PLOS ONE

RESEARCH ARTICLE

Cultivar-specific nutritional status of potato (*Solanum tuberosum* L.) crops

Zonlehoua Coulibali[1], Athyna Nancy Cambouris[2], Serge-Étienne Parent[1]*

1 Department of Soils and Agrifood Engineering, Université Laval, Québec City, Québec, Canada, **2** Quebec Research and Development Centre, Agriculture and Agri-Food Canada, Québec City, Québec, Canada

* serge-etienne.parent.1@ulaval.ca

OPEN ACCESS

Citation: Coulibali Z, Cambouris AN, Parent S-É (2020) Cultivar-specific nutritional status of potato (*Solanum tuberosum* L.) crops. PLoS ONE 15(3): e0230458. https://doi.org/10.1371/journal.pone.0230458

Editor: Paul Esker, Pennsylvania State University, UNITED STATES

Received: September 30, 2019

Accepted: March 1, 2020

Published: March 13, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0230458

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. There is no restriction on sharing of data and/or materials.

Abstract

Gradients in the elemental composition of a potato leaf tissue (*i.e.* its ionome) can be linked to crop potential. Because the ionome is a function of genetics and environmental conditions, practitioners aim at fine-tuning fertilization to obtain an optimal ionome based on the needs of potato cultivars. Our objective was to assess the validity of cultivar grouping and predict potato tuber yields using foliar ionomes. The dataset comprised 3382 observations in Québec (Canada) from 1970 to 2017. The first mature leaves from top were sampled at the beginning of flowering for total N, P, K, Ca, and Mg analysis. We preprocessed nutrient concentrations (ionomes) by centering each nutrient to the geometric mean of all nutrients and to a filling value, a transformation known as row-centered log ratios (clr). A density-based clustering algorithm (*dbscan*) on these preprocessed ionomes failed to delineate groups of high-yield cultivars. We also used the preprocessed ionomes to assess their effects on tuber yield classes (high- and low-yields) on a cultivar basis using k-nearest neighbors, random forest and support vector machines classification algorithms. Our machine learning models returned an average accuracy of 70%, a fair diagnostic potential to detect in-season nutrient imbalance of potato cultivars using clr variables considering potential confounding factors. Optimal ionomic regions of new cultivars could be assigned to the one of the closest documented cultivar.

1 Introduction

Potato cultivars are commonly classified into maturity groups based on the number of days from planting to maturity [1]. Compared to other maturity groups, cultivars with longer maturity generally show yield potential that is similar or higher [2–4] because of higher genetic potential [5] related to higher foliar nitrogen status [6] and root acquisition rate [7]. Hence, nutrient management of potato cultivars often consider the cultivar maturity group. However, nutrient profiles or ionomes [8, 9] may vary among potato cultivars of the same maturity groups because cultivars inherit from a diversity of parents specific traits for nutrient absorption and assimilation [10]. Indeed, White et al. [11] provided evidence of important ionome variations in angiosperm species and stated that plant families could be distinguished by their shoot ionomes. Successful classifications of plant species based on axis-reductions have been

Funding: ZC is partly funded by the Natural Sciences and Engineering Council of Canada (CRDPJ 385199-09 and DG-2254 - https://www.nserc-crsng.gc.ca), the Quebec Ministry of Agriculture, Fisheries and Food (IA216581 - https://www.mapaq.gouv.qc.ca), Centre SEVE (https://centreseve.recherche.usherbrooke.ca/), Patate Dolbec Inc. (https://patatesdolbec.com/), Groupe Gosselin FG (http://gosseling2.com), Agriparmentier Inc., Ferme Daniel Bolduc Inc. (http://fermedanielbolduc.com/), Patate Laurentienne, Ferme Bergeron-Niquet, and Patates Lac-St-Jean (http://plsj.ca/). There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript

Competing interests: All the funders (Natural Sciences and Engineering Council of Canada, Quebec Ministry of Agriculture, Fisheries and Food, Centre SEVE, Patate Dolbec Inc., Groupe Gosselin FG, Agriparmentier Inc., Patate Laurentienne, Ferme Bergeron-Niquet, and Patates Lac-St- Jean) have declared that no competing interests exist. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

implemented on compositionally preprocessed plant ionomes [12, 13]. The potato cultivar may also be classified similarly, allowing newly introduced cultivars to benefit from the documented nutrient management of older cultivars. Hence, the foliar ionome, easily collected from field trials, could provide a tool for the fertilization of newly introduced cultivars.

Tissue ionome portrays plant nutritional status [13] under the assumption of causal relationships between plant growth rate and nutrient concentration in a tissue [14, 15]. In survey datasets, reference compositions are those that are nutritionally balanced [12]. Imbalanced ionomes could be rebalanced theoretically through a perturbation operation [16] *i.e.*, a change in tissue composition after nutrient stress has been applied. Any factor impacting yield response to nutrients can perturb leaf composition [17]. Fertilization perturbs soil composition [18] by supplying readily available plant-nutrients [19].

Because nutrients interact in the plant, Baxter [20] suggested that the ionome could be treated as a combination of elements rather than elements taken in isolation. Parent [13] described ionomes as multivariate balance systems of isometric log-ratios [16]. Isometric log-ratios maps vectors of concentrations, which are strictly positive data constrained to the measurement unit that convey only relative information, to a real space of orthonormal coordinates [21]. Indeed, ionomes are intrinsically multivariate: each part cannot be interpreted without being related to the other parts of the whole [22]. Parent and Dafir [23] developed the compositional nutrient diagnosis in plants using row-centered log-ratios (clr). Thereafter, compositional data transformation has been used to preprocess combined nutrients traits of plant species and cultivars [13, 24–26] as well as animal species [27], and human food [28, 29].

The first objective of this study was to identify clusters of potato cultivars based on their leaf ionomes. The second objective was to develop, evaluate and compare the performance of machine learning algorithms in predicting yield categories using ionomes. The third objective was to develop a conceptual workflow to adjust the ionome of potato cultivars using compositional perturbations. Our hypotheses were that (1) nutritionally balanced leaf ionomes of potato cultivars differ among potato cultivars, (2) tuber yield is impacted by specifically leaf compositional traits, and (3) cultivar-specific leaf ionomes could be rebalanced using a perturbation operation.

2 Methodology

2.1 Data set

The data set is a collection of potato surveys, and nitrogen (N), phosphorus (P) and potassium (K) fertilizer trials conducted in the province of Québec (Canada) from 1970 to 2017 (S1 Table) between the US border (45th parallel) and the Northern limit of cultivation (49th parallel). The data set was filtered to remove foliar samples collected too early or too late from the beginning (10%) of flowering, as reported by scouting teams, and where three or more of the five major elements (N, P, K, Ca and Mg) have not been quantified. The complete data set comprised 3382 observations of 151 field trials. Five maturity classes were represented, and we matched the duration from planting to harvest described by the Canadian Food Inspection Agency [1], although the names differed: early season (65–70 days), early mid-season (70–90 days), mid-season (90–110 days), mid-season late (110–130) and late season (130 days and more) cultivars. The number of samples per cultivar and the corresponding maturity classes are reported in S2 Table.

2.2 Diagnostic tissue composition

The potato diagnostic tissue is the first mature leaf (4th leaf from top) sampled at the beginning (10%) of the blooming stage [15, 30]. Twenty to 30 leaves were collected at random in each

plot, composited, dried at 65˚C, ground to pass through a 1 mm sieve, and analyzed for N, P, K, Ca and Mg concentrations after dissolution of combustion. Total N was determined by micro-Kjeldahl or Dumas combustion (Leco CNS-2000 analyzer, St. Joseph, MI, USA). After acid dissolution [31], K, Ca, and Mg concentrations were quantified by atomic absorption spectrometry or inductively coupled plasma spectroscopy (ICP), and P by colorimetry or ICP. We made no distinction between methodologies in the analysis of ionomes.

