Universidade Estadual de Londrina

JOSÉ VINÍCIUS RIBEIRO

Analysis of fertility attributes in Eutroferric Red Latosol and Red Nitosol by portable EDXRF

Londrina, 2024

JOSÉ VINÍCIUS RIBEIRO

Analysis of fertility attributes in Eutroferric Red Latosol and Red Nitosol by portable EDXRF

Dissertation presented to the Postgraduate Program in Physics at the State University of Londrina - UEL, as a partial requirement for obtaining the master's degree.

Advisor: Prof. Dr. Fabio Luiz Melquiades

Londrina, 2024

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

Ribeiro, José Vinícius .

Analysis of fertility attributes in Eutroferric Red Latosol and Red Nitosol by portable EDXRF / José Vinícius Ribeiro. - Londrina, 2024.
131 f. : il.

Orientador: Fábio Luiz Melquiades.

Dissertação (Mestrado em Física) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Programa de Pós-Graduação em Física, 2024.

Inclui bibliografia.

1. pXRF - Tese. 2. Machine Learning - Tese. 3. Fertility Attributes - Tese. I. Melquiades, Fábio Luiz. II. Universidade Estadual de Londrina. Centro de Ciências Exatas. Programa de Pós-Graduação em Física. III. Título.

CDU 53

JOSÉ VINÍCIUS RIBEIRO

Analysis of fertility attributes in Eutroferric Red Latosol and Red Nitosol by portable EDXRF

Dissertation presented to the Postgraduate Program in Physics at the State University of Londrina - UEL, as a partial requirement for obtaining the master's degree

EVALUATION COMITTEE

Advisor: Prof. Dr. Fabio Luiz Melquiades
Universidade Estadual de Londrina

Prof. Dr. Elton Eduardo Novais Alves
OCP Brasil

Prof. Dr. Avacir Casanova Andrello
Universidade Estadual de Londrina

Londrina, 20 de fevereiro de 2024

Dedico este trabalho a todos que, assim como eu, acreditam na ciência e na educação. Que nossa busca pelo conhecimento inspire as novas gerações, guiadas pela esperança de um futuro repleto de avanços e realizações

Agradecimentos

Ao meu orientador, Prof. Dr. Fábio Luiz Melquiades, pela oportunidade, incentivo, paciência, e, principalmente, pelos preciosos ensinamentos compartilhados ao longo destes últimos 4 anos e meio, desde a primeira iniciação cientifica. Obrigado, professor, por sempre estar disponível para me orientar e aconselhar.

Aos demais professores do Laboratório de Física Nuclear Aplicada da UEL, em especial ao Prof. Dr. Felipe Rodrigues, Prof. Dr. Avacir Andrello e Prof. Dr. Eduardo Inocente, pelos conhecimentos compartilhados através das disciplinas, seminários, conversas e pela pareceria nos trabalhos que elaboramos e publicamos juntos.

A todos os amigos do laboratório, que me auxiliaram diretamente nas preparações, medidas e análises, e também indiretamente, através das conversas, cafezinhos, comemorações e momentos de descontração.

Em especial ao João Marcos, Mateus, e a Leticia, que estiveram comigo ao longo das disciplinas, eventos e apresentações. Obrigado pela amizade construída e por toda ajuda, acadêmica ou não. Desde as conversas mais banais sobre a vida e nossos planos, até as discussões importantes sobre nossos estudos e ideias. Parte deste trabalho não seria possível sem a ajuda de vocês.

Aos demais colegas da pós-graduação em física, com os quais compartilhei momentos de estudo e resoluções de listas.

A todos os professores que contribuíram para a minha formação, desde o ensino básico até a pós-graduação.

A CAPES, pela concessão da bolsa durante esses dois anos, a qual foi essencial para a realização deste trabalho.

Ao Laboratório de Física Nuclear Aplicada da UEL, pelo equipamento para as medidas por pXRF e por toda a estrutura que me acolheu desde a iniciação científica em 2019.

Ao Laboratório de Solos do IDR-PR (Antigo IAPAR) onde as análises convencionais foram realizadas

A minha amada namorada Maria Fernanda, pelo incondicional apoio, incentivo, dedicação, companheirismo, força, paciência, compreensão, torcida, amor e carinho, que me mantiveram firme nos melhores e nos piores dias, sem exceção. Grande parte deste trabalho também não seria possível sem a sua ajuda.

Por fim, agradeço a minha família, em especial a minha mãe, Leandra, e minha tia, Leonilda, que sempre foram as unidades mais fundamentais da minha rede de apoio, cuidado e defesa. Uma última grande parte deste trabalho também não seria possível sem ajuda de vocês.

RIBEIRO, José Vinícius. **Analysis of fertility attributes in Eutroferric Red Lato and Nitosol by portable EDXRF**. 2024. 131 pp. Dissertation (Master's Degree in Physics) – Department of Physics, Universidade Estadual de Londrina, Londrina, 2024.

Abstract

The increasing global demand for agricultural products, food, and energy requires a sustainable and secure approach to agricultural production. Soil fertility is essential to increasing agricultural productivity, and understanding the spatial variability of soil fertility attributes is crucial to preserving high soil fertility. Traditional soil analysis methods are time-consuming, expensive, and generate waste. Spectral analysis techniques, such as energy-dispersive X-ray fluorescence (EDXRF), combined with machine learning (ML) methods, offer a rapid and cost-effective alternative for quantifying soil properties. In this regard, the main motivation of this study is to explore data generated by a portable EDXRF (pXRF) spectrometer applied to soil samples in the context of modeling with machine learning algorithms to quantify fertility attributes. From this perspective, this study established optimized conditions in terms of experimental setup (Chapter 2) and minimum number of samples used in the training step (Chapter 3) for the building local ML models to quantify soil fertility attributes. Furthermore, a preliminary *in situ* validation study of the models with optimal conditions was conducted. The results indicated that the use of 15 kV, 12.15 µA and 30 s without filters was the most suitable for the simultaneous quantification of the evaluated fertility attributes. The optimized pXRF condition produced results equivalent to those modeled with benchtop EDXRF data, suggesting a possible *in situ* application of this method. Nevertheless, the preliminary study of extrapolation from laboratory-calibrated models had negative results, indicating that their extrapolation to *in situ* measurements was not recommended in this research context. On the other hand, it was possible to reduce the number of training samples while maintaining performance equivalent to models trained on the full data set. These results highlight the potential of pXRF as a low-cost, rapid, and efficient alternative approach for local soil fertility mapping, providing valuable information for decision-making related to the use of fertilizers and other soil amendment products.

Keywords: pXRF, Machine learning, Data fusion, Soil fertility attributes

RIBEIRO, José Vinícius. **Análise de atributos de fertilidade em Lato e Nitossolo Vermelho Eutroférrico por EDXRF portátil**. 2024. 131 pp. Dissertação (Mestrado em Física) – Departamento de Física, Universidade Estadual de Londrina, Londrina, 2024.

Resumo

A crescente demanda global por produtos agrícolas, alimentos e energia requer uma abordagem sustentável e segura para a produção agrícola. A fertilidade do solo é vital para aumentar a produtividade agrícola, e entender a variabilidade espacial dos atributos de fertilidade do solo é crucial para preservar a fertilidade do solo. Os métodos tradicionais de análise de solo são demorados, caros e geram resíduos. As técnicas de análise espectral, como a fluorescência de raios X por dispersão em energia (EDXRF) combinadas com metodologias de *machine learning* (ML) oferecem uma alternativa rápida e econômica para a quantificação dos atributos do solo. A respeito disso, a principal motivação deste estudo é explorar os dados gerados por um espectrômetro portátil de EDXRF (pXRF) aplicado a amostras de solo no contexto de modelagem com algoritmos de aprendizado de máquina para quantificar atributos de fertilidade. Nessa perspectiva, este estudo estabeleceu condições otimizadas em termos de configuração experimental (Capítulo 2) e número mínimo de amostras utilizadas na etapa de treinamento (Capítulo 3) na construção de modelos locais de ML para quantificar atributos de fertilidade do solo. Adicionalmente, foi realizado um estudo preliminar de validação *in situ* dos modelos com condições ótimas. Os resultados indicaram que o uso de 15 kV, 12.15 μA e 30 s sem filtros é o mais indicado para quantificação simultânea dos atributos de fertilidade avaliados. A condição otimizada de pXRF forneceu resultados equivalentes aos modelados com dados de EDXRF de bancada, sugerindo uma possível aplicação *in situ* deste método. No entanto, o estudo preliminar sobre extrapolação a partir de modelos calibrados em laboratório teve resultados negativos, indicando que a sua extrapolação para medições *in situ* não foi recomendada no contexto desta pesquisa. Por outro lado, foi possível reduzir o número de amostras de treinamento mantendo desempenho equivalente aos modelos treinados com o conjunto completo de dados. Esses resultados destacam o potencial do pXRF como uma abordagem alternativa econômica, rápida e eficiente para o mapeamento local da fertilidade do solo, fornecendo informações valiosas para a tomada de decisões relacionadas ao uso de fertilizantes e outros produtos para correção do solo.

Palavras-chave: pXRF, Aprendizagem de máquina, Fusão de dados, Atributos de fertilidade

Figures list

Tables list

ABBREVIATIONS AND ACRONYMS

BSP	Base Saturation in Percentage
Ca^{2+}	Exchangeable Calcium
CEC	Cation Exchange Capacity
CRM	Certified Reference Material
CV	Coefficient of Variation
EDXRF	Energy Dispersive X-Ray Fluorescence
H+Al	Potential Acidity (H+Al).
iSPA-PLS	Successive Interval Projection Algorithm for Minimals Partial Squares
K^{+}	Exchangeable Potassium
LIBS	Laser-Introduced Breakdown Spectroscopy
LV	Latent Variables
Mg^{2+}	Exchangeable Magnesium
MLR	Multiple Linear Regression
OM	Organic Matter
PCA	Principal Component Analysis
PC's	Principal Components
PLS	Partial Least Squares Regression
P$_M$	Phosphorus Extracted by Mehlich-1,
pXRF	Portable X-ray Fluorescence
RE	Relative Error
RMSEC	Calibration Root Mean Square Error
RMSECV	Cross Validation Root Mean Square Error
RMSEP	Prediction Root Mean Square Error
RPD	Deviation Performance Ratio
SB	Sum of Exchangeable Bases
SOC	Soil Organic Carbon

Vis-NIR	Near and Visible Infrared Spectroscopy
XRF	X-Ray Fluorescence

Summary

CHAPTER 1. INTRODUCTION AND THEORY

1.1 General introduction

Soil is a fundamental natural resource for sustaining life and economic development. It provides several ecosystem services and is the fundamental resource for many human activities [1]. Therefore, knowledge of soil physical and chemical properties and their spatial variability is essential for its sustainable use, planning, and proper management for increased productivity and conservation [2].

The global population is projected to reach 9.8 billion by 2050 (DESA-EN, 2017). How to feed so many people? Several authors have already discussed the importance of developing methodologies to provide and maintain high soil fertility [4], [5]. Without this, problems will arise in several areas, mainly related to the increase in hunger [4]. In this regard, the preservation of high soil fertility plays a crucial role in increasing agricultural production. This requires an adequate spatial characterization of the different soil fertility attributes. They are the parameters that quantify the soil's ability to provide nutrients to plants and play a fundamental role in determining its quality and ability to sustain agricultural production. These attributes help, for example, to make rational decisions about corrections and soil management, nutrients agricultural inputs.

Soil fertility spatial characterization requires a high density of sampling and analysis, which is problematic considering the traditional system of collection and laboratory analysis. Samples are collected manually and analyzed using techniques that, although accurate, require hard preparation, are time-consuming, use reagents, generate waste, and are expensive. Thus, the development of faster and cheaper methods based on green chemistry has become an important area of research [6].

The use of Proximal Soil Sensors (PSS) combined with Machine Learning (ML) tools has proven capable of quantifying soil fertility attributes and may offer more practical, faster and cheaper alternatives compared to conventional methods [7], [8]. The aim is to use this methodology as a complement to increase sampling density and reduce the number of samples conventionally analyzed. These sensors are based on spectroscopic techniques, each one based on a different band of the electromagnetic spectrum and consequently exploring different phenomena of interaction between radiation and matter. The most commonly used spectroscopic techniques in agriculture are near and visible infrared spectroscopy (Vis-NIR)

[9]–[11], laser-induced breakdown spectroscopy (LIBS) [12], [13], and energy dispersive X-ray fluorescence (EDXRF) [6], [14], [15]. These techniques may be used with benchtop equipment, in the laboratory, or adapted for portable or embedded systems, with the possibility of on-line response. There is also the possibility of combining two or more sensors with data fusion methods, aiming to increase modeling accuracy by exploring the synergy between the distinct nature of the responses generated by each technique [16].

Energy Dispersive X-ray Fluorescence (EDXRF) is a well-established analytical technique used to determine the total content of various chemical elements in different types of samples, including soil samples. On the other hand, some studies have shown that it is a suitable sensor for soil fertility assessment in the precision agriculture context [6], [14], [15]. The technique is based on the photoelectric effect, *i.e.*, after the excitation of electrons from the innermost layers of the atoms present in a sample, the de-excitation is manifested with the emission of an electromagnetic wave with a specific frequency for each element, called characteristic X-ray [17], [18]. So, EDXRF spectra enable the total content measurements of the main elements in the soil (*e.g.* Al, Si, P, K, Ca, Mn, Fe, Cu, Zn, etc.) and may be used as input variables for the indirect determination of other fertility attributes by ML algorithms.

Machine Learning is a branch of artificial intelligence (AI) that explores the study and construction of computational algorithms based on learning from data instead of pre-programmed instructions. The primary goal of an ML model is to construct a complex system that learns from a predefined database and ultimately produces prediction, classification, or detection results for one or more properties of interest [19].

Therefore, the main motivation of this study is to explore the data generated by portable EDXRF (pXRF) instruments applied to soil samples in the context of modeling with machine learning algorithms to quantify fertility attributes. Although studies with this objective involving the use of pXRF already exist, few of them aim to optimize the modeling process. From this perspective, this study seeks to establish optimized conditions in terms of experimental setup (Chapter 2) and a number of samples used in the training step (Chapter 3) in the building of local ML models to quantify soil fertility attributes. The ML algorithms used was partial least square regression and predictive analysis of common dimensions. Additionally, a preliminary study on the *in situ* extrapolation of the laboratory-calibrated models with optimal conditions was conducted.

1.2 Objectives

1.2.1 General objectives

Evaluate optimized conditions for developing ML local models using pXRF data to determine soil organic carbon (SOC), phosphorus extracted by Mehlich-1 (P_M), exchangeable calcium (Ca^{2+}), exchangeable magnesium (Mg^{2+}), exchangeable potassium (K^+), sum of exchangeable bases (SB), total cation exchange capacity (CEC), base saturation in percentage (BSP), pH and potential acidity (H+Al).

1.2.2 Specific objectives

In chapter 2, the specific goals are:

- Evaluate the performance of a commercial pXRF routine for main soil elements quantification;
- Build local models using partial least square (PLS) regression and predictive analysis of common dimensions (P-ComDim) to determine the soil fertility attributes;
- Evaluate the individual and synergistic use (by data fusion) of different experimental conditions in pXRF to determine an optimized condition for modeling fertility attributes;
- Compare the predictive performance of models built by pXRF data vs. benchtop EDXRF data;
- Extrapolate laboratory-calibrated models to *in situ* predictions;

In chapter 3, the specific goals are:

- Evaluate the predictive performance of local models as a function of the number of samples used in the modeling training step;
- Evaluate two different methods for reducing each training set to assess the reproducibility of the results;
- Establish a minimum number of sample reductions for model training while maintaining performance equivalent to models with a full set;

1.3 Theory

1.3.1 Soil fertility

One of the main aspects of sustainable agriculture is knowledge of soil fertility, its characteristics and limitations, so that interventions may be made in the most rational way possible. The main tool for soil fertility proper assessment is the chemical analysis of the several variables that influence the growth and development of plants. The values obtained for each variable are valuable indicators of the potential success of future plantings and enable the elaboration of soil fertilization plans for adequate crop nutrition [20], [21].

The soil chemical analysis is based on the principle of determining the levels of nutrients and characteristics that may influence their availability to plants. It is traditionally composed of two steps: extraction and quantification. In the first step, extractor solutions are used. They simulate the plants root action, *i.e.*, extracting the nutrients chemical forms that would be absorbed by the plants. In this stage, a certain extractor volume is agitated with a soil defined volume, changing the nutrients from the solid to the liquid phase (with the equilibrium solution). In the quantification step, the elements contents in the equilibrium solution are quantified. Conventionally, this is done through a different methodology set, such as atomic Absorption Spectrophotometer; UV-Vis Spectrophotometry; Titrimetric; and Potentiometry [20], [21]. Then, the nutrient contents determined by chemical analysis are compared with reference values and the fertility level classification is performed.

There are sixteen chemical elements essential for plant growth. According to the traditional nomenclature, they are divided into non-minerals and minerals. The non-mineral nutrients, found in the atmosphere and water, are carbon (C), hydrogen (H), and oxygen (O). The mineral nutrients, which are provided by the soil, are also traditionally divided into three subgroups, primary, secondary and micronutrients. The primaries contain the nitrogen (N), phosphorus (P), and potassium (K); the secondary is calcium (Ca), Magnesium (Mg), and Sulfur (S). The primary and secondary nutrients set is also called primary and secondary macronutrients, respectively, or just macronutrients. The micronutrients are Boron (B), Chlorine (Cl), Copper (Cu), Iron (Fe), Manganese (Mn), Molybdenum (Mo), and Zinc (Zn) [22]. In the soil, these nutrients are found in their exchangeable forms, also called available or bioavailable.

One or several nutrients may be present in the soil at very low levels, or in a form that the roots cannot absorb, reducing the plant's productive potential. To make available the elements present, the soil must be well managed. However, when nutrients are lacking, they need to be reintroduced. Because they are required in greater quantities by plants in most cases, primary nutrients become deficient in soils more frequently. Secondary nutrients and micronutrients are generally less deficient and used in lower amounts [22], [23].

The nutrient replacements are mainly made using organic matter or mineral chemical fertilizers. Organic matter contains practically all the macro and micronutrients and, in addition, provides better structure to the soil, increasing its fertility. Mineral fertilizers, unlike organic matter, have a high concentration of highly soluble nutrients, which are absorbed quickly by plants or easily leached. The commercial mineral fertilizers are simple (contain one or more macroelements) or compound (mixture of simple fertilizers). Compound fertilizers are known by their formulas, *e.g.*, 4-14-8, 10-10-10, 20-5-20, where the numbers indicate the nitrogen, phosphorus, and potassium percentage (N-P-K) in the fertilizer chemical [23].

From the nutrient management point of view, the appropriate is the existence of balance. The exacerbated addition of one of them without considering the others and the culture characteristics may lead to poor harvests [23].

In the soil fertility study, arises the fertility attribute concept. Soil fertility attributes are the parameters that determine the soil capacity to provide nutrients for plants. They play a key role in determining soil quality and its ability to sustain agricultural production.

The soil cation exchange capacity measured at pH 7 is called Cation Exchange Capacity (CEC). It is one of the most important variables for the soil productive potential interpretation because it indicates the total amount of negative charges that the soil could present if its pH were 7. In turn, these charges may adsorb (retain) the important nutrients of positive charge (K^+, Ca^{2+}, and Mg^{2+}), added to the soil via fertilization or other corrections. The CEC depends on the organic matter content and the amount and type of clay. Soil with a high content of clay and organic matter may retain more exchangeable cations than soil with a low content of these quantities [20], [22]. On the other hand, when the cationic nutrients added via fertilization is greater than the soil CEC, these nutrients (K^+, Ca^{2+}, Mg^{2+}) may be lost by leaching [22]. The CEC may be quantified indirectly through some ions listed in the equation (1),

(1)

The CEC, being a measure of the total amount cations that the soil may retain, is related to the colloids quantity and quality. Colloids are small particles formed during the weathering process in soil formation where some minerals and organic matter are reduced to extremely small particles. Chemical changes further shrink these particles, to the point where they cannot be seen with the naked eye. Most colloids have a net negative charge. This means that it may attract and retain cations. The cations retained in the colloids may be replaced by other cations, they are exchangeable. For example, the calcium may be exchanged for magnesium or potassium. Therefore, the CEC represents the graduation of the ability to release various nutrients, favoring the fertility maintenance for a prolonged period and reducing or preventing the occurrence of toxic effects from the fertilizer's applications [23].

The term "exchangeable bases" or "total exchangeable bases" refers to the sum of the bases (calcium, magnesium, potassium, etc.) in exchangeable form. This parameter varies considerably depending on the characteristics and conditions of the soil. For instance, soils in humid areas contain less exchangeable bases and more exchangeable hydrogen than soils in desert environments [24]. Specifically, the soil sum of exchangeable bases (SB) represents the soil sum of the exchangeable cation contents, except for H^{+}+Al^{3+} (SB = Ca^{2+}+Mg^{2+}+K^{+}). Then, it indicates the colloids negative charges number that are occupied by bases. Thus, the SB, like CEC, is a good soil quality indicator, as it is also related to the soil's ability to retain nutrients and maintain the chemical balance necessary for plant development [20], [23].

The SB is used in the Percent Base Saturation in Percentage (BSP) calculation, which is the percentage of the cation exchange capacity occupied by base cations (K^{+} + Ca^{2+} + Mg^{2+}) in percentage terms [21]. It is calculated using equation (2):

(2)

The BSP may provide indications of the soil liming need. A low BSP value means that there are small amounts of cations, such as Ca^{2+}, Mg^{2+}, and K^{+}, saturating the colloids negative charges, and that most of these are being neutralized by H^{+} and Al^{3+}. This raises the soil acidity level, which may contain aluminum at a toxic level to plants. Most crops show good productivity when soil BSP is between 50% and 80%, together with a pH value between 6.0 and 6.5 [23]. Then, as a good indicator of general soil fertility conditions, Percentage Base Saturation is also used as a complement in the soil's nomenclature, which are divided according into eutrophic soils (high fertility and with BSP ≥ 50%) and dystrophic (low fertility and BSP < 50%) [23]

The soil organic matter (OM) consists of plant and animal residues in the decomposition stages. It is formed by the proportions, on average, of 58% Carbon, 33% Oxygen, 6% Hydrogen, and 3% Nitrogen, Sulfur, and Phosphorus. In the soil, OM undergoes humification, *i.e.*, it is decomposed by soil microorganisms, resulting in the release of essential nutrients for plants. This process involves the breakdown of complex organic matter into simpler compounds such as humic and fulvic acids, carbon dioxide, water, and minerals [20], [22], [23]. The Soil Organic Carbon (SOC) is the total amount of carbon present in OM. Being proportional to the soil organic matter level, the SOC content is a productive potential parameter, because soils with higher SOC values have higher CEC values and greater capacity to supply nutrients to plants, compared to soils with lower SOC contents [20], [22].

The pH measures the H^+ activity present in the soil. When absorbing positively charged nutrients ($K^{+,}$ Mg^{2+}, Ca^{2+}, etc.), plants release H^+ from the roots, which reduces the soil pH and makes it acidic. In the nitrogenous fertilizers reaction with the soil, specifically in nitrification (passage from ammonium to nitrate), there is also H^+ release. In addition to these, other factors such as source material, OM decomposition, rainfall and irrigation may also affect soil acidity. Therefore, the pH value depending on soil management, successive crops, fertilization, and several other factors [20], [22].

In this perspective, pH is indicator of actual soil fertility. The pH of productive soils may vary between 4.0 and 9.0, and soils with a pH greater than 6.5 have a marked reduction in Zn, Cu, Fe and Mn availability. For these reasons, the ideal pH to plant growth is between 6.0 and 6.5. In this pH range there is no presence of toxic form Al^{3+} and there is good availability of nutrients [20], [22].

Also related to acidity levels in soils, potential acidity (symbolized by H+Al in this study) is the sum of exchangeable and non-exchangeable soil acidity. The exchangeable acidity refers to the Al^{3+} and H^+ ions retained on the surface of soil colloids. Non-exchangeable acidity is the covalently bonded H^+ ion associated with negatively charged colloids and aluminum compounds. The potential acidity characterizes the buffering power of soil acidity, and its accurate estimate is mainly used to CEC estimate [25].

1.3.2 EDXRF

Energy Dispersive X-Ray Fluorescence (EDXRF) is an atomic emission technique based on the interaction of X-rays with matter. It exploits the products generated in these interactions to extract multi-elementary information about the chemical composition of the analyzed samples.

The main phenomenon involved in EDXRF is the photoelectric effect. When an X-ray beam interacts with an atom, electrons from the innermost shells of that atom may be ejected, creating vacancies that lead the atom to an unstable state. This interaction is called photoelectric absorption, which occurs exclusively when the incident photon has energy greater than the binding energy of the electron to the atom. After absorption, for atomic stabilization, electrons from the outer orbital shells perform electronic transitions to fill the generated vacancies, releasing the energy difference between the involved orbitals in the X-ray photons form (called characteristic X-ray) with an energy that is unique to the element's atom transitions [17], [18]. Figure 1 presents a schematic illustration of this interaction.

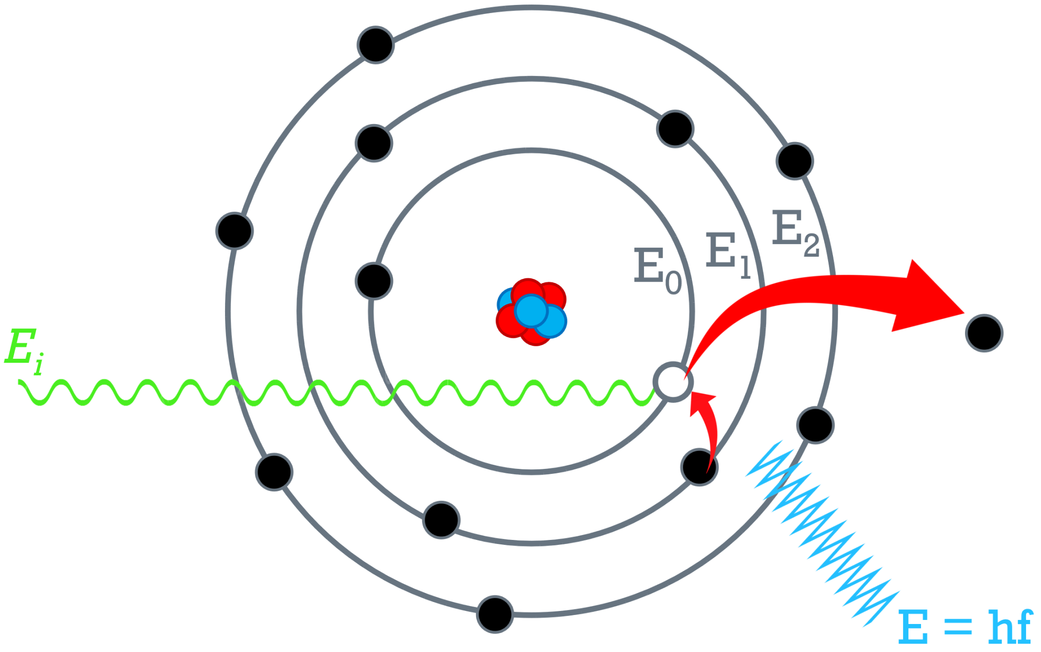

Figure 1 - Schematic representation of the photoelectric effect

The photoelectric effect is most likely to occur for low-energy photons (with a maximum value of 100 keV) and for elements with high atomic number. Its cross section for non-relativistic energies () is estimated by

(3)

is the incident photon energy, is the electron rest energy, and is the fine structure constant [26].

The characteristic X-rays coming from each element are identified according to the orbital shells involved in the stabilization process. In the Siegbahn's nomenclature [17], the vacancy shell determines the prefix name (K, L, M, etc.) while the second part has the suffixes α, β, γ, etc. according to the most intense X-ray emission lines (with higher probabilities) after the electronic transition. For instance, the characteristic X-ray originating from a transition involving the electron from the L to the K shell is called Kα (the most intense), while the one involving the electron from the M to the K shell is called Kβ, etc. On the other hand, the electronic shells are constituted by some subshells with very close energy values. Each sublevel is characterized by a quantum numbers sequence (which give the state quantum information) and the electronic transitions between atomic energy levels are governed by the quantum mechanics laws.

This technique is commonly used in the identification and quantification of the total content of inorganic species from Mg to U. However, depending on the energy range, higher or lower elemental sensitivities are obtained. This occurs for several reasons, including fluorescence yield, characteristic X-ray absorption by the atmosphere, detector sensitivity, and tube power [17]. In addition, the photoelectric effect is not the only interaction suffered by the atom when exposed to ionizing radiation. Radiation scattering by electrons in the analyzed sample is also likely to occur and, depending on the excitation energy, may significantly affect the EDXRF efficiency for some elements. Scattering effects are divided into two classes, elastic (Rayleigh) and inelastic (Compton) scattering [17], [18].

In Rayleigh Scattering, incident photons are elastically scattered throughout the entire atom. In this process, all the electrons in the atom participate coherently, and for this reason, it is also called coherent scattering. During elastic collision, there is no transfer of energy among photons and electrons. Thus, the photons are scattered in a different direction from the original, with the same energy, changing only their linear momentum, *i.e.*, the scattered photons have the same phase as the incident ones. Figure 2 presents a schematic illustration of this interaction.

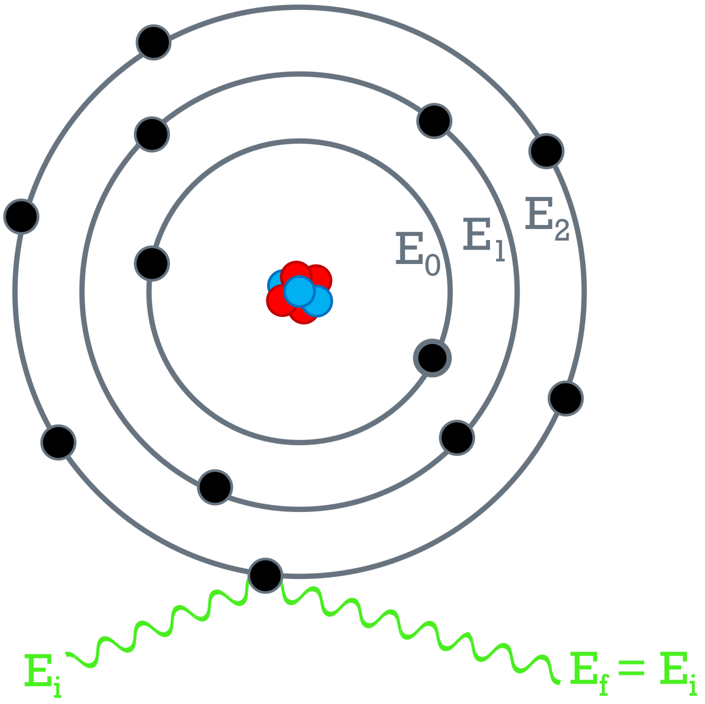

Figure 2 - Schematic representation of the Rayleigh scattering

This coherent scattering occurs mainly at low-energy photons of materials with high Z. Its differential cross section for non-relativistic energies is expressed by equation 4

(4)

corresponds to the electron classical radius, is the atomic scattering factor that depends on the electron cloud charge distribution and is the solid angle for the diffusion angle [17], [27].

Compton Scattering is due to the interaction of a photon with an electron weakly bound to the atom [17]. In this inelastic collision, part of the incident photon energy is transferred to the electron, ejecting it. The photon is then scattered in a different direction from the original and with lower energy, being equal to the incident energy minus the energy transferred to the electron [28]. Considering that the momentum and energy of the proton plus electron system must be conserved, the scattered photon energy may be determined by equation 5,

(5)

and is the angle at which the incident photon was scattered [28].

The Compton Effect is predominant when the binding energy of orbital electrons becomes negligible compared to the incident photon energy. Under these conditions, the electrons are considered free in the interaction process. Thus, the occurrence probability of this scattering is more significant for high-energy photons (from 100 keV to 1.02 MeV) and low Z elements [29]. The differential cross section for the Compton interaction of a photon with a free electron with respect to the solid angle may be described by the Klein-Nishina formula (equation 6),

(6
)

For the interaction with bonded electrons the effects of atomic bonding must be considered and the Klein-Nishina formula are corrected with the multiplication by the incoherent scattering function *S(x,Z)* [29]. Figure 3 presents a schematic illustration of this interaction.

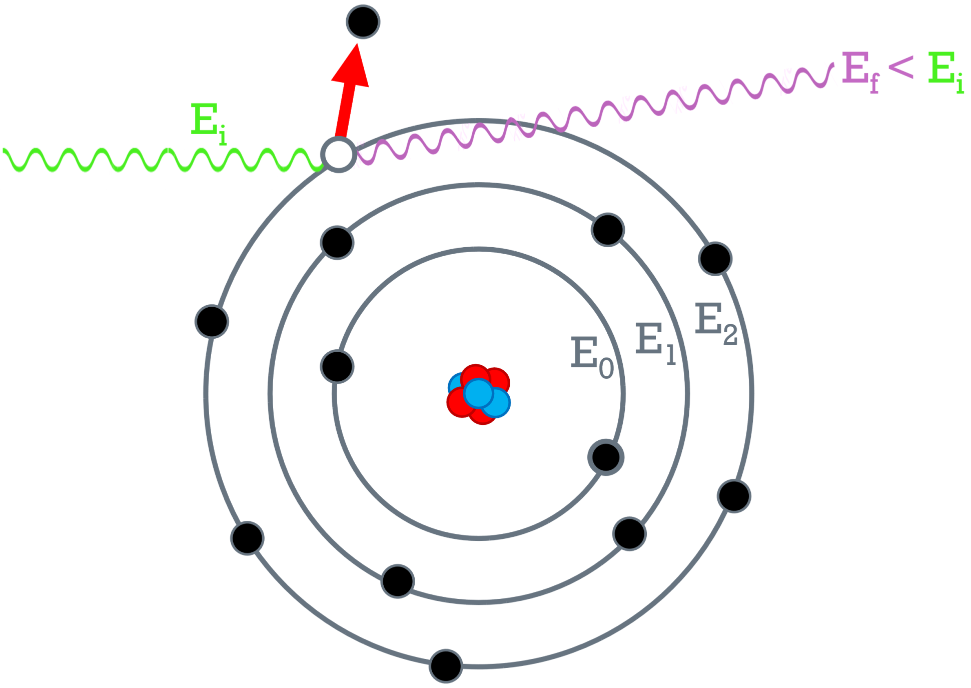

Figure 3 - Schematic representation of the Compton scattering

The interactions probabilities of an atom exposed to a photons beam may be expressed by the cross-section of each involved phenomenon. Therefore, the probability of interaction of a photon with an atom per square centimeter of matter is given by the sum of the cross sections per atom of the predominant phenomena, according to equation 8 [17]

(8)

Figure 4 presents an estimate of the relative contribution of photoelectric absorption and compton and rayleigh scatterings for 3, 15, 30 and 50 keV (this energy range is usually used in EDXRF). These data were extracted from the NIST XCOM software: https://physics.nist.gov/PhysRefData/Xcom/html/xcom1.html. Apart from hydrogen, the probability of scattering effects occurring in interactions between photons and atoms is small for 3 keV, the dominant effect being photoelectric absorption. However, there is also the possibility of radiation falling on the sample without any interaction. On the other hand, for 15 keV, the Compton scattering relative contribution was mainly due to H, C, N, and O (ranging from 1% to 19%). In this energy range, the contribution to Rayleigh Scattering ranged from 12% to 1%, having the highest values for C, N, O and Zr. As the incident photons energy increases to 30 and 50 keV, the scattering contribution also increases and reaches significant values. This impairs the detection of the elements that compose the soil's macro and micronutrients. For example, at 50 keV, Al and Si present a smaller % contribution to photoelectric absorption (47 and 53%, respectively) compared to the summed scattering (53 and 47%, respectively).

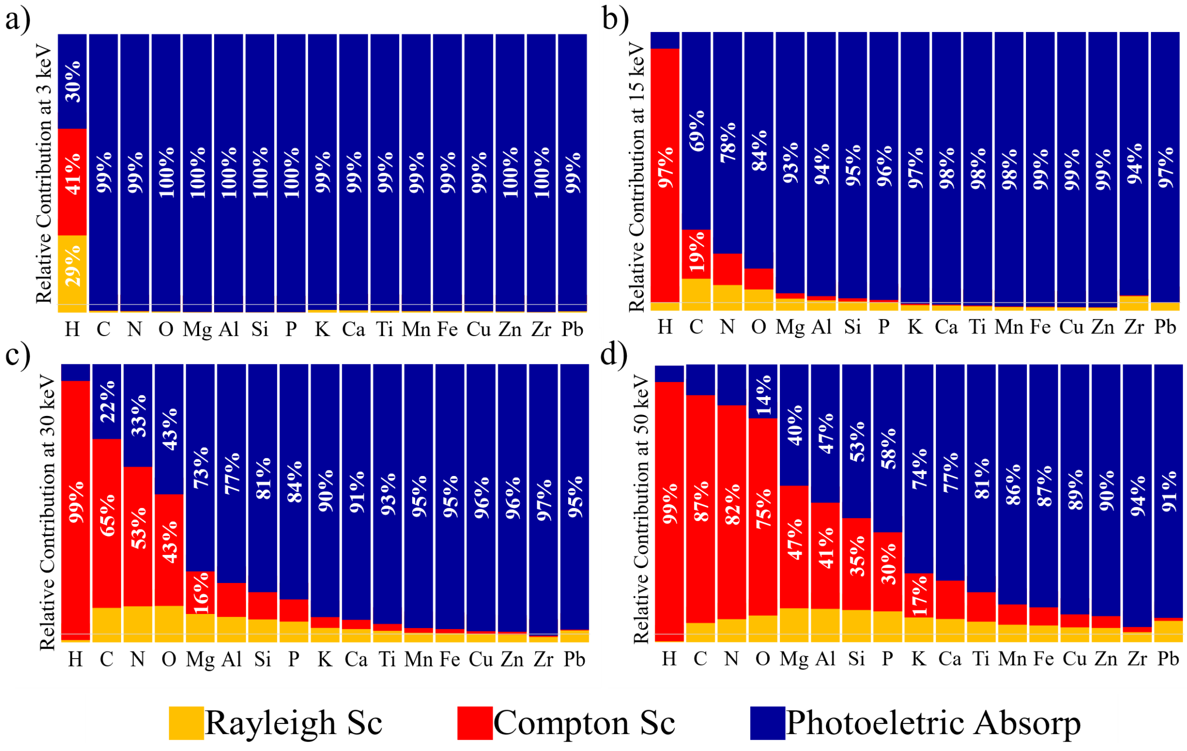

Figure 4 - Estimation of the relative contribution of Photoelectric Absorption, Compton and Rayleigh Scattering to the energy of 3 keV (a), 15 keV (b), 30 keV (c) and 50 keV (d).

Therefore, the scattering effects contribution is different according to the incident photons energy. In general, for the same atom, low energies result in a greater probability of photons being scattered coherently than incoherently (Compton). As the energy increases, the contribution from Rayleigh Scattering is overcome by Compton Scattering in such a way that it may be more significant than the photoelectric absorption, depending on the evaluated atom. Thus, the EDXRF precision to quantify this element may be impaired.

In this perspective, an important parameter to be considered in EDXRF is the fluorescence yield, defined as the number of characteristic X-rays effectively emitted in relation to the number of vacancies produced in each atomic shell [30]. Figure 5 shows the fluorescence yield (ω) of the K, L and M shells as a function of the Z atomic number. It is noted that the K shell has a low fluorescence yield for Z<20 (approximately 0.15) while for the L shell, Z<60 (approximately 0.18).

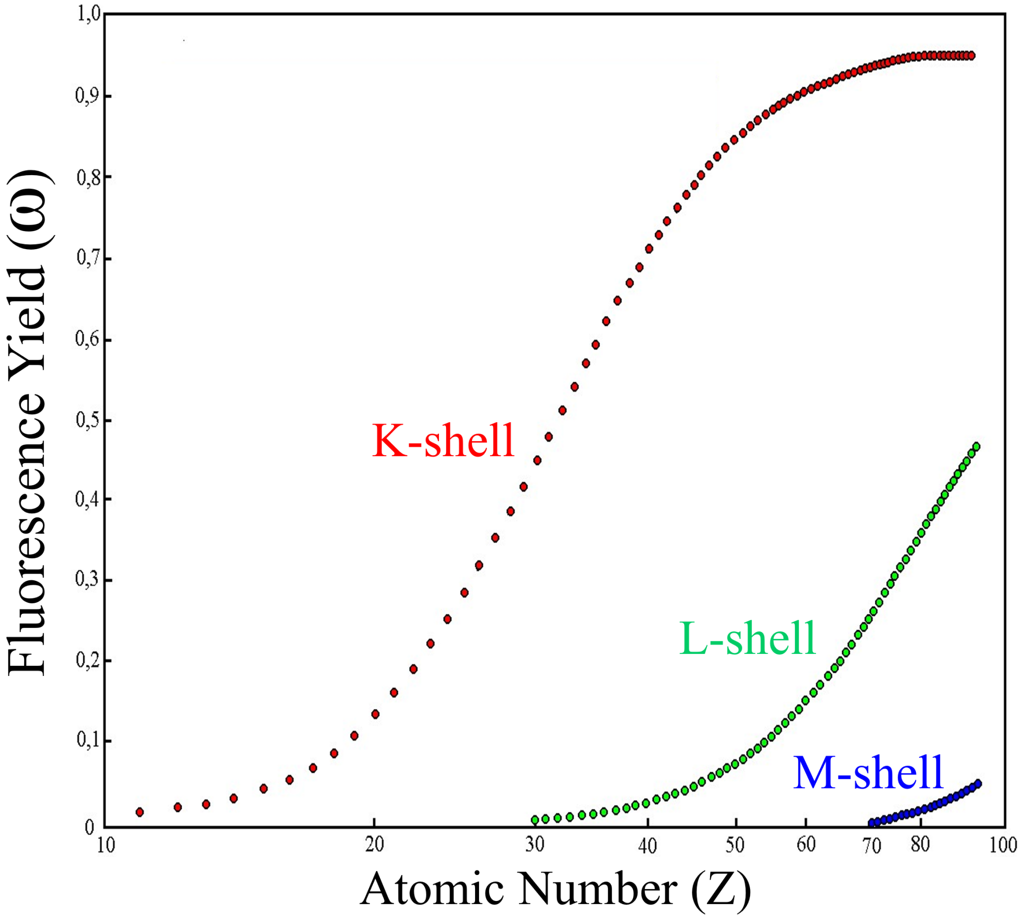

Figure 5 - Fluorescence yield of the K, L and M shells versus Z atomic number. Adapted from [31]

Furthermore, in the characteristic X-ray emission process, the excess energy may be transferred to an electron in an atom outer shell, ejecting it. This phenomenon is known as the Auger effect and is more relevant for elements with low atomic numbers [17].

