plants

MDPI

Article

Nutrient Diagnosis of *Eucalyptus* at the Factor-Specific Level Using Machine Learning and Compositional Methods

Betania Vahl de Paula [1,*], Wagner Squizani Arruda [1], Léon Etienne Parent [1,2], Elias Frank de Araujo [3] and Gustavo Brunetto [1]

1 Departemento dos Solos, Universidade Federal de Santa Maria, Av. Roraima, 1000-Camobi, Santa Maria-RS 97105-900, Brazil; wagnersquizani@hotmail.com (W.S.A.); leon-etienne.parent@fsaa.ulaval.ca (L.E.P.); brunetto.gustavo@gmail.com (G.B.)
2 Department of Soils and Agrifood Engineering, Laval University, Quebec, QC G1V 0A6, Canada
3 Soil and Management Researcher of CMPC-Cellulose Rio Grandense, Rua São Geraldo 1680-Guaíba–RS, Brazil; elias.araujo@cmpcrs.com.br
* Correspondence: behdepaula@hotmail.com; Tel.: +55-5532177117

Received: 21 July 2020; Accepted: 13 August 2020; Published: 18 August 2020

Abstract: Brazil is home to 30% of the world's *Eucalyptus* trees. The seedlings are fertilized at plantation to support biomass production until canopy closure. Thereafter, fertilization is guided by state standards that may not apply at the local scale where myriads of growth factors interact. Our objective was to customize the nutrient diagnosis of young *Eucalyptus* trees down to factor-specific levels. We collected 1861 observations across eight clones, 48 soil types, and 148 locations in southern Brazil. Cutoff diameter between low- and high-yielding specimens at breast height was set at 4.3 cm. The random forest classification model returned a relatively uninformative area under the curve (AUC) of 0.63 using tissue compositions only, and an informative AUC of 0.78 after adding local features. Compared to nutrient levels from quartile compatibility intervals of nutritionally balanced specimens at high-yield level, state guidelines appeared to be too high for Mg, B, Mn, and Fe and too low for Cu and Zn. Moreover, diagnosis using concentration ranges collapsed in the multivariate Euclidean hyper-space by denying nutrient interactions. Factor-specific diagnosis detected nutrient imbalance by computing the Euclidean distance between centered log-ratio transformed compositions of defective and successful neighbors at a local scale. Downscaling regional nutrient standards may thus fail to account for factor interactions at a local scale. Documenting factors at a local scale requires large datasets through close collaboration between stakeholders.

Keywords: compatibility intervals; Euclidean distance; Humboldtian loci; centered log ratios; machine learning

1. Introduction

Eucalyptus plantations cover 20×10^6 ha worldwide to provide raw material for wood, paper, biofuel, firewood, and charcoal [1]. Brazil is the world leader, producing *Eucalyptus* on 6×10^6 ha with an average yield of 36 m^3 ha^{-1} $year^{-1}$ [2]. While *Eucalyptus* is adapted to low-fertility soils, nutrient supply, especially N and K [3–6] can limit stand productivity [1,7,8]. Fertilization was found to increase wood production of *Eucalyptus grandis* by 28% and irrigation by another 30% to reach potential outcome of 83 m^3 ha^{-1} $year^{-1}$, where the most yield-impacting factors are set at near optimum levels [9].

Eucalyptus seedlings are heavily fertilized at planting to prevent nutrient deficiency and non-uniform tree growth until canopy closure [8]. Thereafter, fertilization aims to recharge the soil–plant system with nutrients where initial inputs appeared ineffective. Fertilization decisions are

usually taken based on soil and tissue tests. Plant tissue tests are thought to integrate the effects of growth factors on crop performance [10]. Regional tissue standards [11] have been developed to guide fertilization of *Eucalyptus* seedlings [8,12,13] and of trees more than 6 years of age [14]. No standards have been developed for trees of intermediate age.

Tissue tests are generally interpreted using general nutrient concentration ranges or nutrient ratios. First, the statistical treatment of concentration values to generate intervals may lead to biased or wrong results [15]. In addition, the concept of statistically-derived ranges has been recently challenged by a concept of "compatibility intervals" to avoid taking wrong dichotomous decisions on rejection [16]. Furthermore, regional nutrient ratios or product standards and expressions have been elaborated based on heroic assumptions such as universality and timeless nutrient norms, and function additivity [17].

The suitability of downscaling regional standards for application at a local scale where myriads of factor interactions occur has been minimally addressed [18]. Errors on interactions [19] that involve environmental factors, genetics, nutrients and time may reduce diagnostic efficiency because factor effects are averaged across factors at a regional scale. Several factors can affect plant growth [20]. Soil type, climatic conditions [11], clone nutrient-use efficiency [21,22], and management factors such as stand quality, tree spacing, fertilization, and even tree pruning and thinning [8] vary widely, leading to contrasting fertilizer recommendations for *Eucalyptus* stands [9]. Nevertheless, data sets must be well documented to customize nutrient diagnosis at the specified combination of factors.

Humboldtian principles of quantitative biogeography require integrating data collected in living systems [23]. Humboldtian patterns can be extracted using methods of artificial intelligence to solve complex problems that are beyond human capabilities [24,25]. A heuristically simple factor-specific diagnostic approach is to compare defective and successful Humboldtian loci across a set of features using compositional and classification or regression machine learning (ML) methods [18,26,27]. In such a case, the assumption that factors other than the ones being addressed are equal or at optimum levels [28] is replaced by the assumption that documented factors other than the ones being addressed are comparable. Only non-documented factors must be assumed to be equal.

Compositional data are strictly positive data with constraints such as closure to measurement unit or scale, missing values, data censoring, ethical data collection, data merging, levelling of different datasets from various sources, sample design [29], accuracy of measurements, and handling of zeroes [30]. To handle numerical constraints, compositional data should be log-ratio transformed before conducting statistical analyses [15,31]. Machine learning methods can also unravel complex patterns in data [24,25]. Machine learning (ML) and compositional data analysis (CoDa) methods thus provide unprecedented tools to conduct factor-specific nutrient diagnosis and verify the relevance of downscaling regional standards at a local scale.

We hypothesized that (1) the productivity of *Eucalyptus* following plantation depends on tissue composition and local features, (2) log-ratio transformations increase the accuracy of ML models, and (3) regional diagnosis can be downscaled reliably to factor-specific levels. Our objective was to customize tissue nutrient diagnosis of young *Eucalyptus* trees at a local scale.

2. Materials and Methods

2.1. Data Set

The data set comprised 1861 observations on young *Eucalyptus* trees across eight clones, 48 soil types (with predominance of types Typic Hapludalf and Udorthent), and 148 locations in southern Brazil. Most trees (97%) were between 0.9 and 1.1 years old following plantation. The clones were *Eucalyptus* spp. (*E. benthamii, E. saligna, E. dunnii, E. urophylla, E. urophylla S. T. Blake, E. urophylla* × *E. globulus, E. urophylla* × *E. grandis, E. urophylla Blake* × *E. grandis Hill, E. camaldulensis* × *E. grandis* × *E. urophylla*) collected on the Coastal Region of Rio Grande do Sul, Southern Brazil. Tree seedlings had been grown in 100 mL containers for 5–6 months to reach a root collar diameter of 3–4 mm and plant height of 30–40 cm before planting [8]. Tree spacing was 3 m by 3 m for an average plantation density of 1100 trees ha^{-1}.

