agriculture

MDPI

Article

Estimation of Soil Cations Based on Visible and Near-Infrared Spectroscopy and Machine Learning

Yiping Peng [1,†], Ting Wang [1,2,†], Shujuan Xie [3], Zhenhua Liu [1], Chenjie Lin [1], Yueming Hu [4], Jianfang Wang [1] and Xiaoyun Mao [1,*]

[1] College of Natural Resources and Environment, South China Agricultural University, Guangzhou 510642, China
[2] Dongguan Institute of Surveying and Mapping, Dongguan 523000, China
[3] Guangdong Academy of Social Sciences, Guangzhou 510635, China
[4] College of Tropical Crops, Hainan University, Haikou 570228, China
* Correspondence: xymao@scau.edu.cn; Tel.: +86-020-8528-0296
† These authors contributed equally to this work.

Abstract: Soil exchange cations are a basic indicator of soil quality and environmental clean-up potential. The accurate and efficient acquisition of information on soil cation content is of great importance for the monitoring of soil quality and pollution prevention. At present, few scholars focus on soil exchangeable cations using remote sensing technology. This study proposes a new method for estimating soil cation content using hyperspectral data. In particular, we introduce Boruta and successive projection (SPA) algorithms to screen feature variables, and we use Guangdong Province, China, as the study area. The backpropagation neural network (BPNN), genetic algorithm–based back propagation neural network (GABP) and random forest (RF) algorithms with 10-fold cross-validation are implemented to determine the most accurate model for soil cation (Ca^{2+}, K^{+}, Mg^{2+}, and Na^{+}) content estimations. The model and hyperspectral images are combined to perform the spatial mapping of soil Mg^{2+} and to obtain the spatial distribution information of images. The results show that Boruta was the optimal algorithm for determining the characteristic bands of soil Ca^{2+} and Na^{+}, and SPA was the optimal algorithm for determining the characteristic bands of soil K^{+} and Mg^{2+}. The most accurate estimation models for soil Ca^{2+}, K^{+}, Mg^{2+}, and Na^{+} contents were Boruta-RF, SPA-GABP, SPA-RF and Boruta-RF, respectively. The estimation effect of soil Mg^{2+} (R^2 = 0.90, ratio of performance to interquartile range (RPIQ) = 3.84) was significantly better than the other three elements (Ca^{2+}: R^2 = 0.83, RPIQ = 2.47; K^{+}: R^2 = 0.83, RPIQ = 2.58; Na^{+}: R^2 = 0.85, RPIQ = 2.63). Moreover, the SPA-RF method combined with HJ-1A HSI images was selected for the spatial mapping of soil Mg^{2+} content with an R^2 of 0.71 and RPIQ of 2.05. This indicates the ability of the SPA-RF method to retrieve soil Mg^{2+} content at the regional scale.

Keywords: soil cations; VIS-NIR spectroscopy; feature screening; machine learn algorithm

Citation: Peng, Y.; Wang, T.; Xie, S.; Liu, Z.; Lin, C.; Hu, Y.; Wang, J.; Mao, X. Estimation of Soil Cations Based on Visible and Near-Infrared Spectroscopy and Machine Learning. *Agriculture* **2023**, *13*, 1237. https://doi.org/10.3390/agriculture13061237

Academic Editors: Liujun Zhu, Jeffrey Walker and Carsten Montzka

Received: 11 April 2023
Revised: 8 June 2023
Accepted: 8 June 2023
Published: 13 June 2023

1. Introduction

Soil exchange cations are a basic indicator of soil quality and environmental purification potential, providing a guarantee of the quality of plant growth [1]. However, the rapid expansion of urbanization and industrialization has reduced the competitiveness of agricultural development and altered original land use patterns. Cultivation patterns, drainage and irrigation systems and tillage practices have also changed. This has greatly affected the quantity, nature and natural distribution patterns of fertilizers and crop residues in the soil, resulting in a disruption of the original balance of the soil exchange cation content, with varying degrees of impact on soil quality [2,3]. In addition, soil cation exchange has been shown to be an important predictor of soil quality and has the potential to sequester contamination in the environment [1]. For example, the soil cation exchange capacity can influence the adsorption and desorption of metal ions such as copper, zinc and lead [4,5],

while it is significantly correlated with the sequestration of organic pollutants such as atrazine, diquat and paraquat, among other elements [6]. Therefore, in order to improve soil management and pollution control, an efficient monitoring approach for soil exchange cations must be determined.

Traditional methods for determining the soil exchange cation content generally rely on field sampling and laboratory chemical analysis, which have limitations, including high costs, long lead times, low efficiency, damaging sampling, and so on [7]. The development of remote sensing technology has provided a powerful means to efficiently monitor soil property information. However, current studies that estimate the soil exchangeable cation content using remote sensing technology employ multispectral images. Such images have a low amount of spectral information and hardly capture sensitive spectral features, thus affecting estimation accuracy. The emergence of hyperspectral technology can overcome this constraint as it is able to capture weak signals in the soil and account for fine soil features. Thus, quantitative research on soil properties using hyperspectral technology has currently become a hot research topic [8–13].

Several scholars have attempted inverse soil cations using hyperspectral techniques, making substantial progress [14–17]. Viscarra Rossel et al. (2006) compared the predictive power of four soil spectral ranges (visible, VIS; near-infrared, NIR; mid-infrared, MIR; and VIS-NIR-MIR) for a variety of soil properties (e.g., exchangeable calcium (Ca), organic carbon (OC), cation exchange capacity (CEC) and exchangeable potassium (K)). The authors found that NIR spectroscopy was more accurate for exchangeable aluminum and potassium, with a root mean square error (RMSE) of 0.88 and 1.84 mmol(+)/kg, respectively [14]. Li et al. (2011) explored the ability to predict soil cation exchange and exchangeable potassium, calcium and magnesium content within the drip line of the litchi canopy using NIR spectroscopy. The proposed approach was able to effectively predict soil cation exchange and exchangeable calcium content, yet predictions of exchangeable potassium and magnesium content were poor [15]. Gras et al. (2014) compared three spectral collection methods (scanning the soil surface, primitive or smooth cores captured with an auger and soil blocks from fragmented cores) and combined multiple spectral treatments to predict soil properties, including total nitrogen, organic matter, calcium carbonate, exchangeable potassium and fast-acting phosphorus. Good predictions were achieved for soil exchangeable potassium (residual prediction deviation (RPD) = 2.7–3.2; R^2 = 0.85–0.90), while available phosphorus was poorly predicted (RPD = 1.6–1.8; R^2 = 0.59–0.70) [16]. Despite the advancements made, current studies suffer from a relatively homogeneous approach to feature selection, poor model portability and limited research on soil cation inversion.

