UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL

NUTRIENT DIAGNOSIS OF ORANGE CROPS APPLYING COMPOSITIONAL DATA ANALYSIS AND MACHINE LEARNING TECHNIQUES

Danilo Ricardo Yamane
Engenheiro Agrônomo

2018

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

NUTRIENT DIAGNOSIS OF ORANGE CROPS APPLYING COMPOSITIONAL DATA ANALYSIS AND MACHINE LEARNING TECHNIQUES

Danilo Ricardo Yamane

Orientador: Prof. Dr. Arthur Bernardes Cecílio Filho

Coorientador: Prof. Dr. Léon-Etienne Parent

Coorientador: Prof. Dr. Danilo Eduardo Rozane

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutor em Agronomia (Produção Vegetal)

2018

Y19n

Yamane, Danilo Ricardo

Nutrient Diagnosis of orange crops appyling compositional data analysis and machine learning techniques / Danilo Ricardo Yamane. -- Jaboticabal, 2018

65 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal

Orientador: Arthur Bernardes Cecílio Filho

Coorientador: Léon-Etienne Parent

1. centred log ratio. 2. citrus nutrition. 3. isometric log ratio. 4. k-NN. 5. multivariate analysis. I. Título.

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: NUTRIENT DIAGNOSIS OF ORANGE CROPS APPLYING COMPOSITIONAL DATA ANALYSIS AND MACHINE LEARNING TECHNIQUES

AUTOR: DANILO RICARDO YAMANE
ORIENTADOR: ARTHUR BERNARDES CECILIO FILHO
COORIENTADOR: LEON-ÉTIENNE PARENT
COORIENTADOR: DANILO EDUARDO ROZANE

Aprovado como parte das exigências para obtenção do Título de Doutor em AGRONOMIA (PRODUÇÃO VEGETAL), pela Comissão Examinadora:

Prof. Dr. ARTHUR BERNARDES CECILIO FILHO
Departamento de Produção Vegetal / FCAV / UNESP - Jaboticabal

Prof. Dr. RENATO DE MELLO PRADO
Departamento de Solos e Adubos / FCAV / UNESP - Jaboticabal

Pesquisadora Dra. AMANDA HERNANDES
Coordenadoria de Assistência Técnica Integral / CATI - Campinas/SP

Pesquisador Dr. LUIZ ANTONIO JUNQUEIRA TEIXEIRA
Centro de Solos e Recursos Ambientais / Instituto Agronômico de Campinas / Campinas,SP

Dr. ROBERTO BOTELHO FERRAZ BRANCO
Manejo e Tratos Culturais, Horticultura, Ecofisiologia e Nutrição de Plantas Cultivadas / APTA - Polo Regional Centro Leste - Ribeirão Preto

Jaboticabal, 29 de novembro de 2018

DADOS CURRICULARES DO AUTOR

Danilo Ricardo Yamane – nascido no dia 3 de dezembro de 1986, na cidade de Bebedouro, Estado de São Paulo, Brasil, filho de Ricardo Iudi Yamane e Sônia Regina Tomiosso Yamane. Cursou o Ensino Fundamental e o primeiro e segundo ano do Ensino Médio na Escola Espaço Livre, em Bebedouro, SP. Concluiu o ensino médio no Liceu Albert Sabin, em Ribeirão Preto, SP, no ano de 2004. Em fevereiro de 2005, ingressou no Curso de Engenharia Agronômica na Universidade de São Paulo - Escola Superior de Agricultura "Luiz de Queiroz" (USP/ESALQ) – Câmpus de Piracicaba, SP. Durante a graduação, fez estágio no Laboratório de Epidemiologia no Departamento de Fitopatologia e Nematologia (USP/ESALQ), sob orientação da Profa. Dra. Lilian Amorim; e no Centro de Citricultura Sylvio Moreira (IAC), sendo bolsista de iniciação científica pelo programa PIBIC/CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), sob orientação do Prof. Dr. Dirceu de Mattos Junior. Realizou residência agronômica na University of Florida (Citrus Research and Education Center), nos Estados Unidos, sob orientação do Prof. Dr. James Syvertsen. Em agosto de 2010, obteve o título de Engenheiro Agrônomo. Iniciou em março de 2011, o Curso de Mestrado em Agronomia (Produção Vegetal) na Universidade Estadual Paulista – Faculdade de Ciências Agrárias e Veterinárias (UNESP/FCAV) – Câmpus de Jaboticabal, sob orientação do Prof. Dr. José Eduardo Corá, sendo bolsista da FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo). No dia 13 de junho de 2013, submeteu-se à banca para a defesa da Dissertação e obteve o título de Mestre. Em março de 2015, ingressou no Programa de Pós-graduação em Agronomia (Produção Vegetal), a nível de Doutorado, na Universidade Estadual Paulista – Faculdade de Ciências Agrárias e Veterinárias (UNESP/FCAV) – Câmpus de Jaboticabal, com bolsa da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), sob orientação do Prof. Dr. Arthur Bernardes Cecílio Filho. Realizou Doutorado Sanduíche no Exterior, na Université Laval, Canadá, com bolsa PDSE (Programa de Doutorado Sanduíche no Exterior) da CAPES, sob orientação do Prof. Dr. Léon-Etienne Parent e Prof. Dr. Serge-Étienne Parent.

“Lembremo-nos de que o homem interior se renova sempre. A luta enriquece-o de experiência, a dor aprimora-lhe as emoções e o sacrifício tempera-lhe o caráter”

Chico Xavier

Aos meus pais,
Ricardo Iudi Yamane e Sônia Regina Tomiosso Yamane,
à minha irmã,
Daniela Fernanda Yamane,
aos meus avós,
Takeshi Yamane e Maria Arikawa Yamane,
à minha namorada,
Karoline Pereira dos Reis,
e aos Profs. Drs. William Natale e Arthur Bernardes Cecílio Filho
pelo apoio, inspiração e carinho

Dedico.

AGRADECIMENTOS

A **Deus e a Nossa Senhora Aparecida,** por todas as oportunidades a mim proporcionadas, por ter colocado pessoas “fantásticas” no meu caminho, e por me conceder saúde, motivação e disposição para concluir meus objetivos com êxito durante o curso de Doutorado.

Ao **Prof. Dr. William Natale**, por ter me recebido na UNESP de braços abertos, concedendo-me a oportunidade da realização do presente trabalho. Agradeço pelos ensinamentos, conselhos, confiança e relação de amizade, contribuindo de forma valiosa para a minha formação pessoal e profissional. Muito obrigado por ser um exemplo de humildade, competência, caráter e dedicação a me inspirar ao longo da minha caminhada. Parabéns por priorizar o atendimento das demandas dos produtores agrícolas no campo e pela sua eximia dedicação a formação de recursos humanos de alto nível para a sociedade. O desenvolvimento e a sustentabilidade da agricultura, em especial da Fruticultura, deve muito aos “frutos” colhidos decorrente dos trabalhos e estudos conduzidos pelo senhor e sua equipe!

Ao meu orientador **Prof. Dr. Arthur Bernardes Cecílio Filho**, pela atenção e importantes ensinamentos em relação ao desenvolvimento da minha carreira, pela confiança e por ter sempre me apoiado ao longo desses anos, viabilizando a concretização de grandes sonhos e conquistas durante o curso de Doutorado, que muito contribuíram para o meu aprimoramento pessoal e profissional.

Ao meu coorientador **Prof. Dr. Léon-Etienne Parent** e ao seu filho **Prof. Dr. Serge-Étienne Parent** por terem me recebido na Université Laval (Canadá) durante o Programa de Doutorado Sanduíche no Exterior. Obrigado por terem compartilhado comigo seus conhecimentos e inspiração contagiante pelo desenvolvimento de ciência de alta qualidade. Agradeço ainda por todo o suporte que me concederam, possibilitando que eu fizesse desse período de vivência no Canadá, um dos mais enriquecedores e incríveis de toda a minha existência.

Ao meu coorientador **Prof. Dr. Danilo Eduardo Rozane**, pela atenção, ensinamentos e apoio durante o desenvolvimento do presente trabalho, os quais foram essenciais diante dos grandes desafios a serem superados.

Ao grande amigo **Dr. Rodrigo Hiyoshi Dalmazzo Nowaki**, pelo companheirismo e parceria. Obrigado por todo o aprendizado decorrente de nossas inúmeras conversas e discussões de ideias. Com a sua presença e apoio a caminhada foi bem mais enriquecedora e gratificante.

À **dona Marisa Natale**, a quem tive o privilégio de conhecer e conviver. Obrigado por torcer por mim! Obrigado por todo o seu carinho, apoio, atenção e confiança. Cultivamos ao longo desses anos uma amizade valiosa com um fluxo de energia muito positivo. Meus sinceros agradecimentos pelas palavras inspiradoras e "energizantes", que muito contribuíram para que eu aproveitasse da melhor forma possível as oportunidades de aprendizado e crescimento que tive durante o processo de doutoramento. Um dos maiores valores do ser humano acredito que seja a predisposição em ajudar e querer o bem ao próximo, e foi exatamente o que a senhora fez por mim!

À **Universidade Estadual Paulista – UNESP (Câmpus de Jaboticabal),** em especial ao **Programa de Pós-Graduação em Agronomia (Produção Vegetal)** pela excelente estrutura e corpo docente qualificado oferecido para a realização do curso de Doutorado.

Aos **docentes da UNESP/FCAV**, que muito contribuíram para a minha formação pessoal e profissional.

À **Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES),** pela concessão da bolsa de Doutorado no país, e bolsa de Doutorado Sanduíche no Exterior (Processo número: 88881.133948/2016-01). O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Aos meus avós, **Takeshi Yamane** e **Maria Arikawa Yamane**, que vieram do Japão para o Brasil na esperança de ter uma vida melhor, e neste país escreveram uma linda história de dedicação à agricultura, evidenciando que a persistência é o caminho para o êxito. Obrigado por todo o apoio e amor, vocês foram e sempre serão a minha principal inspiração em construir uma carreira ligada a agricultura. Gratidão eterna!

Ao meu **pai (Ricardo Iudi Yamane)**, minha **mãe (Sônia Regina Tomiosso Yamane)** e minha **irmã (Daniela Fernanda Yamane),** pelo carinho, amor e apoio durante o curso de Doutorado, viabilizando a concretização dessa conquista. Obrigado em especial aos meus pais, por sempre terem investido em minha Educação. As experiências mais enriquecedoras no âmbito social, cultural, intelectual, científico e espiritual em minha vida foram possíveis graças à Educação de qualidade que tive acesso durante toda a minha formação escolar e acadêmica, que me abriu portas para um mundo completamente novo e fascinante.

À **Karoline Pereira dos Reis,** por estar ao meu lado em todos os momentos. Obrigado pela compreensão, apoio e paciência durante a minha caminhada. Obrigado ainda por me inspirar a alcançar meus objetivos, e por compartilhar comigo sonhos e planos de um futuro de plena felicidade, paz e amor.

Ao meu amigo **Marcelo Donizeti Brigati**, que por meio do seu exemplo de fé, dedicação e superação diante dos desafios que a vida lhe impôs, tem me inspirado continuamente a lutar pelos meus sonhos e ideais, nunca desistindo.

Ao amigo **Réza Jamali** pela sua amizade e companheirismo durante a minha estadia na Université Laval. Muito obrigado pelo conhecimento que me proporcionou com nossas trocas de ideias e experiências, e pelos bons momentos de descontração e diversão. Com você muito aprendi sobre a vida, e o sacrifício da "luta" em busca de um futuro melhor.

Ao meu tio **Roberto Yochio Yamane** e à minha prima, **Fabiana Scaramal Yamane**, que conduziram a administração das propriedades agrícolas da minha família, viabilizando que eu me dedicasse ao curso de Doutorado.

À **Natalia Barreto Meneses**, pela amizade e discussões com o nosso grupo de pesquisa, que contribuíram para o desenvolvimento do presente trabalho.

Aos **Profs. Drs. Renato de Mello Prado, Luis Antônio Junqueira Teixeira, Roberto Botelho Ferraz Branco** e à **Dra. Amanda Hernandes**, pela participação na minha banca de defesa da Tese de Doutorado, e pelas valiosas sugestões apresentadas para a melhoria do presente trabalho.

Ao **Prof. Dr. Renato de Mello Prado** e **Profa. Dra. Gilmara Pereira da Silva**, pela participação no Exame Geral de Qualificação e contribuição no presente trabalho, e pelos importantes ensinamentos durante o curso.

Ao coordenador do Programa de Pós-Graduação em Agronomia (Produção Vegetal), **Prof. Dr. Rouverson Pereira da Silva**, pela disponibilidade e auxílio nos momentos em que necessitei.

Aos **funcionários da biblioteca** e da **seção de Pós-Graduação,** pelo excelente atendimento.

A todos os demais **amigos,** que de alguma forma contribuíram para a realização do presente trabalho.

Muito Obrigado!!

SUMMARY

Page

NUTRIENT DIAGNOSIS OF ORANGE CROPS APPLYING COMPOSITIONAL DATA ANALYSIS AND MACHINE LEARNING TECHNIQUES

ABSTRACT – Efficient nutrient management is crucial to attain high fruit productivity. Results of tissue analysis are commonly interpreted using critical nutrient concentration ranges (CNCR) and Diagnosis and Recommendation Integrated System (DRIS) on orange crops. Nevertheless, both methods ignore the inherent properties of compositional data class, not accounting adequately for nutrient interactions and varietal influence on plant ionome. Therefore, effective modeling tools are needed to rectify biases and incorporate genetic effects on nutrient composition. The objective of this study was to develop an accurate diagnostic approach to evaluate the nutritional status across orange (*Citrus sinensis*) canopy varieties using compositional data analysis and machine learning algorithms. We collected 716 foliar samples from fruit-bearing shoots in plots of non-irrigated commercial orange orchards (“Valencia”, “Hamlin”, “Pera”, “Natal”, “Valencia Americana” and “Westin”) distributed across São Paulo state (Brazil), analyzed N, S, P, K, Ca, Mg, B, Cu, Zn, Mn and Fe, and measured fruit yields. Sound nutrient balances were computed as isometric log-ratios (*ilr*). Discriminant analysis of *ilr* values differentiated the nutrient profiles of canopy varieties, indicating plant-specific ionomes. Diagnostic accuracy of nutrient balances reached 88% about cutoff yield of 60 Mg ha^{-1} using *ilrs* and a k-nearest neighbors classification, allowing the development of reliable nutritional standards at high fruit yield level. Citrus growers from São Paulo state should adopt the concept of yield-limiting nutrient balances, where groups of nutrients are optimally balanced. Supplying more Ca as lime or gypsum materials, reducing the P and K fertilizer applications and enhancing soil B fertilization could re-establish the [Mg | Ca], [Ca, Mg | K], [P | N, S], [K, Ca, Mg | N, S, P] and [B | N, S, P, K, Ca, Mg] balances in orange orchards yielding less than 60 Mg ha^{-1}. The software “CND-Citros” can assist citrus growers, agronomy engineers and technicians to diagnose the nutrient status of orange crops based on the proposed method, using the results of leaf chemical analysis.

Keywords: centred log ratio, citrus nutrition, isometric log ratio, k-NN, multivariate analysis

DIAGNÓSTICO DE NUTRIENTES NA CULTURA DA LARANJEIRA APLICANDO ANÁLISE COMPOSICIONAL DOS DADOS E TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

RESUMO – O manejo eficiente de nutrientes é crucial para atingir alta produtividade de frutos. Resultados da análise do tecido são comumente interpretados usando faixas críticas de concentração de nutrientes (CNCR) e Sistema Integrado de Diagnose e Recomendação (DRIS) em culturas de laranja. No entanto, ambos os métodos ignoram as propriedades inerentes à classe dos dados composicionais, não considerando adequadamente as interações de nutrientes e a influência varietal na composição nutricional da planta. Portanto, ferramentas eficazes de modelagem são necessárias para corrigir vieses e incorporar efeitos genéticos na avaliação do estado nutricional. O objetivo deste estudo foi desenvolver uma abordagem diagnóstica precisa para avaliar o estado nutricional de variedades de copa de laranjeira (*Citrus sinensis*), usando a análise composicional dos dados e algoritmos de inteligência artificial. Foram coletadas 716 amostras foliares de ramos frutíferos em pomares comerciais de laranjeiras não irrigadas ("Valência", "Hamlin", "Pera", "Natal", "Valencia Americana" e "Westin") distribuídos pelo estado de São Paulo (Brasil), analisadas as concentrações de N, S, P, K, Ca, Mg, B, Cu, Zn, Mn e Fe, e avaliadas as produções de frutos. Balanços de nutrientes foram computados como relações-log isométricas (ilr). Análises discriminantes dos valores de ilr diferenciaram os perfis de nutrientes das variedades de copa, indicando composições nutricionais específicas. A acurácia diagnóstica dos balanços de nutrientes atingiu 88% com a produtividade de corte correspondente a 60 t ha^{-1}, utilizando-se ilrs e o algoritmo de classificação knn, o que possibilitou o desenvolvimento de padrões nutricionais confiáveis para a obtenção de elevado nível de produtividade de frutos. Os citricultores do estado de São Paulo devem adotar o conceito de balanços de nutrientes, onde grupos de nutrientes estão equilibrados de maneira ideal. Fornecer mais Ca através de calcário ou gesso, reduzir as aplicações de fertilizantes P e K, e aumentar a fertilização de B via solo pode reequilibrar os balanços [Mg | Ca], [Ca, Mg | K], [P | N, S], [K, Ca, Mg | N, S, P] e [B | N, S, P, K, Ca, Mg] em pomares de laranjas com produtividade inferior a 60 t ha^{-1}. O software "CND-Citros" pode auxiliar os citricultores, engenheiros agrônomos e técnicos a diagnosticar o estado nutricional das lavouras de laranja com base no método proposto, utilizando os resultados da análise química das folhas

Palavras-chave: relação logarítmica centrada, nutrição citrus, razão log isométrica, k-NN, análise multivariada

1. INTRODUCTION

Brazil is the world leader in orange (*Citrus sinensis*) production with a total area grown estimated at 682.167 ha (Agrianual, 2018). The state of São Paulo is the main orange producing region and holds approximately 63% of the national harvested area (Agrianual, 2018). However, the Brazilian citrus industry faces major challenges to maintain this prominent position, mainly due to nutritional imbalances and phytosanitary issues that have severely impaired fruit yield. Moreover, economic problems with fruit commercialization have affected the revenues of citrus growers. Therefore, it is required to expand current knowledge to reduce production costs, increase yield and target sustainability.

There is a growing interest in Brazil to develop tissue nutrient standards to avoid over-fertilization and nutrient imbalance of horticultural crops to achieve high fruit yields (Nowaki et al., 2017) and incorporate genetic effects using tools of ionomics (Rozane et al., 2015). The mineral composition of an organism or ionome is related to its genetic make-up, a combination of interactive traits (Baxter, 2015).

Tissue analytical data are commonly interpreted using critical nutrient concentration ranges (CNCR) that neglect nutrient interactions (Bates, 1971), and the Diagnosis and Recommendation Integrated System (DRIS) (Beaufils, 1973), which incorporates dual nutrient interactions in an appealing yet geometrically deficient manner. Both CNCR and DRIS ignore the inherent properties of compositional data class, such as non-normal distribution, redundancy of information and sub-compositional incoherence (Bacon-Shone, 2011). Those sources of bias may lead to wrong inferences and conflicting interpretations (Parent et al., 2012; Parent et al., 2013ab). The development of more effective numerical tools to diagnose nutrient status is thus needed to extract complex relationships in the data set of orange crops to recommend more efficient fertilization management programs in orchards.

Compositional Nutrient Diagnosis (CND) method overcomes defaults of the critical level approach, and rectifies DRIS geometry (Parent and Dafir, 1992; Parent et al., 2012; Parent et al., 2013ab), to conduct unbiased multivariate analysis as formerly suggested by Holland (1966). Using partitioning methods, accuracy in calibration exceeded generally 80% (Parent et al., 2012; Parent et al., 2013ab; Rozane et al., 2015; Nowaki et al., 2017) although data partitioning of the fertigated banana data set returned 72% due to luxury consumption or leaf contamination (Deus et al., 2018). Complementarily, Machine Learning (ML) tools could provide additional benefits over traditional statistical methods for prediction and classification purposes (Pietersma et al., 2003) by increasing the accuracy of nutrient diagnosis (Souza et al., 2016), accounting for the numerous varieties of oranges grown in Brazil.

