horticulturae

MDPI

Review

Compositional and Machine Learning Tools to Model Plant Nutrition: Overview and Perspectives

Léon Etienne Parent

Department of Soils and Agrofood Engineering, Université Laval, Québec, QC G1V 0A6, Canada; leon-etienne.parent@fsaa.ulaval.ca

Abstract: The *ceteris paribus* assumption that all features are equal except the one(s) being examined limits the reliability of nutrient diagnosis and fertilizer recommendations. The objective is to review machine learning (ML) and compositional data analysis (CoDa) tools to make nutrient management feature specific. The average accuracy of the ML methods was 84% across the crops. The additive and orthogonal log ratios of CoDa reduce a *D*-parts soil composition to *D*-1 variables, alleviating redundancy in the predictive ML models. Using a Brazilian onion (*Allium cepa*) database, the combined CoDa and ML methods returned crop response patterns, allowing feature-specific fertilizer recommendations to be made. The centered log ratio (*clr*) diagnoses plant nutrients as a compositional nutrient diagnosis (CND). Using a Quebec database of vegetable crops, the mean variance of *clr* variables ($\overline{VAR}$) allowed comparing total variance among species and growth stages. While *clr* is the summation of equally weighted dual log ratios, dual nutrient log ratios may show unequal importance regarding crop performance. The RReliefF scores or gain ratios can provide weighting coefficients for each dual log ratio. The widely contrasting coefficients of weighted log ratios (*wlr*) improved the accuracy of the ML models for a Quebec muck onion database. The ML models, $\overline{VAR}$ and *wlr*, are advanced tools to improve the accuracy of nutrient diagnosis.

Keywords: compositional data analysis; nutrient diagnosis; fertilizer recommendations; machine learning; model accuracy; centered log ratio (*clr*); weighted log ratio (*wlr*); crop performance

Academic Editor: Domenico Ronga

Received: 13 December 2024
Revised: 25 January 2025
Accepted: 26 January 2025
Published: 3 February 2025

Citation: Parent, L.E. Compositional and Machine Learning Tools to Model Plant Nutrition: Overview and Perspectives. *Horticulturae* **2025**, *11*, 161. https://doi.org/10.3390/horticulturae11020161

1. Introduction

Nutrient diagnosis relies on soil and plant tissue analyses, crop surveys, field trials, deficiency symptoms, and growth factors impacting crop productivity. The results of a limited number of fertilizer trials conducted under the *ceteris paribus* assumption provide the basic information to build fertilizer recommendation models [1]. This assumption bypasses the complexity of agroecosystems too easily. As databases get larger and more diversified, modern mathematical and statistical tools can simultaneously process several growth-impacting features.

Supervised machine learning (ML) models relate a target variable to a set of explanatory variables, such as soil physical, chemical, and biological properties, soil classification and hydrology, plant tissue composition, sensor data, and weather conditions, as well as soil and crop management features documented in a database. This avoids relying entirely on the *ceteris paribus* assumption for data collected in different years and at different locations. The explanatory variables are related to plant yield and quality [2] or to defense mechanisms (trophobiosis) [3–5] as target variables. The ML models generally outperform

parametric models as the training size increases [6]. The accuracy of crop ML models ranged from 0.65 to 0.93, averaging 0.84, depending on the crop, the size and diversity of the database, the ML model, the target variable, and the selected features (e.g., [7–12], amongst others).

Compositional data analysis (CoDa) tools have been developed to process compositions statistically [13]. Compositional data, such as geological data, are subject to spurious correlations [14] because, due to sum closure, as in a ternary diagram, there are *D*-1 degrees of freedom for a *D*-parts composition [15]. In agronomy, such data include tissue analyses [16] and soil [17] and water [18] compositions.

Tissue testing is thought to integrate all the effects of growth-impacting factors and their interactions [19]. The first tissue diagnostic model combined N, P, and K interactively in a ternary diagram [20]. Later, the critical nutrient approach diagnosed nutrients separately [21] and did not consider the myriads of nutrient interactions impacting plant metabolism and performance [22–25]. Ratios are convenient expressions for nutrient dilution [26] and interactions [27]. The diagnosis and recommendation integrated system (DRIS) expanded the interactive approach by integrating several dual ratios into nutrient indices [28]. The DRIS procedure has been modified on several occasions [29] and was revisited by CoDa tools such as compositional nutrient diagnosis (CND) [16]. Although each dual ratio may impact crop performance differentially, the DRIS and CND weighted equally the dual ratio expressions integrated in their formulation.

The objectives of this review were (1) to present CoDa tools for use in crop ML models, and (2) to account for the importance of each nutrient dual log ratio in CoDa expressions using ML tools.

2. Methods

2.1. Closure of Compositions

The CoDa concept addresses the intrinsically multivariate, strictly positive, compositional data close to an entity [13]. A composition made of D components, $x_1.x_2.\ldots.x_D$, and adding up to κ (=1, 100%, 1000 g kg^{-1}, …) is defined by the simplex S^D as follows:

$$S^D = \left\{ x = [x_1, x_2, \ldots . x_D] | x_i > 0,\ i = 1\ to\ D; \sum_{i=1}^{D} x_i = \kappa \right\}$$

The filling value x_D is computed as the difference between κ (generally the measurement unit) and the sum of the quantified components. The tissue dry matter simplex contains all the information on tissue components, including the one that dilutes the others (x_D) into the residual biomass. The x_D comprises C, H, O, and elements that have not been quantified. There are *D*-1 degrees of freedom in a *D*-parts composition [15] because one component can be computed as the difference between κ and the sum of the others.

The tissue simplex is an enclosure wherein nutrients accumulate and interact. Any change in the proportion of one nutrient in the confined sample space must 'resonate' with the other proportions through interactions and dilution. Examples of 'resonance' among components also exist in soil science. The closed textural diagram includes the proportions of sand, silt, and clay that add up to 100%. The exchangeable cations and acidity are constrained to the cation exchange capacity. The sum of the relative air, water, and solid volumes in soils is also closed at 100%.

The statistical analysis of raw compositional data returns distorted results if not transformed beforehand into log ratios [13]. Indeed, confidence intervals about means may miss the limits of the sample space of proportions (less than zero or more than 100%) after conducting statistical analyses. The sum of the means of the proportions may also differ from 100%, leading to physically absurd results [29,30].