The regional climate is humid temperate subtropical according to the international Köppen-Geiger classification. Winters are moderately cold with frost. Summers are hot with day temperatures most often >30 °C. Rainfall is well distributed throughout the year, with annual accumulations ranging from 1000 mm to >2000 mm [32]. The soil classification was coded at each site according to the Brazilian soil classification system [33]. The data set did not include pest management, pest damage, and meteorology.

2.2. Fertilization

Soil tests were not available, but fertilization followed regional guidelines [11]. At plantation, fertilizers were manually applied in planting holes or grooves, or besides tree seedlings, and then mixed with soil. Fertilization rates varied between 15 and 45 kg N ha^{-1} or more depending on soil organic matter content. The P (0–57 kg P ha^{-1}) and K (0–108 kg K ha^{-1}) fertilization depended on soil P and K tests, respectively. Thereafter, P and K fertilizers were applied at rates of up to 22 kg P ha^{-1} and up to 40 kg K ha^{-1}, respectively, based on regional tissue nutrient standards. Additional N supply of 15–45 kg N ha^{-1} depended on soil organic matter content and wood marginal yield exceeding 40 m^3 ha^{-1} $year^{-1}$. Micronutrient levels could have been impacted by applications of composts, fertilization, fungicides, and lime. Micronutrients were applied as needed at rates of 1 kg B ha^{-1}, 1.5 kg Zn ha^{-1}, and 1 kg Cu ha^{-1}.

2.3. Plant Measurements and Analysis

Yearly, between January and March, plant height was measured using a metric tape. Stem diameter was measured as diameter at breast height (DBH ≈1.3 m in height). Plant height and tree diameter are closely related to the wood volume of *Eucalyptus* [9]. The DBH was thus used as a target variable to run ML models.

Yearly, from February to April, leaves were collected in the middle tier of the annual growth (4th to 5th leaf from branch tip) from at least ten trees per site. Eleven nutrients were analyzed [34]. Foliar N was quantified by micro-Kjeldahl. The S, P, K, Ca, Mg, Zn, Cu, Mn, Fe, and B foliar concentrations were determined by ICP-OES after digestion in a mixture of nitric and perchloric acids.

2.4. Log-Ratio Transformation Techniques

Before the work of Aitchison [15], compositions were addressed using concentrations or pairwise ratios between components x_i and x_j and expressed as x_i/x_j [35]. Pairwise ratios required (1) selecting x_i/x_j or its inverse x_j/x_i based on variance ratios between low-yielding and high-yielding subpopulations, (2) reflective equations, and (3) assumptions on additivity to compute functions and indices [17]. While the logarithmic scale avoids large numbers of decimals [36], log-transformed pairwise ratios allows recovering reflectivity, i.e., $ln\left(x_i/x_j\right) = -ln\left(x_j/x_i\right)$. There are $D \times (D-1)/2$ pairwise log ratios (pwlr) derived from D concentration data that generate redundant information in multivariate models.

The pwlr computed as $ln\left(x_i/x_j\right)$ is also called a log contrast, i.e., $ln\left(x_i/x_j\right) = ln(x_i) - ln\left(x_j\right)$. The composition is closed to some total by computing a filling value (Fv) between the total and the sum of quantified components. The pwlr values for a given nutrient can be compressed into a single centered log ratio (clr) [26], as follows for tissue N:

$$clr_N = ln\left(\frac{N}{G}\right) = ln\left(\sqrt[11]{\frac{N}{N} \times \frac{N}{P} \times \frac{N}{K} \times \frac{N}{Mg} \times \frac{N}{Ca} \times \frac{N}{B} \times \frac{N}{Cu} \times \frac{N}{Zn} \times \frac{N}{Mn} \times \frac{N}{Fe} \times \frac{N}{Fv}}\right) = \frac{1}{11} \times \left[ln\left(\frac{N}{N}\right) + ln\left(\frac{N}{P}\right) + \ldots ln\left(\frac{N}{Fv}\right)\right] = \frac{1}{11}\left[ln\left(\frac{N}{P}\right) + \ldots ln\left(\frac{N}{Fv}\right)\right] \quad (1)$$

where N is the tissue nitrogen concentration, and G is the geometric mean across components (including the nutrient itself and the filling value), all expressed using the same measurement unit or scale. The computation of G does not accept missing data unless imputed or approximated from detection limits [30]. The clr transformation provides a solid mathematical ground for the integration of dual ratios [37] and avoids assumptions on additivity and reflectivity as required for Diagnosis and

Recommendation Integrated System (DRIS) computations [38]. The clr transformation can account for all dual nutrient interactions and therefore reduces the inter-relationships among nutrients compared to raw concentrations as shown by the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy in principal component analysis [39].

The D-part compositions can be compressed into D-1 isometric log ratios or orthonormal balances [31], the exact number of degrees of freedom available in compositions [40]. The orthonormal balances between selected subsets of components at the numerator and denominator are computed as follows:

$$ilr_i = \sqrt{\frac{rs}{r+s}} ln\left(\frac{G_N}{G_D}\right) \quad (2)$$

where r and s are numbers of components at the numerator and denominator, respectively, and G_N and G_D are geometric means of components at the numerator and denominator, respectively.

Orthogonality is a concept of linear independence [41]. The *ilr* transformation is the most appropriate log ratio transformation technique to conduct multivariate analysis of compositional data, avoiding spurious correlations and singular matrix [42]. While orthonormal balances can be arranged into meaningful combinations in line with the objectives of the study [43], multivariate distances and the results of multivariate analysis remain the same whatever the arrangement of components into balances, due to orthogonality between *ilr* variables.

2.5. Regional Diagnosis

The *clr* indices are computed from mean and standard deviation of *clr* values for the nutritionally balanced subpopulation as follows [37]:

$$I_i = \frac{\left(clr_i - clr_i^*\right)}{SD_i^*} \quad (3)$$

where I_i is the *clr* index of nutrient i, clr_i is the *clr* value for the diagnosed specimen, and clr_i^* and SD_i^* are the mean and standard deviation of nutrient i used as references. Nutrient indices are ranked in the order of their limitation to yield from the most negative to the most positive *clr* index. To assign a probability level to D-parts compositions, Compositional Nutrient Diagnosis (CND) indices may be added up to a squared multivariate distance distributed like a proximate χ^2 variable with D−1 degrees of freedom [44].

2.6. Local Diagnosis

To conduct nutrient diagnoses at a factor-specific level, the Euclidean distance ϵ between two D-part compositions can be computed across *clr* values as follows [26]:

$$\epsilon = \sqrt{\sum_{i=1}^{D}\left(clr_i - clr_i^*\right)^2} = \sqrt{\sum_{i=1}^{D}\left(ilr_i - ilr_i^*\right)^2} \quad (4)$$

where clr_i and clr_i^* or ilr_i and ilr_i^* represent high yield and nutrient balance TN compositions, respectively. Successful TN specimens are productive specimens showing a small Euclidean distance from the diagnosed specimen. Because $\sum_{i=1}^{D} clr_i = 0$, nutrients can be ranked in the numerical order of *clr* differences from the most negative (relative shortage) to the most positive (relative excess).