Therefore, this study proposes a method that integrates feature screening and model construction for the quantitative inversion of soil cations. The objectives of the study are as follows: (1) determine the feature variables using Boruta and successive projection (SPA) algorithms; (2) quantify the relationship between feature variables and soil cation content using the backpropagation neural network (BPNN), genetic algorithm–based back propagation neural network (GABP) and random forest (RF) algorithms; and (3) explore the feasibility for the spatial mapping of soil cations using HJ-1A HSI imagery.

2. Materials and Methods

2.1. Study Area

Guangdong Province is located in the southernmost part of mainland China (20°09′–25°31′ N, 109°45′–117°20′ E, Figure 1) and exhibits a subtropical monsoon climate with abundant water and climatic resources. It is subject to the combined effects of crustal movements, lithology, folding and fracture tectonics and external forces, resulting in a complex and diverse landscape with mountains, hills, plateaus and plains. The terrain is generally high in the north and low in the south, with mountains and high hills in the north and plains and tablelands in the south. Since Guangdong's reform and opening up, urbanization and industrialization have a negative impact on the overall level of soil quality in the province,

and the problem of soil pollution has become increasingly prominent. Therefore, in this study, the cultivable land in Guangdong Province was selected as the study area.

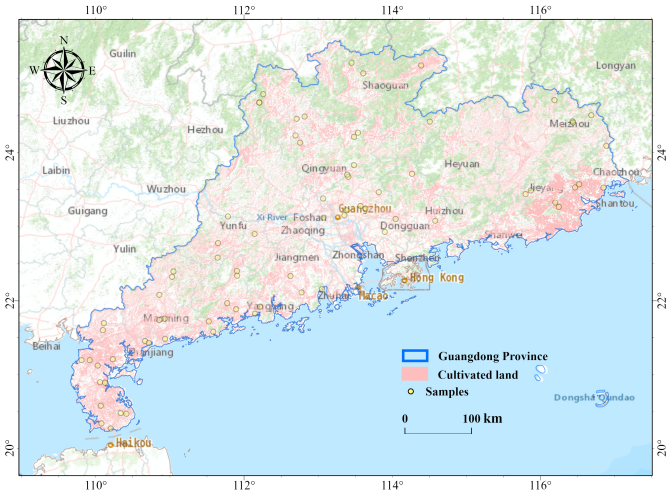

Figure 1. Study area (20°09′–25°31′ N, 109°45′–117°20′ E) used for model development and the spatial distribution of soil samples.

2.2. Data and Pre-Processing

2.2.1. Soil Sampling and Pre-Processing

To ensure a uniform distribution of soil samples, a grid distribution method at a 50 × 50 km resolution was used to select the sampling points, with a total of 75 soil samples (Figure 1). Five points were collected in an "X" shape at the topsoil (0–20 cm) to create a mixed sample. Each sample weighed approximately 300 g. The soil samples were dried naturally, and impurities such as stones, plant stems and plastic products were removed, milled and passed through a 0.2 mm soil sieve. Each soil sample was subsequently divided into two separate samples for the chemical analysis of the soil cation content and the determination of the soil's spectral reflectance, respectively. In addition, 20 soil samples were collected in Conghua District (Figure 2) to verify the feasibility of the model for the spatial mapping of the soil cation content at the regional scale. The soil sample collection and pre-processing procedures follow the above-described steps.

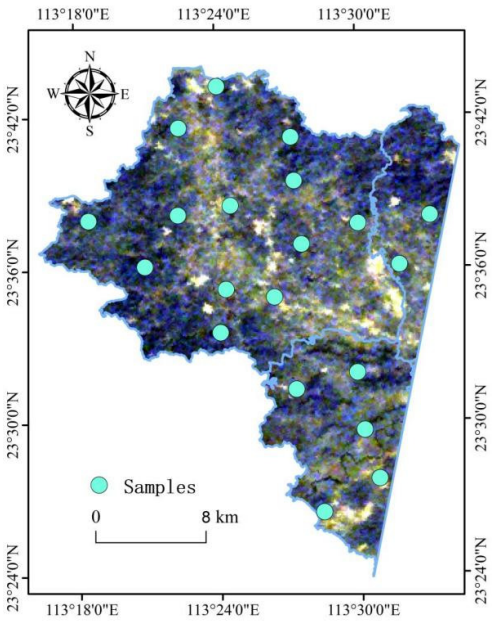

Figure 2. HJ-1A false color composite image and sample point distribution for mapping verification.

(1) Determination of soil cation content

An inductively coupled plasma-optical emission spectrometer (ICP-OES) was used to determine soil Ca^{2+}, K^+, Mg^{2+} and Na^+ contents [18]. Table 1 reports the descriptive statistics of the measured data. The mean values of soil Ca^{2+}, K^+, Mg^{2+} and Na^+ were 3.37 cmol/kg, 0.20 cmol/kg, 0.50 cmol/kg and 0.10 cmol/kg, respectively, and all variables exhibited high variability (CV > 50%) [19].

Table 1. Statistics of the soil cations contents (unit: cmol kg^{-1}).

Soil Cations	Min.	Max.	Mean	SD	CV (%)
K^+	0.03	0.78	0.20	0.15	75
Ca^{2+}	0.24	14.90	3.37	2.91	86
Mg^{2+}	0.08	3.00	0.50	0.58	116
Na^+	0.02	0.46	0.10	0.08	80

Note: SD, standard deviation; CV, coefficient of variation.

