of the predictions was calculated according to Equation (2):

$$Agreement = \frac{(Tp + Tn)}{(Tp + Tn + Fp + Fn)} \quad (2)$$

where Tp is the total number of positive predictions, Tn is the total number of negative predictions, Fp is the number of false positive predictions, and Fn is the number of false negative predictions. Positive predictions refer to values below the threshold and negative predictions refer to values above the threshold.

3. Results

3.1. Descriptive Statistics of PXRF Measurements of the National Set of Soil Samples

Thirteen elements proved to be useful as explanatory variables of element concentrations in the 1833 samples (Table 2). The element closest to the threshold of < 10% NaN readings was Cs (9.8% NaN), followed by barium (Ba) (3.9%), lead (Pb) (2.2%), vanadium (V) (1.4%), manganese (Mn) (0.4%), and Zn (0.2%). The remaining elements shown in Table 2 had no NaN readings. Descriptive statistics of the elements used to calibrate the MLR, RF, and MARS models are also presented in Table 2. The descriptive statistics minimum, maximum, mean, median, and standard deviation (SD) were calculated after removal of samples with NaN values in any of the included variables, which resulted in exclusion of 313 samples out of the original 1833 samples (i.e., 1520 samples were used for modelling). The majority of the samples excluded had readings below the LOD for Cs.

Table 2. Descriptive statistics of the elements used for modelling after removal of samples with "not a number" (NaN) classification in any of the elements included (n = 1520). Minimum, maximum, mean, median, and standard deviation (SD) are presented as mg kg^{-1}, where values < 1000 were rounded to the closest integer and values > 1000 to three significant digits. Rec = mean recovery rates from four measurements based on reference standard 2709a from the National Institute of Standards and Technology (NIST) (%); Rec-SD = standard deviation of the four recovery rates (%).

Element	Minimum	Maximum	Mean	Median	SD	Rec	Rec-SD
Pb	8	146	19	18	7	63	10.8
Cs	10	56	33	34	9	970	33.1
Zn	16	518	72	67	32	92	2.1
V	33	411	93	90	30	123	18.6
Rb	32	181	104	100	26	83	0.8
Sr	71	378	142	132	49	92	0.8
Zr	71	955	251	240	77	65	0.9
Ba	197	1140	491	487	98	87	2.4
Mn	124	6000	542	481	345	97	2.7
Ti	1630	6890	3860	3880	765	114	1.8
Ca	2980	196,000	11,100	9710	9390	105	1.5
Fe	4370	93,000	21,500	19,300	9760	84	0.6
K	11,400	36,200	24,100	24,300	4180	96	1.3

A total of 99% of Cd measurements and 55% of Cu measurements were NaN, indicating that this PXRF device cannot be used for direct measurement of Cd and Cu at the concentration range found in Swedish agricultural soil. The lowest concentration of Cu measured was approximately 20 mg kg^{-1}, indicating that this is perhaps the lowest possible Cu concentration that can be measured with this PXRF device.

The PXRF device measured values similar to known concentrations of the included elements in NIST 2709a (Table 2). Concentrations of some elements, such as Cs and Pb, were overestimated and underestimated, respectively. However, the stability of the measurements, as shown by the standard deviation of the recovery rates, shows that the PXRF measurements can be used for modelling, as the coefficients in the calibrated models will be valid over time. Measurements of Cs had the least stability according to the standard deviation of the recovery rates, but still only fluctuated by 1–3 mg kg^{-1} (see Supplementary Materials).

3.2. Descriptive Statistics of the National and Farm Datasets

Descriptive statistics of lab-analyzed Cu, Zn, and Cd concentrations for the national dataset (calibration and cross-validation data) and the farm dataset (validation data) are shown in Table 3.

Table 3. Descriptive statistics of lab-analyzed copper (Cu), zinc (Zn), and cadmium (Cd) for the calibration data (national dataset, n = 1520) and validation data (farm dataset, n = 179). Minimum, maximum, mean, median, and standard deviation (SD) are presented as mg kg^{-1} rounded to the closest integer, apart from those for Cd.