2.7. Statistical Analysis

The *clr* biplot was drawn using the freeware Codapack 2.02.21 (http://ima.udg.edu/codapack/) to document the relative contribution of nutrient concentrations to tissue compositions. The ML classification models were run using the freeware Orange vs. 3.24 (Bioinformatics Lab, Ljubljana, Slovenia) by relating crop yield (target variable) to growth-impacting features. Overfitting due to too

many features could be handled by ML models [45]. Nevertheless, this is a key issue in ML because the size and number of features differ between concentration, pairwise log ratio (pwlr), centered log ratio (clr) and isometric log ratio (ilr) expressions and this may impact the model accuracy [35].

The *Eucalyptus* population was partitioned into low- and high-yielding subpopulations based on a critical DBH of 4.3 cm as an economically viable yield target. The random forest (RF), neural network (NN), naïve Bayes, support vector machine (SVM), KNN, Adaboost, and stochastic gradient decent (SGD) models were tested in cross-validation. The results of ten successive runs were averaged after randomly removing 10% of the data. Model accuracy was assessed by area under the curve (AUC). An AUC between 0.7 and 0.9 is informative [46]. The contribution of features to model accuracy can be assessed by removing one feature at the time. The confusion matrix of the machine learning model classified specimens into four quadrants as follows [47]:

True negative specimens (TN): high productivity and adequate nutrient balance (negative response to fertilization). They are located in the upper left quadrant of the confusion matrix.

False negative specimens (FN): low productivity despite adequate nutritional balance (negative response to fertilization, some other factor limiting yield). They are located in the lower left quadrant of the confusion matrix.

False positive specimens (FP): high productivity despite nutrient imbalance (contamination, sub-optimal concentration, excess or luxury consumption of some nutrient). They are located in the upper right quadrant of the confusion matrix.

True positive specimens (TP): low productivity and nutritional imbalance (positive response to fertilization). They are located in the lower right quadrant of the confusion matrix.

Classification accuracy (CA) was computed as follows [44]:

$$CA = \frac{TN + TP}{TN + FN + TP + FP} \quad (5)$$

Data partitioning followed principles of data interpretation similar to those used for the human response to drugs in clinical biology [46]. Data partitioning in the confusion matrix avoids merging balanced and imbalanced specimens at high yield level as in DRIS [17,38]. Nutrient imbalance of high yielding specimens is due to over-fertilization leading to luxury consumption of nutrients that should be avoided, or to nutrient contamination that unduly increases the variation of nutrient levels and could bias nutrient diagnosis. Nutrient compatibility intervals [16] at a high yield level were computed from TN quartiles. While FN specimens are also nutritionally balanced and could be considered to compute reference values at a regional scale, they do not provide realistic yield targets as shown in the data set at a local scale. Successful TN specimens are local references to correct defective compositions at the specified combination of factors.

3. Results

3.1. Descriptive Statistics and Exploratory Analyses

There was a large variation in tissue compositions (Table 1).

The *clr* biplot showed that Zn, B and Fe contributed the most to total variance of *Eucalyptus* tissue compositions (Figure 1), indicating wide variation in soil genesis (Fe) and management decisions such as applications of fungicides and organic residues. The large variation in Zn and B levels may have been impacted by composts, fertilization, fungicides, and liming.

Table 1. Ranges of tissue nutrient concentrations and filling values for 1861 young *Eucalyptus* trees in Southern Brazil.

Component	Minimum	Median	Maximum
		g kg^{-1}	
N	9.1	21.9	38.8
P	0.5	1.3	3.3
K	1.2	9.4	19.6
Mg	1.0	2.6	7.5
Ca	2.7	8.6	34.9
S	0.4	1.5	5.1
B	0.011	0.038	0.105
Cu	0.001	0.008	0.036
Zn	0.006	0.018	0.129
Mn	0.066	0.964	4.954
Fe	0.002	0.076	0.594
Filling value	925.6	952.1	973.3

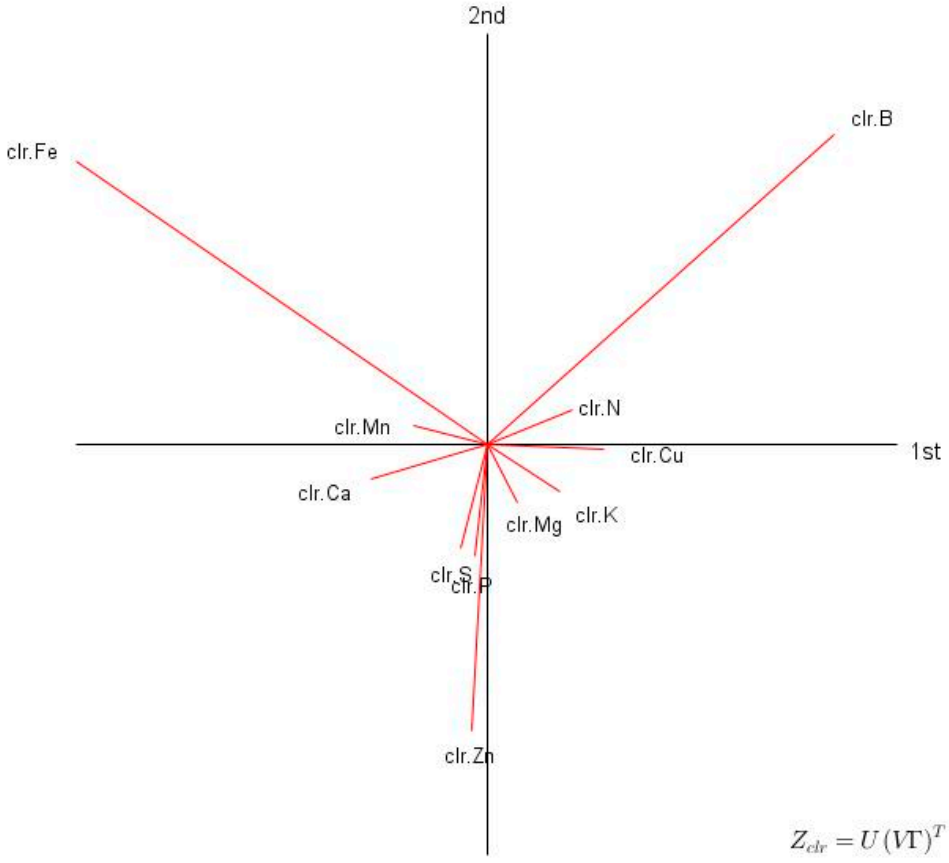

Figure 1. Clr biplot (CoDapack v2.02.21) of tissue compositions of 1861 young *Eucalyptus* trees in southern Brazil.

3.2. Machine Learning Models

The RF, naïve Bayes and NN models were found to be informative (Table 2). Adaboost, SVM, KNN and SGD were not informative (AUC < 0.7).

The RF model was preferred to the naïve Bayes model to avoid assumptions on feature independence in interactive Humboldtian living systems such as *Eucalyptus* ecosystems. While the RF model can deal with over-fitting of partition trees, it may be affected by the choice of the expression [35]. The raw concentration and *clr* expressions returned the highest accuracies (Table 3). The raw

concentration expression is preferable because the model is not affected by missing or zero values that impair computing log ratios.

Table 2. Accuracies of ML classification models for the *Eucalyptus* data set (1861 observations) for nutrient concentrations and other features in cross-validation, using diameter at breast height (DBH) as the target variable at a BDH cut-off of 4.3 cm between high- and low-yielding trees.

