Modèle de fertilisation NPK localisé de la pomme de terre (Solanum tuberosum L.) au Québec

Thèse

Zonlehoua Coulibali

Doctorat en sols et environnement
Philosophiæ doctor (Ph. D.)

Québec, Canada

Modèle de fertilisation NPK localisé de la pomme de terre (*Solanum tuberosum* L.) au Québec

Thèse

Zonlehoua COULIBALI

Sous la direction de :

Serge-Étienne Parent

Athyna N. Cambouris

Résumé

Les cultures à haute valeur ajoutée comme la pomme de terre (*Solanum tuberosum* L.), sont de bons candidats pour l'adoption de l'agriculture de précision en raison des coûts de production particulièrement élevés. Les quantités de fertilisants requises alimentent le défi permanent de la recherche de l'optimisation de fertilisation spécifique à chaque agroécosystème. La modélisation fournit une trousse d'outils pour l'aide à la décision. En ce qui concerne la fertilisation, le rendement est habituellement relié à des doses variables d'un fertilisant à l'aide de fonctions simples (quadratique, linéaire ou linéaire-quadratique, Mitscherlich ou autres). Même si ces fonctions ne devraient être utilisées que pour décrire le comportement des données expérimentales, elles ont été largement utilisées pour prédire les doses optimales de fertilisants. Ce projet de recherche a proposé un modèle de recommandation des doses optimales d'azote (N), de phosphore (P) et de potassium (K) pour la culture de pomme de terre en exploitant les techniques d'autoapprentissage. Dans une première partie, il a d'abord été question de regrouper les cultivars sur la base de la composition chimique de la feuille diagnostique en utilisant une classification non supervisée. Ce regroupement a permis de montrer que les cultivars étudiés sont associés à une composition ionomique spécifique. Ensuite, dans une perspective de prédiction de catégories de rendement en fonction de la composition foliaire, les algorithmes des k plus proches voisins (KNN), des forêts aléatoires (RF) et des machines à vecteurs de supports (SVM) ont montré un potentiel de diagnostic acceptable, avec une précision de 70 %, pour détecter un déséquilibre nutritionnel en cours de saison. Enfin, le vecteur de perturbation de l'espace de composition d'Aitchison pourrait être un bon indicateur pour détecter la présence et l'ampleur d'un déséquilibre nutritionnel en cours de saison. Dans la deuxième partie, les modèles d'autoapprentissage utilisant les algorithmes des KNN, des RF, des réseaux neuronaux (NN) et des processus gaussiens (PG), ont prédit le rendement et le poids spécifique en fonction des conditions expérimentales de façon pratiquement similaire avec des coefficients de détermination (R^2) supérieurs à celui du modèle de Mitscherlich. Les R^2 étaient de 0,52 pour les KNN, de 0,59 pour les RF, de 0,49 pour les NN, de 0,58 pour les PG et de 0,37 pour le modèle de Mitscherlich. Les R^2 des modèles de prédiction de la balance des tubercules de taille moyenne (R^2 = 0,60 – 0,69) et du poids spécifique (R^2 = 0,58 – 0,67) étaient plus élevés comparés aux R^2 de la balance des tubercules de grande taille (R^2 = 0,55 – 0,64) et du rendement vendable. Des dissemblances importantes sont apparues entre les modèles dans le rendu des courbes de

réponse et la prédiction des doses agroéconomiques optimales de N, P et K. Ces prédictions étaient spécifiques au site. Les processus gaussiens étaient plus appropriés en raison de leur capacité d'élaborer des surfaces de réponse lisses et des recommandations probabilistes.

Abstract

High-value crops, like potato (*Solanum tuberosum* L.), are good candidates for the adoption of Precision Agriculture because of the high cost of inputs. The large amount of potato fertilizers requirement makes it economically and environmentally important for producers to determinate site-specific fertilizers dosages. Crop performance responses to fertilizer inputs have yet been modeled using simple functions like quadratic, linear-plateau or Mitscherlich. Even though they should only be used to describe experimental data, such models are used to predict optimal fertilizer doses considering the cost of the fertilizer and crop sales. As large amounts of data are being assembled in repeated observational data sets, *machine learning* models can become useful to predict and detect patterns in data without hardly presuming how a response curve should behave. This project generated models recommending optimal economic doses of nitrogen (N), phosphorus (P) and potassium (K) for potato crops using *machine learning* techniques. We assessed the validity of cultivar grouping as new predictive feature, and predicted potato tuber yields using foliar ionomes. A density-based clustering algorithm (dbscan) failed to delineate groups of high-yield cultivars linked to specific cultivar ionomic composition. Algorithms of *k*-nearest neighbors (KNN), random forests (RF) and support vector machines (SVM) showed a fair diagnostic potential to detect in-season nutritional imbalance, with practically similar accuracy of 70%. The perturbation vector of Aitchison compositional space appeared a good indicator in detecting the presence and magnitude of a nutritional imbalance. Moreover, machine-learning models using KNN, RF, neural networks (NN) and Gaussian processes (GP) algorithms returned almost similar coefficients of determination (R^2) superior to that of the Mitscherlich model. The R^2 values were 0.52, 0.59, 0.49 and 0.58 respectively for the KNN, RF, ANN, and GP, and 0.37 with the Mitscherlich model to predict marketable yield. The models were somewhat more efficient to predict medium-size tubers (R^2 = 0.60 – 0.69) and tuber specific gravity (R^2 = 0.58 – 0.67) than large-size tubers (R^2 = 0.55 – 0.64) and marketable yield. Disagreements appeared between models in site-specific response curves and optimal economic or agronomic N, P, and K doses prediction. However, GP models stood up as the most promising algorithm due to its built-in ability to develop smooth response surfaces and recommendations within a probabilistic risk-assessment framework.