(2) Soil spectral reflectance measurements

Soil spectra reflectance values were measured using an AvaField portable spectrometer (Avantes, Inc., Apeldoorn, Holland) with a spectral resolution of 0.6 nm and 6 nm in the 300–1100 nm and 1100–2500 nm ranges, respectively. Prior to measuring soil sample spectra, an AvaField portable spectrometer was calibrated using a standard white board (LS-WS250-R made of polytetrafluoroethylene (PTFE) with a response wavelength range of 250–2500 nm). The spectra were collected by placing the soil samples in a black box to reduce the influence of the external environment. A 50 W halogen lamp was employed as the measurement light source, and the spectra were measured by the vertical contact of a probe with a 10° field of view connected by an optical fiber. The soil spectrum of each sample was collected five times, and anomalous curves were removed using AvaReader software. The average of the remaining reflectance data was spectrally smoothed using Savitzky–Golay smoothing with a window size of 10, and the smoothed data were taken as the spectral reflectance (R) of the sample. In addition, the first derivative (FD), second derivative (SD) and reciprocal logarithmic (RL = $\lg 1/\rho$, ρ is reflectance) transformations were determined for the soil spectral data to attenuate the effect of the background noise and the signal intensity variation due to scattering and absorption from the soil surface. We refer readers to the previous literature for more details on the three spectral transformations [20].

2.2.2. Image Acquisition and Pre-Processing

To extend the application of the proposed method at the regional scale, satellite imagery was collected to map the soil cation content. The model constructed in this paper was based mainly on soil spectra, the crops in Guangdong were mainly rice, and the rice had been harvested in late October. Therefore, an HJ-1A HSI image was acquired on 30 October 2017, which reflected soil information. The spectral range, spectral resolution, spatial resolution and swath of the image were 459–956 nm, 4.32 nm, 100 m and 50 km, respectively. The collected image was subjected to noise and geometric distortion due to the atmospheric environment and the sensor itself. Thus, we pre-processed the image using absolute radiometric brightness conversion, atmospheric correction, geometric correction and mixed-pixel decomposition. The atmospheric correction of the HJ-1A image was performed using the FLAASH model. Its geometric accuracy correction was carried out using a quadratic polynomial calculation model and a cubic convolution interpolation method. The correction error was within 0.5 pixels. Mixed pixel decomposition was performed using the fully constrained least squares (FCLS) spectral unmixing model.

2.3. Methods

2.3.1. Screening of Characteristic Bands

Redundant information and noise are usually more present in full-band spectra compared to multispectral data, which affects the accuracy of the estimation model. Screening

the characteristic bands is a key step in constructing a prediction model, and it can improve the speed and accuracy of the model [21]. However, due to the weak spectral information and the difficulties in capturing it, the optimal characteristic bands tend to be poorly determined by traditional statistical analysis methods alone. In this study, SPA and Boruta algorithms were introduced to screen the characteristic bands of soil cations. The following sections provide a brief summary of the two algorithms.

(1) Successive projection algorithm

The successive projection algorithm (SPA) is a forward-loop feature variable selection method that can filter the wavelength positions of interest from overlapping spectral information and reduce the effect of co-linearity. It is widely adopted to filter feature wavelengths [22,23]. SPA selects the wavelength with the largest projection vector as the initial wavelength by projecting the wavelength onto other wavelengths, and then it determines the characteristic bands based on a calibration model. The initial iteration vector is denoted as $xk(0)$, the variables to be extracted are denoted as N, and the spectral matrix is denoted as column J. The SPA algorithm was performed in Python software via the SPA package. The steps of the algorithm are summarized as follows [24]:

1. Randomly select a column of the spectral matrix (column j) and assign column j from the modelling set to x_j, denoted as $xk(0)$.
2. Denote the set of vector positions of the other columns (excluding the jth column) as s, where $s = \{j, 1 \le j \le J, j \notin [k(0), \ldots k(n-1)]\}$.
3. Calculate the projection of x_j onto the remaining column vectors, $Px_j = x_j - \left(x_j^T x_{k(n-1)}\right)\left(x_{k(n-1)}^T x_{k(n-1)}\right)^{-1}, j \in s$.
4. Extract the spectral wavelengths with a maximum projection vector, $k(n) = \arg\left(\mathrm{Max}\left(\| Px_j \|\right)\right), j \in s$.
5. Let $x_j = Px_j, j \in s$.
6. Accumulate n and return to step 2 if $n < N$.
7. Denote the extracted variables as $\left\{x_{k(n)} = 0, \ldots, N-1\right\}$, corresponding to $k(0)$ and N in each cycle. Following this, construct a multiple linear regression analysis model to obtain the root mean square error of cross-validation (RMSECV) for the modelling set, and the smallest RMSECV value corresponding to $k(0)$ and N as the characteristic wavelength groups is obtained.

(2) Boruta algorithm

The Boruta algorithm is a feature filtering method based on two core concepts, namely, shaded features and binomial distribution. It offers a fundamental gauge for the importance of each variable, denoted as the Z-score. The Z-scores of the original variables are compared with the expected Z-scores of randomly selected features generated by random permutations, and only variables with Z-scores greater than all randomly selected features are selected to build the classifier [25]. In this study, the Boruta algorithm was used to select significantly important and uncorrelated spectral variables to estimate soil cation (Ca^{2+}, K^+, Mg^{2+} and Na^+) contents. The Boruta algorithm was performed in Python software via the BorutaPy package. The key steps of the algorithm are as follows [26]:

1. Read the original soil reflectance spectral matrix (denoted as R) and then randomly disrupt the column order of the matrix to generate a new soil reflectance spectral matrix (denoted as S). Following this, associate S with R according to the sample point identifier (ID) to obtain a new feature matrix (denoted as N).
2. Input N into the training model to obtain the R and S importance levels (Z-score).
3. Extract the maximum value of the Z-score in S ($maxS$) and record the characteristic bands in R with a Z-score greater than $maxS$.
4. Repeat step 3 for the determination and marking of the importance of the characteristic bands.

5. Remove the unimportant bands and repeat the above process until all the characteristic bands have been marked.

2.3.2. Model Construction

The BPNN, GABP and RF algorithms were used to determine the relationship between the characteristic bands and soil cations contents, and the corresponding accuracy assessments were performed. In the following, we briefly summarize each algorithm.

(1) Back Propagation Neural Network

The back propagation neural network (BPNN) model is a multilayer feed-forward network trained according to the error backpropagation algorithm and consists of input layer, hidden layer and output layer. It uses the gradient descent method and backpropagation algorithm to iteratively adjust the weights and biases of the network until training ends when the predicted amount is as close as possible to the actual value. The learning process consists of two components: the forward propagation of the input signal and the backward propagation of the error. We refer readers to the previous literature for more details [27]. The BPNN algorithm was performed in Python software via the TensorFlow package.