Expression	AUC	CA	TN	FN	FP	TP
Random Forests	0.787	0.718	521	219	315	816
Neural Networks	0.778	0.705	548	271	278	764
Naïve Bayes	0.793	0.715	614	318	212	717
Support Vector Machine	0.544	0.529	-	-	-	-
KNN	0.589	0.570	-	-	-	-
Adaboost	0.636	0.641	-	-	-	-
Stochastic Gradient Decent	0.674	0.679	-	-	-	-

AUC = area under the curve (≥0.7 required); CA = classification accuracy; TN = true negative; FN = false negative; FP = false positive; TP = true positive.

Table 3. Comparison of accuracies from nutrient expressions using the RF model to process the Eucalyptus tree data set (1861 observations) in cross-validation.

Nutrient Expression	Area Under Curve	Classification Accuracy
Raw concentration data	0.787	0.718
Pairwise log ratios	0.721	0.664
Centered log ratios	0.785	0.706
Isometric log ratios	0.776	0.701

3.3. Nutrient Intervals at a Regional Scale

Regional nutrient standards can be assessed from TN quartiles, which are specimens showing high yield and adequate nutrient balance. For this reason, TN specimens are considered the reference compositions for diagnostic purposes at a local scale. Where the number of TN specimens is too small, FN specimens could also be considered at a regional scale. Among the 529 TN specimens, 40 were outside the target age range of 0.9–1.1-years-old and were thus discarded, leaving 489 TN specimens to compute the TN quartile compatibility intervals concentration (Table 4).

Table 4. State concentration ranges and quartile compatibility intervals (0.25, 0.75) nutrient values of 489 TN specimens of 0.9–1.1-year-old *Eucalyptus*.

Nutrient	State (Gatiboni et al. [11])		True Negative Quartiles (25, 75)	
	Lower bound	Upper bound	Lower bound	Upper bound
		$g\ kg^{-1}$		
N	15.0	20.0	17.0	25.3
P	1.0	1.3	1.0	1.4
K	9.0	13.0	7.2	11.5
Mg	6.0	10.0	2.3	3.2
Ca	5.0	8.0	7.0	10.2
S	1.5	2.0	1.2	1.8
		$mg\ kg^{-1}$		
B	30	50	6	12
Cu	7	10	14	21
Zn	35	50	60	96
Mn	400	600	34	54
Fe	150	200	679	1281

The present Brazilian standards for *Eucalyptus* [11] overlapped the across-factor quartile ranges of the TN specimens for N, P, K, and Ca, and but were out of range for Mg and micronutrients. As shown in Table 5, state standards and TN quartiles returned similar diagnoses 12 times out of 22 attempts, indicating a high risk of wrong fertilization decisions. The decision to implement corrective measures is thus influenced by the choice of specific boundaries for compatibility intervals. Machine learning model prediction and compositional analysis tools can avoid diagnosing nutrient levels using fixed compatibility intervals.

Table 5. Nutrient concentrations of fictive specimens diagnosed by state standards and TN quartiles in Table 4 (italicized diagnoses are similar).

Nutrient	Site #1	Site #2	Site #1		Site #2	
	g kg^{-1}		State Standards	TN Quartiles	State Standards	TN Quartiles
N	27.1	15.0	*High*	*High*	Normal	Low
P	1.4	1.3	High	Normal	*Normal*	*Normal*
K	8.8	8.2	Low	Normal	Low	Normal
Mg	1.5	3.8	*Low*	*Low*	Low	High
Ca	3.9	21.2	*Low*	*Low*	*High*	*High*
S	1.7	1.4	*Normal*	*Normal*	Low	Normal
	mg kg^{-1}			Diagnosis		
B	48.0	1.3	Normal	High	*Low*	*Low*
Cu	4.7	17.9	*Low*	*Low*	High	Normal
Zn	14.7	151.8	*Low*	*Low*	*High*	*High*
Mn	452.3	73.8	Normal	High	Low	High
Fe	66.9	1614.4	*Low*	*Low*	*High*	*High*

3.4. Regional vs. Local Diagnosis

Regional diagnosis is conducted by computing *clr* indices from *clr* means and standard deviations of TN specimens, assuming that factors other than the nutritional ones are equal or at near-optimum levels at a regional scale (Table 6). At a local scale, uncontrollable factors (e.g., soil profile) or ones that are difficult to control (e.g., P, Cu, Zn and Fe accumulation in soil) could be accounted for by site analogy. The ML prediction model compares factor-specific defective compositions to the closest TN specimens sharing the same features. The criterion for closeness between compositions is the Euclidean distance at the specified combination of factors. We selected ten close TN specimens to conduct nutrient diagnosis at the specified combination of factors.

Table 6. Centered log ratio statistics for 0.9–1.1-year-old TN specimens of *Eucalyptus* as regional nutrient standards across features.

Nutrient	489 TN Specimens	
	Mean	**Standard Deviation**
N	2.9050	0.3048
P	0.0726	0.2618
K	2.1387	0.2878
Mg	0.8880	0.2270
Ca	2.0454	0.2962
B	−4.8438	0.4189
S	0.3053	0.3024
Cu	−4.0882	0.3872
Zn	−2.7212	0.4089
Mn	−3.2629	0.3338
Fe	−0.2132	0.4754
Fv	6.7743	0.1453

Factor analogy between defective and successful specimens at the specified combination of factors is assessed in the TN data set to diagnose nutrient problems in defective specimens. At the clone × age interaction level, AUC of the RF model was 0.71 and the model was still informative. The RF model predicted that the probabilities for the diagnosed specimens in Table 5 to be classified as high yielders were 48% and 36% at sites #1 and #2, respectively, indicating a need for corrective measures.

We selected successful neighbors showing DBH > 5 cm at the clone × age interaction level and where nutrient requirements were most parsimonious to minimize cost of adjusting nutrient management. Regional diagnosis using *clr* standards in Table 5 and local diagnosis of the two defective compositions are illustrated in the form of histograms in Figures 2 and 3. Close successful neighbors reached a DBH of 5.43–5.44 cm compared to 4.06 cm and 1.71 at sites #1 and #2, respectively, indicating high potential to boost plant growth using appropriate corrective measures.