To our best knowledge, there exists no report in the literature combining the CND approach (tested for orange crops by Rozane et al. (2015), using a small survey data set) and ML algorithms (tested by Souza et al. (2016), using an experimental guava data set - *Psidium guajava*) for diagnosing the nutritional status of orange crops. For this purpose, it is crucial to have access to a large, representative and reliable database, composed of leaf mineral analysis and crop productivity. After deriving nutritional standards, the relatively complex index computations to diagnose nutrients is out-of-reach of the end-user. Therefore, the development of a software that performs nutrient diagnosis of citrus plants, accounting for several nutritional factors that limit crop performance would be useful.

We hypothesized that yield was limited primarily by nutrient imbalances rather than individual nutrient concentrations, and that compositional data analysis combined with machine learning techniques could reach high diagnostic accuracy after accounting for varietal effects.

The objective of this study was to develop an unbiased and accurate variety-specific diagnostic method to diagnose the nutritional status of rainfed orange crops, combining compositional data analysis and machine learning methods. The specific objectives were to: (1) conduct a discriminant analysis of orange varietal ionomes; (2) elaborate a classification technique to identify the reference group at high yield level; (3) develop nutritional standards for rainfed orange commercial orchards; (4) compute a misbalance index; (5) classify nutrients in the order of their limitation to crop performance; (6) diagnose the nutritional status of surveyed misbalanced Brazilian orange crops; and (7) develop a software to diagnose the nutrient status of orange crops based on the proposed method, using the results of leaf mineral analysis.

2. REVIEW OF LITERATURE

2.1. Historical background

Among several factors, the nutritional status is relevant due to its direct impact on the productivity of orange trees (Quaggio et al., 2003, Quaggio et al., 2011). Parsimonious supply of fertilizers can increase crop yield, while reducing production costs, mitigating the risks of water contamination (Parent and Natale, 2008), and thus leading to more sustainable systems.

The classes of soils predominant in the areas cultivated with citrus in Brazil are the Ultisols and Oxisols, which commonly present high acidity and low natural fertility, limiting crop yield potential (Corá et al., 2005). Soil chemical analysis is commonly used as a tool to characterize soil fertility and investigate its effects on the nutritional status of plants. However, Rozane et al. (2015) pointed out, based on Smith et al. (1997), that standard soil analysis (0-20 cm layer) presents limitations for fruit trees with a deep root system, as in the case of citrus, because the plants exploit nutrients in deeper layers in the soil than the one effectively evaluated in the sampling.

Consequently, imbalances in plant nutrition of deep-rooted trees (Smith, 1985) are commonly quantified by conducting leaf chemical analysis (Jones and Case, 1990). Tissue analysis allows a direct nutritional diagnosis because the plant is the soil nutrient extractor (Beaufils, 1973). The newly mature leaves, due to their compositional stability, are the most adequate organs for the representation of the nutritional status of citrus plants (Mattos Jr. et al., 2012).

Different methods can assess plant nutritional status using the results of tissue analysis: the critical concentration (Bates, 1971), DRIS (Diagnosis and Recommendation Integrated System) (Beaufils, 1973; Walworth and Sumner, 1987), CND (Compositional Nutrient Diagnosis) (Parent and Dafir, 1992; Modesto et al., 2014; Nowaki et al., 2017; Parent, 2011; Parent et al., 2013ab), and, more recently, machine learning techniques (Souza et al., 2016).

2.2. Critical concentrations and sufficiency ranges

Plant nutritional status has been traditionally diagnosed from the chemical analysis of individual nutrients (Parent and Natale, 2008) using critical concentration or sufficiency ranges, interpreting concentrations as nutrient deficiency, sufficiency, luxury consumption, or excess (Bates, 1971) (Figure 1). The critical ranges are established assuming, under *ceteris paribus*, that nutrients other than that one being studied are not limiting to productivity, and do not interact significantly when present at adequate levels (Parent, 2011).

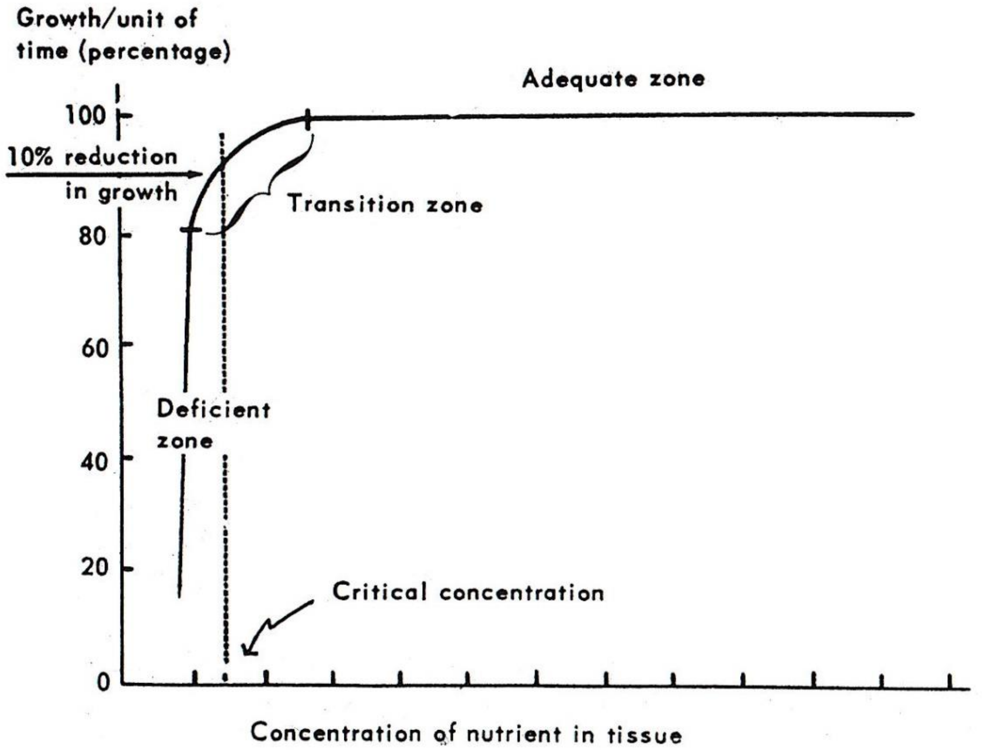


Figure 1. Illustration of the concept of critical nutrient content in plant tissue (Ulrich and Hills,1967).

Nevertheless, since tissue analytical results are confined to a closed space (the unit – g kg^{-1} – or scale – dry matter - of measurement), there must be a resonance effect in the simplex due to composition change (Parent, 2011) affecting other components of the system and inducing a negative covariance or inherently false correlation (if the concentration of one component increases, another one must automatically decrease due to closure, and inversely) (Parent et al., 2013a).

For a given nutrient, the critical concentrations may vary because of the interference of other elements in their absorption and/or assimilation by the plant (Parent and Natale, 2008). Therefore, the use of the critical concentration ranges to interpret plant nutritional status is limited, not accounting adequately for interrelationships among nutrients (Bates, 1971; Parent and Dafir, 1992; Parent et al., 2012), which result from antagonism, synergism and luxury consumption (Parent, 2011).

Another limiting aspect is that the reliability of the nutritional assessment is more consistent when the soil, climate and management conditions of the crop under diagnosis are similar to the conditions in which the critical concentrations or the sufficiency ranges were established (Partelli et al., 2014; Silva et al., 2005; Urano et al., 2007). Consequently, to develop reliable nutritional standards, it is necessary to carry out calibration experiments in several places to ensure a wide representation of the soils, climates and yield potentials of crop varieties. This is often not feasible due to limited human and financial resources, and because it requires a long period of time for conducting the experiments. Moreover, the critical concentration method does not return an order of nutrient limitations to crop yield (Baldock and Shulte, 1996), which is important to assist nutrient management at high yield level.

2.3. Diagnosis and Recommendation Integrated System (DRIS)

The nutritional diagnosis was re-engineered with the development of the DRIS method by Beaufils (1973). Nutrient ratios have been traditionally used in agronomy, aiming to incorporate nutrient interactions on the analysis (Walworth and Sumner, 1987). DRIS takes into account binary interactions (dual ratios), in which relationships between two nutrients in the sample to be diagnosed are compared to DRIS standards (ratio means and coefficients of variation) obtained in a high productivity population (Walworth and Sumner, 1987). It computes D x (D-1) / 2 dual ratios and their corresponding ratio functions, and then integrates the functions into D indexes (Parent et al., 2013b). DRIS standards for orange were developed in USA from 3161 observations (Beverly, 1987), in Venezuela from 1019 observations (Rodríguez et al.,

1997) and in Brazil from 47-50 observations (Hernandes et al., 2014; Mourão Filho and Azevedo, 2003).

The DRIS method shows some remarkable advantages in comparison to the critical concentration and the sufficiency range approach, such as: continuous scale, easy interpretation, ordering from the least to the most limiting nutrients, and an index that reflects nutritional balance (Baldock and Shulte, 1996). On the other hand, DRIS has poor mathematical background (Parent, 2011; Parent et al., 2013b), mainly attributed to: a-) the impossibility of excluding false positive specimens (Rozane et al., 2013); b-) arbitrariness in the choice of high productivity subpopulation (Parent and Natale, 2008; Khiari et al., 2001b; Walworth and Sumner, 1987); c-) the computation of D x (D-1)/2 possible dual ratios and D indices from a D-parts composition while there are D-1 degrees of freedom (Aitchison and Greenacre, 2002); d-) incompatibility with multivariate analysis methods (Parent and Dafir, 1992; Parent and Natale, 2008); and, e-) results expressed as indices that are not linearly independent from each other but are thought to be additive (Mourão Filho, 2004). Those numerical limitations might lead to biased diagnosis (Mourão Filho, 2004; Parent et al., 2013b).

2.4. Compositional Nutrient Diagnosis (centered log ratio technique)

Mathematically sound tools for nutrient diagnosis based on Aitchison (1986) pioneer work on Compositional Data Analysis were developed as Compositional Nutrient Diagnosis (CND). The first version of CND, named centered log ratio (CND-clr) was proposed by Parent and Dafir (1992), allowing the unbiased ordering of nutritional imbalances (Parent and Dafir, 1992; Parent et al., 1994). The centered log ratio was computed as (Aitchison, 1986):

$$\ln (x / g (x))$$

Where x is nutrient concentration and g (x) is the geometric mean of the compositional vector (Parent and Dafir, 1992). The CND method recognizes that nutrient concentration data belong to the compositional data class (Parent and Dafir, 1992; Parent et al., 1994) (i.e. strictly positive data restricted between zero and the unit of measurement), where each part cannot be interpreted in isolation because they are intrinsically related to each other (Aitchinson, 1986).

Compared to the critical concentration approach, CND-clr represents a significant improvement, considering nutrient ratios as in DRIS but in a geometrically sound manner (Parent and Dafir, 1992). Furthermore, it allows using data from commercial crops to develop nutritional standards and compute indices (Partelli et al., 2014; Urano et al., 2007).

The CND-*clr* indices showed low correlation with concentration values (Parent et al., 1994; Khiari et al., 2001a). Correlations with crop yield was higher using CND-*clr* indices than nutrient concentrations (Khiari et al., 2001a). Moreover, there was little agreement between optimal ranges estimated by CND-*clr* and sufficiency ranges for the following crops: flooded rice crops (Wadt et al., 2012), Pêra sweet orange (Camacho et al., 2012; Dias et al., 2013ab), soybean (Urano et al., 2007), cotton (Serra et al., 2010b) and irrigated beans (Partelli et al., 2014). Optimal concentration ranges obtained with CND-*clr* showed lower amplitudes than sufficiency ranges applying conventional methods (Camacho et al., 2012; Dias et al., 2013a; Partelli et al., 2014; Serra et al., 2010b; Urano et al., 2007; Wadt et al., 2012). This might confer to CND-*clr* higher efficiency in nutritional diagnosis (Camacho et al., 2012; Partelli et al., 2014), since a smaller amplitude of the sufficiency range of nutrients can provide greater accuracy in the interpretation of nutritional results (Serra et al., 2010b), increasing the probability of identification of deficiency, luxury consumption or toxicity in the plants under diagnosis (Dias et al., 2013a, Partelli et al., 2014).

Because it preserves the Euclidean distances (Modesto et al., 2014), the centered log ratio used in CND-*clr* rectifies DRIS geometry (Parent and Dafir, 1992), hence allowing to conduct multivariate analysis as formerly suggested by Holland (1966). Besides, CND-*clr* differs from DRIS in the following aspects: a-) it enables computing a Mahalanobis distance to identify outliers in the compositional data sets (Filzmoser and Hron, 2008; Parent et al., 2009; Rozane et al., 2013); b-) the selection of the reference population at high yield level can be determined using cubic function of cumulative variance for individual nutrient (Khiari et al., 2001b) (Figure 2), although binary partitioning methods using the Mahalanobis distance are more conclusive (Parent, 2012); c-) it presents more simplified calculations (Parent et al., 1994); d-) it is proved to be useful to compute nutrient indices to classify nutrients unbiasedly in the order of their limitations to crop performance (Parent, 2011), and e-) it allows conducting exploratory biplot analysis (Egozcue and Pawlowsky-Glahn, 2011). Nutrient diagnosis using CND-*clr* is available online at species level (http://www.registro.unesp.br/#!/sites/cnd/), assisting on the definition of appropriate fertilization management to achieve high yield on different crops. Brazil is the world-leading country in that matter.

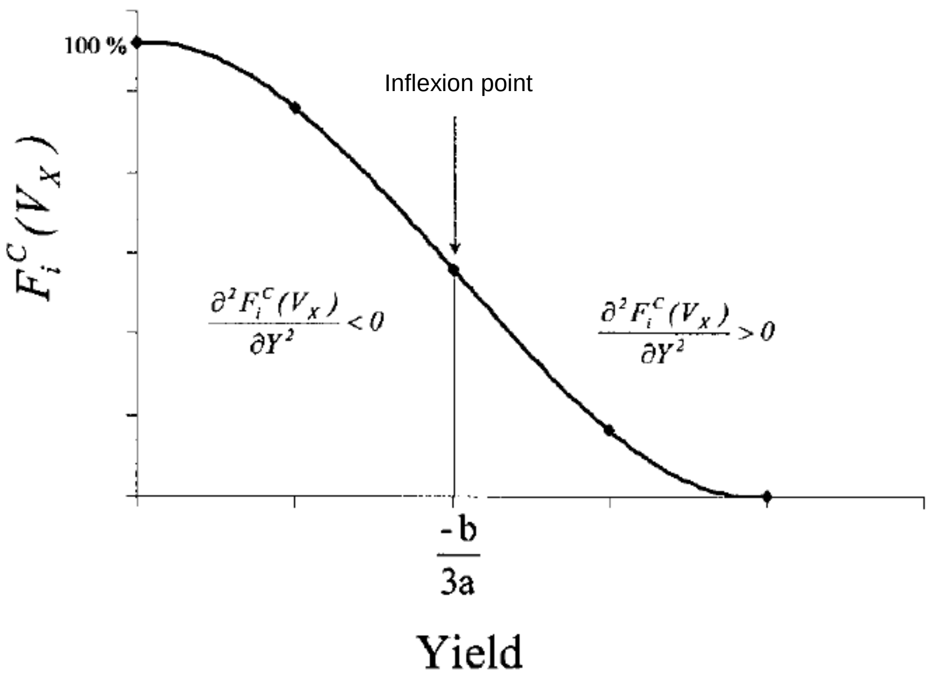


Figure 2. Selection of the reference population at high yield level on the CND-clr approach, using cubic function of cumulative variance ($F^c_i(V_X)$) for individual nutrient (Adapted from Khiari et al., 2001b).

The specific numerical properties of CND-*clr* might lead to different nutrient diagnoses compared to DRIS results, as reported for eucalyptus (Silva et al., 2004), sweet corn (Khiari et al., 2001a) and potato crops (Khiari et al., 2001c). Compared to DRIS, CND-*clr* nutritional indices commonly return the higher coefficients of correlation with yield (Khiari et al., 2001ac), and greater diagnostic potential (Kumar et al., 2003). In contrast, DRIS might lead to conflicting interpretations when correlated with yield (Parent et al., 2012).

Despite its previously described advantages in comparison to CNCR and DRIS, the centred log ratio transformation (CND-*clr*) (Parent and Dafir, 1992) presents some mathematical drawbacks that should be taken into account and corrected. The application of CND-*clr* method results in a number of D clr variables for matrix rank D-1 (Aitchison and Greenacre, 2002). The extra degree of freedom (Parent et al., 2013b) induces redundant information (Modesto et al., 2014) that results in a singular covariance matrix (Modesto et al., 2014; Parent el al., 2013b). One *clr* variable must thus be excluded to perform multivariate analysis (Parent et al., 2013b), hence losing information on the system. Furthermore, diagnostic accuracy of CND-*clr* might be decreased by variations in tissue nutrient concentrations due to contamination by fungicides or foliar nutrient sprays (Parent et al., 2013b). Log ratios are affected by outliers (Filzmoser et al., 2009; Filzmoser and Gschwandtner, 2013) that alter the geometric mean across concentrations (Parent et al., 2012).

Indeed, statistics computed across compositional data class, such as nutrient concentrations (Marchand et al., 2013; Parent et al., 2013b), require an adequate transformation of the data to avoid errors when conducting statistical analysis, which result in spurious correlations and wrong inferences (Bacon-Shone, 2011). Results are biased by: a-) scale dependence, b-) non-normal distribution, and c-) redundancy (Bacon-Shone, 2011; Parent et al., 2013ab).

The scale-dependency refers to the variation of the covariance structure, depending on the unit of measurement on which the composition is scaled (Parent et al., 2013ab). This problem has been hidden by the “k” factor when computing nutrient standards in DRIS. Appropriate transformation should be applied on the concentrations data to avoid sub-compositional incoherence (Parent et al., 2013b), and enable a coherent interpretation of the results of multivariate analysis (Aitchison, 1986; Egozcue and Pawlowsky-Glahn, 2005).

The non-normal distribution of concentration data is attributed to the specific numerical properties of compositional data that are relative to each other and constrained to closure (e.g. unit of measurement), therefore, data distribution is intrinsically logistic-normal (Bacon-Shone, 2011). Consequently, an adequate treatment of the data is required to reach normal distribution and project compositional data into a real space before conducting statistical analysis (Parent et al, 2013b).

Confidence intervals about means must not be allowed to reach below 0 and above 100% or measurement unit, which is physically absurd. The log ratio allows scanning the whole real space (±α).

The redundancy of information is related to closure where one component can be computed by difference between measurement unit and the sum of analytical data (Parent et al, 2013a). There should be D-1 degrees of freedom in a D-part ionome (Parent et al., 2013b) because the matrix has rank D-1 (Aitchison and Greenacre, 2002). Redundant information and the associated spurious correlations occur with raw concentration data, dual ratios and centred log ratios (Parent et al., 2013b).

2.5. Compositional Nutrient Diagnosis (isometric log ratio technique)

The nutrient balance concept (CND-*ilr*) (Parent, 2011; Parent et al., 2013ab; Marchand et al., 2013; Modesto et al., 2014; Nowaki et al., 2017) using D-1 linearly independent isometric log ratios (Egozcue et al., 2003) is based on the most adequate method to perform multivariate analysis on compositional data (Filzmoser et al., 2009; Filzmoser and Hron, 2011). The CND-*ilr* generates scale-invariant variables, that generally returns normally distributed data to project them into a real space (Aitchison, 1986; Egozcue et al., 2003; Parent et al., 2012; Parent et al., 2013a), and avoids redundancy by designing a system of meaningful orthonormal balances defined in a sequential binary partition (SBP), without losing any information on the system (Parent et al., 2013a).

The SBP is defined by a (D-1) x D matrix, in which the group numerator (parts labeled “+1”) is contrasted with the group denominator (parts labeled “-1”) in each ordered row, whereas the part labeled “0” is not considered on the balance between parts (Parent et al., 2013ab). The partition of components is performed sequentially at every ordered row, incorporating two contrasts at the time until subsets (-1) and (+1) contain a single part (Egozcue et al., 2003). However, meaningful balances may start at the lower level of the dendogram using well recognized (in the scientific literature) dual nutrient ratios that are further balanced one against others (Parent et al., 2013ab).

Although there are $D! \times (D-1)!/2^{D-1}$ SBPs possible for a D-parts composition, the selected structure of SBP does not influence the final outcome of linear multivariate analyses (Parent et al., 2013b). Modifying the design of SBP only causes a rotation of orthogonal axes from the origin of the coordinates of a scatter (Parent et al., 2012; Parent et al., 2013b), not affecting the Euclidean distance matrix of *ilr* (Modesto et al., 2014). The SBP could be defined based on a pre-exploratory biplot analysis of *clr* variables (Aitchison and Greenacre, 2002), prior knowledge of nutrient interactions (Parent et al., 2012; Parent et al., 2013ab), specific hypothesis (Nowaki et al., 2017), and plant physiology. The interpretation of the final results (Parent, 2011; Parent al., 2013b) reflects the way the analyst conceives the system (Modesto et al., 2014).