1.3.2.1 XRF fundamental equation for quantitative analysis

Based on the measurement of the characteristic X-rays intensities (number of X-rays detected per unit of time) emitted by the elements that constitute the sample, EDXRF is considered as a quali-quantitative method. Synthetically, the analysis consists of three phases: excitation, dispersion, and detection of characteristic X-rays [31].

The sample excitation may be done in several ways: by accelerated particles such as electrons, protons, or ions; by alpha particles, negative beta particles or gamma rays emitted by radionuclides, and the most used process today, excitation by an X-rays beam generated in X-rays tubes [17], [30], [31]. In detection, the X-rays are processed through electronic pulses produced in an appropriate detector. These pulses are directly proportional to the X-ray energies, hence the dispersive energy nomenclature [17], [18], [31].

Equation 9, involving the intensity of a characteristic X-ray line (*e.g.*, Kα or Lα) and the concentration of an element in the sample is called the XRF fundamental equation [31], [32]. This equation is the basis for the quantitative XRF method.

(9)

represents the net intensity of the ith characteristic X-ray (cps), the ith concentration (g cm^2), the ith elemental sensitivity (cps g^{-1} cm^2) and the absorption factor for the element of interest or analyte (dimensionless). The index is used to refer to the X-ray characteristic of ith energy. Thus, the net areas of characteristic X-ray lines in the XRF spectrum are directly proportional to elemental concentrations.

The absorption factor is given by the equation:

(10
)

where is the density (g cm^{-3}), is the matrix thickness (cm), and is the matrix total absorption coefficient (cm^2 g^{-1}), given by the equation:

(11
)

where and are the matrix absorption coefficients (cm^2 g^{-1}) for the energies of the incident radiation and the characteristic X-ray, respectively, and and are the angles of the incident and emerging radiation, respectively, in relation to the surface Sample [17], [31], [32].

The elemental sensitivity for the K-lines X-ray characteristics is given by the equation:

(12
)

where represents the absorption coefficient of the element for the photoelectric effect on the incident radiation energy (cm^2 g^{-1}), the fluorescence yield for K-line X-rays (fraction), the fraction of photons emitted as Kα X-rays, the jump ratio in the K absorption cutoff, the geometric factor and the detector efficiency for the characteristic X-ray emitted by the analyte [17], [31], [32].

While the absorption factor depends on the density, thickness and total absorption coefficient of the sample, the elemental sensitivity depends on several other parameters. Because of these complex dependencies, elemental sensitivity is commonly obtained indirectly by constructing sensitivity versus Z atomic number curves using certified reference materials samples of pure or simple compounds.

Finally, after adequate XRF system calibration based on the equation 9 parameters for the element of interest, the quantitative analysis is performed by reading the net intensities in the spectrum.

1.3.2.2 XRF instrumentation and portability

The EDXRF system consists of a source for samples excitation, a detector that identifies and separates the characteristic X-rays, a multichannel board that records the obtained spectrum and the electronics necessary for powering the system, amplification of the signals coming from the detector, processing, and analysis of the generated data. Figure 6 shows an illustration of an EDXRF system with X-ray tube excitation.

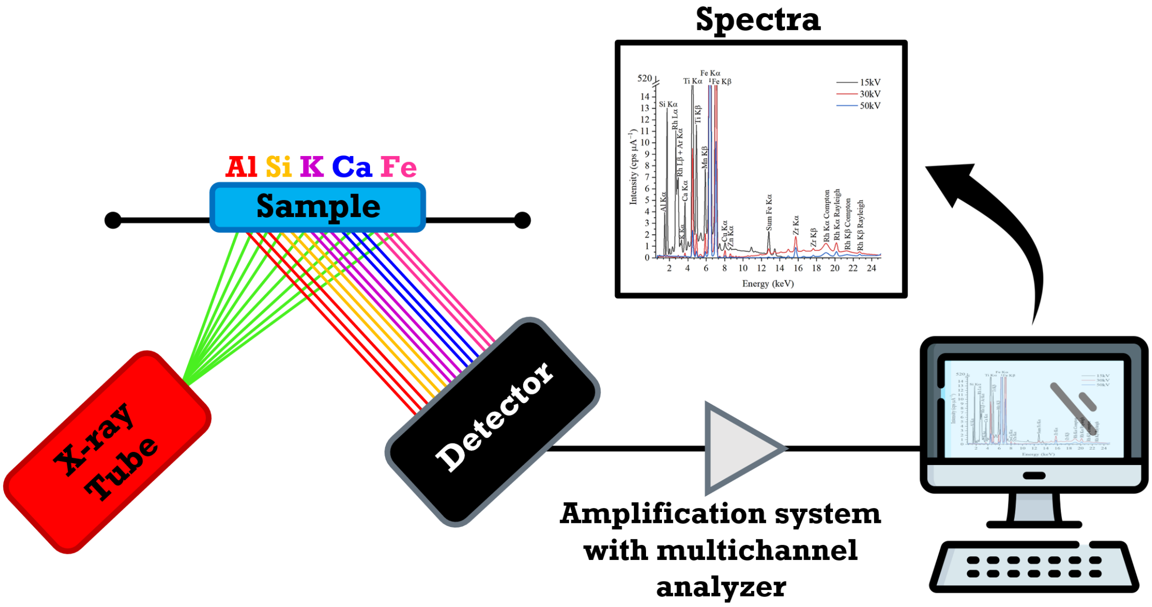

Figure 6 - Schematic illustration with the basic instrumentation of an EDXRF system with X-ray tube excitation

The X-ray tube main components are the cathode and anode. The cathode consists of a filament (usually tungsten) that is heated for thermionic emission of electrons in a vacuum chamber. The current value is adjusted according to the desired electrons flow. The increase in the current increases the filament temperature, also increasing the electrons emitted number from the cathode to the anode. For this reason, the current in the X-ray tube is fundamental for controlling the dead time generated in the EDXRF measurement process [30], [33].

In the vacuum chamber, the electrons are accelerated, by a high voltage, towards a target made of a high-purity metal (such as Cr, Ag, W, Mo, Rh, and Pd), the anode [33]. The interaction of these electrons with the anode produces X-rays of continuous energy (*bremsstrahlung*) and characteristic lines with specific values for the target material. With the exclusive increase of the tube voltage, the energy range, the maximum intensity, and the energy corresponding to the continuous spectrum maximum value also increase [18]. At the same time, the greater the current intensity applied to the cathode, the more intense the radiation beam produced in the tube. Figure 7 illustrates the behavior of the spectrum emitted by an X-ray tube with an Rh anode when changing its voltage and current.

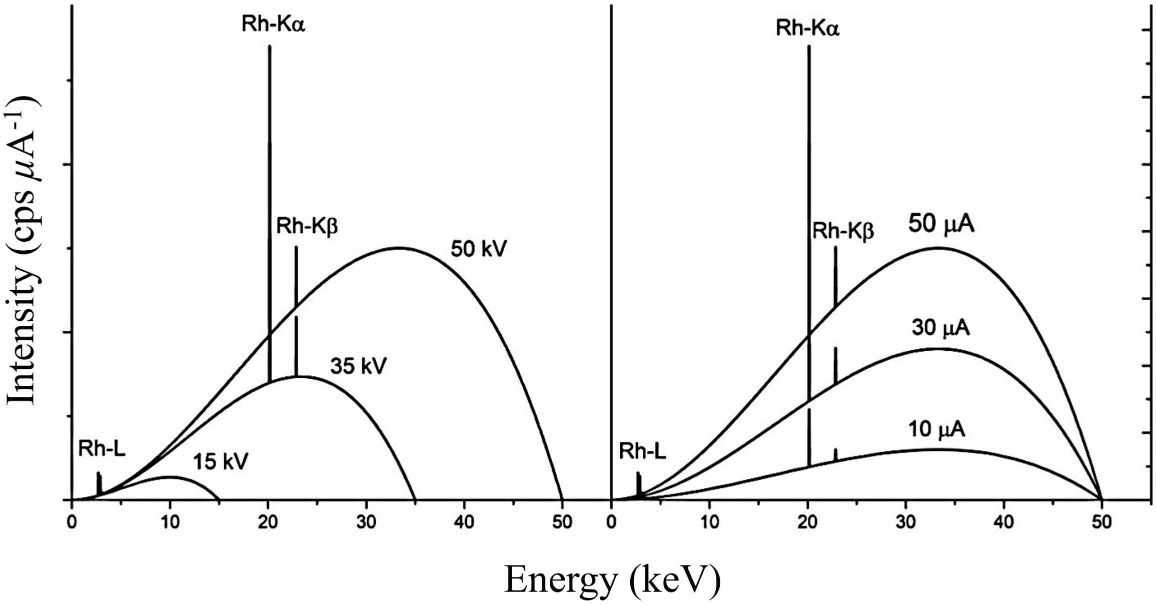

Figure 7 – Illustration of an X-ray spectrum emitted by an Rh-targeted tube with voltage and current variation. Adapted from [30]

Therefore, with adequate voltage and current adjustments, the tube X-ray intensities distribution may be altered to favor some interest energy range and, consequently, increasing the measurement sensitivity for elements with absorption edges close to this range.

The detection is performed by a semiconductor detector. Currently, silicon detectors such as the Si-PIN and the SDD (Silicon Drift Detector) are the most used in commercial XRF systems because they are cooled by the Peltier effect, not requiring external cooling with liquid nitrogen as in the case of Si(Li) detectors [33], [34].

When the characteristic X-rays arrive at the detector, they generate electronic pulses, which are separated according to their amplitudes, which are proportional to their energies. Then, this data is sent through the electronics associated with the measurement system to, in the final step, generate a spectrum with X-ray intensities versus energy.

It is important to mention that due to the characteristics of the detection system used, some peaks, called artifacts, may arise in the EDXRF spectrum. In contrast to scattering (elastic and inelastic) and the *bremsstrahlung*, which are only due to the interaction of radiation with matter, artifacts come from sources other than exclusively the elements in the sample [17]. The main artifacts that arise in the EDXRF spectra are the escape peak and the sum peak.

On Si semiconductor detectors, an incident characteristic X-ray of energy higher than the Si-K absorption edge produces Si-K characteristic X-rays (Si Kα and Kβ) due to the X-ray fluorescence of the atoms that constitute the detector. These X-rays are absorbed into the detector volume and contribute to the overall charge collected for the original incident X-ray photon. However, there is a finite probability that the generated Si-K X-rays escape the detector volume. The highest probability of escape arises near to the front of the detector crystal and, in such cases, the detected energy will be reduced according to the energy of the escaped Si-K X-ray [17], [18]. These are called escape peaks. This artifact may happen with X-rays characteristic of all elements, but due to the probability of occurrence, not all of them will have peaks with intensities high enough to be observed in the XRF spectrum. The escape probability is highest for lines closest to the Si-K absorption edge [17]. For instance, if the characteristic X-rays of the Fe Kα are the highest intensity lines in the spectrum of an analyzed sample, escape peaks of energy equivalent to the subtraction between the Fe Kα and Si Kα lines will arise with greater intensity (4.665 keV).

Sum peaks arise when two characteristic X-rays produce electronic pulses in the detector so close in time that the processing electronics fail to recognize them as two separate events and record them as the same characteristic X-ray with an energy equal to the sum of the two [17], [18]. Although this artifact may seem unlikely to occur because the processing time scales of the XRF instruments are very small (on the order of microseconds, 10^{-6} *s*), the times involved in the processes of fluorescence and characteristic X-ray emission are much shorter, on the order of femtoseconds (10^{-15} *s*) [17]. The probability of sum peaks occurring is greater for the more intense lines of the EDXRF spectrum. For instance, in the spectra of soil samples where the Fe Kα, Fe Kβ, and Ti Kα lines are intense, sum peaks corresponding to Fe Kα + Fe Kα (12.810 keV), Fe Kα + Fe Kβ (13.464 keV), and Fe Kα + Ti Kα (10.917 keV) will arise (see Figure 12).

On the other hand, the advancement and development of new technologies allowed the EDXRF system miniaturization, creating portable X-ray fluorescence (pXRF) devices. These equipment's are typically gun-like instruments that work on the same principles as laboratory-based EDXRF spectrometers [35], [36]. Figure 8, is a simple illustration of how this hand-held device may be built.

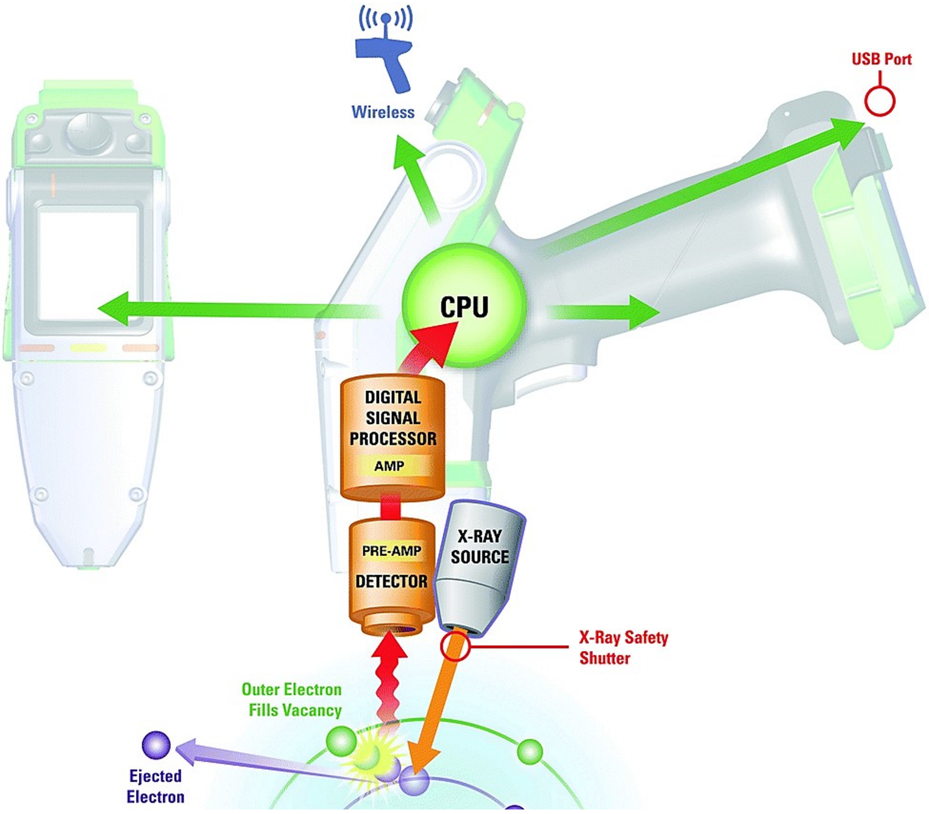

Figure 8 - Schematic illustration of a portable XRF system. Adapted from [36]

These devices have portability as their great advantage. With it, *in situ* measurements are possible. However, that is not the only difference between them and benchtop EDXRF. They may also have faster signal processing systems and are generally less expensive, however, they may be less robust, *i.e.*, presented lower resolution.

1.3.2.3 Primary filters use

In XRF, a primary beam filter acts as an X-ray absorber. It sits between the X-ray tube and the sample to modify the output X-ray tube spectrum to which the sample is exposed. For example, aluminum acts as a single absorption filter, having a single absorption edge at 1.56 keV. Figure 9, adapted from [17], provides the scattered excitation spectrum provided by a silver target X-ray tube operated at 15 kV. The plot shows the effect of two thicknesses of an Al primary beam filter. It is noted that, from the value corresponding to Al K-edge, energy for which the greatest absorption of X-rays by the filter occurs, the spectrum is significantly reduced, and the values start to grow again at 4.5 keV for the fine filter and 5.5 for the thick filter. In this way, the tube's Ag L-lines are completely absorbed. This occurs because of Al high mass absorption coefficient for the employed energy. The bremsstrahlung maximum value at lower energy provided by the spectrum filtered through thin Al yields good peak/background (P/B) ratios for elements in the S to V range (2.3-5 keV) because the absence of the Ag L-lines from the tube removes their spectral interference and avoids the need for these X-rays to consume counting capacity. The thick aluminum filter ‘‘pinches’’ the low-energy side of the bremsstrahlung maximum value, providing high P/B ratios for the elements with lines in the range 3-8 keV [17]. In XRF equipment excited with Rh target tube, Al filters produce similar changes in the spectrum because the Rh L-lines (Lα 2.697 and Lβ 2.834 keV) are close to Ag (Lα 2.983 and Lβ 3.150 keV).

The proper combination of kV and an absorption filter may provide an energy region in which the P/B ratio is optimal. On the low-energy side of this region, the excitation is suppressed, which allows more of the system’s counting capacity to be used in the region itself. On the high-energy side of the region, the excitation from the tube is tuned to provide a high integrated intensity above the absorption edges energy lines of the interest element [17], [18].

The thick Cu filter is another example of a used filter. The Cu filter used is generally thick enough to completely absorb the X-ray tube anode K-lines, which are of high energy (for Ag target, Kα is 22.163 keV and Kβ is 24.941 while for Rh target, Kα is 20.216 keV and Kβ is 22.724 keV). Such an arrangement provides for the effective determination of heavier elements. For example, the quantified of Ag and Cd using a silver anode X-ray tube [17].

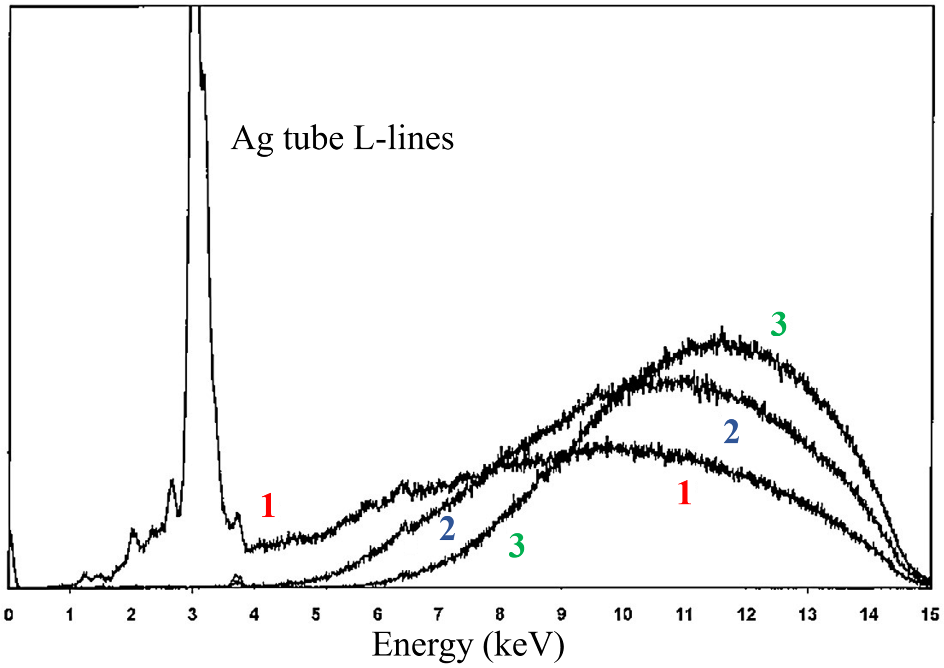

Figure 9 - Scattered excitation spectrum provided by a silver target x-ray tube operated at 15 kV. The plots show the effect of two thicknesses of an Al primary beam filter: (1) unfiltered; (2) thin Al filter; (3) thick Al filter. Adapted from [17]

1.3.3 Multivariate analysis

In general terms, multivariate analysis refers to all statistical methods that simultaneously analyze multiple variables in a dataset under investigation. In situations where data are correlated and with a high variable number, analysis by univariate methods is complex and laborious, on the other hand, multivariate analysis may be a facilitating approach. In this perspective, the multivariate analysis associated with spectroscopic techniques is advantageous because it may improve the results accuracy, modeling in the interferents presence, and extract subtle information that is not explicit [37], [38].

Among the different multivariate analysis subareas, two are highlighted: exploratory analysis and machine learning (or multivariate) calibration. As an exploratory method, principal component analysis (PCA) was used in this work. In addition, for modeling fertility attributes through XRF data, partial least squares linear regression (PLS) and predictive analysis of common dimensions analysis (P-ComDim) were also performed.

1.3.3.1 Principal components analysis

Principal component analysis is an unsupervised exploratory method used to project n-dimensional data into a lower-dimensional space. The information contained in the original data space (n samples versus p variables) is compressed by linear combinations of variables to a space of uncorrelated axes (called PC's), generally of 2 or 3 order, friendly in the results visualization and interpretation. PCA seeks to capture the greatest possible amount of variability present in the data through these orthogonal PCs, which have variance explains in descending order. Thus, by extracting the main sources of data variation, the PCA allows visualization and interpretation of subtle differences among the samples in variables terms, facilitating the hidden pattern identifications [38].

Geometrically, PCA is a projection technique, in which the original dataset matrix is projected onto the PCs subspace. Matrix-wise, the PCA decomposes the original data matrix (***X***) into three new matrices, the scores matrix (***T***), the loadings matrix (***P***) and the residuals (***E***). One way to do this is to perform a singular value decomposition (SVD) on ***X***, resulting in

(13

)

The junction of the first ***T*** column with the first ***P*** row is the principal component number one (PC 1), the junction of the second ***T*** column with the second ***P*** row is the PC 2 and so on. In this way, the data matrix is separated into a series of PCs plus the residual, which contains all information not explained by the data set PCs and is called noise. Due to this, PCA is able to separate the important from redundant information.

The PCA helps in the elaboration of general hypotheses from the dataset. This is done by investigating the scores and loads matrices, simultaneously. By plotting scatterplots of PC scores against each other, clusters between the samples may be verified. Doing the same with the loadings, clusters between the variables are presented. Thus, analyzing the two plots together, the variables responsible for the correlations among the samples are identified.

Beyond to the SVD, there are other algorithms to calculate the ***T*** and ***P*** matrices from ***X***. Among the main ones are the correlation (or covariance) matrix diagonalization and the NIPALS algorithm [39].

1.3.3.2 Partial least square regression

The Partial least square regression (PLS) is a machine learning (or multivariate) algorithm that establishes a quantitative association between a matrix **X** of independent data (spectra matrix) and a vector **y** of dependent data (fertility attributes matrix) so that the maximum covariance between **X** and **y** is reached [40]. Thus, the PLS maintains a compromise between explaining the variance in **X** and obtaining the highest correlation with **y** [38].

Initially, a SVD is performed on ***X*** and ***Y,*** independently,

(14
)

(15
)

T and ***U*** are the orthogonal scores matrices, the projections of the original data in the smallest dimension space defined by the number of principal components (PC's) of ***X*** and ***Y***, respectively; ***P*** and ***Q*** are the loadings matrices, the relative contribution coefficients in each PC of ***X*** and ***Y***, respectively; ***E*** and ***F*** are the error matrices of ***X*** and ***Y***, respectively and the subscript *i* represents the number of PC's ranging from 1 to *A*.

After that, a linear association between the ***X*** and ***Y*** scores is sought with the construction of the regression vector ***b***

(16
)

In this process, the dependents variables ***Y*** are projected on to the space of each PC in ***X***. Geometrically, the ***Y*** are incorporated in ***t*** scores by rotations in each PC of ***X*** to reach the maximum covariance with the ***Y***. Due to this, there is a loss of orthogonality of the PC's, which start to receive the terminology of latent variables (LV's).

This is done with several steps to obtain each LV in an iterative process in which the projection of the samples on the loadings for determining the scores and the adjustment between the scores of ***X*** and ***Y*** are optimized at the same time.

Table 1 - Steps of the NIPLS algorithm for the PLSR method. Adapted from [38]

Step	**Description**	**Equation**
1	Preprocess ***X*** and ***Y***	
2	Calculate the ***W*** matrix (which is the projection of the ***X*** variables onto ***Y***$_{i-1}$)	
3	Normalize ***W***	
4	Calculate the ***t*** scores of ***X***	
5	Calculate the ***p*** loadings of ***X***	
6	Calculate the ***q*** loadings of ***Y***	
7	Remove the effects of the ith factors from the original matrices. For *i* = 1, 2, ..., A	
8	Repeat steps 2 to 8 until A LV's	
9	Calculate the ***b*** regression vector	

Although others exist today, the NIPALS algorithm, proposed by Wold [39], was the first PLS algorithm used to build regression models of this class. The Table 1 presents a summary of the steps performed by NIPALS [38], starting from the ***X*** and ***Y*** matrices and determining the regression vector ***b***. For this, a weight factor matrix (***W***) is considered. The restriction of highly correlated ***t*** and ***y*** is ensured by making ***W*** proportional to the covariance between ***X*** and ***Y***.

1.3.3.3 Predictive analysis of common dimensions

The ComDim (Common Dimensions) methodology is a non-supervised multi-block analysis aiming to extract common information in data generated by different techniques [41]. These data are represented by matrices $\mathbf{X_1}$**,** $\mathbf{X_2}$**, ...**$\mathbf{X_b}$, where *b* indicates the number of techniques used. The ComDim analysis was designed to assess the associations between individuals and variables within a multiblock setting where several variables, organized in blocks, are measured on the same individuals [41]. The Predictive ComDim (P-ComDim) is a supervised extension of the ComDim method in which a response data vector (**y**) is considered. The aim is to investigate the association between ***y*** and $\boldsymbol{X_1}$***, …,*** $\boldsymbol{X_b}$. Latent variables (in this case, common dimensions (CD)) are derived and used to highlight the relationships between ***y*** and X_1*,…,* X_b. The latent variables could also be used in a regression model to predict **y** from $\boldsymbol{X_1}$***, …,*** $\boldsymbol{X_b}$.

Thus, a singular value decomposition (SVD) on the matrices are performed (*i* ranges from 1 to *b*). This "common singular value decomposition" (because it is related to $\boldsymbol{X_b}$ and ***y***) will involve not only a component common to the X-blocks (***t scores***) but also a component in the y-space (***u scores***). First, ***t*** and ***u*** are considered to be length 1 and after the SVD they are updated according to the weight assigned to each X-block. After that, a new SVD is performed and the components are updated again. This algorithm is repeated until the difference in residuals between the last two iterations is less than a pre-specified threshold (normally tending to zero). With this, the first common dimension is obtained. Before proceeding with the same steps to obtain the second common dimension, a deflation of the X-blocks and the y-block with respect to ***t*** is performed. Then, the same algorithm run again with $\boldsymbol{X_1}$***, …,*** $\boldsymbol{X_b}$ and **y** deflated [42].

For each common dimension extracted, there is a set of weights (called saliences), scores, and loadings. Salience is the weight assigned to each data block for constructing a specific CD and may be interpreted as the block variance represented in that CD. The scores are projections of the samples in the CD common space, and the loadings are the weights of the variables contained in each data block [43]–[45]. In the end, a multiple linear regression (MLR) between the y-block and the common dimension scores is built.

Formally, the P-ComDim algorithm runs as follows to calculate de salience of *i*-th X-block in a first CD ():

Table 2 - Steps of the P-ComDim algorithm to determine the ith salience ()

Step	**Description**	**Equation**
1	Randomly initialize $\boldsymbol{t}^{(1)}$ and $\boldsymbol{u}^{(1)}$, both of unit length	
2	Calculate the i-th salience by	
3	Update ***u*** and ***t*** successively	
4	Reiterate the algorithm starting from Step 2 until	

To determine the *i*-th salience of second CD), a deflation of the X-blocks and the Y-block with respect to the ***t*** performed is and , respectively (is the identity matrix). Thereafter, the same algorithm is run anew on these deflated data tables instead of the original data tables. The same process extends in a natural way for the determination of subsequent *i*-th saliences of other CDs.

1.4 References

[1] J. A. M. Demattê *et al.*, “The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges,” *Geoderma*, vol. 354, p. 113793, Nov. 2019, doi: 10.1016/j.geoderma.2019.05.043.

[2] D. H. Wall and U. N. Nielsen, “Biodiversity and ecosystem services: is it the same below ground,” *Nature Education Knowledge*, vol. 3, no. 12, p. 8, 2012.

[3] United Nations Department of Economic and Social Affairs, “World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100,” *New York: United Nations Department of Economic and Social Affairs, Population Division.* 2017. Accessed: Jul. 16, 2023. [Online]. Available: https://www.un.org/development/desa/en/news/population/world-population-prospects-2017.html

[4] A. McBratney, D. J. Field, and A. Koch, “The dimensions of soil security,” *Geoderma*, vol. 213, pp. 203–213, Jan. 2014, doi: 10.1016/j.geoderma.2013.08.013.

[5] J. A. M. Demattê, A. C. Dotto, L. G. Bedin, V. M. Sayão, and A. B. e Souza, “Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact,” *Geoderma*, vol. 337, pp. 111–121, Mar. 2019, doi: 10.1016/j.geoderma.2018.09.010.

[6] F. R. dos Santos, J. F. de Oliveira, E. Bona, J. V. F. dos Santos, G. M. C. Barboza, and F. L. Melquiades, “EDXRF spectral data combined with PLSR to determine some soil fertility indicators,” *Microchemical Journal*, vol. 152, p. 104275, Jan. 2020, doi: 10.1016/j.microc.2019.104275.

[7] R. A. Viscarra Rossel and J. Bouma, “Soil sensing: A new paradigm for agriculture,” *Agric Syst*, vol. 148, pp. 71–74, Oct. 2016, doi: 10.1016/j.agsy.2016.07.001.

[8] G. Archbold Taylor *et al.*, “pH Measurement IoT System for Precision Agriculture Applications,” *IEEE Latin America Transactions*, vol. 17, no. 05, pp. 823–832, May 2019, doi: 10.1109/TLA.2019.8891951.

[9] M. T. Eitelwein, T. R. Tavares, J. P. Molin, R. G. Trevisan, R. V. de Sousa, and J. A. M. Demattê, “Predictive Performance of Mobile Vis–NIR Spectroscopy for Mapping Key Fertility Attributes in Tropical Soils through Local Models Using PLS and ANN,” *Automation*, vol. 3, no. 1, pp. 116–131, Feb. 2022, doi: 10.3390/automation3010006.

[10] J. V. Fontenelli *et al.*, “Evaluating the synergy of three soil spectrometers for improving the prediction and mapping of soil properties in a high anthropic management area: A case of study from Southeast Brazil,” *Geoderma*, vol. 402, p. 115347, Nov. 2021, doi: 10.1016/j.geoderma.2021.115347.

[11] D. Wang *et al.*, “Synthesized use of VisNIR DRS and PXRF for soil characterization: Total carbon and total nitrogen,” *Geoderma*, vol. 243–244, pp. 157–167, Apr. 2015, doi: 10.1016/j.geoderma.2014.12.011.

[12] T. R. Tavares *et al.*, “Multi-Sensor Approach for Tropical Soil Fertility Analysis: Comparison of Individual and Combined Performance of VNIR, XRF, and LIBS Spectroscopies,” *Agronomy*, vol. 11, no. 6, p. 1028, May 2021, doi: 10.3390/agronomy11061028.

[13] T. R. Tavares *et al.*, "Laser-Induced Breakdown Spectroscopy (LIBS) for tropical soil fertility analysis," *Soil Tillage Res*, vol. 216, p. 105250, Feb. 2022, doi: 10.1016/j.still.2021.105250.

[14] F. R. dos Santos, J. F. de Oliveira, G. M. C. Barbosa, and F. L. Melquiades, "Comparison between energy dispersive X-ray fluorescence spectral data and elemental data for soil attributes modelling," *Spectrochim Acta Part B At Spectrosc*, vol. 185, p. 106303, Nov. 2021, doi: 10.1016/j.sab.2021.106303.

[15] F. R. dos Santos, J. F. de Oliveira, E. Bona, G. M. C. Barbosa, and F. L. Melquiades, "Evaluation of pre-processing and variable selection on energy dispersive X-ray fluorescence spectral data with partial least square regression: A case of study for soil organic carbon prediction," *Spectrochim Acta Part B At Spectrosc*, vol. 175, p. 106016, Jan. 2021, doi: 10.1016/j.sab.2020.106016.

[16] D. Xu *et al.*, "Multi sensor fusion for the determination of several soil properties in the Yangtze River Delta, China," *Eur J Soil Sci*, vol. 70, no. 1, pp. 162–173, Jan. 2019, doi: 10.1111/ejss.12729.

[17] R. E. Van Grieken and A. A. Markowicz, *Handbook of X-ray spectrometry*, 2nd ed., vol. 29. New York, 2002.

[18] R. Jenkins, "X-Ray Fluorescence Spectrometry," in *Handbook of Analytical Techniques*, Weinheim, Germany: Wiley-VCH Verlag GmbH, 1988, pp. 753–766. doi: 10.1002/9783527618323.ch23.

[19] T. M. Mitchell, *The discipline of machine learning*, vol. 9. Carnegie Mellon University, School of Computer Science, Machine Learning~…, 2006.

[20] L. C. Prezotti and A. M. Guarçoni, *Guia de interpretação de análise de solo e foliar*. Vitória: Incaper, 2013.

[21] L. F. SOBRAL, M. C. de V BARRETO, A. J. Da Silva, and J. L. Dos Anjos, "Guia prático para interpretação de resultados de análises de solos.," 2015.

[22] A. S. Lopes, "Manual internacional de fertilidade do solo," *Piracicaba: Potafos*, pp. 79–85, 1998.

[23] C. C. Ronquim, "Conceitos de fertilidade do solo e manejo adequado para as regi\~oes tropicais.," 2010.

[24] A. D. Day and K. L. Ludeke, "Exchangeable Bases," 1993, pp. 27–29. doi: 10.1007/978-3-642-77652-6_7.

[25] P. C. Teixeira, G. K. Donagemma, A. Fontana, W. G. Teixeira, and others, "Manual de métodos de análise de solo." Embrapa Bras\'\ilia, DF, 2017.

[26] I. Kaplan, "Nuclear physics," *(No Title)*, 1963.

[27] W. R. Leo, "Statistics and the treatment of experimental data," *Techniques for nuclear and particle physics experiments: A how-to approach*, pp. 81–113, 1994.

[28] R. Eisberg and R. Resnick, *Quantum physics of atoms, molecules, solids, nuclei, and particles*. 1985.

[29] R. Cesareo, A. L. Hanson, G. E. Gigante, L. J. Pedraza, and S. Q. G. Mathaboally, "Interaction of keV photons with matter and new applications," *Phys Rep*, vol. 213, no. 3, pp. 117–178, Apr. 1992, doi: 10.1016/0370-1573(92)90086-F.

[30] R. Klockenkämper and A. Von Bohlen, *Total-reflection X-ray fluorescence analysis and related methods*. John Wiley & Sons, 2014.

[31] V. F. Nascimento Filho, "Técnicas analíticas nucleares de fluorescência de raios X por dispersão de energia (EDXRF) e por reflexão total (TXRF)," *Piracicaba: Escola Superior de Agricultura Luiz de Queiroz*, 1999.

[32] R. M. C. Silva, V. F. Nascimento Filho, and C. R. Appoloni, "Fluorescência de Raios X por dispersão em energia," *LFNATEC-Publicação Técnica do Laboratório de Física Nuclear Aplicada*, vol. 8, no. 1, 2004.

[33] K. Janssens, "X-Ray Fluorescence Analysis," *Handbook of Spectroscopy: Second, Enlarged Edition*, pp. 449–506, 2014.

[34] M. West *et al.*, "Atomic Spectrometry update-X-ray fluorescence Specctrometry," *J Anal At Spectrom*, vol. 27, pp. 1603–1644, 2012.

[35] C. R. Ward *et al.*, "In-situ inorganic analysis of coal seams using a hand-held field-portable XRF Analyser," *Int J Coal Geol*, vol. 191, pp. 172–188, Apr. 2018, doi: 10.1016/j.coal.2018.03.012.

[36] P. J. Potts and M. Sargent, "In situ measurements using hand-held XRF spectrometers: a tutorial review," *J Anal At Spectrom*, vol. 37, no. 10, pp. 1928–1947, 2022, doi: 10.1039/D2JA00171C.

[37] N. Nagata, M. I. M. S. Bueno, and P. G. Peralta-Zamora, "Métodos matemáticos para correção de interferências espectrais e efeitos interelementos na análise quantitativa por fluorescência de raios-X," *Quim Nova*, vol. 24, no. 4, pp. 531–539, Aug. 2001, doi: 10.1590/S0100-40422001000400015.

[38] M. M. C. Ferreira, *Quimiometria: conceitos, métodos e aplicações*. Editora da Unicamp, 2015. doi: 10.7476/9788526814714.

[39] H. Wold, "Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach," *J Appl Probab*, vol. 12, no. S1, pp. 117–142, Sep. 1975, doi: 10.1017/S0021900200047604.

[40] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Anal Chim Acta*, vol. 185, pp. 1–17, 1986, doi: 10.1016/0003-2670(86)80028-9.

[41] A. El Ghaziri, V. Cariou, D. N. Rutledge, and E. M. Qannari, "Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of (K + 1) datasets," *J Chemom*, vol. 30, no. 8, pp. 420–429, Aug. 2016, doi: 10.1002/cem.2810.

[42] V. Cariou, D. Jouan-Rimbaud Bouveresse, E. M. Qannari, and D. N. Rutledge, "ComDim Methods for the Analysis of Multiblock Data in a Data Fusion Perspective," 2019, pp. 179–204. doi: 10.1016/B978-0-444-63984-4.00007-7.

[43] L. M. de Aguiar, D. Galvan, E. Bona, L. A. Colnago, and M. H. M. Killner, "Data fusion of middle-resolution NMR spectroscopy and low-field relaxometry using the Common Dimensions Analysis (ComDim) to monitor diesel fuel adulteration," *Talanta*, vol. 236, p. 122838, Jan. 2022, doi: 10.1016/j.talanta.2021.122838.

[44] E. M. Qannari, I. Wakeling, P. Courcoux, and H. J. H. MacFie, "Defining the underlying sensory dimensions," *Food Qual Prefer*, vol. 11, no. 1–2, pp. 151–154, Jan. 2000, doi: 10.1016/S0950-3293(99)00069-5.

[45] G. Y. F. Makimori and E. Bona, “Commercial Instant Coffee Classification Using an Electronic Nose in Tandem with the ComDim-LDA Approach,” *Food Anal Methods*, vol. 12, no. 5, pp. 1067–1076, May 2019, doi: 10.1007/s12161-019-01443-5.

CHAPTER 2. OPTIMIZATION OF PXRF INSTRUMENTATION CONDITIONS AND MULTIVARIATE MODELING IN SOIL FERTILITY ATTRIBUTES DETERMINATION

2.1 Introduction

Soil stands as one of the humanity's most important resources. Its chemical and granulometric characteristics influence several factors such as water dynamics, climate, organisms, forests, carbon cycles, and more. Apart from its undeniable environmental significance, soils serve as the foundation for food production [1].

It is projected that the global population will reach 9.8 billion by 2050, leading to a continuous increase in demand for agricultural products, primarily for food. The current world scenario where environmental imbalances are increasingly overwhelming means that this increase in production cannot be carried out anyway [2], [3]. This concern was discussed in McBratney *et al*. [4], where the authors introduced the term “soil security” for indicate that it is imperative to take care of soils or we will likely have problems in several areas, including the people's hunger [1]. A potential solution to promote soil-friendly food security lies in the optimization of agronomic productivity in a sustainable manner, such as achieving higher yields from the same cultivable land.

Soil fertility plays a crucial role in improving agricultural production. Preservation of soil fertility and productivity depends on several factors, including periodic analysis [5] and comprehensive knowledge of the spatial variability of fertility attributes [6]. Periodic analysis is fundamental in the evaluation and monitoring of chemical and physical attributes of agronomic interest. Access to high-density information regarding the spatial variability of soil fertility attributes is crucial in the decision-making process to promote optimal nutritional conditions for plants. For instance, adjusting agricultural inputs and implementing liming practices based on local changes in soil properties may lead to both economic and productive benefits.

