Why we should use balances and machine learning to diagnose ionomes

Serge-Étienne Parent[1]

[1]Laval University

May 22, 2020

Abstract

The performance of a plant can be predicted from its ionome (concentration of elements in a living tissue) at a specific growth stage. Diagnoses have yet been based on simple statistical tools by relating a Boolean index to a vector of nutrient concentrations or to unstructured sets of nutrient ratios. We are now aware that compositional data such as nutrient concentrations should be carefully preprocessed before statistical modeling. Projecting concentrations to isometric log-ratios confer a Euclidean space to compositional data, similar to geographic coordinates. By comparing projected nutrient profiles to a geographical map, this perspective paper shows why univariate ranges and ellipsoids are less accurate to assess the nutrient status of a plant from its ionome compared to machine learning models. I propose an imbalance index defined as the Aitchison distance between an imbalanced specimen to the closest balanced point or region in a reference data set. I also propose and raise some limitations of a recommendation system where the ionome of a specimen is translated to its closest point or region where high plant performance is reported. The approach is applied to a data set comprising macro- and oligo-elements measured in blueberry leaves from Québec, Canada.

1 Introduction

Concentrations of elements found in plant tissues (also called the ionome by Salt et al. (2008)) has been linked to crop yield for nearly two centuries. Such concept emerged from the *law of the minimum* and the law of the optimum (de Wit, 1992). In the 70s, E. R. Beaufils found that using nutrients alone was somewhat unstable (Walworth and Sumner, 1987), so he developed DRIS (diagnosis and recommendation integrated system) with an arguably naive mathematical framework, which was difficult to compute at that time. Using tools of compositional data analysis (Aitchison, 1986), Parent and Dafir (1992) used mathematically sound centered log-ratios and developed the CND, for compositional nutrition diagnosis.

Since about 2011, I have published several papers on data analysis in plant ionomics (listed on my ORCID profile). It began with the application of nutrient balances in leaves of several fruit crop species (Parent et al., 2013a). The approach was to free nutrient concentrations from their 0 to 100% constrained space using a transformation technique named the isometric log-ratio (Egozcue, 2003). Indeed, concentrations are compositional data, and not transforming this kind of data will likely return biased results (Aitchison, 1986). Recent works in biology confirmed such problems (Silverman et al., 2017; Mandal et al., 2015; Friedman and Alm, 2012; Morton et al., 2017; Rivera-Pinto et al., 2018; Jeanne et al., 2019).

Even when compositions are preprocessed with log-ratios, how a *healthy state* is defined has a fundamental importance on health diagnosis and on the correction measures to be prescribed. Plant health can be defined as its yield (e.g. Parent, 2013) or its resistance to a disease (Nicolas et al., 2019). And when the *healthy state* is defined, we must define its borders in the ionomic space. This task can be approached with machine learning

techniques, which can detect complex patterns better than any other statistical techniques I explored (e.g. in Parent, 2013). What's more: machine learning algorithms are nowadays quite easy to use.

Predicting the healthy state of a plant would be useless without an accurate recommendation system. The correction measures needed to recover health can be summarized by a translation from an imbalanced nutrient status to a balanced one. A translation from a compositional vector to another is called a *perturbation*. However, interpreting a perturbation is still a difficult cognitive task.

This article describes the logic behind the approach I recommend to diagnose plant ionomes. I show why we should not use confidence intervals and linear statistics, and how machine learning and perturbation of well documented Humboldtian agroecosystems can successfully address nutrient imbalance. I apply the proposed approach to a data set comprising macro- and oligo-elements measured in blueberry leaves from Québec, Canada.

2 From nutrient concentrations to nutrient balances

Transforming concentrations to balances between concentrations not only frees compositional data from their total sum constraint: it also offers a sound framework to interpret the ionome. To transform concentrations to nutrient balances, we need to apply a function to concentration data and, optionally, a balance scheme. As shown in Figure 1, the only rule to design a valid balance scheme is to **split groups of variables sequentially until each group contain a single part, like a strictly bifurcating tree**. So long as this rule is respected, the scheme can be anything you want, but balances can be organized so that they can be interpreted.

The isometric log-ratio (*ilr*) function computes a set of balances by applying the natural log of the geometric mean of concentrations on the right-hand-side of the balance scheme divided by the geometric mean of those on the left-hand-side, multiplied by a normalizing coefficient, i.e.

$$ilr_j = \sqrt{\frac{r_j s_j}{r_j + s_j}} ln \left(\frac{g(c_j^+)}{g(c_j^-)} \right),$$

where, for the j^{th} balance in $[1...D-1]$ where D is the lenght of the compositional vector, r_j is the number of parts on the left-hand-side, s_j is the number of parts on the right-hand-side, c_j^- is the compositional vector at the left-hand-side, c_j^+ is the compositional vector at the right-hand-side, and $g()$ is the geometric mean function. This operation can be performed using the `compositions` R package or the `scikit-bio` Python package.