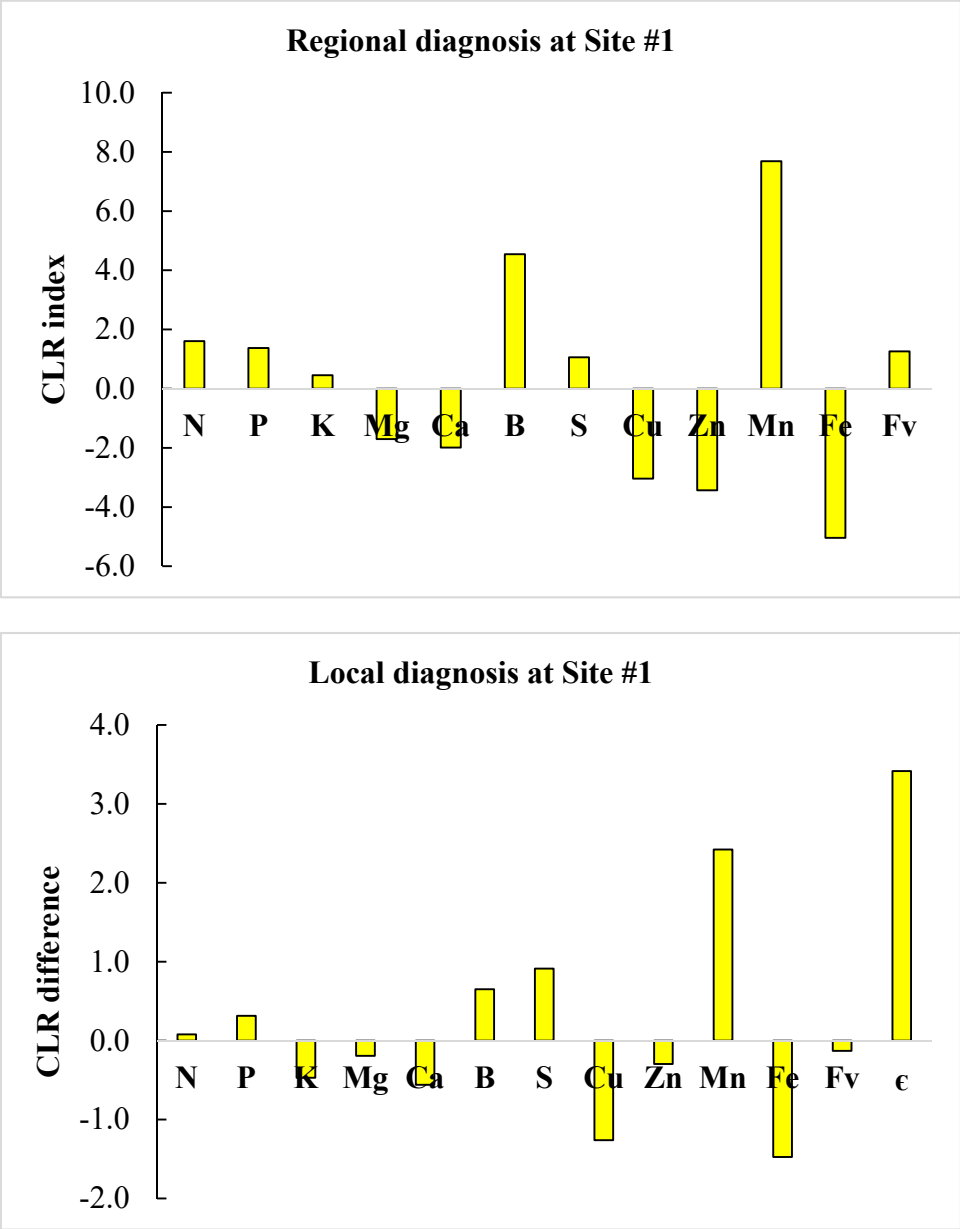

Figure 2. Comparison between regional (**top**) and local (**bottom**) nutrient diagnoses at Site #1 using centered log ratios (CLR) of regional TN standards or a successful local neighbor as measured by the Euclidian distance (ϵ).

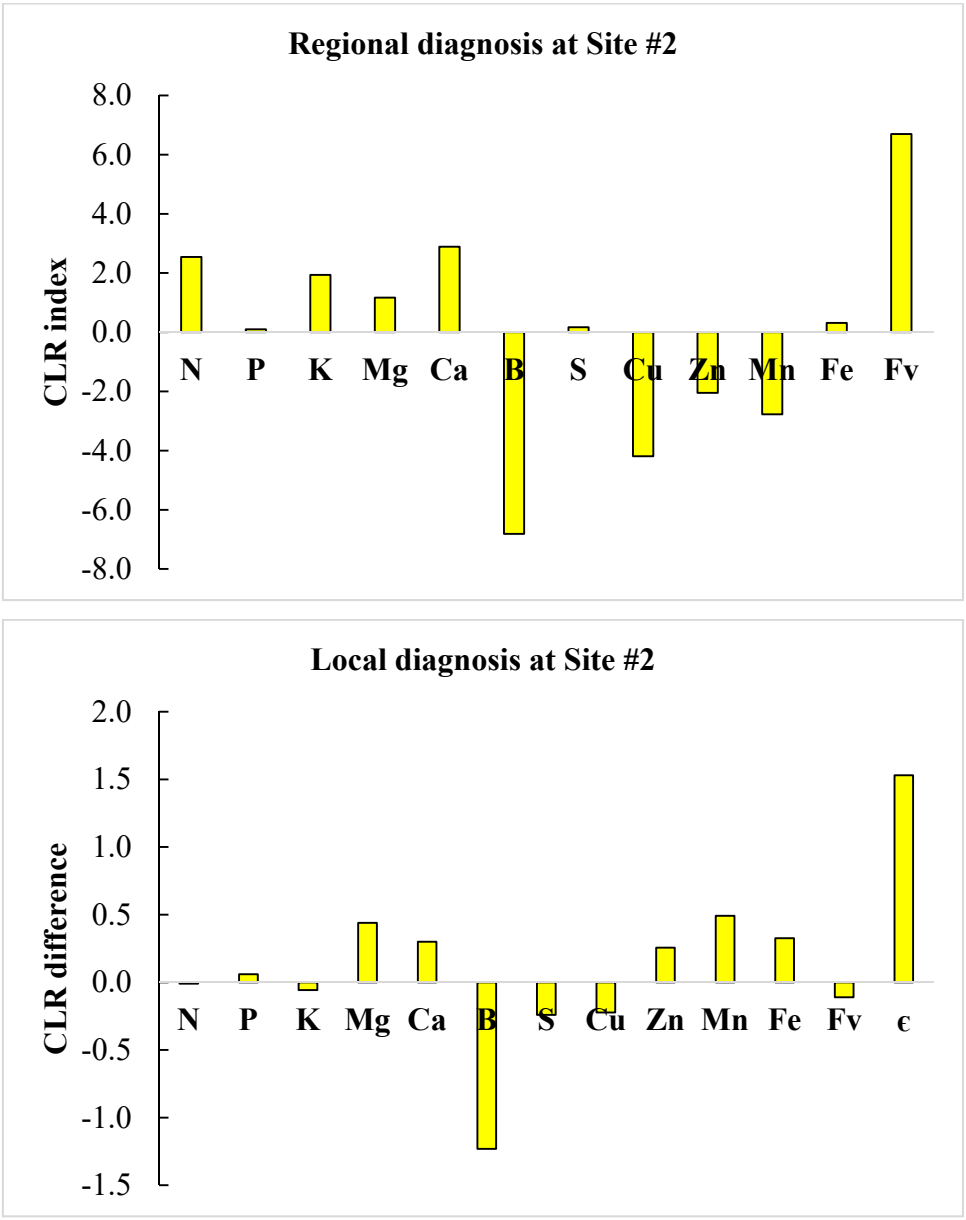

Figure 3. Comparison between regional (**top**) and local (**bottom**) nutrient diagnoses at Site #2 using centered log ratios (CLR) of regional TN standards or a successful local neighbor as measured by the Euclidian distance (ϵ).