Lab-Analyzed Element	Minimum	Maximum	Mean	Median	SD
National dataset					
Cu	2	130	14	11	10
Zn	6	557	61	56	33
Cd	0.04	4.07	0.20	0.17	0.17
Farm dataset					
Cu	3	77	22	17	19
Zn	22	135	72	67	30
Cd	0.06	1.60	0.37	0.21	0.38

The national and farm datasets differed in their frequency distributions of concentrations of Cu, Zn, and Cd. The mean and median showed that the farm dataset generally had higher concentrations of Cu, Zn, and Cd than the national dataset. For example, the national dataset contained five samples with Cu concentrations above 60 mg kg^{-1}, while the farm dataset contained 19. Similarly, eight samples in the national dataset had Cd concentrations above 1 mg kg^{-1}, while there were 25 such samples in the farm dataset. There were, therefore, more samples with higher concentrations of Cu and Cd in the farm dataset than in the national dataset, even though the farm dataset was much smaller. In the national dataset, high concentrations were, therefore, less common in the case of Cu and Cd. For Zn there were 139 samples with concentrations above 100 mg kg^{-1} in the national dataset, while there were 41 in the farm dataset. This implies that the Zn concentrations measured on the selected farms were more similar to those in the national dataset than the measured concentrations of Cu and Cd

3.3. Cross-Validation

In Figure 2, cross-validated leave-one-out predictions of concentrations from the MLR, RF, and MARS models for each element are plotted against lab-analyzed concentrations for the national dataset. The cross-validation results showed that it was possible to predict concentrations beneath the LOD for Cu, which was approximately 20 mg kg^{-1}. However, the RF models could not predict concentrations as low as those predicted by the continuous MLR and MARS models, as is apparent for Cu and Cd predictions with the RF models (Figure 2). For instance, the RF model for Cu could only predict concentrations down to approximately 5 mg kg^{-1}, while the MLR and MARS models could predict lower concentrations. There was no major visual difference between the performance of the MLR and MARS models for Cu (Figure 2). All three models for Zn imposed a fit close to the 1:1 line. The MLR model for Cd produced errors in the higher range of concentrations, while the RF and MARS models predictions at higher concentrations exhibited negative bias. However, the MARS model for Cd gave smaller errors at lower concentrations than the MLR model for Cd (Table 4).

For ease of comparison, the same range of values is shown in the farm-scale validation (Figure 3) and in the national-scale validation (Figure 2). This means that some values outside the range are not shown in Figure 2 (1, 20, and 5 values for Cu, Zn, and Cd, respectively).

Table 4. Validation statistics from the cross-validation of the multiple linear regression (MLR), random forest regression (RF), and multivariate adaptive regression spline (MARS) models for copper (Cu), zinc (Zn), and cadmium (Cd). R^2 = coefficient of determination; MAE = mean absolute error (mg kg^{-1}); ROI = range of interest (0–20 mg kg^{-1} for Cu and 0–0.5 mg kg^{-1} for Cd).

Model	R^2	MAE	R^2-ROI	MAE-ROI
Cu-MLR	0.58	3.87	0.06	3.00
Cu-RF	0.63	3.48	0.20	2.69
Cu-MARS	0.59	3.72	0.04	2.94
Zn-MLR	0.92	5.60	-	-
Zn-RF	0.86	5.93	-	-
Zn-MARS	0.92	5.63	-	-
Cd-MLR	0.49	0.065	−0.17	0.057
Cd-RF	0.48	0.053	0.40	0.043
Cd-MARS	0.70	0.054	0.20	0.047

Validation statistics revealed that the MARS models generally performed best except for with Cu, for which the RF model performed best, based on R^2 and MAE (Table 4). Cross-validation revealed two problems with the continuous models MLR and MARS, especially in the ROI for Cu and Cd. The first was the impact on strange predictions of concentrations for certain samples (e.g., in terms of R^2, the MLR model for Cu exhibited little accuracy). However, removal of a single poorly predicted sample resulted in an increase in R^2 from 0.06 to 0.13. For the MLR model for Cd, R^2 increased from −0.17 to 0.01 with removal of the same sample. The second problem involved predictions below

0 mg kg^{-1}. The numbers of samples with predicted concentrations below 0 mg kg^{-1} were 10 and 8 for the MLR and MARS models for Cu, respectively. The numbers of samples with Cd predictions below 0 mg kg^{-1} were 12 and 1, respectively, for the MLR and MARS models. Hence, predictions below 0 mg kg^{-1} were uncommon. In the ROI, there were 1196 and 1487 samples for Cu and Cd, respectively.

These problems of predictions of outlier samples and predicted negative concentrations were not observed with the discrete predictions of the RF model, as it cannot extrapolate beyond the calibration data.