In general, the key soil fertility attributes analyzed are soil organic carbon (SOC), available nutrients (Ca, Mg, K, and P), and fertility indicators such as cation exchange capacity (CEC), the sum of exchangeable bases (SB), base saturation percentage (BSP), pH, and potential acidity (H+Al) [7]. Traditionally, the samples are manually collected and sent to laboratories where acidic or basic extractions are carried out [8]. More exactly, an analysis may take about 3 to 15 days for delivering results [9]. In addition, the volume of chemical substances involved in these analyses is high. For instance, according to Demattê *et al*. [1], to determine the soil organic matter (OM), the wet combustion is predominantly applied. This methodology uses 0.196 g of dichromate ($Cr_2O_7^{2-}$), 1.20 g of ammonium ferrous sulfate hexahydrate ($Fe(NH_4)_2(SO_4)_26H_2O$), and 5 mL of sulfuric acid (H_2SO_4) for only one sample. Considering the potential of 600 million soil samples to be analyzed globally per year, around 840 thousand kg of dichromate and ammonium ferrous sulfate and 3 million L of sulfuric acid will be consumed. Besides, for OM, it is estimated a cost of US$ 5.00 per sample, with an annual expenditure of US$ 2.5 million. Consequently, the analysis of a large number of samples require hard preparation, are destructive, expensive, time-consuming, use reagents and may generate a considerable amount of waste, which must be correctly disposed. Thus, the development of faster and cheaper methodologies based on green chemistry has become a priority [8], [10].

Spectral-analytical techniques have been used as an alternative methodology to assess soil properties quickly, economically, and environmentally friendly [11], [12]. Techniques such as Near Infrared Spectroscopy (NIR), Laser Induced-Breakdown Spectroscopy (LIBS), and Energy Dispersive X-ray Fluorescence (EDXRF), when associated with multivariate analysis, have shown their ability to quantify soil fertility attributes. Although they are not established as a reference, they already demonstrate their potential and occupy their space in research focused on this area [10], [12]–[18]. In this perspective, EDXRF has been successfully employed in the analyses of several soil attributes, among which are those previously mentioned [10], [17], [19]–[26].

EDXRF is a non-destructive atomic emission technique based on photoelectric effect for the characteristic X-rays production after atomic deexcitation. It provides the advantageous possibility of multielement analysis, quickly and at low cost [27], [28]. With an appropriate detector, the elements present in the sample are identified and quantified by evaluating the peak intensities in the spectra. The determination of elements lighter than Al is affected by the low fluorescent yield, and due to the low characteristic X-ray energies, the efficiency of Si semiconductor (widely used in EDXRF instruments) to detect them is also low. Consequently, the EDXRF is adequate to quantify the total inorganic concentration of elements with $Z > 13$. On the other hand, most EDXRF equipment is operated at an energy range in which the photoelectric effect is not the only predominant interaction phenomenon

[29], [30]. The radiation scattering by electrons in the analyzed sample is also likely to occur. Elements with atomic number lower than 8 have higher cross-sections for Compton scattering. Thus, the Compton scattering may provide information about light elements and the general composition of complex organic samples [31].

In most published studies which use EDXRF associated with multivariate analysis to predict soil attributes, the data matrix used consists in a set of elemental concentration values obtained from an instrument quantitative routine. However, the spectral data is considered a dataset with more information than a set of concentration values and may provide better performance and interpretability in the modeling process. Notwithstanding, several operating conditions are possible to be applied in the measurement of an X-ray spectrum in a soil sample. To efficiently quantify the total concentration of light and heavy elements in the same sample, different measurement conditions are required. As an example: 15 kV in the X-ray tube for light elements (from Na to Sc) and 50 kV for heavy elements (Ti to U)[32]. It implies in a longer measurement time. Conversely, the use of portable EDXRF (pXRF) equipment in this methodology is advantageous because they allow for faster measurements and lower operating costs when compared to the benchtop EDXRF. They also provide the promising possibility of *in situ* measurements. *In situ* prediction based in laboratory-calibrated models is an ambitious and challenging proposal that still needs to be widely assessed.

Although there are already studies using pXRF with multivariate analysis to predict soil fertility attributes [21], [33], [34], few of them compare benchtop EDXRF versus pXRF performances. In this sense, this research was motivated by the study about the redundant information in the spectra from three different operating conditions in a portable EDXRF equipment, aiming to find an optimized measurement condition to predict soil attributes and compare it with the modeling at EDXRF benchtop data. For that, spectra generated at 15, 30 and 50 kV in the X-ray tube and elemental concentrations determined by a commercial quantitative pXRF routine of soil samples were modeled with the Partial Least Squares Regression (PLS) to predict soil fertility attributes and the optimized experimental condition was evaluated.

Considering that data from different measurements conditions will be evaluated, data fusion approach are interesting tools to be applied. Data fusion seeks to exploit the synergism among disparate data to improve predictive accuracy [35]. The combination of data may be carried out basically at three levels: low, mid, and high. Low- and mid-level was used. In the low-level fusion, the model is built with data from all sources concatenated in a single matrix. Thus, the concatenated model performance allows the synergy evaluation among the different data sources in the property of interest prediction. The mid-level fusion requires further steps. Firstly, extracts the relevant features from each data source separately. After that, the concatenation process is performed, and the new dataset generated is used for modeling. There are different ways to extract the relevant information, *e.g.*, through principal component analysis (PCA) scores [35]. Although still little explored with XRF spectral data, the Predictive Analysis of Common Dimensions (P-ComDim) is an alternative to mid-level fusion methods. In this approach, the common information shared by each spectral class is identified, enabling the interpretation of its relevance for the prediction attribute of interest [36].

From this perspective, the spectra data from the three different measurements conditions were fused at the low- and mid-level approaches before the modeling. The worse performance of the fused compared to the individual models indicates the lack of complementarity between the data, allowing to discard the requirement to measure with three experimental conditions together.

Facing these motivations, the hypotheses elaborated in this study were: i) The pXRF spectra data modeled with PLS generate models capable of predicting soil fertility attributes with good performance and confidence; ii) Models with fused data provide models with equivalent performance than models built with individual datasets, enabling the optimized use of a single experimental condition for modelling; iii) pXRF spectra produce results equivalent to those constructed with benchtop EDXRF spectra. The optimized models' performance was compared with those reported by Dos Santos *et al.* [18], [26], in which the same sample area was explored.

Finally, with the optimal conditions for modeling established, a preliminary study assessing *in situ* extrapolation of laboratory-calibrated models was conducted.

2.2 Methodology

2.2.1 Soil Sampling and Reference Analyses

The study was carried out in an agricultural area at Ribeirão Vermelho basin in Cambé municipality, Paraná State, Brazil. The area comprises the latitudes between 23°09'59.70" and 23°09'50.14" S and longitudes 51 °14'42.27" and 51°14'56.87"W. The region is located in the north of Paraná State and covers 12 ha.

The soils were classified as Eutrophic Red Latosol and Eutrophic Red Nitosol [37] or Rhodic Ferralsol and Rhodic Nitisol [38], respectively. These soils are highly weathered from basalt and and has a clay texture (760 g kg^{-1}), whose predominance is kaolinite, iron and aluminum oxyhydroxides. The average slope of the area is 0.09 m m^{-1}, where the altitude varies between 544 m and 514 m. No-tillage has been used for over 17 years with scarification carried out on average every five years. During this period soybeans were cultivated in the summer and corn in the winter. Oats were sometimes grown in winter.

394 soil samples were collected at three depths (0–5 cm, 5–10 cm and 10–20 cm). In the laboratory, the samples were dried at 40 ºC for 48 h and ground in a 2 mm sieve for the soil attributes chemical analysis and EDXRF measurements.

The analyzed soil fertility attributes were: exchangeable macronutrients (Ca^{2+}, Mg^{2+}, and K^{+}), available phosphorus extracted by Mehlich-1 (P_M), pH, potential acidity (H+Al), soil organic carbon (SOC), cation exchange capacity (CEC), sum of exchange bases (SB) and base saturation percentage (BSP) [39].

The SOC contents were determined by the dichromate oxidation method (Walkley-Black). Exchangeable contents of Ca^{+2} and Mg^{+2} was extracted by KCl 1 mol L^{-1} solution, measured by atomic absorption spectrometry while flame photometry and UV-Vis molecular absorption spectrophotometry using Mehlich-1 extractor solution were used for the K^{+} and the P_M quantifications, respectively. For the pH scale, the potentiometer in a $CaCl_2$ solution at a ratio of 1:2.5 was used. By adding the SMP buffer solution, the pH-SMP readings were performed to determine the H+ Al. With these values, the SB ($Ca^{2+} + Mg^{2+} + K^{+}$), CEC ($SB + H^{+} + Al^{3+}$) and BSP ($100 \times SB \times CEC^{-1}$) were estimated.

All conventional analyses were carried out in the IDR Soil Analysis Laboratory in Londrina, Paraná, Brazil, following the recommendations of Pavan *et al.* (1992) [39]. The values of fertility attributes quantified by conventional methods were organized in vectors **y** and were used as dependent variables in the models.

2.2.2 pXRF Analysis

Approximately five grams of soil in loose powder form were accommodated in polyethylene XRF cups (nº.1530, Chemplex Industries Inc., USA) sealed with Mylar film for XRF scanning in the Tracer 5i (Bruker Inc.) handheld equipment. This amount of sample resulted in around 10 mm sample thickness, being considered an infinitely thick sample for the X-ray energy range considered.

The instrument was operated in the GeoExploration mode. It is an empirical package which is factory calibrated using several standard soil samples covering geologically relevant ranges for 48 elements from Mg to U. The output of the Geoexploration mode is a table with the elemental concentration and their uncertainties. Measurements in this mode are carried out in three phases: (i) 30 kV and 13 μA with 25 μm of Ti and 300 μm of Al fiters (30kV condition); (ii) 50 kV and 22.85 μA with the aforementioned filters plus a 75 μm of Cu filter (50kV condition); (iii) 15 kV and 12.15 μA without filters (15kV condition). The elements of $Z < 20$ are quantified using the 15 kV spectrum with a correction for Compton scattering. The excitation was performed in air atmosphere using primary X-rays generated from a Rh tube for 30 seconds for each phase. A Silicon Drift Detector (SDD) was used, and the spectra were recorded in 2048 channels. The spectra were normalized by tube current and detector live time, resulting in spectra units of counts of photons per second per microampere (cps μA^{-1}).

All samples were measured three times in different sample portions, shaking the XRF cup before each measurement. A mean spectrum was calculated for each sample. The pXRF spectra generated by each phase (measurement condition) and the elemental concentrations were organized into individual matrices X and were used as independent variables to build the PLS individual models.

To evaluate the GeoExploration quantitative mode performance, the elemental concentration of standard reference materials (SRMs) was determined. These SRMs were manufactured by the International Atomic Energy Agency and are constituted of clay soil (PTXRF-IAEA13), river clay sediment (PTXRF-IAEA09), plastic clay (IPT32) and São Simão clay (IPT42). All these SRMs were measured ten times in the pXRF equipment. The average values, propagated errors and recoveries against certified values are shown in the tables below.

Table 3 - Mean ± propagated error (n=10) and recovery between certified concentrations and quantified by pXRF of SRM IPT32

Element	Mean ± Propagated Error	Recovery
	g/kg	
Mg	7±6	301%
Al	174±10	116%
Si	225±10	94%
P	0.31±0.11	54%
K	5.4±0.2	81%
Ca	1.24±0.08	102%
Ti	10±0.3	112%
Fe	25.7±0.6	106%

Table 4 - Mean ± propagated error (n=10) and recovery between certified concentrations and quantified by pXRF of SRM PTXRF-IAEA09

Element	Mean ±Propagated Error	Recovery
	g/kg	
Mg	8±5	93%
Al	69±6	115%
Si	288±14	89%
P	0.87±0.14	79%
S	0.33±0.11	102%
K	17.7±0.7	91%
Ca	9.1±0.3	66%
Ti	4.79±0.19	111%
Fe	33.1±0.8	112%
	mg/kg	
Cr	77±36	71%
Mn	710±128	86%
Co	8±5	60%
Ni	32±5	85%
Cu	12±3	61%
Zn	70±6	73%
Ga	11±5	82%
As	12±3	93%
Rb	107±5	100%
Sr	108±5	102%
Y	33±4	104%
Zr	369±16	122%
Nb	17±5	108%
Ba	401±80	98%

La	41±34	113%
Pb	8.2±1.8	22%
Th	10±7	92%

Table 5 - Mean ± propagated error (n=10) and recovery between certified concentrations and quantified by pXRF of SRM IPT42

Element	Mean ± Propagated Error	Recovery
	g/kg	
Mg	6±6	506%
Al	204±9	120%
Si	232±7	97%
P	0.42±0.11	136%
K	4.01±0.15	103%
Ca	0.51±0.05	142%
Ti	6.08±0.14	106%
Fe	7.93±0.19	104%

Table 6 - Mean ± propagated error (n=10) and recovery between certified concentrations and quantified by pXRF of SRM PTXRF-IAEA13

Element	Mean ± Propagated Error	Recovery
	g/kg	
Mg	8±6	71%
Al	87±7	87%
Si	236±11	90%
P	0.36±0.11	295%
S	0.14±0.08	165%
K	1.34±0.09	114%
Ca	10.9±0.3	71%
Fe	60±1	111%
	mg/kg	
Cr	34±30	224%
Mn	1021±60	69%
Co	14±8	79%
Ni	10±4	109%
Cu	26±4	48%
Zn	42±5	64%
Ga	16±5	65%
Rb	13±2	84%
Sr	151±9	101%
Y	27±3	86%
Zr	122±8	90%

The average recovery among the elements present in the SRMs, in %, were: Mg (243), Al (109), Si (93), P (141), S (133), K (97), Ca (95), Ti (110), Cr (155), Mn (80), Fe (109), Co (70), Ni (97), Cu (55), Zn (81), Ga (74), Rb (92), Sr (101), Y (95) and Zr (106). The elements with concentration recoveries between 80 and 120% were used in the modeling with concentration results.

2.2.3 Multivariate Analysis Methods

As an exploratory analysis, the PCA with the fertility attributes by conventional methods and pXRF spectral data was performed to evaluate similar and distinct characteristics among the responses generated by the same sample set under different experimental measurements. In terms of the different variables obtained in the sample set studied, different clusters may be generated. They reveal non-explicit patterns of associations among the samples. This information is useful for PLS modeling because it may be evidence of the optimal conditions sought.

In the PCA with fertility attribute data, the different orders of variables magnitude were adjusted using autoscaling as pre-processing. In this pre-processing, the sample matrix is mean centered (each value is subtracted by the mean calculated for each variable) and divided by the standard deviation calculated for each variable. Therefore, as the standard deviations are proportional to the magnitudes of the data for each variable, they are normalized to a common scale, however, maintaining the variance patterns between the samples. With a similar objective, in the spectral data PCA, the pre-processing Poisson scaling combined with Mean Center (Poisson + MC) was used as preprocessing. Analogous to autoscaling, Poisson scaling divides the sample spectral matrix by the square root of the mean calculated for each variable.

In the multivariate modeling, six calibration strategies have been performed: (i) using 15kV condition; (ii) using 30kVcondition; (iii) using 50kV condition; (iv) using elemental concentration data reported by GeoExploration mode (CNC); (v) concatenation of 15 kV, 30 kV, and 50 kV spectra (CNT); and (vi) using scores of the P-ComDim analysis (PCD). The methodological steps followed in this study are shown in Figure 10.

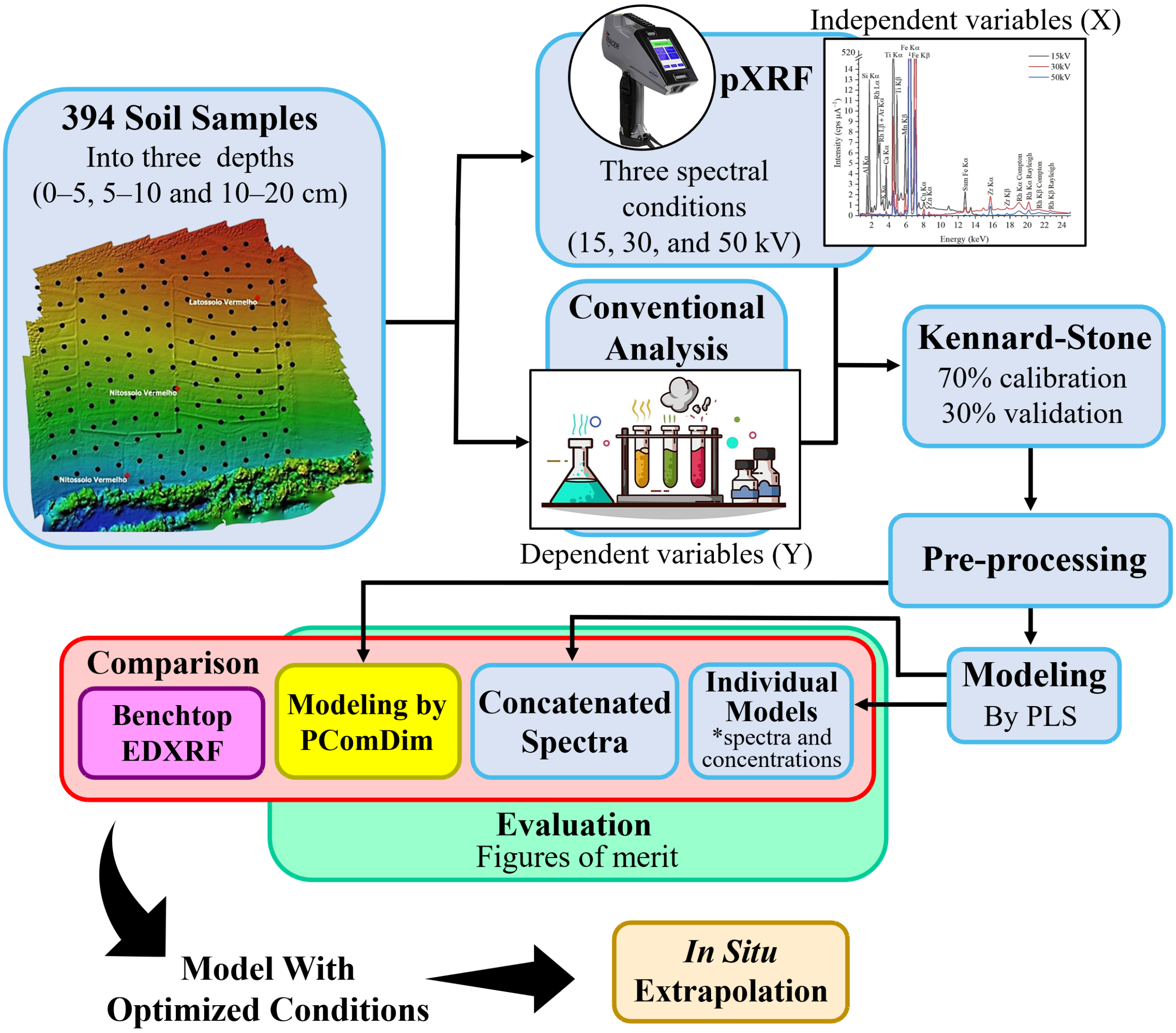

Figure 10 - Methodology schematic design.

The complete dataset was divided, with the Kennard-Stone algorithm, into a calibration (70% of samples) and prediction (30% of samples) sets. This algorithm was applied to the spectral data matrix obtained at 15kV condition. As a result, 276 samples were selected for the calibration set while 118 were selected for the prediction set.

The Kennard-Stone algorithm is used as a method for selecting a representative sample subset from a larger set. It was developed in 1969 by Kennard and Stone and is widely used in analytical chemistry and other areas involving multivariate analysis. This method uses distances for each samples pair to select the samples that will belong to the subset [40], [41]. The algorithm works according to the steps:

i. A distance matrix among all samples of the original data set is calculated. This distance matrix may be built using different metrics, the Euclidean distance is the most used.
ii. Two initial samples are randomly chosen.
iii. In each subsequent iteration, the distances between the remaining samples and the selected set are calculated. The sample that has the greatest distance from the already selected set is chosen.
iv. The sample selected in step iii) is included in the subset.
v. Steps 3 and 4 are repeated until the desired number of samples has been selected.

2.2.3.1 Individual Models

PLS was used to build the prediction models with individual spectra data and concentration data. To properly select the number of latent variables (LV) included in the PLS models, cross-validation with contiguous blocks (using sets of 10 in 10 samples) was used. The number of LV was chosen according to the minimum root mean square error of cross-validation (RMSECV). The identification and removal of outliers in all models was performed in agreement to the criteria of data with extreme leverage and unmodeled residuals. This method is explained in detail in [42] reference. All PLS modeling were performed in MATLAB R2022a (The MathWorks Inc., Natick, USA) and RStudio software's.

Two parts of all spectra were cut. The first corresponds to a region with a signal added by the equipment manufacturers while the second corresponds to a background region without characteristic peaks. As a result, 0.98 to 15.00 keV, 0.98 to 25.00 keV and 0.98 to 25.00 keV were the energy ranges used for the 15, 30 and 50kV instrument conditions, respectively. Thus, the 15kV condition had 701 variables while the 30 and 50kV conditions had 1201.

The 15kV, 30kV and 50kV models were pre-processed using Poisson + MC. It was chosen after evaluation the optimization results reported by Dos Santos *et al.* [26]. Poisson pre-processing weight the data according to its particular uncertainty [43] and is justified because in the XRF data, the statistical uncertainty in each variable is governed by Poisson statistics [32], where the absolute uncertainty at any given data point is larger as the number of counts represented by that point increases [44]. Thus, all the dataset variables are equalized, reducing the impact of variables that present great magnitude and mask information in the other variables with small magnitude.

In the CNC model, the elements with concentration recoveries between 80 and 120% were used in the PLS models. The elements considered were Al, Si, K, Ca, Ti, Fe, Ni, Rb, Sr, Y and Zr. To normalize the different orders of concentration data, autoscaling was the pre-processing applied.

2.2.3.2 Fusion Models

Two data fusion approaches were used to build the fusion models. In the low-level approach, spectra were concatenated and PLS was performed (CNT). Before being concatenated and pre-processed, each spectrum was normalized by the Fe peak intensity to minimize differences among the characteristic peaks generated by the three configurations in the X-ray tube. The concatenation order was performed with 15 kV, 30 kV and 50 kV. Then, Poisson + MC pre-processing was applied.

The mid-level fusion was performed with the P-ComDim approach (PCD). The PCD was applied in the 15 kV, 30 KV and 50 kV XRF spectra pre-processed with Poisson + MC to verify the existence of common information shared by them. Moreover, it was also evaluated which spectral configuration most contributes to the prediction each fertility attribute. The number of common dimensions (CDs) was chosen according to the total accumulated variance and the best R^2 and root mean square error of prediction (RMSEP) values. The Graphical User Interface for P-ComDim analyzes used in this study was developed and coded for Matlab software by Mishra *et al*. [45] and may be downloaded at https://github.com/puneetmishra2/Multi-block.git.Through the results generated by the data fusion approaches, the veracity of the second hypothesis may be attested.

2.2.3.3 Comparison Parameters

The model's performance was evaluated and compared by root mean square error (RMSE), square correlation coefficients (R^2), and the ratio of performance to deviation (RPD):

(17
)
(18
)
(19
)

where represents the *i*-th soil fertility attributes determined by conventional analyses, the fertility attribute determined by the model, the mean of the soil fertility attributes determined by conventional analyses and RMSEP the prediction RMSE.

As reported by Viscarra Rossel *et al.* [46]: (1) excellent models (RPD > 2.5), (2) very good models (2.5 > RPD > 2.0), (3) good model (2.0 > RPD > 1.8), fair (1.8 > RPD > 1.4), and very poor model (RPD <1.4).

The equivalence between the models was performed in two ways. At the first, the relative performance of the models was evaluated in percentage terms using the relative improvement (RI), calculated according to equation 20,

(20
)

where RMSEP$_1$ is the RMSEP of the model taken as reference and RMSEP$_2$ is the value to be compared.

At the second, the statistical significance of the difference between the models' accuracy was evaluated by a randomization test with 0.05 significance [47], [48]. This test compares the RMSEP values evaluating the null hypotheses in which the predictive accuracy is equal (RMSEP$_A$= RMSEP$_B$) and the alternative hypothesis in which the predictive accuracy is not equal (RMSEP$_A$ $\neq$ RMSEP$_B$). This is performed according to five steps: (i) Calculate the difference between the model's prediction squared residuals; (ii) Calculate the mean squared difference of the model residuals; (iii) Randomly multiplicate the residuals difference by +1 or -1; (iv) Calculates the mean squared difference of the residuals after (iii); And (v) repeat steps (iii) and (iv) k-times. The steps (iii) and (iv) were repeated 1999 times and the p-value was compared with the critical p-value at 95% confidence (0.05).

The experimental condition that produced the best individual models was identified according to performance comparison parameters. After that, the equivalence between models in the optimized experimental condition and fusion models was evaluated.

Moreover, the optimized condition models were also compared in terms of performance and equivalence with models built with spectral data from benchtop EDXRF, reported by Dos Santos *et al.* [18], [26]. The benchtop EDXRF measurements were carried out in the Shimadzu (EDX720 model) equipment, which has a 50 W X-ray tube with Rh target. The comparison is possible because these benchtop PLS models were built with the samples of the same area and class of soil. Pareto scaling pre-processing and 15 kV in samples irradiation was used. The compared SOC model was built with iSPA variables selection [26] while the full spectrum was used for the other soil attributes [18]. These were the models with the highest performance reported.

2.2.3.4 Figures of Merit

To attest the performance and quality of the optimized model, the following figures of merit were calculated: accuracy, systematic errors (bias), linearity, sensitivity (SEN), inverse of analytical sensitivity (γ^{-1}), limit of detection (LOD) and limit of quantification (LOQ). The figures of merit definitions used follow the recommendations of the IUPAC (International Union of Pure and Applied Chemistry) and the standard E1655–0017 of the ASTM (American Society for Testing and Materials) [49], [50]. The equations used are found in the references [42], [47], [51].

The accuracy is the agreement degree between the mean of a repeated measurements number and a reference value [52], [53]. It was evaluated by RMSEP, RPD because they are estimated from external validation samples. In addition to these parameters, the relative error in the predicted values in percentage, RE(%), eq. 21, was also used,

(21)

where is the number of prediction samples.

IUPAC defines bias as the non-random systematic errors that are calculated by the difference between the population mean and the true value [50]. The presence of significant bias in the models was evaluated by applying a t-test for the validation samples at the 95% confidence level, as equations 22, 23 and 24, as recommended by ASTM [49].

(22)

(23)

(24)

If the t-calculated by eq. 24 is greater than the t-critic at levels of freedom with a 95% significance level, there is an indication that the systematic errors present in the model are significant.

The linearity evaluation in multivariate calibration models cannot be performed by analysis of variance (ANOVA) as in univariate models. Alternatively, this is done by checking the non-existence of trends in the calibration model residuals. Qualitatively, visual inspection of the residual's scatterplot is an alternative. However, this is a subjective way and may lead to errors. Quantitatively, this may be performed using the Durbin-Watson (DW) test, according to Ref. [47], [51]. This test evaluates the autocorrelations between successive residuals in the augmented partial residuals plot (APARP) establishing the hypotheses: there is no correlation between them (null hypothesis) and there is correlation between them (alternative hypothesis). In this case, the probability associated with the DW statistics (eq. 25) was estimated by p-value. The p-values greater than 0.05 indicates the absence of non-linearities at the 95% confidence level.

(25
)

is the ith residual and is the number of calibration samples.

The sensitivity *(SEN)* determines the fraction of the signal responsible for the addition of a concentration unit of the property of interest [54]. In multivariate calibration models, one of the ways to quantify *SEN* is through the regression vector *(**b**)*, according to eq. 26,

(26
)

The represents the Euclidean norm [54].

Another parameter related to the model sensitivity is the inverse of the analytical sensitivity (γ^{-1}). This quantity provides an estimate of the minimum difference that is discernible by the model considering random experimental noise as the only error source. Thus, this figure of merit is directly related to the method resolution and, consequently, the number of decimal places to be used to express the forecast values [54], [55].

To provide sufficient conditions for the reliable calculation of the experimental noise *(δ_X)*, a background region of the XRF spectrum without the existence of characteristic peaks may be used [17], [56]. In this case, δ_X was quantified by the standard deviation of the region corresponding to the 9-10.5 keV range for the 15 kV spectrum of the calibration samples set. In its value possession, the γ^{-1} was calculated according to eq. 27,

(27
)

The LOD is the lowest concentration of the interest substance that may be detected, but not necessarily quantified. On the other hand, the LOQ represents the lowest concentration of the interest substance measured with a maximum uncertainty of 10% [50]. For PLS models, LOD and LOQ are determined using equations 28 and 29, respectively.

(28
)

(29
)

2.2.4 In situ measurements

In the preliminary study on extrapolation of models with optimized conditions for *in situ* predictions, 18 samples from the same study area were measured *in situ*. They were selected to cover the entire area. The soil was drilled between 0 and 10 cm with a hand auger and minimally flattened with a trowel. Then, Tracer 5i in the 15kV condition of Geoexploration mode adjusted for 4 s were used for each spectral reading, based on the results reported by Tavares *et al.* [57]. Measurements were made in triplicate and mean spectra were calculated. Figure 11 shows some photographs recorded during the measurement process.

In addition, approximately 30 g of soil was collected from all sampling points for new spectral readings in the laboratory and another 300g was sent for conventional analysis of fertility attributes in IDR Soil Analysis Laboratory, Londrina, Paraná, Brazil.

In the laboratory, the samples were dried at 40 ºC in oven and the moisture content was determined. Subsequently, they were prepared in the same way as the original set of 394 samples, *i.e.*, macerated and sieved to a particle size < 2 mm. Spectral measurements were taken in triplicate at the end of each step using 4 and 30 s in pXRF spectrometer. The mean spectra were calculated. This resulted in five sample sets: 18 samples measured *in situ* with 4 s (InSitu), 18 dry samples measured in the laboratory with 4 s (Dry4) and 30 s (Dry30), 18 dry samples with particle size < 2 mm measured in the laboratory with 4 s (Dry2mm4) and 30 s (Dry2mm30). To simplify understanding, these five sets will be called in-situ-sets.

The spectra of in-situ-sets were normalized by the tube current and detector live time, pre-processed with Poisson + MC and, finally, used as extra validation of SOC and CEC laboratory-calibrated models with the optimized conditions. These results allow for the assessment of the generalization power of models calibrated in the laboratory to field conditions, where sample preparation is not controlled, and measurement time is scarce.

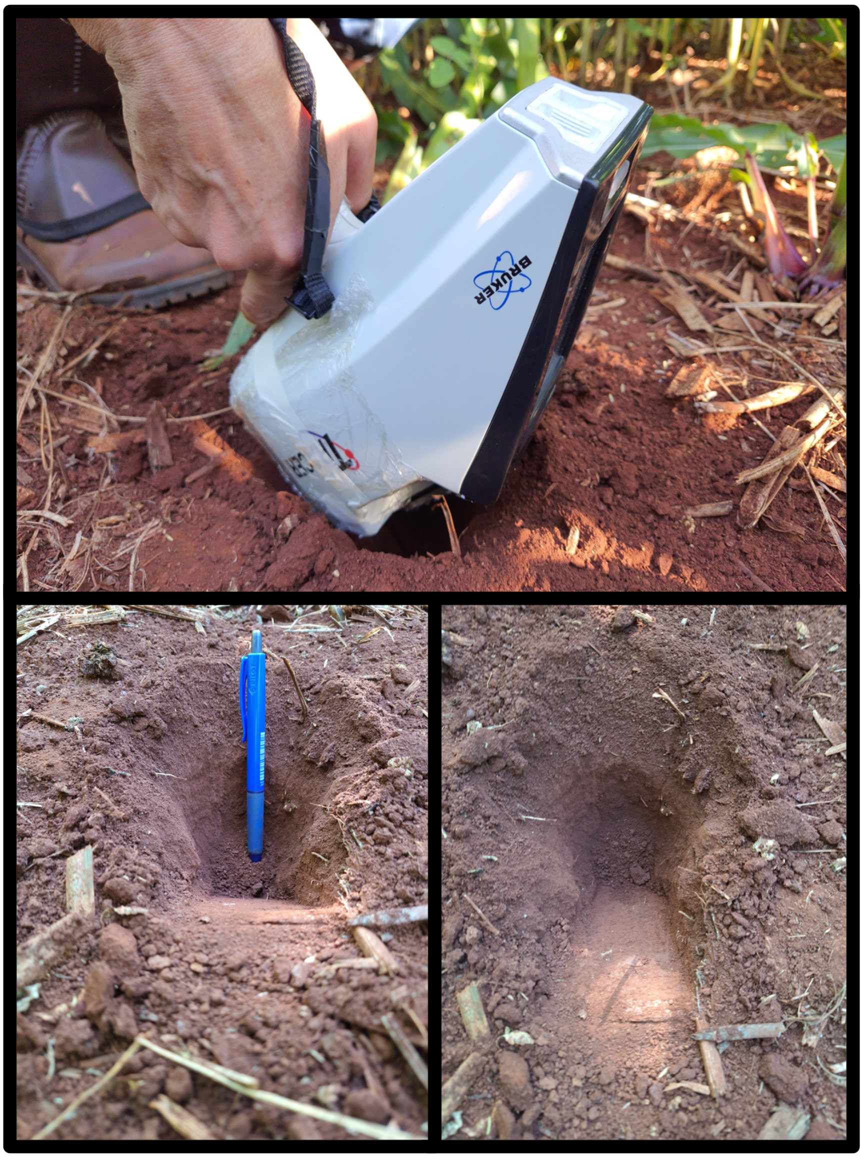

Figure 11 - Photos recorded during the *in situ* measurement process

It is important to mention that due to the initial conditions and ideas of this research, the 18 samples used for the preliminary study on the possibility of *in situ* modeling were collected after several harvests and inter-harvests in the study region. During this period, the area was drastically altered by fertilization, plowing, agricultural machinery traffic, and other agronomic activities.

2.3 Results and Discussion

2.3.1 Conventional Analysis

Soil fertility attributes descriptive statistics quantified by conventional analyzes are presented in Table 7. The results demonstrate that the Kennard-Stone algorithm separated the samples from the calibration and prediction set preserving the data representativeness. The

K^+ and P_M values show high variability (CV > 60%) while pH and CEC are low (CV < 12%). The CV values for the other attributes are of medium variability in the study area (12% < CV < 60%) [58]. This range in the attributes is naturally caused by the fertility gradient on the Ferralsoil profile, commonly observed in areas under no-tillage management where the highest fertility is observed on the surface layers, decreasing with increase of the depth [59]. Conversely, the inputs and fertilizers used may also alter soil fertility attributes. Thus, the high CV values for P_M and K^+ may be also related to the in-line application of fertilizers, as well as the adsorption and mobility of these elements [60]. Furthermore, the values for fertility attributes quantified in Ferrasols managed with a no-tillage system by Briedis [61] are similar to those presented.

Table 7 - Descriptive statistics of soil attributes determined by conventional methods. Ca, K, Mg and P concentration values represent the bio-disponible fraction.

Attribute	Set of Samples	Mean ± SD	CV (%)	Median	Minimum	Maximum
SOC (g kg^{-1})	All	19.8±4.3	21.8	20.0	7.6	30.5
	Cal	19.8±4.5	22.5	20.4	7.6	30.5
	Pred	19.9±4.0	20.3	19.9	9.7	29.3
CEC (cmol$_c$ kg^{-1})	All	13.7±1.5	11.3	14.0	9.3	17.9
	Cal	13.8±1.6	11.4	14.0	9.3	17.9
	Pred	13.5±1.4	10.7	13.5	9.8	17.2
SB (cmol$_c$ kg^{-1})	All	8.3±1.9	22.7	8.0	2.1	12.6
	Cal	8.4±1.9	22.2	8.7	2.2	12.6
	Pred	7.9±1.9	23.6	8.2	2.1	11.5
Ca^{2+} (cmol$_c$ kg^{-1})	All	6.0±1.5	25.2	6.0	1.4	9.2

	Cal	6.1±1.5	24.5	6.4	1.4	9.2
	Pred	5.7±1.5	26.4	5.7	1.4	8.9
K$^+$ (cmol$_c$ kg^{-1})	All	0.24±0.16	66.78	0.00	0.05	1.15
	Cal	0.3±0.2	64.6	0.2	0.05	1.2
	Pred	0.20±0.14	70.34	0.16	0.05	0.7
Mg^{2+} (cmol$_c$ kg^{-1})	All	2.0±0.6	28.5	2.0	0.7	3.7
	Cal	2.0±0.6	29.0	2.0	0.7	3.7
	Pred	2.0±0.6	27.2	2.0	0.7	3.5
P$_M$ (mg kg^{-1})	All	14.0±10.0	71.6	12.0	0.4	69.8
	Cal	13.4±8.9	66.7	12.3	0.4	60.3
	Pred	15.3±12.1	78.8	12.0	1.3	69.8
pH	All	5.3±0.4	7.7	5.0	4.2	6.5
	Cal	5.3±0.4	7.6	5.2	4.2	6.5
	Pred	5.2±0.4	8.0	5.2	4.2	6.5
BSP (%)	All	60±10	16.1	61.0	18.0	80.0
	Cal	60.2±9.3	15.4	61.5	19.9	79.7
	Pred	58.1±10.2	17.6	60.4	17.7	77.2
H+Al (cmol$_c$ kg^{-1})	All	5.5±1.1	19.7	5.0	2.7	9.7
	Cal	5.4±1.1	19.7	5.3	2.7	9.0
	Pred	5.6±1.1	19.9	5.6	2.9	9.7

SD = Standard deviation; CV = coefficient of variation; All = complete dataset with 394 samples; Cal = calibration set; Pred = prediction set.

2.3.2 pXRF Results

The average pXRF spectrum from all soil samples is shown in Figure 12. It is noted that the spectral response is different according to XRF instrumental setup. When 15 kV is applied to the tube, the X-rays energy generated is not enough to adequately excite elements heavier than Fe. As a result, this energy range is more efficient to excite lighter elements with Z <26 [32]. Only the Rh scattering L-lines for X-ray tube are verified, however, they overlap the Kα-line of Ar (2.958 keV), caused by the fluorescence of this element, which is present in the ambient air between the sample and the detector. Beyond that, the 15kV condition shows the highest background because none filter was used.

On the other hand, at 30kV and 50kV conditions, the fluorescence of the elements heavier than Ti is favored and the Rh scattering K-lines peaks from the tube are evident. The Rh peaks, mainly Rh from Compton scattering, are caused by interactions with light elements (such as H, C, N, O), which result in high scattering and low absorption of incident radiation

[16], [62]. Furthermore, the background is lower under these two conditions due to the use of filters.

Based on the different responses of the conditions analyzed in the spectrum, it was possible to identify the Al, Si, P, S, K, Ca, Ti, Mn, Fe, Cu, Zn, Rb, Y and Zr presence, with the Fe K-lines presented higher values. Consequently, the spectra at 15kV and 30kV conditions showed the regions between 10 and 15 kV affected by Fe sum peaks. The 50kV condition did not show the Fe peak sum due to the Cu filter use. The high intensity of the Fe peak is related to the fact that this element is the main forming component of the soil under study and has a high content [63], to the favorable fluorescence yield for the photoelectric absorption of this element, and to the voltage used in the X-ray tube.

The descriptive statistics of the concentration results generated directly from Geoexploration routine are presented in Table 8. The concentrations of Fe, Al and Si were quantified in higher levels. This may be attributed to soil characteristics, including the soil profile (rich in iron and aluminum oxyhydroxides), high mineral fraction (the basis of the structure of most clay minerals is Si [64]), and weathering processes. Soils in an advanced state of weathering show intense loss of silica (composed mainly of SiO2), with a resulting accumulation of insoluble iron and aluminum oxides [65], the most frequent being goethite (FeOOH), hematite (Fe_20_3) and gibbsite (AlO_3H_3) [66]. In lower orders, the contents of Ti, Mn, Ca, and K were also quantified in larger quantities in relation to the remaining elements. In the analyzed soil, these may be related to several sources, such as the clay fraction mineralogy, soil texture, organic matter, fertilizers, to the matrix rock that formed the soil and resistance to weathering and leaching.