(2) Genetic Algorithm–Based Back Propagation Neural Network

The genetic algorithm–based back propagation neural network (GABP) is a genetic algorithm that optimizes the BPNN in terms of weight assignment and threshold setting. The key concept of the algorithm is to transform the original weights and thresholds of the BPNN into chromosomes in the genetic algorithm using real number encoding. We refer readers to the previous literature for more details [28]. The GABP algorithm was performed in Python software via the TensorFlow package.

(3) Random Forest

Random forest (RF) is an ensemble learning algorithm that can effectively simulate the non-linear relationship between characteristic indicators and explanatory variables, with the advantages of a high generalization ability and robustness, a low number of tuning parameters and high training speed [29,30]. It generates many equal-sized new samples from the sample itself using bootstrap sampling based on put-back random sampling. Two-thirds of the original sample is typically selected to build the decision tree. The remaining data are used as out-of-bag data (OOB) and are not involved in the model training; instead, they are used to validate the decision tree. A decision tree built from numerous bootstrap samples is then combined to improve the predictive power. Finally, the results of all decision trees are synthesized, and the average of their predictions is used as the final result [31]. The RF algorithm was performed in Python software via the Sklearn package.

In this study, the above high-accuracy estimation methods were applied to mapping the contents of the soil Mg^{2+} using the HJ-1A HSI image for the Conghua district at the regional scale. However, the model could not be used directly on the HJ-1A HSI image because the spectral resolution of the image is coarser than the measured spectrum of the AvaField portable spectrometer. Therefore, to match the spectral resolution of HJ-1A HSI data, ENVI's spectral resampling procedure was used to perform the spectral resampling of soil spectral measurements collected using the AvaField portable spectrometer. Moreover, these resampling spectral variables from HJ-1A HSI data were applied to map the spatial distribution of the soil cation contents. This involved two specific steps: (1) determining the characteristic bands using the optimal feature screening algorithm and (2) using the most accurate estimation models identified to perform soil Mg^{2+} spatial mapping.

2.3.3. Accuracy Metrics

The coefficient of determination (R^2) and the ratio of performance to interquartile range (RPIQ) were used to assess the performance of the estimation models. RPIQ is defined as the ratio of IQ to RMSECV. IQ is the difference between the third quartile

and the first quartile of the measured value. RMSECV is the root mean square error of cross-validation [32,33].

3. Results

3.1. Determining the Characteristic Bands for the Estimation of Soil Cation Contents

3.1.1. Characterization of Soil Spectra

The Pearson correlation coefficient between soil cation content and spectral data was used to initially derive the spectral characteristics and sensitive bands of soil cations in order to provide a reference for the input variables of the subsequent feature screening algorithm (Figure 3). The correlation between FD and the content of the four soil cations is generally significantly higher than that of other spectral transforms (SD, RL and R). The correlation coefficient curves of the soil cations with FD and SD fluctuate sharply, while those with R and RL are relatively smooth. The curves exhibit sharp changes around 1400 nm, 2000 nm and 2200 nm. This may be attributed to the absorption features of the water molecule hydroxide groups occurring near 1400 nm and 2000 nm and the presence of clay mineral absorption features near 2200 nm.

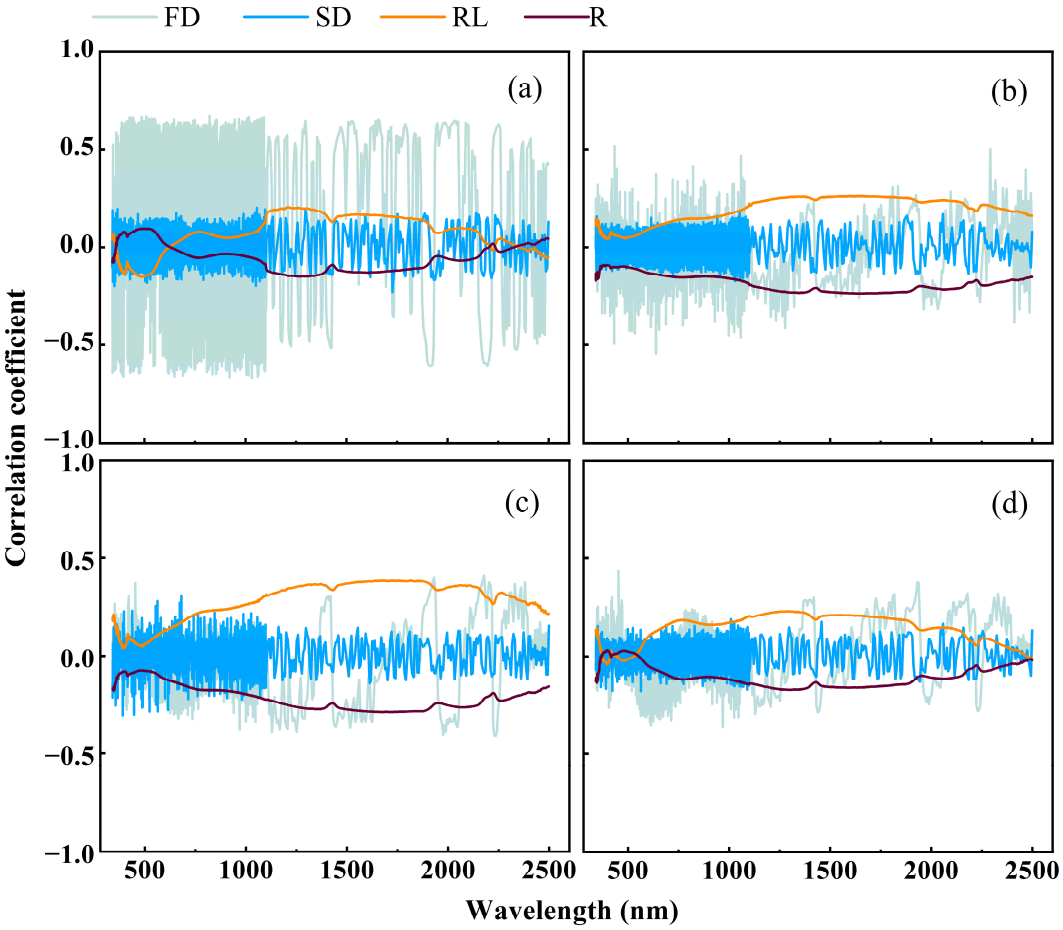

Figure 3. Correlation coefficients between the spectral variables and concentrations of (**a**) Ca^{2+}, (**b**) K^{+}, (**c**) Mg^{2+} and (**d**) Na^{+}.