The nutritional imbalances in CND-*ilr* are shown by the dissimilarity between two compositions, applying the Mahalanobis distance across selected *ilr* coordinates of ionomes (Parent et al., 2013b). The Mahalanobis distance is a multivariate distance between the reference compositional standards and the diagnosed composition (Parent et al., 2013b) in the inclined hyper-ellipsoidal scatters of plant ionomes (Parent et al., 2012). The lower the distance, the more balanced is the diagnosed specimen because its nutrient composition profile gets closer to that of the reference group (Modesto et al, 2014).

To derive diagnostic compositional nutrient standards, it has been suggested to use a binary classification in a scatter diagram relating *ilr* values (X) to a performance index (Y) that splits the crops into low and high yield (i.e. cut-off yield value) about a predictor delimiter (i.e. Mahalanobis distance) (Parent et al., 2013b). Once delimiters are set and the classification model optimized, four quadrants are partitioned (Nowaki et al., 2017; Parent et al., 2013b; Swets, 1988), namely : I-) True negative (TN) = high yielding crops correctly identified as balanced (below predictor critical index); II-) True positive (TP) = low yielding crops correctly identified as imbalanced (above predictor critical index); III-) False negative (FN) = low yielding crops incorrectly identified as balanced (below predictor critical index) ; IV-) False positive (FP) = high yielding crops incorrectly identified as imbalanced (above predictor critical index).

The TN specimens correspond to observations showing adequate nutrient status and classified as reference population from which the nutritional norms are derived after excluding outliers (Parent et al., 2012; Parent et al., 2013b). In TP specimens, at least one nutrient is nutritionally imbalanced (Parent et al., 2013b). The FN quadrant contains observations where non-nutritional factors impacted crop productivity (Parent et al., 2013b). The FP observations suggest the occurrence of luxury consumption, contamination or abnormally high nutrient use efficiency by the plant (Parent et al., 2013b).

In a more interpretable approach, nutrient balances might be represented by a dendogram or mobile-and-fulcrums design where nutrient levels located in weighing pans are equilibrated at fulcrum (Parent et al., 2013ab) (Figure 3). This model design facilitates interpreting final results, integrating the balance and the concentration domains to extract useful information to guide nutrient management and achieve high crop performance (Parent et al., 2013ab).

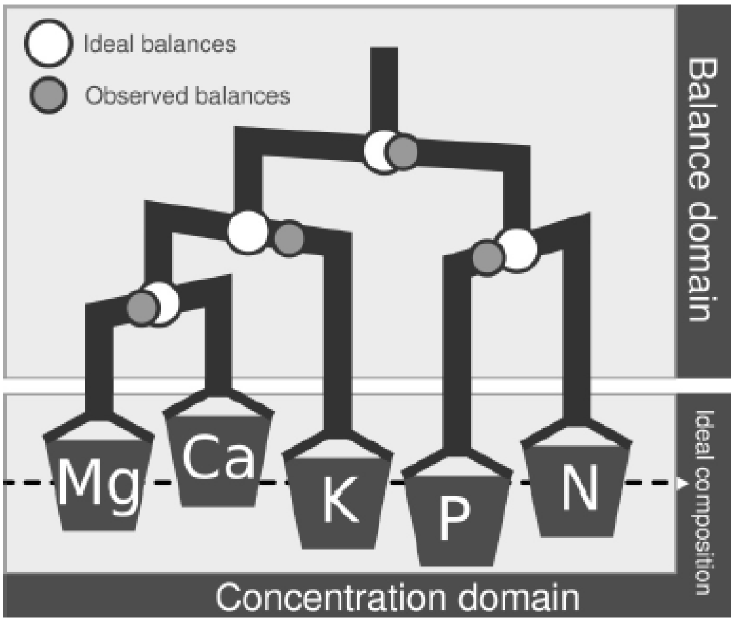


Figure 3. Illustration of balances represented by a dendogram or mobile-and-fulcrums design where nutrient levels located in weighing pans are equilibrated at fulcrum, considering the following nutrients: Mg, Ca, K, P e N (Modesto et al., 2014).

With the exception of fertigated banana where the number of FP specimens was abnormally high (Deus et al., 2018) and irrigated cranberry where the number of FN was high (Marchand et al., 2013), isometric log ratios (*ilr*) (Parent, 2011) returned accuracy higher than 80% applying common data partitioning procedures (Modesto et al., 2014; Nowaki et al., 2017; Parent et al., 2013b; Rozane et al., 2015). The CND-*ilr* was successively applied to plant nutrient diagnosis in several crops: maize (Modesto et al., 2014); kiwifruit (Parent et al., 2013a), apple (Parent et al., 2013a), mango (Parent et al., 2013ab), cloudberry (Parent et al., 2013a), lowbush blueberry (Parent et al., 2013a), cranberry (Parent et al., 2013a); tomato (Nowaki et al., 2017) and guava (Souza et al., 2016).

Considering that nutrient-specific corrective measures might influence all nutritional balances associated with it in a complex interactive system (Parent et al., 2013b), CND-*ilr* provides a holistic view of the changes in nutrient availability in soil-plant system, taking into account all nutrients and their balances in the multivariate analysis (Holland, 1966; Parent et al., 2012; Parent et al., 2013ab).

Nevertheless, the low accuracy (59%) of CND-ilr to classify nutrients in cranberry (Marchand et al., 2013) suggests the importance of accounting for varietal effects. Rozane et al. (2015) found evidence of varietal effects on nutrient composition of the orange diagnostic leaf. Therefore, effective numerical tools are required to extract complex relationships in the data set, especially accounting for varietal effects on nutrient compositions of orange crops to increase the diagnostic accuracy.

2.6. Machine learning methods

Machine Learning (ML) algorithms have great potential to improve the use of CND-*ilr* on nutrient diagnosis of plants. This technique has been developed from a multidisciplinary field, integrating statistics, computer science, artificial intelligence and expert systems (Ma et al., 2014), and might be defined as a general term representing a wide set of models to identify patterns in a data set and make predictions (Witten et al., 2011). Fundamentally, ML elaborates algorithms enabling the computer to extract expert knowledge from unstructured information, which is then applied to generate associations and generalizations on new data (Romero et al., 2013), supporting decisions in agriculture (Chaudhary et al., 2016; Natarajan et al., 2016; Pantazi et al., 2016). Despite such advantages, only a limited number of studies has been carried out to apply ML in plant science.

The supervised ML technique requires five main phases to be implemented (Ma et al., 2014; Pietersma et al., 2003; Shekoofa et al., 2014): I-) extract or collect features (predictors) to be studied with class label assigned; II-) preprocess raw data to improve the representation and quality of the data set; III-) train the model applying an adequate

ML algorithm; IV-) use the ML model to predict the class label of a new observation, based on its features input; and V-) evaluate the generalization performance of the learned model and its validity.

The choice of an optimal ML model depends on the specific problem to be solved, not being possible to provide general recommendations in advance (Behmann et al., 2014). The classification performance attained with ML algorithm might be affected by the type of data preprocessing (Witten and Frank, 2000). In addition, each type of ML model presents specific parameter configurations, which may influence the relationship between imputed covariates and final results, being required to tune these parameters to optimize model performance (Brungard et al., 2015).

Therefore, several ML tests might be run to compare multiple models for achieving optimal results (Pietersma et al., 2003). Nevertheless, it is important to remark that the selection of a new ML algorithm configuration should not rely only on its performance, but also on some other practical characteristics such as: classifier comprehensibility and interpretability, time effort to learn and understand the method, and requirement of specialized software to run the analysis (Kampichler et al., 2010).

Because there are several available ML modeling functions to be explored, it is a challenge to keep track of all the configuration and details of each function (Kuhn, 2008). Indeed, building predictive models in R (R Development Core Team, 2017), using the Caret package, might assist the analyst to take advantage of available ML tools, since it provides a unified interface to test different ML algorithms, allowing to: a-) split the data into training and test sets; b-) eliminate the differences of the synthase configuration between several functions for training and predicting models; c-) test different combinations of pre-processing techniques on the data set, d-) semi-automate the optimization of values of tuning parameters to select the “ideal” model across parameters, e-) use resampling to assist diagnosing over-fitting, f-) estimate model performance, and, g-) conduct a fair comparison between multiple models based on the same training and test sets (Kuhn, 2008).

The ML training (model building) of a supervised model is performed based on an approximation of the function, which establishes the predictive relationship between input covariates (features) and desired outcomes (class, label) (Hastie et al., 2001). More specifically for model training, a set of observations is provided with known membership classes (labels), each one described by a vector of n features (also called variables, attributes or covariates). Posteriorly, the classification model is capable of assigning specific classes to new unseen observations based on previous specific patterns recognized in the feature vectors of the training set.

The performance of the classification model might be evaluated on test sets (validation). A common practice has been to random split the data into one part for model training, and another part for testing. However, single-split hold-outs of the dataset might be detrimental to model generalizability. Instead, the use of cross-validation (CV) has emerged as a statistically robust technique (Kuhn, 2008). This approach reduces overfitting (i.e. the model fits the training data very well, but shows poor performance on testing); and the variability observed when only one split of training and validation test is applied (Li et al., 2014).

The CV technique randomly splits data into n exclusive subsets or folds, each of which maintains similar distribution of classes as in the original dataset (Ma et al., 2014; Pietersma et al., 2003). Thereafter, it applies n successive iterations, where each fold is used once for testing model performance, then fitted on data from the remaining folds (Pietersma et al., 2003; Scott et al., 2013). Indeed, the CV approach performs model training using all samples, and tests it through multiple rounds (resampling) to assess its performance (Kuhn, 2008).

At termination of CV, the average performance (predictive power) is computed based on all data partitions by aggregating the outcomes of each hold-out sample set (Kuhn, 2008; Scott et al., 2013). The suitable tuning parameters associated with the best outcomes in the performance estimation (usually it is considered the largest accuracy) are assigned, and the final model is refitted on the entire training set (Kuhn, 2008). Nevertheless, depending on the use of the model, an external sample set for validation might be required in order to evaluate its performance on data that were not used in the training data set (Kuhn, 2008).

ML provides several benefits over traditional statistical methods for prediction and classification purposes (Pietersma et al., 2003), because it requires few statistical assumptions, building linear and non-linear models, adapts flexibly to a broad range of complex data (Behmann et al., 2014), even where data do not follow the same distribution patterns (Shekoofa et al., 2014), and accommodates both continuous and categorical variables (McQueen et al., 1995). ML allows automating the difficult and time consuming process of acquisition of knowledge from complex dataset for the development of expert systems (Langley and Simon, 1995).

The ML technique provided feasible solutions, especially for data scientists extracting information from Big Data, being robust enough to overcome the limitations of imperfect data such as: heterogeneity, complexity, noise, high dimensionality, irrelevance and incompleteness, which have challenged traditional statistical methods (Ma et al., 2014). In agriculture, several ML models have been successfully applied for soil classification and mapping (Brungard et al., 2015; Kirkwood et al., 2016), image analysis to identify fruit growth stages (Li et al., 2014), diseases diagnosis (Chaudhary et al., 2016), crop yield prediction (Natarajan et al., 2016; Pantazi et al., 2016; Romero et al., 2013), and the assessment of nutritional status of fruit crop (Souza et al., 2016). In contrast, much of conventional statistics is strictly applied to continuous, numeric data, and to perform tests to analyze relationships which have been hypothesized a *priori* (McQueen et al., 1995).

Before applying ML to nutrient diagnosis, Souza et al. (2016) transformed nutrient concentrations into isometric log ratios (CND-*ilr*) to correct any source of bias intrinsically embedded into compositional data (i.e. redundancy, non-normal distribution and scale-dependency). Posteriorly, they applied the k nearest neighbor (knn) ML classification algorithm, reaching foliar diagnostic accuracy of 93% that allowed to develop reliable nutritional standards at high yield level. Such accuracy was to our best knowledge the highest ever achieved.

The knn method is one of the most commonly used ML algorithm in science, producing relatively simple, fast and applicable models to solve a wide range of problems in different fields with high level of comprehensibility and interpretability. The knn can be combined with CND-ilr approach, relying on the Euclidean distance. The Euclidean distance consists in a geometric distance of similarity used in multivariate analysis as follows:

$$d = [(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2]^{1/2}$$

Where d is the distance between two points in n dimensional space and with coordinates ($x_1, x_2, \ldots, x_n$), ($y_1, y_2, \ldots, y_n$) (Ferreira et al., 2015). The knn method tests several values of k (number of the nearest neighbors) and selects the ones with the smallest error rate (Rencher, 2002). The distances are ranked, and the class of the diagnosed observation is predicted based on the class of its k nearest neighbors which the observation is assigned to (Ferreira et al., 2015; Li et al., 2014). In practice, the model assumes that the diagnosed sample present similar patterns of the k-closest observations in a multidimensional space, taking into account the features under study.

Considering that data may follow different distribution patterns (Drummond et al., 2002; Gautam et al., 2006), a hyper-space enclosing high-yielding crops (represented by a “hyper-blob” or “hyper-cloud” delineated by peninsulas and bays) returns a greater accuracy on nutrient diagnosis compared to the “hyper-ellipsoids” delineated by Mahalanobis distance assuming multinormal distribution (Parent et al., 2013b). While the Mahalanobis distance is easily computed as global nutrient imbalance index from the critical “hyper-ellipsoid” (Nowaki et al., 2017), a novel numerical method is needed to computed a multivariate distance from the high-yielding “hyper-blob” islands.

2.7. Changing paradigm

As anticipated by Parent et al. (2013a), there is a need for changing the former univariate paradigm based on the Laws of Minimum and Optimum (De Wit, 1992) by the interactive concept of nutrient balances whereby groups of elements must be balanced to optimize plant growth and yield, taking into account all nutrients in multivariate analysis after removing the sources of bias (Holland, 1966; Parent et al., 2012; Parent et al., 2013ab). The mathematical concept behind CND-ilr method can be combined with ML techniques in a novel approach to provide a more efficient nutritional diagnosis of the plant, overcoming the numerical limitations inherent to compositional data class and incorporating the effects of genetic variability on nutrient composition.

3. MATERIAL AND METHODS

3.1 Data sets

Data were collected from 2012 to 2014 in 7-15 year-old non-irrigated commercial orange orchards in São Paulo state, Brazil. Orchards are considered in full production six years following planting (Mourão Filho and Azevedo, 2003). The predominant soils in the surveyed areas are classified as Oxisols and Ultisols (EMBRAPA, 2013). The climatic regime is classified as Aw according to the Köppen-Geiger classification (tropical climate with a winter dry season and rainfall concentrated in the summer season).

The dataset comprised six rootstocks (Rangpur lime, Sunki mandarin, Cleopatra mandarin, Volkamer lemon, Swingle citrumelo and Trifoliate) and six canopy varieties (Hamlin, Valencia, Valencia Americana, Westin, Pera and Natal). The rootstock was assumed to have little effect on nutrient composition based on a pre-exploratory study (data not shown). Planting density varied between 220 and 830 plants per ha.

The original dataset was composed by 716 observations collected on plots of orchards distributed across Center-south region of São Paulo state, including the following cities: Espírito Santo do Turvo, Lucianópolis, Lençóis Paulistas, Botucatu, Iaras, Anhembi, São Manuel, Itapetininga, Nova Europa and Reginópolis. Outliers among *ilrs* were discarded at the 0.01 level using the R “mvoutlier” package (Filzmoser and Gschwandtner, 2013). The final dataset comprised 675 observations from canopy varieties as follows: Valencia (n=336), Hamlin (n=134), Pera (n=153), Natal (n=39), Valencia Americana (n=11) and Westin (n=2).

3.2. Crop management

The mineral fertilization at production phase is intended to offset crop nutrient removal and losses from the system, to meet nutrient demand for fruit development and the growth of leaves, branches and roots (Quaggio et al., 2005). The P and K application rates of 0-160 kg ha^{-1} P_2O_5 and 0-200 kg ha^{-1} K_2O depended on soil chemical analysis and expected yield (Quaggio et al., 2010). The N rate varying from

70 to 240 kg ha^{-1} N was assessed from leaf nutrient concentration and expected yield (Quaggio et al., 2010). Average nutrient exportation per ton of fresh fruit amount to 1.2 kg N, 0.18 kg P, 1.54 kg K, 0.57 kg Ca, 0.12 kg Mg, 0.09 kg S, 1.6 g B, 0.39 g Cu, 2.1 g Fe, 0.38 g Mn and 0.40 g Zn (Mattos Jr. et al, 2003b). Fertilizers were split-applied at three occasions during the wet season (spring/summer). The B, Zn and Mn were mainly supplied as soluble salts by foliar sprays during the plant growth cycle in the spring and summer (Quaggio et al., 2010). Some orchards received 2 kg ha^{-1} B per year as soil application (Quaggio et al., 2010).

The Ca and Mg were supplied between April and June using dolomitic limestone to target 70% base saturation of the cation exchange capacity (CEC) and Mg concentration of at least 4 mmol$_c$ dm^{-3} in the 0-0.20 m layer (Quaggio et al., 2010). The CEC was computed as the sum of cations K, Ca, Mg and exchangeable acidity (H + Al) determined as $(H + Al) = 10^{(7.76+1.053pH_{SMP})}$ using the Shoemaker-McLean-Pratt buffer pH (pH_{SMP}) method (pH$_{SMP}$) (Quaggio et al., 1985). Lime requirement (LR in Mg ha^{-1}) was quantified as $LR = [CEC \times (BS_2 - BS_1)]/(10 \times RNPL)$ (van Raij et al., 1997), where BS$_1$ is the original soil base saturation computed as $100 \times (K^+ + Ca^{2+} + Mg^{2+})/CEC$ on a molar basis, BS$_2$ is the targeted soil base saturation, and RNPL is limestone neutralizing power (van Raij et al., 1997). Other management practices were carried out as recommended for orange crops (Mattos Jr. et al., 2005).

3.3. Tissue analysis

Twenty-five plants were randomly selected per plot for leaf chemical analysis. Four mature leaves approximately six month-old were sampled per tree from fruit-bearing shoots (as the third or fourth leaf from fruit) when fruit size was 2-4 cm (Quaggio et al., 2010), for a total of 100 leaves samples composited per plot.

Leaves were gently washed successively with distilled water, detergent solution (0.1%), solution with hydrochloric acid (0.3%) and deionized water. Samples were oven-dried at 65°C for 48-96 hours and then grounded to less than 2 mm. Leaf samples were analyzed for nitrogen (N), sulfur (S), phosphorus (P), potassium (K), calcium (Ca), magnesium (Mg), boron (B), copper (Cu), zinc (Zn), manganese (Mn), and iron (Fe) using standard procedures (Jones Jr. and Case, 1990) for the determination of the total nutrient contents. Foliar N was determined by micro-Kjeldahl. For other nutrients, samples were digested in a mixture of nitric and perchloric acids (Jones Jr. and Case, 1990).

3.4. Evaluation of fruit yield

The production of each plot was obtained by harvesting and weighing the fruits of all plants and, considering the spacing between plants and the area of the plot, the productivity per area (Mg ha^{-1}) was calculated.

3.5. Compositional transformation

The tissue compositional simplex (S^D) comprised *D* intrinsically multivariate components described as follows (Aitchison, 1986):

$$S^D = \left\{x = [x_1. x_2. \ldots . x_D] | x_i > 0. i = 1.2. \ldots . D; \sum_{i=1}^{D} x_i = \kappa\right\}$$

where x_i is the i^{th} non-overlapping component within the unit of measurement κ (e.g., mg kg^{-1}). The vector comprised concentrations of N, P, K, Ca, Mg, S, Cu, Zn, Mn, Fe, B, and a filling value x_D computed as a difference between κ and quantified components. The inclusion of the filling value enables back-transforming the ilr values into concentration values with familiar units of measurement (Parent et al., 2013b).