At the end, so long as the tree is bifurcating, the design of the tree doesn't matter so much to run multivariate analyses since the isometric log-ratio transformation explodes data from a closed to a real orthonormal space (orthogonal axes with equal scales). Switching from one balance design to another only rotates from the origin the axes across the cloud of data points. Euclidean distances between vectors of balances remain the same no matter how you designed the tree.

In this article, I'm *de facto* ruling out approaches based on concentrations and unorganized ratios and from now on, I will use balances computed with isometric log-ratios as variables on which diagnoses are performed.

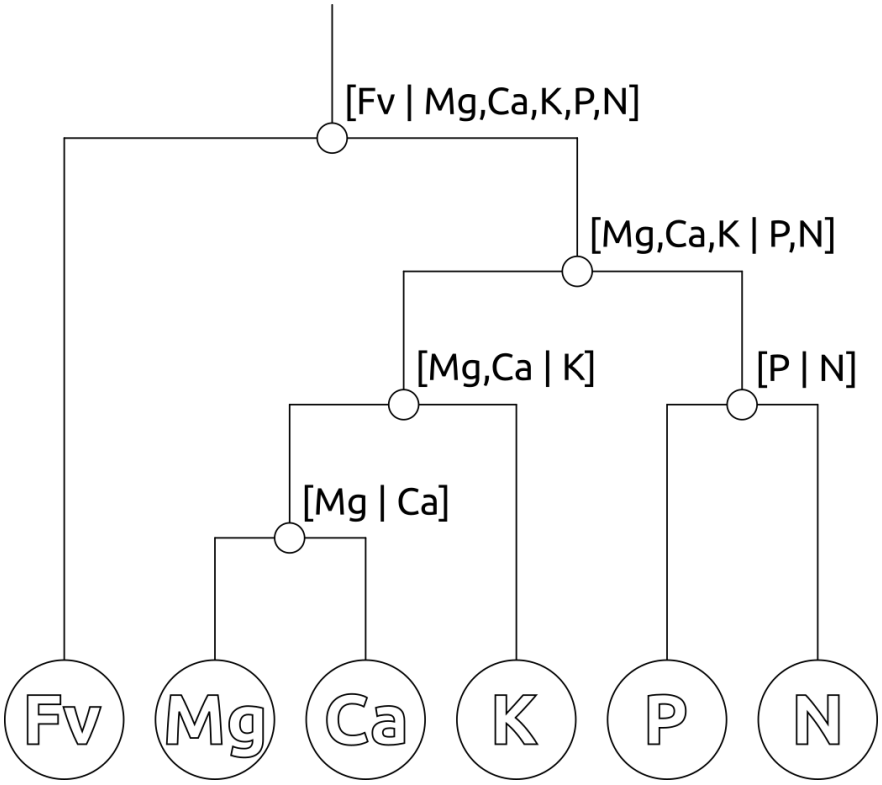

Figure 1: Strictly bifurcating balance tree of nutrients defining balance variables based on the valence of ions in the soil solution. Fv is the filling value, computed by subtracting the sum of all nutrients from the unit of measurement. Balances are the nodes in the scheme and concentrations are the weights at the bottom.

3 The ionome as a map

Once mapped with isometric log-ratios, ionomic data are *coordinates*(Pawlowsky-Glahn et al., 2015) and analogous to coordinates in geographical maps, like the archipelago of the Îles-de-la-Madeleine, Québec, Canada (Figure 2).

Now let's assume that yellow points (land) are *healthy* ionomes and that blue points (water) are *unhealthy*. Indeed, **there is no reason why the delimitation of a health status should return a regular shape**.

Of course, health is not binary. In particular, if we use yield as a health index, the problem could be approached as a regression rather than a classification. A regression would add some flexibility to the approach, because the threshold that triggers an action could be decided by the professional. Although they should not be ruled out in the first place, in my experience regressions have often returned inaccurate predictions.

I thus prefered to use a yield threshold: when a predicted yield-category is low, a red flag is raised and correction measures should be applied. Hence, the boolean approach with healthy/unhealthy becomes analogous to the land/water dichotomy.

Because I work in prediction mode, I randomly assigned the points to training (70%) or testing (30%) sets to respectively fit and evaluate our future models.