2.3 Processing nutrient composition to nutrient balances

The compositional space [16] of the leaf tissue comprised five nutrients (N, P, K, Mg, Ca) and undetermined components amalgamated into a filling value (Fv) computed by difference between the measurement unit and the sum of quantified nutrients. Tissue components were preprocessed using the row-centered log-ratio transformation, as follows [23]:

$$clr_i = ln\left(\frac{x_i}{g(x)}\right) \quad (1)$$

where x_i is raw concentration of the i^{th} component and $g(x)$ is the geometric mean across components including the filling value.

2.4 Clustering cultivars

Yield thresholds are useful for decision-making. Because tuber yield potential varies widely among cultivars, we processed by discretizing tuber yields into low- and high-productivity categories [12] by ranking the marketable yield in ascending order within a given cultivar, and selecting the yield corresponding to the 65^{th} percentile as cut-off between the two subgroups. Hence, each cultivar had its cut-off with respect to its yield potential as shown in S2 Table. The high-yielding subpopulation ionomes were used to assess cultivars clustering ability. This subgroup comprised 1190 occurrences (after the exclusion of 144 outliers) across 151 trials and 47 cultivars. A density-based clustering method [32] was used to delineate cultivar groups of similar compositions using clr variables.

2.5 Ionome effect and yield prediction

Machine learning algorithms can either regress to predict continuous variables or classify to predict categories [33]. Tuber yield categories were predicted using clr variables and information on ionomic groups of the full data set (high and low yielders *i.e.* 3382 rows). Three machine learning algorithms were compared: k-nearest neighbors, random forest and support vector machines.

We estimated the relative influence of variables in the model and their rank by examining how can prediction error increases when data for a variable is permuted while all others are left unchanged [34, 35]. A variable can score a zero or too small value compared to others. Deleting such variable from the dataset should not impact on the results. The random forest model was used for feature selection to assess importance of each clr variable in predicting tuber yield, but none of the variable was removed.

The data were split into training (75%) and testing (remaining 25%) sets at cultivar level i.e., for each cultivar the samples were randomly separated according to these proportions. The performance of the classification models was assessed using accuracy computed with the testing set. Applied to the context, the four quadrants defined by Swets [36] in binary system diagnosis to delineate the response classes are presented in the contingency table (Table 1).

Table 1. Term definitions used for the study.

		Observed yield	
		Low (unbalanced)	**High (balanced)**
Predicted yield	Low	**True positive (TP)**: observed low-yielders correctly predicted as low-yielders.	**False positive (FP)**: observed high-yielders incorrectly predicted as low-yielders.
	High	**False negative (FN)**: observed low-yielders incorrectly predicted as high-yielders.	**True negative (TN)**: observed high-yielders correctly predicted as high-yielders.

As in medical sciences, the *negative* term is used in cases where no intervention is needed after diagnosis.

https://doi.org/10.1371/journal.pone.0230458.t001

The accuracy is the proportion of correctly-predicted instances:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (2)$$

2.6 Rebalancing a composition: The enchanting islands

A compositional perturbation is a translation in the compositional space [37, 38]. A perturbation vector expressed as clr values contains a series of deltas (differences). Once back-transformed into the compositional space, the perturbation vector alters a composition through a perturbation ($\oplus$) operation as follows [37]:

$$A \oplus \mathrm{B} = [a_1, a_2, \ldots, a_D] \oplus [b_1, b_2, \ldots, b_D] = \mathcal{C}(a_1 \times b_1, a_2 \times b_2, \ldots, a_D \times b_D) \quad (3)$$

where a D-part composition A is perturbed ($\oplus$) by a D-part composition B, and $\mathcal{C}$ is the closure operator to constant sum.

We used the testing set to display the effect of a perturbation across the simplex. We selected two elements (N and P) and simulated an increase of their initial (observed) clr values by 20% (theoretically). The observed (ionome of the instance) and new clr vector (perturbed ionome) were back-transformed into N, P, K, Ca, Mg and Fv compositional space for comparison using familiar concentration units.

The high yielders of the training set correctly diagnosed as balanced (true negative specimens) by the most accurate model were used as the reference subpopulations. The clr values of these reference specimens were used as reference nutritional status at high yield potential. A potato nutrient imbalance index was computed as a distance from the closest high-yielding specimen using the Aitchison distance, *i.e.* the Euclidean distance between compositions using clr-transformed concentrations [39]. For any misbalanced or new specimen of a given cultivar, the closest true negative (closest reference composition) was identified as the sample with the minimum Aitchison distance from the new composition. The nutrient clr differences defining the Aitchison distance may be considered as apparently excess or deficiency of the nutrient requiring correcting measures in a multivariate and compositional data perspective [40]. Hence, the clr space of nutrient components (N, P, K, Ca, Mg) was described not as an ellipsoidal hyper-space [41] but as islands of high-yielding specimens dispersed in the hyper-space of differently yielding specimens. The closer is a specimen from the enchanting island, the higher its chance to become a high-yielder [40]. The clr-difference was converted into a perturbation vector between two nutrient compositions expressed as familiar nutrient concentrations.

2.7 Statistical analysis

Statistical computations were performed in the R statistical environment version 3.6.1 [42]. Compositional data analysis was conducted using the R *compositions* package version 1.40–2 [43]. Multivariate outliers were removed for robust multivariate analysis [44] using the

Mahalanobis distance at a 0.01 level of significance with the R *mvoutlier* package version 2.0.9 [45]. The clustering operation were performed using *dbscan* package version 1.1–3 [32]. Linear discriminant analysis (LDA) was conducted using the R *ade4* package version 1.7–13 [46] which allows computing linear combinations of clr coordinates that best discriminate cultivars ionomes centroids. Supervised analysis was conducted using the *caret* package version 6.0–84 [47]. Our results are reproducible by using the R computation codes and data given as supplementary information and available online in a GitHub repository (https://git.io/Jvt2r).

3 Results

3.1 Cluster analysis

The data set used for clustering is described in S2 Table. The AC Chaleur cultivar showed the lowest tuber marketable yield cut-off (65th percentile) at 17.4 Mg ha^{-1} and Red-Maria, the highest at 64.6 Mg ha^{-1}. Average marketable yield was 40.5 Mg ha^{-1} for high yielders and 24.8 Mg ha^{-1} for low yielders. In comparison, average potato tuber yields in Canada and Québec were 31.2 Mg ha^{-1} and 32.2 Mg ha^{-1} respectively, in 2018 [48].