Compositional data was expressed as isometric log ratios (*ilr*) to avoid numerical biases (Aitchison, 1986; Parent et al., 2013b) as it was done by Parent et al. (2013b). The orthonormal balance concept that computes isometric log ratios (ilr), developed by Egozcue et al. (2003) and Egozcue and Pawlowsky-Glahn (2005), is considered as the most suitable to conduct robust multivariate analysis of compositional data (Filzmoser and Hron, 2011). The ilr transformation applies the geometric means to properly average ratios (Fleming and Wallace, 1986) and to extract D-1 informative variables from D-part compositions (Aitchison and Greenacre, 2002). The *D*-1 ilrs were computed between non-overlapping subsets of components as follows (Egozcue et al., 2003):

$$ilr_j = \sqrt{\frac{n_j^+ n_j^-}{n_j^+ + n_j^-}}\, ln \frac{g(c_j^+)}{g(c_j^-)}$$

where n_j^+ and n_j^- are the numbers of components at numerator and denominator, respectively, $g(c_j^+)$ and $g(c_j^-)$ are the geometric means for components at numerator and denominator, respectively, and $\sqrt{\frac{n_j^+ n_j^-}{n_j^+ + n_j^-}}$ is a normalization coefficient.

We presented balances as [-1 or components in denominator group | +1 or components in the numerator group]. The part labeled "0" was excluded from the contrast. If the denominator (-1) group of the log ratio presents higher values of concentration, the balance is negative and leans to the left side of the vertical bar, and vice-versa, as occurs in algebra (Malavolta, 2006; Parent al., 2013b).

3.6. Discriminant analysis and balance arrangements

The *ilr* transformed data allows conducting an unbiased discriminant analysis (DA) of plant ionomes (Parent et al., 2013ab). This approach avoids numerical constraints due to non-normal distribution, scale-dependency and redundancy, which might lead to wrong interpretations when DA is performed using raw or ordinary log-transformed concentration (Parent et al., 2013ab).

Nutrients were integrated into balances using a sequential binary partition (SBP) (Parent et al., 2013b). The SBP is a structured system partitioned into subsystems of dimensionally reduced spaces of compositions where nutrient concentrations interact. There are $(D! \times (D-1)!)/2^{D-1}$ possible balance designs in a D-part composition, some being more meaningful and informative than others (Pawlowsky-Glahn et al., 2011). *Ad hoc* balances among components might be defined based on local knowledge, scientific literature, management or a specific hypothesis (Nowaki et al., 2017) to facilitate interpreting the results (Parent et al., 2013b).

The SBP in Table 1 was elaborated considering current knowledge on nutrient management and interactions in plants (Malavolta, 2006; Rozane et al., 2015). We contrasted macronutrients and B, because it interacts with macronutrients (Malavolta, 2006). Macronutrients cations (K, Ca, Mg) were then contrasted with macronutrients anion (N, S, P) to represent charge balance in the cells (Parent et al., 2013b). Macronutrients cations were categorized as monovalent and divalent ions, where K, Ca and Mg compete (Marschner, 2011). Macronutrients anions (N, S, P) were contrasted taking into account nutrients associated with energy (P) or the synthesis of proteins (N, S) (Parent et al., 2013b). The partitions [Cu + Zn] vs. [Mn + Fe], [Cu] vs. [Zn], and [Mn] vs. [Fe] were associated with fungicide sprays (Cu), soil genesis (Fe) or foliar sprays (Mn, Zn). All nutrients were contrasted with the filling value as a measure of nutrient dilution or accumulation in tissues (Jarrell and Beverly, 1981).

Table 1. Sequential binary partition (SBP) of the orange leaf components including the filling value (F_v) between the sum of analytical results and the unit of measurement.

ilr	Balance	N	S	P	K	Ca	Mg	B	Cu	Zn	Mn	Fe	F_v	r	s
1	[Cu, Zn, Mn, Fe | N, S, P, K, Ca, Mg, B]	1	1	1	1	1	1	1	-1	-1	-1	-1	0	7	4
2	[B |N, S, P, K, Ca, Mg]	1	1	1	1	1	1	-1	0	0	0	0	0	6	1
3	[K, Ca, Mg | N, S, P]	1	1	1	-1	-1	-1	0	0	0	0	0	0	3	3
4	[P| N, S]	1	1	-1	0	0	0	0	0	0	0	0	0	2	1
5	[S|N]	1	-1	0	0	0	0	0	0	0	0	0	0	1	1
6	[Ca, Mg | K]	0	0	0	1	-1	-1	0	0	0	0	0	0	1	2
7	[Mg | Ca]	0	0	0	0	1	-1	0	0	0	0	0	0	1	1
8	[Mn, Fe| Cu, Zn]	0	0	0	0	0	0	0	1	1	-1	-1	0	2	2
9	[Zn | Cu]	0	0	0	0	0	0	0	1	-1	0	0	0	1	1
10	[Fe| Mn]	0	0	0	0	0	0	0	0	0	1	-1	0	1	1
11	[Fv| N, S, P, K, Ca, Mg, B, Cu, Zn, Mn, Fe]	1	1	1	1	1	1	1	1	1	1	1	-1	11	1

ilr: isometric log ratio coordinate of orthonormal balance; Fv: filling value; r and s: numbers of components in plus (+) and minus (−) groups, respectively.

3.7. Machine learning model

The objective of machine learning is basically to establish a useful approximation of the function that predicts the outcome of interest (target variable) based on the relationship with input covariates (features or predictors) (Hastie et al., 2001). On the present study, a supervised model was applied, considering that the label information (target variable) was provided, and a classification method was used as it is suitable for discrete label values (Behmann et al., 2014).

The data were arranged into a matrix (X_{ij}) where each row ($x_i = [x_{i1}, x_{i2} \ldots, x_{ij}]$) is a sample composed of j predictors. The input covariates were derived from *ilr*-transformed data and specification of the variety of the canopy. We discretized yield between low and high yielders across a confusion matrix. Each sample (x_i) was associated with an outcome (class label named either “high yield” or “low yield” (y_i), based on a cut-off yield value of 60 Mg ha^{-1}). The yield delimiter was set at 60 Mg ha^{-1} because this value not only optimized classification accuracy but also corresponded to the economic productivity target of citrus growers. In comparison, cutting yields between low and high levels were suggested to be 40-60 Mg ha^{-1} for “Valencia” grafted on three rootstocks (Mourão Filho and Azevedo, 2003; Mourão Filho, 2005), 22 Mg ha^{-1} for “Pera” grafted on “Cleopetra tangerine” (Hernandes et al., 2014) and 55 Mg ha^{-1} for “Valencia”, “Hamlin”, “Pera”, and “Natal” grafted on several rootstocks (Rozane et al., 2015).

Considering that the choice of an optimal ML model depends on the specific pattern of the dataset and on the problem to be solved (Behmann et al., 2014; Kampichler et al., 2010), and that classification performance achieved by the algorithm might be influenced by data preprocessing (Scott et al., 2013) and/or the tuning of model parameters (Pietersma et al., 2003), prior research into methodological aspects was required. We selected the k nearest neighbors (knn) ML algorithm to be used on the present study, as it has been applied with great success to solve a wide range of complex problems in different fields, including to perform the nutrient diagnosis of fruit crops (Souza et al., 2016). The knn method tests several values of k (number of the nearest neighbors) and selects the ones with the smallest error rate (Rencher, 2002). The distances are then ranked, and the class of the diagnosed observation is predicted based on the class of its k nearest neighbors, which the observation is assigned to (Ferreira et al., 2015; Li et al., 2014).

To simplify model training and tuning across modeling techniques and to evaluate their performance through resampling, we followed the guidelines of Kuhn (2008) to build predictive models in R using the caret package. We implemented the following procedures in modeling: a-) tested different combinations of data pre-processing to optimize the model (using "zv" to identify zero-variance predictors; "center" and "scale" to provide simple location and to scale the transformations of each predictor; and "spatial Sign" to project the predictor values for each sample onto a unit circle, applying $x^* = x/||x||$(Kuhn, 2008)); and b-) customized the functions to: (1) expand the set of models evaluated, increasing the size of the default grid of tuning parameters, (2) specify 10-fold-cross validation as the type of resampling to be used; and (3) modify the method to measure model performance against the area under curve (AUC) of the receiver operating characteristic (ROC) procedure, given that the AUC is regarded as an adequate performance index to measure accuracy of classification models (Swets, 1988).

The estimation of the resampling performance was computed by aggregating the results of each hold-out sample set (Kuhn, 2008). The function selects the values of model tuning parameters associated with the best performance calculated from the resampling (Kuhn, 2008). Model training across samples and resampling (e.g., cross-validation) to evaluate the model's efficacy is the most appropriate statistical test for ML data modeling (Kuhn, 2008). The final model is then refit based on the entire training set (Kuhn, 2008).

The knn algorithm configuration which attained the highest diagnostic accuracy was selected as the optimal model for the present study. Classification performance of knn model was optimized by pre-processing the data (using "zv" to identify zero-variance predictors; "center" and "scale" to provide simple location and to scale the transformations of each predictor; and "spatialSign" to project the predictor values for each sample onto a unit circle, applying $x^* = x/||x||$(Kuhn, 2008)); and using a number of 11 nearest neighbors (kmax = 11), a distance of two, and a kernel set at optimum.

3.8. Classification of specimens using machine learning method

After defining the optimal ML model configuration as previously described, we split the data set into training and testing sets, creating random splits within each class (high yield, low yield), preserving the overall class distribution (Kuhn, 2008). Posteriorly, the ML algorithm was applied and the model was fitted using three repeats of 10-fold cross validation on training data. A proportion of 75% of the data was used for model training, and of 25% to evaluate the performance of the model (validation). Depending on how the model is applied, an external validation sample set might be required, so that the evaluation of the performance of the model could be done using data not used for modeling (Kuhn, 2008), since machine learning algorithms may easily cause overfitting in training data (Weiss and Kulikowski, 1991).

After validation, the model was applied to classify the observations of the whole data set (n = 675) to obtain a confusion matrix, used to provide summaries for classification models (Kuhn, 2008). Then, quadrants were partitioned into four subpopulations (Parent et al., 2013b; Souza et al., 2016): (1) true negative (TN) specimens (high performance crops correctly identified as nutritionally balanced), which have been considered the reference population at high yield level (Marchand et al., 2013; Parent et al., 2013b; Modesto et al., 2014), (2) true positive (TP) specimens (low performance crops correctly identified as nutritionally imbalanced), (3) false negative (FN) specimens (low performance crops incorrectly identified as nutritionally balanced, type II error) and (4) false positive (FP) specimens (high performance crops incorrectly identified as nutritionally imbalanced, type I error). Classification performance of specimens in quadrants are calculated as follows (Parent et al., 2013b):

- Accuracy, computed as (TN+TP)/(TN+FN+TP+FP), is the probability that an observation is correctly identified as balanced or imbalanced;

- Sensitivity, computed as TP/(TP+FN), is the probability that a low-performance observation is imbalanced.

- Specificity, computed as TN/(TN+FP), is the probability that a high-performance observation is balanced;

- The negative predictive value (NPV), computed as TN/(TN+FN), is the probability that a balanced diagnosis returns a high performance;

- The positive predictive value (PPV), computed as TP/(TP+FP), is the probability that an imbalanced diagnosis returns a low performance;

3.9. Nutrient balance diagnosis and limiting order of nutrients to fruit yield

In order to provide complementary information on the relationship between nutrient imbalance and yield, we elaborated a compositional dendrogram as balance scheme (Nowaki et al., 2017) where confidence intervals (P = 0.05) about *ilr* means of true negative (TN) specimens are presented at fulcrums, and concentration centroids of back transformed *ilr* means of TN and true positive (TP) specimens are shown in buckets.

To establish the order of yield-limiting nutrients, we considered the TN and TP quadrants previously obtained from the ML classification of specimens. Then, we computed the clr transformed data which are useful to classify nutrient indices in the order of nutrient limitation to crop performance (Parent, 2011). The clr of each component (clr_X) was computed as the ratio of nutrient concentration of the (x) to the geometric mean of the D-part composition (G) as follows (Parent and Dafir, 1992):

$$clr_x = ln\frac{x}{G}$$

The geometric mean (G) was computed as follows:

$$G = ([\mathrm{N}]\mathrm{x}\ [\mathrm{P}]\mathrm{x}\ [\mathrm{K}]\mathrm{x}\ [\mathrm{Ca}]\mathrm{x}\ [\mathrm{Mg}]\mathrm{x}\ [\mathrm{S}]\mathrm{x}\ [\mathrm{B}]\mathrm{x}\ [\mathrm{Cu}]\mathrm{x}\ [\mathrm{Fe}]\mathrm{x}\ [\mathrm{Mn}]\mathrm{x}\ [\mathrm{Zn}]xFv)^{\frac{1}{D}}$$

where F_v is the non-quantified components computed as:

$$F_v = 1\ 000\ 000 - (\text{sum of quantified nutrient concentrations in mg/kg})$$

We determined the shortest "pathway" from the diagnosed composition to the TN ionome zone (high performance crops correctly identified as nutritionally balanced), by identifying the 10-Nearest Neighbors (10-NN) in the TN multidimensional "hyper-blob" space. The 10-NN were those showing the shortest Euclidean distances, computed between the diagnosed composition of vector X (x_1 to x_n = clr means of TP specimens) and the reference vector Y among TN specimens (y_1 to y_n = clrs of TN specimen), as follows:

$$d = [(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2]^{1/2}$$

where d is the distance between two points in n dimensional space and with coordinates (x_1, x_2, …, x_n), (y_1, y_2, …, y_n). The 10-NN were considered as the closest target compositions to reach high fruit productivity (i.e. reference population) to compute clr indices as follows:

$$I_{clr_x} = clr_x - clr_x^*$$

where I_{clr_x} is the x^{th} *clr* index, clr_x represents the composition of the specimen to be diagnosed (in our study: clr mean of TP group) and clr_x^* is the nutritional standard, corresponding to the mean of the 10-NN among TN specimens. Positive (+) and negative (-) indices indicate relative excess and deficiency, respectively. The highest value of the indice (in module) indicate the most limiting nutrient to crop performance.

3.9.1. Statistical analysis

Statistical computations were conducted in the R statistical environment (R Core Team, 2017). Compositional data transformations were computed using the R "compositions" package (van den Boogaart et al., 2014). Outliers among *ilrs* were removed using "mvoutlier" package. Discriminant analysis (DA) to compare the nutrient balances of orange varieties was performed using the R "aded4" package

(Chessel et al., 2011). Machine learning modeling was conducted using the R “caret” package (Kuhn, 2008). We applied the t-test ($p < 0.05$) to compare ilr means of true negative (TN) specimens with those of true positive (TP), false negative (FN) and false positive (FP) specimens. “Critical” nutrient concentration ranges were approximated after back-transforming *ilr* values associated with 95% confidence intervals ($p \leq 0.05$) of TN and TP specimens. We extracted minimum and maximum concentrations to represent univariate TN and TP ranges (Nowaki et al., 2017).

3.9.2. Development of software “CND-Citros”

Based on the proposed method, the software "CND-Citros" was developed in the programming language C # (Csharp) and Html associated with JavaScript, which will enable the assessment of the nutritional status of orange trees, from the CND concept integrated with ML technique, computing nutritional indices for each nutrient and a misbalance index, to classify nutrients in the order of their limitation to crop performance, using the results of the leaf chemical analysis.

Due to programming limitations, however, we could not yet incorporate on the software “CND-Citros” the option to specify the canopy variety, and the function to compute the multivariate distance to identify the 10-nearest TN neighbors from the composition to be diagnosed, located in bays or peninsulas of high-yielding ionome zone (still under development). Therefore, the clr nutrient imbalance indices to define the limiting order were computed across the barycenter of the reference group (TN specimens defined by ML algorithm).

For the elaboration of the algorithms in the C # language was used the Visual Studio Professional® tool. The application produced is compatible with mobile devices (notebooks and tablets) and microcomputers with the Windows operating system. For this same language, the Xamarin Studio® tool was used, and the software is compatible with smartphones and tablets with the Android® operating system.

For the elaboration of the algorithms in the languages Html and JavaScript, was used the tool Microsoft Expression Web 4®. The developed software is compatible with any device connected with the internet, that presents browsers such as Google Chrome®, Mozilla Firefox®, Opera® and Safari®. For these same languages, the webOS TV Eclipse Ide® tool was used, with the applications produced compatible with the webOS® operating system, currently common in electronic devices as smart TVs.

Because of the multiplatform purpose of the "CND-Citros" software, the project was structured in the Unified Modeling Language (UML). The UML consists of the elaboration of a plan or diagram for each software, where concepts and symbols are inserted, allowing a logical view of the whole process, facilitating its construction. The diagram allows to represent the way the software works, as well as the input and output information elements that the software processes.

In order to prevent the copy and / or reverse engineering of the project by third parties, and to ensure the free distribution of the material to all national and international public, the software will be registered with the National Institute of Industrial Property (INPI), as done for CND -Goiaba® (UNESP, 2013a) and CND-Manga® (UNESP, 2013b).

The distribution of the software / application will be done through institutional electronic addresses such as those of Universities, just like what already exists. It will also be possible to access the material through the Play Store and specific books with included CDs, allowing the consultation by citrus growers, agronomy engineers and technicians.

4. RESULTS AND DISCUSSION

4.1. Discriminant analysis

The discriminant analysis (DA) was performed across isometric log ratios (*ilrs*) of six varieties of orange (“Valencia”, “Pera”, “Hamlin”, “Natal”, “Valencia Americana” and “Westin”) grown in the state of São Paulo, Brazil (Figure 4).

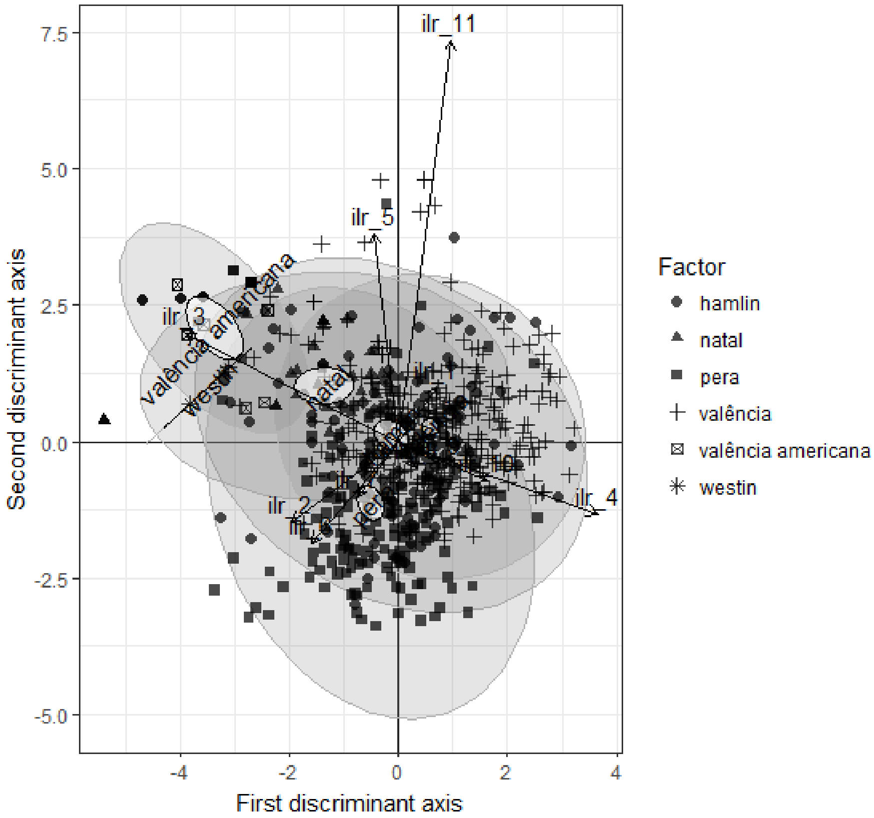


Figure 4. Discriminant analysis of foliar compositions of six varieties of orange grown in the state of São Paulo, Brazil: “Valencia” (336 obs.), “Pera” (153 obs.), “Hamlin” (134 obs.), “Natal” (39 obs.), “Valencia Americana” (11 obs.) and “Westin” (2 obs.). Large semitransparent ellipses represent regions that include 95% of the theoretical distribution of canonical scores for each species. Smaller plain white ellipses represent confidence regions about means of canonical scores at 95% confidence level.

The *ilr*-based DA of plant nutrient compositions averaged across orange varieties showed that distributions of semitransparent ellipses overlapped, whereas the confidence regions about means (smaller white ellipses) differed significantly between varieties of citrus, hence indicating plant-specific ionomes (Figure 4), possibly due to differences of genotype (Parent et al., 2013a). The DA of ionomes of four varieties of mango (“Palmer”, “Tommy”, “Espada” and “Haden”) (Parent et al., 2013b) as well as of different fruit species (kiwifruit, orange, mango, guava, cloudberry, lowbush blueberry and cranberry) (Parent et al., 2013a) also showed the influence of the genotypic variability on foliar nutrient composition. The results indicate varietal effects on nutrient balances of fruit crops tied to specific diagnostic standards.