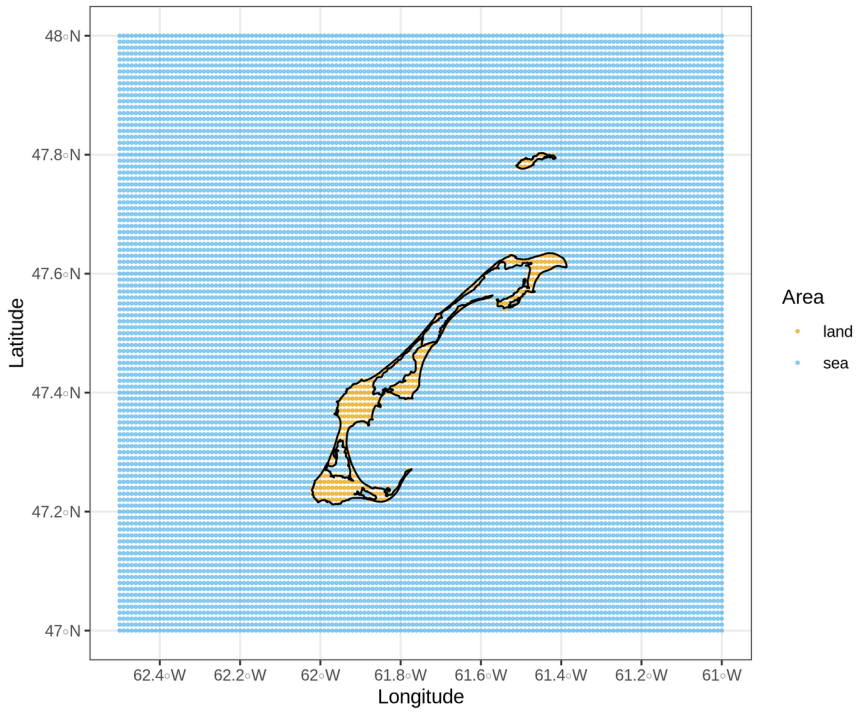

Figure 2: Grid of land and water points around the Îles-de-la-Madeleine archipelago, Québec, Canada.

4 Using univariate ranges

Ionomic ranges are commonly used to differentiate groups of organisms. They are usually reported in terms of concentrations (Neugebauer et al., 2018), unorganized ratios (Bittsánszky et al., 2016) or log-ratios (Prater et al., 2018), centered log-ratios (Barłóg, 2014) or isometric log-ratio balances (Parent et al., 2013b). Balance ranges to diagnose health (of land) with 95% confidence level on the training set would look like the shade region in figure 3.

We are lucky here that the mean (black dot) fall on a land area, but only 8% of the testing set enclosed in the shade region is water (negative predictive value): a pretty bad approximation if you were a parachutist expecting to fall on land! Worse, the map metaphor relies on only two dimensions. As you add dimensions (the tree in figure 1 has six balances), the proportion of the healthy hypershape in the hypercube will decrease considerably (Nowaki et al., 2017). Moreover, selecting an alternative balance scheme would rotate the axes and greatly affect the range-based diagnosis. Although results conform with the literature, they are biased and likely inaccurate.

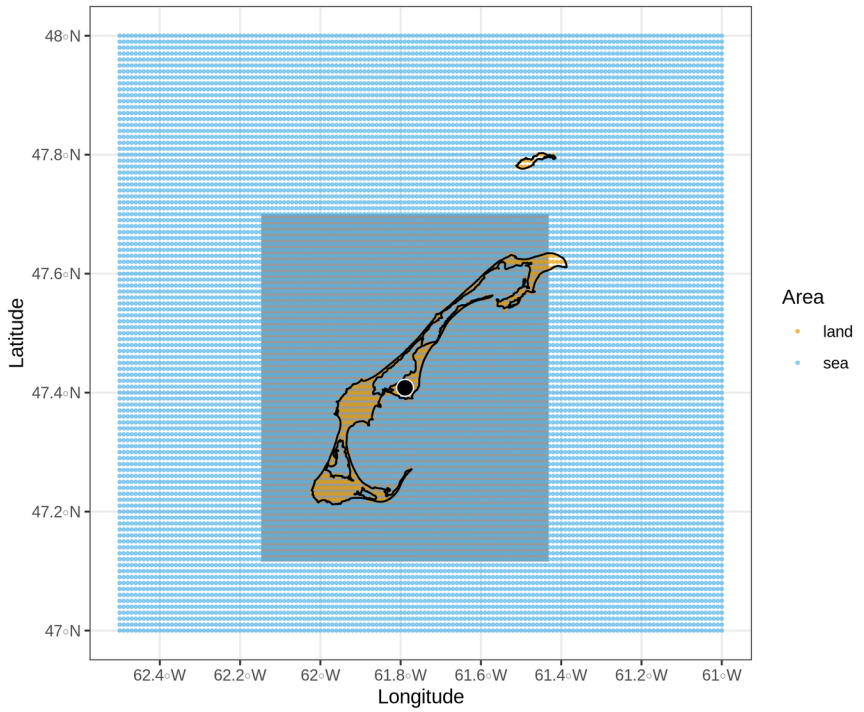

Figure 3: Land-water diagnosis with 95% univariate ranges

5 Using ellipsoids

In a paper on mango ionomics (Parent, 2013), I analyzed the ionome with an in-house classification technique based on the Mahalanobis distance to classify specimens between high and low crop yields. The algorithm iterated to isolate true high yielders (high yields correctly diagnosed as high yields) as reference. Let's see if the Mahalanobis distance is useful (shaded region in figure 4).