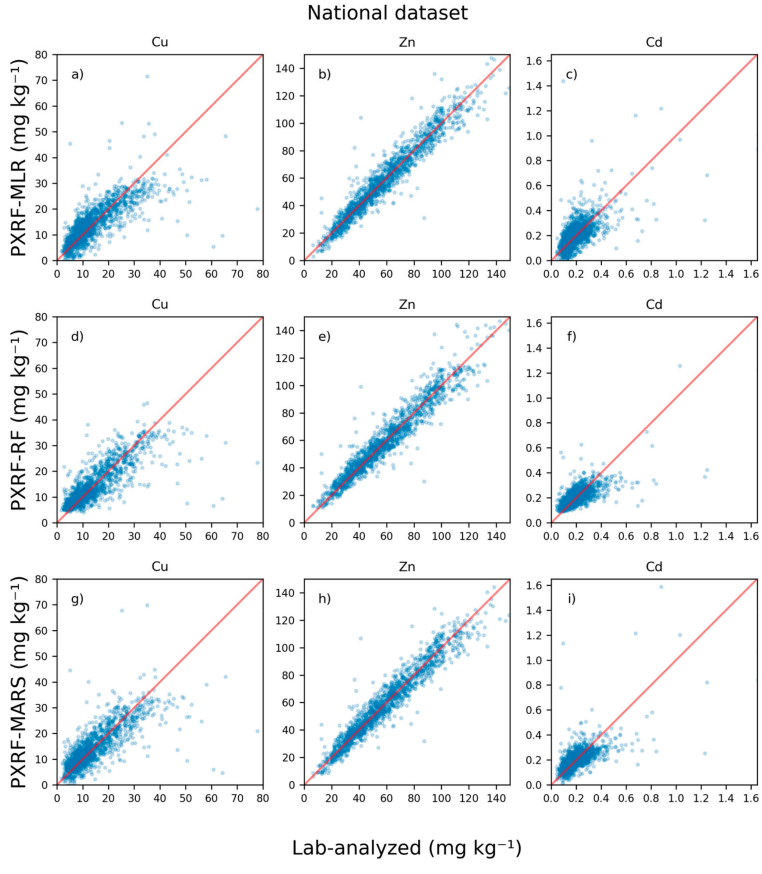


Figure 2. Concentrations of copper (Cu), zinc (Zn), and cadmium (Cd) predicted from portable X-ray fluorescence (PXRF) measurements using multiple linear regression (MLR), random forest regression (RF), and multivariate adaptive regression splines (MARS) for national-scale data using leave-one-out cross-validation compared with 7M HNO_3 extraction and inductively coupled (ICP) analysis. The symbols are semi-transparent to show point density.

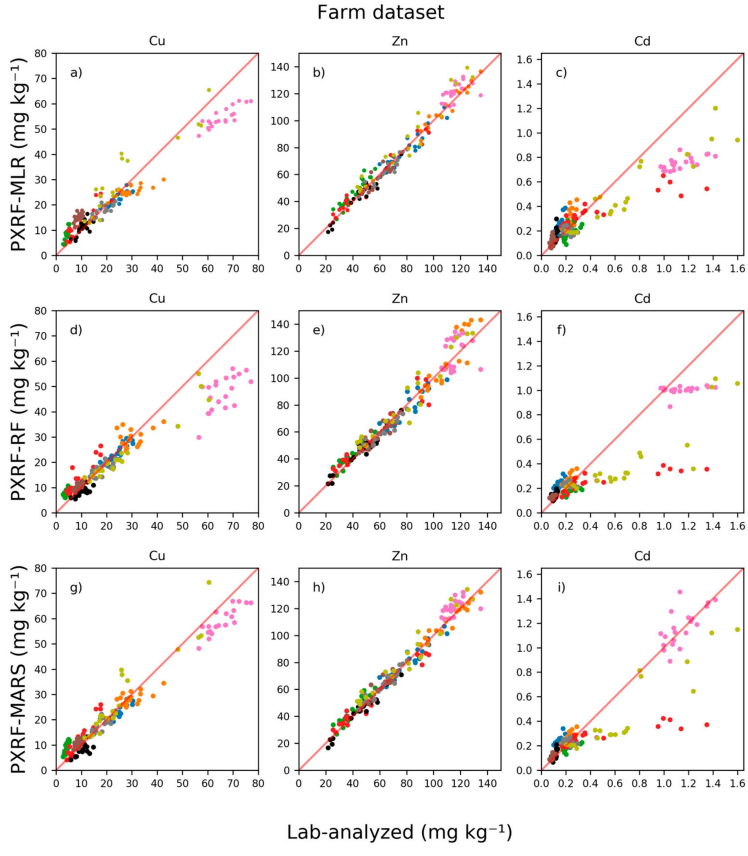


Figure 3. Concentrations of copper (Cu), zinc (Zn), and cadmium (Cd) predicted from portable X-ray fluorescence (PXRF) measurements using multiple linear regression (MLR), random forest regression (RF), and multivariate adaptive regression splines (MARS) on the farm dataset compared with 7M HNO_3 extraction and inductively coupled (ICP) analysis. The models were calibrated at the national scale and applied on the farm dataset. Each color represents a specific farm.