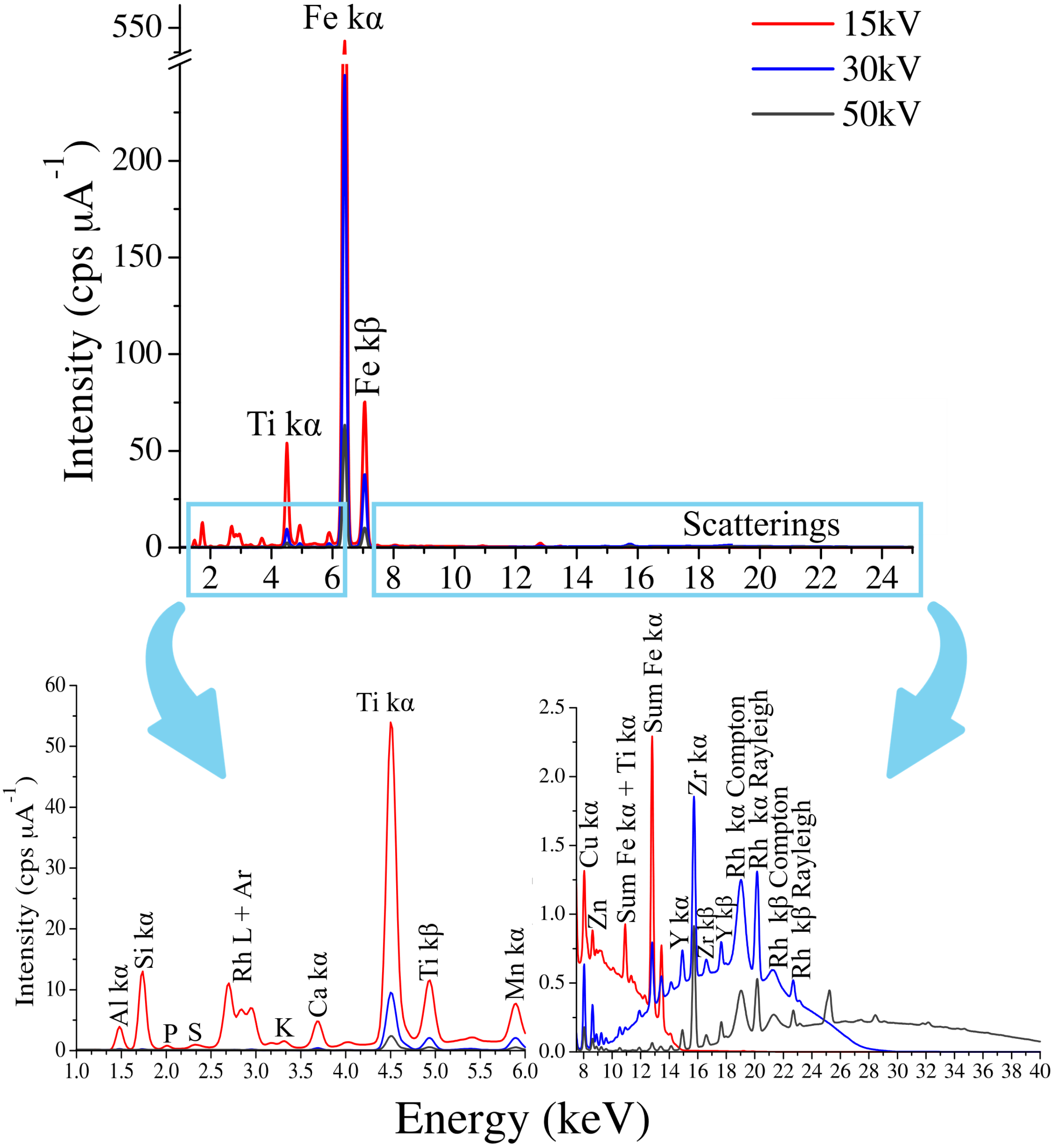

Figure 12 - Mean spectra of the 394 soil samples using three experimental conditions in the pXRF spectrometer.

Table 8 - Descriptive statistics of the total concentration per element determined by pXRF

Element	Mean ± SD	Med	CV	Max	Min	Q25	Q75	Kurt	Skew

				g/kg					
Al	100±8	100	8%	126	76	95	105	0.1	-0.1
Si	98±10	99	10%	127	67	91	104	0.04	-0.3
K	0.68±0.13	0.67	19%	1.23	0.39	0.59	0.78	0.08	0.44
Ca	1.7±0.6	1.6	35%	4.4	0.6	1.3	2.0	2.2	1.2
Ti	22±3	22	14%	33	15	20	24	1.0	0.5
Mn	1.6±0.3	1.5	19%	2.4	0.9	1.3	1.8	-0.5	0.5
Fe	188±13	190	7%	219	151	182	196	0.3	-0.6
				mg/kg					
Zn	94±11	93	12%	144	71	87	101	0.7	0.7
Rb	14±4	13	29%	24	8	11	18	-0.8	0.6
Sr	12±3	12	25%	35	6	10	14	5.8	1.5
Y	42±7	41	17%	68	23	36	47	0	0.2
Zr	226±13	226	6%	259	181	217	235	0.2	-0.3

SD = Standard deviation, Med = median, CV = coefficient of variation, Var = Variance, Max = Maximum, Min = Minimum, Q25 = First quartile, Q75 = Third quartile, Kurt = Kurtosis, Skew = Skewness

2.3.3 Exploratory Analysis

Figure 13, Figure 14, Figure 15, and Figure 16 present the PCAs with the samples set of fertility attributes and spectra under different experimental conditions terms. Regardless of the data matrix used, all PCAs demonstrated that Kennard-Stone separated the prediction and calibration samples set maintaining data representativeness in both sets. This result is a reflection of the clusters non-existence in the data set terms along all PCAs, Figure 13 (a), Figure 14 (a), Figure 15 (a), and Figure 16 (a).

Figure 13 shows the scores and loadings plots from PCA built with soil fertility attribute dataset ($\boldsymbol{X_{394x10}}$) determined by conventional methods. While PC1 explained 54.1% of the total data variance, PC2 explained 16.6%. The remaining PCs do not demonstrate any consistent separation and do not capture significant data variability values, so they were not included. According to the scores plot with soil class separations in Figure 13 (b), it was possible to observe a samples separation tendency. Samples classified as Red Latosol are located in the negative PC2 direction while those classified as Red Nitosol are in the PC2 positive direction. Through the loadings plot (Figure 13d), the attributes responsible for the samples separation tendency in relation to the soil class were P_M, Mg^{2+}, pH, SOC, and K^+ in the Red Latosol samples while the BSP, SB, H+Al, Ca^{2+}, and CEC favored the Red Nitosol samples separation. According to its positioning, the attribute with the highest correlation with the Red Latosol samples was the P_M which, on the other hand, presented smaller correlations with the Red Nitosol samples. It is important to emphasize that the Red Latosol is positioned in an area with lower slopes compared to the Red Nitosol, and the P_M in the Red Latosol samples was quantified by conventional analysis, on average, with higher values than to the Red Nitosol ones. Thus, a possible cause for the observed separation tendency may be related to the water flow in rainy events, which are greater in areas with steeper slopes, causing greater P_M losses in the Red Nitosol compared to the Red Latosol by soil transport [67]–[69]. In this context, the higher P_M values in the Red Latosol samples may indicate that this region acted as a deposit area for phosphorus and, possibly, for organic matter, reducing the adsorption capacity and increasing the P availability [69]. This hypothesis is supported by the SOC presence among the attributes correlated with the Red Latosol samples, since the SOC content is directly proportional to the soil organic matter content.

Conversely, the exchangeable Ca and CEC contents are the attributes with the highest correlation with the Red Nitosol samples. These attributes were found in higher values in these samples class and also contribute more significantly than the others to the sample separation in the score plot.

The score plot with class separations by depth of Figure 13(c) did not show any consistent cluster. This is evidence that the fertility attribute values have a uniform behavior in terms of the three soil depths in the analyzed area.

The scores and loadings plots from the PCA with the spectral data matrix of the 15kV ($\boldsymbol{X}_{394x701}$), 30kV ($\boldsymbol{X}_{394x1201}$) and 50kV ($\boldsymbol{X}_{394x1201}$) experimental conditions are shown in the Figure 14, Figure 15, and Figure 16, respectively. For 15kV, the PC1 captured 74.47% of the total data variability while PC2 contained 10.58%. For 30kV, 53.75% of the data variance was captured by PC1 and 18.27 by PC2. At 50kV, PC1 and PC2 captured 52.85 and 14.62% of the total data variability, respectively. The other PCs in each figure showed redundant information and therefore were not reported.

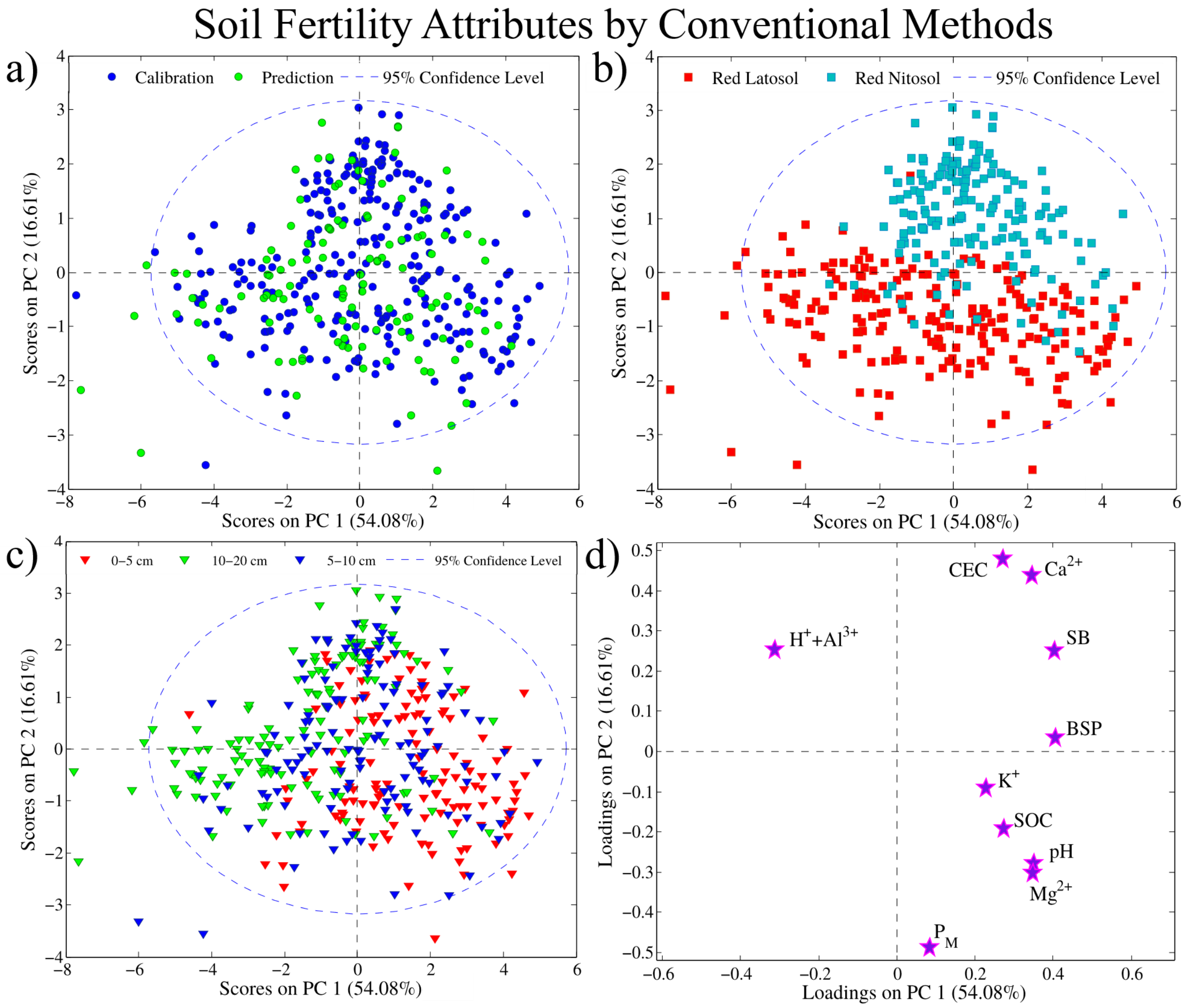

Figure 13 - Scores (a, b, and c) and loadings (d) plots from the PCA with soil fertility attributes determined by conventional methods.

Analyzing the PCAs scores plots with spectral data, there is no samples clustering by depth for all experimental conditions studied (Figure 14c, Figure 15c, and Figure 16c). However, it is noted the existence of a separation tendency among the samples due to the soil class in the 30kV and 50kV conditions, Figure 15 (b) and Figure 16 (b), respectively. In both plots, the Red Nitosol samples are positioned on the positive PC2 axis, while those from Red Latosol are on the negative PC2 axis. This is the same trend seen in the PCA with fertility attributes of Figure 13(c).

Through loading plots of Figure 15(d) and Figure 16 (d), it may be observed that, in spectral generated at 30 and 50kV condition terms, the variables that most contribute to the Red Nitosol samples differentiation compared to Red Latosol are those with greater absolute intensity in PC2. Therefore, it is the Ti K-lines, together with the Mn Kα peak and the scattering region (Compton and Rayleigh) when in the 30kV condition. In 50kV condition terms, the Ti K-line and the scattering region are the more intense variables.

Although the PC1 at 30kV and 50kV condition has the greatest variance, the data does not show separation trends along its axis. In the loading plots (Figure 15d and Figure 16d), the Fe K-lines are the variables with the greatest intensity. In this context, this may be evidence that samples of Red Latosol and Red Nitosol have similar Fe contents.

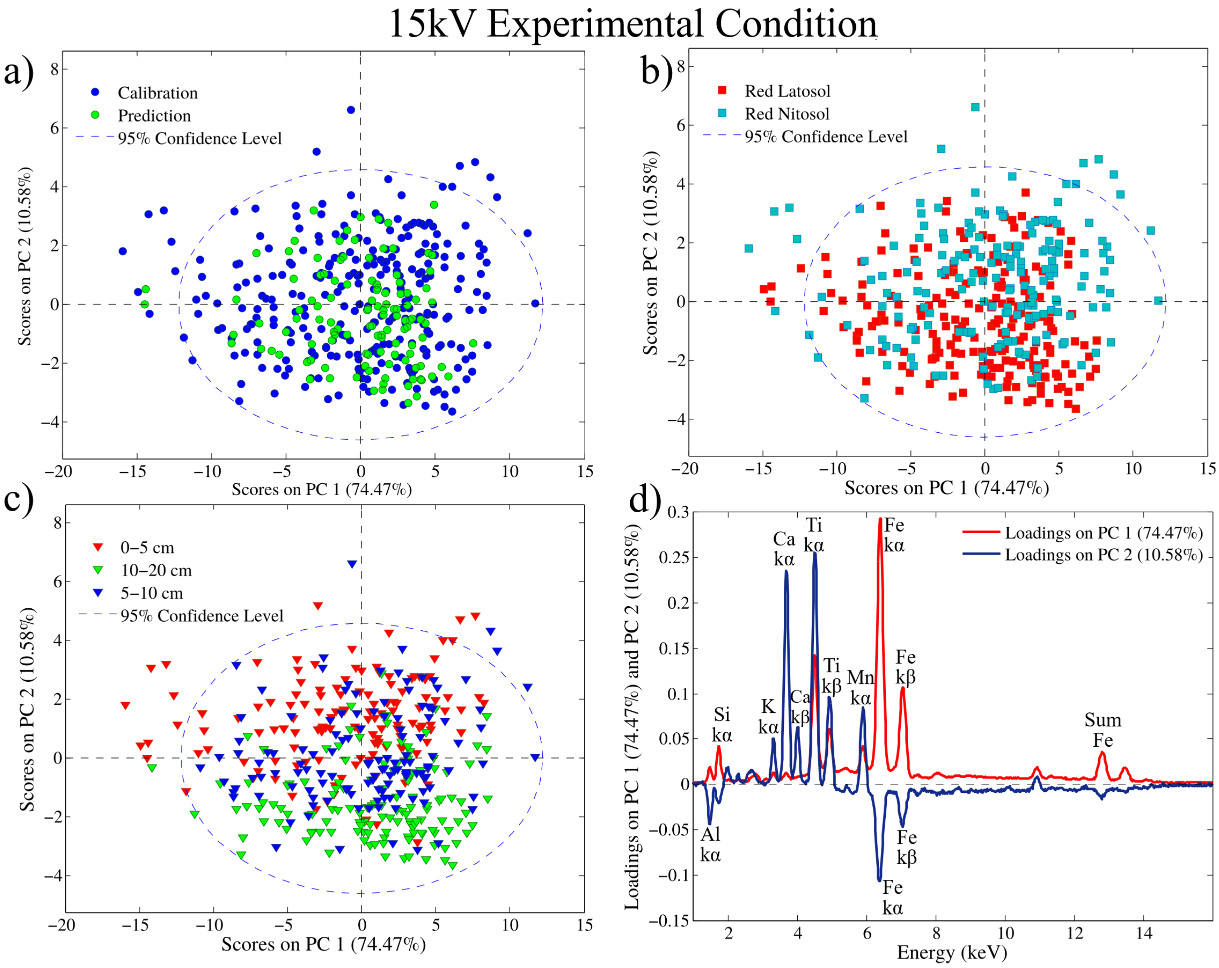

Figure 14 - Scores (a, b, and c) and loadings (d) plots from the PCA with spectra at 15kV experimental condition

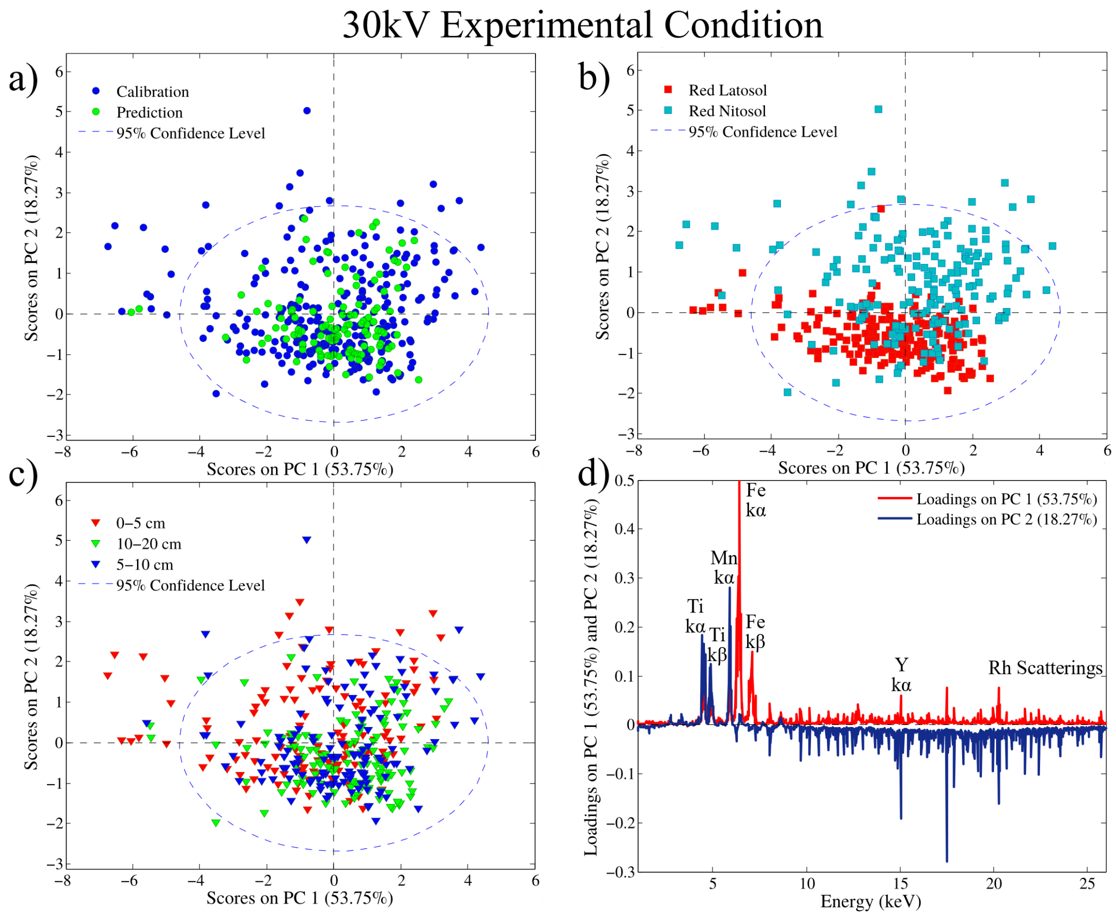

Figure 15 - Scores (a, b, and c) and loadings (d) plots from the PCA with spectra at 30kV experimental condition

In opposition to the behaviors found in the 30kV and 50kV conditions, the 15kV PCA scores (Figure 14b) do not show any significant clustering in soil class terms. The Red Latosol and Red Nitosol samples form a single cluster, indicating similarities when analyzed at 15kV condition. With 15 kV in the X-ray tube without filters, the fluorescence of lighter elements is favored [32]. This explains the presence of Al, Si, K, and Ca K-lines among the variables with greater intensity in the loadings plot (Figure 14d). In addition to these variables, as in the 30kV and 50kV conditions, the Ti, Mn, and Fe K-lines also have high values along PC1 and PC2 loadings.

In 30kV and 50kV conditions terms, the fluorescence of the elements heavier than Ti and the Rh scattering K-lines peaks from the tube are favored, but the elements lighter than Ti are not detected. As a result, the exploratory analysis provided evidence that the samples were separated by soil classification according to the absence of elements lighter than Ti, in the

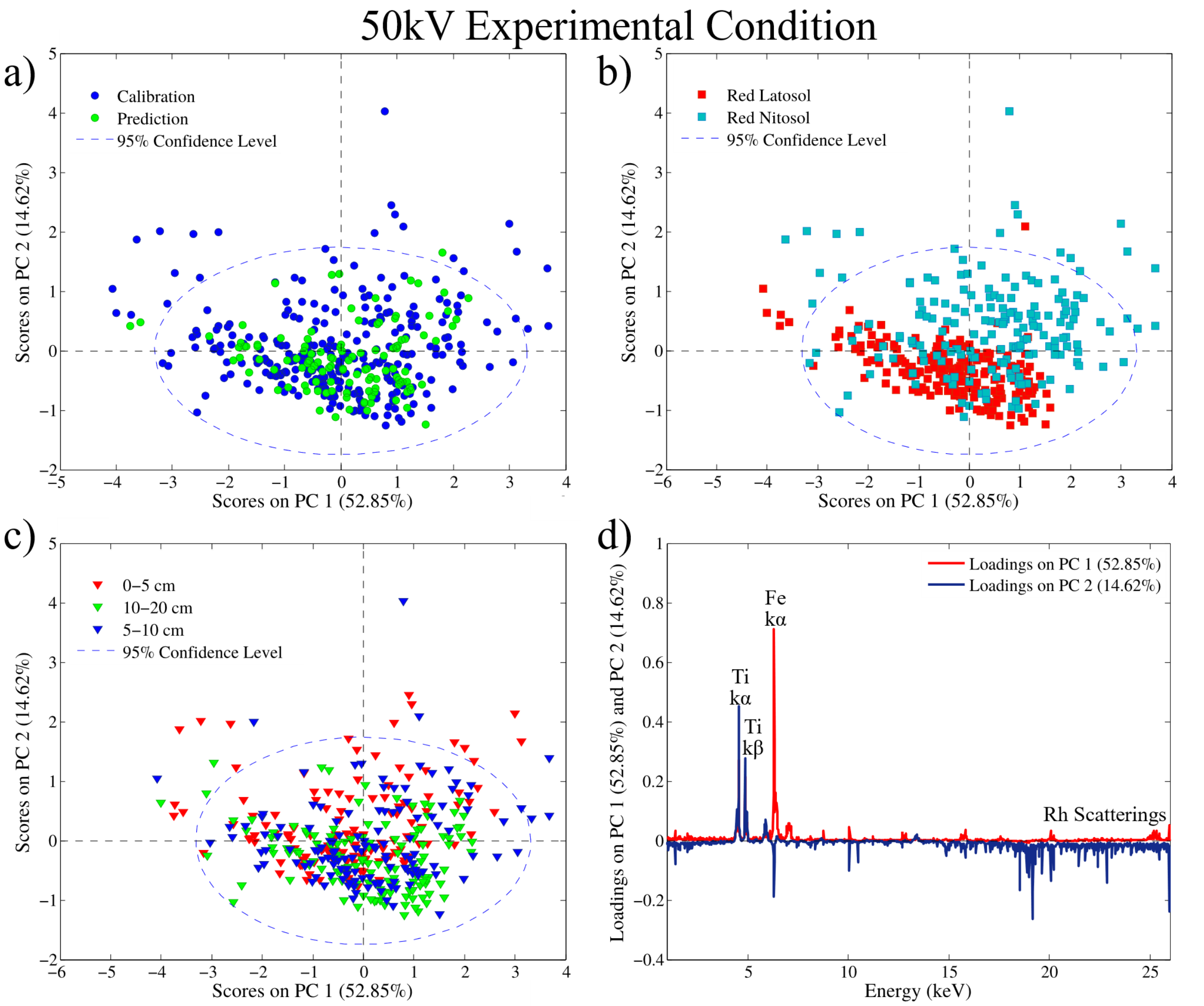

Figure 16 - Scores (a, b, and c) and loadings (d) plots from the PCA with spectra at 50kV experimental condition

spectral cases. The elements lighter than Ti identified in the sample set are Al, Si, P, K, and Ca. These elements are directly correlate with most fertility attributes that will be quantified by PLS. For instance, Al, P, K and Ca total contents directly influence the exchangeable forms Ca^{2+}, K^{+}, available P and potential acidity (H+Al). The CEC, SB and BSP, which are attributes obtained from them, are also influenced. Therefore, these findings are evidence that the 15kV condition may be more advantageous in PLS modeling with spectral data to simultaneously quantify the 10 proposed fertility attributes, because it may allow the fertility attributes in the Red Latosol and Red Nitosol samples to be quantified with similar performance.

2.3.4 Prediction Models Comparison

Figure 17 shows the RMSEP, RPD and pred R² results. It compares all the models for each fertility attribute. The prediction results were considered very good to poor depending on the fertility attribute (R² pred from 0.13 to 0.81 and RDP from 1.0 to 2.3). The models with highest pred R² and RDP values were SOC, CEC, SB, Ca^{2+}, K^{+} and Mg^{2+}. They also showed the lowest values for RMSEP.

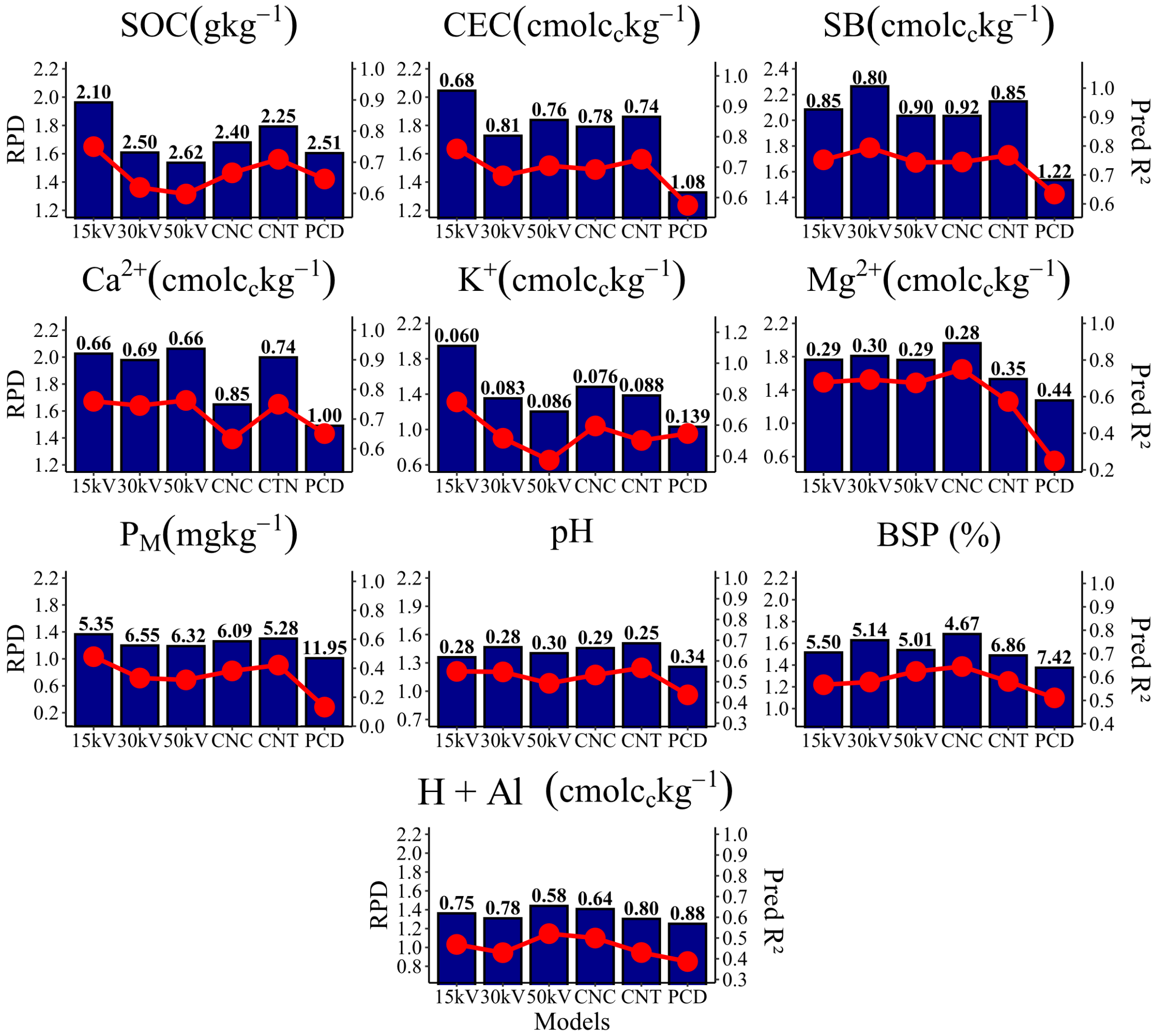

Figure 17 - RPD (blue bars and right vertical scale), prediction R² (red circles and left vertical scale) and RMSEP (bold values at top of bars) for all prediction models.

In addition to Figure 17, full details of models at individual, low- and mid-level fusion approaches are presented in Table 9, Table 10, and Table 11, respectively. In general, the most successful approaches have been with 15kV, 30kV and 50kV, individually. CNC models had intermediate performance, *i.e.*, better than data fusion, however, less than or equal to individual modeling. Data fusion produced the worst results, with the low-level approach performing better than the mid-level.

The 15kV were the best for predicting SOC, CEC, and K^+, presenting Pred R^2 (RMSEP) values of 0.75 (2.10), 0.76 (0.68) and, 0.75 (0.06), respectively. For SB, the best model was obtained at 30kV, with Pred R^2 (RMSEP) of 0.81 (0.80). For Ca^{2+}, the 50kV approach produced bests results, Pred R^2 (RMSEP) of 0.76 (0.66). In both situations (SB and Ca^{2+} best approaches), the 15kV spectra produced similar results. On the other hand, the CNC was best for Mg^{2+} modeling, Pred R^2 (RMSEP) of 0.75 (0.28).

According to the RPD intervals stablished by Viscarra Rossel *et al.* 2006 [46], the 15kV SOC, 15kV CEC, 30kV SB, 50kV Ca^{2+} and CNC Mg^{2+} models were considered very good for prediction (RPD greater or equal to 2.0) while the 15kV K^+ model presented RPD of 1.9, considered good for predictions (Table 9 and Figure 17). Quantitative analyses are reliable for these attributes. These findings support the first hypothesis, indicating that the pXRF spectra data modeled with PLS generate models capable of predicting soil fertility attributes with good performance and confidence.

In the remaining attributes, the 15kV strategy produced the best performances for P_M while CNT, CNC and 50kV was the best for pH, BSP and H+Al, respectively. These four spectral models showed lower pred R^2, RPD, and higher RMSEP results when compared to the previous. The RPD values ranged from 1.7 and 1.4, indicating fair predictions, therefore, the P_M, pH, BSP, and H+Al results may be used in the qualitative evaluation and correlation analysis. These soil attributes depend on many factors, including soil texture, mineralogy, organic matter, soil depth, and water availability. The magnitudes variation of these parameters in the soil is not accurately captured by the pXRF spectrum. In the pH and BSP cases, the lowest CV values in the conventional analysis (Table 7) attest to the low variability of these attributes captured by the samples, making their prediction even more difficult by the adopted approaches. Conversely, although total P signal is detected by XRF, the modeling of P_M is challenging for several reasons. One reason may be associated with the low photoelectric cross-section and P fluorescent yield, as well as its low concentration in the soil compared to other elements such as Si and Al. Another reason may be the interference of other elements in P quantification. Phosphorus emits characteristic X-rays at two main energies: 2.013 keV (Kα) and 2.139 keV (Kβ). More abundant elements in soil, such as Al and Si, may absorb these characteristic X-rays before reaching the detector, since they have absorption edges close to these energies (1.559 and 1.839 keV for Al and Si respectively). Additionally, the P emission K-lines may suffer interference from the Zr emission L-lines, and the Ca-K escape peak. Finally, the most critical reason is that P_M generally does not have a significant correlation with total phosphorus. Other studies with similar methodologies also have shown difficulty in predicting P_M, even with XRF and NIR data, and presented findings similar to those of this work [70],[10], [18], [26].

Table 9 - Detailed results of PLS individual modeling approach. Bold values indicate the best PLS individual modeling strategy

Models	Attribute	Samples	LV	RMSEC	R^2 Cal	RMSECV	R^2 CV	RMSEP	R^2 Pred	RPD
15kV	**SOC (g kg^{-1})**	**383**	**7**	**1.70**	**0.84**	**2.30**	**0.73**	**2.10**	**0.75**	**2.0**
30kV		380	5	1.93	0.80	2.83	0.59	2.50	0.62	1.6
50kV		383	5	1.99	0.79	2.83	0.58	2.62	0.60	1.5
CNC		383	8	2.25	0.75	2.45	0.70	2.40	0.67	1.7
15kV	**CEC (cmol$_c$ kg^{-1})**	**369**	**6**	**0.75**	**0.76**	**0.93**	**0.64**	**0.68**	**0.76**	**2.0**
30kV		377	4	0.81	0.72	0.96	0.62	0.81	0.67	1.7
50kV		377	4	0.75	0.76	0.98	0.60	0.76	0.70	1.8
CNC		372	7	0.82	0.71	0.95	0.62	0.78	0.69	1.8
15kV	**SB (cmol$_c$ kg^{-1})**	368	6	0.72	0.85	0.83	0.80	0.85	0.77	2.1
30kV		**368**	**6**	**0.60**	**0.89**	**0.95**	**0.74**	**0.80**	**0.81**	**2.3**
50kV		372	5	0.71	0.85	0.94	0.74	0.90	0.76	2.0
CNC		372	6	0.77	0.81	0.83	0.78	0.92	0.76	2.0
15kV	**Ca^{2+} (cmol$_c$ kg^{-1})**	368	7	0.5	0.88	0.65	0.80	0.66	0.76	2.0
30kV		371	7	0.42	0.92	0.84	0.66	0.69	0.75	2.0
50kV		**367**	**6**	**0.41**	**0.92**	**0.75**	**0.73**	**0.66**	**0.76**	**2.1**
CNC		373	5	0.78	0.71	0.83	0.68	0.85	0.63	1.6
15kV	**K^+ (cmol$_c$ kg^{-1})**	**368**	**8**	**0.05**	**0.86**	**0.09**	**0.58**	**0.06**	**0.75**	**1.9**
30kV		363	6	0.05	0.85	0.12	0.32	0.08	0.51	1.4
50kV		368	4	0.10	0.41	0.12	0.20	0.09	0.37	1.2
CNC		369	7	0.09	0.62	0.09	0.57	0.08	0.59	1.5
15kV	**Mg^{2+} (cmol$_c$ kg^{-1})**	366	4	0.31	0.69	0.34	0.65	0.29	0.68	1.8
30kV		372	4	0.32	0.69	0.36	0.62	0.30	0.69	1.8
50kV		369	6	0.20	0.69	0.37	0.61	0.29	0.68	1.8
CNC		**367**	**7**	**0.30**	**0.73**	**0.32**	**0.69**	**0.28**	**0.75**	**2.0**
15kV	**P_M (mg kg^{-1})**	**377**	**7**	**4.55**	**0.7**	**6.35**	**0.43**	**5.35**	**0.48**	**1.4**
30kV		374	4	5.79	0.44	6.75	0.27	6.55	0.33	1.2
50kV		375	4	5.55	0.48	7.01	0.21	6.32	0.32	1.2
CNC		371	7	5.43	0.50	5.84	0.42	6.09	0.38	1.3
15kV	**pH**	389	4	0.24	0.58	0.26	0.51	0.21	0.48	1.4
30kV		384	4	0.24	0.62	0.27	0.53	0.28	0.55	1.5
50kV		387	5	0.21	0.72	0.29	0.46	0.3	0.49	1.4
CNC		**374**	**6**	**0.18**	**0.75**	**0.2**	**0.7**	**0.29**	**0.53**	**1.5**
15kV	**BSP (%)**	373	6	4.23	0.76	4.90	0.68	5.50	0.57	1.5
30kV		370	4	4.86	0.65	5.48	0.56	5.14	0.58	1.6
50kV		370	5	3.83	0.80	5.18	0.63	5.01	0.62	1.5
CNC		**366**	**5**	**4.69**	**0.69**	**5.00**	**0.64**	**4.67**	**0.64**	**1.7**
15kV	**H+Al (cmol$_c$ kg^{-1})**	379	6	0.52	0.74	0.78	0.44	0.75	0.47	1.4
30kV		379	4	0.68	0.52	0.82	0.34	0.78	0.43	1.3
50kV		**367**	**5**	**0.51**	**0.74**	**0.75**	**0.44**	**0.58**	**0.52**	**1.4**
CNC		368	6	0.65	0.52	0.74	0.38	0.64	0.50	1.4

Cal = calibration; Pred=prediction; LV=latent variables; CV=cross-validation.

Regarding the CNC models, the better were SB, P_M, pH, BSP and H+Al. However, the concentration data models showed higher RMSEP and lower RPD and R^2 values in most soil attributes compared to the individual spectra modeling, except for Mg^{2+} and BSP (Table 9). The presence of different characteristics in the samples, such as different soils, and other organic and mineral materials, affects the total elemental concentration measurements, but the spectral data are more sensitive to these variations. This complete complex matrix representation existing in a soil sample provided by spectral data may be a possible reason for its better performance compared to concentration ones.

The best CNT models were for the SB, with Pred R^2 (RMSEP) of 0.79 (0.8), and the Ca^{2+}, with Pred R^2 (RMSEP) of 0.75 (0.7). The SOC and CEC CNT models presented Pred R^2 (RMSEP) of 0.71 (2.2) and 0.73 (0.7), respectively. According to RPD results, these were the only models classified as suitable for quantitative analysis. The remaining models showed RPD too many values were below 1.5, being suitable for qualitative and correlation analysis (Table 10). In conclusion, the use of concatenated spectra resulted in predictions similar or lower than models with individual spectra attested by the RPD values ranging from 1.3 to 2.1.

The PCD models saliences and regression coefficients values of each CD for each soil attribute are presented in Figure 18. According to these results, it is possible to verify that all spectral conditions have similar salience values in each common dimension (CD) for all soil attributes. This indicates that the common information generated by the different conditions is distributed across the CDs. In addition, the similar saliences demonstrate the existence of redundant information contained in the spectra obtained at 15, 30, and 50 kV. Consequently, there are no regression coefficients with values much higher than the others in each CD.

Table 10 - Detailed results of low-level fusion modeling. The RI and randomization tests p-values (at 95% confidence) result comparing the 15kV models for each soil attribute.

Attribute	Samples (LV)	RMSEC	R^2 Cal	RMSECV	R^2 CV	RMSEP	R^2 Pred	RPD	RI
SOC (g kg^{-1})	383 (6)	1.6	0.87	2.5	0.67	2.2	0.71	1.6	-7.0
CEC (cmol$_c$ kg^{-1})	383 (7)	0.5	0.89	1.0	0.61	0.7	0.73	1.3	-8.2
SB (cmol$_c$ kg^{-1})	374 (6)	0.6	0.91	0.9	0.79	0.8	0.79	1.5	0.4
Ca^{2+} (cmol$_c$ kg^{-1})	379 (6)	0.5	0.90	0.7	0.75	0.7	0.75	1.5	-12.7
K^+ (cmol$_c$ kg^{-1})	374 (6)	0.1	0.75	0.1	0.39	0.1	0.50	1.0	-47.0
Mg^{2+} (cmol$_c$ kg^{-1})	381 (4)	0.3	0.69	0.4	0.62	0.3	0.57	1.3	-19.1
P_M (mg kg^{-1})	378 (5)	5.4	0.57	7.3	0.25	5.3	0.42	1.3	1.2
pH	375 (4)	0.2	0.71	0.2	0.64	0.3	0.57	1.3	-16.2

BSP (%)	378 (4)	0.2	0.71	0.2	0.64	0.3	0.57	1.4	-24.7
H+Al (cmol$_c$ kg^{-1})	386 (4)	0.7	0.53	0.8	0.33	0.8	0.43	1.3	-6.7

p_{critic} = 0.05; LV= latent variables; Cal = calibration; Pred=prediction; CV=cross-validation.

Table 11 - Detailed results of mid-level fusion. RI are comparisons with the 15kV models

Attribute	**RMSEC**	**R^2 Cal**	**RMSEP**	**R^2 Pred**	**RPD**	**RI**
SOC (g kg^{-1})	1.6	0.87	2.5	0.65	1.6	-22.34
CEC (cmol$_c$ kg^{-1})	1.0	0.63	1.1	0.57	1.3	-59.47
SB (cmol$_c$ kg^{-1})	0.6	0.91	1.2	0.65	1.5	-42.89
Ca^{2+} (cmol$_c$ kg^{-1})	0.5	0.91	1.0	0.65	1.5	-52.89
K^{+} (cmol$_c$ kg^{-1})	0.1	0.36	0.1	0.25	1.0	-118.46
Mg^{2+} (cmol$_c$ kg^{-1})	0.2	0.87	0.4	0.55	1.3	-51.70
P$_M$ (mg kg^{-1})	4.3	0.76	11.9	0.13	1.0	-123.28
pH	0.2	0.76	0.3	0.44	1.3	-6.62
BSP (%)	3.1	0.89	7.4	0.51	1.4	-34.89
H+Al (cmol$_c$ kg^{-1})	0.6	0.72	0.9	0.39	1.3	-18.70

p_{critic} = 0.05; LV= latent variables; Cal = calibration; Pred=prediction; CV=cross-validation.