The ellipse is shaped by the covariance matrix and is independent from the balance scheme. The ellipse encloses only 15% of land points and completely miss the top island (*Rocher-aux-oiseaux*). Again, increasing the number of dimensions only makes the prediction worse.

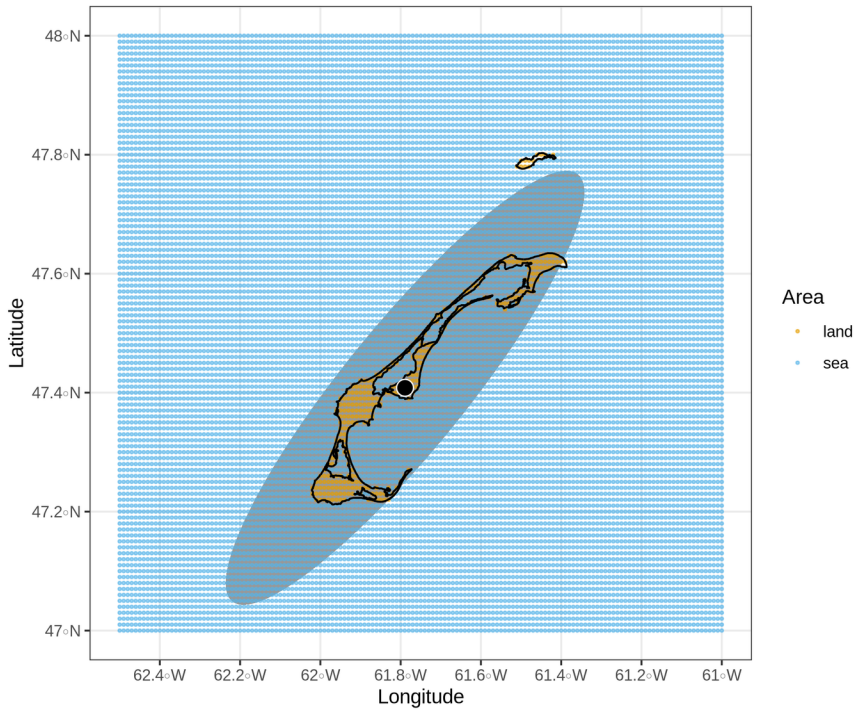

Figure 4: Land-water diagnosis with Mahalanobis distance

6 Using machine learning

Machine learning is a concept comprising hundreds of computing techniques that aim to detect patterns in data sets. Because, with the island example, we aim at predicting territory categories (just like we would have done with yield categories), we are looking for a supervised classification algorithm.

The *k-nearest neighbors* (KNN) algorithm looks for the k nearest training samples surrounding an observation whose outcome is to be predicted, then outputs the most likely outcome according to the ones of these k nearest training samples. These values can be categories (most common category of the k nearest points) or continuous values (the mean of the outcome of the k nearest points). A KNN model based on an Euclidean distance matrix is appropriate for isometric log-ratios, which are orthonormal. However, other machine learning algorithms could be used (Tolosana-Delgado et al., 2019).

With KNNs, I coarsely fitted a model predicting land/water loci according to Longitude and Latitude coordinates and their associated categories (land or water), with 3 neighbors. The probability of land closely overlaps the land perimeter in figure 5.

A relatively simple machine learning technique such as KNNs, could detect a complex pattern with a negative predictive value of 74%. But most importantly, this pattern is driven by the data, not by a prior geometry. If I were going for a parachute jump, I would prefer KNNs rather than ranges or ellipsoids!

In case that one diagnoses imbalance (or if the model predicts that the parachutist will fall in the sea), it

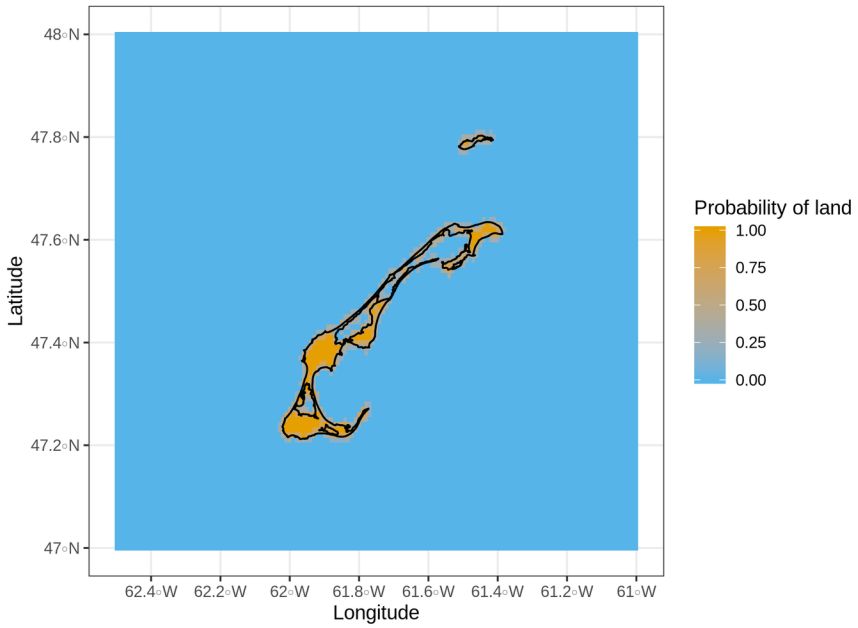

Figure 5: Land-water prediction using KNNs

would be useful to know how far the ionome is from the closest healthy island (or the closest path to swim towards the coast). I can compute a health index using the Euclidean distance from the closest healthy point (figure 6).