The DA of *ilr*-transformed data has been considered as an appropriate and unbiased mathematical technique to evaluate the effects of genetic variability on plant nutrient signatures (Parent et al., 2013ab). Such DA avoids numerical biases inherent to compositional data due to redundancy, scale-dependency and non-normal distribution, allowing a robust interpretation of the results in the multivariate analysis across plant ionomes (Parent et al., 2013ab).

Nutrient management of orange crops in Brazil relies on information obtained from trials conducted across a limited number of varieties (Quaggio et al., 1998, 2003, 2005, 2006, 2011). This knowledge is then transferred to other citrus varieties (Quaggio et al., 2005; Rozane et al., 2015) despite evidences of specific nutrient varietal profiles (Mourão Filho and Azevedo, 2003; Parent et al., 2013a; Rozane et al., 2015) (Figure 4). The results of the present study (Figure 4), however, suggest that the use of general nutritional standards on the evaluation of nutritional status of specific orange varieties might lead to inaccurate diagnosis and wrong interpretation of the outcomes.

There is a need to develop variety-specific nutritional norms, or classify nutrient profiles of orange varieties into homogenous groups for nutrient management purposes (Rozane et al., 2015). Nevertheless, such approaches require sufficient data at varietal level, which would likely not be feasible on the short run due to limited human, time and financial resources. Therefore, the development of more effective numerical tools is required to extract complex relationships and identify specific nutritional patterns in the already available data set, leading to an increase of the accuracy of nutrient diagnosis of orange crops to attain high fruit yield with an effective and low-cost approach.

4.2. Classification of specimens

The k-nearest neighbour model run across *ilr* coordinates and canopy varieties performed well in the classification of orange crops. The observations (n = 675) were classified in a confusion matrix using the ML algorithm (Figure 5). Model performance is shown for the whole and the validation data sets, respectively (Table 2).

n = 675	Predicted: HY	Predicted: LY
Actual: HY	TN = 147	FP = 56
Actual: LY	FN = 28	TP = 444

Figure 5. Classification of observations of the whole data set (n = 675) using the k-nearest neighbour machine learning model fitted on the training data set and considering the cut-off yield value of 60 Mg ha^{-1} (Actual and Predicted yield; HY = high yield, LY = low yield; TN = true negative; TP = true positive; FN = false negative; FP = false positive).

Table 2. Performance indices obtained with the classification of observations using the k-nearest neighbor machine learning model trained using the training data set, considering a cut-off yield value of 60 Mg ha^{-1}.

Data set	Accuracy	Specificity	Sensitivity	NPV†	PPV†
Whole (n = 675)	0.88	0.72	0.94	0.84	0.89
Validation (n = 168)	0.85	0.68	0.92	0.77	0.87

† NPV = negative predictive value; PPV = positive predictive value.

Most observations were correctly diagnosed either as nutritionally balanced (TN) or imbalanced (TP), with 85-88% accuracy (Table 2), indicating high potential of the proposed method as an advanced procedure to diagnose the nutrient status of orange orchards. In comparison, Souza et al. (2016) also attained high diagnostic accuracy (93%) in a planned 5-year experiment on single guava (*Psidium gujava*) variety using isometric log ratios and k-nearest neighbour (knn) classification, allowing the development of reliable nutritional standards at high yield level. Several machine learning models have been successful in soil classification and mapping (Brungard et al., 2015; Kirkwood et al., 2016), image analysis to identify fruit growth stages (Li et al., 2014), diseases diagnosis (Chaudhary et al., 2016), and crop yield prediction (Natarajan et al., 2016; Pantazi et al., 2016; Romero et al., 2013), providing consistent solutions to complex problems in different fields of agriculture.

In our survey data set, a total of 147 observations were classified as TN (21.8% of the whole data set) (Figure 5) by the ML algorithm. TN quadrant is theoretically composed of samples with high performance, correctly identified as nutritionally balanced (Parent et al., 2013b). The TN subpopulation is considered the reference population to develop nutrient standards (Marchand et al., 2013; Modesto et al., 2014; Parent et al., 2013b). The number of specimens in the TN quadrant (Figure 5) was adequate to assure statistical significance and validity according to previous studies (Marchand et al., 2013; Modesto et al., 2014; Parent et al., 2013b).

The performance of the classification model was robust and consistent (Table 2), allowing to elaborate reliable TN nutritional standards at high yield level. The TN quadrant was composed of “Valencia” (n = 90) and “Hamlin” (n = 57) only, indicating lower yield potential for other varieties (“Valencia Americana”, “Westin”, “Pera” and “Natal”). Nevertheless, while TN standards appeared most appropriate for “Valencia” and “Hamlin”, the ML method incorporated the varietal effects shown in Figure 4 into nutrient standards, since input covariates (predictors) on the model were derived from *ilr*-transformed data and specification of the variety of the canopy.

The TP group (n = 444) represented 66% of the whole data set (Figure 5). TP specimens are composed of low performance crops, correctly identified as imbalanced (Parent et al., 2013b). Indeed, at least one nutrient may have caused imbalance leading to yield reduction on the observations of this quadrant. The classification model returned 56 false positive (FP) specimens (8.3% of the data set) (Figure 5), possibly due to luxury consumption, leaf contamination (Nowaki et al., 2017), or exceptionaly high nutrient use efficiency (Parent et al., 2013b).

The large proportion of imbalanced specimens (TP + FP) representing 74% of the population (Figure 5), indicated that nutrient levels, nutrient management or soil properties were not optimal at most production sites, requiring special attention from the citrus manager. There were 28 observations (4.1% of the data set) classified as false negative (FN) (Figure 5), where non-nutritional factors (not included in the ML model), such as soil quality, climate, rooting depth, pests and diseases (Mourão Filho and Azevedo, 2003) could have influenced crop yield. Parent et al. (2013b), for example, found that soil fertility interacted with varietal effect on the performance of mango (*Mangifera indica*).

The ML model returned 68% to 72% specificity and 92% to 94% sensitivity across the testing and the whole data sets, respectively (Table 2). Most observations classified as imbalanced presented yield lower than cut-off yield of 60 Mg ha^{-1} (PPV =

87% on the validation data set and 89% on the whole data set) (Table 2). On the other hand, more than three quarters of specimens classified as balanced yielded more than cut-off yield (NPV = 77% on the validation data set and 84% on the whole data set) (Table 2). Therefore, the proposed method based on isometric log ratio transformation integrated with ML technique constitutes a robust technique to classify specimens and assess the nutritional status of plants, improving the accuracy of nutrient diagnosis (Souza et al., 2016).

The *ilr* transformation (Egozcue et al., 2003) is the most adequate method to conduct multivariate analysis on compositional data (Filzmoser et al., 2009; Filzmoser and Hron, 2011), hence to diagnose plant nutrient status (Marchand et al., 2013; Modesto et al., 2014; Nowaki et al., 2017; Parent, 2011; Parent et al., 2013ab). Compositional techniques correct errors in the statistical analysis of compositional data (Bacon-Shone, 2011), avoiding spurious correlations and wrong inferences on final results, otherwise biased by redundancy, scale dependence and non-normal distribution (Bacon-Shone, 2011; Parent et al., 2013ab).

Redundancy is rectified by designing a system of balances into a sequential binary partition (SBP), where balances are orthogonally arranged, enabling to project D components into an Euclidean space of D-1 non-overlapping orthonormal log contrasts that are geometrically suited to conduct multivariate analysis (Egozcue et al., 2003; Filzmoser et al., 2009; Parent et al., 2013ab). Scale dependency is corrected by rationing components expressed on the same unit of measurement or scale (Parent et al., 2012). Non-normal distribution is addressed using log-ratio transformations that project data from the compositional space constrained between 0 and 100% into the real space between –α and +α (Aitchison, 1986; Egozcue et al., 2003; Parent et al., 2012; Parent et al., 2013a).

Complementarily, ML methods enable to account for genetic variability affecting the ionome of plant species (Figure 4), since they can incorporate both numerical and categorical data (McQueen et al., 1995) into linear and non-linear models (Behmann et al., 2014) for prediction and classification purposes (Pietersma et al., 2003). ML handles a broad range of complex data (Behmann et al., 2014) and identify specific patterns (Witten et al., 2011), even when the data do not follow the same distribution (Shekoofa et al., 2014). Expert knowledge is thus extracted from unstructured information (Romero et al., 2013), providing numerical benefits over traditional statistical methods (Pietersma et al., 2003).

Previously, a partitioning procedure was combined with the Mahalanobis distance to generate a confusion matrix (Marchand et al., 2013; Modesto et al., 2014; Nowaki et al., 2017; Parent et al., 2013b), assuming multi-normal data distribution. A critical Mahalanobis distance was then computed as a multivariate measure of dissimilarity between compositions in the multi-normal space of “hyper-ellipsoids” (Parent et al., 2012). Lower accuracy (59%) has been reported on nutrient diagnosis applying only isometric log-ratio with common data partition procedures (Marchand et al., 2013). The results were associated with a larger number of false negative (FN) specimens due to non-nutritional factors involved in yield reduction, including genetic variability. In contrast, ML methods return a space of cloud-like critical “hyper-blobs” delineated by bays and peninsulas, which accounts for varietal influence on plant ionome, enabling an accurate classification of specimens on the nutrient diagnosis of the surveyed orange orchards, regardless of data distribution (Table 2, Figure 5).

4.3. Comparing nutrient balances between (TN) and other (TP, FN, FP) specimens

The t-tests detected significant differences between the nutrient balances of TN and other groups (TP and FP) (Table 3). When contrasting the TN and the TP specimens, significant differences ($p < 0.01$) occurred on the following balances: [K, Ca, Mg | N, S, P]; [B | N, S, P, K, Ca, Mg]; [P | N, S]; [Ca, Mg | K]; [Mg | Ca]; [Zn | Cu] and [Fe | Mn] (Table 3). Such results indicate the main sources of potential nutrient misbalances that might have impaired the crop yield on TP group, as will be further discussed on the next section.

There were no significant differences between the balances of the TN and the FN specimens (Table 3), since both groups are theoretically considered as nutritionally balanced (Parent et al., 2013b). The comparison between TN and FP groups showed significant differences in the [K, Ca, Mg | N, S, P], [Ca, Mg | K], [Mg | Ca] and [Zn | Cu] ($p < 0.01$) balances (Table 3), pointing out potential sources of nutrient imbalances, despite not affecting fruit yield, as probably related to luxury consumption, leaf contamination (Nowaki et al., 2017) or exceptionall high nutrient use efficiency (Parent et al., 2013b) on FP specimens.

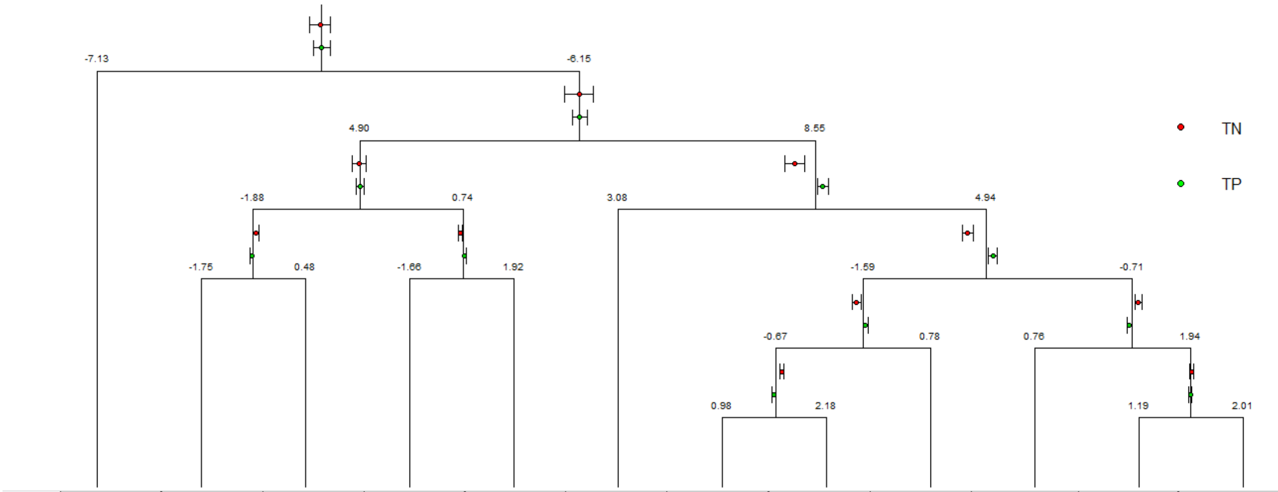
Table 3. Results of t-tests comparing ilr means of true negative (TN) specimens with those of true positive (TP), false negative (FN) and false positive (FP) specimens.

ilr	**Balances**	**TN[1]**	**TP**	**FN**	**FP**
1	[Cu, Zn, Mn, Fe | N, S, P, K, Ca, Mg, B]	6.66	6.66^{ns}	6.73^{ns}	6.69^{ns}
2	[B | N, S, P, K, Ca, Mg]	3.97	4.11**	3.99^{ns}	4.04^{ns}
3	[K, Ca, Mg | N, S, P]	-1.25	-1.17**	-1.27^{ns}	-1.18**
4	[P | N, S]	1.55	1.48**	1.53^{ns}	1.52^{ns}
5	[S | N]	1.60	1.59^{ns}	1.59^{ns}	1.61^{ns}
6	[Ca, Mg | K]	0.09	0.17**	0.16^{ns}	0.22**
7	[Mg | Ca]	1.66	1.57**	1.62^{ns}	1.56**
8	[Mn, Fe | Cu, Zn]	-0.56	-0.55^{ns}	-0.53^{ns}	-0.51^{ns}
9	[Zn | Cu]	0.07	0.22**	0.00^{ns}	0.29**
10	[Fe | Mn]	-0.59	-0,68**	-0.55^{ns}	-0.65^{ns}
11	[Fv | N, S, P, K, Ca, Mg, B, Cu, Zn, Mn, Fe]	-6.67	-6,67^{ns}	-6.64^{ns}	-6.68^{ns}

[1]Reference population; ^{ns}, *, **: not significant different, significant difference at 0.05 and 0.01 probability, respectively, compared to the TN group.

4.4. Compositional dendrogram of True-Balanced (TN) and True-Misbalanced (TP) specimens

Nutrient imbalances can be illustrated by a statistically unbiased and coherent model using a mobile design (Figure 6) that integrates balance and concentration domains into a coherent understanding of nutrient inter-relationships for nutrient management at high yield level (Parent et al., 2013ab). Balances shown at fulcrums are impacted by nutrient levels in buckets (Modesto et al., 2014; Nowaki et al., 2017; Parent et al., 2013b). Confidence intervals (P = 0.05) are shown about the TN and TP *ilr* means at fulcrums. Concentration values in buckets are *ilr* means back-transformed to concentration centroids (Figure 6). Relative deficiency, adequacy, luxury consumption or excess of nutrients (Parent et al., 2013b) can be assessed for specimens outside the critical TN. The balances were elaborated according to current knowledge on nutrient management and interactions (Malavolta et al., 2006; Rozane et al., 2015). Seven balance means differed significantly ($p < 0.05$) between the TN and TP specimens (Figure 6).



	Fv (%)	Fe (%)	Mn (%)	Zn (%)	Cu (%)	B (%)	Mg (%)	Ca (%)	K (%)	P (%)	S (%)	N (%)
TN	91,5040	0,0125	0,0058	0,0052	0,0060	0,0104	0,3676	3,8056	1,3413	0,1257	0,2658	2,5500
TP	91,6498	0,0142	0,0055	0,0045	0,0075	0,0092	0,3860	3,4714	1,4053	0,1466	0,2800	2,6200

Figure 6. Compositional dendrogram representation of balances for true negative (TN) and true positive (TP) specimens. Confidence intervals (P=0.05) about ilr means of true negative (TN) specimens are at fulcrums and concentration centroids are in buckets.

There were marked differences in Ca concentration misbalancing the [Mg | Ca] balance due to apparent Ca shortage relative to Mg in the TP group (Figure 6). Macronutrient cations were categorized as monovalent and divalent ions, where K, Ca and Mg compete (Marschner, 2011). The [Ca, Mg | K] imbalance might be attributed to relative deficiency of Ca combined with K relative excess (Figure 6) driving K-Ca antagonism in the TP group (Marschner, 2011; Mattos Jr. et al., 2004; Quaggio et al., 2011) due to K over-fertilization and insufficient supply of calcium.

The P concentration differed markedly between TPs and TNs (Figure 6). The [P | N, S] balance is associated with the balance between energy production and protein synthesis (Loladze and Elser, 2011). The [P | N, S] imbalance leaned significantly toward the P side of the TPs due apparently to relative P excess (Figure 6), indicating excessive P fertilization. Irrigation may exacerbate P excess because water supply is a means to enhance P use efficiency by increasing the contact between roots and soil P reserves (Barber, 1995; Nowaki et al., 2017), hence requiring even more attention to the P management in irrigated orchards.

Macronutrient cations (K, Ca, Mg) were contrasted with macronutrients anion (N, S, P) to represent charge balance in the cells (Parent et al., 2013b). The [K, Ca, Mg | N, S, P] balance significantly leaned toward the cationic side for TN compared to TP specimens, mainly due to relatively higher Ca and lower P concentrations in TN specimens (Figure 6). We contrasted B and macronutrients because B interacts with macronutrients (Malavolta, 2006). The [B | N, S, P, K, Ca, Mg] balance was significantly lower in TN compared to TP specimens, as related to relative B deficiency combined with excess of P and K in the TPs (Figure 6). In such case, rebalancing orange nutrition requires lowering P and K and enhancing B.

The partitions [Cu + Zn] vs. [Mn + Fe], [Cu] vs. [Zn], and [Mn] vs. [Fe] were associated with fungicide sprays (Cu), soil genesis (Fe) or foliar sprays (Mn, Zn). Relatively high Fe and low Mn levels among TP observations affected the [Fe | Mn] balance, where the TN group significantly leaned toward the right (positive side) (Figure 6). The [Zn | Cu] balance of TN specimens was significantly lower compared to TP, which might be attributed to relatively low Zn or high Cu concentrations in the TP group (Figure 6). Supplemental P may also have contributed to the [Zn | Cu] imbalance by decreasing leaf Zn level (Figure 6) in TP due to P:Zn antagonism (Olsen, 1972; Sumner and Farina, 1986). There were no significant differences between TN and TP groups for the following balances: [Cu, Zn, Mn, Fe | N, S, P, K, Ca, Mg, B], [S | N], [Mn, Fe | Cu, Zn] and [Fv | N, S, P, K, Ca, Mg, B, Cu, Zn, Mn, Fe]. Indeed, such balances likely did not affect orange fruit yield.

All nutrients were contrasted with the filling value as a measure of nutrient dilution or accumulation in tissues (Jarrell and Beverly, 1981). As a result, we verified that nutrient balances were more important than absolute values, due to the fact that there was no significant difference between TN and TP groups for the [Fv | N, S, P, K, Ca, Mg, B, Cu, Zn, Mn, Fe] balance. According to Parent et al. (2013a), one should consider a concept of growth-limiting nutrient balances, whereby groups of nutrients must be optimally balanced, rather than addressing the concept of growth limited by individual nutrient concentrations (univariate perspective). Such approach contrasts with previous concepts that assume that all other factors but the ones being varied by fertilization are equal (Liebig's "Law of Minimum") or "optimal" (Liebscher's "Law of optimum"). The level of any nutrient can be adjusted to the levels of others within the closed interactive system of foliar composition by combining nutrients into the isometric log ratio or orthonormal balance expressions, as conducted on the present study.

4.5. Comparison of orange nutrient standards with literature

To enable comparisons with other studies, we computed apparent concentration ranges (Table 4), using the minimum and maximum values of the confidence interval ($p \leq 0.05$) of back-transformed TN and TP balances. Concentration ranges must be interpreted with care because of the multivariate nature of compositional data and their joint distribution (Nowaki et al., 2017). Several concentration centroids in the Quaggio et al. (2010) sufficiency ranges for orange crops in Brazil (Table 4) such as Ca, K, N, Fe and B differed markedly from those shown in Figure 6, indicating difficulties in defining the optimum nutrition of orange plants in terms of individual nutrient concentration values.