The *dbscan* clustering function looked for dense regions in the clr-space, and detected a single cluster of cultivars *i.e.*, cultivars were scattered without any particular dense region. A principle components analysis allowed to map cultivars and nutrients in the biplot shown in Fig 1. The principle components scores mapped on the distance biplot (Fig 1A) showed no particular pattern allowing groups partition. The clr correlation loadings (Fig 1B) showed a negative

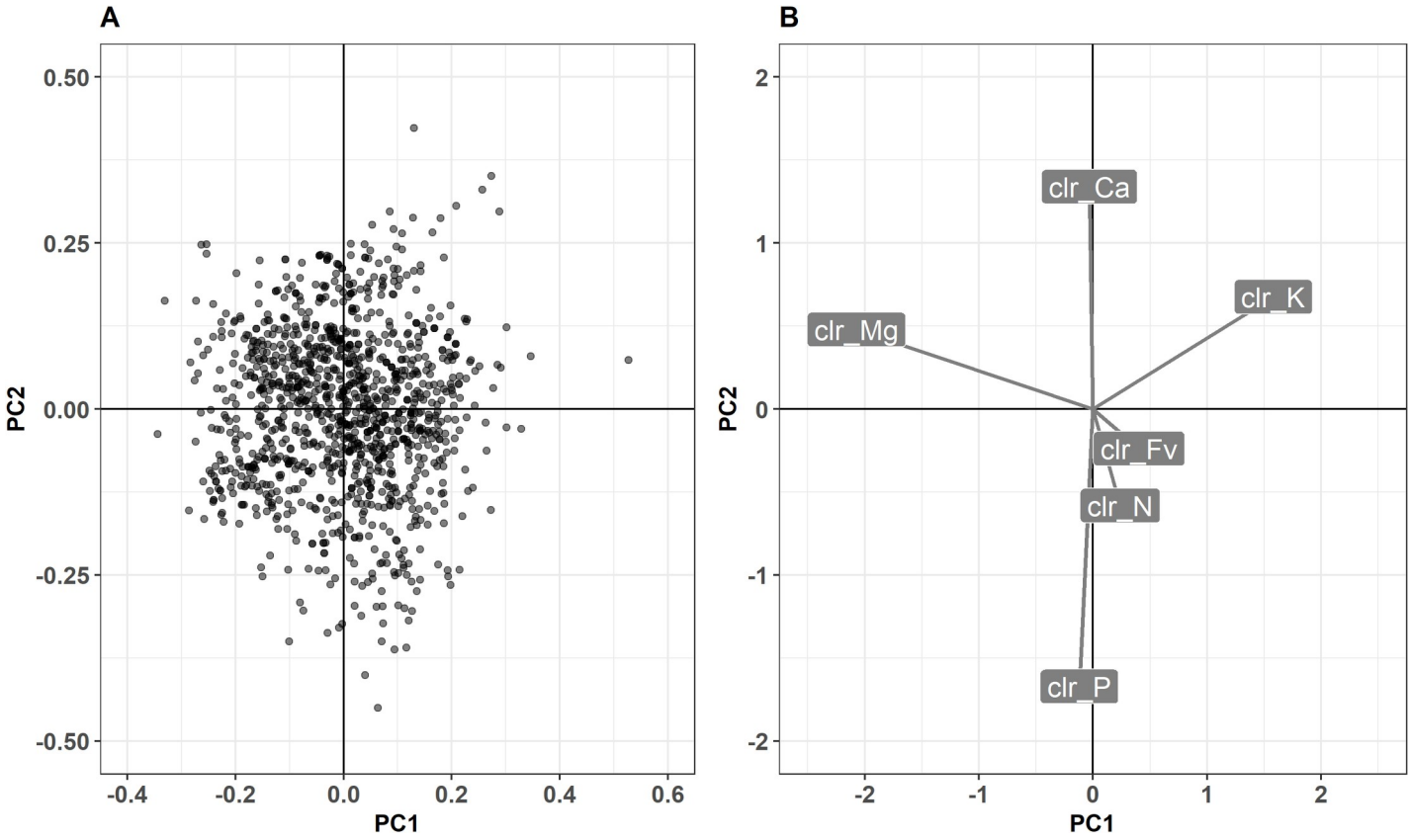

Fig 1. Principle components biplot of potato ionome showing (A) scores in distance scaling and (B) loadings in correlation scaling.

https://doi.org/10.1371/journal.pone.0230458.g001

relationship between K and Mg, P and Ca, and positive relationship between N and P in agreement to concentration changes with time as the plant matures [49]. Discrepancies between cultivars were driven mainly by Mg and K on the first axis, and by P and Ca on the second axis (right hand side plot).

3.2 Predicting tuber yield

Classification models assigned explanatory clr variables to two categorical tuber marketable yield: high- and low-yielders. The random-forest algorithm allowed to rank the importance of variables in the model. The clr of nitrogen appeared to be the most discriminant variable between tuber yield categories, followed by the amalgamated unknown components (Fv), then Ca, Mg and, finally, P.

After splitting data into training (75%) and testing (25%) data sets, we used a ten-fold cross-validation process that sequentially splits the training data set into ten parts, using nine parts for calibration and the remainder for validation. The *k*-nearest neighbours, the random forest and the support vector machine models returned practically similar predictive accuracies (although slightly lower for the support vector machine algorithm), with a mean accuracy of 70% representing 591 successful and 254 unsuccessful cases classification with the testing set. The null hypothesis for a random classifier *i.e.*, non-informative classification consisting of an equal distribution of 50% successful and 50% unsuccessful cases was rejected after a χ^2 homogeneity test ($\chi^2 = 69.135$, $p < 2.2\ 10^{-16}$). Since all the models returned practically similar accuracy over the testing set, predictions with the *k*-nearest neighbors model were used for interpreting. There was high variation in model fit by cultivar as shown in Fig 2. The accuracy at testing varied from 25% for Estima and Waneta, to 100% for Ambra, Carolina, Dark Red