At site #1, there was relative Mn excess at both regional and local scales. At the regional scale, B and N ranked second and third in relative excess, while Fe, Zn and Cu ranked in a descending order of relative nutrient shortage. At the local scale, S and B ranked second and third in relative excess, while Fe and Cu ranked in a descending order of relative nutrient shortage. As a result, Fe, Zn and Cu should be added following regional diagnosis at site #1, while only Fe and Cu would be required following local diagnosis. At site #2, B appeared to be in relative shortage at both regional and local scales. At the regional scale, N and Ca showed relative excess, while B and Cu ranked in a descending order of relative nutrient shortage. At the local scale, only B appeared to limit yield. As a result, B and Cu should be added and N reduced following regional diagnosis at site #2, while only B would be required according to local diagnosis.

4. Discussion

4.1. ML Model

The AUC of the RF model that included features available in the data set was 0.78, indicating that the model was informative. The accuracy of the RF classification model was 0.72 compared to more than 0.80 for most tested crops [48]. Raw concentrations with no need to impute missing values returned higher model accuracy than log ratios. On the other hand, zero or missing values make it impossible to compute log ratios, potentially reducing the size of the data set available to run ML models if imputation is not possible or there are too many zeroes in the data set.

Compared to compositional models that report nutrient interactions as ratios or multi-ratios, ML models address factor interactions as combinations of factors at a given geographical scale. This is different from the definition of factor interactions in statistical models. Errors on interactions occur when comparing means of main effects where interactions were significant or reporting means at the interaction level where the interactions were not significant [19]. While ecological patterns result from myriads of interactive processes, most statistical models can solve only a limited number of interactions between factors [49]. In ML models, the concept of significance is replaced by an assessment of increased accuracy after adding potentially contributing factors whatever their size effect or significance. The minimum number of combined factors to reach high model accuracy is the minimum data set required to solve the problem under study with smallest effort on data collection.

In statistical analysis, claiming 'statistically non-significant' differences does not mean that there was no difference at all, leading to potential conflictual conclusions [16]. Confidence intervals should thus be renamed "compatibility intervals" to embrace uncertainty on interpretation. In comparison, ML methods include growth-impacting factors, avoiding the accept/reject "dichotomania" of either adding or removing features based on significance to assess factor contribution to model accuracy.

Critical concentration ranges bear different meanings. They can be presented as statistically derived intervals such as boxplots and confidence intervals, or as physiological response patterns to nutrient additions where critical boundaries are defined arbitrarily at 90–95% maximum yield. Boxplots are easily derived from regional crop surveys where nutrient treatments are not varied systematically, by assigning tissue nutrient compositions to yield classes. Tissue nutrient thresholds require varying doses in one-nutrient or factorial experiments, but such trials are site-specific and expensive. In both cases, concentration ranges are fixed values leading dichotomous decisions. Claiming that some nutrients of the diagnosed specimen fall outside the "critical concentration range" does not mean that the specimen is nutritionally imbalanced. It merely reflects some incompatibility between diagnosed concentrations and the statistically or physiologically derived concentration ranges.

It appears nonsensical that 50% of the TN specimens in the present study would fall outside the boundaries delineated by boxplots for diagnostic purposes. It is even more surprising to find just one TN specimen surviving after diagnosing the whole TN data set across all compatibility intervals in Table 4, an insignificant success rate (one out of 489 observations!). Regional compatibility intervals also proved to be a complete failure (zero success). Indeed, current critical nutrient ranges are assemblages of separately derived concentration ranges pasted together to generate a "Frankenstein-built" diagnostic tool that denies nutrient interactions. Indeed, assuming normal data distribution within normalized critical ranges, it can be shown geometrically that diagnosing by nutrient compatibility intervals collapses in the Euclidean hyper-space as more nutrients are being diagnosed to fully capture nutrient imbalance [50]. While there is a false belief that crossing the threshold of statistical significance is sufficient as a proof [16], it is similarly a false belief that crossing critical concentration ranges is enough to demonstrate nutrient imbalance. This is why critical nutrient concentration ranges (compatibility intervals) should be abandoned for diagnostic purposes as strongly impacted by errors on nutrient interactions.

While nutrients interact with each other, *clr* or *ilr* variables can project them into the Euclidean hyper-space of plant nutrients to avoid disastrous conclusions. The compositional methods view

nutrient compositions as entities, i.e., unique combinations of nutrients in a tissue. Nutrients interact between them in several ways [17,51], and this can be handled by log ratio transformations [37]. The distance between two equal-length compositions is computed as a Euclidean distance using *clr* or *ilr* variables. The *clr* differences can rank nutrients in the order of their limitation to yield.

To allow trustful downscaling of nutrient diagnostic methods, regional diagnosis across factors must be coherent with diagnosis at a local scale where myriads of factor combinations occur. Growers solved this problem intuitively by conducting side-by-side comparisons between unhealthy and nearby healthy specimens. Compositional methods provide a quantitative compositional diagnostic approach by comparing defective to successful neighboring compositions at factor levels shared by the defective and successful specimens. Such side-by-side comparison also provides trustful attainable yields under the specified combination of factors. As shown by the discrepancy between regional and local diagnoses, the factor-specific approach could control errors attributable to factor dissimilarity potentially affecting crop yield at the local scale.

4.2. Compositions as Unique Combinations of Nutrients

Nutrient acquisition by plants depends on environmental factors such as soil properties, soil water content, and climatic conditions [52,53]. Nutrient combinations leading to high-yields under successful conditions at the specified factor levels may change as controllable growth-limiting factors are alleviated. While the Law of the Maximum relies on tens of growth factors and countless factor interactions [20], «Jardins do Eden», where all factors are at their optimum levels, are rarely encountered. On the other hand, «ilhas encantadas» (enchanting islands) [18,26,27], where controllable factors are close to their optima under given combinations of uncontrollable factors, can be documented as successful Humboldtian loci where several yield-limiting factors have been handled adequately by local growers.

At a local scale, under a given combination of uncontrollable and controlled factors shared by neighboring defective and successful specimens, assumptions on factors being equal or at optimum levels can be considerably reduced. Parent [26] depicted growers searching for maximum yield from a set of controllable growth-limiting factors as compositional parachutists trying to land on the nearest enchanting island by manipulating D-1 suspension lines at a time to avoid falling into the surrounding turbulent sea. Where low yield, DBH or plant vigor is observed and nutrient imbalance is suspected, the objective is to reach high nutrient-use efficiency by adopting reliable corrective measures already implemented in the successful neighborhood. To generate large, trustful, and informative data sets to conduct nutrient diagnoses at a local scale, a close and ethical collaboration is required between researchers and stakeholders [54].

5. Conclusions

The present Brazilian nutrient concentration ranges for Mg, Mn, Fe and Zn differed markedly from compatibility intervals derived from the TN specimens in the data set. Moreover, denying nutrient interactions, nutrient concentration ranges collapsed in the Euclidean space as more nutrients are added. Indeed, only one TN specimen survived after diagnosing 489 TN specimens across eleven nutrient compatibility intervals bounded by the TN quartiles. Although easy to interpret, dichotomous decisions inherited from the past using critical nutrient concentration ranges should be replaced by tools of machine learning and compositional data analysis.

The ML model showed that the productivity of young *Eucalyptus* trees depended not only on mineral nutrition but also on local features such as clone, soil type, location, and tree age. Raw concentrations returned higher model accuracy and were not affected by missing values compared to log-ratios. As a result, log-ratio transformations are solely required in data post-processing to integrate nutrient interactions in the diagnostic nutrient-ranking heuristic model.