Table 4. Concentration ranges for the true negative (TN) and true positive (TP) specimens confidence intervals about *ilr* values ($p \leq 0.05$) back-transformed to concentration values

Nutrients	**TN[1]**		**TP**		***Reference**	
	LL	**UL**	**LL**	**UL**	**LL**	**UL**
			g kg^{-1}			
N	25.1	25.9	25.9	26.5	23	27
S	2.6	2.7	2.7	2.9	2.0	3.0
P	1.2	1.3	1.4	1.5	1.2	1.6
K	12.8	14.1	13.8	14.3	10	15
Ca	37.2	39.0	34.0	35.4	35	45
Mg	3.5	3.8	3.8	4.0	3.0	4.0
			mg kg^{-1}			
B	98	110	89	95	80	160
Cu	54	67	69	82	10	20
Zn	46	57	43	47	35	50
Mn	54	62	52	58	35	50
Fe	120	130	136	148	50	120

[1]Reference population; TN = true negative; TP = true positive, LL = lower limit; UL = upper limit. *Current Brazilian nutrient concentration standards (Quaggio et al., 2010).

For most nutrients (N, P, K, Ca, Mg, S and B), concentration ranges of TN specimens were within sufficiency ranges reported by Quaggio et al. (2010) (Table 4), although they were narrower. Concentration ranges for Mn and Zn were also narrower (Table 4), but they were higher compared to adequate levels suggested by Quaggio et al. (2010). Smaller sufficiency ranges, as obtained on the present study (Table 4) might increase the efficiency of the nutritional diagnosis (Camacho et al., 2012; Dias et al., 2013a; Partelli et al., 2014; Serra et al., 2010b). The Cu levels were much higher in our data set (Table 4). Concentrations above 20 mg kg^{-1} Cu are generally associated with contamination caused by Cu-based fungicides (Quaggio et al., 2010). Such management may lead to oxidative stress induced by Cu toxicity; potentially affecting plant growth (Hippler et al., 2018) and, consequently, impairing fruit yield of orange orchards.

Differences on the concentration ranges reported in the present research in comparison with results of literature can be attributed to the fact that: (1) only the TN group was used to compute concentration ranges, excluding FP specimens; and (2) concentration ranges tended to become narrower combining more information on nutrient compositions and interactions (Nowaki et al., 2017).

There is a growing interest in Brazil to develop tissue nutrient standards to avoid over-fertilization and nutrient imbalances in horticultural cropping systems to achieve high fruit yield (Nowaki et al., 2017). Nevertheless, tissue analytical data are still commonly interpreted using only critical nutrient concentration ranges (CNCR) (Table 4), that neglect nutrient interactions (Bates, 1971), resulting in antagonism, synergism, or neutrality (Parent, 2011). The CNCR is erroneously established, assuming under *ceteris paribus*, that nutrients other than that one being studied, are not limiting factors to productivity and do not interact significantly if present at adequate levels (Parent, 2011). In addition, such approach ignores the inherent properties of compositional data class, such as non-normal distribution, redundancy of information and sub-compositional incoherence (Bacon-Shone, 2011), not incorporating the effects of genetical variability on leaf nutrient composition (Rozane et al., 2015). Indeed, those sources of bias may lead to wrong inferences and conflicting interpretations (Parent et al., 2012; Parent et al., 2013ab; Rozane et al., 2015).

Therefore, the nutritional standards currently used in Brazil to assess the nutritional status of orange crops (Table 4) should be revisited by the compositional balance concept approach, where groups of elements are balanced to optimize plant growth and yield (Figure 6, Table 4). In this context, the Compositional Nutrient Diagnosis (CND) methods allow overcoming the restraints on critical level approach (Parent and Dafir, 1992; Parent et al., 2012; Parent et al., 2013ab), by conducting unbiasedly multivariate analysis as formerly suggested by Holland (1966). Complementarily, Machine Learning (ML) tools might improve the accuracy of CND diagnosis (Souza et al., 2016) after accounting for varietal effect on nutrient ionomes.

4.6. Order of nutrient limitations in TP specimens

The order of nutrient limitations to crop yield allows detecting the main sources of nutrient misbalance as decision-making tool to adjust fertilizer recommendations to local conditions. Nutrient imbalance indices were previously computed across *ilr* coordinates as the distance between an observation and the barycenter of a reference group (Parent et al., 2012; Parent et al., 2013b; Modesto et al., 2014; Nowaki et al., 2017) using a mobile design (Figure 3). However, because ML methods return a critical “hyper-blob” rather than a critical “hyper-ellipsoid” made of TN specimens, a numerical method is needed to compute the multivariate distance from the high yielding ionome zone, since composition to be diagnosed may be located close to a “hyper-blob” peninsula or to a “hyper-blob” bay.

The ten TN neighbors in bays or peninsulas of the critical “hyper-blob” nearest to the specimen under diagnosis were selected as those showing the shortest Euclidean distances from the diagnosed composition. Nutritional clr indices returned an order of nutrient limitations in the diagnosed TP specimens (Figure 7). The adjustment of nutrient management to re-establish nutritional equilibrium follows the order of nutrient limitations as inherited from DRIS, but in a numerically unbiased manner (Figure 7). It becomes less complex and challenging using the mobile design (Figure 6) integrated with the order of nutrient limitations (Figure 7), as it allows interpreting balances at nutrient levels to assist on the definition of adequate sources and rates of fertilizers to be applied at different occasions.

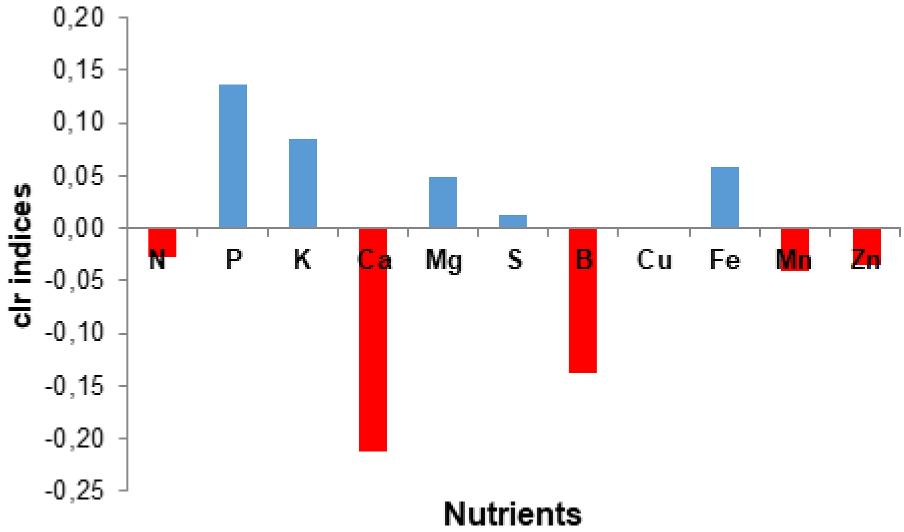


Figure 7. Order of nutrient limitation to crop performance at surveyed sites (TP specimens). Positive and negative *clr* indices indicate relative nutrient excess or deficiency, respectively, in relation to the 10-nearest neighbor TN specimens.

A slight relative nitrogen deficiency was observed in the present study (Figure 7), although it did not significantly affect nutritional balances associated with this nutrient (Figure 6), hence likely not impairing crop productivity on TP group. Citrus plants store large amounts of nitrogen in the biomass, which can be redistributed via phloem to the developing organs that constitute drains of nutrients and photo-assimilates, such as new leaves and fruits (Mattor Jr. et al., 2003a). For this reason, even in cases of relative deficiency (Figure 7), fruit production might not be affected on the short-term (Quaggio et al., 2010), although trees may undergo a gradual reduction of canopy density and growth, which, consequently, will lead to a decrease of fruit production in future seasons (Quaggio et al., 2010).

Nitrogen has been considered as the most important nutrient for defining citrus productivity (Alva and Paramasivam, 1998; Mattos Jr. et al., 2005). The nitrogen application increases leaf density (Smith, 1969) and regulates photosynthetic rate (Kato, 1986). As an integrative response to N supply, there is an increase on the assimilation rate of carbon dioxide and, consequently, on the production potential of citrus trees (Lea-Cox and Syvertsen, 1996). Leaf nitrogen level has been successfully used as an indicator to adjust the N rates on citrus crops in Brazil (Quaggio et al.,1998; Quaggio et al., 2010). Due to its prominent importance on plant metabolism, N supply has been closely monitored by citrus growers to avoid nutritional imbalances associated with this nutrient (Figures 6 and 7), which might potentially impair crop performance.

The Cu and S apparently were also not limiting in TP specimens (Figures 6 and 7), as these nutrients are commonly supplied via foliar application of Cu and S-based products to abate higher disease and pest infestation levels in citrus orchards.

On the other hand, Ca appeared to be the most frequently limiting nutrient in the surveyed orchards (Figure 7). The Ca performs essentially three important functions: structural, enzymatic regulator and as a secondary messenger (Malavolta, 2006). Several studies showed positive response of citrus to liming (Anderson, 1987; Quaggio 1991; Quaggio et al., 1992b), which have been partly attributed to Ca supply, indicating the critical importance of this nutrient in tropical soils (Quaggio et al., 1992b). Orange plants have high demand for Ca, the nutrient accumulating the most in the tree biomass, even more than N (Mattos Jr. et al., 2003b). Nevertheless, calcium uptake is limited to young roots, in which the endoderm cell walls are not suberized. As a result, even medium to high soil Ca levels might lead to Ca deficiency in the plant (Figure 7), and, consequently, impair citrus root development resulting in yield losses (Auler et al., 2011). Furthermore, excessive K applications may have contributed to decrease leaf Ca concentration due to antagonism with Ca in the foliar tissues of TP specimens (Figures 6 and 7), leading to yield reduction (Mattos Jr. et al., 2004; Quaggio et al., 2011). Indeed, applying lime or gypsum to supply Ca to the TP group, combined with a reduction of K fertilization, could contribute to re-establish the main sources of misbalances in the plant, namely: [Mg | Ca], [Ca, Mg | K] and [K, Ca, Mg | N, S, P] balances (Figure 6). Adding Ca as lime also increases soil pH, potentially enhancing the availability of other nutrients in the soil, which may increase crop productivity in general.

The second limitation among TP specimens appeared to be relative P excess (Figure 7). Supplemental P may have misbalanced the [P | N, S] and [K, Ca, Mg | N, S, P] balances in TP specimens, as well as the [Zn | Cu] balance by decreasing leaf Zn level (Figures 6 and 7). Citrus plants show a significant response of root-shoot growth (Zambrozi et al., 2013) and fruit yield to P fertilization in soils with low P availability (Quaggio et al., 1998; Quaggio et al., 2006). Considering that tropical soils are inherently low in P due to strong sorption of phosphate ions by aluminum and iron oxyhydroxides (Roy et al., 2016), the proportion of P fertilizer that could be acquired by the plant from the soil is generally considered to be low (Novais, 1977; Baligar and Benett, 1986ab). Hence, growers tend to adopt a supplemental P fertilization, regardless of soil-test P, especially where yield level is low. However, the use of soil analysis is important to predict the response of citrus plants to P fertilization to adjust an optimal rate (Quaggio et al., 1998), avoiding the indiscriminate application of this nutrient, which may lead to P over-fertilization (Nowaki et al., 2017) and, consequently, yield reduction due to nutritional imbalances (Figures 6 and 7).

The reduction of leaf Zn in TPs (Figure 6) was possibly caused by the P:Zn antagonism. High P/Zn ratios in corn leaf tissue with the application of high rates of P led to Zn deficiency and decreased yield (Sumner and Farina, 1986). Co-precipitation reactions of P and Zn might occur in conducting vessels affected by P excess, inhibiting Zn transport from root to shoot, and causing metabolic disorders in the cells due to P/Zn imbalance (Olsen, 1972). Those nutritional imbalances caused by an apparent P excess observed in the present study (Figures 6 and 7) could explain previous results reported on the literature, in which fruit yield of citrus trees did not generally respond to P additions where soil test P levels were higher than 20 mg resin-P dm^{-3} (Quaggio et al., 1998; Quaggio et al., 2006), whereas yield decrease may occur where soil test P exceeded 40 mg resin-P dm^{-3} (Quaggio et al., 2002). Fertilizer P rate should thus be adjusted based on soil analysis, expected yield (Quaggio et al., 2010), and also the general plant nutritional status.

Boron appeared to be the third most limiting nutrient shown as relative deficiency in TP specimens (Figure 7). Boron plays a role in the formation of the xylem, root growth and water transportation to the canopy (Mesquita et al., 2016). Boron deficiency can thus affect citrus productivity (Quaggio et al., 2003; Liu et al., 2014; Wang et al., 2015). This occurs especially in tropical regions (Mattos Jr. et al., 2017), as attributed to low B availability in the soil due to low levels of organic matter (Goldberg and Suarez, 2011). In addition, tropical climatic conditions can favor the occurrence of boron deficiency. Long periods of drought stress might impair B uptake, since absorption and transport of this nutrient through the plant xylem depends on the transpiration rate (Brown and Shelp, 1997), whereas excessive rainfall may also limit

B absorption, as B has high mobility in a predominantly non-ionic form (boric acid) in the soils of the tropics (Quaggio et al., 2003; Mattos Jr. et al., 2017). Indeed, it is required periodic supply of B on tropical agro-systems. Nevertheless, B is commonly applied as foliar sprays on orange orchards (Quaggio et al., 2003; Boaretto et al., 2011; Mattos Jr. et al., 2012). Because of low B mobility in the phloem of citrus (Boaretto et al., 2008), the redistribution of this nutrient in the plant is limited (Boaretto et al., 2011; Liu et al., 2012), making foliar fertilization inefficient to meet the demand of new flushes of shoot growth, contributing to the observed results (Figures 6 and 7). Boron addition via the soil is a more efficient management strategy to supply B to citrus trees, with positive response on crop productivity (Quaggio et al., 2003; Boaretto et al., 2011; Mattos Jr. et al., 2017). Therefore, soil rather than foliar B fertilization could re-establish the [B | N, S, P, K, Ca, Mg] balance (Figure 6), potentially increasing fruit yield in the TP group.

The fourth limitation in the TP group was relative K excess (Figure 7). Potassium is the nutrient most exported through fruit harvest (Mattos Jr. et al., 2003b). Citrus fruit yield and quality are greatly influenced by K fertilization (Mattos Jr. et al., 2004; Alva et al., 2006; Quaggio et al., 2006; Quaggio et al., 2011). It is thus common to use concentrated K fertilizer formulations in citrus management programs to reach high yield but not at the expense of high-quality fruits. This may lead to K excess in the soil, especially where soil analysis is not taken into account in the prescription (Quaggio, 1996); or where low fruit load decreases K removal (Embleton et al., 1973; Mattos Jr. et al., 2003), reducing K requirement.

The response of orange trees to K fertilization depends on soil test K, and, in the case of high-K soils, excessive K additions may even reduce fruit yield (Quaggio et al., 1998; Quaggio et al., 2006). In the present study, K over-fertilization may have decreased leaf Ca concentration, leading to yield reduction in response to nutritional imbalances (Figures 6 and 7). High potassium rates impair calcium absorption, characterizing the effect of competitive inhibition (Bernardi et al., 2000; Marques et al., 2011). In an experiment with potassium fertilization in 'Pêra' and 'Valência' orange trees grafted on Rangpur lime, Quaggio et al. (2011) observed that the increase in leaf K content was accompanied by a decrease in Ca concentration in the leaves, where high doses of potassium fertilizer were applied. Foliar Ca levels were lower than those considered adequate for citrus. This effect was attributed to the equilibrium of exchangeable cations in the soil, which is proportional to the ratio of K activity in relation to the divalent cations in the soil solution. Increasing rates of potassium fertilizer lead to increase on the K activity in the soil solution, impairing Ca uptake. Similar effects of the ionic antagonism of K with Ca, in which the Ca foliar contents were reduced by the K doses, were also reported in Valencia orange seedling grafted on Rangpur lime (Bernardi et al., 2000) and in Murcott tangor grafted on Rangpur lime (Mattos Jr. et al., 2004). Magnesium showed relative excess in TP specimens (Figure 7), contributing even more to [Mg | Ca] imbalance (Figure 6). A reduction of K fertilization, and Ca additions could likely relief the [Ca, Mg | K] balance in specimens of the TP group (Figure 6). Note that the relationship between soil exchangeable K, Ca, Mg and acidity has been revisited using a novel cationic balance concept based on clr and ilr variables where soil exchangeable K can be balanced against exchangeable Ca and Mg (Parent et al., 2012), or against exchangeable Ca, Mg, acidity and the filling value between the sum of cations (1 g per mol of exchangeable acidity) and the unit of measurement, generally mg kg^{-1} (Parent and Parent, 2017).

The relative Fe excess among TPs (Figure 7) is likely related to soil genesis, given that soil is considered to be a large reservoir of Fe (Parent et al., 2013b), and that this nutrient is not added as fertilizer in Brazilian citrus crops. It should be considered, however, that this relative excess of Fe may not impair fruit yield, because not all iron species in the tissue performs biological function, as they can be in an insoluble or complexed form.

The Mn showed relative deficiency (Figure 7), presumably caused by Mn shortage in the soil due to low Mn levels in parent materials (Hippler et al., 2015) and high Mn sorption by colloids (Quaggio et al., 2010). Moreover, the Mn is often supplied to citrus groves via foliar fertilization (Mattos Jr. et al., 2012). Because Mn presents low mobility in the phloem, its redistribution from mature tissues to young flushes of shoot growth is limited (Boaretto et al., 2003), requiring regular foliar sprays to meet crop demand at every new flushes. The Mn deficiency likely occurred in TP specimens where orchards were not managed properly. Soil application of Mn using soluble sulfate fertilizers to re-establish the [Fe | Mn] balance among TP specimens (Figure 6) may meet the Mn requirement of imbalanced orange trees, especially in soils with low Mn sorption capacity (Hippler et al., 2015).

4.7. Nutrient diagnosis of orange orchards using the software “CND-Citros”

We developed the software “CND-Citros”, aiming to facilitate the implementation of the mathematical calculations proposed on the present work and to improve interpretation of the results of foliar chemical analysis. The software allows the assessment of the nutritional status of orange crops through the application of compositional data analysis integrated with k-nearest neighbors ML classification algorithm, defining the order of nutrient limitations to crop yield.

The results obtained on the software “CND-Citros” express the mineral composition of nutrient levels on the plant tissue in the form of relative values, which is the basic numerical information to establish the diagnosis of the nutritional status of plants. The software assists on the definition of appropriate fertilization recommendations, enabling to reduce production costs, increase fruit yield and target production sustainability, especially in crops grown under edaphoclimatic conditions similar to those of the present study. To use the software 'CND-Citros' for free, access http://www.registro.unesp.br/sites/cnd/ (best viewed in browsers "Google Chrome"; “Mozilla Firefox”; “Opera” and “Safari”) and follow the instructions below (using a hypothetical example):

1. Transcribe in the corresponding fields the nutrient concentrations obtained on the chemical analysis of the leaves (macronutrients : g.kg^{-1}, micronutrients : mg.kg^{-1}) (Figure 8).

Figure 8. Corresponding fields to be filled in with the nutrient concentration data obtained on the chemical analysis of the leaves.

2. After filling in all required fields with foliar nutrient concentrations, click on the ''CALCULAR'' buttom. The balance index for each nutrient will be calculated, as well as the CND-r², which represents the sample's nutritional imbalance index (Figure 9).

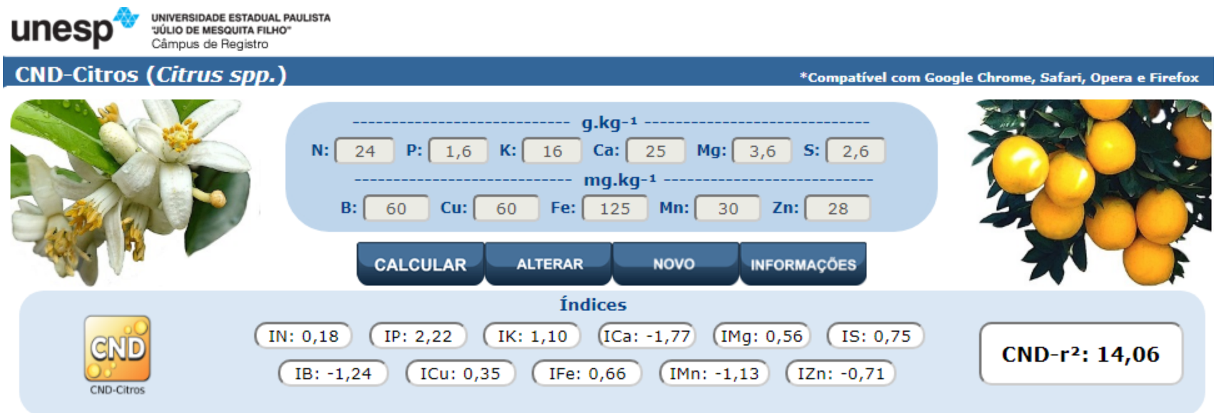


Figure 9. Results of the balance index for each nutrient, and the CND-r², which represents the sample's nutritional imbalance index.