Log-ratio transformations, like additive log ratios (*alr*), dual log ratios (*dlr*), centered log ratios (*clr*), orthogonal log ratios (*olr*) such as isometric log ratios (*ilr*) and pivot log ratios (*plr*), and the summated or amalgamated log ratio (*slr*), were propounded to free compositional data from their constrained sample space [31,32]. Log ratios project raw compositional data into the real space ($\pm\infty$) required to conduct statistical analyses. Indeed, $log(a/b) \rightarrow \infty \; if \; a \rightarrow \infty$ or $b \rightarrow 0$, and conversely.

The geometric mean is most often used to compute log ratios, but it is impacted by errors of measurement, outliers, and results below the detection limits. Data below the detection limit may be replaced by a small value like 2/3 of the detection limit, or they can be amalgamated with other components [31,32]. Nevertheless, including compositional data that are moderately variable but informative contributes to solving complex systems [33].

2.2. Machine Learning Models

The ML decision tree models are non-parametric, have few parameters and good scalability, and can detect multivariate effects among variables in high-dimensional databases [34,35]. Several features, such as soil chemical, physical, and biological properties, weather conditions, management practices, soil hydrology, and sensor data can be documented and processed and related to a target variable. This avoids relying solely on the *ceteris paribus* assumption of equal or optimal growing conditions. To facilitate model adoption, features should be easy to collect by stakeholders and verified for their capacity to generalize. The models surveyed in this review are decision tree models, like random forest and boosting models, the Gaussian [7], and the KNN (k-nearest neighbors) [12] models.

2.3. Databases

The first objective of this paper was addressed using a Brazilian onion (*Allium cepa*) database retrieved from the web [8]. Crop yield was related to soil variables previously reduced to *D*-1 additive or orthogonal log ratios. For the second objective, crop yield was related to tissue analyses using a Quebec vegetable database, where total N was analyzed by micro-Kjeldahl, and other nutrients were quantified by plasma emission spectroscopy after acid digestion [36].

3. CoDa Tools to Alleviate Redundancy in ML Crop Response Models

3.1. Log-Ratio Transformations

3.1.1. Additive Log-Ratio (*alr*) Transformation

The *alr* is defined as follows:

$$alr = ln(x_i/x_j)$$

where x_i is any component, and x_j is a reference component common to all components. If $x_j = x_D$, and x_D is large, the *alr* transformation resembles the ordinary logarithmic transformation [37]. The *alr* is easy to compute and interpret, but its geometry is non-Euclidean. The KNN may require Euclidean geometry.

Nitrogen (N) has been used as the reference component to monitor fertilization in tree nurseries [38]. The *alr* is also useful to model soil compositional data. The clay may be used as a common denominator for soil organic matter, soil test P, and soil test K because kaolinite clay contributes to cation exchange capacity and to P fixation by oxyhydroxides in tropical and subtropical soils and interacts with soil organic matter to build soil structure [39].

3.1.2. Orthogonal Log-Ratio (*olr*) Transformations

Orthogonal log ratios are useful in running multivariate analyses unbiasedly [31,32]. The isometric log ratio (*ilr*) and its simplified version, the pivot ratio (*plr*), are orthonormal

log ratios that return *D*-1 variables, the exact number of degrees of freedom available in a *D*-parts composition. The *olr* variables have Euclidian geometry. The *ilr* is defined as follows [40]:

$$ilr_k = \sqrt{\frac{r_k s_k}{r_k + s_k}} ln\left(\frac{\sqrt[r_k]{\prod_{i=1}^{r_k} x_i}}{\sqrt[s_k]{\prod_{j=1}^{s_k} x_j}}\right)$$

where r_k and s_k are the numbers of components in the numerator and denominator, respectively; i and j refer to components in the numerator and denominator, respectively; $\sqrt{\frac{r_k s_k}{r_k+s_k}}$ is a normalization coefficient; $\sqrt[r_k]{\prod_{i=1}^{r_k} x_i}$ and $\sqrt[s_k]{\prod_{j=1}^{s_k} x_j}$ are geometric means of the components in the numerator and denominator, respectively.

The number of combinations of components in the *ilr* expressions is enormous: $(2D-2)!/\left[2^{D-1}(D-1)!\right]$ [41]. Thanks to orthogonality, the results of multivariate analyses remain the same whatever the combination of components. The number of combinations is reduced to $D!$ using the *plr* [42]. The *plr* sequentially contrasts one component against the others, facilitating the interpretation compared to the *ilr*.

The *ilr* transformation has been applied to plant nutrition studies [28]. However, *ilr*-transformed data are difficult to interpret. In soil science, compositions comprise a smaller number of variables, making the interpretation easier. It has been shown that *ilr* axes are related to the fractal dimension of soil aggregates, providing a more detailed description of aggregate disruption or building [17].

The *olr* variables are conceptualized in a sequential binary partition (SBP) [40]. Soil compositional data impacting soil structure as soil quality indicators can be contrasted as 'balances' between soil organic matter and textural components; then, clay as a binding agent can be contrasted against silt and sand; the remaining contrast is silt against sand (Table 1). The SBP is illustrated in Figure 1.

Table 1. The sequential binary partition of the compositional soil simplex made of organic matter, sand, silt, and clay proportions adding up to 100%.

ilr	OM	Clay	Silt	Sand	r	s	Equation (*ilr*)
1	1	−1	−1	−1	1	3	$\sqrt{\frac{1\times3}{1+3}} ln\left(\frac{MO}{\sqrt[3]{Sand\times Silt\times Clay}}\right)$
2	0	1	−1	−1	1	2	$\sqrt{\frac{1\times2}{1+2}} ln\left(\frac{Clay}{\sqrt[2]{Silt\times Sand}}\right)$
3	0	0	1	−1	1	1	$\sqrt{\frac{1\times1}{1+1}} ln\left(\frac{Silt}{Sand}\right)$

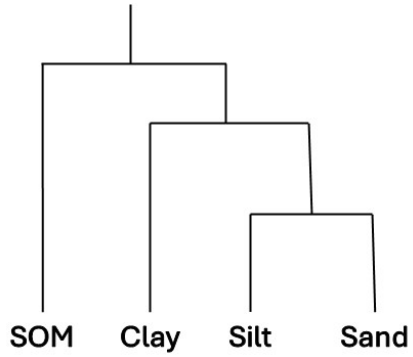

Figure 1. Balance scheme to compute *ilr* contrasting organic matter (SOM) with textural components, then clay with silt and sand, and, finally, silt with sand.

3.1.3. Summated Log-Ratio (*slr*) Transformation

Parts can be amalgamated to avoid dealing with zeroes or to facilitate the interpretation of the results based on domain knowledge [31,32]. For example, silt and clay could be amalgamated as soil's capacity to protect organic C physically [43].