Regional and local nutrient diagnoses of defective specimens may differ. As a result, downscaling regional nutrient standards to a local scale could be hazardous and could explain the large variation

in fertilization regimes in Brazilian *Eucalyptus* ecosystems, where environmental and managerial factors vary widely. Local scale diagnosis by factor analogy is viable to reach potential yield levels. Factor-specific diagnosis has the advantage over regional diagnosis that local factors can be kept similar in every aspect but factors that have been controlled in the successful neighborhood.

Although the local diagnostic approach is appealing to avoid error on interactions, it is highly demanding in well-documented and trustful data. Meteorological data, pest management and soil quality tests could be further documented to increase *Eucalyptus* model accuracy. Commitment to share relevant information is essential to build large data sets and return accurate predictions. A close, trustful, and ethical collaboration is thus necessary between stakeholders to customize and validate tissue nutrient diagnosis of *Eucalyptus* trees at a local scale.

Author Contributions: B.V.d.P. organized the data for modelling, ran models, co-wrote the paper. W.S.A. collected and updated the data set. L.E.P. elaborated models, co-wrote the paper. E.F.d.A. collected data and metadata. G.B. revised the paper. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge the financial support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), CMPC Impresa, and the Natural Science and Engineering Research Council of Canada (NSERC-2254).

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

References

1. Laclau, J.P. *Eucalyptus 2018: Managing Eucalyptus Plantations under Global Changes*, 1st ed.; IUFRO, Ed.; Montpellier: Le Corum, France, 2018; ISBN 978-2-87614-743-0.
2. IBÁ Relatório Anual 2017. Available online: https://iba.org/images/shared/Biblioteca/IBA_RelatorioAnual2017.pdf (accessed on 10 July 2019).
3. Clearwater, M.J.; Meinzer, F.C. Relationships between hydraulic architecture and leaf photosynthetic capacity in nitrogen-fertilized Eucalyptus grandis trees. *Tree Physiol.* **2001**, *21*, 683–690. [CrossRef]
4. Graciano, C.; Goya, J.F.; Frangi, J.L.; Guiamet, J.J. Fertilization with phosphorus increases soil nitrogen absorption in young plants of Eucalyptus grandis. *For. Ecol. Manag.* **2006**, *236*, 202–210. [CrossRef]
5. Laclau, J.-P.; Almeida, J.C.R.; Goncalves, J.L.M.; Saint-Andre, L.; Ventura, M.; Ranger, J.; Moreira, R.M.; Nouvellon, Y. Influence of nitrogen and potassium fertilization on leaf lifespan and allocation of above-ground growth in Eucalyptus plantations. *Tree Physiol.* **2008**, *29*, 111–124. [CrossRef]
6. Gazola, R.d.N.; Buzetti, S.; Teixeira Filho, M.C.M.; Gazola, R.P.D.; Celestrino, T.D.S.; Silva, A.C.D.; Silva, P.H.M.D. Potassium fertilization of eucalyptus in an entisol in low-elevation cerrado. *Rev. Bras. Ciência do Solo* **2019**, *43*. [CrossRef]
7. Hubbard, R.M.; Ryan, M.G.; Giardina, C.P.; Barnard, H. The effect of fertilization on sap flux and canopy conductance in a Eucalyptus saligna experimental forest. *Glob. Chang. Biol.* **2004**, *10*, 427–436. [CrossRef]
8. Viera, M.; Ruíz Fernández, F.; Rodríguez-Soalleiro, R. Nutritional prescriptions for eucalyptus plantations: Lessons learned from spain. *Forests* **2016**, *7*, 84. [CrossRef]
9. Stape, J.L.; Binkley, D.; Ryan, M.G.; Fonseca, S.; Loos, R.A.; Takahashi, E.N.; Silva, C.R.; Silva, S.R.; Hakamada, R.E.; Ferreira, J.M.d.A.; et al. The Brazil eucalyptus potential productivity project: Influence of water, nutrients and stand uniformity on wood production. *For. Ecol. Manag.* **2010**, *259*, 1684–1694. [CrossRef]
10. Munson, R.D.; Nelson, W.L. Principles and Practices in Plant Analysis. In *Soil Testing and Plant Analysis*; Westerman, R.L., Ed.; Soil Science Society of America, Inc.: Madison, WI, USA, 1990; pp. 359–387.
11. Gatiboni, L.C.; Da Ros, C.O.; Stahl, S.; Araújo, E.F. Adubação de Eucalipto. In *Manual de Calagem e Adubação do RS/SC*; Silva, L.S., Ed.; Comissão de Química e Fertilidade: Alegre, Brazil, 2016; pp. 245–247.
12. Ferreira, E.V.d.O.; Novais, R.F.; Médice, B.M.; de Barros, N.F.; Silva, I.R. Leaf total nitrogen concentration as an indicator of nitrogen status for plantlets and young plants of Eucalyptus clones. *Rev. Bras. Ciência do Solo* **2015**, *39*, 1127–1140. [CrossRef]
13. de Morais, T.C.B.; Prado, R.D.M.; Traspadini, E.I.F.; Wadt, P.G.S.; de Paula, R.C.; Rocha, A.M.S. Efficiency of the CL, DRIS and CND Methods in assessing the nutritional status of eucalyptus spp. rooted cuttings. *Forests* **2019**, *10*, 786. [CrossRef]