3.1.2. Screening of Characteristic Bands for Soil Cations

The results of the soil spectral characterization identify FD as the most sensitive spectral transformation to soil cations. Thus, in this study, only FD was used as the input data of SPA and Boruta for the screening of characteristic bands (Table 2). For the SPA algorithm, the number of bands selected for soil Ca^{2+}, K^{+}, Mg^{2+} and Na^{+} was 8, 5, 5 and 6, respectively. The characteristic bands for Ca^{2+} were concentrated between 600 and 1100 nm

and those for K^+ were mainly distributed between 450 and 900 nm. The characteristic bands for Mg^{2+} were more evenly distributed in the range of 500 to 2000 nm, while those for Na^+ generally surpassed 1000 nm. For Boruta, the number of characteristic bands for Ca^{2+}, K^+, Mg^{2+} and Na^+ was 23, 8, 10 and 10, respectively, exceeding the number of bands screened by the SPA algorithm, while the distribution of the screened bands was also less consistent.

Table 2. Characteristic bands for the four soil cations based on SPA and Boruta.

Model	Soil Cation	Number	Characteristics Bands
SPA	Ca^{2+}	8	FD_{800}, FD_{638}, FD_{805}, FD_{785}, FD_{371}, FD_{1073}, FD_{1392}, FD_{896}
	K^+	5	FD_{455}, FD_{808}, FD_{2360}, FD_{474}, FD_{854}
	Mg^{2+}	5	FD_{1624}, FD_{1937}, FD_{664}, FD_{513}, FD_{1076}
	Na^+	6	FD_{1937}, FD_{1017}, FD_{596}, FD_{1088}, FD_{2253}, FD_{443}
Boruta	Ca^{2+}	23	FD_{651}, FD_{1649}, FD_{1816}, FD_{776}, FD_{1066}, FD_{1321}, FD_{1257}, FD_{690}, FD_{768}, FD_{1656}, FD_{1643}, FD_{1000}, FD_{1784}, FD_{661}, FD_{960}, FD_{654}, FD_{1765}, FD_{1790}, FD_{903}, FD_{1937}, FD_{620}, FD_{1752}, FD_{974}
	K^+	8	FD_{1962}, FD_{2184}, FD_{2190}, FD_{566}, FD_{647}, FD_{892}, FD_{650}, FD_{651}
	Mg^{2+}	10	FD_{1476}, FD_{1482}, FD_{1515}, FD_{2235}, FD_{969}, FD_{1128}, FD_{1930}, FD_{1975}, FD_{2178}, FD_{744}
	Na^+	10	FD_{1128}, FD_{969}, FD_{1367}, FD_{968}, FD_{577}, FD_{1309}, FD_{1994}, FD_{1476}, FD_{594}, FD_{1930}

3.2. Determining an Accurate Model for the Estimation of Soil Cation Contents

The characteristic bands (Table 2) screened by the Boruta and SPA algorithms were used as independent variables, and the soil cation content was employed as the dependent variable. The estimation models of soil cation content were then constructed using BPNN, GABP and RF, as shown in Figure 4. The most accurate estimation models for soil Ca^{2+}, K^+, Mg^{2+} and Na^+ contents were determined as Boruta-RF, SPA-GABP, SPA-RF and Boruta-RF, respectively. The predictions of soil Mg^{2+} (R^2 = 0.90, RPIQ = 3.84) content were markedly better than those of the other three elements (Ca^{2+}: R^2 = 0.83, RPIQ = 2.47; K^+: R^2 = 0.83, RPIQ = 2.58; Na^+: R^2 = 0.85, RPIQ = 2.63). This may be attributed to the stronger dependence of exchangeable Mg ions in organic matter and clay, which are more responsive in the NIR spectral range [34].

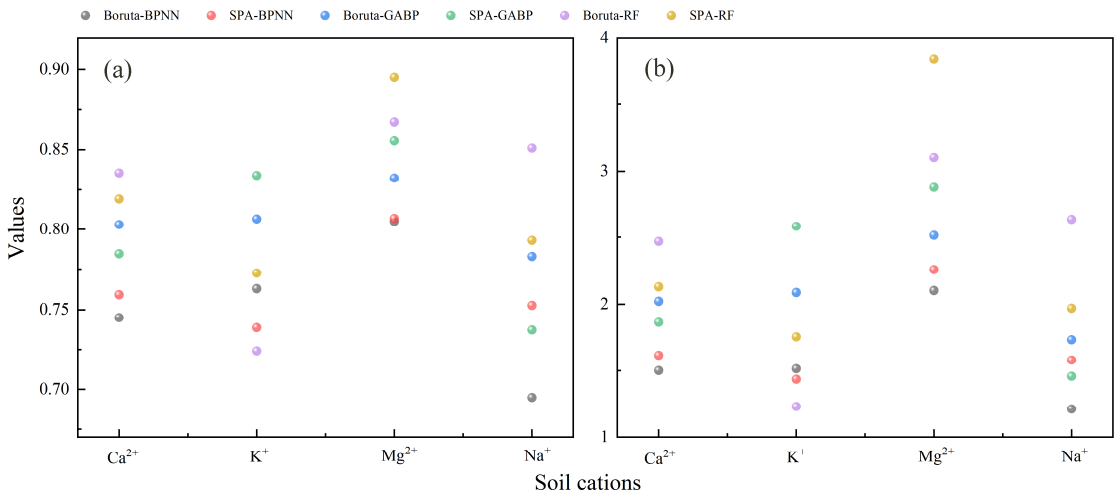

Figure 4. Scatterplots of the (**a**) coefficient of determination (R^2) and (**b**) ratio of performance to interquartile range (RPIQ) values from the estimation models with 10-fold cross-validation.