3.4. Validation at the Farm Scale

The MLR, RF, and MARS models of concentrations for each element in the farm-scale validation are compared with the lab-analyzed concentrations in Figure 3, where each farm is represented by a specific color (see full equations for the MLR models in the Supplementary Materials). Validation statistics are shown in Table 5. All models were able to predict below the LOD of the PXRF device for Cu, as also seen in the cross-validation. At lower concentrations, the MLR model for Cu had a general positive bias, while at higher concentrations it had a general negative bias. The RF and MARS models for Cu also exhibited negative bias for predictions at higher concentrations, with MARS having the least negative bias. However, at lower concentrations the RF and MARS models exhibited less positive bias in predictions than the MLR model. Farms with lower concentrations showed a smaller spread in predicted values with the RF model for Cu (i.e., the farms represented by green symbols) compared with the MLR and MARS models (Figure 3). However, as in the cross-validation, the RF model could not predict concentrations as low as those predicted by the MLR and MARS models for Cu, which resulted in a positive bias in predictions for farms with low concentrations of Cu. All models, though

especially the MARS model, were able to predict variations in Cu concentrations on certain farms with ranges of Cu concentrations.

Table 5. Validation statistics from the farm dataset of the multiple linear regression (MLR), random forest regression (RF), and multivariate adaptive regression spline (MARS) models for copper (Cu), zinc (Zn), and cadmium (Cd). R^2 = coefficient of determination; MAE = mean absolute error (mg kg^{-1}); ROI = range of interest (0–20 mg kg^{-1} for Cu and 0–0.5 mg kg^{-1} for Cd).

Model	R^2	MAE	R^2-ROI	MAE-ROI
Cu-MLR	0.90	4.40	0.12	3.56
Cu-RF	0.84	4.51	0.54	2.43
Cu-MARS	0.94	3.21	0.47	2.72
Zn-MLR	0.96	4.40	-	-
Zn-RF	0.94	5.40	-	-
Zn-MARS	0.97	4.00	-	-
Cd-MLR	0.74	0.121	0.34	0.052
Cd-RF	0.74	0.109	0.44	0.050
Cd-MARS	0.80	0.087	0.50	0.043

The MLR model for Zn was able to predict throughout the range with relatively high accuracy and perhaps a slight positive model bias. The MLR model for Cd showed similar problems to the MLR model for Cu, as high concentrations could not be predicted and there were errors in prediction at lower concentrations. In general, all models showed equally good performance, with more or less bias in the predictions in some cases.

The MLR and RF models for Cd exhibited positive bias in the predictions at lower concentrations, while the MARS model exhibited the least positive bias. However, as can be seen from Figure 3, three specific farms were difficult to predict. One farm, colored red, had Cd concentrations ranging from about 0.5 to 1.4 mg kg^{-1}. The Cd concentrations on this farm could not be accurately predicted by any model tested in this study. However, sites on the farm exhibiting the highest concentrations of soil Cd, colored pink, were those most accurately predicted by the MARS model.

Validation metrics of the prediction at the farm scale are presented in Table 5. These include validation metrics on how well the Cu and Cd models performed in the ROI, for which there were 102 and 140 samples for Cu and Cd, respectively.

Based on the metrics, the nationally calibrated MARS models performed best of the models tested in the farm-scale validation for Cu, Zn, and Cd (Table 5). However, within the ROI the RF model for Cu performed better than the MARS model for Cu. This can also be inferred from Figure 3, where predictions for each farm with lower concentrations showed less spread. The MARS model performed better for the whole range than the RF model for Cu. All the models for Zn at the farm scale performed very similarly, but with the MARS model the performance was the best, as also observed in the cross-validation. Using the Zn values measured on the farms with the PXRF device compared with lab-analyzed values resulted in R^2 = 0.81, which was lower than that of the MLR, RF, and MARS models for Zn (Table 5). The best model for predicting Cd, for the whole range and within the ROI, was the MARS model, as also found in the cross-validation.