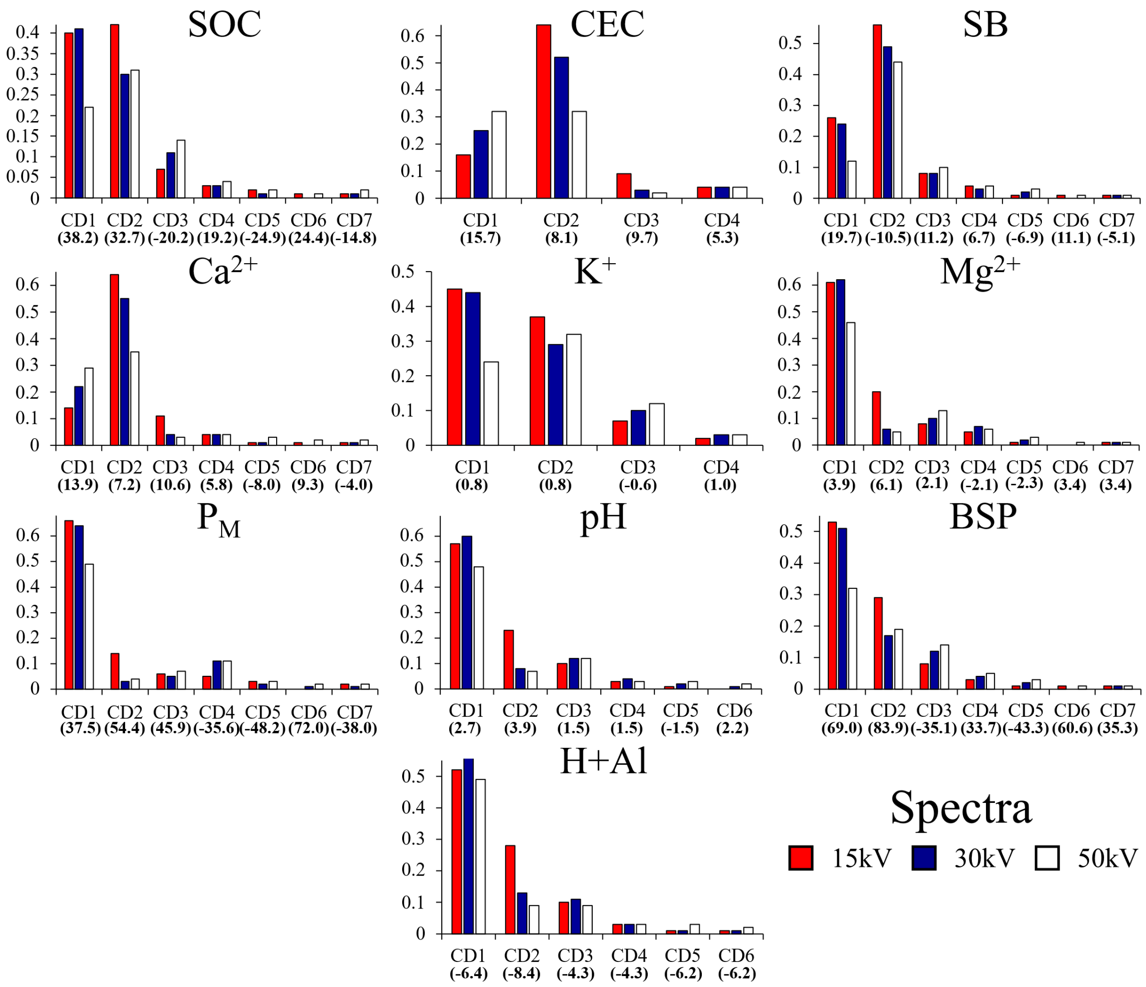

Figure 18 - Saliences (vertical bars) and regression coefficient values (below the CDs on the X axis) of the P-ComDim models

Reflecting the redundancy of spectra information under different experimental conditions, the PCD models are inferior to all other models, showing the largest RMSEP and the smallest R² (Figure 17). The RPD varied between 1.0 and 1.6 (Table 11), with better results for the SOC, SB, and Ca^{2+} models, classified as fair. The other RPD attributes were below 1.4, being considered very poor models.

Evaluating all the model's performance in Figure 17 it is verified that the 15 kV spectra produced the best models for a greater number of fertility attributes than the other experimental conditions. Therefore, 15kV condition may be a general optimized condition. To attest it, the equivalence between the 15kV model and all other models were performed with randomization tests. Table 12 presents the random test p-values.

Table 12 - Random Test p-values comparing 15kV models with all others

Attribute	15kV x 30kV	15kV x 50kV	15kV x CNC	15kV x CTN	15 kV x PCD

SOC	0.004	0.001	0.01	0.03	0.001
CEC	0.03	0.01	0.03	**0.2**	0.001
SB	**0.3**	**0.2**	**0.2**	**0.4**	0.001
Ca^{2+}	**0.2**	**0.4**	0.002	**0.1**	0.001
K^{+}	0.001	0.001	0.01	0.01	0.01
Mg^{2+}	**0.3**	**0.4**	**0.4**	0.02	0.001
P$_M$	**0.06**	**0.06**	**0.1**	**0.4**	0.001
pH	0.03	**0.2**	**0.9**	**0.9**	**0.07**
BSP	**0.2**	**0.2**	**0.07**	0.04	0.01
H+Al	0.01	**0.3**	**0.06**	**0.2**	0.04

* Bold indicates values greater than critical (0.05 for 95% confidence).

The randomization test at 95% of confidence level demonstrates that 15kV condition is considered equivalent to the best ones in the SB, Ca^{2+}, Mg^{2+}, pH, BSP and H+Almodels. For the remaining attributes, 15kV had already produced the best performance results. This may be justified because, unlike other spectral configurations, the use of 15 kV without filters in the X-ray tube is advantageous for soil analysis, as it favors the fluorescence of light elements [32], which are correlated with most of the analyzed soil fertility attributes. Although the spectra at 15kV compared to those at 30kV and 50kV lack the K-lines associated with Compton and Rayleigh Scatterings, which may carry information about the sample complex organic composition, this absence is compensated for by the presence of absorption peaks associated with Al, Si, P, K, and Ca (Figure 12). Therefore, these elements may be responsible for the superior performance of the individual 15kV models, as suggested by the results of the exploratory analysis.

It is important to note that the Rh scattering L-lines are present in the 15kV spectra, but, they overlap the Kα-line of Ar, hampering its usefulness in modeling. Thus, the 15kV models were considered the optimized condition to be used as a reference in the pXRF spectrometer to predict all the fertility attributes for comparisons in the next discussions.

The concatenated use of the spectra was only beneficial for the P$_M$ modeling compared to the individual 15kV model, with a 1.2% improvement. However, this was not reflected in a higher RPD value, not being enough to classify them as good and suitable for quantitative analysis [46]. In all other attributes the RI values were negative (Table 10). Similar results were found in previous studies using different spectroscopic sensors but with the same low-level approach [6].

The PCD models compared to the 15kV model produced only negative RI values (Table 11). These negative values indicate that there was no relative improvement with this multiblock modeling. Despite the unsatisfactory results of the regression with P-ComDim, the methodology allowed verifying the non-existence of complementarity among the spectral data in the different measurement conditions from the modeling fertility attributes point of view.

The worse results of the data fusion models compared to the individuals allow discarding the need to measure under three experimental conditions. Thus, the use of only one experimental condition, in this case 15 kV, 12.15 µA, 30 s, without filters in the X-ray tube, produced reliable prediction results for the soil fertility attributes evaluated. These findings are evidence that directly corroborated the second hypothesis. The optimized use of a single experimental condition may produce a shorter time interval for measurements, a considerably smaller amount of data for manipulation and less processing power to manage the models.

2.3.5 Figures of Merit

Once stablished the optimized reference models, the figures of merit for the 15kV models were calculated (Table 13). Through the values of |Bias| and t_{Bias} it was observed that only the K^+ model showed significant bias (t_{bias}>t_{critic}). The LOD and LOQ values were above the lowest values obtained by conventional analysis only for the K^+ and P_M models (Table 13). For the other fertility attributes, the entire range of values obtained by the reference method is valid.

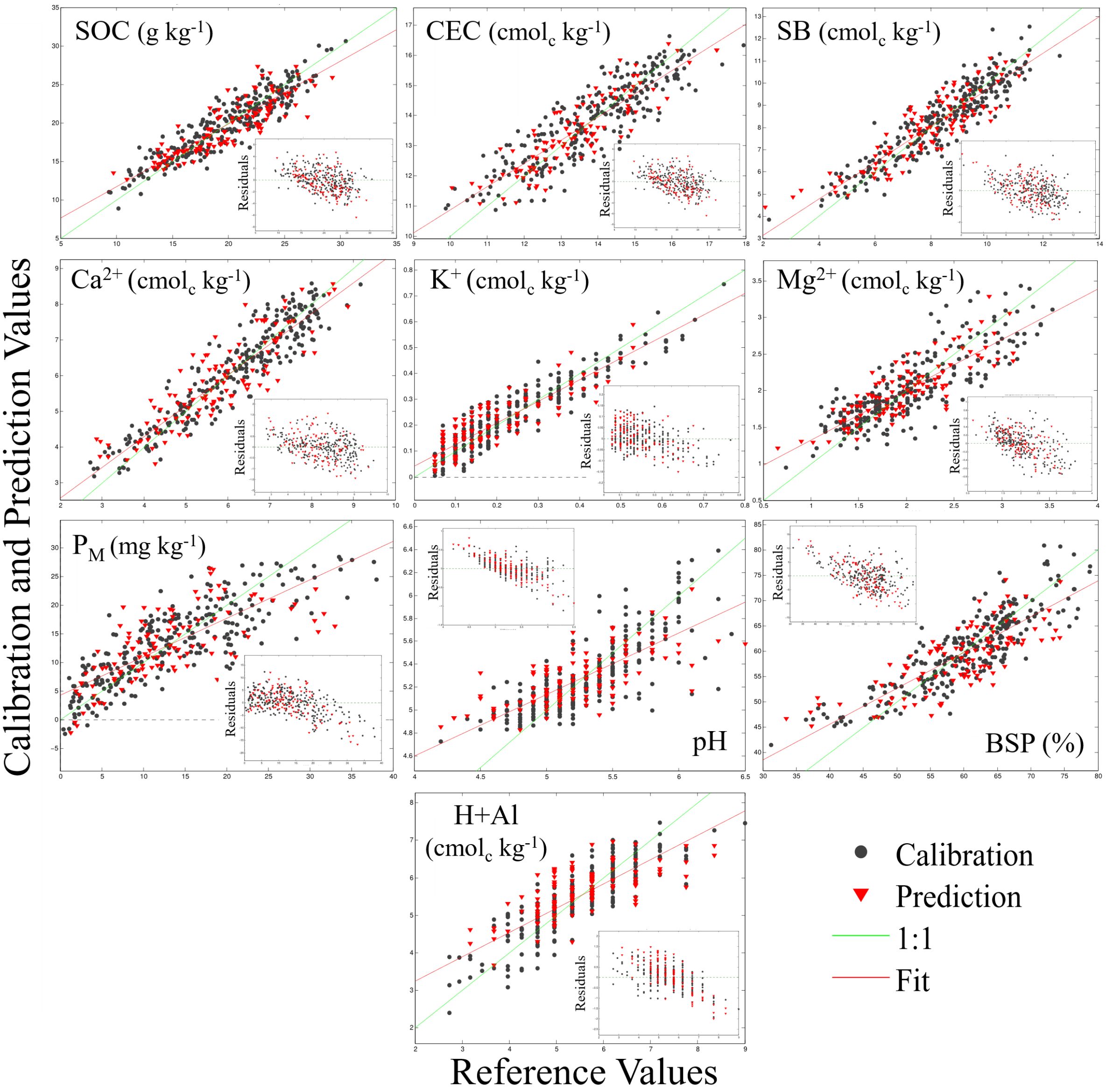

Figure 19 - Scatter plots of reference values versus predicted values (larger boxes) and residuals versus reference values (smaller boxes) for PLS 15kV models

Besides the RMSE and RPD values, RE(%) was also evaluated to certify the models accuracy. The K$^+$ and P$_M$ models showed the highest RE(%), 28 and 34%, respectively. The other models presented values below 14%.

Conversely, the scatterplots for the 15kV models of all soil attributes (Figure 19) show a linear pattern between predicted and reference data. It is also observed that the residuals are randomly distributed around zero. Although pH, BSP and H+Al models show slightly biased behavior in the residuals plot, the visual analyzes are qualitative ways of analyzing linearity, which was verified quantitatively through the DW test. The DW p-values at 95% confidence level did not indicate a lack of linearity in anyone model (Table 13). Thus, linear multivariate calibration methods such as PLS is enough for modeling soil fertility attributes with pXRF spectral data.

Table 13 - Figures of merit of the 15kV PLS.

Attribute	DW Test p-value	Bias	t_{bias}	SEN	γ^{-1}	RE (%)	LOQ	LOD
SOC (g kg^{-1})	0.82	0.37	1.95	0.3	0.2	10	1.83	0.60
CEC (cmol$_c$ kg^{-1})	0.15	0.03	0.50	0.6	0.1	5	0.99	0.33
SB (cmol$_c$ kg^{-1})	0.5	0.07	0.85	0.7	0.1	11	0.89	0.29
Ca^{2+} (cmol$_c$ kg^{-1})	0.5	0.05	0.65	0.4	0.08	11	0.80	0.27
K^+ (cmol$_c$ kg^{-1})	0.94	0.01	2.46	0.15	0.04	28	0.40	0.13
Mg^{2+} (cmol$_c$ kg^{-1})	0.98	0.03	0.93	1.63	0.04	14	0.38	0.13
P_M (mg kg^{-1})	0.85	0.06	0.02	0.22	0.14	36	1.44	0.47
pH	0.12	0.03	1.28	1.9	0.03	6	0.33	0.11
BSP (%)	0.41	0.50	0.94	0.66	0.09	9	0.94	0.31
H+Al (cmol$_c$ kg^{-1})	0.31	0.10	1.15	0.24	0.26	13	2.56	0.84

$p_{critic} = 0.05$; $t_{critic} = 1.98$

2.3.6 VIP Scores

The spectra energy range or peaks with higher importance in the prediction of each fertility attribute for the 15kV models were evaluated by the informative vector called the variable influence on projection (VIP) scores (Figure 20).

In the VIP scores, the variables with intensities greater than unity contribute significantly to the modeling. According to them (Figure 20), the Ca- kα characteristic X-ray peak was the most important variable in SOC, SB, Ca^{2+}, Mg^{2+}, P_M, pH, BSP and H+Al models. The CEC and K^+ models had Mn- kα, and K- kα peaks, respectively, as the most important variables. However, the peak Ca-kα is the second most important variable in these models. These findings agree with the correlation analysis carried out in previous studies by Dos Santos *et al.* [18]. This was expected because Ca is directly or indirectly involved with most of the evaluated attributes. The pH variation over time happens according to soil management, successive crops, and fertilizations. When plants absorb positively charged nutrients (including Ca^{2+}), they release H^+ from the roots into the soil solution, which reduces the pH. The Ca^{2+} and Mg^{2+} contents are directly related to soil acidity. Generally, acidic soils have low Ca^{2+} and Mg^{2+} levels and soils with good fertility have higher Ca^{2+} and Mg^{2+} levels [71]. Moreover, Ca^{2+} is directly considered in CEC, SB, and BSP calculations.

This significant importance of Ca content in different fertility attributes modeling has also been reported in other studies conducted in different Brazilian tropical soils [18],

[19], [72]–[75]. Because a significant part of the mineral composition is quartz, micas, kaolinite and Fe and Al-oxides, Brazilian tropical soils may lack Ca in their crystalline structure. According to most authors, this may explain the Ca content in soils, since 95% of its speciation consists of the exchangeable form Ca^{2+} [18].

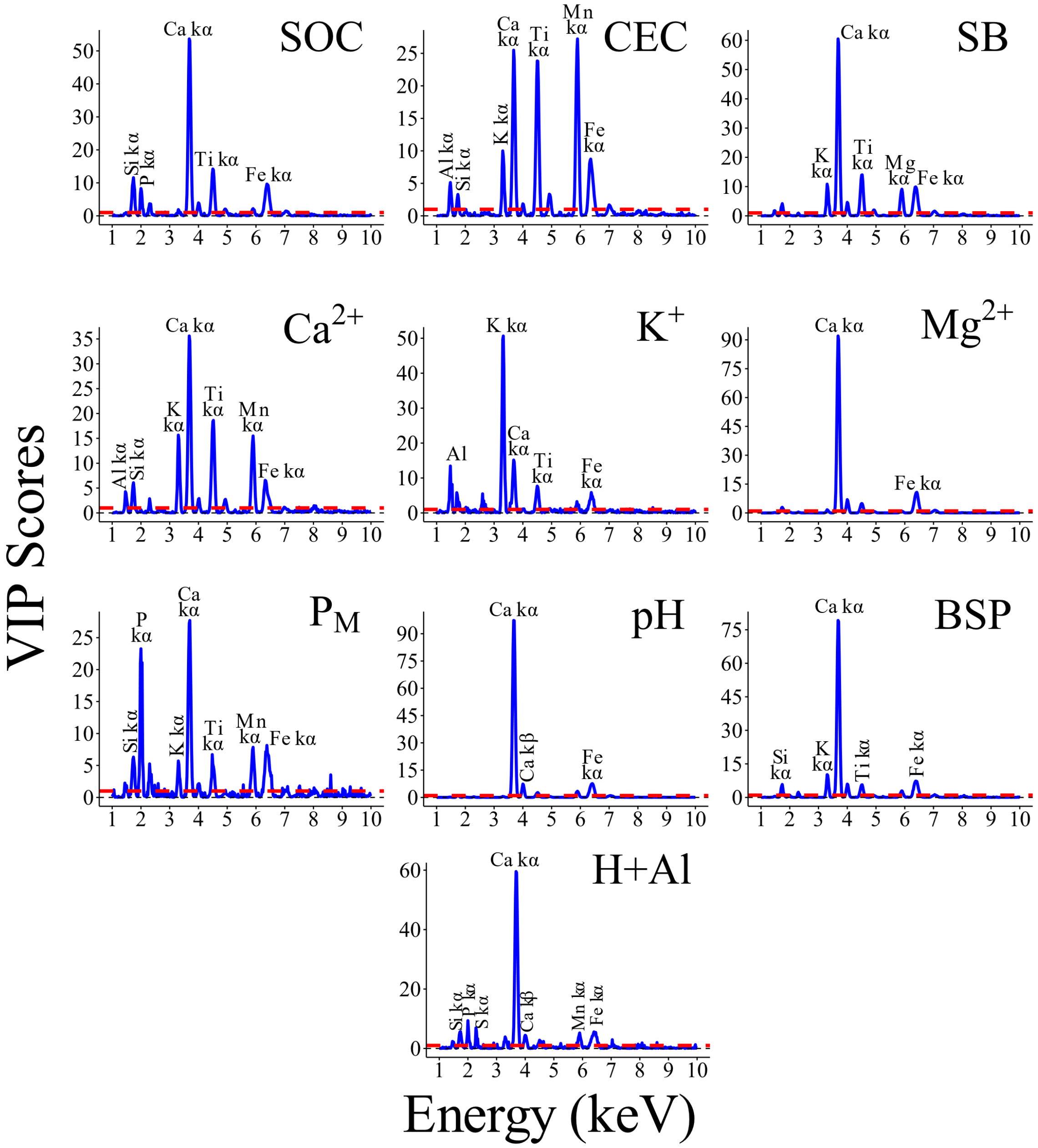

Figure 20 - Variable influence on projection (VIP) scores of 15kV PLS models. Values greater than unity (red dotted line) indicate variables that significantly contributed to modeling

2.3.7 Comparison Between pXRF and Benchtop EDXRF Models

The prediction results of the EDXRF Benchtop models reported by Dos Santos *et al.* [18], [26] are shown in Table 14. Comparing the results of the 15kV models with them, it is possible to verify that the values of Pred R², RMSEP and RPD were similar, with a slight superiority for the pXRF models in the case of CEC, K^+, Mg^{2+} and P_M.

Table 14 - Prediction results from PLS calibration models with benchtop EDXRF spectral data reported by Dos Santos *et al.* in [18], [26]. The randomization test regarding the 15kV pXRF model is presented.

Attribute	Pred R²	RMSEP	RPD	Random Test p-value
SOC (g kg^{-1})	0.79	1.90	2.2	**0.31**
CEC (cmol$_c$ kg^{-1})	0.67	0.85	1.7	0.01
SB (cmol$_c$ kg^{-1})	0.82	0.77	2.4	**0.15**
Ca^{2+} (cmol$_c$ kg^{-1})	0.81	0.63	2.3	**0.32**
K^+ (cmol$_c$ kg^{-1})	0.65	0.075	1.7	**0.93**
Mg^{2+} (cmol$_c$ kg^{-1})	0.64	0.35	1.7	0.02
P_M (mg kg^{-1})	0.39	5.72	1.2	**0.19**
pH	0.57	0.26	1.5	**0.16**
BSP (%)	0.71	5.2	1.8	**0.45**
H+Al (cmol$_c$ kg^{-1})	0.55	0.72	1.5	**0.27**

* Bold indicates values greater than critical (0.05 for 95% confidence).

A randomization test at 95% confidence level comparing the 15kV models' prediction values with the benchtop EDXRF values was performed. It is notable that, among all soil attributes analyzed, only the CEC and Mg^{2+} pXRF models were not considered statistically equivalent to the models built with benchtop EDXRF spectra. In these cases, the pred R², RMSEP and RDP values indicated a small superiority of the pXRF models over the benchtop EDXRF models [18]. Although the difference in the model's performance is small, it may be responsible for the non-equivalence in the randomization test.

Although further studies are needed, it is possible to theorize that one of the reasons for the pXRF and benchtop EDXRF models equivalence may be related to the technology involved in both compared equipment. At first, the Shimadzu EDX 720 benchtop EDXRF equipment may be considered better because it is more robust, however, the electronics associated with signal processing are more modern in Tracer 5i pXRF equipment used in this study, allowing processing to be performed in a shorter time interval. This makes pXRF capable of detecting a greater number of counts in a shorter time, which leads to a richer spectrum of information. Another possible reason is the measurement geometry. In the case of pXRF, the distance traveled by the X-rays characteristic of the sample to the detector is smaller, reducing the chances of attenuation in the air, an effect that reduces the number of counts that arrive for detection. Therefore, even being less robust, pXRF could produce results equivalent to the compared benchtop.

In general terms, the performance equivalence of the models built with the equipment evaluated in this study are evidence that corroborate the third hypothesis, indicating that PLS modeling with spectra from portable XRF equipment generate models as accurately and with good performance as the methodology with benchtop equipment. Since the benchtop equipment remains in the laboratory, a great advantage of using pXRF is the mobility and speed associated with the measurement step. After calibrating and validating the models, soil attributes measurements with good accuracy may be obtained only with the spectral reading of new soil samples prepared in the same procedure as before [14]. As soil chemical mapping needs to be renewed periodically due to the influence of agronomic practices and the dynamics of the soil-plant-atmosphere system [76], this methodology may allow the soil fertility spatial variability to be quickly determined and with a smaller number of samples sent for conventional analysis. The decrease in sampling density based only on conventional laboratory analysis is advantageous from a financial, environmental, and logistical point of view. On the other hand, with spatially accurate information about soil fertility attributes, fertilizers and limes application may be optimized, increasing agricultural productivity, and reducing input costs.

However, it is important to clarify that the generalization of the equivalence between the pXRF models with the EDXRF benchtop in all circumstances is not recommended, since the models built in this work are local in relation to space (sample collection area and soil types), in time (management, dynamics of use and state of the soil at the time of collection), in the experimental conditions and in the equipment tested.

2.3.8 Extrapolation of optimized models for in situ measurements

The positive results of laboratory calibration with the portable XRF spectrometer highlight the potential of *in situ* measurements and modeling. Although *in situ* modeling is an ambitious and challenging proposal that still needs to be widely evaluated, a preliminary study on the extrapolation of laboratory-calibrated models to *in situ* measurements was conducted. Table 15 shows the moisture content of the 18 samples measured *in situ,* which was determined by weighing them before and after drying. The values ranged from 11.7 to 7.8%, and the average moisture content of all samples was 10%.

As the 15kV spectra of the 276 samples used to calibrate the models was established as optimal conditions, the models calibrated with them were used to evaluate the *in situ* predictions. Figure 21 provides a plot of the calibration mean spectra and the mean spectra of in-situ-sets samples (InSitu, Dry4, Dry30, Dry2mm4, and Dry2mm30 sets). It is noted that the spectra have different intensities across all variables. The *in situ* spectra have the lowest intensity compared to the others, followed by the dry sample spectra (4 and 30 s). The most intense spectra have dry samples and particle size < 2 mm (4 and 30 s). This difference in intensity may be due to physical matrix effects. Measurements made directly in the field suffer from the effects of moisture, roughness, irregularities, and different particle sizes. For example, the irregularity of the surface and the lack of granulometry control complicate the uniform X-ray irradiation, causing shadows on smaller particles when they are in the proximity of larger particles [77]–[79]. On the other hand, because the X-ray fluorescence signal emitted from the sample surface is a function of its composition, moisture and SOC may cause a dilution effect, resulting in lower count rates. This is because the increase in lighter elements in the soil (such as H, C, and O) increases the Rayleigh and Compton scattering of the spectrum [78], [79]. Therefore, the spectra also differ according to the variation in moisture and SOC content in the soil at different sampling depths (*e.g.*, higher SOC values are found closer to the surface).

In addition, the noise in spectra collected for 4 s is higher than that in 30 s spectra (Figure 21). In general, this reduces the accuracy of the measurements. However, for elements with intense K-lines (such as Fe, Ti, and even Si), this noise may not be significantly interfered. Tavares *et al.* [57] show that although reducing the measurement time of XRF analysis reduces the precision of the data, it is still possible to obtain excellent prediction performances in fertility attributes modeling with drastic reductions (for example, from 90 to 2 s).

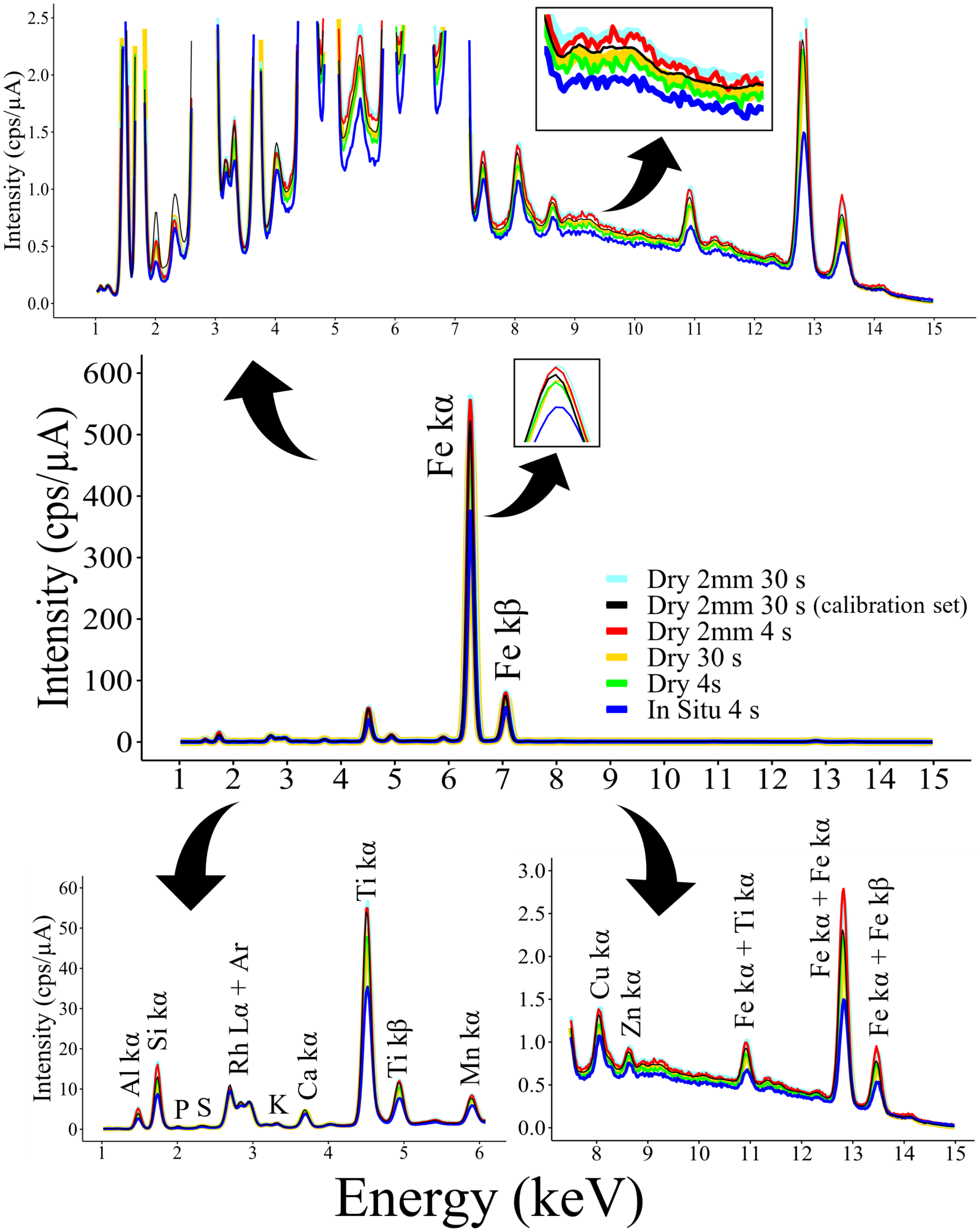

Figure 21 - Comparison among the mean spectra used to models calibration obtained with optimized conditions (at 15kV), and in-situ-sets spectra

Table 15 - Moisture of samples collected and measured *in situ*

Sample	Collected (g)	Dry (g)	Moisture (%)
1	21.9	19.6	9.4
2	23.6	21.4	8.4
3	27.0	24.1	9.7
4	39.1	34.1	11.7
5	33.3	29.6	10.5
6	34.1	30.0	11.0
7	29.1	25.8	10.2
8	27.7	24.6	10.3
9	34.2	30.8	9.5
10	43.3	38.5	10.7
11	33.8	30.1	10.6
12	33.7	30.9	7.8
13	31.8	28.3	10.5
14	31.0	27.9	9.5
15	45.6	40.5	10.7
16	61.7	55.1	10.4
17	38.0	33.4	11.4
18	24.8	22.3	8.4

To gain perspective on the results, a PCA was performed with the spectra of in-situ-sets samples and from the optimized 15kV condition (calibration set). To avoid false behavior due to spectral misalignment, the spectra of the in-situ-sets were aligned using the mean spectrum calculated among the calibration samples (15kV condition) as a reference. After that, all spectra were preprocessed with Poisson + MC. Figure 22 shows the scores and loadings of the first two principal components, which together sum to 99.3% of the explained variance.

Except for the *in situ* spectra, all other spectral sets form individual clusters in the scores. The two sets measured at 30 s (dry and dry with grain size) are positioned on the positive axis of PC 2, while those obtained at 4 s (dry and dry with grain size) are on the negative axis. The *in situ* spectra are distributed along PC 2 in the negative range of PC 1.

Moreover, the cluster of 15kV condition spectra (calibration set) is located approximately at the origin of the PC 1 and PC 2 axes (point 0,0). This behavior may be attributed to the alignment performed.

The loading plot (Figure 22) indicates that the variables that most contributed to these clusters are the Fe and Ti-K lines. This is also evident from the mean spectra plot

(Figure 21), *i.e.*, the intensities of the Fe-K and Ti-K lines are different for each sample set. The existence of multiple clusters is not a good result from a modeling point of view, as it indicates that the spectral sets are significantly different and may generate Bias.

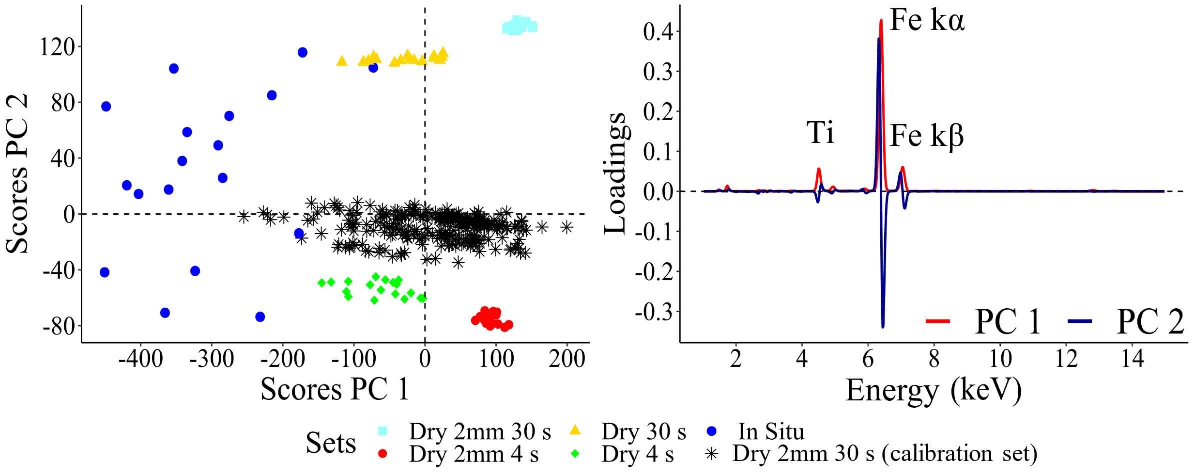

Figure 22 - Scores and loadings from the PCA with spectra collected and measured *in situ* and from the optimized 15kV calibration set. PC 1 and PC 2 have explained variance of 79.6 and 19.7 %, respectively.

The extrapolation prediction results of the SOC and CEC models calibrated with the optimized condition (15kV spectra) are presented in Table 16 and Figure 23. All models performed poorly, being considered very poor models (RPD < 1.4 [46]). The predictions are biased (t_{Bias} >> t_{critic}) and inaccurate (high RMSEP values).

These results indicate that, although SOC and CEC modeling under controlled sample preparation conditions is promising, extrapolation to *in situ* measurements is challenging and complex. Therefore, the use of calibration and prediction sample sets separated by large time intervals is not recommended. The SOC and CEC models did not show sufficient robustness to produce reliable prediction results in the scenario where temporal variations in the sample set occur. This may be due to the different soil management and preparation practices used for cultivation in the study area. Negative results were obtained even in the same preparation conditions of the calibration and prediction samples. The significant difference in spectra complicates the assessment of the true impact of the reduction in SOC and CEC modeling prediction performance caused by matrix effects present in samples measured under *in situ* conditions.

Table 16 - Predictive performance of extrapolation using models calibrated with spectra at 15kV (optimized condition) to quantify SOC and CEC *in situ.* t_{critic} = 1.98

	R² Pred	RMSEP	RPD	Bias Pred	t_{Bias}
			SOC models (g kg^{-1})		
InSitu*	0.03	7.15	0.92	6.11	6.79
Dry4	0.12	11.96	0.45	9.25	5.03
Dry30	0.13	12.55	0.44	9.89	5.28
Dry2mm4	0.12	12.12	1.11	11.73	15.79
Dry2mm30	0.14	12.78	1.06	12.38	15.98
			CEC models (cmolc kg^{-1})		
InSitu*	0.03	1.77	0.66	-1.11	3.31
Dry4	0.05	2.13	0.64	-1.60	4.61
Dry30	0.04	2.16	0.63	-1.61	4.62
Dry2mm4	0.04	2.15	0.63	-1.62	4.63
Dry2mm30	0.05	2.16	0.63	-1.61	4.62

*Sample sets: 18 samples measured *in situ* with 4 s (InSitu), 18 dry samples measured in the laboratory with 4 s (Dry4) and 30 s (Dry30), 18 dry samples with particle size < 2 mm measured in the laboratory with 4 s (Dry2mm4) and 30 s (Dry2mm30). Pred (prediction)

Conversely, although the results of this study are negative, *in situ* modeling still needs to be further explored through other approaches. For example, models calibrated and validated with data from samples from the same area and collection period. Another approach for cases of large sample sets: samples collected at different periods from the same area may be mixed in the calibration and validation sets, providing more variability for modeling.

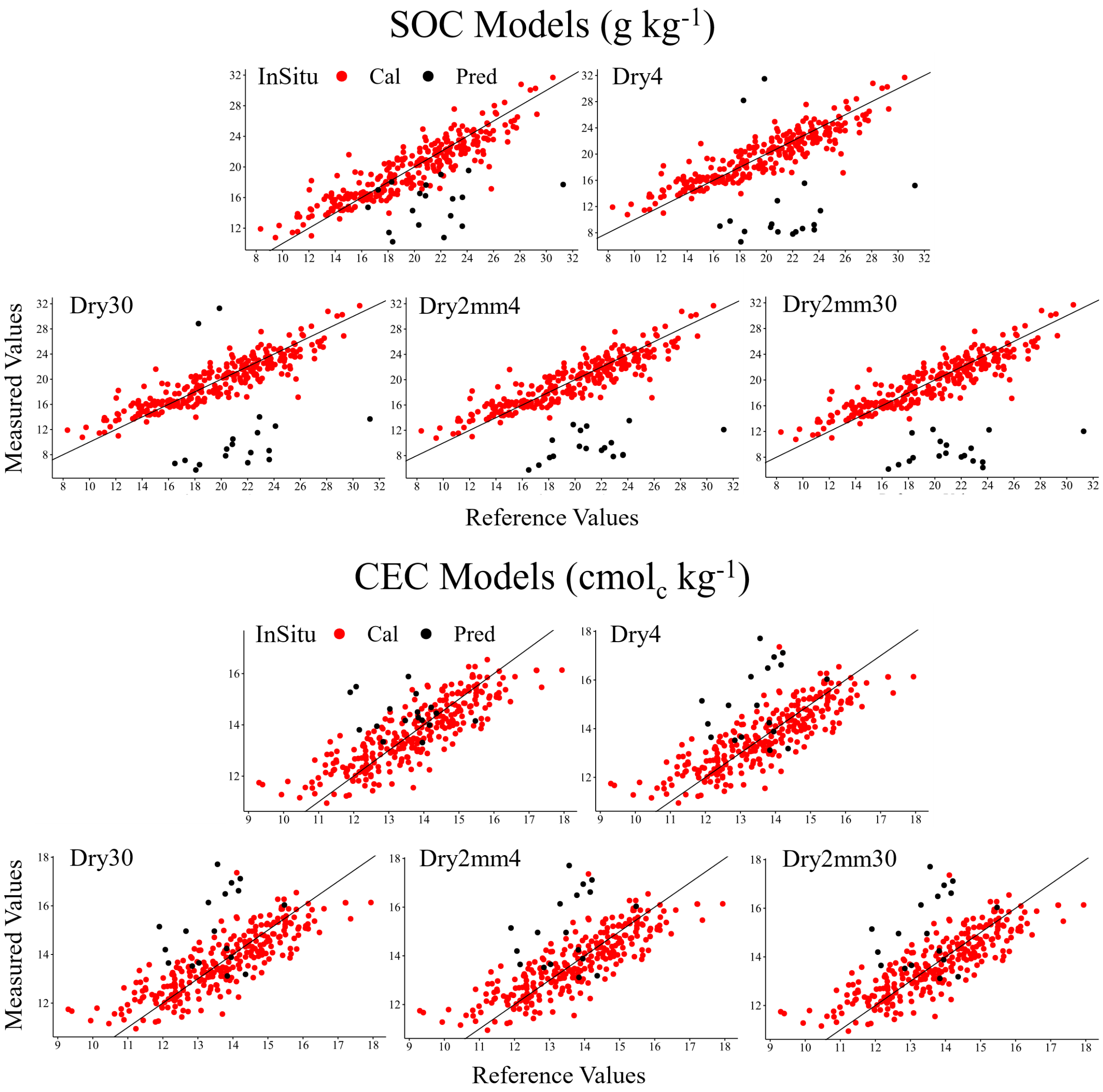

Figure 23 - Scatter plots of reference values versus predicted values of models extrapolated to in situ predictions. Sample sets: 18 samples measured *in situ* with 4 s (InSitu); 18 dry samples measured in the laboratory with 4 s (Dry4) and 30 s (Dry30); 18 dry samples with particle size < 2 mm measured in the laboratory with 4 s (Dry2mm4) and 30 s (Dry2mm30). Cal (calibration); Pred (prediction)

2.4 Conclusion

The present study evaluated the optimized instrumental condition for soil fertility attributes employing a pXRF spectrometer and soil samples. Exploring the different possibilities of results from a commercial pXRF routine, four data matrices were used in the multivariate modeling. Three of them with raw spectra and one with elemental concentration data. The use of raw spectra measured with 15 kV and 12.15 µA without filters as experimental conditions, modeled with PLS allows the best performance for the simultaneous quantification of the 10 soil fertility attributes evaluated. By the established RPD ranges [50], the SOC, CEC, SB, and Ca^{2+} models were considered very good for quantitative prediction while the K^{+} and Mg^{2+} models were considered good for quantitative predictions. The P_M, pH, BSP, and H+Al models presented fair predictions, which may be used for evaluation and correlation. These results are reliable for local applications in the context of the soil samples evaluated in this study. However, they demonstrate the pXRF application potential combined with multivariate calibration, which may be applied in different new contexts to extract quantitative and qualitative information on soil fertility.