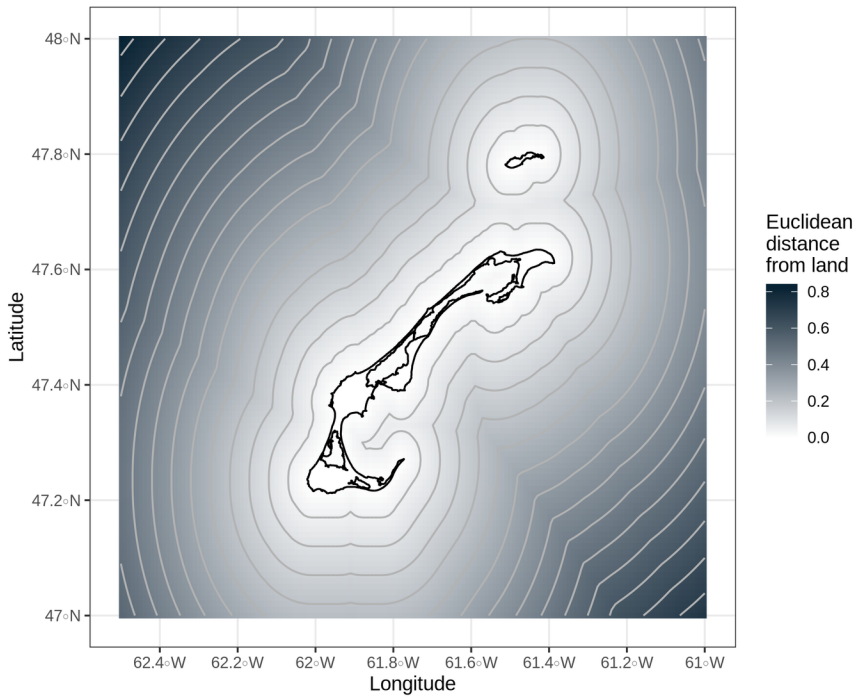

Figure 6: Diagnosis by distance from the land area

7 Recommendation system

Once the model predicts that a specimen indicates low yields due to nutrient imbalance, what should we do? By computing the imbalanced index, we spotted the closest balanced point. Moving from an imbalanced point to this closest balanced point (or the mean vector of the k nearest balanced points) is a translation operation described by a vector of deltas, one on each isometric log-ratio coordinate. In the compositional data jargon, such vector is known as a perturbation vector. The isometric log-ratio perturbation vector can be backtransformed to the concentration scale. The perturbation operator is applied to compositions as follows:

$$[y_1, y_2, ..., y_n] = [x_1, x_2, ..., x_n] \bigoplus [p_1, p_2, ..., p_n] = \mathcal{C}\left([x_1 \times p_1, x_2 \times p_2, ..., x_n \times p_n]\right)$$

where, with n components, the composition x is perturbed by the composition p returning the translated composition y.

A convenient way to compute the perturbation vector is to compute the difference between the coordinates of the reference point and the coordinates of the imbalanced one, then back-transform the balances to compositions according to the bifurcating tree. The inverse of isometric log-ratios could be computed in a spreadsheet, but I suggest the use of coding with your favorite language (most scientists use either R or Python - both the `compositions` R package and the `scikit-bio` Python package have an inverse isometric log-ratio function).

For example, if the observed ionome is $[0.04, 0.02, 0.015, 0.005, 0.008, 0.912]$ and I identify the target ionome as $[0.05, 0.02, 0.015, 0.01, 0.008, 0.897]$, the first and fourth components are increased at the expense of the last one. The perturbation from the observation to the target is $[0.1728, 0.1382, 0.1382, 0.2765, 0.1382, 0.1360]$. This perturbation vector contains all the information needed to move from the imbalance specimen to the target, but is difficult to interpret.

To overcome this difficulty, we proposed in (Parent et al., 2013b) to interpret the differences directly in a balance dendrogram. Still, a high degree of abstraction is needed to visualize a composition in a multidimensional mobile of parts in equilibrium. Some authors, like (de et al., 2018), proposed to diagnose ionomes using the centered log-ratio form. A centered log-ratio is a log ratio between a part and the geometric mean of the whole composition: interpreting a centered log-ratio is itself a cognitive challenge. We could imagine several ways to represent a perturbation vector in more insightful manners. The best option I have found to date is to ratio the actual components with the targeted ones.