3.5. Testing Performance for Fertilization and Sewage Sludge Fertilization

Confusion matrices of predictions in relation to actual concentrations above or below threshold concentrations for Cu, Zn, and Cd in the cross-validation are presented in Table 6. The thresholds are based on the recommendations for Cu fertilization and permissible levels of soil Cu, Zn, and Cd concentrations, above which sewage sludge application is prohibited [4,7]. The models used were those identified as the best based on the coefficient of determination, presented in Table 4. Thus, the RF model was used for Cu, MLR was used for Zn, and MARS was used for Cd.

Table 6. Confusion matrices for classifications above and below thresholds for copper (Cu) fertilization and sewage sludge application for Cu, zinc (Zn), and cadmium (Cd) using the best models for each element in the cross-validation. Swedish recommendations suggest that there is risk of Cu deficiency if the Cu concentration in the soil is below 8 mg kg^{-1}, while sewage sludge application is prohibited if the concentrations of Cu, Zn, and Cd exceed 40, 100, and 0.4 mg kg^{-1}, respectively.

Cu Fertilization		Lab-Analyzed		Total
		Below Threshold	Above Threshold	
Predicted	Below Threshold	224	70	294
	Above Threshold	200	1026	1226
Total		424	1096	
Cu Sewage Sludge		**Lab-Analyzed**		**Total**
		Below Threshold	**Above Threshold**	
Predicted	**Below Threshold**	1490	27	1517
	Above Threshold	2	1	3
Total		1492	28	
Zn Sewage Sludge		**Lab-Analyzed**		**Total**
		Below threshold	**Above Threshold**	
Predicted	**Below Threshold**	1337	21	1358
	Above Threshold	44	118	162
Total		1381	139	
Cd Sewage Sludge		**Lab-Analyzed**		**Total**
		Below Threshold	**Above Threshold**	
Predicted	**Below Threshold**	1437	49	1486
	Above Threshold	18	16	34
Total		1455	65	

The level of agreement between predicted and lab-analyzed values was 82% when predicting whether a soil was Cu-deficient or not (Table 6). However, there was higher accuracy in predicting soils that were not Cu-deficient in the national dataset (94% correctly classified) compared with those that were Cu deficient (53% correctly classified) (see Figure 2).

Assessment of samples regarding suitability for sewage sludge application revealed high agreement between predicted and lab-analyzed values for Cu, Zn, and Cd (98%, 95%, and 95%, respectively). This was especially true for predictions below the respective threshold. Most of the samples had concentrations below the permissible level for sewage sludge application (shown in Figure 2).

4. Discussion

The results in this study demonstrated that an approach based on PXRF measurements coupled with machine learning algorithms is capable of predicting concentrations of Cu, Zn, and Cd in non-organic (SOM < 20%; Table 1) Swedish agricultural soils that can be used for risk assessments. An interesting finding was that concentrations of elements that are difficult or impossible to measure directly with the PXRF device, such as Cu and Cd, can be indirectly predicted with predictor elements present in measurable concentrations in Swedish agricultural soil (shown in Table 2). For example, it was found that MLR modelling of Zn was better than only using direct measurements of Zn made with the PXRF device. However, the relatively accurate results obtained with the MLR model for Zn were attributable to some degree to PXRF-measured Zn being included as an explanatory variable. Cd concentrations were most difficult to predict accurately, as was evident for certain farms in the

farm-scale validation, for which medium and high concentrations could not be predicted without substantial errors. Hence, predictions of lower concentrations can be deemed more accurate.

The method presented for creating predictive models from PXRF measurements is a valid option (especially for Cu and Zn) when a dense sampling scheme is needed to create high-resolution maps of Cu, Zn, and Cd showing within-field variation. This can be a powerful tool in precision agriculture and for regional or national soil monitoring and mapping projects.

4.1. Cu Deficiency

There are certain ranges of Cu concentrations that are especially interesting for Swedish agriculture. According to Swedish recommendations [4], a soil is deemed to be at risk of Cu deficiency when the concentration is below 6–8 mg kg^{-1}. Indications of soil Cu status could be obtained using the MARS and RF models for Cu, where the predictions could be used to assess whether a soil is at risk of being Cu-deficient or not, considering the high model agreement and MAE in the ROI (Tables 5 and 6). Using PXRF, Hu et al. [14] obtained accurate measurements of Cu comparable to those in laboratory analysis (R^2 = 0.67). In this study, we achieved substantially higher R^2 relative to laboratory analysis when predicting Cu (up to R^2 = 0.94). Hence, using PXRF measurements for prediction appears promising. However, the MAE in the ROI was 2–3 mg kg^{-1} depending on the model used, so predictions around the threshold of 6–8 mg kg^{-1} should be viewed with caution and complementary conventional laboratory analysis should be conducted. For example, it should be noted that the results presented in Table 6 are binary, while the input data for this classification were not. This means that if a sample is predicted to have a concentration of 8.1 mg kg^{-1} (i.e., slightly above the threshold), the prediction will be classified as incorrect. For example, if a predicted sample was deemed to be correctly predicted up to 9 mg kg^{-1}, the number of correctly predicted samples increased from 224 to 294 with the RF model for Cu.