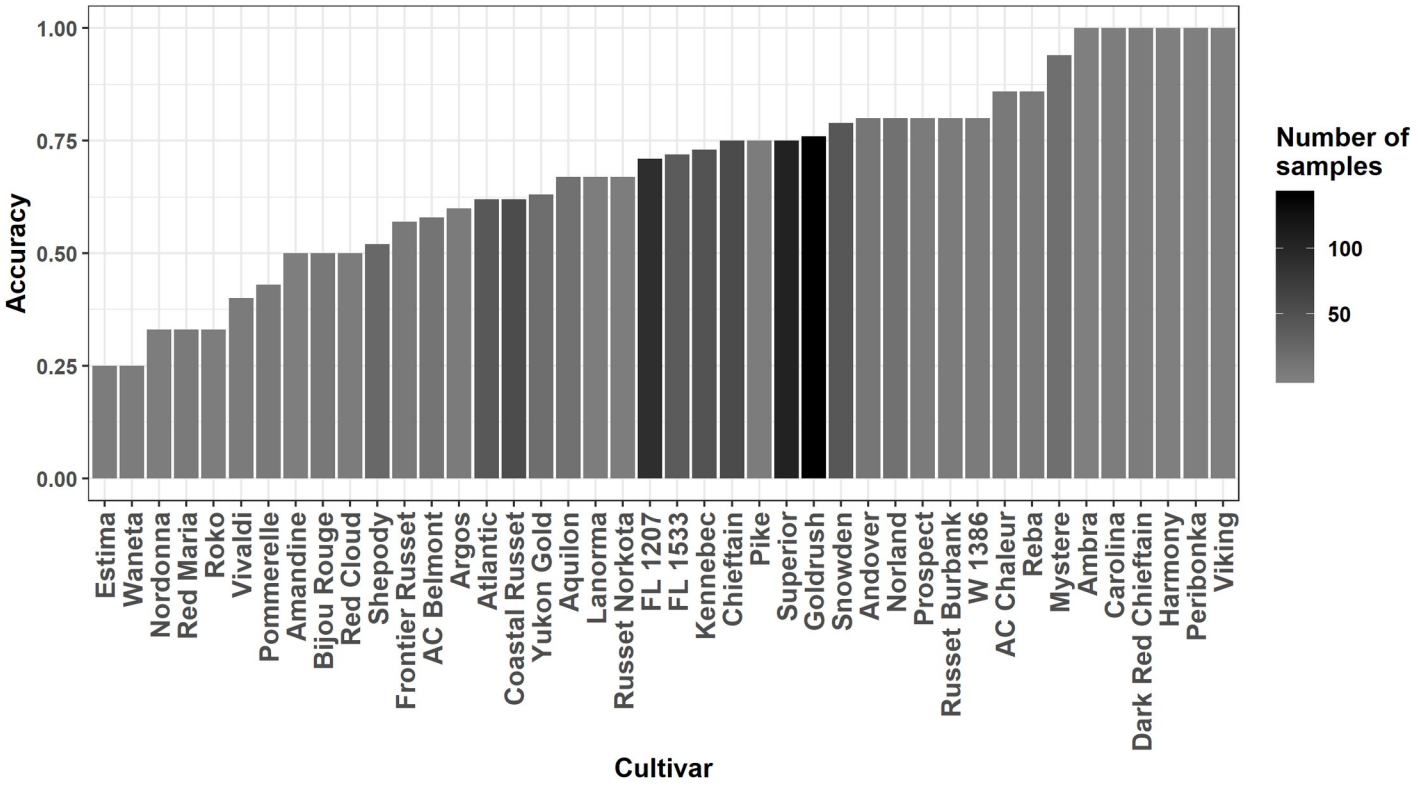

Fig 2. The k nearest neighbors model evaluation accuracies for cultivars.

https://doi.org/10.1371/journal.pone.0230458.g002

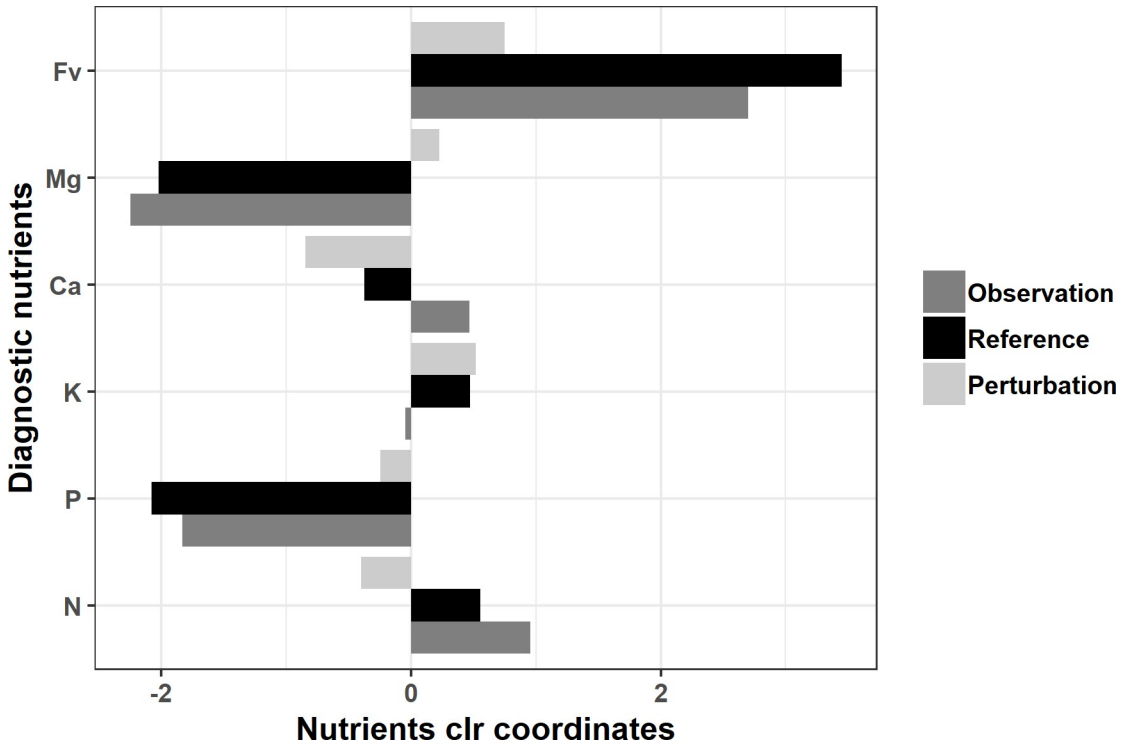

Fig 3. Perturbation vector example mapped using the most imbalanced sample. The most imbalanced observation nutrient composition was (0.0601, 0.0037, 0.0355, 0.0032, 0.0048. 0.8919), the nearest reference composition was (0.0561, 0.0036, 0.0603, 0.0052, 0.0184, 0.8565), the corresponding perturbation vector was (0.0919, 0.0965, 0.1696, 0.1629, 0.3832, 0.0959) for N, P, K, Mg, Ca and Fv respectively. The Aitchison distance computed between the observation and its associated true negative was 1.135.

https://doi.org/10.1371/journal.pone.0230458.g003

Chieftain, Harmony, Peribonka and Viking. All these cultivars had small sample sizes in the dataset, as shown in the S2 Table.

3.3 Ionome perturbation

The true negative specimens (correctly diagnosed as balanced) comprising 783 occurrences in the training data set provided the clr reference values required to compute the Aitchison distance, which is equal to the Euclidean distance across clr-transformed compositions. The S3 Table displays mean values for each cultivar. Using the Aitchison metric, the closest true negative specimen was set as the reference composition for each imbalanced specimen. In the clr-space, the difference between the reference and the imbalanced compositions returns a perturbation vector. The Fig 3 shows the imbalanced sample with the highest Aitchison distance from its reference and the perturbation to apply as a translation to reach a balanced ionome.

4 Discussion

4.1 Clustering potato cultivars

The Canadian Food Inspection Agency classified potato cultivars broadly into maturity groups based on the time elapsed between planting and maturity [1]. However, nutrient requirements, especially nitrogen, vary widely between cultivars of the same maturity group. In New Brunswick (Canada), Zebarth et al. [50] recommended 200–208 kg N ha^{-1} for Russet Norkota (early-season cultivar) and Russet Burbank (late-season cultivar), 190 kg N ha^{-1} for Superior (early-mid-season cultivar) and Goldrush (mid-season cultivar), 175 kg N ha^{-1} for Shepody (mid-

season), 135 kg N ha^{-1} for early cultivars for the table market, 160–180 kg N ha^{-1} for other mid-season, 180–200 kg N ha^{-1} for other late, and 135–160 kg N ha^{-1} for low N requirement cultivars. Such large discrepancies within the same cultivar maturity group was attributed to differential foliar gene expression [6] and root development [7]. Hence, information additional to maturity grouping is needed to assess nutrient requirements of potato cultivars. Huang and Salt [51] reported that ionomics allows the discovery of genes controlling natural variation in the plant ionome and for Salt et al. [9], ionomics could capture information about the functional state of an organism driven by genetic and environmental factors. The content of plant tissue reflects what the plant can absorb from the soil and for each nutrient, there is a correlation between its concentration and yield. Moreover, since tissue analysis is also carried out to observe the effect of fertilizer applications, and for determining the in-season or next season nutrient requirement [52, 53], ionomes could be useful in discriminating potato cultivars. Indeed, using a small data set of eight potato cultivars, Hernandes et al. [10] showed that foliar nutrient profiles varied widely among cultivars of the same maturity group. According to Parent et al. [12], variations in ionomes could be interpreted only partly as genotypic effect, and phenotypic plasticity can also be driven by nutrient supply capacity specific to agroecosystems while breeding programs are conducted under relatively luxurious environments to reach high productivity. The N, Mg and K clr values, that dominated principal components (Fig 1), could reflect the abilities of individual cultivars to acquire and use those nutrients more efficiently [54, 55]. Natale et al. [56] provided evidence that in general macronutrient contents differ among species and cultivars and within the same species for fruit trees. For N, K and Ca, this range is wider because of higher requirement of these elements by plants, and narrower for P, Mg and S, indicating smaller demand for the latter.