3. The graphics in form of radar (left) and bars (right) will also be shown, representing the nutritional indexes calculated for each nutrient, assisting to define the limiting order of nutrients to crop performance. It may be interpreted that in relation to the high-yield population (i.e. TN reference population), all indexes in green color are in 'excess', as phosphorus (P), potassium (K), and so on, and that every index in red are in 'defficiency', such as calcim (Ca), boron (B), etc. On the other hand, the indexes which tend to zero might be considered as "suitable", such as nitrogen (N) (Figure 10).

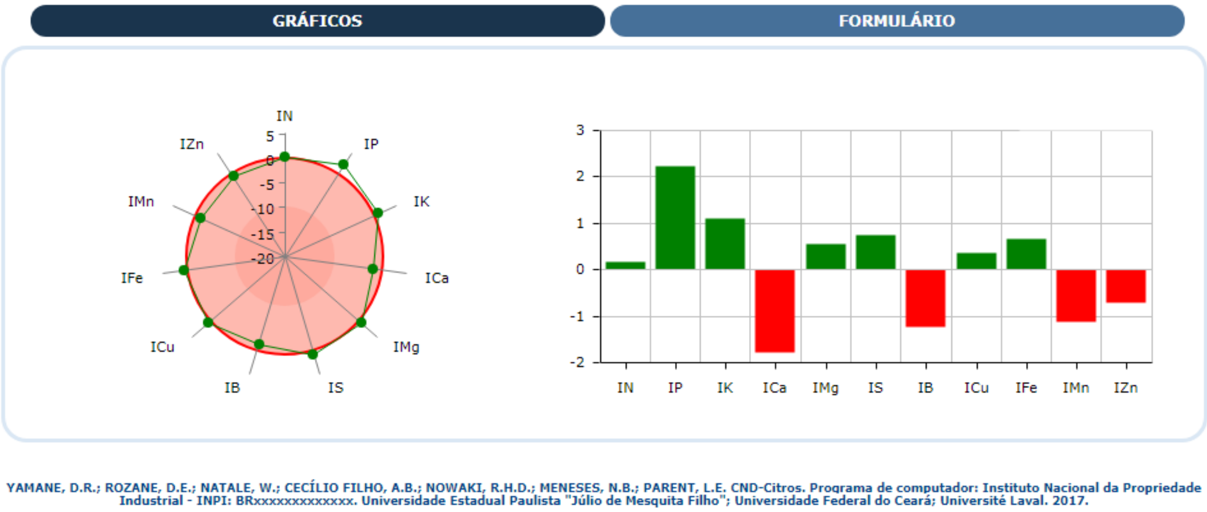
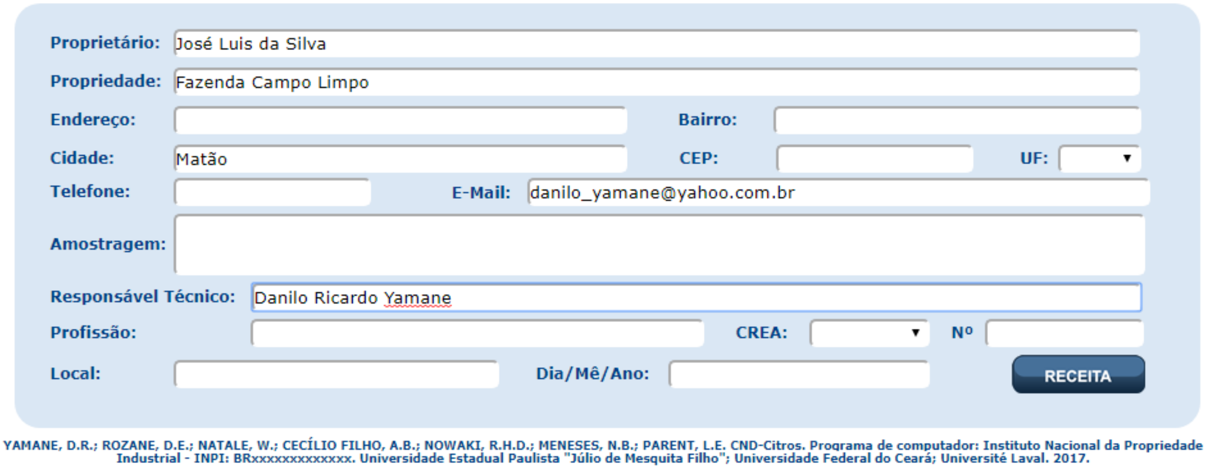


Figure 10. Graphics elaborated in the form of radar (left) and bars (right), representing the nutritional indexes calculated for each nutrient.

4. After analyzing the graphs the user is able to click on “FORMULÁRIO” bottom (see Figure 10), and fill in the fields to create the history of the plots of the orchard, whose details are described in the image below (Figure 11).



Proprietário: José Luis da Silva
Propriedade: Fazenda Campo Limpo
Endereço:
Bairro:
Cidade: Matão
CEP:
UF:
Telefone:
E-Mail: danilo_yamane@yahoo.com.br
Amostragem:
Responsável Técnico: Danilo Ricardo Yamane
Profissão:
CREA:
Nº
Local:
Dia/Mê/Ano:
RECEITA

YAMANE, D.R.; ROZANE, D.E.; NATALE, W.; CECÍLIO FILHO, A.B.; NOWAKI, R.H.D.; MENESES, N.B.; PARENT, L.E. CND-Citros. Programa de computador: Instituto Nacional da Propriedade Industrial - INPI: BRxxxxxxxxxxxxx. Universidade Estadual Paulista "Júlio de Mesquita Filho"; Universidade Federal do Ceará; Université Laval. 2017.

Figure 11. Fields to be filled in to create the report about the history of plant`s nutritional status of the citrus orchard.

5. Finally, it’s possible to click on "RECEITA" (see Figure 11), to generate a complete report of the diagnosis, as illustrated (Figure 12) :

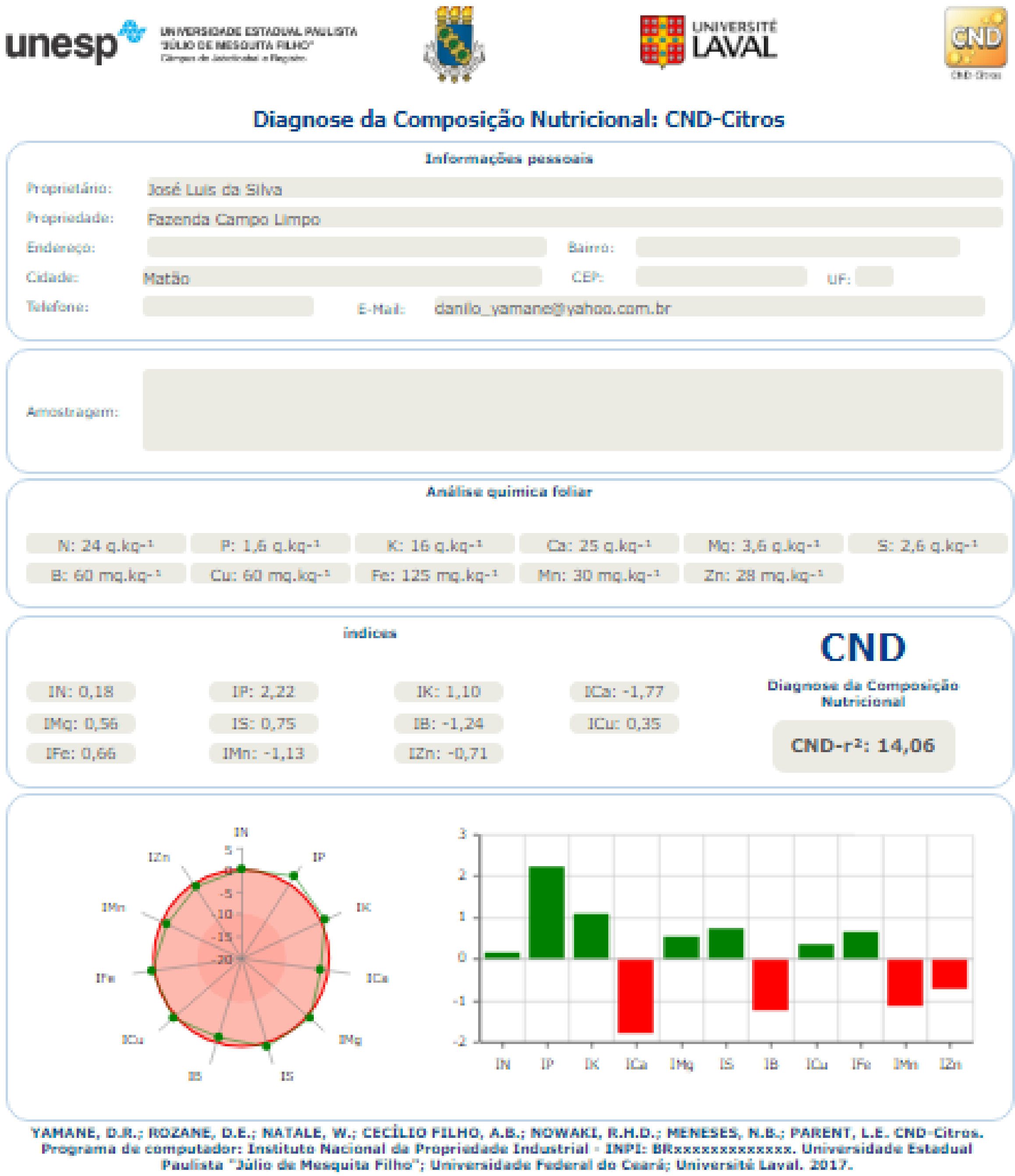


Figure 12. Final report of the diagnosis, where is possible to view the summary of all the information related to the assessment of nutritional status of plant, based on the compositional data analysis integrated with machine learning technique.

5. CONCLUSIONS

- The ilr-transformed data allowed to conduct an unbiased discriminant analysis of orange foliar ionomes. Canopy varieties formed distinct groups due to specific nutrient profiles, indicating that the use of general standards to assess the nutrient status of specific orange varieties, as currently done in Brazil, may return inaccurate diagnosis and wrong interpretation of the outcomes.

- Leaf nutrient status could be diagnosed and monitored accurately using compositional data analysis combined with k-nearest neighbors ML classification algorithm. This research is the first ever to use “hyper-bob” of log ratio expressions to describe the reality of data distribution approximated formerly by “hyper-ellipsoids”. In contrast with conventional methods, the *ilr* transformation rectified any source of bias intrinsically embedded into compositional data (i.e. redundancy, non-normal distribution and scale-dependency), and the ML technique incorporated varietal effects in nutrient diagnosis, enabling to reach 88% diagnostic accuracy in a “hyper-bob”.

- The order of nutrient limitation to crop performance for specimens located outside high-yielding ionome zone delineated by the critical “hyper-blob”, allowed interpreting balances at nutrient levels to adjust fertilizer recommendations to attain high fruit productivity at local scale.

- In the surveyed orange orchards of São Paulo state, several nutrients could be readjusted to yield more than 60 Mg ha^{-1}. Nevertheless, current critical nutrient concentration ranges used in Brazil appears too wide and inaccurate for diagnostic purposes because it does not account for the nutrient interactions and genetic variability impacting the foliar ionome of plant species.

- Citrus growers should hopefully adopt the concept of yield-limiting nutrient balances, where groups of nutrients are optimally balanced. The more familiar conventional approaches to perform the nutrient diagnosis address only the concept of yield limited by individual nutrient concentration, under the traditional over-optimistic assumptions that other nutrients are equal or at adequate levels under the laws of minimum or optimum, respectively, avoiding interactions among nutrients and leading to large “adequate” intervals rather than targeting adequate nutrient balances.

- Our ML-balance diagnosis of misbalanced São Paulo orange orchards suggests that supplying more Ca as lime or gypsum materials, reducing the P and K fertilizer applications, and enhancing soil B fertilization could re-balance nutrients in the low-yielding orange orchards, namely through re-establishing the [Mg | Ca], [Ca, Mg | K], [P | N, S], [K, Ca, Mg | N, S, P] and [B | N, S, P, K, Ca, Mg] balances.

- The software “CND-Citros” can assist citrus growers, agronomy engineers and technicians to diagnose the nutrient status of orange crops based on the proposed method, using the results of leaf chemical analysis. Plant diagnosticians should consider though, that corrective measures involving one nutrient might affect all balances connected with it in the closed system of the foliar tissue, requiring to validate the ensuing plant reaction to the new fertilization regimes. The softwares of other fruit crops available could also be revisited using the ML-balance concept elaborated in this research.

6. REFERENCES

AGRIANUAL 2018: anuário da agricultura brasileira (2017). São Paulo: FNP Consultoria & Comércio p. 237-244.

Aitchison J (1986) **The statistical analysis of compositional data**. Londres: Chapman and Hall. 416 p.

Aitchison J, Greenacre M (2002) Biplots of compositional data. **Journal of the Royal Statistical Society. Series C, Applied statistics** 51:375-392.

Alva AK, Mattos Jr. D, Paramasivam S, Patil B, Dou H, Sajwan K (2006) Potassium management for optimizing citrus production and quality. **International Journal of Fruit Science** 6:3-43.

Alva AK, Paramasivam S (1998) Nitrogen management for high yield and quality of citrus in sandy soils. **Soil Science Society of America Journal** 62:1335-1342.

Anderson CA (1987) Fruit yields, tree size, and mineral nutrition relationship in Valencia orange trees as affected by liming. **Journal of Plant Nutrition** 10:1907-1916.

Auler PAM, Neves CSVJ, Fidalski J, Pavan MA (2011) Calagem e desenvolvimento radicular, nutrição e produção de laranja 'Valência' sobre porta-enxertos e sistemas de preparo do solo. **Pesquisa Agropecuária Brasileira** 46:254-261.

Bacon-Shone J (2011) A short history of compositional data analysis. In.: Pawlowsky-Glahn V, Buccianti A (Eds.) **Compositional data analysis:** theory and applications. Nova Iorque: John Wiley and Sons, p. 3-11.

Baldock JO, Schulte EE (1996) Plant analysis with standardized scores combines DRIS and sufficiency range approaches for corn. **Agronomy Journal** 88:448-456.

Baligar VC, Bennett OL (1986a) Outlook on fertilizer use efficiency in the tropics. **Fertilizer Research** 10:83-86.

Baligar VC, Bennett OL (1986b) NPK-fertilizer efficiency – a situation analysis for the tropics**. Fertilizer Research** 10:147-164.

Barber S A (1995) **Soil Nutrient Bioavailability**. New York: Wiley. 414 p.

Bates TE (1971) Factors affecting critical nutrient concentrations in plant and their evaluation: a review. **Soil Science** 112:116-130.

Baxter I (2015) Should we treat the ionome as a combination of individual elements, or should we be deriving novel combined traits? **Journal of Experimental Botany** 66:2127-2131.

Beaufils ER (1973) **Diagnosis and recommendation integrated system (DRIS)**: a general scheme for experimentation and calibration based on principles developed from research in plant nutrition. Pietermararitzbur: University of Natal. 132 p. (Soil Science Bulletin, 1).

Behmann J, Mahlein AK, Rumpf T, Romer C, Plumer L (2014) A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. **Precision Agriculture** 16: 239-260.

Bernardi ACC, Carmello QAC, Carvalho AS (2000) Macronutrientes em mudas de citros cultivadas em vasos em resposta à adubação NPK. **Scientia Agricola** 57:761-767.

Beverly RB (1987) Modified DRIS method for simplified nutrient diagnosis of 'Valencia' oranges. **Journal of Plant Nutrition** 10:1401-1408.

Boaretto AE, Muraoka T, Boaretto RM (2003) Absorção e translocação de Mn, Zn e B aplicados via foliar em citros. **Laranja** 24:177-197.

Boaretto RM, Quaggio JA, Mattos Jr. D, Muraoka T, Boaretto AE (2011) Boron uptake and distribution in field grown citrus trees. **Journal of Plant Nutrition** 34:839-849.

Boaretto RM, Quaggio JA, Mourão Filho FAA, Giné MF, Boaretto AE (2008) Absorption and mobility of boron in young citrus plants. **Communications in Soil Science and Plant Analysis** 39:2501-2514.

Brown P, Shelp BJ (1997) Boron mobility in plants. **Plant Soil** 193:85-101.

Brungard CW, Boettinger JL, Duniway MC, Wills SA, Edwards Jr. TC (2015) Machine learning for predicting soil classes in three semi-arid landscapes. **Geoderma** 239-240: 68-83.

Camacho MA, Silveira MV, Camargo RA, Natale W (2012) Faixas normais de nutrientes pelos métodos ChM, DRIS e CND e nível crítico pelo método de distribuição normal reduzida para laranjeira-Pera. **Revista Brasileira de Ciência do Solo** 36:193-200.

Chaudhary A, Kolhe S, Kamal R (2016) A hybrid ensemble for classification in multiclass datasets: an application to oilseed disease dataset. **Computers and Electronics in Agriculture** 124:65-72.

Chessel D, Dufour AB, Dray S (2011) **ade4: Analysis of Ecological Data: Exploratory and Euclidean Methods in Environmental Sciences**. R package version 1.4-17. Available at: http://CRAN.R-project.org/package=ade4

Corá JE, Silva GO, Martins Filho MV (2005) Manejo do solo sob citros. In: Mattos Jr. D, Negri JD, Pio RM, Pompeu Jr. J (Eds.) **Citros**. Campinas: Instituto Agronômico, p. 345-368.

De Wit CT (1992) Resource use efficiency in agriculture. **Agricultural Systems** 40:125–151.

Deus JAL, Neves JCL, Corrêa MCM, Parent SÉ, Natale W, Parent LE (2018) Balance design for robust foliar nutrient diagnosis of “Prata” banana (Musa spp.). **Scientific Reports** (in press).

Dias JRM, Tucci CAF, Wadt PGS, Silva AM, Santos JZL (2013a) Níveis críticos e faixas de suficiência nutricional em laranjeira-pêra na Amazônia Central obtidas pelo método DRIS. **Acta Amazonica** 43:239-246.

Dias JRM, Wadt PGS, Tucci CAF, Santos JZL, Silva SV (2013b) Normas DRIS multivariadas para avaliação do estado nutricional de laranjeira ‘Pera’ no estado do Amazonas. **Revista Ciência Agronômica** 44:251-259.

Drummond ST, Sudduth KA, Joshi A, Birrell SJ, Kitchen NR (2002) Statistical and neural network methods for site-specific yield prediction. **Transactions of the American Society of Agricultural Engineers** 47:5-14.

Egozcue JJ, Pawlowsky-Glahn V (2011) Basic Concepts and Procedures. In.: Pawlowsky-Glahn V, Buccianti A (Eds.) **Compositional Data Analysis**: Theory and Applications. New York: John Wiley and Sons, p. 12–28.

Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. **Mathematical Geology** 37:795-828.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric log-ratio transformations for compositional data analysis. **Mathematical Geology** 35:279-300.

Embleton TW, Reitz HJ, Jones WW (1973) Citrus Fertilization. In.: Reuther W (Ed.) **The Citrus Industry**. Riverside: University of California, p. 122-182.

EMBRAPA (2013). **Sistema Brasileiro de Classificação de Solos**. Rio de Janeiro: Embrapa Solos. 356 p.

Ferreira JEV, Costa CHSC, Miranda RM, Figueiredo AF (2015) The use of K nearest neighbor method to classify the representative elements. **Educación química** 26:195-201.

Filzmoser P, Gschwandtner M (2013) '**Mvoutlier'**: Multivariate Outlier Detection Based on Robust Methods. R package.

Filzmoser P, Hron K (2011) Robust Statistical Analysis. In.: Pawlowsky-Glahn V, Buccianti A (Eds.) **Compositional Data Analysis**: Theory and Applications. New York: John Wiley and Sons, p. 59-72.

Filzmoser P, Hron K, Reimann C (2009) Univariate statistical analysis of environmental (compositional) data: problems and possibilities. **Science of the Total Environment** 407:6100–6108.

Filzmoser P, Hron K (2008) Outlier detection for compositional data using robust methods. **Mathematical Geosciences** 40:233-248.