4.2. Sewage Sludge Application

The prediction models could be used to determine whether sewage sludge may be applied in an agricultural field. For example, the best model for predicting soil Cd concentrations had an MAE of 0.04 mg kg^{-1} in the ROI in the farm-scale validation with the MARS model, which makes it possible to determine whether an agricultural soil is at risk of excessive Cd concentrations. The results showed that Zn predictions were of high accuracy and good model agreement (Table 6). López-Núnez et al. [25] showed a similar high accuracy of predicted Zn in organic amendments with a linear model. Hence, predictions of Zn concentrations with PXRF appear highly suitable. This study showed that the MLR model is sufficiently accurate to predict whether sewage sludge application is permissible in relation to soil Zn concentrations. A similar level of agreement was found in the Cu and Cd predictions. However, most samples in the national dataset had Cu, Zn, and Cd concentrations below the threshold where sewage sludge application is legal. Hence, concentrations above the legal limits can be deemed as outliers in the distribution and the predictions at these concentrations should be viewed with caution. Similar to the results in Table 6 mentioned earlier, there might be a need to perform a conventional laboratory analysis when predicted concentrations are close to the thresholds for sewage sludge application.

4.3. Data, Model Selection, PXRF Methodology, and Variable Selection

The MAE in the national cross-validation was generally lower than in the farm-scale validation and the descriptive statistics showed that the farm dataset was somewhat unrepresentative of the national dataset. Thus, the farm-scale dataset can be regarded as rather difficult to predict accurately, since the farms included are quite unique in terms of their high Cd and Cu values (see Table 3). Hence, as shown by the results, predictions of mid- and high-range concentrations of Cd should be viewed with caution. The results indicated that other predictors from other sensors may be needed when there is little variation in concentrations measured by PXRF. However, some farms with varying

concentrations of Cu and Cd were predicted with accuracy, which indicates that the farm dataset is difficult to validate against in some cases.

The results showed that when using the models presented, some caution is needed. For instance, the RF models cannot predict concentrations as low as those predicted by the continuous MLR and MARS models. However, in rare instances MARS and MLR can predict non-sensical concentrations. The RF algorithm benefits greatly from a uniform distribution of concentrations in the calibration dataset in order to create classes throughout the range. In the present study, an insufficient number of classes was constructed by the RF model in the higher ranges for Cd. This implies that the accuracy of the RF model could be improved with more samples, so that more variations in soil Cd could be accounted for. Overall, the continuous models tested in this study appear more interesting as they allow more extrapolation in predictions.

A simple linear model such as MLR can be very effective, as seen with the predictions of Zn, and in some instances Cu. Non-linear models such as RF and especially MARS can be better overall options, as there is lower associated error in the predictions. Hence, depending on the range of concentrations to be predicted, either RF or MARS might be more or less suitable. For example, the RF model for Cu performed the best out of the Cu models at lower concentrations, but was unable to makes predictions as accurately as the MARS model at higher concentrations.

It should be noted that the PXRF measurements and the models were made on processed samples. This means that these models might not be appropriate when using PXRF in the field due to the sensitivities of the method to soil matrix factors, such as moisture and particle size distribution [26]. The parameterized models are, thus, calibrated for a specific soil matrix type, which in the present case was dry, homogeneous, and fine-grained soil. The PXRF device was used as a small, nimble laboratory device that is easy to use and provides ample amounts of data in terms of measured elements, in a shorter time, and at a lower cost than conventional laboratory analysis, even when used in an ex situ setting [14,27]. Based on the results in this study, the method will be tested in the future on a larger dataset of soil samples to create maps of the modelled elements. Hence, this study provides excellent groundwork for a future where these models are the foundation in mapping of soil Cu, Zn, and Cd concentrations in Sweden.

In the present study, no feature selection of predictor elements was performed. This was because with national models, the relationship between predictor elements and the target element in question can vary in space. For instance, when the present analysis was performed with half the dataset, particular relationships between elements were more prevalent, while these relationships were not present when the whole dataset was used. However, if regional models are to be created in future studies, feature selection might be necessary, as certain relationships depend on the soil type and underlying geology.