Conversely, the Relative Improvement and p-values comparing the optimized condition with data-fusion models demonstrated that a single experimental condition employment is sufficient to produce accurate and reliable results in this study context. Optimally utilizing this single experimental condition results in shorter time intervals for measurements, a significantly smaller amount of data to manipulate, and less processing power required to manage the models.

Additionally, the performance of the modeling with pXRF was considered equivalent to the EDXRF benchtop equipment by the randomization test, revealing a potentially possibility for the *in situ* application of this method. Nevertheless, the preliminary study on extrapolation of laboratory-calibrated models showed negative results, indicating that their extrapolation to *in situ* measurements is challenging and complex. Therefore, it is not recommended to use calibration and prediction sample sets collected separately between wide time intervals in the same study area. The SOC and CEC models did not demonstrate sufficient robustness to produce reliable prediction results in this scenario.

In conclusion, the present study provided subsidies regarding the pXRF experimental condition choice in soil fertility attributes modeling. These findings demonstrated that the methodology may be employed in local soil analysis to speed up results in terms of measurement time and processing power to turn agile the decision related to the use of fertilizers and other products for soil correction. Now that the optimized conditions are established, further studies are necessary to effectively evaluate the performance of this approach under a wider range of soil variability, especially if *in situ* application will be considered.

2.5 References

[1] J. A. M. Demattê, A. C. Dotto, L. G. Bedin, V. M. Sayão, and A. B. e Souza, “Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact,” *Geoderma*, vol. 337, pp. 111–121, Mar. 2019, doi: 10.1016/j.geoderma.2018.09.010.

[2] United Nations Department of Economic and Social Affairs, “World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100,” *New York: United Nations Department of Economic and Social Affairs, Population Division.* 2017. Accessed: Jul. 16, 2023. [Online]. Available: https://www.un.org/development/desa/en/news/population/world-population-prospects-2017.html

[3] T. Rodríguez-Espinosa, J. Navarro-Pedreño, I. G. Lucas, and M. Belén Almendro-Candel, “Land recycling, food security and Technosols,” *Journal of Geographical Research*, vol. 4, no. 3, pp. 44–50, Aug. 2021, doi: 10.30564/jgr.v4i3.3415.

[4] A. McBratney, D. J. Field, and A. Koch, “The dimensions of soil security,” *Geoderma*, vol. 213, pp. 203–213, Jan. 2014, doi: 10.1016/j.geoderma.2013.08.013.

[5] S. H. G. Silva *et al.*, “Soil texture prediction in tropical soils: A portable X-ray fluorescence spectrometry approach,” *Geoderma*, vol. 362, p. 114136, Mar. 2020, doi: 10.1016/j.geoderma.2019.114136.

[6] J. V. Fontenelli *et al.*, “Evaluating the synergy of three soil spectrometers for improving the prediction and mapping of soil properties in a high anthropic management area: A case of study from Southeast Brazil,” *Geoderma*, vol. 402, p. 115347, Nov. 2021, doi: 10.1016/j.geoderma.2021.115347.

[7] M. A. Gomes and H. F. Filizola, *Indicadores físicos e químicos de qualidade de solo de interesse agrícola*, 1st ed. Jaguariúna: Embrapa Meio Ambiente, 2006.

[8] T. H. Waiser, C. L. S. Morgan, D. J. Brown, and C. T. Hallmark, “In Situ Characterization of Soil Clay Content with Visible Near-Infrared Diffuse Reflectance Spectroscopy,” *Soil Science Society of America Journal*, vol. 71, no. 2, pp. 389–396, Mar. 2007, doi: 10.2136/sssaj2006.0211.

[9] J. A. M. Demattê *et al.*, “The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges,” *Geoderma*, vol. 354, p. 113793, Nov. 2019, doi: 10.1016/j.geoderma.2019.05.043.

[10] F. R. dos Santos, J. F. de Oliveira, E. Bona, J. V. F. dos Santos, G. M. C. Barboza, and F. L. Melquiades, “EDXRF spectral data combined with PLSR to determine some soil fertility indicators,” *Microchemical Journal*, vol. 152, p. 104275, Jan. 2020, doi: 10.1016/j.microc.2019.104275.

[11] A. Gredilla, S. Fdez-Ortiz de Vallejuelo, N. Elejoste, A. de Diego, and J. M. Madariaga, “Non-destructive Spectroscopy combined with chemometrics as a tool for Green Chemical Analysis of environmental samples: A review,” *TrAC Trends in Analytical Chemistry*, vol. 76, pp. 30–39, Feb. 2016, doi: 10.1016/j.trac.2015.11.011.

[12] T. R. Tavares *et al.*, “Multi-Sensor Approach for Tropical Soil Fertility Analysis: Comparison of Individual and Combined Performance of VNIR, XRF, and LIBS Spectroscopies,” *Agronomy*, vol. 11, no. 6, p. 1028, May 2021, doi: 10.3390/agronomy11061028.

[13] S. M. O'Rourke, B. Minasny, N. M. Holden, and A. B. McBratney, "Synergistic Use of Vis-NIR, MIR, and XRF Spectroscopy for the Determination of Soil Geochemistry," *Soil Science Society of America Journal*, vol. 80, no. 4, pp. 888–899, Jul. 2016, doi: 10.2136/sssaj2015.10.0361.

[14] C. Guerrero, R. Zornoza, I. Gómez, and J. Mataix-Beneyto, "Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy," *Geoderma*, vol. 158, no. 1–2, pp. 66–77, Aug. 2010, doi: 10.1016/j.geoderma.2009.12.021.

[15] T. R. Tavares, J. P. Molin, S. H. Javadi, H. W. P. de Carvalho, and A. M. Mouazen, "Combined Use of Vis-NIR and XRF Sensors for Tropical Soil Fertility Analysis: Assessing Different Data Fusion Approaches," *Sensors*, vol. 21, no. 1, p. 148, Dec. 2020, doi: 10.3390/s21010148.

[16] F. L. Melquiades and F. R. dos Santos, "Preliminary Results: Energy Dispersive X-Ray Fluorescence and Partial Least Squares Regression for Organic Matter Determination in Soil," *Spectroscopy Letters*, vol. 48, no. 4, pp. 286–289, Apr. 2015, doi: 10.1080/00387010.2013.874532.

[17] F. Morona, F. R. dos Santos, A. M. Brinatti, and F. L. Melquiades, "Quick analysis of organic matter in soil by energy-dispersive X-ray fluorescence and multivariate analysis," *Applied Radiation and Isotopes*, vol. 130, pp. 13–20, Dec. 2017, doi: 10.1016/j.apradiso.2017.09.008.

[18] F. R. dos Santos, J. F. de Oliveira, G. M. C. Barbosa, and F. L. Melquiades, "Comparison between energy dispersive X-ray fluorescence spectral data and elemental data for soil attributes modelling," *Spectrochim Acta Part B At Spectrosc*, vol. 185, p. 106303, Nov. 2021, doi: 10.1016/j.sab.2021.106303.

[19] R. Andrade *et al.*, "Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains," *Geoderma*, vol. 357, p. 113960, Jan. 2020, doi: 10.1016/j.geoderma.2019.113960.

[20] T. R. Tavares *et al.*, "Effect of X-Ray Tube Configuration on Measurement of Key Soil Fertility Attributes with XRF," *Remote Sens (Basel)*, vol. 12, no. 6, p. 963, Mar. 2020, doi: 10.3390/rs12060963.

[21] T. R. Tavares *et al.*, "Assessing Soil Key Fertility Attributes Using a Portable X-ray Fluorescence: A Simple Method to Overcome Matrix Effect," *Agronomy*, vol. 10, no. 6, p. 787, Jun. 2020, doi: 10.3390/agronomy10060787.

[22] A. Sharma, D. C. Weindorf, T. Man, A. A. A. Aldabaa, and S. Chakraborty, "Characterizing soils via portable X-ray fluorescence spectrometer: 3. Soil reaction (pH)," *Geoderma*, vol. 232–234, pp. 141–147, Nov. 2014, doi: 10.1016/j.geoderma.2014.05.005.

[23] A. Rawal *et al.*, "Determination of base saturation percentage in agricultural soils via portable X-ray fluorescence spectrometer," *Geoderma*, vol. 338, pp. 375–382, Mar. 2019, doi: 10.1016/j.geoderma.2018.12.032.

[24] A. F. dos Santos Teixeira *et al.*, "Tropical soil pH and sorption complex prediction via portable X-ray fluorescence spectrometry," *Geoderma*, vol. 361, p. 114132, Mar. 2020, doi: 10.1016/j.geoderma.2019.114132.

[25] F. R. dos Santos, J. F. de Oliveira, G. M. C. Barbosa, and F. L. Melquiades, "Comparison between energy dispersive X-ray fluorescence spectral data and elemental data for soil

attributes modelling," *Spectrochim Acta Part B At Spectrosc*, vol. 185, p. 106303, Nov. 2021, doi: 10.1016/j.sab.2021.106303.

[26] F. R. dos Santos, J. F. de Oliveira, E. Bona, G. M. C. Barbosa, and F. L. Melquiades, "Evaluation of pre-processing and variable selection on energy dispersive X-ray fluorescence spectral data with partial least square regression: A case of study for soil organic carbon prediction," *Spectrochim Acta Part B At Spectrosc*, vol. 175, p. 106016, Jan. 2021, doi: 10.1016/j.sab.2020.106016.

[27] G. W. D. Ferreira *et al.*, "Assessment of iron-rich tailings via portable X-ray fluorescence spectrometry: the Mariana dam disaster, southeast Brazil," *Environ Monit Assess*, vol. 193, no. 4, p. 203, Apr. 2021, doi: 10.1007/s10661-021-08982-7.

[28] G. F. C. Lima, C. C. Bento, A. H. Horn, E. D. Marques, and H. B. Filho, "Geochemical signature and environmental background of bottom sediments in a tropical aquatic system: the Três Marias Reservoir, Brazil," *Environ Monit Assess*, vol. 193, no. 2, p. 85, Feb. 2021, doi: 10.1007/s10661-021-08876-8.

[29] R. Jenkins, "X-Ray Fluorescence Spectrometry," in *Handbook of Analytical Techniques*, Weinheim, Germany: Wiley-VCH Verlag GmbH, 1988, pp. 753–766. doi: 10.1002/9783527618323.ch23.

[30] Y. Mizuno and Y. Ohmura, "Theory of X-Ray Raman Scattering," *J Physical Soc Japan*, vol. 22, no. 2, pp. 445–449, Feb. 1967, doi: 10.1143/JPSJ.22.445.

[31] F. M. Verbi, E. R. Pereira-Filho, and M. I. M. S. Bueno, "Use of X-Ray Scattering for Studies with Organic Compounds: a Case Study Using Paints," *Microchimica Acta*, vol. 150, no. 2, pp. 131–136, Jun. 2005, doi: 10.1007/s00604-005-0352-5.

[32] R. E. Van Grieken and A. A. Markowicz, *Handbook of X-ray spectrometry*, 2nd ed., vol. 29. New York, 2002.

[33] L. Benedet *et al.*, "Rapid soil fertility prediction using X-ray fluorescence data and machine learning algorithms," *Catena (Amst)*, vol. 197, p. 105003, Feb. 2021, doi: 10.1016/j.catena.2020.105003.

[34] R. Andrade *et al.*, "Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains," *Geoderma*, vol. 357, p. 113960, Jan. 2020, doi: 10.1016/j.geoderma.2019.113960.

[35] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, and O. Busto, "Data fusion methodologies for food and beverage authentication and quality assessment – A review," *Anal Chim Acta*, vol. 891, pp. 1–14, Sep. 2015, doi: 10.1016/j.aca.2015.04.042.

[36] A. El Ghaziri, V. Cariou, D. N. Rutledge, and E. M. Qannari, "Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of (K + 1) datasets," *J Chemom*, vol. 30, no. 8, pp. 420–429, Aug. 2016, doi: 10.1002/cem.2810.

[37] H. G. dos Santos *et al.*, *Sistema brasileiro de classificação de solos*, 2. Rio de Janeiro: EMBRAPA-SPI, 2006.

[38] FAO, "IUSS working group WRB. World reference base for soil resources 2014, International soil classification system for naming soils and creating legends for soil maps," *World Soil Resources Reports No. 106*, 2014.

[39] M. A. Pavan, M. de F. Bloch, H. da C. Zempulski, M. Miyazawa, and D. C. Zocoler, *Manual de análise química de solo e controle de qualidade*, vol. 76. Iapar Londrina, 1992.

[40] R. W. Kennard and L. A. Stone, "Computer Aided Design of Experiments," *Technometrics*, vol. 11, no. 1, pp. 137–148, Feb. 1969, doi: 10.1080/00401706.1969.10490666.

[41] R. de A. Ferreira, G. Teixeira, and L. A. Peternelli, "Kennard-Stone method outperforms the Random Sampling in the selection of calibration samples in SNPs and NIR data," *Ciência Rural*, vol. 52, no. 5, 2022, doi: 10.1590/0103-8478cr20201072.

[42] P. Valderrama, J. W. B. Braga, and R. J. Poppi, "Variable Selection, Outlier Detection, and Figures of Merit Estimation in a Partial Least-Squares Regression Multivariate Calibration Model. A Case Study for the Determination of Quality Parameters in the Alcohol Industry by Near-Infrared Spectroscopy," *J Agric Food Chem*, vol. 55, no. 21, pp. 8331–8338, Oct. 2007, doi: 10.1021/jf071538s.

[43] M. R. Keenan and P. G. Kotula, "Optimal scaling of TOF-SIMS spectrum-images prior to multivariate statistical analysis," *Appl Surf Sci*, vol. 231–232, pp. 240–244, Jun. 2004, doi: 10.1016/j.apsusc.2004.03.025.

[44] M. R. Keenan and P. G. Kotula, "Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images," *Surface and Interface Analysis*, vol. 36, no. 3, pp. 203–212, Mar. 2004, doi: 10.1002/sia.1657.

[45] P. Mishra *et al.*, "MBA-GUI: A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing," *Chemometrics and Intelligent Laboratory Systems*, vol. 205, p. 104139, Oct. 2020, doi: 10.1016/j.chemolab.2020.104139.

[46] R. A. Viscarra Rossel, R. N. McGlynn, and A. B. McBratney, "Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy," *Geoderma*, vol. 137, no. 1–2, pp. 70–82, Dec. 2006, doi: 10.1016/j.geoderma.2006.07.004.

[47] A. C. Olivieri, "Practical guidelines for reporting results in single- and multi-component analytical calibration: A tutorial," *Anal Chim Acta*, vol. 868, pp. 10–22, Apr. 2015, doi: 10.1016/j.aca.2015.01.017.

[48] H. van der Voet, "Comparing the predictive accuracy of models using a simple randomization test," *Chemometrics and Intelligent Laboratory Systems*, vol. 25, no. 2, pp. 313–323, Nov. 1994, doi: 10.1016/0169-7439(94)85050-X.

[49] E1655-05 ASTM, "Standard Practices for Infrared Multivariate Quantitative Analysis," *Conshohocken*, 2012.

[50] L. A. Currie, "Nomenclature in evaluation of analytical methods including detection and quantification capabilities1Adapted from the International Union of Pure and Applied Chemistry (IUPAC) document 'Nomenclature in Evaluation of Analytical Methods including Detection and Quantification Capabilities', which originally appeared in Pure and Applied Chemistry, 67 1699–1723 (1995) © 1995 IUPAC. Republication permission granted by IUPAC.1," *Anal Chim Acta*, vol. 391, no. 2, pp. 105–126, May 1999, doi: 10.1016/S0003-2670(99)00104-X.

[51] V. Centner, O. E. de Noord, and D. L. Massart, "Detection of nonlinearity in multivariate calibration," *Anal Chim Acta*, vol. 376, no. 2, pp. 153–168, Dec. 1998, doi: 10.1016/S0003-2670(98)00543-1.

[52] ABNT.NBR ISO 5725-1:2018, “Exatidão (veracidade e precisão) dos métodos e dos resultados de medição - Parte 1: Princípios gerais e definições,” *Rio de Janeiro*, 2018.

[53] INMETRO - INSTITUTO NACIONAL DE METROLOGIA NORMALIZAÇÃO E QUALIDADE INDUSTRIAL, “Vocabulário Internacional de Metrologia – Conceitos fundamentais e gerais e termos associados,” *Rio de Janeiro*, 2012.

[54] P. Valderrama, J. W. B. Braga, and R. J. Poppi, “Estado da arte de figuras de mérito em calibração multivariada,” *Quim Nova*, vol. 32, no. 5, pp. 1278–1287, 2009, doi: 10.1590/S0100-40422009000500034.

[55] M. H. Ferreira, J. W. B. Braga, and M. M. Sena, “Development and validation of a chemometric method for direct determination of hydrochlorothiazide in pharmaceutical samples by diffuse reflectance near infrared spectroscopy,” *Microchemical Journal*, vol. 109, pp. 158–164, Jul. 2013, doi: 10.1016/j.microc.2012.03.008.

[56] F. L. Melquiades, G. G. Bortoleto, L. F. S. Marchiori, and M. I. M. S. Bueno, “Direct Determination of Sugar Cane Quality Parameters by X-ray Spectrometry and Multivariate Analysis,” *J Agric Food Chem*, vol. 60, no. 43, pp. 10755–10761, Oct. 2012, doi: 10.1021/jf302471b.

[57] T. R. Tavares, J. P. Molin, E. E. N. Alves, F. L. Melquiades, H. W. P. de Carvalho, and A. M. Mouazen, “Towards rapid analysis with XRF sensor for assessing soil fertility attributes: Effects of dwell time reduction,” *Soil Tillage Res*, vol. 232, p. 105768, Aug. 2023, doi: 10.1016/j.still.2023.105768.

[58] A. W. Warrick and D. R. Nielsen, “Spatial Variability of Soil Physical Properties in the Field,” in *Applications of Soil Physics*, Elsevier, 1980, pp. 319–344. doi: 10.1016/B978-0-12-348580-9.50018-3.

[59] J. M. Reichert *et al.*, “Conceptual framework for capacity and intensity physical soil properties affected by short and long-term (14 years) continuous no-tillage and controlled traffic,” *Soil Tillage Res*, vol. 158, pp. 123–136, May 2016, doi: 10.1016/j.still.2015.11.010.

[60] L. A. Santi, J. E. Fiorin, K. T. Cocco, M. R. Cherubin, M. T. Eitelwein, and T. J. C. Amado, *Distribuição horizontal e vertical de fósforo e potássio em área manejada com ferramentas de agricultura de precisão*. Revista Plantio Direto, 2012.

[61] C. Briedis *et al.*, “Can highly weathered soils under conservation agriculture be C saturated?,” *Catena (Amst)*, vol. 147, pp. 638–649, Dec. 2016, doi: 10.1016/j.catena.2016.08.021.

[62] F. M. Verbi, E. R. Pereira-Filho, and M. I. M. S. Bueno, “Use of X-Ray Scattering for Studies with Organic Compounds: a Case Study Using Paints,” *Microchimica Acta*, vol. 150, no. 2, pp. 131–136, Jun. 2005, doi: 10.1007/s00604-005-0352-5.

[63] U. Schwertmann and R. M. Taylor, “Iron Oxides,” 2018, pp. 379–438. doi: 10.2136/sssabookser1.2ed.c8.

[64] M. Mauad, H. Grassi Filho, C. A. C. Crusciol, and J. C. Corrêa, “Teores de silício no solo e na planta de arroz de terras altas com diferentes doses de adubação silicatada e nitrogenada,” *Rev Bras Cienc Solo*, vol. 27, no. 5, pp. 867–873, Oct. 2003, doi: 10.1590/S0100-06832003000500011.

[65] G. Uehara, “Acric properties and their significance to soil classification,” in *International Soil Classification Workshop, 8*, 1988.

[66] L. R. F. Alleoni and O. A. Camargo, “Óxidos de ferro e de alumínio e a mineralogia da fração argila deferrificada de latossolos ácricos,” *Sci Agric*, vol. 52, no. 3, pp. 416–421, Dec. 1995, doi: 10.1590/S0103-90161995000300002.

[67] W. H. Schlesinger and E. S. Bernhardt, *Biogeochemistry: an analysis of global change*. Academic press, 2013.

[68] G. B. Sousa, M. V. Martins Filho, and S. S. R. Matias, “Perdas de solo, matéria orgânica e nutrientes por erosão hídrica em uma vertente coberta com diferentes quantidades de palha de cana-de-açúcar em Guariba - SP,” *Engenharia Agrícola*, vol. 32, no. 3, pp. 490–500, Jun. 2012, doi: 10.1590/S0100-69162012000300008.

[69] J. V. F. Dos Santos, “FORMAS DE FÓSFORO EM UMA TOPOSSEQUÊNCIA DE SOLOS ORIUNDOS DO BASALTO,” *VI Reunião Paranaense de Ciência do Solo-RPCS*. INSTITUTO AGRONÔMICO DO PARANÁ (IAPAR), Ponta Grossa , 2019.

[70] B. T. Ribeiro, S. H. G. Silva, E. A. Silva, and L. R. G. Guilherme, “Portable X-ray fluorescence (pXRF) applications in tropical Soil Science,” *Ciência e Agrotecnologia*, vol. 41, no. 3, pp. 245–254, Jun. 2017, doi: 10.1590/1413-70542017413000117.

[71] L. C. Prezotti and A. M. Guarçoni, *Guia de interpretação de análise de solo e foliar*. Vitória: Incaper, 2013.

[72] T. R. Tavares *et al.*, “Effect of X-Ray Tube Configuration on Measurement of Key Soil Fertility Attributes with XRF,” *Remote Sens (Basel)*, vol. 12, no. 6, p. 963, Mar. 2020, doi: 10.3390/rs12060963.

[73] A. F. dos Santos Teixeira *et al.*, “Tropical soil pH and sorption complex prediction via portable X-ray fluorescence spectrometry,” *Geoderma*, vol. 361, p. 114132, Mar. 2020, doi: 10.1016/j.geoderma.2019.114132.

[74] S. H. G. Silva, A. F. dos S. Teixeira, M. D. de Menezes, L. R. G. Guilherme, F. M. de S. Moreira, and N. Curi, “Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence spectrometer (pXRF),” *Ciência e Agrotecnologia*, vol. 41, no. 6, pp. 648–664, Dec. 2017, doi: 10.1590/1413-70542017416010317.

[75] Á. J. G. de Faria *et al.*, “Soils of the Brazilian Coastal Plains biome: prediction of chemical attributes via portable X-ray fluorescence (pXRF) spectrometry and robust prediction models,” *Soil Research*, vol. 58, no. 7, p. 683, 2020, doi: 10.1071/SR20136.

[76] J. Wetterlind, B. Stenberg, and M. Söderström, “Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models,” *Geoderma*, vol. 156, no. 3–4, pp. 152–160, May 2010, doi: 10.1016/j.geoderma.2010.02.012.

[77] N. Nagata, M. I. M. S. Bueno, and P. G. Peralta-Zamora, “Métodos matemáticos para correção de interferências espectrais e efeitos interelementos na análise quantitativa por fluorescência de raios-X,” *Quim Nova*, vol. 24, no. 4, pp. 531–539, Aug. 2001, doi: 10.1590/S0100-40422001000400015.

[78] L. Löwemark *et al.*, “Normalizing XRF-scanner data: A cautionary note on the interpretation of high-resolution records from organic-rich lakes,” *J Asian Earth Sci*, vol. 40, no. 6, pp. 1250–1256, Apr. 2011, doi: 10.1016/j.jseaes.2010.06.002.

[79] J. R. Bacon, O. T. Butler, W. R. L. Cairns, J. M. Cook, R. Mertz-Kraus, and Julian. F. Tyson, "Atomic Spectrometry Update – a review of advances in environmental analysis," *J Anal At Spectrom*, vol. 34, no. 1, pp. 9–58, 2019, doi: 10.1039/C8JA90044B.

CHAPTER 3. SYSTEMATIC STUDY OF TRAINING SAMPLES SET REDUCTION IN PLS SPECTRAL MODELS FOR SOIL FERTILITY ATTRIBUTES DETERMINATION

3.1 Introduction

Soil is one of the most important natural resources to sustain life. Its chemical and physical properties directly influence several factors such as climate dynamics, organisms, forests, carbon cycles, and even the economic development of countries [1], [2]. The main reason for this undeniable importance lies in the fact that soil is the most fundamental basis for food production [1]. Currently, it is estimated that around 600 million traditional wet soil analyses are carried out worldwide each year to maintain optimal planting conditions [1]–[3]. The wet methods have been used for more than 100 years and are the reference for characterizing soil fertility attributes. They are important in determining soil quality and the ability to sustain agricultural production [2]. However, this traditional approach suffers from the use of chemical reagents and is time-consuming [2], [4]. One set of fertility attributes analysis may take 3 to 15 days to delivering results, which is not adequate in the current scenario where specific soil fertility management using precision agriculture (PA) is becoming widespread [1], [5]. Furthermore, environmental concerns related to the proper disposal of waste generated and the total cost of these analyses are also important. It is estimated that there is a global consumption of 3 million L of sulfuric acid and 840 thousand kg of dichromate and ammoniacal ferrous sulfate only for organic matter analysis, generating an annual cost of approximately US$ 2.5 million [1]. These issues have led to the development of faster and cheaper methodologies based on green chemistry, which has occupied a significant space in research aimed at this cause [6].

Over the last 2 decades, many researchers have explored the use of proximal soil sensors (PSS) based on spectroscopic methods combined with machine learning (ML) algorithms, especially vis-NIR, to successfully determine various soil attributes [3], [7]–[15]. This approach is promising as a complement to local soil fertility assessment and may reduce the sample set sent for conventional wet treatments. Since the development of a ML model requires a training and a validation set, the idea is to use traditional methods just to determine the fertility attributes of the minimum number of samples required for model training with a desired level of accuracy [16]. Subsequently, only the spectral reading of new soil samples from the same study area, prepared in the same procedure as the model validation is required to estimate soil fertility attributes [16], [17]. As soil fertility needs to be assessed periodically due to the influence of agronomic practices and the dynamics of the soil-plant-atmosphere system, this proposed method may allow a rapid determination of fertility attributes with a reduced number of samples sent for conventional analysis [16], [18].

PSS based on energy dispersive X-ray fluorescence (EDXRF) have also been shown to be able to estimate soil attributes in a fast and non-destructive way, without reagent demand and waste generation [6], [19]–[28]. EDXRF is a qualitative and quantitative analytical technique for assessing chemical elements content in samples. It is founded on the photoelectric effect, *i.e.*, high energy X-rays ionize a material, the electrons ejected from an inner layer of the atom leave vacancies that are filled by higher energy electrons from the outer layers to stabilize the atom. The energy difference between the orbitals involved is manifested in the emission of an electromagnetic wave with specific energy for each element, called characteristic X-rays [6], [29], [30]. EDXRF raw spectra may be used to determine the total content of the main elements in the soil (*e.g.*, Al, Si, P, K, Ca, Mn, Fe, Cu, Zn, etc.) and as input variables for the indirect estimation of the fertility attributes by ML algorithms [6], [28].

Machine learning is a subfield of computer science that seeks to build algorithms which combine mathematical and statistical techniques with artificial intelligence (AI) [31]–[33]. The main aim of a ML model is to build a complex system that learns from a pre-defined database and establishes a mapping pattern to classify or quantify a property of interest [33]–[35]. In this process, a model that generalizes the information is created so that new data (never analyzed by the algorithm) may be accurately and reliably classified or quantified according to the desired parameter. Thus, the interaction process of a ML algorithm is divided into three phases: data organization, training (also called calibration), and validation (also called prediction or evaluation) [33], [36]–[40]. Some of the ML algorithms used with spectroscopic data from soil samples are Partial Least Squares Linear Regression (PLS), Predictive Common Dimensions Analysis (P-ComDim), Support Vector Machines (SVM), Random Forest (RF) and, Artificial Neural Networks (ANN) [3], [6], [7], [13], [15], [25]–[28], [41]–[43].

To develop robust ML models from PSS data, the training (calibration) database must be representative of the study area. Commonly, a huge number of training samples are employed [3], [16], [28], [44], [45]. Managing with such sample sets may become laborious and computationally costly, in addition increases the processing time of the ML algorithm training step. On the other hand, since results from reference methods are essential to attest to the reliability and accuracy of the ML models, reducing the number of training samples is advantageous from a financial, environmental, and logistical point of view. From this perspective, this research aimed at the systematic study of the PLS ML model performances for local estimation of soil fertility attributes (Soil Organic Carbon, Cation Exchange Capacity, Sum of Exchangeable Bases, exchangeable Mg, K and Ca) as a function of the number of training samples. The models were built using pXRF spectral data from soil samples, which were separated between training and validation sets. Then, the training set was subsequently reduced 3 times. The hypothesis to be corroborated is: reduced sets of training samples generate models with performance comparable to models with the whole training set.

3.2 Materials and methods

3.2.1 Soil sampling, study area, and conventional analyses

394 soil samples were collected at three depths (0-5 cm, 5-10 cm, and 10-20 cm) in a basalt-derived soil toposequence from an agricultural field located in the municipality of Cambé, Paraná State, Brazil (23◦09′ 59.70′ ’S, 51◦14′ 42.27′ ’W to 23◦09′ 50.14′ ’S, 51◦14′ 42.27′ ’W). The study area is managed under the no-tillage system with soybean and corn as predominant crops. The soils are classified as Red Eutroferric Latosol (Rhodic Ferralsol) at the highest altitude (between 544 and 529 m) and Red Eutroferric Nitosol (Rhodic Nitosol) at the lowest altitude (between 529 and 514 m) [46], [47]. As a result of the advanced weathering degree, these soils have high levels of iron oxides, aluminum, 1:1 clay minerals (kaolinite), and quartz [48], [49].

Dried at 40ºC for 48 hours, the samples were ground and sieved to a 2 mm granulometry. The conventional analyses were carried out following the recommendations of Pavan *et al.* (1992) [50]. Soil Organic Carbon (SOC) was quantified by dichromate oxidation method; exchangeable contents of Ca (Ca^{2+}) and Mg (Mg^{2+}) were extracted by KCl 1 mol L^{-1} solution and determined by atomic absorption spectrometry while the K exchangeable (K^{+}) content was extracted using Mehlich-1 solution and determined by flame spectroscopy. The H

+ Al content was estimated by the SMP extractor, and these results led to the Cation Exchange Capacity (CEC) and Sum of Exchangeable Bases (SB) indicators calculation (CEC =SB + H + Al and SB = $Ca^{2+} + Mg^{2+} + K^{+}$).

3.2.2 pXRF measurement

Approximately 5 g of soil samples in loose powder form were accommodated in polyethylene XRF cups (no.1530, Chemplex Industries Inc., USA) sealed with Mylar film for XRF scanning in the Tracer 5i (Bruker Inc.) handheld device. This amount of samples results in around 10 mm sample thickness, being considered an infinitely thick sample for the X-ray energy range used. The instrumental conditions used were 15 kV, 12.15 μA without primary filters. These conditions were selected in accordance with study by Ribeiro *et al.* [28] on the best experimental condition for modeling fertility attributes. The authors argue that when 15 kV is applied to the X-ray tube, the X-rays energy generated is efficient to excite the lighter elements [30] and these elements are directly correlated with the fertility attributes evaluated. The measurements were taken in triplicate, shaking the XRF cup before each measurement. Then, the average spectra of each replicate sample were calculated. The excitation was performed in air atmosphere using primary X-rays generated from the Rh tube during 30 seconds for each measurement. The equipment has a Silicon Drift Detector (SDD), and detections were recorded on 2,048 channels. The spectra were normalized by tube current and detector live time, resulting in spectra units of photons counts per second per microampere (cps μA^{-1}). Two parts of all spectra were cut. The first corresponds to a region with a signal added by the equipment manufacturers while the second corresponds to a background region without characteristic peaks. As a result, 0.98 to 15.00 keV were the final energy range, corresponding to 701 variables.

Raw spectra were utilized for modeling because they are the raw output of the XRF measurement, *i.e.*, its intrinsic error is the lowest possible compared to elementary data (concentration or intensity). The raw spectra were also used in conjunction with the fertility attributes determined by traditional methods in Pearson correlation matrix calculation to qualitatively analyze the interrelationships among these variables and gain perspective on the modeling results.

3.2.3 Data subset and training strategies

The whole dataset (n = 394) was representatively separated into an external validation (or prediction) dataset (n = 118) and a full training (or calibration) set (n = 276) in the traditional 70-30 % proportion. Then, the full training set was also reduced into three new training sets with 206, 136, and 66 samples. These set separation processes were performed in two ways to evaluate their impact on performance results: i) by the Kennard-Stone (KS) [51], [52] algorithm with mean-centered spectra, and ii) by random sample selection.

These dataset splits enabled the creation of eight training strategies for modeling: full training set split by KS (276-KS), full training set randomly separated (276-R), 206 samples set split by KS (206-KS), 206 samples set randomly separated (206-R), 136 samples set split by KS (136-KS), 136 samples set randomly separated (136-R), 66 samples set split by KS (66-KS), and 66 samples set randomly separated (66-R). In addition, to facilitate discussions, all models with sample sets separated by KS will be indicated by -KS suffix while those with randomly sample sets separated by -R . The validation set (n = 118) separated by KS was used to validate all -KS models, while the validation set (n = 118) separated randomly was used to validate all -R models. The two different methods of reducing each training set allow for a more honest assessment of model's performance, *i.e.*, assure the results reproducibility.

3.2.4 Predictive models and evaluation metrics

Spectral data of each training subset were used to build the calibration models for SOC, CEC, SB, Ca^{2+}, $K^{+,}$ and Mg^{2+} by the machine learning method of partial least square regression (PLS). It was chosen because it is a linear algorithm widely used in soil analysis by XRF data (which studies show to be linear in predicting fertility attributes) [6], [13], [21], [27], [53], [54], and because it is less cumbersome than the others (Support Vector Machine, Random Forest, etc.), having only one hyperparameter to be adjusted during the analysis, the number of latent variables [55], [56]. This contributes to optimizing the analysis in terms of computational time and cost. In addition, PLS has an advantage related to its easy interpretability, *e.g.*, easy identification of the variables that most contributed to the modeling (with the VIP score and the regression vector), outliers' detection, and the vast possibility of determining different figures of merit that assess not only overall modeling performance, but also quantification and detection limits, resolution, and linearity [57].

PLS establishes a quantitative correlation between a matrix $\boldsymbol{X}$ of independent data (spectra matrix, in this case) and a vector $\boldsymbol{y}$ of dependent data (fertility attributes) so that the maximum covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$ is reached. This is done by projecting $\boldsymbol{X}$ onto orthogonal factors, namely latent variables (LV) [27], [28], [55]. There are several algorithms used for PLS analysis, among which the most famous is NIPALS, proposed by Wold [58]. For more detailed descriptions of PLS modeling, see references [55]–[58].

According to the results by Dos Santos *et al.* [54], Poisson scaling and mean center were the pre-processings used in the PLS modeling. Poisson pre-processing weight the data according to its particular uncertainty. It is suitable because, in the XRF spectra, the statistical uncertainty in each variable is governed by Poisson statistics, where the absolute uncertainty at any given data point is larger as the number of counts represented by that point increases [28], [30], [59], [60]. Thus, all the dataset variables are equalized, reducing the impact of variables that present great magnitude and mask information in the other variables with small magnitude [28]. The optimal number of LVs was determined through 10-fold cross-validation, aiming to minimize the root mean square error of cross-validation (RMSECV). The detection and removal of outliers were based on data with extreme leverage and unmodeled residuals in spectral data, as recommended by ASTM E1655-00 [57].

The model's performance was evaluated through the root mean square error of prediction (RMSEP), determination coefficients (R^2) of prediction and the ratio of performance to deviation (RPD). The RPD is the ratio of the conventional analysis standard deviation (reference method results) to the RMSEP. As reported by Viscarra Rossel *et al.* [61] and Chang *et al.* [62]: excellent models, RPD > 2.5; very good models, 2.5 > RPD > 2.0; good model, 2.0 > RPD > 1.8; fair, 1.8 > RPD > 1.4; and very poor model, RPD <1.4. Quantitative analyzes are reliable when RPD is above 1.8.

Complementarily, the following figures of merit were also determined: accuracy by relative error in the predicted values in percentage (RE), presence of systematic errors by Bias and t_{bias} with a 95% confidence level ($\alpha = 5\%$), sensitivity (SEN), inverse of the analytical sensitivity (), limit of detection (LOD) and limit of quantification (LOQ). The definitions used follow the recommendations of the IUPAC (International Union of Pure and Applied Chemistry) and the standard E1655–0017 of the ASTM (American Society for Testing and Materials) [63], [64].

The model's linearity was assessed by Durbin-Watson (DW) test according to Ref [65], [66]. The probability associated to the DW statistics (Eq. 30) was estimated and p-values greater than p_{critic} (α=5%) indicates absence of non-linearities, is the *i*th residual and *n* is the number of training samples.

The equivalence between the full training set model and the models with reduced training sample sets was evaluated in two ways. The relative performance of the models was assessed in percentage terms by the relative improvement (RI), calculated according to Eq. 31.

the is the RMSEP from the 118 validation samples applied in the model taken as reference, in this case the model with 276 training samples. is the error from the model do be compared.

The models equivalence also was assessed by the Random Test with a 95% confidence level (α = 5%). In this test, it is assumed in the null hypothesis that the predictive accuracy of the two compared models is equivalent ($RMSEP_1 = RMSEP_2$), while the alternative hypothesis is that this predictive accuracy is not equivalent ($RMSEP_1 \neq RMSEP_2$) [65].

All pre-processing and multivariate analyses were performed using the R language in RStudio software (R version 4.2.3). The main packages used were: dplyr (data manipulation), mdatools (PLS modeling), and ggplot2 (graphical analysis).

3.3 Results and discussion

3.3.1 Conventional analyses

Table 17 shows the descriptive statistics of the conventional results. Only the CEC presented low variability by the coefficient of variation (CV < 12%) in the study area. SOC, SB, Mg^{2+}, and Ca^{2+} presented a medium variability (12% < CV <60%) while K^+ CV values are highly variable [67]. The variation of these fertility attributes in the soil are related to several factors, including the presence of different fractions of organic matter, such as fulvic acid, humic acid, and humina, which may impact the soil fertility attributes. Other factors are application of different types of fertilizers, the behavior of water and sediment losses, differences in the fertility and acidity levels, and the gradient fertility on the Ferralsoil profile, commonly observed in areas under no-tillage management, with higher fertility on surface layers (0.00–0.05 m) decreasing with an increase of the depth (0.10–0.20 m) [27], [68]–[70].

Table 17 - Descriptive statistics for the full dataset (*n*=394) from the conventional methods analysis

Attribute	SOC (g kg^{-1})	CEC (cmol$_c$ kg^{-1})	SB (cmol$_c$ kg^{-1})	Mg^{2+} (cmol$_c$ kg^{-1})	K$^+$ (cmol$_c$ kg^{-1})	Ca^{2+} (cmol$_c$ kg^{-1})
Mean + SD	20 ± 4	13.7 ± 1.5	8.3 ± 1.9	2.0 ± 0.6	0.24 ± 0.16	6.0 ± 1.5
CV (%)	22	11.3	22.7	28.5	66.78	25.2
Median	20	13.8	8.4	2.0	0.18	6.2
Minimum	8	9.3	2.1	0.7	0.05	1.4
Maximum	31	17.9	12.6	3.7	1.15	9.2
Kurtosis	-0.5	-0.3	-0.1	-0.2	3.53	-0.5
Skewness	-0.3	-0.2	-0.5	0.5	1.57	-0.4

3.3.2 pXRF analyses

Figure 24 shows the mean pXRF spectra for the full soil samples dataset. It is notable that the X-ray tube operated at 15 kV produces photons whose energies are not sufficient to properly excite the elements heavier than Fe (Z<26). The Rh scattering L-lines for the tube target are clearly identified, however, they overlap the Ar kα-line (2.958 keV), caused by the fluorescence of this element, which is present in the ambient air between the sample and the detector.

It was possible to identify the presence of Al, Si, P, S, K, Ca, Ti, Mn, Fe, Cu and Zn in which the Fe and Ti K-lines are the most intense (Figure 24). The Fe peak high intensity is related to the fact that this element is the main component that forms the Red Latosol and Red Nitosol under study [71] and to the favorable fluorescence yield for the photoelectric absorption of this element with the experimental condition used.