To cluster is to recognize that objects are sufficiently similar to be put in the same group, and to identify distinctions or separations between groups of objects [57, 58]. Based on the assumptions of differential genotypic potential, root development, nutrient requirements, nutrient uptake and use efficiency, the goal was to discover interesting structures in the N, P, K, Mg and Ca contents of the diagnosis tissue in order to decipher dissimilarities between cultivars [33]. However, the process failed to discriminate groups of cultivars along the clr coordinates. Hernandes et al. [10] reached similar results with overlapping nutrient profiles between cultivar groups depending on isometric log ratio (ilr) coordinates. They found similar nutrient profiles between cultivars groups along some ilr coordinates and very different ones along others. While ionome dissimilarities are not numerically compelling, they could assist classifying new cultivars into appropriate ionomic group to benefit from costly fertilizer trials conducted on cultivars of the same group.

4.2 Tuber marketable yield prediction

The P content of the diagnostic leaf did not appear useful in predicting potato tuber yield classes. Other elements (N, K, Ca and Mg) showed important contribution of their clr values to the prediction quality metric, especially N, which is directly related to photosynthesis [59]. Since the fertilization trials were conducted over a time span of 47 years (1970–2017), the question arises whether the different methodology of quantifying P (colorimetry/ICP) may have contributed to depreciating this variable in predicting tuber yield classes. The ICP method is shown to be faster and to give higher results for total phosphorus content in 'soil' extracts in comparison to the colorimetric method. However, there are exceptions and controversial results [60–62]. Ivanov et al. [61] found that the two methods for total P determination in plant material were highly correlated, and the results were generally within 5% to 10% of one another. Moreover, Valkama et al. [63] reported that, although agricultural practices, soil

conditions and analytical techniques have undergone substantial changes over time, the differences between old and recent experiments in yield responses to P application were not statistically important. For all these reasons, we consider the two analytical methods equally relevant to the analysis. The low importance of the P clr variable in predicting tuber yield classes may come from its correlation with Ca. Globally, the selection of relevant features is achieved, by first checking the correlation between features and response to select the features that have correlation above a selected level (*e.g.*, 0.5). Then, the independent variables need to be uncorrelated with one another. If some features are correlated, only one is kept. The process selected the *clr_Ca* variable (alphabetical order) instead of *clr_P* since these features are correlated as shown in Fig 1B. In this study no element was discarded from the process relative to its importance.

The tested algorithms (*k*-nearest neighbours, random forest and support vector machine) returned similar accuracies in the prediction of yield classes using clr variables as predictors and showed fair diagnostic potential to detect nutrient imbalance. The correctly predicted high and low yielders reached 70% in the testing data set. The models classified more accurately the yield categories compared to a random classifier [64]. Specimens classified as false negatives (*i.e.*, low yielders incorrectly classified as high yielders) are attributable to limiting conditions other than N, P, K, Mg, and Ca nutrition: soil physical and chemical properties [65, 66], fertilization [67], management failures, diseases [68] or weather events [69] impacting plants growth and yield potential. False positive specimens (*i.e.*, high yielders incorrectly classified as low yielders) indicate luxury consumption when nutrient concentrations are higher [12, 70], or other particularly favorable growth conditions. The confusion matrix built for cultivars revealed poor predictive accuracy for certain cultivars (*i.e.*, 25% for Estima and Waneta) and conversely an accuracy of 100% for others (*i.e.*, Ambra, Peribonka) as shown in Fig 2. These cultivars involved mainly small sample sizes (only one, two or three high-yielders and five, six or lightly more low-yielders). The problems of small-data in machine learning are numerous, but mainly revolve around over-fitting. The training and testing datasets division could only aggregate observations of one class in the training set so that the model would train to always predict this dominant class [71]. The model could also memorize labels, which is not ideal for generalizing from new data. Brownlee [72] explained that imbalanced classifications (one or less examples in a minority class for hundreds or more examples in the other) pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class. This results in models that have poor predictive performance, specifically for the minority class. The controversial accuracy level for some cultivars (especially low level) could also come from other yield limiting factors specific to the experiments but not involved in this study, as for false positive specimens. Our model was not effective for these cultivars treated separately.

The differential nutrition of potato cultivars could be addressed objectively using mineral analysis of the diagnostic leaf. More data are needed for poorly documented cultivars. Moreover, dedicated models could be trained for cultivars for which sufficient data are available (e.g., Goldrush, Superior, FL 1207, Chieftain). Other algorithmic, sampling and quality measurement approaches could further be implemented to deal with the problems of small-data and unequal distribution of classes [71, 72]. One could extend the predictors to the experimental conditions (soils, weather data), fitting a site-and-cultivar-specific nutrients diagnosis model.

4.3 Perturbation vector for fertilizer recommendation

Rational fertilization requires information on the nutrients that are available in the soil, and the nutritional status of the plant [14] as portrayed by the diagnostic tissue composition [14,

15]. However, the diagnosis of deficiency and toxicity of mineral nutrients may be complicated in field-grown plants where more than one mineral nutrient is deficient or where there is a deficiency of one nutrient and simultaneously toxicity of another [14]. The scientific principle behind tissue analysis is that healthy plants contain predictable concentrations of analytical nutrients [73]. The values are compared to established norms for inadequate, adequate and excess levels. However, Parent et al. [13] proved that this concept of growth-limiting nutrient concentrations supported by the *Law of minimum* and illustrated by Liebig's barrel, should be replaced by a concept of growth-limiting nutrient balances illustrated by a pan balance design, where groups of elements are balanced optimally against each other in weighing pans.

The difference between two equal-length compositional vectors can only be computed using tools of compositional data analysis. The perturbation vector concept applied to foliar tissue diagnosis returns a scaling operator [21] that when applied to an imbalanced composition translates it (theoretically) into a balanced composition with high yield potential (*i.e.*, true negative). Although the closure of the simplex implies that a perturbation on the clr of a specific nutrient is methodologically not a change in proportion of a single nutrient, perturbations expressed in the clr space appear suitable for interpretation. Indeed, the difference measured between clr values of the diagnosed sample and reference (true negative) specimen can be ranked using the sign of that difference [10, 74, 75], hence indicating which components are at excessive or deficient levels. As provided by Parent [40], K and Mg were apparently deficient while N, P and Ca were apparently in excess compared to the closest *reference* specimen (Fig 3). Using the same approach, ionomes of newly introduced cultivars with unknown nutrient requirements could be assigned to the cultivars of known nutrient requirements showing the closest ionomes.

A perturbation as the one shown in Fig 3 should not be interpreted as shifts of individual components, since the operation on a single component resonates on the whole simplex [40]. For instance, an offset in the simplex $\boldsymbol{S}$ (N, P, K, Ca, Mg, Fv) composition following the increase by 20% (theoretically) of N and P clr values is displayed on Fig 4. The K, Ca and Mg concentrations seemed more stable with respect to the others. Although P clr values have been increased, P proportion decreased globally for the new equilibrium of the simplex. The offset was higher for the selected components followed by the filling value (Fv).