5. Conclusions

- Predictive models using PXRF measurements were created and found to be applicable at farm and national scales;
- The models were able to predict concentrations of Cu, Zn, and Cd in non-organic Swedish agricultural soils at both national and farm levels, but with varying amounts of error;
- Non-linear models proved most suitable for predicting concentrations of Cu and Cd, while the linear model for Zn yielded predictions with the same level of accuracy as the non-linear models;
- The accuracy of predictions means that the models created can be used to assess the risk of Cu deficiency. However, complementary laboratory analysis is advisable if predicted concentrations are close to the threshold value;
- The same applies for models created to assess whether an agricultural soil is eligible for sewage sludge application based on its Cu, Cd, and Zn concentrations.

Supplementary Materials: The following are available online at http://www.mdpi.com/1424-8220/20/2/474/s1: data from PXRF measurements on reference standards with the PXRF device used in this study and equations for the MLR models.

Author Contributions: Conceptualization, K.A., K.P., and M.S.; methodology, K.A., K.P., and M.S.; software, K.A.; validation, K.A.; formal analysis, K.A.; investigation, K.A.; data curation, K.A., O.A., and J.E.; writing—original draft preparation, K.A.; writing—review and editing, K.A., K.P., M.S., J.E., and O.A.; supervision, K.P., M.S., and J.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Västra Götaland Region (VGR) and the Swedish University of Agricultural Sciences (SLU), grant number: RUN 2018-00141.

Acknowledgments: We would like to thank Elin Laxmar, who produced part of the soil samples and data from PXRF-measurements used here during her master thesis work; and Lantmännen and the Swedish Environmental Protection Agency for funding the monitoring program from which the national data set was supplied.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bongiovanni, R.; Lowenberg-Deboer, J. Precision Agriculture and Sustainability. *Precis. Agric.* **2004**, *5*, 359–387. [CrossRef]
2. Mertens, J.; Smolder, E. Zinc. In *Heavy Metals in Soils: Trace Metals and Metalloids in Soil and their Bioavailability*; Alloway, B., Ed.; Springer: Dordrecht, The Netherlands, 2013; pp. 465–493.
3. Oorts, K. Copper. In *Heavy Metals in Soils: Trace Metals and Metalloids in Soil and their Bioavailability*; Alloway, B., Ed.; Springer: Dordrecht, The Netherlands, 2013; pp. 367–394.
4. Swedish Board of Agriculture. *Rekommendationer för Gödsling och Kalkning 2019 (In English: Recommendations for Fertilizing and Liming 2019)*; Swedish Board of Agriculture: Jönköping, Sweden, 2019.
5. Alloway, B.J. Soil factors associated with zinc deficiency in crops and humans. *Environ. Geochem. Health* **2009**, *31*, 537–548. [CrossRef] [PubMed]
6. Söderström, M.; Eriksson, J. Gamma-ray spectrometry and geological maps as tools for cadmium risk assessment in arable soils. *Geoderma* **2013**, *192*, 323–334. [CrossRef]
7. Eriksson, J.; Mattson, L.; Söderström, M. *Tillståndet i svensk åkermark och gröda, data från 2001–2007 (Current status of Swedish arable soils and cereal crops, data from the period 2001–2007)*; Swedish Environmental Protection Agency: Stockholm, Sweden, 2010.
8. Swedish Environmental Protection Agency. *Statens Naturvårdsverks Föreskrifter om Ändring i Kungörelsen (SNFS 1994:2) med Föreskrifter om Skydd för Miljön, Särskilt Marken, när Avloppslam Används i Jorbruket (The Swedish Environmental Protection Agencies Changes on When Sewage Sludge Can Be Used in Agriculture)*; Statens Naturvårdsverks Författningsssamling: Stockholm, Sweden, 1998.
9. Weindorf, D.C.; Bakr, N.; Zhu, Y. Chapter One—Advances in Portable X-ray Fluorescence (PXRF) for Environmental, Pedological, and Agronomic Applications. In *Advances in Agronomy*; Sparks, D.L., Ed.; Academic Press: San Diego, CA, USA, 2014.
10. Rouillon, M.; Taylor, M.P. Can field portable X-ray fluorescence (pXRF) produce high quality data for application in environmental contamination research? *Environ. Pollut.* **2016**, *214*, 255–264. [CrossRef] [PubMed]
11. Lemiére, B. A review of pxrf (field portable X-ray fluorescence) applications for applied geochemistry. *J. Geochem. Explor.* **2018**, *188*, 350–363. [CrossRef]
12. Weindorf, D.C.; Zhu, Y.; Chakraborty, S.; Bakr, N.; Huang, B. Use of portable x-ray fluorescence spectrometry for environmental quality assessment of peri-urban agriculture. *Environ. Monit. Assess.* **2012**, *184*, 217–227. [CrossRef] [PubMed]
13. Weindorf, D.C.; Chakraborty, S. Portable X-ray Fluorescence Spectrometry Analysis of Soils. In *Methods of Soil Analysis*; Soil Science Society of America: Madison, WI, USA, 2016.
14. Hu, W.; Huang, B.; Weindorf, D.C.; Chen, Y. Metals Analysis of Agricultural Soil via Portable X-ray Fluorescence spectrometry. *Environ. Contam. Toxicol.* **2014**, *92*, 420–426. [CrossRef] [PubMed]