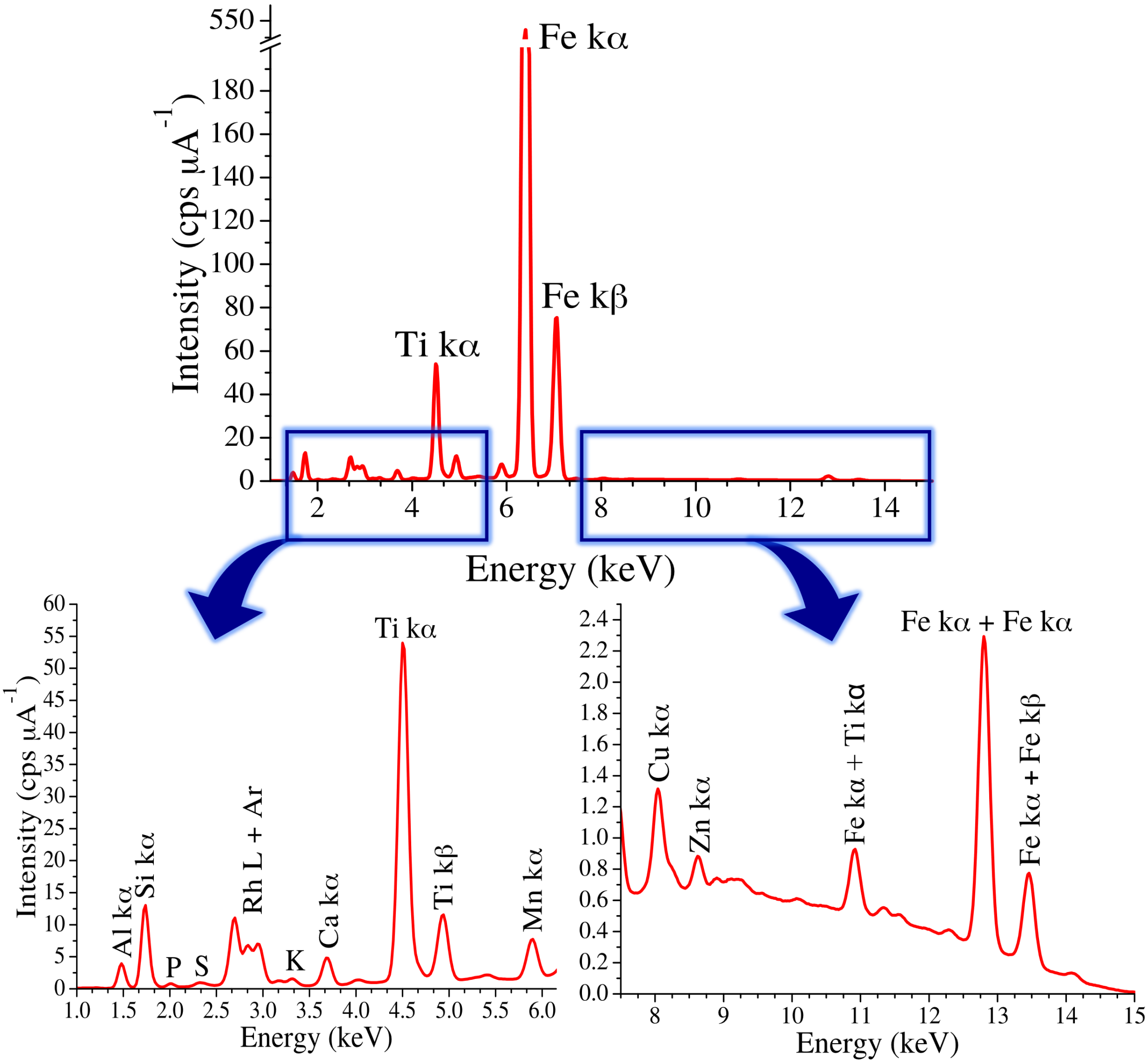

Figure 24 - Mean pXRF spectra of full soil dataset (*n*=394)

3.3.3 Correlations between fertility attributes and XRF spectra

Figure 25 shows Pearson's correlogram between the soil fertility attributes by conventional methods and pXRF spectra for the full dataset. These interrelationships aid to understand reasons of successful indirect determination of soil attributes by only spectral signatures in the modeling data, *e.g.*, predictions of SOC, CEC, and SB, which do not have emission lines in XRF spectra. Thus, the Ca emission lines is the variable most strongly correlated with all fertility attributes simultaneously. In the SOC and Mg^{2+} cases, the P and S emission lines, although very less intense, also exhibited positive correlations, while the Si emission lines exhibited negative correlations. CEC, SB and Ca^{2+} also showed positive correlations with the K and Mn emission lines and negative correlations with those of Al. K^+ was the attribute that presented the least intense correlations. In addition to Ca, it also exhibited positive correlations with K and P emissions and negative correlations with Al and Si.

The higher associations are among each soil attribute. The Ca^{2+} has the higher correlations, mainly with CEC and SB. Mg^{2+} is most correlated with SOC and SB. CEC, in addition to Ca^{2+}, is close to SB. K^{+} has the correlation coefficients closest to zero between all the soil attributes. So, K^{+} is the most independent fertility attribute from the other attributes and the pXRF data.

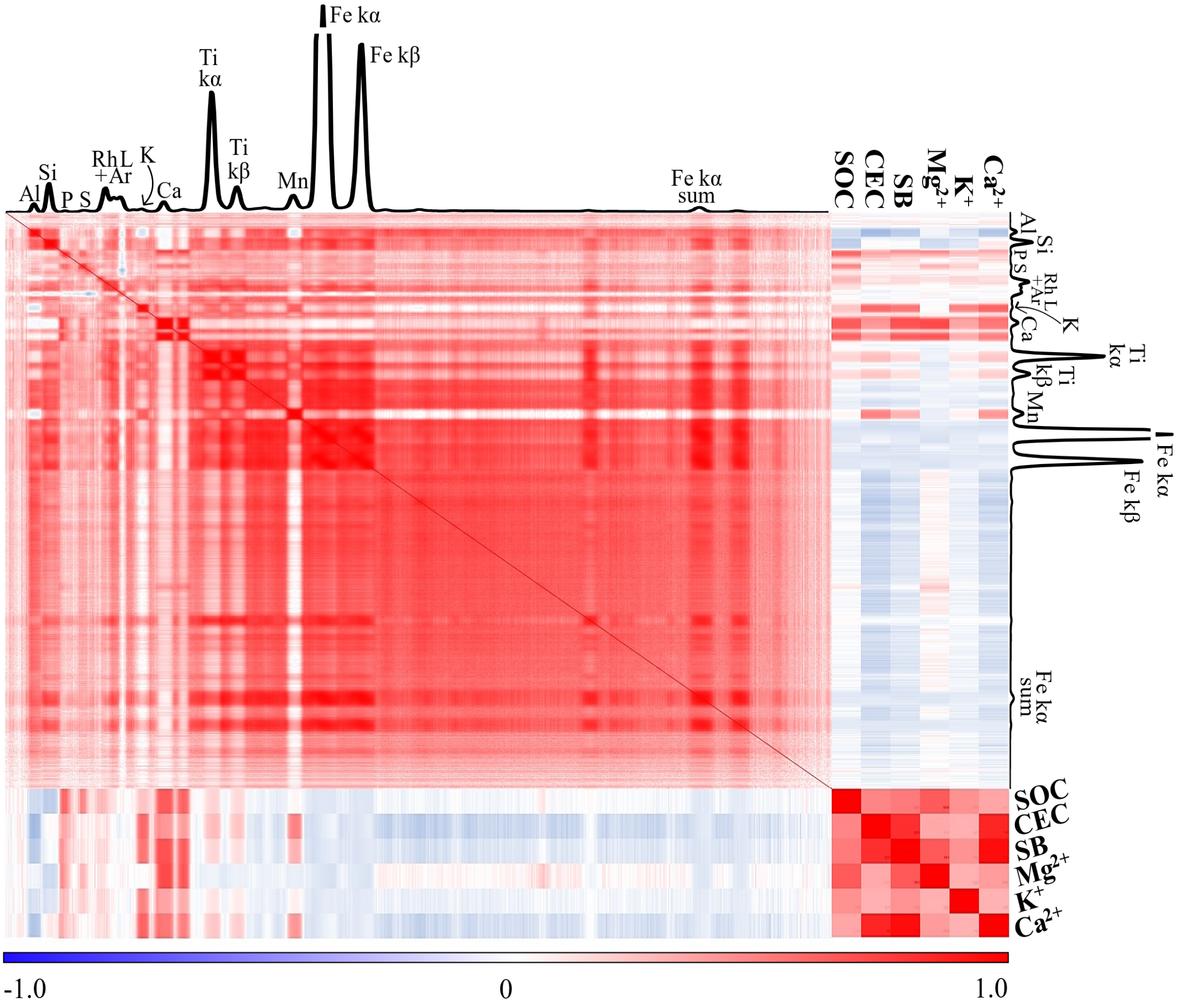

Figure 25 - Pearson correlogram among the soil fertility attributes by conventional methods and XRF spectra for the full soil dataset (*n*=394).

3.3.4 Predictive models results

The prediction results for the fertility attributes as function of the different calibration strategies are shown in Table 18. The results indicate that the models trained with 276 samples showed the best prediction performance for all attributes simultaneously, succeeded by the models trained with 206, 136, and 66 samples. In each context, both separation methods (-KS and -R) produced similar results, with a slight superiority for the models with randomly separated sets. Complete details of these models are presented in Table 19 and Table 20.

Table 18 – Prediction performance metrics of fertility attributes models

Attribute	**R² Pred**	**RMSE Pred**	**RPD**	**R² Pred**	**RMSE Pred**	**RPD**
		-KS models			**-R models**	
			276 training samples			
SOC	0.75	2.10	1.96	0.78	1.92	2.12
CEC	0.76	0.68	2.05	0.73	0.75	1.93
SB	0.77	0.85	2.08	0.81	0.80	2.28
Mg^{2+}	0.68	0.29	1.82	0.70	0.29	1.89
K^{+}	0.75	0.06	1.95	0.73	0.08	1.97
Ca^{2+}	0.76	0.66	2.03	0.80	0.64	2.23
			206 training samples			
SOC	0.75	1.91	2.03	0.77	2.00	2.09
CEC	0.73	0.65	1.94	0.74	0.77	2.00
SB	0.73	0.86	1.97	0.83	0.75	2.44
Mg^{2+}	0.68	0.27	1.82	0.70	0.30	1.83
K^{+}	0.68	0.07	1.84	0.68	0.09	1.81
Ca^{2+}	0.66	0.76	1.92	0.81	0.63	2.32
			136 training samples			
SOC	0.74	1.93	1.99	0.78	1.92	2.12
CEC	0.70	0.74	1.84	0.71	0.79	1.88
SB	0.68	0.92	1.84	0.80	0.82	2.23
Mg^{2+}	0.70	0.26	1.84	0.65	0.34	1.71
K^{+}	0.53	0.08	1.69	0.65	0.09	1.70
Ca^{2+}	0.73	0.73	1.97	0.78	0.68	2.15
			66 training samples			
SOC	0.67	2.04	1.76	0.75	2.03	2.16
CEC	0.68	0.70	1.90	0.67	0.85	1.82
SB	0.68	0.89	1.81	0.77	0.84	2.10
Mg^{2+}	0.64	0.29	1.69	0.58	0.36	1.56
K^{+}	0.33	0.10	1.58	0.59	0.10	1.57
Ca^{2+}	0.67	0.75	1.84	0.76	0.70	2.09

SOC unit, g kg^{-1}; CEC, SB, Mg^{2+}, K^{+}, Ca^{2+} units, $cmol_c kg^{-1}$; -KS, models with samples set separated by KS; -R, models with randomly separated sample sets; Pred, prediction; KS, Kennard-Stone

The fertility attributes estimated with the higher accuracy were SOC, CEC, SB and, Ca^{2+}. For SOC predictions, the models with 276, 206, and 136 training samples separated by KS showed similar results, with 1.96 RPDs values (good model), 2.03 (very good model), and 1.99 (good model), respectively, considered adequate for quantitative analysis. The SOC 66-KS model had the highest RMSEP and the lowest R^2 and RPD than the others -KS SOC models, classified as fair. Conversely, all SOC -R models showed RPD greater than 2.00, being classified as very good models for quantitative analysis. Independent of the separation method, all models for the CEC, SB, and Ca^{2+} predictions presented RPD greater than 1.8 (suitable for quantitative analysis), with the highest values for the 276-R and 206-R models. The 276-KS, 276-R, 206-KS, 206-R, and 136-KS Mg^{2+} models presented similar results (1.8 RDP value). The remaining Mg^{2+} models had an RPD less than 1.8 (unsuitable for quantitative analysis) and, together with the K^+ models, they had the lowest performance. The K^+ models presented the worst performance results compared to the other attributes. Only the 276 and 206 models (independent of the separation method, which presented similar prediction results) were considered good for predictions. These findings support establishment of the methodology, indicating that the pXRF raw spectra data modeled with PLS generate local models capable of predicting fertility attributes in the context of this study with good performance and confidence in R^2, RMSEP and RPD terms [61], [62]. Similar results have been reported in studies similar methods to estimate attributes from the XRF spectrum [6], [20], [27], [28], [72].

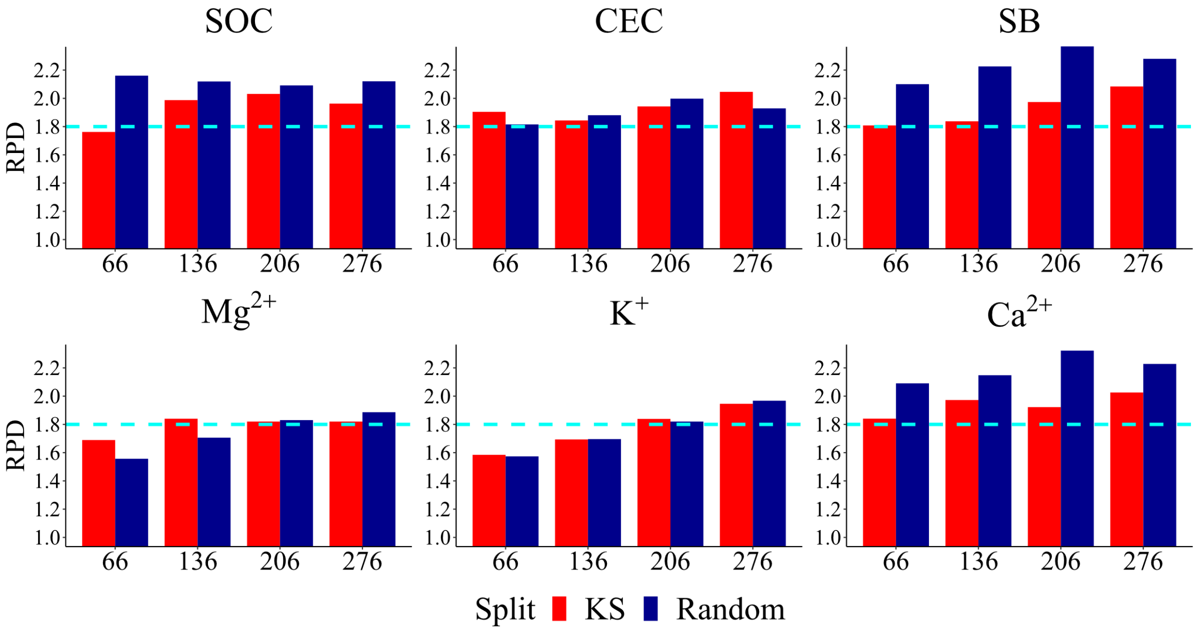

Figure 26 - RPD values for soil attribute prediction models as a function of the 276, 206, 136, and 66 calibration sets. Values higher than 1.8 line classify the model as suitable for quantitative analysis [61], [62]

Figure 26 shows a RPD values comparison for all models. It is noted that the models with 276, 206, and 136 training samples presented adequate performance for quantitative analysis in most cases, independent of the split set methods. As the 276 training sample models are the best, these results highlight a lower limit on samples number for the training step without significant performance loss in RPD terms, which may be between 206 and 136 samples. However, the analysis over RPD alone is not sufficient to completely evaluate the success of modeling with reduced training sets. It is a metric for assessing modeling accuracy. The presence of systematic errors, linearity, resolution, and detection and quantification limits of the models still need to be verified. These figures of merit will allow estimation of the real minimum number of training samples required to quantify each fertility attribute.

Table 19 - Detailed results of PLS modeling of fertility attributes with sets separated by KS.

Models	LV	R^2 Training	R^2 CV	R^2 Pred	RMSE Training	RMSE CV	RMSE Pred	RPD
				276-KS				
SOC	7	0.84	0.73	0.75	1.73	2.26	2.05	1.96
CEC	6	0.76	0.64	0.76	0.75	0.93	0.68	2.05
SB	6	0.85	0.8	0.77	0.72	0.83	0.85	2.08
Mg^{2+}	4	0.69	0.65	0.68	0.31	0.34	0.29	1.82
K^+	8	0.86	0.58	0.75	0.05	0.09	0.06	1.95
Ca^{2+}	7	0.88	0.8	0.76	0.5	0.65	0.66	2.03
				206-KS				
SOC	6	0.77	0.68	0.75	2.11	2.48	1.91	2.03
CEC	6	0.73	0.61	0.73	0.78	0.93	0.65	1.94
SB	6	0.83	0.77	0.73	0.77	0.89	0.86	1.97
Mg^{2+}	5	0.75	0.66	0.68	0.28	0.33	0.27	1.82
K^+	7	0.79	0.52	0.68	0.06	0.1	0.07	1.84
Ca^{2+}	6	0.8	0.74	0.73	0.67	0.77	0.73	1.92
				136-KS				
SOC	6	0.83	0.72	0.74	1.93	2.53	1.93	1.99
CEC	6	0.78	0.58	0.7	0.71	0.98	0.74	1.84
SB	6	0.81	0.72	0.68	0.76	0.92	0.92	1.84
Mg^{2+}	5	0.75	0.62	0.7	0.29	0.36	0.26	1.84
K^+	8	0.88	0.37	0.53	0.05	0.11	0.08	1.69
Ca^{2+}	6	0.79	0.69	0.66	0.67	0.81	0.76	1.97
				66-KS				
SOC	6	0.81	0.49	0.67	1.72	2.81	2.04	1.76
CEC	6	0.88	0.57	0.68	0.54	1.03	0.7	1.9
SB	6	0.89	0.69	0.68	0.61	1.02	0.89	1.81

Mg^{2+}	5	0.86	0.62	0.64	0.22	0.37	0.29	1.69
K^{+}	7	0.86	0.36	0.33	0.05	0.12	0.1	1.58
Ca^{2+}	6	0.87	0.63	0.67	0.54	0.91	0.75	1.84

Pred=prediction or validation; LV=latent variables; CV=cross-validation. SOC unit = g kg^{-1}, CEC, SB, Mg^{2+}, K^{+}, and Ca^{2+} units = $cmol_c kg^{-1}$.

Table 20 - Detailed results of PLS modeling of fertility attributes with sets randomly separated.

Models	**LV**	**R^2 Training**	**R^2 CV**	**R^2 Pred**	**RMSE Training**	**RMSE CV**	**RMSE Pred**	**RPD**
				276-R				
SOC	5	0.76	0.7	0.78	2.11	2.36	1.92	2.12
CEC	7	0.85	0.57	0.73	0.6	1	0.75	1.93
SB	6	0.78	0.73	0.81	0.85	0.94	0.8	2.28
Mg^{2+}	5	0.74	0.68	0.7	0.28	0.31	0.29	1.89
K^{+}	8	0.83	0.55	0.73	0.06	0.1	0.08	1.97
Ca^{2+}	6	0.82	0.77	0.8	0.65	0.72	0.64	2.23
				206-R				
SOC	5	0.79	0.72	0.77	1.94	2.24	2	2.09
CEC	6	0.74	0.66	0.74	0.74	0.85	0.77	2
SB	6	0.85	0.8	0.83	0.75	0.85	0.75	2.44
Mg^{2+}	5	0.75	0.67	0.7	0.28	0.33	0.3	1.83
K^{+}	7	0.74	0.43	0.67	0.07	0.1	0.08	1.82
Ca^{2+}	6	0.85	0.8	0.81	0.61	0.7	0.63	2.32
				136-R				
SOC	6	0.84	0.72	0.78	1.82	2.4	1.92	2.12
CEC	6	0.73	0.56	0.71	0.77	0.98	0.79	1.88
SB	6	0.81	0.72	0.8	0.85	1.02	0.82	2.23
Mg^{2+}	4	0.62	0.54	0.65	0.36	0.4	0.34	1.71
K^{+}	6	0.65	0.33	0.65	0.08	0.11	0.09	1.7
Ca^{2+}	6	0.8	0.71	0.78	0.71	0.87	0.68	2.15
				66-R				
SOC	5	0.88	0.51	0.75	1.48	2.99	2.03	2.16
CEC	3	0.69	0.59	0.67	0.74	0.85	0.85	1.82
SB	4	0.74	0.62	0.77	1.05	1.29	0.84	2.1
Mg^{2+}	3	0.72	0.66	0.58	0.34	0.38	0.36	1.56
K^{+}	6	0.79	0.34	0.59	0.09	0.15	0.1	1.57
Ca^{2+}	6	0.83	0.53	0.76	0.64	1.07	0.7	2.09

Pred=prediction or validation; LV=latent variables; CV=cross-validation. SOC unit = g kg^{-1}, CEC, SB, Mg^{2+}, K^{+}, and Ca^{2+} units = $cmol_c kg^{-1}$.

3.3.5 Figures of merit

The scatter plots in Figure 27 and Figure 28 show the correlations between reference values and prediction results for the -R and -KS models, respectively. It is visually notable that the reduction of the training sample numbers leads to a decrease in models' performance, even if the appropriate number of LV is selected. It is more evident especially with 66 training sample models, which do not have a robust calibration to accurately quantify the same validation set (such as Mg^{2+} and K^{+}), predicting samples with values that are further away from the 1:1 line. In addition, both figures also show the sensitivity (SEN), the inverse of the analytical sensitivity (γ^{-1}), p-value of the Durbin-Watson test (DW p-value) and the relative error in the predicted values in percentage (RE%). The SEN determines the fraction of the signal responsible for adding a concentration unit of fertility attributes while γ^{-1} provides an estimate of the minimum difference that is discernible by the model considering random experimental noise as the only error source [57]. So, although there may be some fluctuations, the SEN and γ^{-1} values of -R and -KS models are close when compared with respect to the same training samples number. Since these parameters are directly related to the method resolution [57], these results indicate that the use of both strategies to separate sample sets in modeling is similar in this terms. Besides, only the -R models presented DW p-values lower than the p_{critic} (α=5%) in some cases, they were: 136 training sample models for the K^{+} and Ca^{2+} estimations, and 66 training sample model for the K^{+} estimation. They presented a lack of linearity, indicating that the PLS methodology is not suitable for quantifying these attributes with these sets of measurements. Due to the lack of linearity, the K^{+} models exhibited the highest RE(%) among all others, being close to 30% regardless of the sample sets separation methods. The RE(%) values of the -R and -KS models are similar and, in most cases, increased with the decrease in the training sample numbers. Mg^{2+} models presented the second highest RE values (14.5% on average). The other fertility attributes models presented lower RE(%) values, below 12.38%.

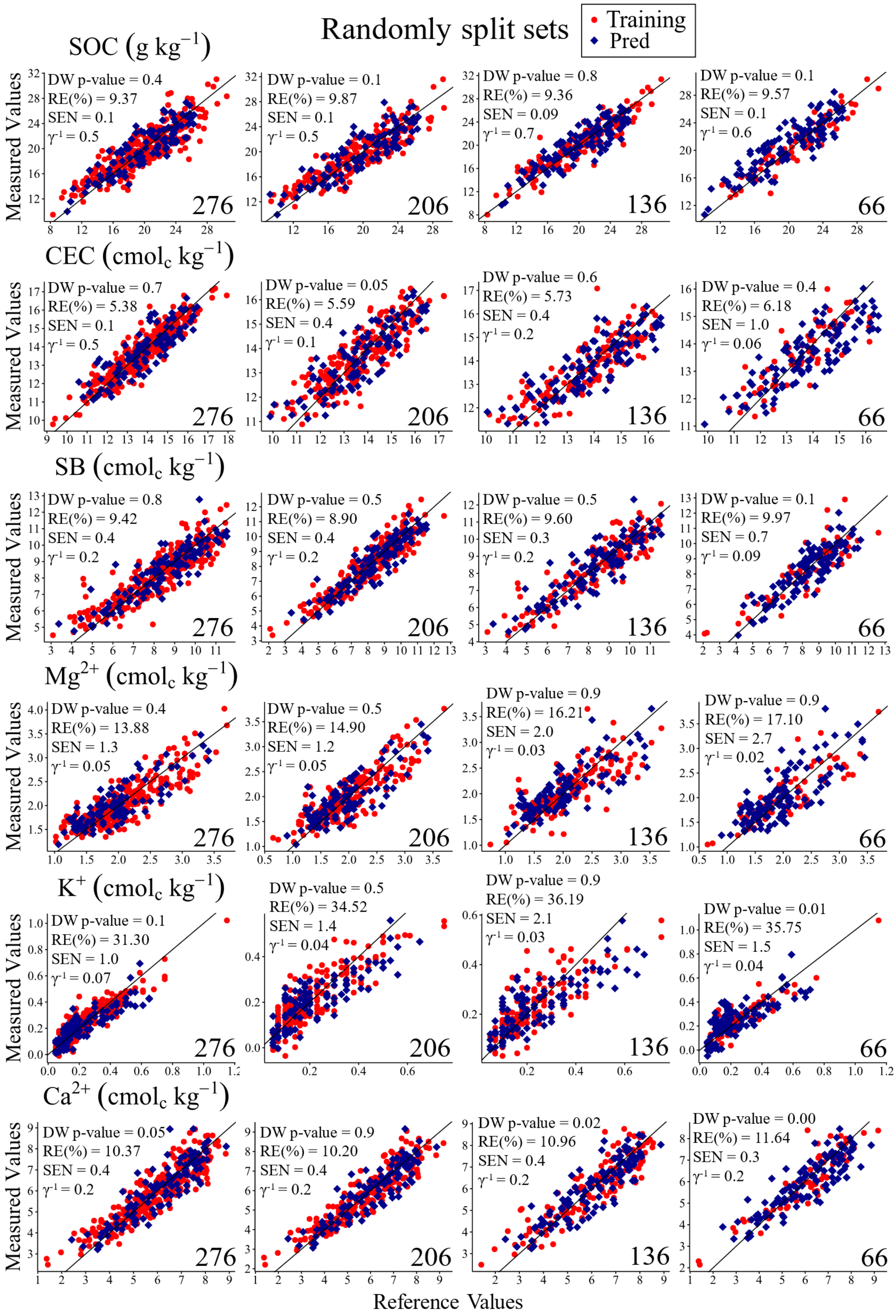

Figure 27 – PLS models scatter plots of reference values versus predicted values and figures of merit from the different training sample sets. Split sets by random choice. Pred, prediction or validation samples; DW, Durbin-Watson Test; p_{critic}=0.05 (α=5%); RE(%), relative error in the predicted values in percentage; SEN, sensitivity, γ^{1}, inverse of the analytical sensitivity

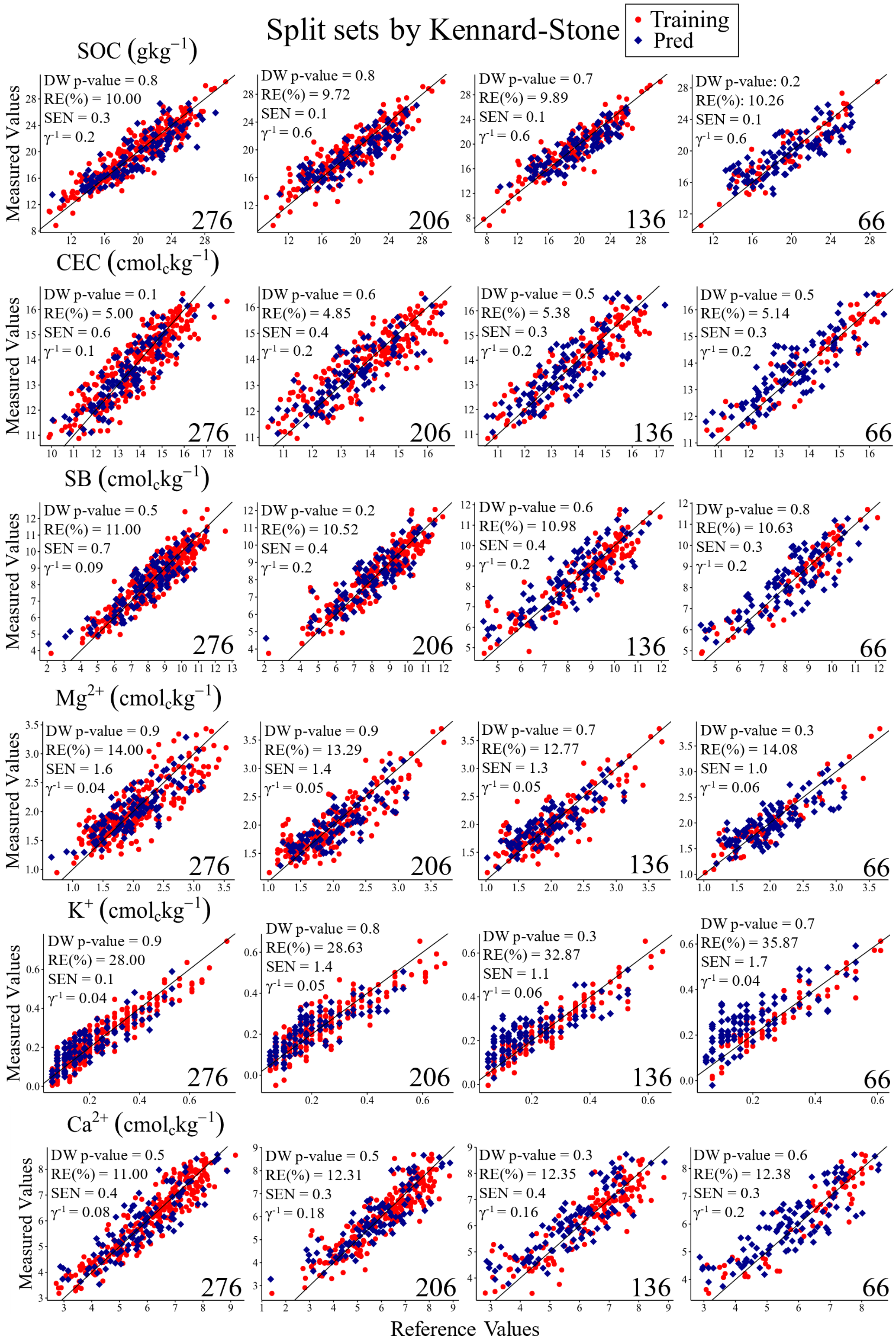

Figure 28 - PLS models scatter plots of reference values versus predicted values and figures of merit from the different training sample sets. Split sets by Kennard-Stone. Pred, prediction or validation samples; DW, Durbin-Watson Test; p_{critic}=0.05 (α=5%); RE(%), relative error in the predicted values in percentage; SEN, sensitivity, γ^{-1}, inverse of the analytical sensitivity

Table 21 – Some figures of merit for all PLS models.

	Pred Bias	t_{Bias}	LOD	LOQ	Pred Bias	t_{Bias}	LOD	LOQ
	-KS models				**-R models**			
				276 training samples				
SOC	-0.37	1.95	0.60	1.83	-0.02	0.13	1.61	4.89
CEC	0.03	0.50	0.33	0.99	0.04	0.50	1.65	5.00
SB	0.07	0.85	0.29	0.89	-0.03	0.38	0.56	1.71
Mg^{2+}	-0.03	0.93	0.38	0.13	-0.06	1.97	0.15	0.46
K^+	-0.01	**2.46**	0.40	0.13	0.01	1.83	0.22	0.65
Ca^{2+}	0.05	0.65	0.27	0.80	0.00	0.07	0.57	1.71
				206 training samples				
SOC	0.29	1.52	1.89	5.72	0.07	0.38	1.70	5.16
CEC	-0.08	1.15	0.57	1.72	0.08	1.07	0.49	1.48
SB	-0.16	1.86	0.57	1.73	0.09	1.25	0.57	1.72
Mg^{2+}	0.04	1.57	0.15	0.46	-0.01	0.39	0.17	0.50
K^+	-0.02	**2.60**	0.15	0.45	0.02	**2.19**	0.14	0.44
Ca^{2+}	-0.11	1.52	0.59	1.80	0.08	1.29	0.58	1.75
				136 training samples				
SOC	0.23	1.19	2.03	6.14	0.00	0.02	2.36	7.14
CEC	-0.07	1.01	0.68	2.05	0.06	0.81	0.55	1.67
SB	-0.23	**2.58**	0.55	1.65	-0.06	0.80	0.61	1.85
Mg^{2+}	-0.03	0.95	0.15	0.47	-0.04	1.39	0.10	0.31
K^+	-0.04	**5.85**	0.19	0.57	0.01	1.09	0.10	0.30
Ca^{2+}	-0.34	**4.89**	0.52	1.59	0.02	0.36	0.58	1.77
				66 training samples				
SOC	0.03	0.13	2.13	6.44	-0.75	**4.20**	2.05	6.21
CEC	-0.24	**3.65**	0.79	2.38	0.24	**3.08**	0.19	0.58
SB	-0.14	1.60	0.79	2.39	0.11	1.43	0.28	0.86
Mg^{2+}	0.02	0.60	0.21	0.64	-0.04	1.32	0.08	0.23
K^+	-0.06	**8.30**	0.12	0.37	-0.01	1.02	0.14	0.41
Ca^{2+}	-0.23	**3.27**	0.66	1.99	0.11	1.60	0.72	2.19

t_{critic} = 1.98 (α=5%); SOC unit = g kg^{-1}, CEC, SB, Mg^{2+}, K^+, and Ca^{2+} units = $cmol_c$ kg^{-1}; -KS, models with samples set separated by KS; -R, models with randomly separated sample sets; Pred, prediction; Bold values indicate the significant presence of Bias.

The remaining figures of merit are presented in Table 21. All K^+ models exhibited LOQ and LOD values higher than the minimum concentration obtained by the reference analysis of 0.05 $cmol_c$ kg^{-1} (Table 17). Considering Ca^{2+}, only the 276-KS model had a LOQ value below the conventional analysis minimum (1.4 $cmol_c$ kg^{-1}, Table 17). However, the difference between the LOQ and this minimum for other Ca^{2+} models was not large (a maximum of 0.80 $cmol_c$ kg^{-1} in the 66-R model case). Apart from these attributes, all the other models classified as good or very good for quantitative analysis (RPD>1.8 and RPD>2.0 [61], [62], respectively) presented LOQs below the minimum value established by the reference analysis (Table 17). Meanwhile, all models classified as suitable for qualitative analysis (RPD<1.8 [61], [62]) had LODs below the minimum value established by the traditional wet analysis (Table 17).

Having the worst results among the estimated fertility attributes, the K^+ estimation showed significant presence of Bias ($t_{bias} > t_{critic}$) in all -KS models plus the 206-R models (Table 21). A possible cause of these negative results may be associated with the low intensity of the Potassium characteristic X-ray line in the pXRF spectrum (Figure 24), which is the XRF variable that most correlated with K^+ (Figure 25). As a result, the PLS regression algorithm seeks correlations among the characteristic X-rays of other elements to estimate soil K^+. However, the K^+ Pearson coefficients exhibit the weakest correlations with the XRF data among all the fertility attributes (Figure 25), which may underestimate or overestimate their quantification by PLS, generating poor results. Additionally, this issue may also explain the large RE(%) values and lack of linearity of some K^+ quantification models.

Except for K^+, significant prediction Bias was observed only in some attributes of the models with 136 and 66 training samples (Table 21). This result also highlights, as revealed by the RPD, that the interval between 206 and 136 samples contains the maximum limit for sample reduction without significant losses, now in prediction Bias terms. Thus, it remains to define a criterion to determine the ideal reduced sample set for each attribute.

3.3.6 Equivalence between full training sample set models and reduced training sample set models

The criteria of RPD > 1.8, $t_{Bias} < t_{critic}$, and DW p-value > p_{critic} regardless of the separation method used (-KS or -R) were used to determine the minimum sample set for each attribute. These criteria lead to models with adequate performance for quantitative analysis, not biased (α=5%) and with uncorrelated residuals (α=5%). Consequently, the sets with 136 training samples were considered suitable for the soil SOC and CEC estimation, while the sets with 206 training samples were suitable for soil SB, Mg^{2+} and Ca^{2+} estimation (see Table 18, Table 21, Figure 27, and Figure 28). These reduced sets were considered as the best reduced training sample sets. Conversely, since there was a significant prediction Bias in the complete training sample set separated by KS (Table 21), the sample set reduction for K^+ estimation was not considered sufficiently reliable.

So, except for K^+, the equivalence between models with full and best reduced training sample sets was evaluated qualitatively by box plots with the prediction results for each model (Figure 29) and quantitatively by random test (α=5%) and the relative improvement taking the model with full calibration set samples as a reference (Table 22, Table 23, respectively).

The box plots (Figure 29) presents the median symmetrically positioned between the 3rd and 4th quartiles for most attributes, with slight fluctuations. Additionally, the average values being close to the median, and the equal length of the tails (ignoring outliers) corroborate the equivalence between the models evaluated.

The random test p-values comparing the models with full and best reduced training sample sets (Table 22) provide two instances to assess their equivalence. First, all models trained with reduced training samples are equivalent to those trained with the full samples (α=5%), in both separation method used (KS and random).

In the second instance, the results show that the two methods used to separate the sample sets produce results of equivalent performance (α=5%) in the context of comparing models with the same number of training samples (Table 22). These findings provide a more honest assessment of the positive results robustness that indicated the feasibility of quantitative analysis of models for estimating SOC, CEC, SB, Mg^{2+}, and Ca^{2+} attributes, *i.e.*, regardless of the separation method, models with the same number of training samples showed similar performance to estimate the same fertility attributes and were considered equivalent by the random test.

Table 22 - Random Test p-values comparing attributes estimated by full and best reduced training samples models. p_{critic} = 0.05 (α=5%)

	SOC	**CEC**		**SB**	**Mg^{2+}**	**Ca^{2+}**
	276-KS				**276-KS**	
136-KS	0.5	0.8	**206-KS**	0.3	0.3	0.06
	276-R				**276-R**	
136-R	0.5	0.8	**206-R**	0.2	0.2	0.4
	276-KS				**276-KS**	

276-R	0.4	0.2	**276-R**	0.3	0.5	0.5
	136-KS				**206-KS**	
136-R	0.4	0.2	**206-R**	0.1	0.2	0.1

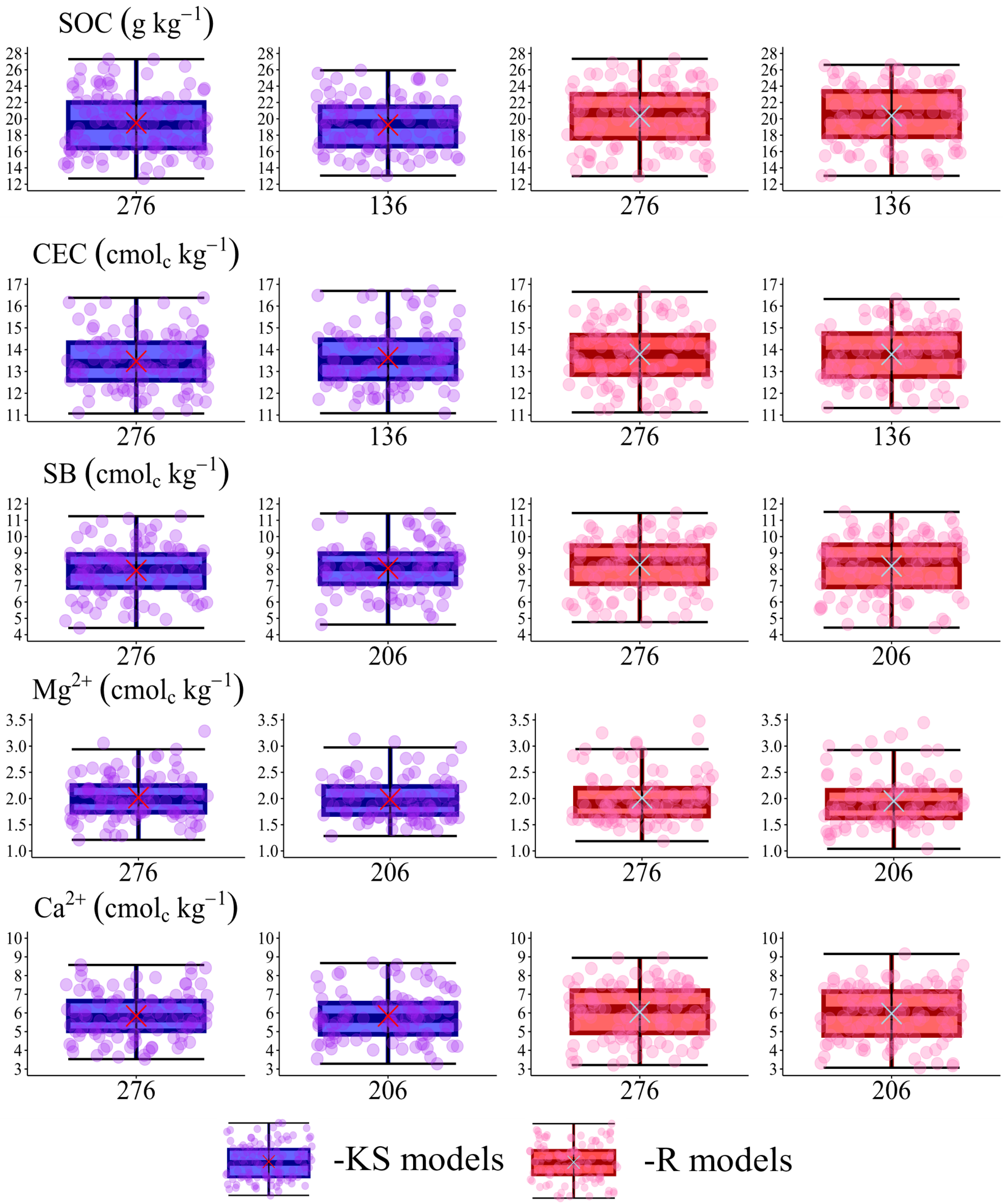

Figure 29 - Boxplots with prediction results of models with full and best reduced training sample sets

Relative Improvement (RI) also comparing the models with full and best reduced training sample sets varied between positive and negative values (ranging from -17.3 to 15.4%), with no apparent trend (Table 23). This pattern was verified in comparisons between models with the same number of training samples but with different separation methods (KS and Random), and in comparisons between models with full vs. best reduced training samples, using the same method for separating the sets.