Table des matières

Liste des tableaux

Liste des figures

Liste du matériel complémentaire

Liste des abréviations

AAC	Agriculture et Agroalimentaire Canada.
AQSSS	Association Québécoise des Spécialistes en Sciences du Sol.
BD	Base de Données.
Ca	Calcium.
CFIA	Canadian Food Inspection Agency (Agence Canadienne d'inspection des aliments).
CRAAQ	Centre de Reference en Agriculture et Agroalimentaire du Québec.
CRD	Centre de Recherche et de Développement.
CUN	Coefficient d'Utilisation de l'azote (N)
DO	Dose Optimale.
GP	Gaussian Processes (Processus gaussiens).
GPS	Global positioning system (système mondial de positionnement)
IPC	International Potato Center.
ISP_1	Indice de saturation en phosphore, rapport (P/Al)-Mehlich-3
K	Potassium.
KNN	k-Nearest Neighbors (*k* plus proches voisins).
MAPAQ	Ministère de l'Agriculture, des Pêcheries et de l'Alimentation du Québec.
Mg	Magnésium.
N	Azote.
NN	Neural Networks (Réseaux de neurones artificiels ou réseaux neuronaux).
NPK	Combinaison azote, phosphore, potassium.

P	Phosphore.
PG	Processus Gaussien (ou GP Gaussian Process en anglais)
R^2	Coefficient de détermination.
RF	Random Forest (forêts aléatoires).
RR	Rendement relatif.
SDI	Shannon Diversity Index (Indice de diversité de Shannon).
SIG	Systèmes d'Information Géographique
SVM	Support Vector Machines (machines à vecteur de support).

Remerciements

La réalisation de ce projet de doctorat a été rendue possible grâce au support inconditionnel de bon nombre de personnes qu'il est indispensable de remercier. Ainsi, je voudrais remercier tous ceux et toutes celles qui, de façon directe ou indirecte, ont contribué à sa réussite.

Je remercie mon directeur de recherche, le professeur Serge-Étienne Parent et ma co-directrice de recherche la Dre Athyna N. Cambouris chercheuse à Agriculture et Agroalimentaire Canada au Centre de Recherche et de Développement de Québec, pour les orientations, la patience et la disponibilité à toutes les étapes du travail. Les suggestions en programmation et graphisme (dans R et Python) et les échanges sur les concepts et techniques d'analyse de données compositionnelles du professeur Serge-Étienne Parent ont particulièrement marqué ce parcours. Mes sincères remerciements pour votre rigueur scientifique et votre inspiration.

Je remercie également le pré-lecteur de cette thèse, le Dr Marc-Olivier Gasser, chercheur en gestion des engrais organiques, des sols et de l'eau à l'Institut de Recherche et de Développement en Agroenvironnement (IRDA) à Québec. Vos commentaires et suggestions ont contribué à l'amélioration de cette thèse.

Je remercie également la grande participation de nombreux collaborateurs en l'occurrence Élizabeth Parent pour son laborieux travail en laboratoire, le professeur Michael Leblanc pour son aide dans la description des profils de sols des sites d'essais, toute l'équipe du professeur Léon-Étienne Parent, tous les conseillers ou producteurs des sites d'essais du projet, mon épouse et ma fille, pour leur aide indispensable dans la réalisation des travaux sur le terrain et la récolte des données.

J'exprime ma plus profonde gratitude aux professeurs Lotfi Khiari et Léon-Étienne Parent de l'Université Laval dont le premier a permis mon inscription à l'Université Laval pour un projet de maîtrise et le second m'a accordé ce projet de doctorat. Si ces projets étaient à refaire, je les ferais à vos côtés.

Enfin, merci à ma petite famille dont mon épouse Katiéné Blandine COULIBALY, nos filles Nongnimé Nina et Manitia Victoire COULIBALY, pour avoir su gérer mes périodes de stress lorsque le contrôle m'échappait. Ce travail a donc été une affaire de famille.

Avant-propos

Ce projet de thèse est partiellement financé par une subvention de recherche et de développement coopératif du Conseil de Recherche en Sciences Naturelles et en Génie du Canada (CRDPJ 385199-09 et DG-2254), du Ministère de l'Agriculture, des Pêcheries et de l'Alimentation du Québec (IA216581), du Centre SÈVE, et la contribution des producteurs ou organisation de producteurs partenaires : Patate Dolbec Inc. (St-Ubalde, QC), Groupe Gosselin FG (St-Augustin-de-Desmaures, QC), Agriparmentier Inc. (Notre-Dame-du-Bon-Conseil, QC), Ferme Daniel Bolduc Inc. (Péribonka, QC), Patate Laurentienne (Notre-Dame-de-la-Paix, QC), Ferme Bergeron-Niquet (Péribonka, QC) et Patates Lac-St-Jean (Péribonka, QC).

Ce document est divisé en trois chapitres et une conclusion générale. Deux des chapitres sont des articles insérés rédigés en anglais.

Le premier chapitre est une revue introductive sur le sujet traité. Ce chapitre situe le sujet de thèse dans son contexte à travers une revue de littérature. Il présente la culture de pomme de terre au Québec et en Amérique du nord, l'importance de la fertilisation en N, P et K et d'autres facteurs d'impact tels que le matériel végétal (la génétique), la météorologie et les propriétés du sol et du sous-sol. Les méthodes de détermination des optima de fertilisation et l'avenue des techniques d'apprentissage automatique explorées pour ce travail y sont également résumées. Ce chapitre définit enfin les objectifs et les hypothèses du projet de recherche.