15. US EPA: Method 6200—Field Portable X-Ray Fluorescence Spectrometry for the Determination of Elemental Concentrations in Soil and Sediment. Available online: Epa.gov/sites/production/files/2015-12/documents/6200.pdf (accessed on 29 December 2019).
16. Fredén, C. *Geology, National Atlas of Sweden*; SNA Publishing: Stockholm, Sweden, 1994.
17. Jones, A.; Montanarella, L.; Jones, R. *Soil Atlas of Europe*; European Commission: Luxemburg, Luxemburg, 2005.
18. Piikki, K.; Söderström, M. Digital soil mapping of arable land in Sweden—Validation of performance at multiple scales. *Geoderma* **2019**, *352*, 342–350. [CrossRef]
19. Eriksson, J.; Dahlin, S.A.; Sohlenius, G.; Söderström, M.; Öborn, I. Spatial patterns of essential trace element concentrations in Swedish soils and crops. *Geoderma Reg.* **2017**, *10*, 163–174. [CrossRef]
20. *SS 28311—Determination of Trace Elements in Soils—Extraction with Nitric Acids*; Swedish Standards Institute (SIS): Stockholm, Sweden, 2017.
21. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer-Verlag: New York, NY, USA, 2009.
22. Pedregosa, F.; Varaquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn.* **2011**, *12*, 2825–2830.
23. Rudy, J. Py-Earth Documenetation. Available online: https://contrib.scikit-learn.org/py-earth/index.html (accessed on 4 July 2019).
24. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290. [CrossRef]
25. López-Núñezm, R.; Ajmal-Poley, F.; González-Pérez, J.A.; Bello-López, M.A.; Burgos-Doménech, P. Quick Analysis of Organic Amendments via Portable X-ray Fluorescence Spectrometry. *Int. J. Environ. Res. Public Health* **2019**, *16*, 4317. [CrossRef] [PubMed]
26. Lahio, J.V.P.; Perämäki, P. Evaluation of portable X-ray fluorescence (PXRF) sample preparation methods. *Geol. Surv. Finl.* **2005**, *38*, 73–82.
27. Wan, M.; Hu, W.; Qu, M.; Tian, K.; Zhang, H.; Wang, Y.; Huang, B. Application of arc emission spectrometry and portable X-ray fluorescence spectrometry to rapid risk assessment of heavy metals in agricultural soils. *Ecol. Indic.* **2019**, *101*, 583–594. [CrossRef]

Acta Universitatis agriculturae Sueciae
Doctoral Thesis No. 2022:33

This thesis investigated digital soil mapping of cadmium (Cd) and copper (Cu), and portable X-ray fluorescence (PXRF) measurements for predicting Cd, Cu and zinc in Swedish agricultural topsoil. Results show that PXRF modelling can be an accurate and fast alternative to conventional laboratory analysis. The Cd map was used to delineate agricultural soil with low concentrations of Cd in winter wheat grain. The Cu map was used to locate agricultural soil highly likely not at risk of Cu deficiency.

Karl Adler received his graduate degree at the Department of Physical Geography and Ecosystem Science, Lund University, and undergraduate degree from the Department of Earth Sciences, Gothenburg University.

Acta Universitatis agriculturae Sueciae presents doctoral theses from the Swedish University of Agricultural Sciences (SLU).

SLU generates knowledge for the sustainable use of biological natural resources. Research, education, extension, as well as environmental monitoring and assessment are used to achieve this goal.

Online publication of thesis summary: https://pub.epsilon.slu.se

ISSN 1652-6880
ISBN (print version) 978-91-7760-941-4
ISBN (electronic version) 978-91-7760-942-1