3.2. CoDa Tools to Run ML Models

The Brazilian database was retrieved from the web [8]. Crop yield was related to soil variables reduced to *D*-1 additive log ratios. The target variable was marketable bulb yield. The categorial variables were stratified to avoid model overfitting. The ML regression model was run using the Orange Data Mining freeware vs. 3.37 (University of Ljubljana, Slovenia).

The R^2 coefficient, root mean square of error (RMSE), and mean absolute error (MAE) are measures of model accuracy, as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|^2 \quad (2)$$

The coefficient of determination (R^2) is interpreted as follows: $R^2 < 0.25$, very weak; $0.25 \leq R^2 < 0.50$, weak; $0.50 \leq R^2 < 0.75$, moderate, and $R^2 > 0.75$, substantial [44]. The Catboost model was accurate (Table 2).

Table 2. Accuracy of the Catboost machine learning (ML) models using NPK fertilization data, pH, and *alr*-transformed soil data.

R^2 Coefficient	Root Mean Square Error	Mean Absolute Error
	Mg marketable onion bulbs ha^{-1}	
0.929	4739	3662

A universality test [45] was conducted to verify the model's capacity to generalize. The set of features from the unseen site was related to marketable bulb yields to predict the crop response pattern (Figure 2). Thereafter, a response function can be fitted to the response pattern to assess the economic optimum rate of fertilization. The followed a polynomial linear plateau model levelling off near 200 kg N ha^{-1}.

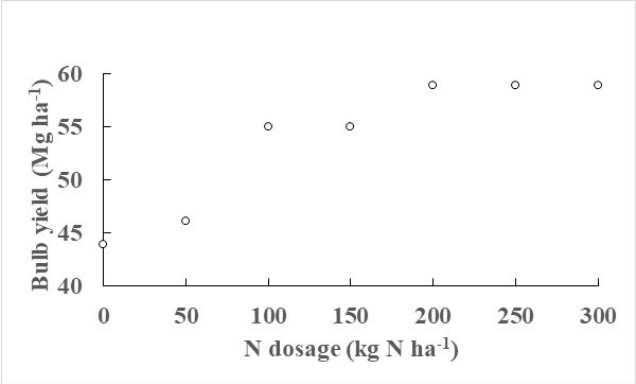

Figure 2. Relationship between onion bulb yield and N dosage predicted by the Catboost model for a site unseen by the model.

The right choice of the response function is critical because several non-linear functions return comparable R^2 coefficients while recommending very different optimum economic doses (OED). Indeed, the OED depends on the slope of the function [46]. The environmental costs of fertilization, such as GHG emissions as nitrous oxide (N_2O) [47] or loss in water quality [48], may be considered as additional constraints of the non-linear functions.

4. Combining CoDa and ML Tools

4.1. The clr Transformation

The centered log ratio (*clr*) is commonly used to diagnose nutrients as a compositional nutrient diagnosis (CND) [16]. The *clr* is the integration of equally weighted dual log ratios. The *clr* is defined as follows [30]:

$$clr_i = ln(x_i / g[x])$$

where $g[x] = (x_1 x_2 \ldots x_D)^{\frac{1}{D}}$ is the geometric mean across components including x_D. There are D *clrs* in a D-parts composition. Their sum is zero.

The *clr* is the mean of D equally weighted dual log ratios, computed as follows for nutrient x_i:

$$ln\left(\frac{x_i}{g(x)}\right) = ln\left(\frac{x_i}{x_1} \times \frac{x_i}{x_2} \times \ldots \times \frac{x_i}{S_D}\right)^{\frac{1}{D}}$$

Hence:

$$ln\left(\frac{x_i}{g(x)}\right) = \frac{1}{D}\left[ln\left(\frac{x_i}{x_1}\right) + ln\left(\frac{x_i}{x_2}\right) + \ldots + ln\left(\frac{x_i}{S_D}\right)\right]$$

The *clr* has Euclidean geometry, as follows [30]:

$$\varepsilon = \sqrt{\sum_{j=1}^{D}\left(clr_j - clr_j^*\right)^2}$$

where ε is the Euclidean distance between two compositional vectors of equal length, clr_j is the jth *clr* value of the diagnosed specimen, and clr_j^* is the corresponding reference *clr* value. This allows the comparison of abnormally to normally growing plants in an otherwise comparable neighborhood to diagnose apparent or hidden deficiency symptoms in plants under the *ceteris paribus* assumption. The nutrient indices are the differences between the clr_j of the abnormal plant and the corresponding clr_j^* of the normal plant. They are illustrated in a histogram to facilitate the interpretation.

Using population statistics (mean and standard deviation) instead of a normal plant as reference, the CND index for nutrient x_i is computed as follows [16]:

$$Index\ x_i = \frac{clr_{x_i} - clr_{x_i}^*}{standard\ deviation_{x_i}^*}$$

4.2. Mean clr Variance Among Species and Sampling Stages

Tissue nutrient concentration and variability in the diagnostic tissue depend on species, cultivars, sample position on the plant, the sampling period, plant age, season, and the cropping system [49] and the application of pesticides containing cationic micronutrients like Cu, Zn, and Mn [28]. The diagnostic tissue should be easy to identify and collect. Total variance across the *clr* variances is a complementary selection criterion for the sampling period. The mean variance $\overline{VAR}$ across the centered log ratio is averaged as follows [31,32]:

$$\overline{VAR} = \frac{1}{D}\sum_{j=1}^{D} VAR(clr_j)$$

The mean *clr* variances of three crops were computed at four growth stages, as shown in Table 3. The smallest mean variance occurred at the flower bud stage for potato (*Solanum tuberosum*), bulb enlargement for onion (*Allium cepa*), and the 8–10 leaf stage for carrot (*Daucus carota*).

Table 3. The *clr* standard deviation at high-yield level of 11 foliar components for potato (*Solanum tuberosum*), onion (*Allium cepa*), and carrot (*Daucus carota*) tissues sampled at four growth stages (data from the Quebec, Canada, database).