14. da Silva, G.G.C.; Neves, J.C.L.; Alvarez, V.V.H.; Leite, F.P. Nutritional diagnosis for eucalypt by DRIS, M-DRIS, and CND. *Sci. Agric.* **2004**, *61*, 507–515. [CrossRef]
15. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B* **1982**, *44*, 139–160. [CrossRef]
16. Amrhein, V.; Greenland, S.; McShane, B. Retire statistical significance. *Nature* **2019**, *567*, 305–307. [CrossRef] [PubMed]
17. Walworth, J.L.; Sumner, M.E. The diagnosis and recommendation integrated system (DRIS). *Adv. Soil Sci.* **1987**, *6*, 149–188. [CrossRef]
18. Betemps, D.L.; de Paula, B.V.; Parent, S.-É.; Galarça, S.P.; Mayer, N.A.; Marodin, G.A.B.; Rozane, D.E.; Natale, W.; Melo, G.W.B.; Parent, L.E.; et al. Humboldtian diagnosis of peach tree (prunus persica) nutrition using machine-learning and compositional methods. *Agronomy* **2020**, *10*, 900. [CrossRef]
19. Umesh, U.N.; Peterson, R.A.; McCann-Nelson, M.; Vaidyanathan, R. Type IV error in marketing research: The investigation of ANOVA interactions. *J. Acad. Mark. Sci.* **1996**, *24*, 17–26. [CrossRef]
20. Wallace, A.; Wallace, G.A. *Horticultural Review*; Janick, J., Ed.; John Wiley & Sons, Inc.: Oxford, UK, 1993; ISBN 9780470650547.
21. Adams, M.; Rennenberg, H.; Kruse, J. Resilience of primary metabolism of eucalypts to variable water and nutrients. In *Eucalyptus 2018: Managing Eucalyptus Plantations under Global Changes*; Montpellier: Le Corum, France, 2018; pp. 100–104, ISBN 978-2-87614-743-0.
22. Aspinwall, M.; Blackman, C.; Resco De Dios, V.; Tjoelker, M.; Tissue, D. Photosynthesis and carbon allocation are both important predictors of genotype productivity responses to elevated CO_2 in Eucalyptus camaldulensis. In *Eucalyptus 2018: Managing Eucalyptus Plantations under Global Changes*; Montpellier: Le Corum, France, 2018; pp. 106–110, ISBN 978-2-87614-743-0.
23. Keppel, G.; Kreft, H. Integration and synthesis of quantitative data: Alexander von Humboldt's renewed relevance in modern biogeography and ecology. *Front. Biogeogr.* **2019**, *11*. [CrossRef]
24. Olden, J.D.; Lawler, J.J.; Poff, N.L. Machine learning methods without tears: A Primer for ecologists. *Q. Rev. Biol.* **2008**, *83*, 171–193. [CrossRef]
25. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*, 1st ed.; Cambridge University Press: New York, NY, USA, 2014; ISBN 978-1-107-05713-5.
26. Parent, S.-É. Why we should use balances and machine learning to diagnose ionomes. *Authorea* **2020**, *1*. [CrossRef]
27. Coulibali, Z.; Cambouris, A.N.; Parent, S.-É. Cultivar-specific nutritional status of potato (Solanum tuberosum L.) crops. *PLoS ONE* **2020**, *15*, e0230458. [CrossRef]
28. de Wit, C.T. Resource use efficiency in agriculture. *Agric. Syst.* **1992**, *40*, 125–151. [CrossRef]
29. Grunsky, E.C.; de Caritat, P. State-of-the-art analysis of geochemical data for mineral exploration. *Geochem. Explor. Environ. Anal.* **2020**, *20*, 217–232. [CrossRef]
30. Martín-Fernández, J.A.; Barceló-Vidal, C.; Pawlowsky-Glahn, V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math. Geol.* **2003**, *35*, 253–278. [CrossRef]
31. Egozcue, J.J.; Pawlowsky-Glahn, V.; Mateu-Figueras, G.; Barceló-Vidal, C. Isometric logratio transformations for compositional data analysis. *Math. Geol.* **2003**, *35*, 279–300. [CrossRef]
32. Köppen, W.; Geiger, G. Klima Der Erde (Map) 1954. Available online: http://koeppen-geiger.vu-wien.ac.at (accessed on 3 March 2020).
33. Santos, H.G. *Sistema Brasileiro de Classificação de Solos*, 5th ed.; Embrapa: Brasilia, Brazil, 2018; ISBN 978-85-7035-800-4.
34. Tedesco, M.J.; Gianello, C.; Bissani, C.A.; Bohnen, H. *Análises de Solo, Plantas e Outros Materiais*; UFRGS: Porto Alegre, Brazil, 1995.
35. Tolosana-Delgado, R.; Talebi, H.; Khodadadzadeh, M.; Van den Boogaart, K.G. On machine learning algorithms and compositional data. In Proceedings of the 8th International Workshop on Compositional Data Analysis, Terrassa, Spain, 3–8 June 2019; Egozcue, J.J., Graffelman, J., Ortego, M.I., Eds.; pp. 172–175.
36. Budhu, M. *Soil Mechanics and Foundations*, 3rd ed.; Wiley: New York, NY, USA, 2010; Volume 1, ISBN 9788578110796.
37. Parent, L.E.; Dafir, M. A theoretical concept of compositional nutrient diagnosis. *J. Am. Soc. Hortic. Sci.* **1992**, *117*, 239–242. [CrossRef]
38. Beaufils, E. *Diagnosis and Recommendation Integrated System (DRIS)*, 1st ed.; University of Natal: Pietermaritzburg, South Africa, 1973.

39. Badra, A.; Parent, L.-É.; Allard, G.; Tremblay, N.; Desjardins, Y.; Morin, N. Effect of leaf nitrogen concentration versus CND nutritional balance on shoot density and foliage colour of an established Kentucky bluegrass (Poa pratensis L.) turf. *Can. J. Plant Sci.* **2006**, *86*, 1107–1118. [CrossRef]
40. Aitchison, J. Principles of compositional data analysis. *Multivar. Anal. Its Appl. IMS Lect. Notes Monogr. Ser.* **1994**, *24*, 73–81.
41. Rodgers, J.L.; Nicewander, W.A.; Toothaker, L. Linearly independent, orthogonal, and uncorrelated variables. *Am. Stat.* **1984**, *38*, 133. [CrossRef]
42. Filzmoser, P.; Hron, K.; Reimann, C. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. *Sci. Total Environ.* **2009**, *407*, 6100–6108. [CrossRef]
43. Egozcue, J.J.; Pawlowsky-Glahn, V. Groups of parts and their balances in compositional data analysis. *Math. Geol.* **2005**, *37*, 795–828. [CrossRef]
44. de Oliveira, C.T.; Rozane, D.E.; de Amorim, D.A.; de Souza, H.A.; Fernandes, B.S.; Natale, W. Diagnosis of the nutritional status of 'Paluma' guava trees using leaf and flower analysis. *Rev. Bras. Frutic.* **2020**, *42*, 1–9. [CrossRef]
45. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
46. Delacour, H.; Servonnet, A.; Perrot, A.; Vigezzi, J.F.; Ramirez, J.M. La courbe ROC (receiver operating characteristic): Principes et principales applications en biologie clinique. *Ann. Biol. Clin. (Paris)* **2005**, *63*, 145–154.
47. Parent, S.-É.; Parent, L.E.; Rozane, D.-E.; Natale, W. Plant ionome diagnosis using sound balances: Case study with mango (Mangifera Indica). *Front. Plant Sci.* **2013**, *4*, 449. [CrossRef] [PubMed]
48. Parent, L.E.; Rozane, D.E.; Deus, J.A.L.; Natale, W. Diagnosis of nutrient composition in Fruit crops: Latest developments. In *Fruit Crops. Diagnosis and Management of Nutrient Constraints*; Srivastava, A.K., Hu, C., Eds.; Elsevier: New York, NY, USA, 2019; p. 400.
49. Ryo, M.; Rillig, M.C. Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere* **2017**, *8*, e01976. [CrossRef]
50. Nowaki, R.H.D.; Parent, S.-É.; Cecílio Filho, A.B.; Rozane, D.E.; Meneses, N.B.; Silva, J.A.; Natale, W.; Parent, L.E. Phosphorus over-fertilization and nutrient misbalance of irrigated tomato crops in Brazil. *Front. Plant Sci.* **2017**, *8*, 825. [CrossRef] [PubMed]
51. Wilkinson, S.R.; Grunes, D.L.; Sumner, M.E. Nutrient interactions in soil and plant nutrition. In *Handbook of Soil Fertility and Plant Nutrition*; Sumner, M.E., Ed.; CRC Press: London, UK, 2000; p. 91.
52. Marschner, P. *Marschner ' s Mineral Nutrition of Higher Plants*, 3rd ed.; Academic Press: London, UK, 2012; ISBN 9780123849052.
53. Barber, S.A. *Soil Nutrient Bioavailability: A Mechanistic Approach*, 2nd ed.; Wiley: New York, NY, USA, 1995.
54. Gibson, K.J.; Streich, M.K.; Topping, T.S.; Stunz, G.W. Utility of citizen science data: A case study in land-based shark fishing. *PLoS ONE* **2019**, *14*, e0226782. [CrossRef]