Table 23 – Relative Improvement (%) comparing attributes estimated by full and best reduced training samples models.

	SOC	**CEC**		**SB**	**Mg^{2+}**	**Ca^{2+}**
	276-KS (reference)			**276-KS (reference)**		
136-KS	8.2	-8.4	**206-KS**	-1.4	6.4	-15.4
	276-R (reference)			**276-R (reference)**		
136-R	-0.001	-5.3	**206-R**	6.3	-3.4	1.6
	276-KS (reference)			**276-KS (reference)**		
276-R	8.6	-10.3	**276-R**	6.0	-11.5	3.0
	136-KS (reference)			**206-KS (reference)**		
136-R	0.4	-7.2	**206-R**	13.1	-11.5	17.3

Overall, the results of this study demonstrate the feasibility of decreasing the calibration samples number used in machine learning models maintaining equivalent performance to predict fertility attributes of 118 samples. The study hypothesis was corroborated by equivalence between the models with full and best reduced training sample sets attested by the random test for all soil attributes, except K^{+}. Employing a reduced number of calibration (training) samples is advantageous from a financial and environmental perspectives. This is owing to the decrease in traditional analysis-related expenses, waste, and reagent usage associated with fewer samples. Additionally, decreasing the amount of samples may minimize the time needed to conduct the analysis. This is significant in scenarios where fast results are desirable.

The 206 training samples models performed as well as the models with full training set for SB, Mg^{2+} and Ca^{2+} estimation while the 136 training sample models for estimating SOC and CEC. Although the performance of the reduced training samples models for some fertility attributes was slightly lower than the full set models in terms of RI, both models showed RPD results that classified them as suitable for quantitative estimation of this fertility attributes [61], [62]. Therefore, this performance loss in some cases may be advantageous depending on the level of accuracy the researcher expects from the results. The advantage of fewer samples may outweigh the loss in performance, especially if the goal of the study is to obtain results with an acceptable level of accuracy, but without the need for high precision.

3.3.7 Variable influence on projection (VIP) Score

One way to evaluate the spectral variables with higher importance in the estimated fertility attributes in PLS models are the informative vector called Variable Influence on Projection (VIP) scores. Figure 30 presents the VIP scores for the models with full and best reduced training sample sets. In general, the most important variables for modeling were the same in the models for each fertility attribute, regardless of the sample separation method (-KS or -R models) and the number of training samples. Nonetheless, slight fluctuations in the intensity values of the variables were observed, especially for elements lighter than Ca (energies below 3.692 keV). VIP scores for the remaining models may be found in Figure 31.

Most of the correlations between soil attributes and spectral data exhibited in the Pearson correlogram (Figure 25) are reproduced in the VIP scores. Ca-kα characteristic X-ray peak was the most important variable in SOC, SB, Ca^{2+}, and Mg^{2+} estimation. CEC had Ca-kα, Ti- kα and Mn-kα as the most important variables, with similar intensities. K^+ had the Potassium emission K lines as the most significant variables for its estimation. Apart from them, several other variables showed similar importance for K^+ estimation, such as Al, Si, Ca, Ti, Fe, and even the L line of Rh scattering, which comes from the scattering of light elements (such as C, O, and H) with photons from the X-ray tube [30], [73]. However, the correlations between K^+ and these elements showed a very poor Pearson coefficient value (Figure 25). This result supports the discussion about the negative results of K^+ prediction in section 3.6.

On the other hand, the high importance of Ca XRF peaks in the estimation of most fertility attributes evaluated may be related to its direct or indirect involvement with these soil parameters. The Ca characteristic X-ray is proportional to its total concentration and is also associated with the exchangeable Ca content (Ca^{2+}). Ca^{2+} is used in the calculations of CEC (CEC =SB + H^++Al^{3+} and SB = Ca^{2+} + Mg^{2+} + K^+) and, because CEC is related to the ability of soil to exchange cations with the soil solution, it affects the availability of nutrients to plants, including Mg^{2+}. Other studies performed with different tropical Brazilian soils have also identified Ca as the most important variable for soil fertility estimation [25], [28], [43], [53], [74]–[76].

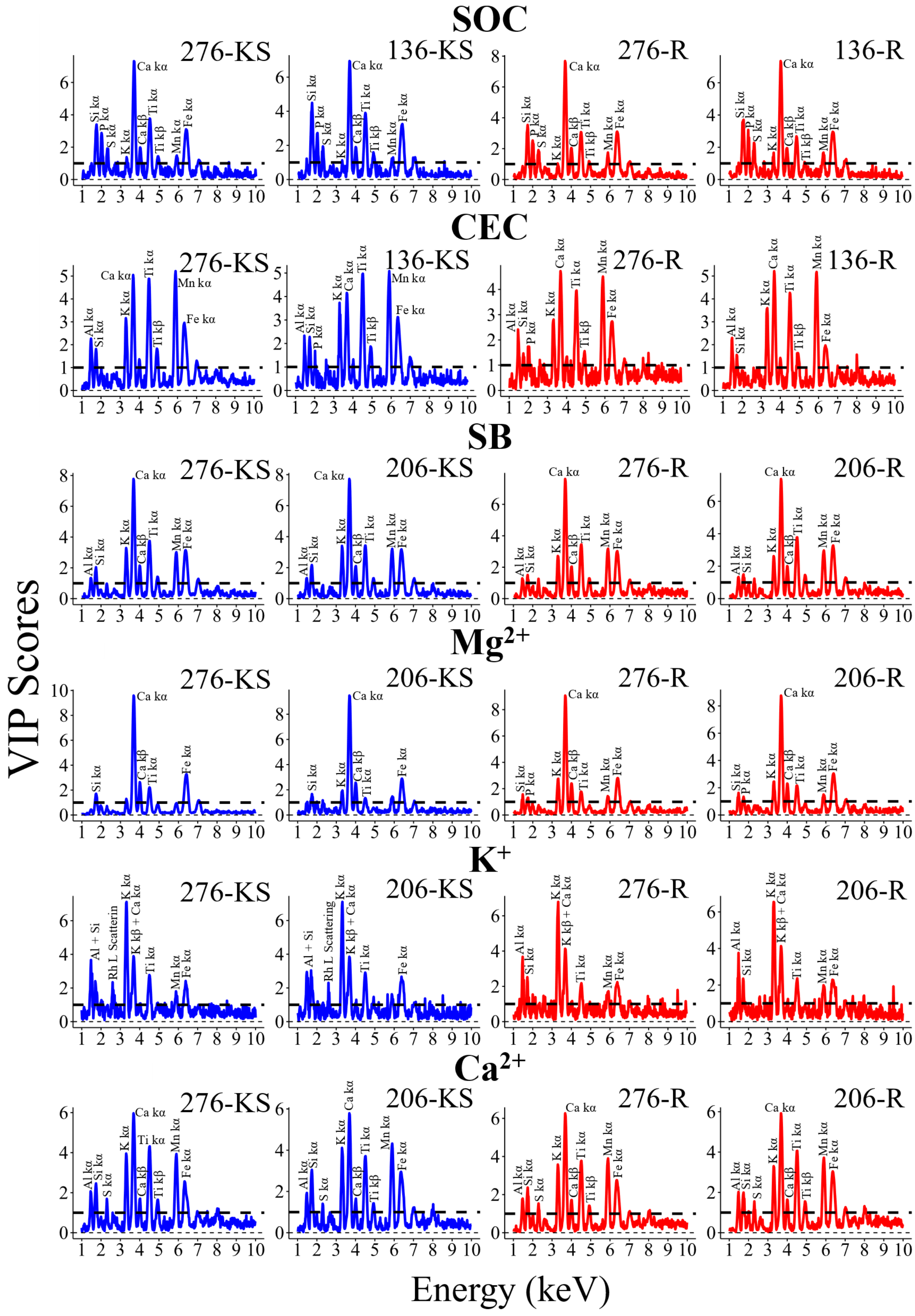

Figure 30 - Variable influence on projection (VIP) scores of the full and best reduced training samples models. Blue and red lines indicate -KS and -R models, respectively. Values greater than unity (black dotted line) indicate variables that significantly contributed to modeling.

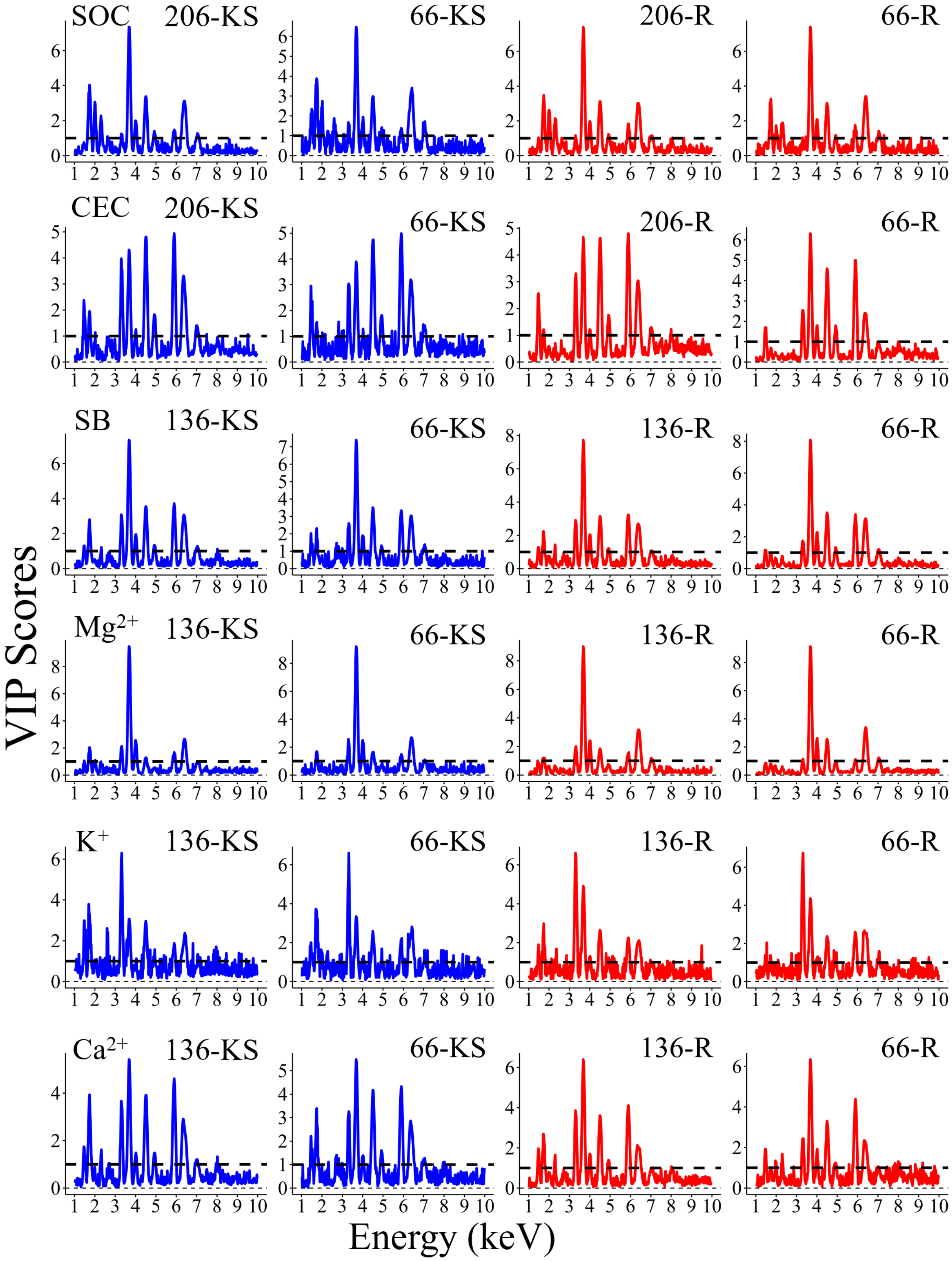

Figure 31 - Variable influence on projection (VIP) scores of the 136 and 66 training samples models. Blue and red lines indicate -KS and -R models, respectively. Values greater than unity (black dotted line) indicate variables that significantly contributed to modeling.

3.4 Conclusion

The results demonstrated that it is possible to reduce the number of training samples while maintaining equivalent performance in PLS models with XRF spectral data. The 136-sample set was ideal for estimating SOC and CEC, while the 206-sample set was best for estimating SB, Mg^{2+} and Ca^{2+}. Both reduced models were considered good or very good for predictive quantitative analysis.

It was not possible to reduce the training set of samples for the indirect quantification of K^+. Although the K^+ model with the full data set showed an RPD > 1.8 (good for predictive quantitative analysis), the models with samples separated by the Kennard-Stone algorithm showed a significant Bias ($\alpha = 5\%$). These results highlight the challenges involved in estimating K^+ from the XRF spectrum, which often presents a low Potassium peak/background ratio and makes modeling difficult. On the other hand, the remaining attributes showed similar performance regarding the methods used to separate the sample sets (Kennard-Stone and Random). This allowed a more honest assessment of the reproducibility results.

Overall, the present study provided subsidies regarding the choice of pXRF training sample set size in soil fertility attributes PLS modeling. Although models with reduced training sets show some performance loss compared to those with a full set in RI terms, this loss may be advantageous depending on the required accuracy results. The fewer samples use may outweigh this loss because the reduction in the number of samples analyzed by conventional wet methods is advantageous from a financial, logistical, and environmental point of view. Therefore, these findings allow bringing XRF closer to local soil fertility mapping in the precision agriculture context. Researchers are encouraged to reduce the training sample set to optimize the use of XRF. This method may accelerate local soil analysis by reducing measurement time and processing power.

3.5 References

[1] J. A. M. Demattê, A. C. Dotto, L. G. Bedin, V. M. Sayão, and A. B. e Souza, “Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact,” *Geoderma*, vol. 337, pp. 111–121, Mar. 2019, doi: 10.1016/j.geoderma.2018.09.010.

[2] J. A. M. Demattê *et al.*, “The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges,” *Geoderma*, vol. 354, p. 113793, Nov. 2019, doi: 10.1016/j.geoderma.2019.05.043.

[3] F. B. de Santana, S. K. Otani, A. M. de Souza, and R. J. Poppi, “Comparison of PLS and SVM models for soil organic matter and particle size using vis-NIR spectral libraries,” *Geoderma Regional*, vol. 27, p. e00436, Dec. 2021, doi: 10.1016/j.geodrs.2021.e00436.

[4] R. A. Viscarra Rossel *et al.*, “A global spectral library to characterize the world’s soil,” *Earth Sci Rev*, vol. 155, pp. 198–230, Apr. 2016, doi: 10.1016/j.earscirev.2016.01.012.

[5] R. Gebbers and V. I. Adamchuk, “Precision Agriculture and Food Security,” *Science (1979)*, vol. 327, no. 5967, pp. 828–831, Feb. 2010, doi: 10.1126/science.1183899.

[6] F. R. dos Santos, J. F. de Oliveira, E. Bona, J. V. F. dos Santos, G. M. C. Barboza, and F. L. Melquiades, “EDXRF spectral data combined with PLSR to determine some soil fertility indicators,” *Microchemical Journal*, vol. 152, p. 104275, Jan. 2020, doi: 10.1016/j.microc.2019.104275.

[7] F. B. de Santana, A. M. de Souza, and R. J. Poppi, “Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters,” *Spectrochim Acta A Mol Biomol Spectrosc*, vol. 191, pp. 454–462, Feb. 2018, doi: 10.1016/j.saa.2017.10.052.

[8] R. A. Viscarra Rossel, S. R. Cattle, A. Ortega, and Y. Fouad, “In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy,” *Geoderma*, vol. 150, no. 3–4, pp. 253–266, May 2009, doi: 10.1016/j.geoderma.2009.01.025.

[9] R. Zornoza, C. Guerrero, J. Mataix-Solera, K. M. Scow, V. Arcenegui, and J. Mataix-Beneyto, “Near infrared spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils,” *Soil Biol Biochem*, vol. 40, no. 7, pp. 1923–1930, Jul. 2008, doi: 10.1016/j.soilbio.2008.04.003.

[10] T. H. Waiser, C. L. S. Morgan, D. J. Brown, and C. T. Hallmark, “In Situ Characterization of Soil Clay Content with Visible Near Infrared Diffuse Reflectance Spectroscopy,” *Soil Science Society of America Journal*, vol. 71, no. 2, pp. 389–396, Mar. 2007, doi: 10.2136/sssaj2006.0211.

[11] J. A. M. Demattê, L. Ramirez-Lopez, K. P. P. Marques, and A. A. Rodella, “Chemometric soil analysis on the determination of specific bands for the detection of magnesium and potassium by spectroscopy,” *Geoderma*, vol. 288, pp. 8–22, Feb. 2017, doi: 10.1016/j.geoderma.2016.11.013.

[12] M. Nocita, A. Stevens, G. Toth, P. Panagos, B. van Wesemael, and L. Montanarella, “Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach,” *Soil Biol Biochem*, vol. 68, pp. 337–347, Jan. 2014, doi: 10.1016/j.soilbio.2013.10.022.

[13] T. R. Tavares *et al.*, “Multi-Sensor Approach for Tropical Soil Fertility Analysis: Comparison of Individual and Combined Performance of VNIR, XRF, and LIBS Spectroscopies,” *Agronomy*, vol. 11, no. 6, p. 1028, May 2021, doi: 10.3390/agronomy11061028.

[14] S. M. O’Rourke, B. Minasny, N. M. Holden, and A. B. McBratney, “Synergistic Use of Vis-NIR, MIR, and XRF Spectroscopy for the Determination of Soil Geochemistry,” *Soil Science Society of America Journal*, vol. 80, no. 4, pp. 888–899, Jul. 2016, doi: 10.2136/sssaj2015.10.0361.

[15] T. R. Tavares *et al.*, “Laser-Induced Breakdown Spectroscopy (LIBS) for tropical soil fertility analysis,” *Soil Tillage Res*, vol. 216, p. 105250, Feb. 2022, doi: 10.1016/j.still.2021.105250.

[16] J. V. Fontenelli *et al.*, “Evaluating the synergy of three soil spectrometers for improving the prediction and mapping of soil properties in a high anthropic management area: A case of study from Southeast Brazil,” *Geoderma*, vol. 402, p. 115347, Nov. 2021, doi: 10.1016/j.geoderma.2021.115347.

[17] C. Guerrero, R. Zornoza, I. Gómez, and J. Mataix-Beneyto, “Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy,” *Geoderma*, vol. 158, no. 1–2, pp. 66–77, Aug. 2010, doi: 10.1016/j.geoderma.2009.12.021.

[18] J. Wetterlind, B. Stenberg, and M. Söderström, “Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models,” *Geoderma*, vol. 156, no. 3–4, pp. 152–160, May 2010, doi: 10.1016/j.geoderma.2010.02.012.

[19] A. Sharma, D. C. Weindorf, T. Man, A. A. A. Aldabaa, and S. Chakraborty, “Characterizing soils via portable X-ray fluorescence spectrometer: 3. Soil reaction (pH),” *Geoderma*, vol. 232–234, pp. 141–147, Nov. 2014, doi: 10.1016/j.geoderma.2014.05.005.

[20] F. L. Melquiades and F. R. dos Santos, “Preliminary Results: Energy Dispersive X-Ray Fluorescence and Partial Least Squares Regression for Organic Matter Determination in Soil,” *Spectroscopy Letters*, vol. 48, no. 4, pp. 286–289, Apr. 2015, doi: 10.1080/00387010.2013.874532.

[21] T. R. Tavares *et al.*, “Assessing Soil Key Fertility Attributes Using a Portable X-ray Fluorescence: A Simple Method to Overcome Matrix Effect,” *Agronomy*, vol. 10, no. 6, p. 787, Jun. 2020, doi: 10.3390/agronomy10060787.

[22] S. H. G. Silva *et al.*, “Soil texture prediction in tropical soils: A portable X-ray fluorescence spectrometry approach,” *Geoderma*, vol. 362, p. 114136, Mar. 2020, doi: 10.1016/j.geoderma.2019.114136.

[23] A. Rawal *et al.*, “Determination of base saturation percentage in agricultural soils via portable X-ray fluorescence spectrometer,” *Geoderma*, vol. 338, pp. 375–382, Mar. 2019, doi: 10.1016/j.geoderma.2018.12.032.

[24] R. Andrade *et al.*, “Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains,” *Geoderma*, vol. 357, p. 113960, Jan. 2020, doi: 10.1016/j.geoderma.2019.113960.

[25] S. H. G. Silva, A. F. dos S. Teixeira, M. D. de Menezes, L. R. G. Guilherme, F. M. de S. Moreira, and N. Curi, “Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence spectrometer (pXRF),” *Ciência e Agrotecnologia*, vol. 41, no. 6, pp. 648–664, Dec. 2017, doi: 10.1590/1413-70542017416010317.

[26] T. R. Tavares, J. P. Molin, S. H. Javadi, H. W. P. de Carvalho, and A. M. Mouazen, "Combined Use of Vis-NIR and XRF Sensors for Tropical Soil Fertility Analysis: Assessing Different Data Fusion Approaches," *Sensors*, vol. 21, no. 1, p. 148, Dec. 2020, doi: 10.3390/s21010148.

[27] F. R. dos Santos, J. F. de Oliveira, E. Bona, G. M. C. Barbosa, and F. L. Melquiades, "Data fusion of XRF and vis-NIR using p-ComDim to predict some fertility attributes in tropical soils derived from basalt," *Microchemical Journal*, vol. 191, p. 108813, Aug. 2023, doi: 10.1016/j.microc.2023.108813.

[28] J. V. Ribeiro, F. R. dos Santos, J. F. de Oliveira, G. M. C. Barbosa, and F. L. Melquiades, "Optimization of pXRF instrumentation conditions and multivariate modeling in soil fertility attributes determination," *Spectrochim Acta Part B At Spectrosc*, vol. 211, p. 106835, Nov. 2024, doi: 10.1016/j.sab.2023.106835.

[29] R. Jenkins, "X-Ray Fluorescence Spectrometry," in *Handbook of Analytical Techniques*, Weinheim, Germany: Wiley-VCH Verlag GmbH, 1988, pp. 753–766. doi: 10.1002/9783527618323.ch23.

[30] R. E. Van Grieken and A. A. Markowicz, *Handbook of X-ray spectrometry*, 2nd ed., vol. 29. New York, 2002.

[31] R. C. Deo, "Machine Learning in Medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, Nov. 2015, doi: 10.1161/CIRCULATIONAHA.115.001593.

[32] Z. Obermeyer and E. J. Emanuel, "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine," *New England Journal of Medicine*, vol. 375, no. 13, pp. 1216–1219, Sep. 2016, doi: 10.1056/NEJMp1606181.

[33] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science (1979)*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.

[34] A. K. Waljee and P. D. R. Higgins, "Machine Learning in Medicine: A Primer for Physicians," *American Journal of Gastroenterology*, vol. 105, no. 6, pp. 1224–1226, Jun. 2010, doi: 10.1038/ajg.2010.173.

[35] A. M. Darcy, A. K. Louie, and L. W. Roberts, "Machine Learning and the Profession of Medicine," *JAMA*, vol. 315, no. 6, p. 551, Feb. 2016, doi: 10.1001/jama.2015.18421.

[36] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMra1814259.

[37] T. Kohonen and T. Honkela, "Kohonen network," *Scholarpedia*, vol. 2, no. 1, p. 1568, 2007, doi: 10.4249/scholarpedia.1568.

[38] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, Apr. 2014, doi: 10.1007/s11036-013-0489-0.

[39] R. Sathya and A. Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification," *International Journal of Advanced Research in Artificial Intelligence*, vol. 2, no. 2, 2013, doi: 10.14569/IJARAI.2013.020206.

[40] G. A. Barreto and L. G. M. Souza, "Adaptive filtering with the self-organizing map: A performance comparison," *Neural Networks*, vol. 19, no. 6–7, pp. 785–798, Jul. 2006, doi: 10.1016/j.neunet.2006.05.005.

[41] M. T. Eitelwein, T. R. Tavares, J. P. Molin, R. G. Trevisan, R. V. de Sousa, and J. A. M. Demattê, “Predictive Performance of Mobile Vis–NIR Spectroscopy for Mapping Key Fertility Attributes in Tropical Soils through Local Models Using PLS and ANN,” *Automation*, vol. 3, no. 1, pp. 116–131, Feb. 2022, doi: 10.3390/automation3010006.

[42] G. Forkuor, O. K. L. Hounkpatin, G. Welp, and M. Thiel, “High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models,” *PLoS One*, vol. 12, no. 1, p. e0170478, Jan. 2017, doi: 10.1371/journal.pone.0170478.

[43] T. R. Tavares *et al.*, “Effect of X-Ray Tube Configuration on Measurement of Key Soil Fertility Attributes with XRF,” *Remote Sens (Basel)*, vol. 12, no. 6, p. 963, Mar. 2020, doi: 10.3390/rs12060963.

[44] D. Wang *et al.*, “Synthesized use of VisNIR DRS and PXRF for soil characterization: Total carbon and total nitrogen,” *Geoderma*, vol. 243–244, pp. 157–167, Apr. 2015, doi: 10.1016/j.geoderma.2014.12.011.

[45] S. Dasgupta, S. Chakraborty, D. C. Weindorf, B. Li, S. H. G. Silva, and K. Bhattacharyya, “Influence of auxiliary soil variables to improve PXRF-based soil fertility evaluation in India,” *Geoderma Regional*, vol. 30, p. e00557, Sep. 2022, doi: 10.1016/j.geodrs.2022.e00557.

[46] H. G. dos Santos *et al.*, *Sistema brasileiro de classificação de solos*, 2. Rio de Janeiro: EMBRAPA-SPI, 2006.

[47] FAO, “IUSS working group WRB. World reference base for soil resources 2014, International soil classification system for naming soils and creating legends for soil maps,” *World Soil Resources Reports No. 106*, 2014.

[48] M. Brossard *et al.*, “Analysis of an illustrative interaction between structural features and earthworm populations in Brazilian ferralsols,” *Comptes Rendus Geoscience*, vol. 344, no. 1, pp. 41–49, Jan. 2012, doi: 10.1016/j.crte.2011.12.001.

[49] G. M. de C. Barbosa, J. F. de Oliveira, M. Miyazawa, D. B. Ruiz, and J. T. Filho, “Aggregation and clay dispersion of an oxisol treated with swine and poultry manures,” *Soil Tillage Res*, vol. 146, pp. 279–285, Mar. 2015, doi: 10.1016/j.still.2014.09.022.

[50] M. A. Pavan, M. de F. Bloch, H. da C. Zempulski, M. Miyazawa, and D. C. Zocoler, *Manual de análise química de solo e controle de qualidade*, vol. 76. Iapar Londrina, 1992.

[51] R. de A. Ferreira, G. Teixeira, and L. A. Peternelli, “Kennard-Stone method outperforms the Random Sampling in the selection of calibration samples in SNPs and NIR data,” *Ciência Rural*, vol. 52, no. 5, 2022, doi: 10.1590/0103-8478cr20201072.

[52] R. W. Kennard and L. A. Stone, “Computer Aided Design of Experiments,” *Technometrics*, vol. 11, no. 1, pp. 137–148, Feb. 1969, doi: 10.1080/00401706.1969.10490666.

[53] F. R. dos Santos, J. F. de Oliveira, G. M. C. Barbosa, and F. L. Melquiades, “Comparison between energy dispersive X-ray fluorescence spectral data and elemental data for soil attributes modelling,” *Spectrochim Acta Part B At Spectrosc*, vol. 185, p. 106303, Nov. 2021, doi: 10.1016/j.sab.2021.106303.

[54] F. R. dos Santos, J. F. de Oliveira, E. Bona, G. M. C. Barbosa, and F. L. Melquiades, “Evaluation of pre-processing and variable selection on energy dispersive X-ray fluorescence

spectral data with partial least square regression: A case of study for soil organic carbon prediction," *Spectrochim Acta Part B At Spectrosc*, vol. 175, p. 106016, Jan. 2021, doi: 10.1016/j.sab.2020.106016.

[55] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Anal Chim Acta*, vol. 185, pp. 1–17, 1986, doi: 10.1016/0003-2670(86)80028-9.

[56] I. S. Helland, "Partial Least Squares Regression and Statistical Models," *Scandinavian Journal of Statistics*, vol. 17, no. 2, pp. 97–114, 1990, [Online]. Available: http://www.jstor.org/stable/4616159

[57] P. Valderrama, J. W. B. Braga, and R. J. Poppi, "Variable Selection, Outlier Detection, and Figures of Merit Estimation in a Partial Least-Squares Regression Multivariate Calibration Model. A Case Study for the Determination of Quality Parameters in the Alcohol Industry by Near-Infrared Spectroscopy," *J Agric Food Chem*, vol. 55, no. 21, pp. 8331–8338, Oct. 2007, doi: 10.1021/jf071538s.

[58] H. Wold, "Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach," *J Appl Probab*, vol. 12, no. S1, pp. 117–142, Sep. 1975, doi: 10.1017/S0021900200047604.

[59] M. R. Keenan and P. G. Kotula, "Optimal scaling of TOF-SIMS spectrum-images prior to multivariate statistical analysis," *Appl Surf Sci*, vol. 231–232, pp. 240–244, Jun. 2004, doi: 10.1016/j.apsusc.2004.03.025.

[60] M. R. Keenan and P. G. Kotula, "Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images," *Surface and Interface Analysis*, vol. 36, no. 3, pp. 203–212, Mar. 2004, doi: 10.1002/sia.1657.

[61] R. A. Viscarra Rossel, R. N. McGlynn, and A. B. McBratney, "Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy," *Geoderma*, vol. 137, no. 1–2, pp. 70–82, Dec. 2006, doi: 10.1016/j.geoderma.2006.07.004.

[62] C.-W. Chang, D. A. Laird, M. J. Mausbach, and C. R. Hurburgh, "Near Infrared Reflectance Spectroscopy–Principal Components Regression Analyses of Soil Properties," *Soil Science Society of America Journal*, vol. 65, no. 2, pp. 480–490, Mar. 2001, doi: 10.2136/sssaj2001.652480x.

[63] L. A. Currie, "Nomenclature in evaluation of analytical methods including detection and quantification capabilities1Adapted from the International Union of Pure and Applied Chemistry (IUPAC) document 'Nomenclature in Evaluation of Analytical Methods including Detection and Quantification Capabilities', which originally appeared in Pure and Applied Chemistry, 67 1699–1723 (1995) © 1995 IUPAC. Republication permission granted by IUPAC.1," *Anal Chim Acta*, vol. 391, no. 2, pp. 105–126, May 1999, doi: 10.1016/S0003-2670(99)00104-X.

[64] E1655-05 ASTM, "Standard Practices for Infrared Multivariate Quantitative Analysis," *Conshohocken*, 2012.

[65] A. C. Olivieri, "Practical guidelines for reporting results in single- and multi-component analytical calibration: A tutorial," *Anal Chim Acta*, vol. 868, pp. 10–22, Apr. 2015, doi: 10.1016/j.aca.2015.01.017.

[66] V. Centner, O. E. de Noord, and D. L. Massart, “Detection of nonlinearity in multivariate calibration,” *Anal Chim Acta*, vol. 376, no. 2, pp. 153–168, Dec. 1998, doi: 10.1016/S0003-2670(98)00543-1.

[67] A. W. Warrick and D. R. Nielsen, “Spatial Variability of Soil Physical Properties in the Field,” in *Applications of Soil Physics*, Elsevier, 1980, pp. 319–344. doi: 10.1016/B978-0-12-348580-9.50018-3.

[68] A. Fontana, M. G. Pereira, A. Loss, T. J. F. Cunha, and J. C. Salton, “Fertility properties and humic fractions in a Rhodic Ferralsol in Brazilian Cerrado,” *Pesqui Agropecu Bras*, vol. 41, no. 5, pp. 847–853, May 2006, doi: 10.1590/S0100-204X2006000500018.

[69] M. A. G. da Silva, A. S. Muniz, J. D. D. V. da Mata, C. Carissimi, and A. C. Cegana, “Amostragem e variabilidade nos atributos de fertilidade em um latossolo sob plantio direto em São Miguel do Iguaçu, Estado do Paraná,” *Acta Sci Agron*, vol. 25, no. 1, Apr. 2003, doi: 10.4025/actasciagron.v25i1.2679.

[70] J. M. Reichert *et al.*, “Conceptual framework for capacity and intensity physical soil properties affected by short and long-term (14 years) continuous no-tillage and controlled traffic,” *Soil Tillage Res*, vol. 158, pp. 123–136, May 2016, doi: 10.1016/j.still.2015.11.010.

[71] U. Schwertmann and R. M. Taylor, “Iron Oxides,” 2018, pp. 379–438. doi: 10.2136/sssabookser1.2ed.c8.

[72] L. Benedet *et al.*, “Rapid soil fertility prediction using X-ray fluorescence data and machine learning algorithms,” *Catena (Amst)*, vol. 197, p. 105003, Feb. 2021, doi: 10.1016/j.catena.2020.105003.

[73] F. M. Verbi, E. R. Pereira-Filho, and M. I. M. S. Bueno, “Use of X-Ray Scattering for Studies with Organic Compounds: a Case Study Using Paints,” *Microchimica Acta*, vol. 150, no. 2, pp. 131–136, Jun. 2005, doi: 10.1007/s00604-005-0352-5.

[74] R. Andrade *et al.*, “Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains,” *Geoderma*, vol. 357, p. 113960, Jan. 2020, doi: 10.1016/j.geoderma.2019.113960.

[75] A. F. dos Santos Teixeira *et al.*, “Tropical soil pH and sorption complex prediction via portable X-ray fluorescence spectrometry,” *Geoderma*, vol. 361, p. 114132, Mar. 2020, doi: 10.1016/j.geoderma.2019.114132.

[76] Á. J. G. de Faria *et al.*, “Soils of the Brazilian Coastal Plains biome: prediction of chemical attributes via portable X-ray fluorescence (pXRF) spectrometry and robust prediction models,” *Soil Research*, vol. 58, no. 7, p. 683, 2020, doi: 10.1071/SR20136.

[77] S. H. G. Silva, A. E. Hartemink, A. F. dos S. Teixeira, A. V. Inda, L. R. G. Guilherme, and N. Curi, “Soil weathering analysis using a portable X-ray fluorescence (PXRF) spectrometer in an Inceptisol from the Brazilian Cerrado,” *Appl Clay Sci*, vol. 162, pp. 27–37, Sep. 2018, doi: 10.1016/j.clay.2018.05.028.

[78] A. M. Brinatti, Y. P. Mascarenhas, V. P. Pereira, C. S. de M. Partiti, and Á. Macedo, “Mineralogical characterization of a highly-weathered soil by the Rietveld Method,” *Sci Agric*, vol. 67, no. 4, pp. 454–464, Aug. 2010, doi: 10.1590/S0103-90162010000400013.

FINAL REMARKS

Chapter 2 evaluated optimized instrumental conditions for soil fertility attributes using a portable X-ray fluorescence (pXRF) spectrometer and soil samples. The study explored different possibilities for results from a commercial pXRF routine, using four differente data matrices in machine learning modeling. The best performance for the simultaneous quantification of the 10 soil fertility attributes evaluated was achieved by using the single raw spectra measured at 15 kV, 12.15 μA and 30 s without filters as experimental conditions, modeled with partial least squares. Comparison between individual spectra and data fusion models demonstrated that a single experimental condition is sufficient to produce accurate and reliable results in the context of this study. Therefore, based on the established RPD ranges, the models for SOC, CEC, SB, and Ca^{2+} were considered very good for quantitative prediction (RPD > 2.0), while the models for K^+ and Mg^{2+} were considered good for quantitative prediction (1.8 < RPD < 2.0). The models for PM, pH, BSP, and H+Al presented fair predictions (RPD < 1.8), useful for evaluation and correlation. These findings are valid for local applications in the context of the soil samples evaluated and demonstrate the potential of pXRF application combined with machine learning calibration in different contexts to extract quantitative and qualitative information about soil fertility.

Additionally, the randomization test showed that the modeling performance with pXRF was equivalent to benchtop EDXRF, suggesting a possible *in situ* application of this method. However, a preliminary study on extrapolation from laboratory calibrated models had negative results, indicating that their extrapolation to *in situ* measurements is challenging and complex. This is not recommended in the scenario where training and validation sample sets are collected over large time intervals.

The study of chapter 3 showed that it is possible to reduce the number of training samples while maintaining equivalent performance in PLS models with XRF spectral data. A set of 136 samples was optimal for estimating SOC and CEC, while a set of 206 samples was best for estimating SB, Mg^{2+}, and Ca^{2+}. Both reduced models were considered good (1.8 < RPD < 2.0) or very good (RPD > 2.0) for quantitative predictive analysis. It was not possible to reduce the training set for indirect K^+ quantification.

In conclusion, the study in Chapter 2 provided information on the best choice of pXRF experimental condition in modeling soil fertility attributes, highlighting the viability of this methodology for local soil analyses, accelerating results in terms of measurement time and capacity for processing. Chapter 3 provides information on the minimum size of the pXRF training sample set for modeling soil fertility attributes. Although models with reduced training sets show some performance loss, this may be advantageous depending on the accuracy required. The reduction in the number of samples analyzed by conventional methods is beneficial from a financial, logistical, and environmental point of view. Therefore, these findings bring pXRF closer to rapid local soil fertility mapping in the context of precision agriculture. The procedure enables agile decisions related to the use of fertilizers and other products for soil correction. Researchers are encouraged to reduce the training sample set and use pXRF optimized conditions to assess soil fertility in different contexts and purposes.

Scientific production during the master's degree.

Paper Published

RIBEIRO, J. V., DOS SANTOS, F. R., DE OLIVEIRA, J. F., BARBOSA, G. M. C., MELQUIADES, F. L. Optimization of pXRF instrumentation conditions and multivariate modeling in soil fertility attributes determination. Spectrochimica Acta Part B Atomic Spectroscopic, vol. 211, p. 106835, Nov. 2024, doi: 10.1016/j.sab.2023.106835.

Papers Submitted

RIBEIRO, J. V., ROCHA, D. R., ANTONELI, V., THOMAZ, E. L. MELQUIADES, F. L. Effect of soil tillage systems and crop seasonality on the sediment geochemical properties transferred into a hydrosystem. Paper Submitted to Soil and Sediments contamination.

LOPES, J. M. F., RIBEIRO, J. V., MELQUIADES, F. L., ANDRELLO, A. C. Gamma-Rays and X-Rays Spectrometries Applied to Evaluate Soil Redistribution. Paper Submited to Journal of Environmental Radioactivity.

Studies Presented at Scientific Events

RIBEIRO, J. V., DOS SANTOS, F. R., FRANCIRLEI, J., BARBOSA, G. MELQUIADES, F. L. pXRF spectral data combined with PLSR to determine SOC, CEC, and SB of soil. INCT-FNA Symposium, 2022, Universidade Federal Fluminense, Niterói, RJ.

RIBEIRO, J. V., DOS SANTOS, F. R., FRANCIRLEI, J., BARBOSA, G. MELQUIADES, F. L. Otimização das condições experimentais de um espectrômetro de pXRF para determinação de carbono orgânico em solo via regressão multivariada. VI Escola de Inverno de Quimiometria, 2023, Universidade de Brasília, Brasília, DF.

RIBEIRO, J. V, CALIXTO, O. P. P., MELQUIADES, F. L. Evaluation of Spectroscopic Sensors as Nutrient Quantification Tools in Animal Forage Samples. II Escola de Fisiologia e Nutrição de Plantas, 2023, Centro de Energia Nuclear na Agricultura, Piracicaba, SP.

RIBEIRO, J. V., FRANCIRLEI, J., BARBOSA, G. MELQUIADES, F. L. Comparison of predictive performance between ANN and PLS for Soil Organic Carbon quantification through EDXRF spectrum. V Escola de Física de Curitiba, 2024, Universidade Federal do Paraná, Curitiba, PR.

Awards at Scientific Events

Best poster: RIBEIRO, J. V., DOS SANTOS, F. R., FRANCIRLEI, J., BARBOSA, G. MELQUIADES, F. L. Otimização das condições experimentais de um espectrômetro de pXRF para determinação de carbono orgânico em solo via regressão multivariada. VI Escola de Inverno de Quimiometria, Universidade de Brasília, 2023, Brasília, DF.

Participation in Scientific Events

INCT-FNA Symposium, Universidade Federal Fluminense, 2022, Niterói, RJ

IV Aplicações na Agropecuária - Espectroscopia Vis Nir, Universidade Estadual de Londrina, 2023, Londrina, PR

Cadeia Produtiva de Alimentos e Produtos Orgânicos, Universidade Estadual de Londrina, 2023, Londrina, PR

VI Escola de Inverno de Quimiometria, Universidade de Brasília, 2023, Brasília, DF

II Escola de Fisiologia e Nutrição de Plantas, Centro de Energia Nuclear na Agricultura, 2023, Piracicaba, SP

V Escola de Física de Curitiba, Universidade Federal do Paraná, 2024, Curitiba, PR