Perturbation (as defined in Eq 3) is the measure of compositional change from one composition to another [37]. Because foliar composition belongs to compositional data family, the Fig 4 illustrates the principle that changing a proportion of such data affects at least another proportion of the simplex [16]. The result displayed variable offsets for other elements, decreasing or increasing to reach another balance in the simplex.

5 Conclusion

Since the concept of compositional data analysis was applied to plant tissues, several studies classified plant species and cultivars using multivariate analysis of nutrients compositions. This study is, to our knowledge, the third (following Parent et al. [49] and Hernandes et al. [10]) to use statistical tools to address the differential nutrition of potato cultivars using combination of nutrient concentrations in the diagnostic leaf, and the first using tools of machine learning to predict tuber marketable yield. The potato ionomes showed some dissimilarities in principle components analysis, but not compelling to separate definite density-based clusters between cultivars on the basis of the clr values. However, the ionome showed a determinant effect on tubers yield. Used as predictors in machine learning tools, clr variables showed diagnostic potential to detect in-season nutrient imbalance to address objectively the differential response of cultivars to fertilization. The perturbation vector of the leaf compositional space

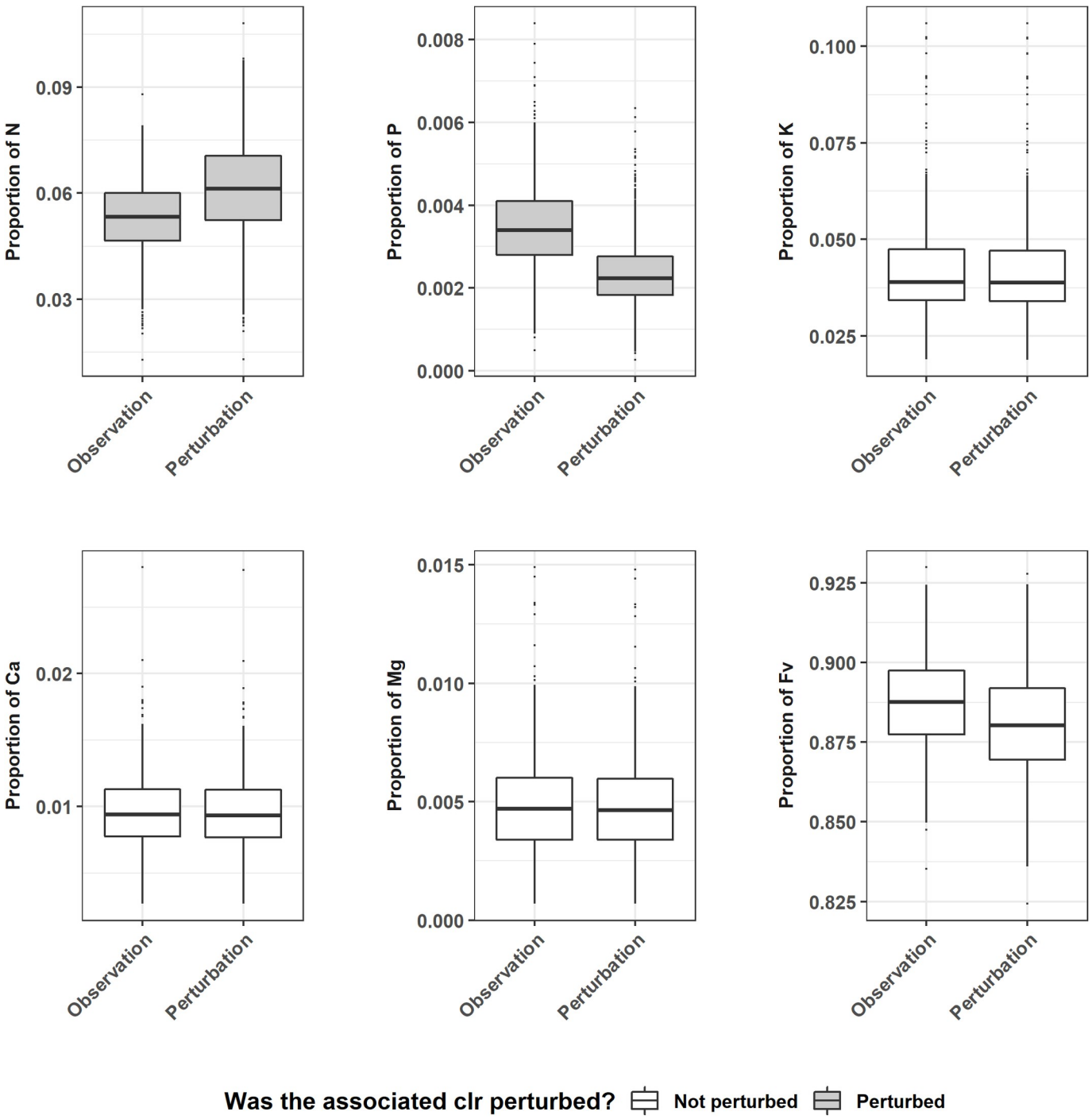

Fig 4. Effect of the perturbation of N and P clr coordinates on the other element proportions. 'Observation' stands for the element's original proportion, 'Perturbation' designates the new proportion after the 'Observed' vector's clr value was offset. Greyed boxplots plot distribution of perturbed elements of the simplex.

https://doi.org/10.1371/journal.pone.0230458.g004

could indicate cultivar sensitivity to fertilization and address specific problems of nutrient imbalance in new cultivars. Tissue testing remains an informative, diagnostic and preventive tool with real-world applications for growers in evaluating the effectiveness of their nutrient

management program. When using the right interpretation, this timely and correct tissue testing helps diagnose the presence and magnitude of suspected nutrient deficiencies. By using the compositional perturbation vector involving interactions among nutrients, our study provided a useful tool in potato precision fertilization in Quebec. The perturbation vector can help identify limiting nutrients requiring correcting measures as a season progresses or for subsequent seasons. Moreover, our study implicitly provided robust multi-nutrient norms for potato crops, gathering more cultivars of different maturity classes than the previous works. These norms are sets of true-negative or nutritionally-balanced compositions per cultivar (enchanting islands) with high-yield potential. More data are needed to fine-tune the models, especially for poorly-documented cultivars. New algorithms, other sampling methods and model quality measures could be tested to deal with the problem of small-data and imbalanced classification. Further studies extending predictive features to site-specific conditions could improve the diagnosis with a site- and cultivar-specific nutrient diagnosis model.

Supporting information

S1 Table. Quebec potato leaves ionome data set. raw_leaf_df.csv file available online in data repository at https://git.io/Jvt2r.
(CSV)

S2 Table. Potato data set used for cluster analysis.
(DOCX)

S3 Table. True negatives mean clr values for cultivars.
(DOCX)

Author Contributions

Conceptualization: Serge-Étienne Parent.

Data curation: Zonlehoua Coulibali, Serge-Étienne Parent.

Formal analysis: Zonlehoua Coulibali, Serge-Étienne Parent.

Investigation: Zonlehoua Coulibali, Serge-Étienne Parent.

Methodology: Zonlehoua Coulibali, Serge-Étienne Parent.

Project administration: Serge-Étienne Parent.

Software: Zonlehoua Coulibali, Serge-Étienne Parent.

Supervision: Athyna Nancy Cambouris, Serge-Étienne Parent.

Validation: Zonlehoua Coulibali, Serge-Étienne Parent.

Visualization: Zonlehoua Coulibali, Serge-Étienne Parent.

Writing – original draft: Zonlehoua Coulibali, Serge-Étienne Parent.