Le chapitre 2 est un article publié dans le journal *PlosOne* et porte sur le diagnostic du statut nutritionnel des cultivars de pomme de terre en cours de culture. Le chapitre 3 est aussi un article publié dans le journal *PlosOne* et propose un outil d'aide à la décision dans la détermination des doses optimales de N, de P et de K spécifiques au site pour un rendement élevé et des tubercules de qualité. La dernière partie est une conclusion générale suivie de recommandations. Le lecteur notera cependant des modifications dans les légendes des figures et le style bibliographique comparées aux articles. Les instructions du journal ont été modifiées pour uniformiser la thèse. Toutes les figures sont insérées dans le texte au lieu d'être renvoyées à la fin.

Je suis l'auteur principal des articles insérés dans la thèse. J'ai participé à certaines activités de laboratoire, mais la plupart des analyses ont été réalisées par Élizabeth Parent, membre de l'équipe de recherche du professeur Léon-Étienne Parent. J'ai extrait la grande partie des données de la base de données des essais de fertilisation de la pomme de terre élaborée dans le cadre d'un projet antérieur du professeur Léon-Étienne Parent (CRDPJ 385199 09). Des données de 12 essais de fertilisation potassique ont été reçues du MAPAQ et ajoutées à la base de données. J'ai enfin coordonné la mise en place, le suivi et la récolte de données de 17 essais N, P et K en 2016 et 2017. La description des profils de sols des sites d'essais de fertilisation a été réalisée avec l'aide du professeur Michael Leblanc alors expert indépendant. Je suis le responsable des analyses statistiques, de l'interprétation et la discussion des résultats, puis de la rédaction des chapitres, sous la direction de mon directeur et de ma codirectrice de recherche. Les co-auteurs des articles sont le professeur Serge-Étienne Parent de l'Université Laval et la Dre Athyna Cambouris d'Agriculture et Agroalimentaire Canada au Centre de Recherche et de Développement de Québec.

Les résultats de ces études ont été présentés en tout ou en partie aux colloques annuels de la pomme de terre en novembre 2016 et 2018 à Lévis (Québec, Canada) et au 33ème congrès annuel de l'Association Québécoise des Spécialistes en Sciences du Sol (AQSSS) en juin 2019 à Duhamel-Ouest en Abitibi-Témiscamingue (Québec, Canada).

Introduction

Les facteurs pouvant influencer le développement, la croissance et une production agricole en quantité et qualité sont en perpétuelle exploration (Cerrato et Blackmer, 1990; Kooman et Haverkort, 1995; Fortin et al., 2011; Morissette et al., 2016). Leur optimisation passe entre autres, par la modélisation c'est-à-dire la représentation mathématique des mécanismes qui régissent les phénomènes naturels partiellement contrôlés ou compris (Tedeschi, 2006). La modélisation revêt trois utilités majeures en agriculture (Angus et al., 1993). La première est qu'elle rend possible les analyses de sensibilité. Le modèle de production peut en effet générer des surfaces de réponse selon les rendements et les prix en présence d'incertitudes. Le producteur peut ainsi voir les conséquences en termes de rendement et de revenus, d'une augmentation ou une réduction de ses intrants selon différentes conditions saisonnières. La deuxième utilité est qu'elle permet d'élaborer des recommandations stratégiques dans un cadre probabiliste. Les modèles de simulation peuvent être exécutés sur de nombreuses saisons pour générer des distributions de profits. Le producteur peut ainsi évaluer les chances d'obtenir des rendements donnés à différents niveaux d'application d'un fertilisant. Enfin, la modélisation offre la possibilité de faire des prédictions tactiques au fur et à mesure que la saison progresse. Le test de nitrate en post levée (Ziadi et al., 2012) et les tests de tissus végétaux peuvent par exemple être utilisés, mais la réponse des derniers reste limitée sur la quantité de fertilisant à appliquer (Angus et al., 1993).

Au Québec, la recommandation de fertilisants se fait en fonction du type de culture, de la biodisponibilité des éléments nutritifs dans le sol et, pour certaines cultures, du groupe textural de la couche de surface de sol (Parent et Gagné, 2010). Dans une approche holistique, la détermination des besoins en fertilisants doit tenir compte du potentiel de croissance spécifique qu'offre aux cultures un site, une zone d'aménagement ou une région. Ce potentiel est lié à un ensemble d'éléments comme les propriétés pédologiques, la qualité des sols, les pratiques culturales et les conditions climatiques (Leblanc, 2016). Ce qui revient à considérer la complexité des interactions entre les conditions environnementales locales, la génétique et le mode de gestion (Qin et al., 2018).

Une base de données (BD) a été élaborée pour documenter les essais agronomiques sur la fertilisation de la pomme de terre menés au Québec depuis 1970 (projet CRDPJ 385199 09). Cette BD a servi à sélectionner un ensemble de facteurs ayant un impact sur le rendement des tubercules en lien avec la fertilisation azotée (Parent et al., 2017). Les facteurs identifiés étaient entre autres, les cultivars préalablement classés selon la durée pour atteindre la maturité, le précédent cultural et des indices calculés des propriétés du sol, du sous-sol et du climat. Le modèle de Parent et al. (2017) n'a pas été testé indépendamment des sites sur lesquels il a été ajusté et son coefficient d'ajustement a été jugé trop faible (R^2 = 0,47 sans les effets aléatoires) pour faire des recommandations spécifiques aux conditions locales.

Les systèmes d'acquisition de données soutiennent le développement d'outils qui permettent de détecter des structures complexes dans des bases de données (Parent et al., 2017) et d'appuyer les décisions dans les exploitations agricoles (Wolfert et al., 2017). Ces données peuvent être des résultats d'expériences rassemblées (Fixen, 2014) pour faire des prédictions et pour prendre des décisions opérationnelles en temps réel (Sabarina et Priya, 2015; Lokers et al., 2016; O'Grady et O'Hare, 2017; Wolfert et al., 2017).

Ce projet de doctorat évalue le potentiel de prédiction de la réponse du rendement et du poids spécifique des tubercules de pomme de terre en fonction des caractéristiques génétiques, agronomiques, pédoclimatiques et du mode de gestion de la fertilisation, et est globalement schématisé à la Figure 0.1.