Crop	Growth Stage	N	P	K	Ca	Mg	B	Cu	Zn	Mn	Fe	x_D †	Mean Standard Deviation
Potato (*Solanum tuberosum*)	Plant 20 cm high	0.315	0.296	0.274	0.393	0.294	0.313	0.461	0.371	0.476	0.535	0.290	0.375
	Flower bud	0.213	0.246	0.124	0.314	0.177	0.176	0.407	0.357	0.475	0.452	0.138	0.305
	10% flowering	0.181	0.233	0.172	0.34	0.210	0.332	0.511	0.564	0.546	0.416	0.144	0.364
	Tuber initiation	0.169	0.270	0.247	0.307	0.275	0.430	0.415	0.584	0.555	0.317	0.138	0.364
Onion (*Allium cepa*)	2–3 leaves	0.269	0.258	0.248	0.167	0.173	0.226	0.426	0.263	0.760	0.496	0.199	0.360
	4–5 leaves (leek stage)	0.241	0.239	0.198	0.168	0.189	0.181	0.336	0.522	0.786	0.521	0.214	0.378
	6–8 leaves	0.208	0.237	0.153	0.165	0.205	0.134	0.505	0.297	0.636	0.365	0.200	0.321
	Bulb enlargement	0.176	0.178	0.198	0.147	0.147	0.172	0.392	0.452	0.585	0.348	0.238	0.309
Carrot (*Daucus carota*)	4–5 leaves	0.196	0.230	0.277	0.163	0.212	0.210	0.304	0.273	0.582	0.616	0.154	0.329
	6–7 leaves	0.144	0.156	0.137	0.125	0.175	0.172	0.252	0.274	0.654	0.366	0.140	0.280
	8–10 leaves	0.129	0.155	0.106	0.137	0.153	0.137	0.359	0.244	0.568	0.330	0.110	0.260
	Root enlargement	0.143	0.190	0.158	0.181	0.234	0.167	0.403	0.343	0.564	0.278	0.160	0.304

† Filling value between measurement unit and the sum of nutrients.

4.3. Weighted log Ratio (*wlr*)

Although mathematically sound, the *clr* weights the dual log ratios equally. However, dual ratios may impact crop performance differentially. We modified the *clr* by assigning a coefficient to each dual log ratio before its integration into a 'weighted log ratio' (*wlr*) formulation. The coefficients account for the importance of each dual log ratio for the target variable. The *wlr* accounts for the importance of each log ratio regarding the target variable, as follows:

$$wlr_{x_i} = \frac{1}{D}\sum_{j=1}^{D} \varphi_j ln\left(\frac{x_i}{x_j}\right), i \neq j$$

where x_i is the common numerator for nutrient i, x_j represents other components, and φ_j is a coefficient to each log-transformed dual ratio. The *clr* is thus a special case of the *wlr*, where all φ_js are equal. The ML models can provide φ_j coefficients as RReliefF scores in a regression mode [50] or gain ratios in a classification mode. To maintain nutrient x_i in the numerator to compute wlr_{x_i} across dual log ratios containing x_i, $\varphi_j ln\left(\frac{x_j}{x_i}\right)$ is multiplied by -1 to relocate x_i in the numerator, recovering $\varphi_j ln\left(\frac{x_i}{x_j}\right)$.

4.4. The Quebec Database to Compare *clr* and *wlr*

The Quebec onion database comprises 275 observations on tissue N, P, K, Ca, and Mg raw concentrations collected at the 'leek' stage (4–5 leaves). With six components, including the filling value, there are $D(D-1)/2 = 15$ dual log ratios. The importance of log-transformed dual log ratios regarding bulb yield classification (< or ≥50 Mg bulbs ha^{-1}) is presented as gain ratios in Figure 3. The dual log ratio P/K showed the highest gain ratio. The Ca/x_D, Ca/Mg, and Mg/x_D dual log ratios were not contributive to yield classification, indicating that their importance is null for classification purposes using that database. Other log-transformed dual log ratios were important. Dual ratios thus unequally impacted onion yield classification using that database because several dual ratios showed negligible importance for bulb yield.

The *wlr* N was poorly related to the corresponding raw N concentration values but closely related to the *clr* N (Figure 4). Raw concentrations inject white noise into the *wlr* by not addressing the "resonance" in the simplex that is attributable to nutrient interactions

and dilution within the confined sample space of the measurement unit. The *clr* N variables injected white noise into the *wlr* by not addressing the unequal importance of dual log ratios regarding crop yield.

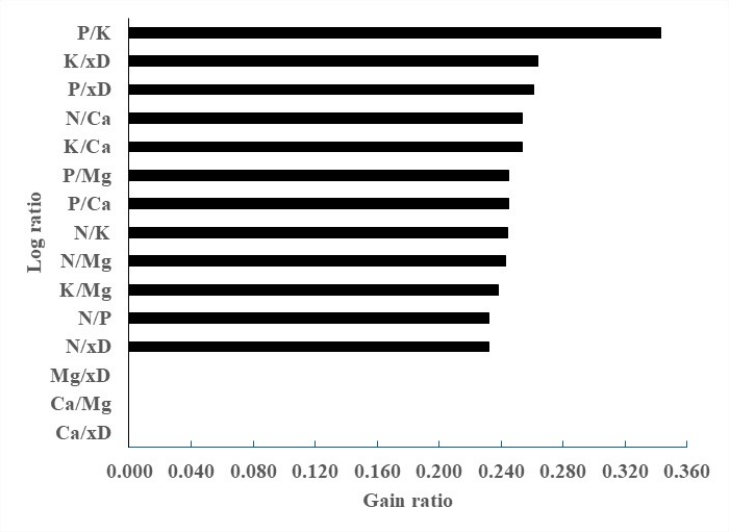

Figure 3. Gain ratio of log-transformed dual ratios regarding bulb yield classification using the Quebec onion (*Allium cepa*) database. The Ca/xD, Ca/Mg, and Mg/xD dual log ratios are unimportant; xD = filling value.