3.3. Regional-Scale Soil Cation Content Estimation Based on HJ-1A Data

To explore the applicability of the spatial mapping of soil cation content using HJ-1A HSI data, the best-estimated soil cation (Mg^{2+}) was selected for the content inversion at the regional scale. All hyperspectral bands were first-order differentially transformed, and the SPA algorithm was then used to filter out the characteristic bands of soil Mg^{2+} based

on the HJ-1A image, including $FD_{460.04}$, $FD_{521.475}$ and $FD_{664.90}$. The SPA-RF model was subsequently combined with HJ-1A HSI images to spatially map the soil Mg^{2+} in Conghua District, Guangzhou City (Figure 5a).

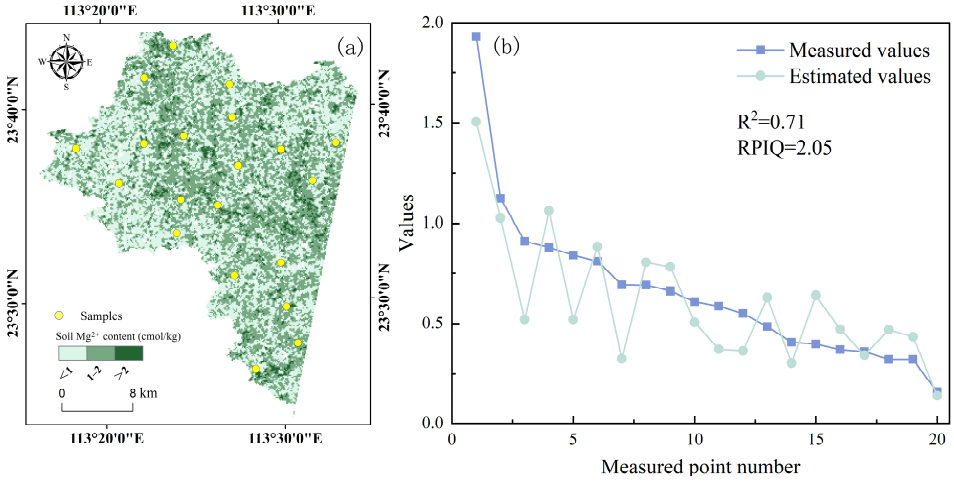

Figure 5. (**a**) Spatial distribution of soil Mg^{2+} content and (**b**) accuracy validation.

Figure 5a reveals that the soil Mg^{2+} content in the study area is mainly concentrated in the range of 0–1 cmol/kg, with a spatial distribution that is generally "high in the center and low in the surroundings". In order to verify the feasibility of the SPA-RF model for the estimation of soil Mg^{2+} content at the regional scale, the measured and predicted values of the soil Mg^{2+} content in the validation samples (cyan points in Figure 2) were compared (Figure 5b). R^2 and RPIQ were determined as 0.71 and 2.05, respectively, indicating a relatively high estimation accuracy and strong ability of the model to reflect the spatial distribution characteristics of soil Mg^{2+} content. However, the estimation accuracy at the regional scale is lower than that at the point scale. This may be due to the impact of atmospheric conditions, soil surface conditions and inconsistent band ranges.

4. Discussion

Soil exchange cations are a fundamental indicator of soil quality and environmental clean-up potential. Moreover, soil cations influence the ionic adsorption and desorption of metal elements and organic pollutants, and they have the potential to sequester pollution. Therefore, the efficient monitoring of soil cation content is of significance for soil management and pollution prevention and control. In this study, soil cation content was quantitatively estimated using hyperspectral techniques by fusing feature screening methods and machine learning modelling approaches.

Current studies relevant to the quantitative estimation of soil nutrients [35,36], soil heavy metals [37,38] and other soil properties [32,39,40] using hyperspectral techniques are relatively mature. However, research on soil exchangeable cations is relatively minimal. In this study, based on the analysis of soil spectral characteristics, novel characteristic-band screening methods (SPA and Boruta) were introduced to optimize the soil cation content estimation model. The optimal algorithm for the characteristic-band screening of soil Ca^{2+} and Na^{+} was identified as Boruta. This may be due to its ability to add randomness to the system and collect results in a random sample set, reducing errors due to random fluctuations and correlations [25].

Among the three estimation models (BPNN, GABP and RF), the RF model was generally more effective than the models for the majority of soil cation (Ca^{2+}, Mg^{2+} and Na^{+}) estimations. This may be due to the greater tolerance of the random forest algorithm relative to outliers and noise. Moreover, the RF model is not prone to overfitting and can better handle data with non-linear relationships compared to the other models [41]. In addition, the estimation model for soil Mg^{2+} was significantly better than that for the other

cations. This study could serve as a reference for government departments involved in agricultural management for soil monitoring, crop cultivation management, and so on.

Atmospheric and soil surface conditions (e.g., vegetation cover), as well as a narrow spectral band range (459–956 nm), resulted in lower-accuracy regional-scale estimates compared to the point scale. Therefore, these factors will be considered when conducting future research. Furthermore, due to time and labor constraints, only 75 soil samples were collected for developing the hyperspectral estimation models of the contents of soil cations. Although sampling was designed according to different levels of soil characteristics and types, the sample size was comparatively small. Hence, the 10-fold CV method [42] was used to validate the reliability of the model. In future work, larger sample sizes will be employed to further validate the proposed method. Additionally, based on the types of soil, training samples and validation samples will further be divided for discussing the representativeness of the points chosen for each type of soil. Finally, it should be pointed out that the model here was constructed using data from a relatively short time period and assumes that the cation is static, and its reliability needs to be validated by using data from varying time periods. In future studies, more samples and high-quality hyperspectral data will be collected during the non-growing period of cultivated land in order to further validate our conclusions.

5. Conclusions

This research study proposes new approaches for the quantitative estimation of soil cations using soil spectral variables. These new approaches combined non-linear screening with estimation algorithms to improve the estimation accuracy of soil cation contents. Research results showed that Boruta-RF, SPA-GABP, SPA-RF and Boruta-RF were optimal for estimating soil Ca^{2+}, K^{+}, Mg^{2+} and Na^{+} contents, respectively, and the proposed method for Mg^{2+} was reliable with an R^2 of 0.71 and RPIQ of 2.05 at the regional scale, indicating that these methods have the potential to effectively select the spectral characteristic bands of soil cations, increasing the accuracy of results. In future studies, we will add more sample data and introduce additional machine learning algorithms (e.g., XGBoost, LASSO, etc.) to improve the estimation accuracy of soil cations.