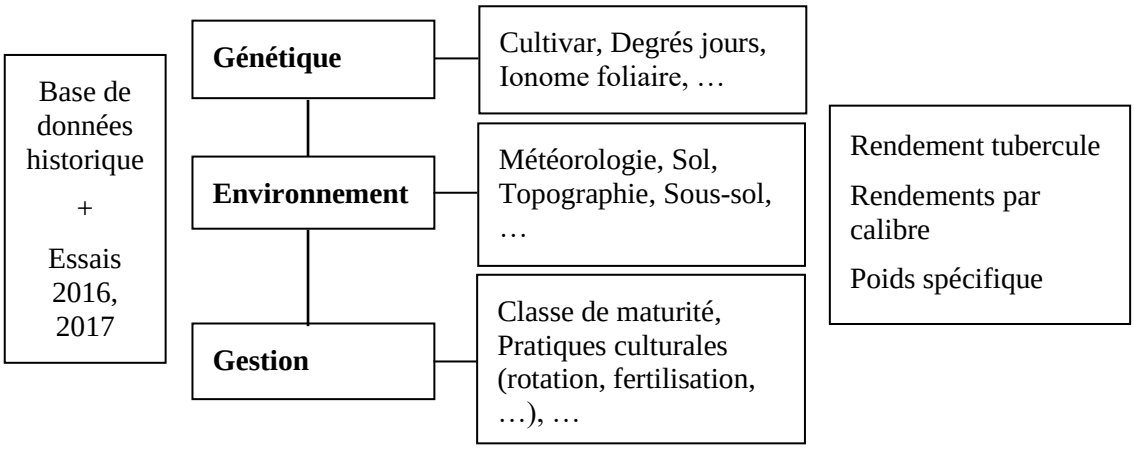


Figure 0.1 : Schéma descriptif du type de variables mises en relation dans l'élaboration du modèle de gestion de la fertilisation N, P, K en production de pomme de terre

Le projet fait appel à l'analyse de données compositionnelles dont le concept est largement traité par Parent et al. (2013a), et à la modélisation prédictive utile pour l'avancement des connaissances agronomiques particulièrement en agriculture de précision. Ce premier chapitre introduit des concepts et des connaissances générales sur la culture de pomme de terre et la modélisation, puis énonce les hypothèses et objectifs du projet.

Chapitre 1 : Revue de littérature

1.1 La culture de pomme de terre au Québec et en Amérique du Nord

1.1.1 Production et productivité

La pomme de terre est une herbacée de la famille des solanacées, vivace par ses tubercules, mais cultivée en culture annuelle (Rossignol et Rousselle-Bourgeois, 1996; Fageria, 2010). Le genre *Solanum* est très vaste avec environ 1000 espèces dans le monde. Les solanacées tubéreuses n'en représentent qu'un dixième (Spire et Rousselle, 1996).

La pomme de terre est cultivée dans plusieurs pays, produite et consommée à grande échelle, à prix abordables et d'accès facile (Zaheer et Akhtar, 2016). La production mondiale totale en 2017 était de 388 millions de tonnes. Le Canada occupait le 18^{ième} rang des pays producteurs avec 4,4 millions de tonnes (AAC, 2019). Le Québec se classe au 5^{ième} rang des provinces productrices de pomme de terre au Canada avec 535 000 tonnes de tubercules produits en 2018 pour un rendement moyen de 32,2 t ha^{-1} (AAC, 2019). La production se répartit entre la table (55 %), la transformation (35 %) et la semence (10 %). L'Asie et l'Europe sont les principales régions productrices au monde cumulant plus de 80 % de la production mondiale. Mais l'Amérique du Nord détient les rendements moyens les plus élevés avec 41,2 t ha^{-1} dépassant l'Europe (17,4 t ha^{-1}), l'Amérique latine (16,3 t ha^{-1}), l'Asie (15,7 t ha^{-1}) ou l'Afrique avec 10,8 t ha^{-1} (IPC, 2008). Cette productivité est le résultat de la mécanisation avancée, du système de production impliquant la rotation avec les céréales et les fourrages améliorant la structure du sol et réduisant l'incidence des maladies, de l'utilisation accrue de l'irrigation, de la sélection variétale et des apports élevés de fertilisants et de pesticides (Allison et al., 2001; Camire et al., 2009).

1.1.2 Cycle végétatif

Le cycle végétatif de la pomme de terre peut être divisé en quatre phases (Figure 1.1), chacune commençant et se terminant par un stade caractéristique de développement (Kooman et Haverkort, 1995; Fageria, 2010). La phase I commence à la plantation et se termine à l'émergence (lorsque 50 % des plants ont émergé). La phase II va de ce début de la croissance foliaire à l'initiation du tubercule. La phase III se termine à la fin de la croissance foliaire où 90 % des assimilats produits chaque jour parviennent aux

tubercules. Durant cette phase, il y a une concurrence pour les assimilats entre les tubercules et le feuillage. La phase IV enfin, va jusqu'à la récolte où tous les assimilats sont répartis sur les tubercules. Ces phases de croissance sont étroitement liées tant dans le temps que dans la physiologie, et la durée de chacune dépend du génotype, de l'environnement et de l'interaction génotype par environnement (Struik, 2007).