To recall the example above, ratioing the target to the observation makes thing much clearer, i.e. $[0.8, 1, 1, 0.5, 1, 1.02]$: the first component of then imbalanced specimen is in shortage by one fifth, the fourth component is half as it should be, all at the expense of the last one. Bingo.

8 Case study with blueberry leaf data

While this paper addresses the whys and how-tos of ionomic diagnosis in agriculture, I will keep the methodology short. The Québec blueberry (*Vaccinium angustifolium*) data set comprises concentrations of nitrogen, phosphorus, potassium, calcium, magnesium, boron, copper, manganese, zinc, iron, molybdenum and aluminium in the diagnostic leaves, related to yield.

In this paper, I will use the R statistical language (R Core Team, 2019) to perform the necessary tasks. The GitLab repository offered as supplementary material provides all the necessary R codes.

I preprocessed the composition by computing the filling value by difference between the total sum and the sum of the quantified components, then computed the isometric log-ratios. I used a k-nearest neighbors algorithm based on Euclidean distances: the balance scheme does not affect the results.

After randomly splitting data between a training (70%) and a testing set (30%) and optimizing the KNN model using the custom tuning algorithms of the `caret` package (Kuhn, 2020), I obtained a model accuracy (proportion of correctly classified samples) of 77% on the testing set.

I created a fictive but plausible imbalanced specimen (table 1): its predicted yield category was a low-yielder (< 5000 kg/ga) with a probability of 73% (proportion of high yielders among the k nearest points). I decided to search for the 10 closest observations in the training set, from which I extracted the median: this strategy will prevent reaching a point too close to the edge of the hyper volume (which I use to name the hyperdblob) of high yielders. This target, found at an Aitchison distance (imbalance index, a value that should be contextualize for a given model) of 0.58, is also shown in table 1 with the associated observation / target ratio.

Table 1 can be appreciated with a collection of slope plots, such as current and target concentrations in figure 7 and ratios in figure 8.

Figure 8 shows that Al, Cu and Mn are apparently in excess, but that in K, Fe and P are apparently deficient. **Importantly**, this interpretation should be done with a multivariate and compositional data perspective in mind. This implies that (1) a univariate or an incomplete multivariate perspective (e.g. focusing on extreme excesses and deficiencies) could miss a high yield region (a parachutist adjusting her fall following only one axis will likely miss the island) and (2) changes of concentrations in a closed system are relative, i.e.

Nutrient	Current concentration (%)	Target concentration (%)	Current / Target ratio
N	1.5	1.7	0.91
P	0.11	0.13	0.86
K	0.42	0.59	0.71
Ca	0.41	0.47	0.88
Mg	0.18	0.18	1
Al	0.08	0.05	1.6
B	0.04	0.04	0.97
Cu	0.01	0.0	1.3
Zn	0.01	0.01	1.2
Mn	0.26	0.21	1.3
Fe	0.01	0.02	0.76
Mo	1.5e-4	1.5e-4	0.99
Fv	97	97	1

Table 1: Fictive specimen whose ionome should be corrected

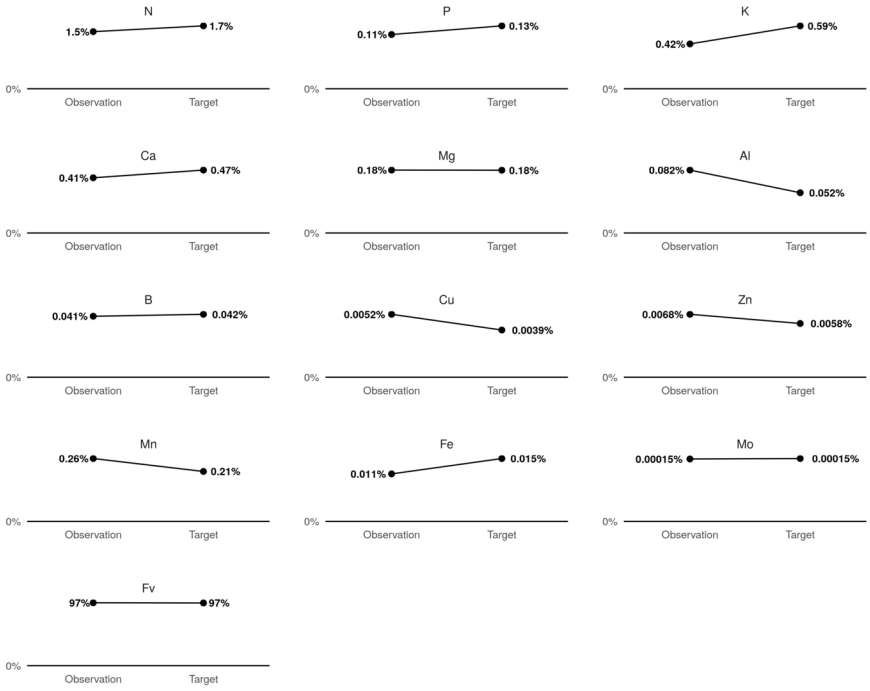

Figure 7: Nutrient concentrations in the observed and in the target samples

increasing the concentration of a component will inevitably decrease the concentration of at least another one.