Writing – review & editing: Zonlehoua Coulibali, Athyna Nancy Cambouris, Serge-Étienne Parent.

References

1. CFIA. Potato plants characteristics, maturity. Canadian Food Inspection Agency: Canadian Food Inspection Agency; 2015 [Available from: http://www.inspection.gc.ca/plants/potatoes/characteristics/eng/1326490397702/1326490477981#mature.

2. Eschemback V, Kawakami J, Melo PEd. Performance of modern and old, European and national potato cultivars in different environments. Horticultura Brasileira. 2017; 35:377–84.
3. Kawakami J, Iwama K, Jitsuyama Y, Zheng X. Effect of cultivar maturity period on the growth and yield of potato plants grown from microtubers and conventional seed tubers. American Journal of Potato Research. 2004; 81(5):327–33.
4. Söğüt T, Öztürk F. Effects of harvesting time on some yield and quality traits of different maturing potato cultivars. African Journal of Biotechnology. 2011; 10(38):7349–55.
5. Saric MR. Theoretical and practical approaches to the genetic specificity of mineral-nutrition of plants. Plant and Soil. 1983; 72(2–3):137–50.
6. Zebarth BJ, Tai HL, Luo SN, Millard P, De Koeyer D, Li XQ, et al. Differential gene expression as an indicator of nitrogen sufficiency in field-grown potato plants. Plant and Soil. 2011; 345(1–2):387–400.
7. Sattelmacher B, Klotz F, Marschner H. Influence of the nitrogen level on root growth and morphology of two potato varieties differing in nitrogen acquisition. Plant and soil. 1990; 123(2):131–7.
8. Lahner B, Gong JM, Mahmoudian M, Smith EL, Abid KB, Rogers EE, et al. Genomic scale profiling of nutrient and trace elements in Arabidopsis thaliana. Nature Biotechnology. 2003; 21(10):1215–21. https://doi.org/10.1038/nbt865 PMID: 12949535
9. Salt DE, Baxter I, Lahner B. Ionomics and the study of the plant ionome. Annual Review of Plant Biology. 2008; 59:709–33. https://doi.org/10.1146/annurev.arplant.59.032607.092942 PMID: 18251712
10. Hernandes A, Parent S-É, Veillette J-P, Parent P, Leblanc M, Roy G, et al. Compositional meta-analysis of the nutrient profile of potato cultivars. 2011.
11. White PJ, Broadley MR, Thompson JA, McNicol JW, Crawley MJ, Poulton PR, et al. Testing the distinctness of shoot ionomes of angiosperm families using the Rothamsted Park Grass Continuous Hay Experiment. New Phytologist. 2012; 196(1):101–9. https://doi.org/10.1111/j.1469-8137.2012.04228.x PMID: 22803633
12. Parent SE, Parent LE, Rozane DE, Natale W. Plant ionome diagnosis using sound balances: case study with mango (Mangifera Indica). Frontiers in plant science. 2013; 4:1–12. https://doi.org/10.3389/fpls.2013.00001
13. Parent SE, Parent LE, Egozcue JJ, Rozane DE, Hernandes A, Lapointe L, et al. The plant ionome revisited by the nutrient balance concept. Frontiers in Plant Science. 2013; 4.
14. Marschner H. Diagnosis of deficiency and toxicity of mineral nutrients. In: Marschner H, editor. Mineral Nutrition of Higher Plants. 2e ed: Academic Press, London; 1995. p. 461–79.
15. Jones JJB, Wolf B, Mills HA. Plant analysis handbook. A practical sampling, preparation, analysis, and interpretation guide: Micro-Macro Publishing, Inc.; 1991.
16. Aitchison J. The statistical analysis of compositional data. London: Chapman and Hall. 1986.
17. Dumenil LC. Relationship between the chemical composition of corn leaves and yield responses from nitrogen and phosphorus fertilizer Iowa State University Capstones; 1958.
18. McKenzie RH, Stewart JWB, Dormaar JF, Schaalje GB. Long-term crop rotation and fertilizer effects on phosphorus transformations: I. In a Chernozemic soil. Canadian Journal of Soil Science. 1992; 72(4):569–79.
19. McKenzie R. Crop nutrition and fertilizer requirements. Alberta Agriculture, Food and Rural Development Lethbridge. 1998:1–7.
20. Baxter I. Should we treat the ionome as a combination of individual elements, or should we be deriving novel combined traits? Journal of Experimental Botany. 2015; 66(8):2127–31. https://doi.org/10.1093/jxb/erv040 PMID: 25711709
21. Pawlowsky-Glahn V, Buccianti A. Compositional data analysis. Theory and applications: A John Wiley & Sons, Ltd, Publication; 2011. 378 p.
22. Tolosana-Delgado R, Van Den Boogart KG. Linear models with compositions in R. In: Pawlowsky-Glahn V, Buccianti A, editors. Compositional data analysis: Theory and applications: (New York: John Wiley and Sons); 2011. p. 356–71.
23. Parent LE, Dafir M. A theoretical concept of compositional nutrient diagnosis. Journal of the American Society for Horticultural Science. 1992; 117(2):239–42.
24. de Deus JAL, Neves JCL, Correa MCD, Parent SE, Natale W, Parent LE. Balance design for robust foliar nutrient diagnosis of "Prata" banana (Musa spp.). Scientific Reports. 2018; 8:1–7. https://doi.org/10.1038/s41598-017-17765-5
25. Nicolas O, Charles MT, Jennie S, Toussaint V, Parent SE, Beaulieu C. The ionomics of lettuce infected by Xanthomonas campestris pv. vitians. Frontiers in Plant Science. 2019; 10:1–10. https://doi.org/10.3389/fpls.2019.00001