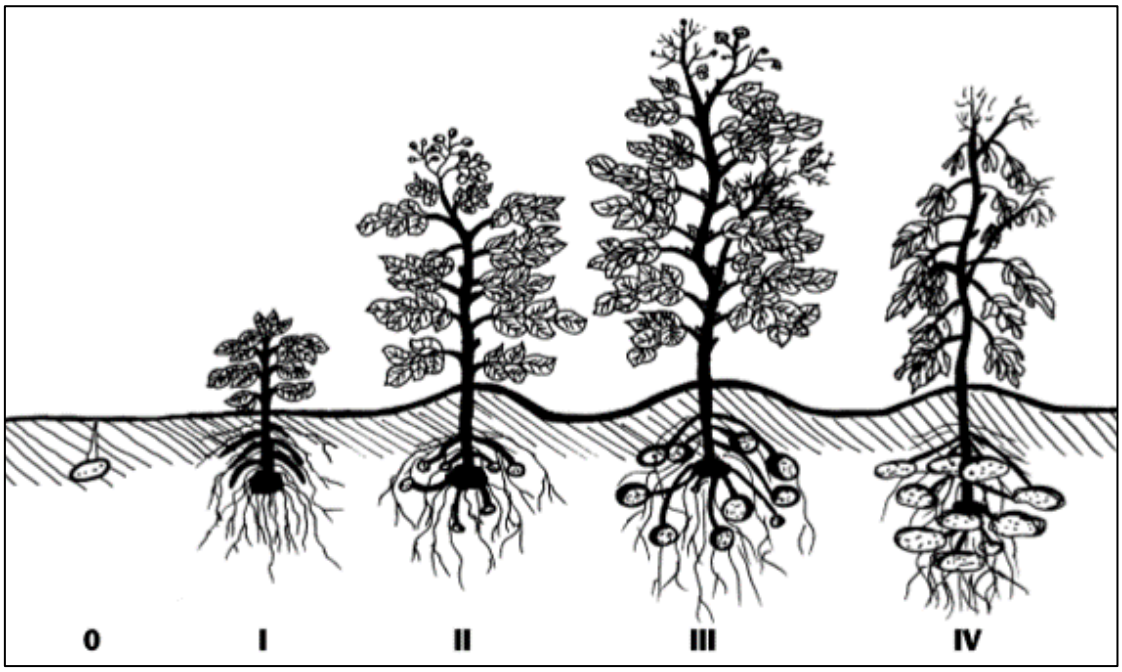


Figure 1.1 : Stades de développement de la pomme de terre (Fraser, 1998)

1.1.3 Classement des variétés

Les variétés de pomme de terre sont classées principalement selon la durée pour atteindre la maturité, le mode d'utilisation des tubercules, l'aptitude au stockage et la coloration de la chair et de la peau.

En fonction du nombre de jours qui sépare la plantation de la récolte, l'agence canadienne d'inspection des aliments (CFIA, 2015) distingue des cultivars très précoces (65 - 70 j), des cultivars précoces (70 - 90 j), des cultivars de mi-saison (90 - 110 j), des cultivars tardifs (110 - 130 j) et des cultivars très tardifs (> 130 j). Les cultivars ayant des durées plus longues présentent un potentiel de rendement généralement plus élevé que ceux à maturité plus courte (Kawakami et al., 2004; Eschemback et al., 2017).

Les tubercules sont produits pour répondre à des normes de qualité liées à une utilisation spécifique, soit pour la table, la transformation, la semence ou la production

d’amidon et autres (Camire et al., 2009; Bansal et Trehan, 2011; Mori et al., 2015). Pour le marché de la transformation en croustilles, on préfère une teneur élevée en matière sèche qui confère de la robustesse et de la croustillance aux tranches. La teneur en sucres réducteurs doit être faible pour éviter de produire des croustilles brun-foncées. Pour le marché de la pomme de terre frite, le pourcentage de matière sèche est moins important que la longueur des tubercules, la fermeté et le poids spécifique. Pour le marché de la table, la teneur en matière sèche n'est pas importante et les gros tubercules sont préférables. Enfin, les tubercules destinées à la semence sont plus petits (Imas et Bansal, 1999).

Les cultivars sont aussi classés en fonction de leur aptitude à conserver les propriétés intrinsèques longtemps au stockage. Certains cultivars doivent être consommés ou transformés peu de temps après la récolte et d'autres conservent leurs propriétés plus longtemps en entrepôt. Les cultivars qui peuvent être entreposés plus longtemps sont avantageux pour l'industrie alimentaire (Camire et al., 2009).

Les préférences des consommateurs peuvent porter sur la couleur de la peau ou de la chair des tubercules. Les couleurs de pelure les plus courantes sont brune, rouge, jaune et rousse avec une chair blanche (Camire et al., 2009). D’autres critères de qualité externes importants s’appliquent tels que le calibre des tubercules, la forme du tubercule (ronde, allongée), la fermeté de la peau du tubercule, les blessures et les défauts extérieurs (Camire et al., 2009; Bansal et Trehan, 2011).

1.2 L’azote, le phosphore et le potassium pour la production

La biomasse de la pomme de terre est constituée à 95% de carbone, d’hydrogène et d’oxygène que la plante absorbe par l'assimilation du dioxyde de carbone et l'absorption d'eau. Le reste de la biomasse est composé d'éléments fournis par l'application d'engrais et absorbés à partir de la solution du sol (Bucher et Kossmann, 2007). La culture de pomme de terre est très exigeante en N, P, K, calcium et magnésium (Van Kempen et al., 1996; Westermann, 2005).

1.2.1 L’azote (N)

L’azote joue un rôle central dans le métabolisme des plantes en tant que constituant des protéines, des acides nucléiques, de la chlorophylle, des coenzymes, des

phytohormones et des métabolites secondaires (Hawkesford et al., 2012). Il est responsable de la croissance végétale (Westermann, 2005). En culture de pomme de terre, l'azote favorise le développement du feuillage, puis la formation et le grossissement des tubercules. En excès, il augmente le rapport tiges/racines qui peut avoir un impact négatif sur l'acquisition d'autres nutriments et de l'eau, et a tendance à privilégier la croissance foliaire au détriment des tubercules dont la maturité est retardée, la teneur en matière sèche diminuée et la teneur en nitrates augmentée (Westermann et Davis, 1992; Marschner, 1995; Van Kempen et al., 1996; Zebarth et al., 2004a; Bucher et Kossmann, 2007).