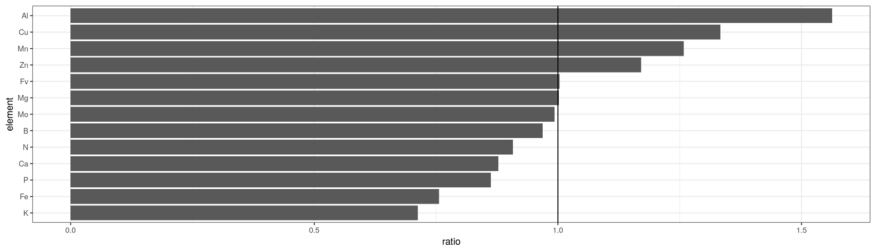

Figure 8: Ratio of the composition observation on the reference

9 Conclusion

All in all, the approach is fairly simple:

1. explode concentrations to balances of concentrations with the isometric log-ratio transformation,
2. identify a healthy state,
3. fit a predictive model the identify the healthy hyper-volume,
4. run the model to predict the yield class of a new ionome,
5. if unhealthy, identify the targeted ionome and
6. express the needed perturbation as a concentration ratios.

A perturbation indicates the change needed, not how to achieve this change, which could be done by identifying what happened in targetted balanced cases or by fertilizer trials. It should be noted that the closest healthy point might not be the easiest to reach, and professional agronomist must adapt results to local conditions and may select a solution not appearing at the closest point. For example, if the concentration of aluminium is too high, this could be corrected by increasing soil pH. But if increasing pH is not a viable solution, the agronomist could look for a high yield zone in the hyper-volume where Al is still high but less damageable due to other interactions.

Now you might think "*hey, this could work for species X too!*" Indeed, there is no reason why it won't.

All computations done for this article can be found in R Jupyter notebooks on GitLab.

References

J. Aitchison. Compositional data: some challenging problems. In *The Statistical Analysis of Compositional Data*, pages 1–23. Springer Netherlands, 1986. doi: 10.1007/978-94-009-4109-0_1. URL https://doi.org/10.1007%2F978-94-009-4109-0_1.

Przemysław Barłóg. Diagnosis of sugar beet (Beta vulgarisL.) nutrient imbalance by DRIS and CND-clr methods at two stages during early growth. *Journal of Plant Nutrition*, 39(1):1–16, sep 2014. doi: 10.1080/01904167.2014.964366. URL https://doi.org/10.1080%2F01904167.2014.964366.

András Bittsánszky, Nikolett Uzinger, Gábor Gyulai, Alex Mathis, Ranka Junge, Morris Villarroel, Benzion Kotzen, and Tamás Kőmíves. Nutrient supply of plants in aquaponic systems. *Ecocycles*, 2(2), 2016. doi: 10.19040/ecocycles.v2i2.57. URL https://doi.org/10.19040%2Fecocycles.v2i2.57.

Deus JAL de, JCL Neves, MCM Corrêa, SÉ Parent, W Natale, and LE Parent. Balance design for robust foliar nutrient diagnosis of Prata banana (Musa spp.). *Sci Rep*, 8:15040, Oct 2018.

C.T. de Wit. Resource use efficiency in agriculture. *Agricultural Systems*, 40(1-3):125–151, jan 1992. doi: 10.1016/0308-521x(92)90018-j. URL https://doi.org/10.1016%2F0308-521x%2892%2990018-j.

J. J. Egozcue. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35(3):279–300, 2003. doi: 10.1023/a:1023818214614. URL https://doi.org/10.1023%2Fa%3A1023818214614.

Jonathan Friedman and Eric J. Alm. Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology*, 8(9):e1002687, sep 2012. doi: 10.1371/journal.pcbi.1002687. URL https://doi.org/10.1371%2Fjournal.pcbi.1002687.

Thomas Jeanne, Serge-Étienne Parent, and Richard Hogue. Using a soil bacterial species balance index to estimate potato crop productivity. *PLOS ONE*, 14(3):e0214089, mar 2019. doi: 10.1371/journal.pone.0214089. URL https://doi.org/10.1371%2Fjournal.pone.0214089.

Max Kuhn. *caret: Classification and Regression Training*, 2020. URL https://CRAN.R-project.org/package=caret. R package version 6.0-85.

Siddhartha Mandal, Will Van Treuren, Richard A. White, Merete Eggesbø, Rob Knight, and Shyamal D. Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health & Disease*, 26(0), may 2015. doi: 10.3402/mehd.v26.27663. URL https://doi.org/10.3402%2Fmehd.v26.27663.