Author Contributions: Conceptualization, Y.P. and T.W.; methodology, Y.P., Z.L. and T.W.; software, S.X. and C.L.; validation, Y.P., S.X. and T.W.; investigation, Y.P. and T.W.; resources, Y.H. and X.M.; data curation, Y.P., S.X. and J.W.; writing—original draft preparation, Y.P. and T.W.; writing—review and editing, Y.P., T.W. and Z.L.; funding acquisition, X.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Guangdong Province, China (No. 2021A1515011643), and Guangdong Province Agricultural Science and Technology Innovation and Promotion Project (No. 2023KJ102).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tang, L.; Zeng, G.M.; Nourbakhsh, F.; Shen, G.L. Artificial neural network approach for predicting cation exchange capacity in soil based on physico-chemical properties. *Environ. Eng. Sci.* **2009**, *26*, 137–146. [CrossRef]
2. Xu, W.W.; Qian, M. Soil Carbon Contents in Relation to Soil Physicochemical Properties in Arid Regions of China. *J. Desert Res.* **2014**, *34*, 1558–1561.
3. Zhang, H.; Zhao, X.M.; Ouyang, Z.C.; Guo, X.; Kuang, L.H.; Ye, Y.C. Effects of Different Farmland Use Types on Soil Nutrients in Jiangxi Province. *Res. Soil Water Conserv.* **2018**, *25*, 53–60.
4. Altin, A.; Degirmenci, M. Lead (II) removal from natural soils by enhanced electrokinetic remediation. *Sci. Total Environ.* **2005**, *337*, 1–10. [CrossRef]
5. Arias, M.; Perez-Novo, C.; Osorio, F.; López, E.; Soto, B. Adsorption and desorption of copper and zinc in the surface layer of acid soils. *J. Colloid Interface Sci.* **2005**, *288*, 21–29. [CrossRef]

6. Liao, K.H.; Xu, S.H.; Wu, J.C.; Ji, S.H.; Lin, Q. Cokriging of soil cation exchange capacity using the first principal component derived from soil physico-chemical properties. *Agric. Sci. China* **2011**, *10*, 1246–1253. [CrossRef]
7. Xia, X.Q.; Ji, J.F.; Chen, J.; Liao, Q.L.; Yang, Z.F. Analysis of soil physical and chemical properties by reflectance spectroscopy. *Earth Sci. Front.* **2009**, *16*, 354–362.
8. Zhang, Z.P.; Ding, J.L.; Zhu, C.M.; Wang, J.Z.; Ma, G.L.; Ge, X.Y.; Li, Z.S.; Han, L.J. Strategies for the efficient estimation of soil organic matter in salt-affected soils through Vis-NIR spectroscopy: Optimal band combination algorithm and spectral degradation. *Geoderma* **2021**, *382*, 114729. [CrossRef]
9. Munnaf, M.A.; Guerrero, A.; Nawar, S.; Haesaert, G.; Meirvenne, M.V.; Mouazen, A.M. A combined data mining approach for on-line prediction of key soil quality indicators by Vis-NIR spectroscopy. *Soil Tillage Res.* **2021**, *205*, 104808. [CrossRef]
10. Vohland, M.; Ludwig, B.; Seidel, M.; Hutengs, C. Quantification of soil organic carbon at regional scale: Benefits of fusing vis-NIR and MIR diffuse reflectance data are greater for in situ than for laboratory-based modelling approaches. *Geoderma* **2022**, *405*, 115426. [CrossRef]
11. Biney, J.K.M.; Blocher, J.R.; Bell, S.M.; Boruvka, L.; Vasat, R. Can in situ spectral measurements under disturbance-reduced environmental conditions help improve soil organic carbon estimation? *Sci. Total Environ.* **2022**, *838*, 156304. [CrossRef] [PubMed]
12. Azizi, K.; Ayoubi, S.; Demattê, J.A.M. Controlling factors in the variability of soil magnetic measures by machine learning and variable importance analysis. *J. Appl. Geophys.* **2023**, *210*, 104944. [CrossRef]
13. Annam, S.; Singla, A. Estimating the concentration of soil heavy metals in agricultural areas from AVIRIS hyperspectral imagery. *Int. J. Intell. Syst. Appl. Eng.* **2023**, *11*, 156–164.
14. Viscarra Rossel, R.A.; Walvoort, D.J.J.; Mcbratney, A.B.; Janik, L.J.; Skjemstad, J.O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **2006**, *131*, 59–75. [CrossRef]
15. Li, D.; Chen, S.S.; Chen, X.Z.; Peng, Z.P. Study of Near infrared Spectroscopy Assessment for Soil Exchangeable K, Ca, Mg and CEC in Lychee Orchard. *Trop. Geogr.* **2011**, *31*, 368–372.
16. Gras, J.P.; Barthès, B.G.; Mahaut, B.; Trupin, S. Best practices for obtaining and processing field visible and near infrared (VNIR) spectra of topsoils. *Geoderma* **2014**, *214–215*, 126–134. [CrossRef]
17. Zhao, D.X.; Arshad, M.; Wang, J.; Triantafilis, J. Soil exchangeable cations estimation using Vis-NIR spectroscopy in different depths: Effects of multiple calibration models and spiking. *Comput. Electron. Agric.* **2021**, *182*, 105990. [CrossRef]
18. Li, M.Z.; Zhang, W.C.; Sun, Z.T.; Zhao, Y.G.; Wang, S.J. Effects of different extractants on determination results of fluorine in saline-alkali soil. *Exp. Technol. Manag.* **2021**, *38*, 33–39.
19. Leone, A.P.; Viscarra-Rossel, R.A.; Amenta, P.; Buondonno, A. Prediction of soil properties with plsr and vis-nir spectroscopy: Application to mediterranean soils from southern italy. *Curr. Anal. Chem.* **2012**, *8*, 283–299. [CrossRef]
20. Cai, T.Y.; Wang, Z.G.; Yang, L.S.; Wang, Q.; Huang, H.J.; Yu, H.Y.; Zhang, C.Z.; Zhang, C.; Liu, P.; Feng, Y.Q.; et al. Hyperspectral inversion model of Zn in high standard farmland soil in Xiping County. *J. Agro-Environ. Sci.* **2022**, *41*, 2223–2231.
21. Zou, X.B.; Zhao, J.W.; Malcolm, J.W.P.; Mel, H.; Mao, H.P. Variables selection methods in near-infrared spectroscopy. *Anal. Chim. Acta* **2010**, *667*, 14–32.
22. Tavares, T.R.; Mouazen, A.M.; Nunes, L.C.; Santos, F.R.D.; Melquiades, F.L.; Silva, T.R.D.; Krug, F.J.; Molin, J.P. Laser-Induced Breakdown Spectroscopy (LIBS) for tropical soil fertility analysis. *Soil Tillage Res.* **2022**, *216*, 105250. [CrossRef]
23. Yang, X.Y.; Bao, N.S.; Li, W.W.; Liu, S.J.; Fu, Y.H.; Mao, Y.C. Soil Nutrient Estimation and Mapping in Farmland Based on UAV Imaging Spectrometry. *Sensors* **2021**, *21*, 3919. [CrossRef]
24. Cheng, Z.; Zhang, L.Q.; Liu, H.Y.; Zhu, A.S. Successive projections algorithm and its application to selecting the wheat near-infrared spectral variables. *Spectrosc. Spectr. Anal.* **2010**, *30*, 949–952.
25. Kursa, M.B.; Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
26. Guo, B.; Bai, H.R.; Zhang, B.; Pei, L.; Zhao, Y.H.; Lei, Y.Z.; Yuan, R.C. Inversion of soil zinc contents using hyperspectral remote sensing based on random forest and continuous wavelet transform in an opencast coal mine. *Trans. CSAE* **2022**, *38*, 138–147.
27. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
28. Koza, J.R.; Rice, J.P. Genetic generation of both the weights and architecture for a neural network. In Proceedings of the IJCNN-91-Seattle International Joint Conference on Neural Networks, Seattle, WA, USA, 8–12 July 1991; Volume 2, pp. 397–404.
29. Verikas, A.; Gelzinis, A.; Bacauskiene, M. Mining data with random forests: A survey and results of new tests. *Pattern Recognit.* **2011**, *44*, 330–349. [CrossRef]
30. Hong, Y.S.; Chen, S.C.; Chen, Y.Y.; Linderman, M.; Mouazen, A.M.; Liu, Y.L.; Guo, L.; Yu, L.; Liu, Y.F.; Cheng, H.; et al. Comparing laboratory and airborne hyperspectral data for the estimation and mapping of topsoil organic carbon: Feature selection coupled with random forest. *Soil Tillage Res.* **2020**, *199*, 104589. [CrossRef]
31. Wu, Z.M.; Li, J.C.; Wang, R.; Shi, L.; Mao, S.; Lv, H.; Li, Y.M. Estimation of CDOM concentration in inland lake based on random forest using Sentinel-3A OLCI. *J. Lake Sci.* **2018**, *30*, 979–991.
32. Tziolas, N.; Tsakiridis, N.; Ben-Dor, E.; Theocharis, J.; Zalidis, G. Employing a Multi-Input Deep Convolutional Neural Network to Derive Soil Clay Content from a Synergy of Multi-Temporal Optical and Radar Imagery Data. *Remote Sens.* **2020**, *12*, 1389. [CrossRef]