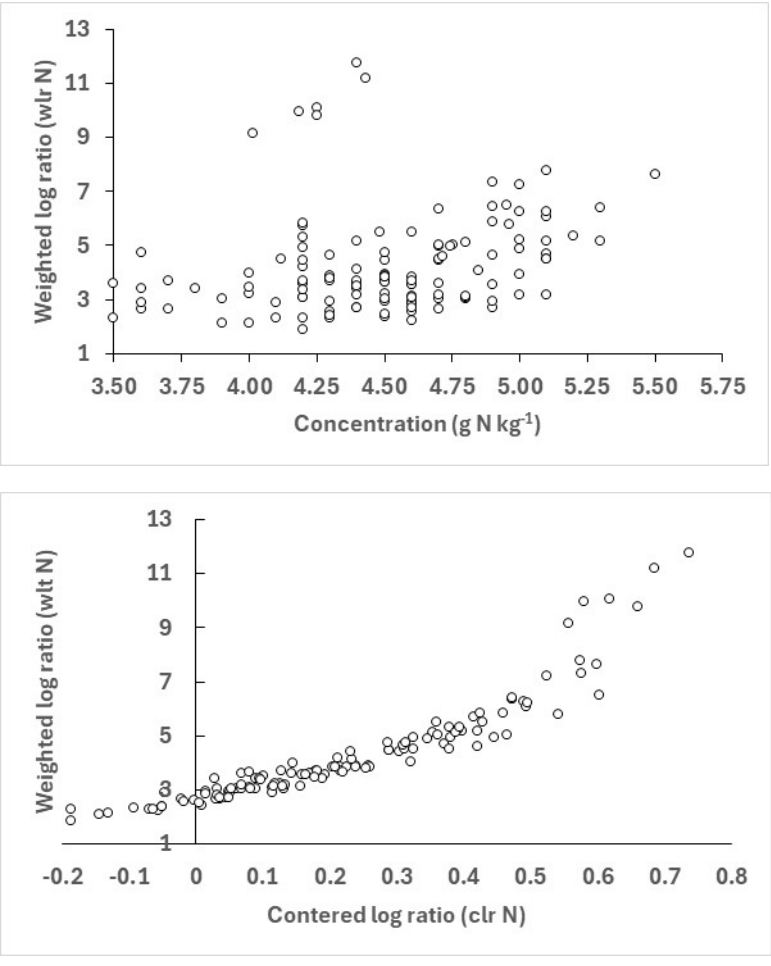

Figure 4. Relationships between weighted log ratio for N (*wlr* N) and N concentrations (upper graph) or centered log ratios for N (*clr* N) (lower graph). White noise is shown by point dispersion.

4.5. Tissue Nutrient Diagnosis Using ML Models

Tissue tests collected in field experiments and during crop surveys can be related to the target variable using classification ML models to derive nutrient standards about a cutoff yield. The specimens are classified in a confusion matrix (Figure 5) as true negative (TN = population of high-yielding and nutritionally balanced specimens), true positive (TP = population of low-yielding and nutritionally imbalanced specimens), false negative (FN = population of low-yielding yet nutritionally balanced specimens), and false positive (FP = population of low-yielding but nutritionally imbalanced specimens due to luxury consumption or contamination).

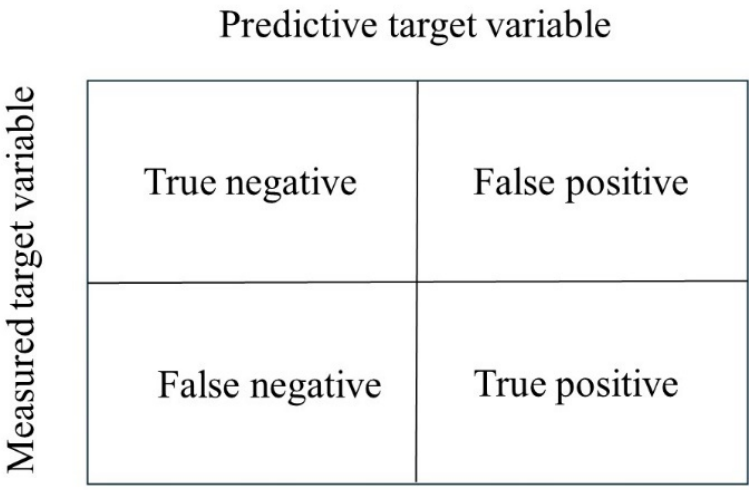

Figure 5. Confusion matrix classifying specimens into four categories.

The accuracy of ML classification models is measured as follows:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

A model may not be so informative if most of the correctly classified specimens are true positive specimens. As a result, true negative specimens are too poorly documented to derive nutrient standards. The area under the curve (AUC) is the probability that the ML model ranks the classification correctly.

In the following examples using the Quebec onion database, ML classification models were run using the Orange Data Mining freeware vs. 3.37 (University of Ljubljana, Slovenia). The accuracy of the classification ML models differed between raw concentrations, with the dual *lr*, *clr*, and *wlr* used as features (Table 4). The classification models were accurate. The *wlr* outperformed raw concentration or the *clr* expressions. Indeed, some dual ratios were much less important than others regarding onion yields (Figure 3). Moreover, the number of TN specimens was higher using the *wlr*.

The means and standard deviations of the *wlr* values for the 109 TN and 132 TP specimens are presented in Table 5. The *wlr* expression showed significant differences between the *TN* and *TP* specimens across the nutrients. In contrast, the *clr* expression showed no significant differences for P, K, and Ca, and significant differences for N, Mg, and s_D. The *wlr* transformation not only enhanced the accuracy of the ML models but also discriminated all means between the TN and TP populations.

Table 4. Accuracy of regression and classification models for the relationships between onion yield or yield class and tissue features. The true negative (TN) specimens are high-yielding and nutritionally balanced specimens.

Model	Area Under Curve	Accuracy	Number of TN Specimens
		Raw concentration	
Random Forest	0.673	0.646	64
Catboost	0.689	0.664	62
		Centered log ratio (*clr*)	
Random Forest	0.657	0.635	68
Catboost	0.686	0.657	61
		Weighted log ratio (*wlr*)	
Random Forest	0.930	0.880	109
Catboost	0.926	0.879	107

Table 5. Means and standard deviations (SD) of weighted log ratios (*wlrs*) and centered log ratios (*clrs*) for 109 true negative (TN) and 132 true positive (TP) specimens classified by the Catboost model.

Log Ratio	TN Specimens		TP Specimens		T Test
	Mean	SD	Mean	SD	
			Weighted log ratio (*wlr*)		
wlr N	0.090	0.055	4.127	1.996	**
wlr P	−0.571	0.076	−1.279	1.202	**
wlr K	0.244	0.038	3.713	1.275	**
wlr Ca	−0.012	0.044	−0.919	0.372	**
wlr Mg	−0.276	0.041	−5.632	1.998	**
wlr xD	0.525	0.024	−0.010	0.102	**
			Centered log ratio (*clr*)		
clr N	0.297	0.189	0.220	0.201	**
clr P	−1.742	0.241	−1.744	0.331	ns
clr K	0.706	0.114	0.712	0.149	ns
clr Ca	−0.305	0.238	−0.259	0.210	ns
clr Mg	−2.158	0.231	−2.080	0.194	**
clr xD	3.202	0.177	3.151	0.200	*

ns, *, **: non-significant and significant at the 0.05 and the 0.01 levels, respectively.