1.2.2 Le phosphore (P)

Le plus grand défi auquel l'agriculture moderne est confrontée en lien avec la qualité du sol est selon Condron (2004), l'établissement et le maintien d'un équilibre approprié entre les applications de P suffisant pour soutenir la production tout en minimisant le transfert du P diffus et les impacts associés sur la qualité de l'environnement. Le P est un élément structurel des acides nucléiques et des adénosines phosphates. En tant que tel, il joue un rôle important dans le transfert d'énergie. Il est également essentiel pour le transfert des hydrates de carbone dans les cellules des feuilles (Hawkesford et al., 2012). Son approvisionnement adéquat stimule la croissance précoce à travers le développement racinaire et accélère la maturité. Il est essentiel à la production des fleurs, des fruits et des graines (Van Kempen et al., 1996; Busman et al., 2002; Westermann, 2005). Le P influence également la synthèse de l'amidon des tubercules (Stark et Love, 2003) et le nombre de tubercules par plant (Rosen et Bierman, 2008).

1.2.3 Le potassium (K)

Contrairement à l'N et au P qui sont des éléments de structure (protéines, acides nucléiques, phospholipides, ATP, *etc.*), le K existe principalement sous forme d'ion libre ou de cation adsorbé (Lindhauer, 1985). Une meilleure disponibilité en K favorise l'expansion foliaire aux premiers stades de la croissance et prolonge la durée de végétation (Bansal et Trehan, 2011). Le K assure diverses fonctions comme l'activation d'enzymes, l'ouverture des stomates, la photosynthèse et l'osmose, la régulation de l'eau dans la plante et la turgescence cellulaire (Dampney et al., 2011; Hawkesford et al., 2012). La disponibilité en K favorise le taux et la durée de grossissement des tubercules, la hauteur

et la vigueur des plants et leur confère une résistance contre la sécheresse, le gel et les maladies. Le K influence également la qualité des tubercules (calibre, teneurs en matière sèche, en amidon, en sucre réducteurs), la qualité à la cuisson, la durée du stockage et limite la sensibilité aux dommages physiques (Perrenoud, 1993; Imas et Bansal, 1999).

Des rendements élevés sont assurés par l'application de doses optimales de ces fertilisants dans des proportions équilibrées. Leur absorption varie en fonction du stade de développement de la culture (Figure 1.2). Le besoin en K de la pomme de terre est plus élevé que celui de la plupart des grandes cultures (Dampney et al., 2011). Il est aussi requis en quantités plus grandes comparé à N et P (Figure 1.2). La quantité de K prélevée est 1,5 fois supérieure à celle de N et 4 à 5 fois supérieure à celle du P (Stark et al., 2004; Bansal et Trehan, 2011).

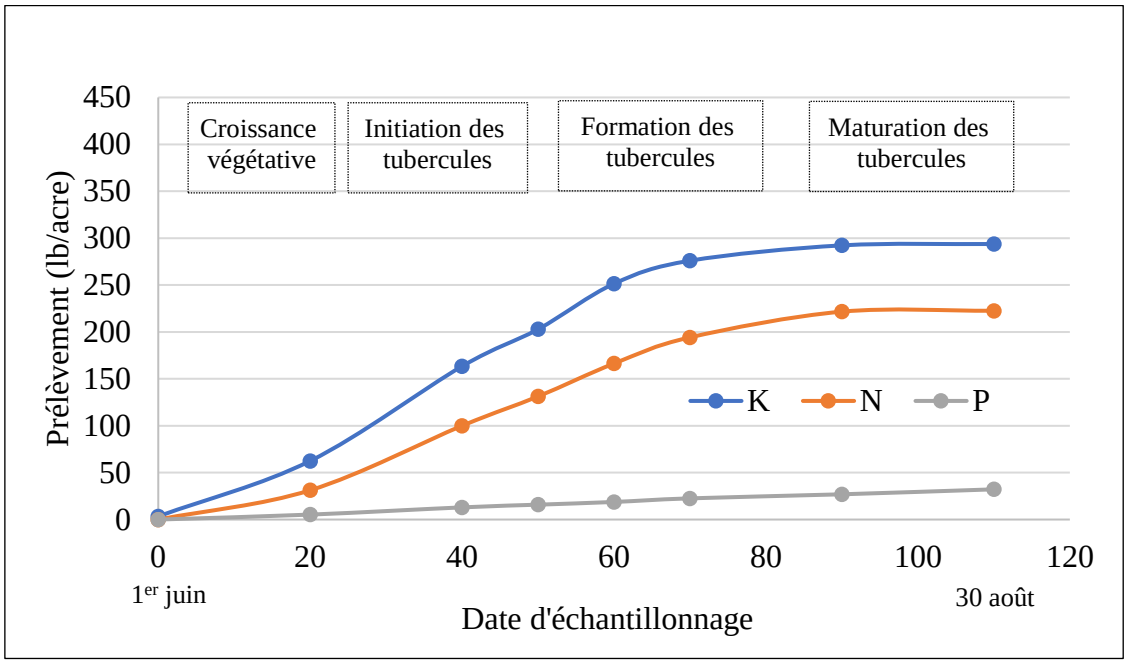


Figure 1.2 : Évolution de l'absorption de N, P et K par la pomme de terre, adapté de Stark et al. (2004)

1.2.4 Les recommandations N, P et K pour la pomme de terre au Québec

La grille de fertilisation du CRAAQ recommande des doses de N variant de 125 à 150 kg N ha^{-1} pour le groupe textural G2 et de 135 à 175 kg N ha^{-1} pour le groupe G3 (Pellerin, 2010). Des ajustements locaux nécessitant l'expertise du praticien sont suggérés (Zebarth et al., 2007). Les recommandations de P dépendent de l'indice de saturation du sol en P (P/Al Mehlich-3) alors que le K est dosé en fonction de sa teneur dans le sol (K