James T. Morton, Jon Sanders, Robert A. Quinn, Daniel McDonald, Antonio Gonzalez, Yoshiki Vázquez-Baeza, Jose A. Navas-Molina, Se Jin Song, Jessica L. Metcalf, Embriette R. Hyde, Manuel Lladser, Pieter C. Dorrestein, and Rob Knight. Balance Trees Reveal Microbial Niche Differentiation. *mSystems*, 2 (1), jan 2017. doi: 10.1128/msystems.00162-16. URL https://doi.org/10.1128%2Fmsystems.00162-16.

K Neugebauer, MR Broadley, HA El-Serehy, TS George, JW McNicol, MF Moraes, and PJ White. Variation in the angiosperm ionome. *Physiol Plant*, Feb 2018.

O Nicolas, MT Charles, S Jenni, V Toussaint, SÉ Parent, and C Beaulieu. The Ionomics of Lettuce Infected by ¡i¿Xanthomonas campestris¡/i¿ pv. ¡i¿vitians¡/i¿. *Front Plant Sci*, 10:351, 2019.

Rodrigo H. D. Nowaki, Serge-Étienne Parent, Arthur B. Cecílio Filho, Danilo E. Rozane, Natália B. Meneses, Juliana A. dos Santos da Silva, William Natale, and Léon E. Parent. Phosphorus Over-Fertilization and Nutrient Misbalance of Irrigated Tomato Crops in Brazil. *Frontiers in Plant Science*, 8, may 2017. doi: 10.3389/fpls.2017.00825. URL https://doi.org/10.3389%2Ffpls.2017.00825.

Léon Etienne Parent and M. Dafir. A Theoretical Concept of Compositional Nutrient Diagnosis. *Journal of the American Society of Horticultural Science*, 117, 1992.

Serge-Étienne Parent. Plant ionome diagnosis using sound balances: case study with mango (Mangifera Indica). *Frontiers in Plant Science*, 4, 2013. doi: 10.3389/fpls.2013.00449. URL https://doi.org/10.3389%2Ffpls.2013.00449.

Serge-Étienne Parent, Léon Etienne Parent, Juan José Egozcue, Danilo-Eduardo Rozane, Amanda Hernandes, Line Lapointe, Valérie Hébert-Gentile, Kristine Naess, Sébastien Marchand, Jean Lafond, Dirceu Mattos, Philip Barlow, and William Natale. The Plant Ionome Revisited by the Nutrient Balance Concept. *Frontiers in Plant Science*, 4, 2013a. doi: 10.3389/fpls.2013.00039. URL https://doi.org/10.3389%2Ffpls.2013.00039.

SÉ Parent, LE Parent, DE Rozane, and W Natale. Plant ionome diagnosis using sound balances: case study with mango (Mangifera Indica). *Front Plant Sci*, 4:449, 2013b.

Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. Coordinate representation. In *Modelling and Analysis of Compositional Data*, pages 32–64. John Wiley & Sons Ltd, feb 2015. doi: 10.1002/9781119003144.ch4. URL https://doi.org/10.1002%2F9781119003144.ch4.

Clay Prater, David E. Scott, Stacey L. Lance, Schyler O. Nunziata, Ryan Sherman, Nathan Tomczyk, Krista A. Capps, and Punidan D. Jeyasingh. Understanding variation in salamander ionomes: A nutrient

balance approach. *Freshwater Biology*, dec 2018. doi: 10.1111/fwb.13216. URL https://doi.org/10.1111%2Ffwb.13216.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

J. Rivera-Pinto, J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle. Balances: a New Perspective for Microbiome Analysis. *mSystems*, 3(4), jul 2018. doi: 10.1128/msystems.00053-18. URL https://doi.org/10.1128%2Fmsystems.00053-18.

David E. Salt, Ivan Baxter, and Brett Lahner. Ionomics and the Study of the Plant Ionome. *Annual Review of Plant Biology*, 59(1):709–733, jun 2008. doi: 10.1146/annurev.arplant.59.032607.092942. URL https://doi.org/10.1146%2Fannurev.arplant.59.032607.092942.

Justin D Silverman, Alex D Washburne, Sayan Mukherjee, and Lawrence A David. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, 6, feb 2017. doi: 10.7554/elife.21887. URL https://doi.org/10.7554%2Felife.21887.

Raimon Tolosana-Delgado, H. Talebi, M. Khodadadzadeh, and K. G. van den Boogaart. On machine learning algorithms and compositional data. 2019. ISBN 978-84-947240-2-2.

J. L. Walworth and M. E. Sumner. The Diagnosis and Recommendation Integrated System (DRIS). In *Advances in Soil Science*, pages 149–188. Springer New York, 1987. doi: 10.1007/978-1-4612-4682-4_4. URL https://doi.org/10.1007%2F978-1-4612-4682-4_4.