The interpretation of the results may also differ between the *wlr* and *clr* indices where the weighting coefficients vary widely. The *wlr* showed the relative excess of N and K and the relative shortage of Mg and x_D in the TP specimens, while P and Ca were close to the zero balance (Figure 6). The *clr* diagnosed the relative excess of K, Ca, and Mg and the relative shortage of N and x_D in the TP specimens, while P was close to the zero balance. In both cases, the x_D was negative, indicating insufficient carbon accumulation relative to other components. Combining CoDa with the gain ratios generated by ML methods is promising to conduct nutrient diagnosis.

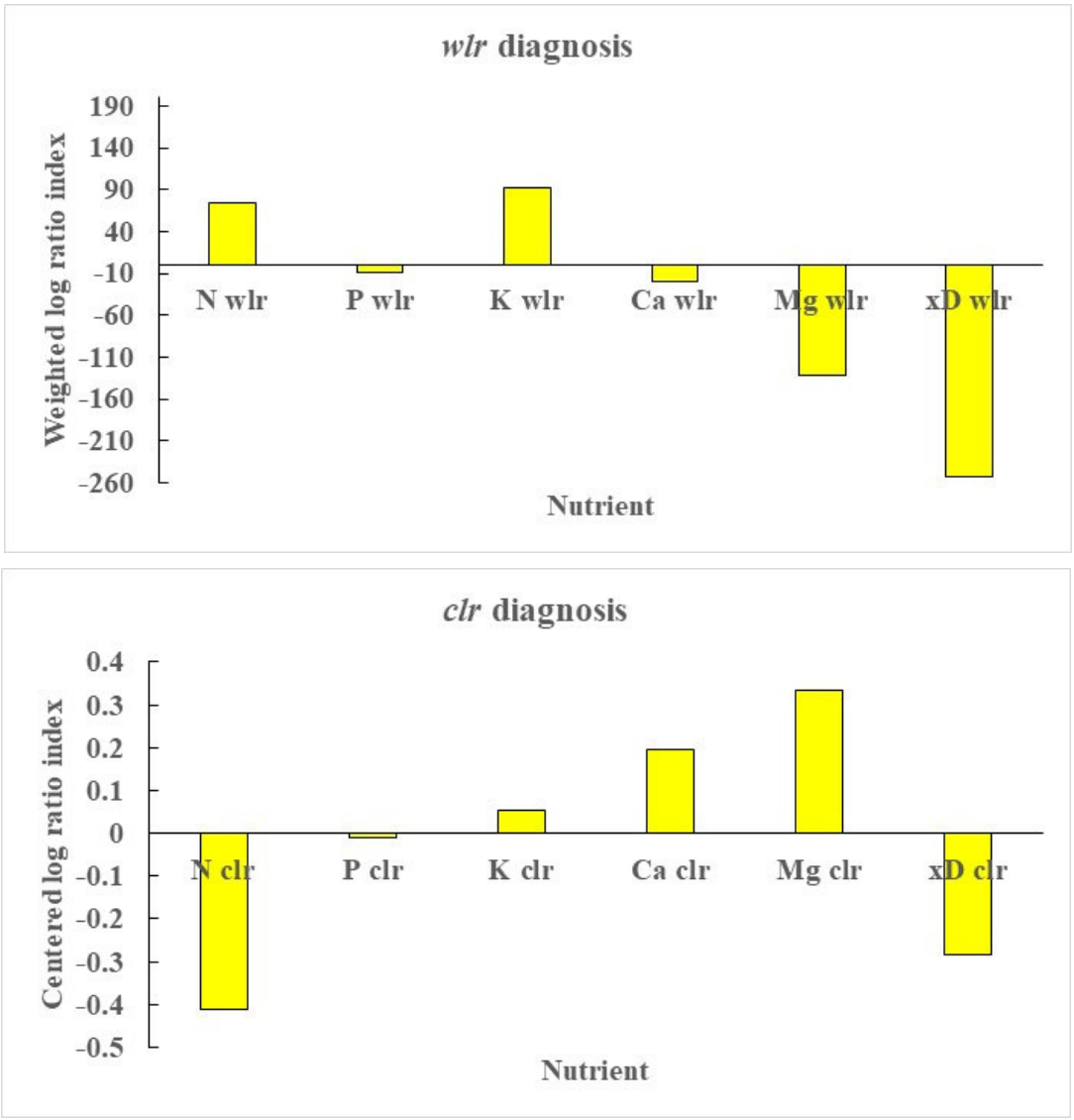

Figure 6. Diagnosis of the mean *wlr* and *clr* values of the true positive (low-yielding and nutritionally imbalanced) population using the *wlr* and *clr* means and standard deviations of the true negative (high-yielding and nutritionally balanced) population as nutrient standards.

5. Conclusions

The additive and orthogonal log ratios reduce information redundancy in compositional data without information loss. Those log ratios return the exact number of degrees of freedom available in a composition for use in ML predictive models. While accurate crop response patterns can be derived from the ML models, the most appropriate non-linear response curve must still be selected carefully, combining economic and environmental costs and judgment on the outcomes.

The centered log ratio is a sound mathematical and nutrient diagnostic tool. The variance of the *clr* provides an additional means to select the proper sampling time and growth stage to elaborate nutrient standards. However, the importance of each dual log ratio regarding the target variable is still neglected in the formulation of the *clr* (as well as the DRIS) but may vary widely depending on the database. The weighted log ratio (*wlr*) formulation proposed in this paper can account for the unequal importance of each dual log ratio regarding crop productivity and plant health. The gain ratios generated by the ML classification methods are specific to each dual log ratio. The *wlr* expression was found to increase the accuracy of the ML models compared to the *clr* expression

using a Canadian onion database. The *wlr* increased ML model accuracy and returned more nutritionally balanced and high-yielding specimens to derive nutrient standards. Nevertheless, the weighting may vary widely among databases and result in different outcomes. The RReliefF scores provide weighting coefficients in the regression mode.

Compared to traditional methods, CoDa and ML tools offer great potential to decrypt yield-impacting features simultaneously, diagnose plant nutrition, and make fertilizer recommendations. This paper presented crop nutrition models in relation to crop productivity. Because crop fertilization and tolerance to pests are also closely related, a key challenge in sustainable agriculture is to search for and maintain the right combination of nutrients, not only to tackle the most limiting nutrients for crop productivity but also to support plant defense mechanisms (trophobiosis).