Mehlich-3). Ces recommandations résultent d'une distinction des sols en classes de fertilité au sein desquelles une dose unique est recommandée. Les doses de P recommandées varient ainsi de 200 kg P_2O_5 ha^{-1} pour les ISP_1 très faibles, à 50 kg P_2O_5 ha^{-1} pour les ISP_1 extrêmement élevés. Pour le K, les recommandations varient de 240 kg K_2O ha^{-1} pour un sol à teneur en K-Mehlich-3 inférieure ou égale à 75 kg ha^{-1}, à 20 kg K_2O ha^{-1} pour un sol à teneur en K-Mehlich-3 supérieure à 450 kg ha^{-1}.

Plus le sol contient l'élément biodisponible, moins les plantes utilisent les fertilisants. Les recommandations sont plus faibles et vice-versa, ce qui limite les risques de lixiviation et d'eutrophisation (Pellerin et al., 2006).

1.3 Autres facteurs d'impact en production de pomme de terre

Les coûts liés à la fertilisation contribuent pour 15 à 20 % des coûts de production (selon La Financière Agricole). Pourtant, les engrais appliqués ont une efficacité globale faible : inférieure à 50 % pour le N, inférieure à 10 % pour le P, et de 40 % pour le K (Baligar et al., 2001). Ces taux de récupération reflètent l'effet des interactions entre plusieurs facteurs.

1.3.1 Le cultivar ou l'influence de la spécificité génétique

La durée des stades de développement dépend du génotype, de l'environnement et de leur interaction (Struik, 2007). Le génotype influence le rendement en tubercules, le nombre de tubercules, la teneur en matière sèche, le poids spécifique, la composition en nutriments et la distribution des calibres (Tekalign et Hammes, 2005; White et al., 2009). Dans les modèles de croissance SUBSTOR-Potato (Griffin et al., 1993), un coefficient génétique spécifique est attribué à chaque cultivar pour le taux d'expansion foliaire, le taux de croissance des tubercules, le déterminisme, et la sensibilité de l'initiation des tubercules au photopériodisme et à la température. Ces variations génotypiques sont liées à la capacité de la plante à intercepter les radiations solaires (Sandaña et Kalazich, 2015), à sa capacité à absorber le P (Thornton et al., 2014; Sandaña et Kalazich, 2015) et l'azote (Saluzzo et al., 1999; Zebarth et al., 2004b; Fontes et al., 2010; White et al., 2011). Les besoins en K sont spécifiques au cultivar et au marché (Dampney et al., 2011; Hüwing, 2012).

Les cultivars hâtifs assimilent moins d'azote que les cultivars de mi-saison et tardifs (Zebarth et al., 2004b). D'importantes variations subsistent également entre les cultivars de même précocité. À ce titre, la fertilisation azotée doit être ajustée à chaque cultivar (Saluzzo et al., 1999) et adaptée aux besoins localement (Fontes et al., 2010). Nyiraneza et al. (2017) ont constaté que la dose de P au rendement maximal différait entre les cultivars AAC Alta Cloud, AAC Alta Strong et Russet Burbank. Selon Firman (2013) et Vreugdenhil et al. (2011), il existe un contrôle génétique non seulement de l'absorption du P par la plante, mais aussi de la densité de plantation à recommander par cultivar. Sur la période de 2005 à 2010 au Royaume-Uni, les doses d'application de K, qui étaient en moyenne de 184 kg K ha^{-1} pour la pomme de terre de table et 175 kg K ha^{-1} pour les pommes de terre précoces et de semence, ont été réduites à 170 et 132 kg K ha^{-1}, respectivement (Dampney et al., 2011). En Basse-Saxe, les recommandations de K sont plus élevées pour les pommes de terre de table et de semence que pour les pommes de terre transformées, avec une différence dépendant du type de sol et de la teneur du sol en K. En sols marécageux à teneur en K très faible, les recommandations varient de 220 à 240 kg K_2O ha^{-1} pour les pommes de terre de table et de semence, et de 160 à 180 kg K_2O ha^{-1} pour la croustille (Hüwing, 2012).

1.3.2 La météorologie

Le plant de pomme de terre a un système racinaire superficiel très sensible au stress hydrique et aux carences nutritionnelles (Diriba, 2017). L'humidité et la température du sol influent sur l'activité microbienne dans la minéralisation de la matière organique (Ellert et Bettany, 1992), sur l'absorption des éléments nutritifs par les racines (Barber, 1995) de même que sur le rendement et la taille des tubercules (Sands et al., 1979; Shahnazari et al., 2007; Ahmadi et al., 2010; Ahmadi et al., 2014). L'humidité du sol dépend de la classe de drainage (série de sol), des précipitations et de l'irrigation. Pour analyser l'impact de ces conditions météorologiques en production de maïs, Tremblay et al. (2012) ont utilisé l'indice de diversité de Shannon (SDI) comme indicateur de la distribution des précipitations pendant la saison de croissance. La pomme de terre cesse de croître à des températures inférieures à 7°C et supérieures à 35°C, la croissance maximale se produisant à 21°C (Sands et al., 1979). Les températures fraîches du printemps réduisent la diffusion naturellement lente des ions phosphate dans le sol (Barber, 1995) et limitent la minéralisation de l'azote sous l'effet combiné de l'humidité.

Une longue saison de croissance (Kawakami et al., 2004), un climat frais et un apport important en eau sont favorables à des rendements élevés. Les besoins en eau vont principalement avec la profondeur du système racinaire et varient selon la période (3 à 4 mm d'eau/jour avant la tubérisation et 5 à 6 mm/jour à partir de la tubérisation). Pour une saison de croissance complète, les besoins totaux atteignent 455 mm (Allison et al., 2001; Camire et al., 2009).