26. Melo GW, Rozane DE, Brunetto G, Lattuada DS. Discriminant analysis in the selection of groups of peach cultivars. In: Mimmo T, Pii Y, Scandellari F, editors. Viii International Symposium on Mineral Nutrition of Fruit Crops. Acta Horticulturae. 12172018. p. 335–42.
27. Prater C, Scott DE, Lance SL, Nunziata SO, Sherman R, Tomczyk N, et al. Understanding variation in salamander ionomes: A nutrient balance approach. Freshwater Biology. 2019; 64(2):294–305.
28. Leite MLC, Prinelli F. A compositional data perspective on studying the associations between macronutrient balances and diseases. European Journal of Clinical Nutrition. 2017; 71(12):1365–9. https://doi.org/10.1038/ejcn.2017.126 PMID: 28853741
29. Leite MLC. Applying compositional data methodology to nutritional epidemiology. Statistical Methods in Medical Research. 2016; 25(6):3057–65. https://doi.org/10.1177/0962280214560047 PMID: 25411321
30. Westermann DT, Davis JR. Potato nutritional management changes and challenges into the next century. American Potato Journal. 1992; 69(11):753–67.
31. Mills HAJJ Walsh LMB, James D, CottenieE A, Faithfull NT, Larrahondo JE, et al. Plant analysis handbook II: a practical preparation, analysis, and interpretation guide: Potash and Phosphate Institute; 1996.
32. Hahsler M, Piekenbrock M, Arya S, Mount D. dbscan: Density based clustering of applications with noise (DBSCAN) and related algorithms. R package version 1.1–3. 2017.
33. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning-with applications in R. New York, NY: Springer; 2013.
34. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002; 2(3):18–22.
35. Breiman L. Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley, CA, USA. 2002; 1:58.
36. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988; 240(4857):1285–93. https://doi.org/10.1126/science.3287615 PMID: 3287615
37. Aitchison J, Ng KW. The role of perturbation in compositional data analysis. Statistical Modelling. 2005; 5(2):173–85.
38. Monna F, Marques AN, Guillon R, Losno R, Couette S, Navarro N, et al. Perturbation vectors to evaluate air quality using lichens and bromeliads: a Brazilian case study. Environmental Monitoring and Assessment. 2017; 189(11).
39. Egozcue JJ, Pawlowsky-Glahn V. Simplicial geometry for compositional data. In: Buccianti A, MateuFigueras GH, GlahnPawlowsky V, editors. Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society Special Publication. 2642006. p. 145–59.
40. Parent SE. Why we should use balances and machine learning to diagnose ionomes. Authorea [Internet]. 2020. Available from: https://www.authorea.com/users/23640/articles/281937-why-we-should-use-balances-and-machine-learning-to-diagnose-ionomes.
41. Hron K. Analytical representation of ellipses in the Aitchison geometry and its application. Acta Universitatis Palackianae Olomucensis Facultas Rerum Naturalium Mathematica. 2009; 48(1):53–60.
42. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2019.
43. Van den Boogaart KG, Raimon T, Bren M. compositions: compositional data analysis. R package version 1.40–1. 2014.
44. Filzmoser P, Hron K. "Robust statistical analysis. Chapter 5. In: Pawlowsky-Glahn V, Buccianti A, editors. Compositional Data Analysis: Theory and Applications: John Wiley and Sons, New York, NY; 2011. p. 59–72.
45. Filzmoser P, Gschwandtner M. mvoutlier: Multivariate Outlier Detection Based on Robust Methods. R package version 2.0.9. 2018.
46. Dray S, Dufour AB. The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software. 2007; 22(4):1–20.
47. Kuhn M, Wing J, Weston S, Williams A. Caret package: classification and regression training Journal of Statistical Software. 2008; 28(5):1–26.
48. Statistics Canada. Area, production and farm value of potatoes 2017 [Available from: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3210035801&pickMembers%5B0%5D=1.6&request_locale=en.
49. Parent LE, Cambouris AN, Muhawenimana A. Multivariate diagnosis of nutrient imbalance in potato crops. Soil Science Society of America Journal. 1994; 58(5):1432–8.
50. Zebarth BJ, Karemangingo C, Scott P, Savoie D, Moreau G. Nitrogen management for potato: general fertilizer recommendations. New-Brunswick Ministry of Agriculture, Fisheries and Aquaculture, Fredericton, NB, Canada. 2007.

51. Huang XY, Salt DE. Plant Ionomics: From Elemental Profiling to Environmental Adaptation. Molecular Plant. 2016; 9(6):787–97. https://doi.org/10.1016/j.molp.2016.05.003 PMID: 27212388

52. Hochmuth GJ, Maynard D, Vavrina C, Hanlon E, Simonne E. Plant tissue analysis and interpretation for vegetable crops in Florida. 2004. p. 1–48.

53. Cottenie A. Soil and plant testing as a basis of fertilizer recommendations. FAO Soils Bulletin. 1980; 38 (2):1–118.

54. White PJ, Wheatley RE, Hammond JP, Zhang K. Minerals, soils and roots. Potato Biology and Biotechnology: Elsevier; 2007. p. 739–52.

55. Giletto CM, Echeverría HE. Critical nitrogen dilution curve in processing potato cultivars. American Journal of Plant Sciences. 2015; 6(19):3144–56.

56. Natale W, Lima Neto AJ, Rozane DE, Parent L-É, Corrêa MCM. Mineral nutrition evolution in the formation of fruit tree rootstocks and seedlings. Revista Brasileira de Fruticultura. 2018; 40(6):(e-133).

57. Legendre P, Legendre L. Cluster analysis. In: Legendre P, Legendre L, editors. Developments in environmental modelling Numerical ecology. 24: Elsevier; 2012. p. 337–424.

58. Borcard D, Gillet F, Legendre P. Numerical ecology with R: Springer; 2018.

59. Andrews M, Raven JA, Lea PJ. Do plants need nitrate? The mechanisms by which nitrogen form affects plants. Annals of Applied Biology. 2013; 163(2):174–99.

60. Sikora FJ, Howe PS, Hill LE, Reid DC, Harover DE. Comparison of colorimetric and ICP determination of phosphorus in Mehlich3 soil extracts. Communications in Soil Science and Plant Analysis. 2005; 36 (7–8):875–87.

61. Ivanov K, Zaprjanova P, Angelova V, Bekjarov G, Dospatliev L, editors. ICP determination of phosphorous in soils and plants. 19th World Congress of Soil Science, Soil Solutions for a Changing World; 2010.

62. Adesanwo OO, Ige DV, Thibault L, Flaten D, Akinremi W. Comparison of Colorimetric and ICP Methods of Phosphorus Determination in Soil Extracts. Communications in Soil Science and Plant Analysis. 2013; 44(21):3061–75.

63. Valkama E, Uusitalo R, Ylivainio K, Virkajarvi P, Turtola E. Phosphorus fertilization: a meta-analysis of 80 years of research in Finland. Agriculture Ecosystems & Environment. 2009; 130(3–4):75–85.

64. Hollander M, Wolfe DA, Chicken E. Nonparametric statistical methods. Third ed. Hoboken, New Jersey: John Wiley & Sons, Inc; 2013. 837 p.

65. Stalham MA, Allen EJ, Herry FX. Effects of soil compaction on potato growth and its removal by cultivation. Research review. 2005(R261):1–60.

66. Boiteau G, Goyer C, Rees HW, Zebarth BJ. Differentiation of potato ecosystems on the basis of relationships among physical, chemical and biological soil parameters. Canadian Journal of Soil Science. 2014; 94(4):463–76.

67. Zebarth BJ, Leclerc Y, Moreau G, Botha E. Rate and timing of nitrogen fertilization of Russet Burbank potato: Yield and processing quality. Canadian Journal of Plant Science. 2004; 84(3):855–63.

68. Rich AE. Potato diseases. New York: Academic Press; 1983. xiv, 238 p p.

69. Herman DJ, Knowles LO, Knowles NR. Heat stress affects carbohydrate metabolism during cold-induced sweetening of potato (Solanum tuberosum L.). Planta. 2017; 245(3):563–82. https://doi.org/10.1007/s00425-016-2626-z PMID: 27904974

70. Parent SE, Parent LE, Rozane DE, Hernandes A, Natale W. Nutrient balance as paradigm of plant and soil chemometrics. Chapter 4. In: Issaka RN, editor. Soil Fertility: Tech Publ, NY; 2012. p. 83–114.

71. Kuhn M, Johnson K. Applied predictive modeling. New York, NY: Springer; 2013. Available from: http://dx.doi.org/10.1007/978-1-4614-6849-3.

72. Brownlee J. Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning. mistery Ml, editor2020. 463 p.

73. Campbell CR. Reference sufficiency ranges for plant analysis in the southern region of the United States. SAAESD, editor: SAAESD; 2000. 134 p.

74. Rozane DE, Mattos Junior Dd, Parent SE, Natale W, Parent LE, editors. Compositional meta-analysis of citrus varieties in the state of São Paulo, Brazil. 4th International Workshop on Compositional Data Analysis; 2011; Saint Feliu de Giuxols, Girona, Spain.

75. Rozane DE, Mattos D, Parent SE, Natale W, Parent LE. Meta-analysis in the selection of groups in varieties of citrus. Communications in Soil Science and Plant Analysis. 2015; 46(15):1948–59.