Funding: This research received no external funding.

Data Availability Statement: The Brazilian garlic database is available at Zenodo DOI https://doi.org/10.5281/zenodo.10615658 (accessed on 10 December 2024). The Quebec onion, carrot, and potato databases are available upon request.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Danhke, W.C.; Olson, R.A. *Soil Test Correlation, Calibration, and Recommendation*; Westerman, R.L., Ed.; Soil Science Society of America Book Series No. 3; Soil Science Society of America: Madison, WI, USA, 1990; pp. 454–471.
2. Nava, G.; Reisser Júnior, C.; Parent, L.E.; Brunetto, G.; Moura-Bueno, J.M.; Navroski, R.; Benati, J.A.; Barreto, C.F. Esmeralda Peach (*Prunus persica*) Fruit Yield and Quality Response to Nitrogen Fertilization. *Plants* **2022**, *11*, 352. [CrossRef] [PubMed]
3. Dordas, C. Role of nutrients in controlling plant diseases in sustainable agriculture. A review. *Agron. Sustain. Agric. Dev.* **2008**, *28*, 33–46. [CrossRef]
4. Chaboussou, F. *Healthy Crops: A New Agricultural Revolution*; Jon Carpenter Publishing: Charlbury, UK, 2004.
5. Martinez, D.A.; Loening, U.E.; Graham, M.C.; Gathorne-Hardy, A. When the Medicine Feeds the Problem; Do Nitrogen Fertilisers and Pesticides Enhance the Nutritional Quality of Crops for Their Pests and Pathogens? *Front. Sustain. Food Syst.* **2021**, *5*, 701310. [CrossRef]
6. Westhues, C.C.; Simianer, H.; Beissinger, T.M. learnMET: An R package to apply machine learning methods for genomic prediction using multi-environment trials data. In *G3*; 2022; 12. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC9635651/pdf/jkac226.pdf (accessed on 25 January 2025). [CrossRef]
7. Parent, S.-É.; Leblanc, M.A.; Parent, A.C.; Coulibali, Z.; Parent, L.E. Site-Specific Multilevel Modeling of Potato Response to Nitrogen Fertilization. *Front. Environ. Sci.* **2017**, *5*, 81. [CrossRef]
8. Hahn, L.; Kurtz, C.; Paula, B.V.d.; Feltrim, A.L.; Higashikawa, F.S.; Moreira, C.; Rozane, D.E.; Brunetto, G.; Parent, L.E. Feature-specific nutrient management of onion (*Allium cepa*) using machine learning and compositional methods. *Nat. Sci. Rep.* **2024**, *14*, 6034. [CrossRef]
9. Meng, L.; Liu, H.L.; Ustin, S.; Zhang, X. Predicting Maize Yield at the Plot Scale of Different Fertilizer Systems by Multi-Source Data and Machine Learning Methods. *Remote Sens.* **2021**, *13*, 3760. [CrossRef]
10. Parent, L.E. Vegetable response to added nitrogen and phosphorus using machine learning decryption and N/P ratios. *Horticulturae* **2024**, *10*, 356. [CrossRef]
11. Ransom, C.J.; Kitchen, N.R.; Camberato, J.J.; Carter, P.R.; Ferguson, R.B.; Fernandez, F.G.; Franzen, D.W.; Laboski, C.A.M.; Myers, D.B.; Nafziger, E.D.; et al. Statistical and machine learning methods evaluated for incorporating soil and weather into maize nitrogen recommendations. *Comput. Electron. Agric.* **2019**, *164*, 104872. [CrossRef]
12. Yamane, D.R.; Parent, S.-É.; Natale, W.; Cecílio Filho, A.B.; Rozane, D.E.; Nowaki, R.H.D.; Mattos Junior, D.d.; Parent, L.E. Site-Specific Nutrient Diagnosis of Orange Groves. *Horticulturae* **2022**, *8*, 1126. [CrossRef]
13. Aitchison, J. *The Statistical Analysis of Compositional Data*; Chapman and Hall: London, UK, 1986.
14. Chayes, F. On correlation between variables of constant sum. *J. Geophys. Res.* **1960**, *65*, 4185–4193. [CrossRef]
15. Aitchison, J. Principles of compositional data analysis. *Multivar. Anal. Its Appl. IMS Lect. Notes—Monogr. Ser.* **1994**, *24*, 73–81.
16. Parent, L.E.; Dafir, M. A Theoretical Concept of Compositional Nutrient Diagnosis. *J. Am. Soc. Hortic. Sci.* **1992**, *117*, 239–242. [CrossRef]
17. Parent, L.E.; de Almeida, C.X.; Parent, S.É.; Hernandes, A.; Egozcue, J.J.; Kätterer, T.; Gülser, C.; Bolinder, M.A.; Andrén, O.; Anctil, F.; et al. Compositional analysis for an unbiased measure of soil aggregation. *Geoderma* **2012**, *179–180*, 123–131. [CrossRef]