Fleming PJ, Wallace JJ (1986) How not to lie with statistics: correct way to summarize benchmark results. **Communications of fhe ACM** 29:218-221.

Gautam R, Panigrahi S, Franzen D (2006) Neural network optimisation of remotely sensed maize leaf nitrogen with a genetic algorithm and linear programming using five performance parameters. **Biosystems Engineering** 95:359–370.

Goldberg S, Suarez DL (2011) Influence of soil solution cation composition on boron adsorption by soils. **Soil Science** 176:80-83.

Hastie T, Tibshirani R, Friedman JH (2001) **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. New York: Springer. 534 p.

Hernandes A, Souza HA, Amorim DA, Natale W, Lavres Jr. J, Boaretto AE, et al. (2014) DRIS norms for Pêra orange. **Communications in Soil Science and Plant Analysis** 45:2853-2867.

Hippler FWR, Boaretto RM, Dovis VL, Quaggio JA, Azevedo RA, Mattos Jr. D (2018) Oxidative stress induced by Cu nutritional disorders in Citrus depends on nitrogen and calcium availability. **Scientific Reports** 8:1641.

Hippler FWR, Boaretto RM, Quaggio JA, Azevedo RA, Mattos Jr. D (2015) Towards soil management with Zn and Mn: estimates of fertilization efficacy of citrus trees. **Annals of Applied Biology** 166:484-495.

Holland DA (1966) The interpretation of leaf analysis. **Journal of Horticultural Sciences** 41:311-329.

Jarrell WM, Beverly RB (1981) The dilution effect in plant nutrition studies. **Advances in Agronomy** 34:197-224.

Jones Jr. JB, Case VW (1990) Sampling, Handling, and Analyzing Plant Tissue Samples. In.: Westerman RL (Ed.) **Soil Testing and Plant Analysis**. Madison: Soil Science Society of America, p. 389-427.

Kampichler C, Wieland R, Calmé S, Weissenberger H, Arriaga-Weiss S (2010) Classification in conservation biology: a comparison of five machine-learning methods. **Ecological Informatics** 5:441-450.

Kato T (1986) Nitrogen metabolism and utilization in citrus. **Horticultural Reviews** 8:181-216.

Khiari L, Parent LE, Tremblay N (2001a) Critical compositional nutrient indexes for sweet corn at early growth stage. **Agronomy Journal** 93:809-814.

Khiari L, Parent LE, Tremblay N (2001b) Selecting the high-yield subpopulation for diagnosing nutrient imbalance in crops. **Agronomy Journal** 93:802-808.

Khiari L, Parent LE, Tremblay N (2001c) The phosphorus compositional nutrient diagnosis range for potato. **Agronomy Journal** 93:815-819.

Kirkwood C, Cave M, Beamish D, Grebby S, Ferreira A (2016) A machine learning approach to geochemical mapping. **Journal of Geochemical Exploration** 167:49-61.

Kuhn, M (2008) Building predictive models in R using the caret package. **Journal of Statistical Software** 28:1-26.

Kumar PSS, Geetha SA, Savithri P, Mahendran PP, Ragunath KP (2003) Evaluation of DRIS and CND indexes for effective nutrient management in Muscat grapevines (*Vitis vinifera*). **Journal of Applied Horticulture** 5:76-80.

Langley P, Simon HA (1995) Applications of machine learning and rule induction. **Communications of fhe ACM** 38:55-64.

Lea-Cox JD, Syvertsen JP (1996) How nitrogen supply affects growth and nitrogen uptake, use efficiency, and loss from Citrus seedlings. **Proceedings of the American Society for Horticultural Science** 121:105-114.

Li H, Lee WS, Wang K (2014) Identifying blueberry fruit of different growth stages using natural outdoor color images. **Computers and Electronics in Agriculture** 106:91-101.

Liu GD, Wang RD, Wu LS, Peng SA, Wang YH, Jiang CC (2012) Boron distribution and mobility in navel orange grafted on citrange and trifoliate orange. **Plant and Soil** 360:123-133.

Liu G, Dong X, Liu L, Wu L, Peng S, Jiang C (2014) Boron deficiency is correlated with changes in cell wall structure that lead to growth defects in the leaves of navel orange plants. **Scientia Horticulturae** 176:54-62.

Loladze I, Elser JJ (2011) The origins of the redfield nitrogen-to-phosphorus ratio are in a homoeostatic protein-to-rRNA Ratio. **Ecology Letters** 14:244-250.

Ma C, Zhang HH, Wang X (2014) Machine learning for big data analytics in plants. **Trends in Plant Science** 19:798-808.

Malavolta E. (2006) **Manual de Nutrição Mineral de Plantas**. São Paulo: Agronômica Ceres. 638 p.

Marchand S, Parent SE, Deland JP, Parent LE (2013) Nutrient signature of Quebec (Canada) cranberry (*Vaccinium macrocarpon* AIT.). **Revista Brasileira de Fruticultura** 35:199-209.

Marques DJ, Broetto F, Silva EC, Carvalho JG (2011) Dinâmica de cátions na raiz e folhas de berinjela cultivada sobre doses crescentes de potássio oriundas de duas fontes. **Idesia** 29:69-77.

Marschner P (2011) **Mineral Nutrition of Higher Plants**. London: Academic Press. 672 p.

Mattos Jr. D, Quaggio JA, Cantarella H, Boaretto RM, Zambrosi FCB (2012). Nutrient management for high citrus fruit yield in tropical soils. **Better Crops** 96:4-7.

Mattos Jr. D, Quaggio JA, Cantarella H, Alva AK (2003b) Nutrient content of biomass components of hamlin sweet orange trees. **Scientia Agricola** 60:155-160.

Mattos Jr. D, Quaggio JA, Cantarella H, Carvalho SA (2004) Superfícies de resposta do tangor ‘Murcott’ à fertilização com N, P e K. **Revista Brasileira de Fruticultura** 26:164-167.

Mattos Jr. D, Boaretto RM, Quaggio JA (2012) Diagnose foliar na cultura do citros. In: Prado RM (Ed.) **Nutrição de plantas**: diagnose foliar em frutíferas. Jaboticabal: FCAV/CAPES/FAPESP/CNPq, p. 175-198.

Mattos Jr. D, Graetz DA, Alva AK (2003a) Biomass distribution and Nitrogen-15 partitioning in citrus trees on a sandy Entisol. **Soil Science Society of America Journal** 67:555-563.

Mattos Jr. D, Negri JD, Pio RM, Pompeu Jr. J (2005) **Citros**. Campinas: Instituto Agronômico e Fundag. 929 p.

Mattos Jr. D, Hippler FWR, Boaretto RM, Stuchi ES, Quaggio JA (2017) Soil boron fertilization: The role of nutrient sources and rootstocks in citrus production. **Journal of Integrative Agriculture** 16:1609-1616.

McQueen RJ, Garner SR, Nevill-Manning CG, Witten IH (1995) Applying machine learning to agricultural data. **Computers and Electronics in Agriculture** 12: 275-293.

Mesquita GL, Zambrosi FCB, Tanaka FAO, Boaretto RM, Quaggio JA, Ribeiro RV, et al. (2016) Anatomical and physiological responses of Citrus trees to varying boron availability are dependent on rootstock. **Frontiers in Plant Science** 7:224.

Modesto VC, Parent SE, Natale W, Parent LE (2014) Foliar nutrient balance standards for maize (*Zea mays* L.) at high-yield level. **American Journal of Plant Sciences** 5: 497-507.

Mourão Filho FAA (2005) DRIS and sufficient range approaches in nutritional diagnosis of “Valencia” sweet orange on three rootstocks. **Journal of Plant Nutrition** 28:691-705.

Mourão Filho FAA (2004) Dris: concepts and applications on nutritional diagnosis in fruit crops. **Scientia Agricola** 61:550-560.

Mourão Filho FAA, Azevedo JC (2003) DRIS norms for Valencia sweet orange on three rootstocks. **Pesquisa Agropecuária Brasileira** 38:85-93.

Natarajan R, Subramanian J, Papageorgiou EI (2016) Hybrid learning of fuzzy cognitive maps for sugarcane yield classification. **Computers and Electronics in Agriculture** 127:147-157.

Novais RF (1977) **Phosphorus supplying capacity of previously heavily fertilized soils**. Doctoral thesis - North Carolina State University, Raleigh.

Nowaki RHD, Parent SE, Cecílio Filho AB, Rozane DE, Meneses NB, Santos da Silva JA, et al. (2017). Phosphorus over-fertilization and nutrient misbalance of irrigated tomato crops in Brazil. **Frontiers in Plant Science** 8:825.

Olsen SR (1972) Micronutrients Interactions. In.: Mortvedt JJ, Giordano PM, Lindsay WL (Eds.) **Micronutrients in Agriculture**. Madison: SSSA, p. 243-264.

Pantazi XE, Moshou D, Alexandridis T, Whetton RL, Mouazen AM (2016) Wheat yield prediction using machine learning and advanced sensing techniques. **Computers and Electronics in Agriculture** 121: 57-65.

Parent SÉ, Parent LE (2017). Balance designs revisit indices commonly used in agricultural science and eco-engineering. In: THE 7TH INTERNATIONAL WORKSHOP ON COMPOSITIONAL DATA ANALYSIS. **Abstracts…**Abbadia San Salvatore, p. 195-227. Disponível em <http://www.compositionaldata.com/codawork2017/ proceedings/ProceedingsBook2017_May30.pdf >

Parent LE (2011) Diagnosis of the nutrient compositional space of fruit crops. **Revista Brasileira de Fruticultura** 33:321–334.

Parent LE, Dafir M (1992) A theoretical concept of compositional nutrient diagnosis. **Journal of the American Society for Horticultural Science** 117:239-242.

Parent LE, Isfan D, Tremblay N, Karam A (1994) Multivariate nutrient diagnosis of the carrot crop. **Journal of the American Society for Horticultural Science** 119:420-426.

Parent LE, Natale W (2008) CND: vantagens e benefícios em culturas de alta produtividade. In.: Prado RM et al. (Eds.) **Nutrição de plantas**: Diagnose foliar em grandes culturas. Jaboticabal: Funep, p. 105-114.

Parent LE, Natale W, Ziadi N (2009) Compositional nutrient diagnosis of corn using the mahalanobis distance as nutrient imbalance index. **Canadian Journal of Soil Science** 89:383-390.

Parent SE, Parent LE, Rozane DE, Hernandes A, Natale W (2012) Nutrient Balance as Paradigm of Soil and Plant Chemometrics. In.: Issaka RN (Ed.) **Soil Fertility**. InTech, p. 83-114.

Parent SE, Parent LE, Egozcue JJ, Rozane DE, Hernandes A, Lapointe L, Hébert-Gentile V, Naess K, Marchand S, Lafond J, Mattos Jr. D, Barlow P, Natale W (2013a) The plant ionome revisited by the nutrient balance concept. **Frontiers in Plant Science** 4:1-10.

Parent SE, Parent LE, Rozane DE, Natale W (2013b) Plant ionome diagnosis using sound balances: case study with mango (*Mangifera Indica*). **Frontiers in Plant Science** 4:1-12.

Partelli FL, Dias JRM, Vieira HD, Wadt PGS, Paiva Jr. E (2014) Avaliação nutricional de feijoeiro irrigado pelos métodos CND, DRIS e faixas de suficiência. **Revista Brasileira de Ciência do Solo** 38:858-866.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2011) Principal Balances. In: 4th INTERNATIONAL WORKSHOP ON COMPOSITIONAL DATA ANALYSIS, eds. J. Egozcue, R. Tolosana-Delgado and M. Ortego (Sant Feliu de Guixols, GI: Codawork), 1-10.

Pietersma D, Lacroix R, Lefebvre D, Wade KM (2003) Performance analysis for machine-learning experiments using small data sets. **Computers and Electronics in Agriculture** 38:1-17.

Quaggio JA, Cantarella H, van Raij B (1998) Phosphorus and potassium soil test and nitrogen leaf analysis as a basis for citrus fertilization. **Nutrient Cycling in Agroecosystems** 52:67-74.

Quaggio JA, Mattos Jr. D, Cantarella H (2005) Manejo da Fertilidade do Solo na Citricultura. In.: Mattos Jr. D, De Negri JD, Pio RM, Pompeu Jr. J (Eds.). **Citros**. Campinas: Instituto Agronômico/Fundag, p. 485-507.

Quaggio JA, Mattos Jr. D, Cantarella H (2006) Fruit yield and quality of sweet oranges affected by nitrogen, phosphorus and potassium fertilization in tropical soils. **Fruits** 61: 293-302.

Quaggio JA, Mattos Jr. D, Cantarella H, Almeida ELE, Cardoso SAB (2002) Lemon yield and fruit quality affected by NPK fertilization. **Scientia Horticulturae** 96:151-162.

Quaggio JA, van Raij B, Malavolta E (1985) Alternative use of the SMP-buffer solution to determine lime requirement of soils. **Communications in Soil Science and Plant Analysis** 16:245-260.

Quaggio JA, Mattos Jr. D, Boaretto RM (2010) Citros. In: Prochnow LI, Casarin V, Stipp SR (Eds.). **Boas práticas para uso eficiente de fertilizantes**: culturas. Piracicaba: IPNI, p. 371-409.

Quaggio JA, Mattos Jr. D, Boaretto RM (2011) Sources and rates of potassium for sweet orange production. **Scientia Agricola** 68:369-375.

Quaggio JA, Mattos Jr. D; Cantarella H, Tank Jr. A (2003) Fertilização com boro e zinco no solo em complementação à aplicação via foliar em laranjeira Pêra. **Pesquisa Agropecuária Brasileira** 38: 627-634.

Quaggio JA (1996) Análise de solo para citros: métodos e critérios para interpretação. In: Donadio LC, Baumgartner JG (Coord.) SEMINÁRIO INTERNACIONAL DE CITROS: nutrição, 4. Bebedouro. Campinas: Fundação Cargill, p. 95-114.

Quaggio JA (1991) **Resposta da laranjeira Valência à calagem e ao equilíbrio de bases num latossolo de textura argilosa**. 107 f. Tese (Doutorado em Solos e Nutrição de Plantas) – USP, Piracicaba.

Quaggio JA, Teófilo Sobrinho J, Dechen AR (1992b) Response to liming of 'Valencia' orange tree on Rangpur lime: effects of soil acidity on plant growth and yield. **Proceedings of the International Society of Citriculture** 2:628-632.

R Core Team (2017). **R: A Language and Environment for Statistical Computing.** Vienna: R Foundation for Statistical Computing.

Rencher AC (2002). **Methods of multivariate analysis** (2nd ed.). New York: John Wiley & Sons. 714 p.

Rodríguez O, Rojas E, Sumner M (1997) Valencia orange DRIS norms for Venezuela. **Communications in Soil Science and Plant Analysis** 28:1461-1468.

Romero JR, Roncallo PF, Akkiraju PC, Ponzoni I, Echenique VC, Carballido JA (2013) Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires. **Computers and Electronics in Agriculture** 96:173-179.

Roy ED, Richards PD, Martinelli LA, Coletta LD, Lins SRM, Vasquez FF, et al. (2016) The phosphorus cost of agricultural intensification in the tropics. **Nature Plants** 43:1-6.

Rozane DE, Mattos Jr. D, Parent SE, Natale W, Parent LE (2015) Meta-analysis in the selection of groups in varieties of Citrus. **Communications in Soil Science and Plant Analysis** 46:1948-1959.

Rozane DE, Natale W, Parent LE, Parent SE, Souza HA, Amorim DA (2013) População de referência no diagnóstico da composição nutricional (CND) em mangueiras. In: CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO, 34., 2013, Florianópolis. **Resumos...** Florianópolis: Epagri/SBCS, 2013. p. 1-3.

Scott IM, Lin W, Liakata M, Wood JE, Vermeer CP, Allaway D, et al. (2013) Merits of random forests emerge in evaluation of chemometric classifiers by external validation. **Analytica Chimica Acta** 801:22-33.

Serra AP, Marchetti ME, Vitorino ACT, Novelino JO, Camacho MA (2010b) Determinação de faixas normais de nutrientes no algodoeiro pelos métodos CHM, CND e DRIS. **Revista Brasileira de Ciência do Solo** 34:105-113.

Shekoofa A, Emam Y, Shekoufa N, Ebrahimi M, Ebrahimie E (2014) Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithm : a new avenue in intelligent agriculture. **PLoS ONE** 9:e97288.

Silva GGC, Neves JCL, Alvarez VVH, Leite FP (2005) Avaliação da universalidade das normas DRIS, M-DRIS e CND. **Revista Brasileira de Ciência do Solo** 29 :755-761.

Silva GGC, Neves JCL, Alvarez VVH, Leite FP (2004) Nutritional diagnosis for eucalipto by DRIS, M-DRIS, and CND. **Scientia Agricola** 61:507-515.

Smith G (1985) **Kiwifruit Nutrition:** Diagnosis of Nutritional Disorders. Wellington North: Southern Horticulture, Agpress.

Smith GS, Asher GJ, Clark CJ (1997) **Kiwi fruit**: nutrition diagnosis of nutritional disorders. Auckland: Horticulture and Food Research Institute.

Smith P (1969) Nitrogen stress and premature leaf abscission in Citrus. **HortScience** 4:326-327.

Souza HA, Parent SE, Rozane DE, De Amorim DA, Modesto VC, Natale W, et al. (2016) Guava waste to sustain guava (Psidium guajava) agroecosystem: nutrient “balance” concepts. **Frontiers in Plant Science** 7:1252.

Sumner ME, Farina MPW (1986) Phosphorus interaction with other nutrients and lime in field cropping systems. **Advances in Soil Science** 5:201-236.

Swets JA (1988) Measuring the accuracy of diagnostic systems. **Science** 240:1285-1293.

Ulrich A, Hills FJ (1967) Principles and practices of plant analysis. In: **Soil testing and plant analysis**. Madison: SSSA, p.11-24. (Special Publications Series).

UNESP (2013a) Universidade Estadual Paulista “Júlio de Mesquita Filho” (São Paulo, Brasil); Université Laval (Ville de Québec, Canadá). Rozane DE, Natale W, Parent LE, Parent SE, Santos EMH. **CND-Goiaba**. BR5120130003792, 18 abr.

UNESP (2013b) Universidade Estadual Paulista “Júlio de Mesquita Filho” (São Paulo, Brasil); Université Laval (Ville de Québec, Canadá). Rozane DE, Natale W, Parent LE, Parent SE, Santos, EMH. **CND-Manga**. BR5120130003806, 18 abr.

Urano EOM, Kurihara CH, Maeda S, Vitorino ACT, Gonçalves MC, Marchetti ME (2007) Determinação de teores ótimos de nutrientes em soja pelos métodos chance matemática, sistema integrado de diagnose e recomendação e diagnose da composição nutricional. **Revista Brasileira de Ciência do Solo** 31:63-72.

van den Boogaart KG, Tolosana-Delgado R, Bren M (2014). **Compositions**: Compositional Data Analysis in R package. Version 1.40-1. Available online at: https://cran.r-project.org/web/packages/compositions/index.html van Raij B, Cantarella H, Quaggio JA, Furlani AMC (1997) **Recomendações de Adubação e Calagem para o Estado de São Paulo**. Campinas: Instituto Agronômico p. 133-136.

Wadt PGS, Anghinoni I, Guindani RHP, Lima AST, Puga AP, Silva GS, Prado RM (2012) Padrões nutricionais para lavouras arrozeiras irrigadas por inundação pelos métodos da CND e Chance Matemática. **Revista Brasileira de Ciência do Solo** 37:145-156.

Walworth JL, Sumner ME (1987) The diagnosis and recommendation integrated system (DRIS). In.: Stewart BA (Ed.) **Advances in soil science** New York: Springer, p. 149-188.

Wang N, Yang C, Pan Z, Liu Y, Peng S (2015). Boron deficiency in woody plants: Various responses and tolerance mechanisms. **Frontiers in Plant Science** 6:916.

Weiss SM, Kulikowski CA (1991) **Computer Systems that Learn:** Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. San Mateo: Morgan Kaufmann. 223 p.

Witten IH, Frank E, Hall MA (2011) **Data Mining:** Practical Machine Learning Tools and Techniques. Burlington: Morgan Kaufmann. 664 p.

Witten IH, Frank E (2000) **Data Mining**: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann. 371 p.

Zambrosi FCB, Mattos Jr. D, Quaggio JA, Cantarella H, Boaretto RM (2013) Phosphorus uptake by young citrus trees in low-P soil depends on rootstock varieties and nutrient management. **Communications in Soil Science and Plant Analysis** 44:2107-2117.