33. Song, K.S.; Li, L.; Tedesco, L.P.; Li, S.; Duan, H.T.; Liu, D.W.; Hall, B.E.; Du, J.; Li, Z.C.; Shi, K.; et al. Remote estimation of chlorophyll-a in turbid inland waters: Three-band model versus GA-PLS model. *Remote Sens. Environ.* **2013**, *136*, 342–357. [CrossRef]
34. Zornoza, R.; Guerrero, C.; Mataix-Solera, J.; Scow, K.M.; Arcenegui, V.; Mataix-Beneyto, J. Near infrared spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils. *Soil Biol. Biochem.* **2008**, *40*, 1923–1930. [CrossRef]
35. Bao, N.S.; Wu, L.X.; Ye, B.Y.; Yang, K.; Zhou, W. Assessing soil organic matter of reclaimed soil from a large surface coal mine using a field spectroradiometer in laboratory. *Geoderma* **2017**, *288*, 47–55. [CrossRef]
36. Mohamed, E.S.; Baroudy, A.A.E.; El-beshbeshy, T.; Emam, M.; Belal, A.A.; Elfadaly, A.; Aldosari, A.A.; Ali, A.M.; Lasaponara, R. Vis-NIR Spectroscopy and Satellite Landsat-8 OLI Data to Map Soil Nutrients in Arid Conditions: A Case Study of the Northwest Coast of Egypt. *Remote Sens.* **2020**, *12*, 3716. [CrossRef]
37. Nyarko, F.; Tack, F.M.G.; Mouazen, A.M. Potential of visible and near infrared spectroscopy coupled with machine learning for predicting soil metal concentrations at the regional scale. *Sci. Total Environ.* **2022**, *841*, 156582. [CrossRef] [PubMed]
38. Pyo, J.C.; Hong, S.M.; Kwon, Y.S.; Kim, M.S.; Cho, K.H. Estimation of heavy metals using deep neural network with visible and infrared spectroscopy of soil. *Sci. Total Environ.* **2020**, *741*, 140162. [CrossRef]
39. Greenberg, I.; Seidel, M.; Vohland, M.; Ludwig, B. Performance of field-scale lab vs in situ visible/near- and mid-infrared spectroscopy for estimation of soil properties. *Eur. J. Soil Sci.* **2022**, *73*, e13180. [CrossRef]
40. Chen, Y.T.; Gao, S.B.; Jones, E.J.; Singh, B. Prediction of Soil Clay Content and Cation Exchange Capacity Using Visible Near-Infrared Spectroscopy, Portable X-ray Fluorescence, and X-ray Diffraction Techniques. *Environ. Sci. Technol.* **2021**, *55*, 4629–4637. [CrossRef]
41. Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.* **2012**, *13*, 1063–1095.
42. Hermansen, C.; Norgaard, T.; Jonge, L.; Moldrup, P.; Müller, K.; Knadel, M. Predicting glyphosate sorption across New Zealand pastoral soils using basic soil properties or Vis-NIR spectroscopy. *Geoderma* **2020**, *360*, 114009. [CrossRef]