18. Lopez, J.; Parent, L.E.; Tremblay, N.; Gosselin, A. Sulfate accumulation and Ca balance in hydroponic tomato culture. *J. Plant Nutr.* **2002**, *25*, 1585–1597. [CrossRef]
19. Munson, R.D.; Nelson, W.L. *Principles and Practices in Plant Analysis*; Westerman, R.L., Ed.; Soil Testing and Plant Analysis; Soil Science Society of America Book Series No. 3; Soil Science Society of America: Madison, WI, USA, 1990; pp. 359–387.
20. Lagatu, H.; Maume, L. Le diagnostic foliaire de la pomme de terre. *Ann. De L'école Natl. D'agronomie De Montp.* **1934**, *22*, 50–158.
21. Ulrich, A.; Hills, F.J. Principles and practices of plant analysis. In *Soil Testing and Plant Analysis. Part II*; Stelly, M., Hamilton, H., Eds.; Soil Science Society of America: Madison, WI, USA, 1967; pp. 359–387.
22. Wilkinson, S.R.; Grunes, D.L.; Sumner, M.E. Nutrient interactions in soil and plant nutrition. In *Handbook of Soil Science*; Sumner, M.E., Ed.; CRC Press: Boca Raton, FL, USA, 2000; pp. 89–112.
23. Fageria, N.K.; Baligar, V.C. Improving nutrient use efficiency of annual crops in Brazilian acid soils for sustainable crop. *Commun. Soil Sci. Plant Anal.* **2001**, *32*, 1303–1319. [CrossRef]
24. Courbet, G.; Gallardo, K.; Vigani, G.; Brunel-Muguet, S.; Trouverie, J.; Salon, C.; Ourry, A. Disentangling the complexity and diversity of crosstalk between sulfur and other mineral nutrients in cultivated plants. *J. Exp. Bot.* **2019**, *70*, 4183–4196. [CrossRef]
25. Prevot, P.; Ollagnier, M. Law of the minimum and balanced mineral nutrition. In *Plant Analysis and Fertilizer Problems*; Reuther, W., Ed.; American Institute of Biological Sciences: Herndon, VA, USA, 1961; pp. 257–277.
26. Jarrell, W.M.; Beverly, R.B. The dilution effect in plant nutrition studies. *Adv. Agron.* **1981**, *34*, 197–224.
27. Walworth, J.L.; Sumner, M.E. The Diagnosis and Recommendation Integrated System (DRIS). *Adv. Soil Sci.* **1987**, *6*, 149–188. [CrossRef]
28. Beaufils, E.R. Diagnosis and recommendation integrated system (DRIS). In *Soil Science Bulletin 132*; University of Natal: Pietermaritzburg, South Africa, 1973.
29. Diaz-Zorita, M.; Perfect, E.; Grove, J.H. Disruptive methods assessing soil structure. *Soil Tillage Res.* **2002**, *64*, 3–22. [CrossRef]
30. Weltje, G.J. Ternary sandstone composition and provenance: An evaluation of the 'Dickinson model'. In *Compositional Data Analysis in the Geosciences: From Theory to Practice*; Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V., Eds.; Special Publication 264; Geological Society: London, UK, 2006; pp. 79–99.
31. Greenacre, M.; Gunsky, E.; Bacon-Shone, J.; Erb, I.; Quinn, T. Aitchison's Compositional Data Analysis 40 Years On: A Reappraisal. *Stat. Sci.* **2023**, *38*, 386–410. [CrossRef]
32. Greenacre, M. Compositional data analysis. *Annu. Rev. Stat. Its Appl.* **2021**, *8*, 271–299. [CrossRef]
33. Mert, M.C.; Filzmoser, P.; Hron, K. Error Propagation in Isometric Log-ratio Coordinates for Compositional Data: Theoretical and Practical Considerations. *Math. Geosci.* **2016**, *48*, 941–961. [CrossRef]
34. Chlningaryan, A.; Sukkarieh, S.; Whelan, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Comput. Electron. Agric.* **2018**, *151*, 61–69. [CrossRef]
35. Huynh-Thu, V.A.; Geurts, P. Unsupervised Gene Network Inference with Decision Trees and Random Forests. In *Gene Regulatory Networks; Methods in Molecular Biology*; Sanguinetti, G., Huynh-Thu, V., Eds.; Humana Press: New York, NY, USA, 2019; Volume 1883.
36. Benton-Jones, J.; Case, W.W. Sampling, handling, and analyzing plant tissue samples. In *Soil Testing and Plant Analysis*, 3rd ed.; Westerman, R.L., Ed.; Book Ser. 3; Soil Science Society of America: Madison, Wi, USA, 1990; pp. 389–4427.
37. Filzmoser, P.; Hron, K.; Reimann, C. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. *Sci. Total Environ.* **2009**, *407*, 6100–6108. [CrossRef]
38. Ingestad, T.A. Definition of optimum nutrient requirements in birch seedlings. III. Influence of pH and temperature of nutrient solution. *Physiol. Plants* **1979**, *46*, 31–35. [CrossRef]
39. Santos, H.G.D.; Jacomine, P.K.T.; Anjos, L.H.C.D.; Oliveira, V.A.D.; Lumbreras, J.F.; Coelho, M.R.; Almeida, J.A.; Araujo Filho, J.C.D.; Oliveira, J.B.D.; Cunha, T.J.F. *Brazilian Soil Classification System*; Embrapa Publ.: Brasilia, Brazil, 2018; p. 20.
40. Egozcue, J.J.; Pawlowsky-Glahn, V. Groups of parts and their balances in compositional data analysis. *Math. Geosci.* **2005**, *37*, 795–828. [CrossRef]
41. Pawlowsky-Glahn, V.; Egozcue, J.J.; Tolosana-Delgado, R. Principal balances. In Proceedings of the 4th International Workshop on Compositional Data Analysis (Codawork 2011), San Feliu de Guixols, Spain, 1–5 June 2011.
42. Hron, K.; Filzmoser, P.; de Caritat, P.; Fišerová, E.; Gardlo, A. Weighted pivot coordinates for compositional data and their application to geochemical mapping. *Math. Geosci.* **2017**, *49*, 777–796. [CrossRef]
43. Hassink, J. Preservation of Plant Residues in Soils Diflfering in Unsaturated Protective Capacity. *Soil Sci. Soc. Am. J.* **1996**, *60*, 487–491. [CrossRef]
44. Ravelojaona, N.; Jégo, G.; Ziadi, N.; Mollier, A.; Lafond, J.; Karam, A.; Morel, C. STICS Soil–Crop Model Performance for Predicting Bsinclairiomass and Nitrogen Status of Spring Barley Cropped for 31 Years in a Gleysolic Soil from Northeastern Quebec (Canada). *Agronomy* **2023**, *13*, 2540. [CrossRef]
45. Sinclair, T.R.; Seligman, N. Criteria for publishing papers on crop modeling. *Field Crops Res.* **2000**, *8*, 165–172. [CrossRef]

46. Cerrato, M.E.; Blackmer, A.M. Comparison of models for describing maize yield response to nitrogen fertilizer. *Agron. J.* **1990**, *82*, 138–143. [CrossRef]
47. Government of Canada. Carbon Pollution Pricing Systems Across Canada. Available online: https://www.canada.ca/en/environment-climate-change/services/climate-change/pricing-pollution-how-it-will-work.html (accessed on 9 December 2024).
48. Galvez-Cloutier, R.; Sanchez, M. Trophic Status Evaluation for 154 Lakes in Quebec, Canada: Monitoring and Recommendations. *Water Qual. Res. Can.* **2007**, *4*, 252–268. [CrossRef]
49. Bould, C.; Bradfield, E.G.; Clarke, G.M. Leaf analysis as a guide to the nutrition of fruit crops. I. General principles, sampling techniques and analytical methods. *J. Sci. Food Agric.* **1960**, *11*, 229–242. [CrossRef]
50. Robnik-Šikonja, M.; Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Mach. Learn. J.* **2003**, *53*, 23–69. [CrossRef]