1.3.3 Les propriétés du sol et du sous-sol

L'absorption du phosphore par la plante dépend largement du contact étroit entre les racines et les particules de sol. Ce contact dépend de la texture et de la teneur en eau du sol (Barber, 1995). En raison d'un système de racines fines superficielles (Iwama, 2008; Bolinder et al., 2015) avec une biomasse moins développée dans les sols compactés (Stalham et al., 2005; Boiteau et al., 2014), la pomme de terre est très sensible aux carences en nutriments ou au stress hydrique (Diriba, 2017). Par conséquent, l'acquisition des éléments nutritifs par la culture est régulée par la santé du sol. Aussi, les ions phosphate sont moins biodisponibles lorsque le pH du sol est faible (Truog, 1947) car ils sont fortement adsorbés par les oxydes hydratés du sol (Sample et al., 1980). Leblanc et al. (2016) ont montré qu'on peut créer un ensemble de classes de sols représentatives basées sur des données morphologiques pour refléter les propriétés du sol en culture de pomme de terre. Ce regroupement numérique a fourni une base quantitative pour intégrer les descriptions de profils de sol dans le modèle de réponse de la pomme de terre à la fertilisation azotée (Parent et al., 2017).

1.4 Risque environnemental lié à la fertilisation

La pomme de terre est cultivée principalement sur des sols à texture grossière où le risque de lessivage du nitrate est élevé (Giroux, 1982; Levallois et al., 1998). Ces sols sont le plus souvent acides et bien pourvus en oxydes hydratés de fer et d'aluminium qui limitent la disponibilité du P (Khiari et al., 2000). En raison de son système racinaire superficiel, la culture nécessite un sol bien pourvu en P comparativement à la plupart des cultures (Hopkins et al., 2014; Thornton et al., 2014). Mais le maintien d'un niveau élevé en P disponible dans le sol augmente les coûts de production et constitue un risque pour la qualité des eaux de surface. Des critères agro-environnementaux basés sur l'analyse de sol et liant le pouvoir de rétention du P aux besoins de la culture, limitent son impact

environnemental (Khiari et al., 2000; Pellerin et al., 2006). Contrairement à l'N et au P, le K n'a pas d'effet délétère connu sur la qualité des eaux. Il n'induit pas d'eutrophisation dans les rivières et les lacs (IPI, 2018).

1.5 L'optimisation de la fertilisation

En général, la recommandation de fertilisants tient compte du type de culture, de la biodisponibilité des éléments nutritifs et pour certaines cultures, du groupe textural de la couche de surface de sol. La dose optimale (DO) de fertilisant est déterminée en ajustant un modèle statistique liant un indicateur de performance (souvent le rendement) à une dose de fertilisant. Pour certaines cultures, la méta-analyse permet de développer des modèles descriptifs intégrant l'influence de plusieurs facteurs. Les travaux de Tremblay et al. (2012), Xie et al. (2013) et Tremblay et Breault (2014) ont conduit à l'élaboration de *FieldApex* (anciennement SCAN), logiciel qui calcule la dose optimale d'azote pour le maïs selon les conditions météorologiques et la texture du sol. En exploitant les données des stations météorologiques (Xie et al., 2013), les caractéristiques du sol de surface et du sous-sol (Parent et al., 2017), les cartes pédologiques, les analyses spectroscopiques et les technologies d'agriculture de précision, les champs peuvent être subdivisés en zones de gestion uniforme où les recommandations de fertilisants sont uniformes (Cambouris et al., 2006). Par ailleurs, plus de trente modèles mécanistes ont été développés pour ajuster la fertilisation azotée, la gestion de l'irrigation ou l'impact du changement climatique sur la croissance et le développement de la pomme de terre (Raymundo et al., 2014). Les approches suivantes ont permis l'élaboration des grilles de fertilisation actuelles.

1.5.1 L'approche du bilan prévisionnel pour l'azote

L'estimation de la quantité d'azote disponible pour la culture est plus complexe due à sa grande mobilité dans le sol. La dose optimale varie considérablement selon les systèmes de rotation et la fertilité azotée du sol (Giroux et Lemieux, 2006). Le bilan prévisionnel d'azote est combiné avec les indicateurs de fertilité azotée du sol (comme le test des nitrates dans la couche 0 – 20 cm deux semaines après le semis) pour déterminer la dose optimale (Giroux et Lemieux, 2006; Zebarth et al., 2007). Le bilan prévisionnel évalue les besoins en azote associés à un rendement visé et détermine la proportion de l'azote qui sera comblée par le sol et par l'engrais, à laquelle on applique un coefficient

d'utilisation (CUN). Le besoin en azote qui est le produit du rendement visé (Rdt) par les exportations (Exp), doit être couvert par la fourniture du sol (S_N) et le fertilisant apporté (dose N) selon la formule (Giroux et Lemieux, 2006) :

$$dose\ N(kg/ha) = \frac{\boldsymbol{Rdt}\left(t\ ha^{-1}\right) x\ \boldsymbol{Exp}\left(kg\ N\ t^{-1}\right) - \boldsymbol{S_N}\left(kg\ ha^{-1}\right) x\ \boldsymbol{Pr}\ (\%)}{\boldsymbol{CUN}\ (\%)} \qquad \text{(Équation 1.1)}$$

Le CUN varie considérablement selon la dose et les systèmes de rotation, soit de 30 à 70% pour la pomme de terre (Parent, 2014). Le prélèvement relatif (Pr) renseigne sur la capacité des sols à fournir l'azote. Ces coefficients sont déterminés expérimentalement (Giroux et al., 2000; Giroux et Lemieux, 2006) :

$$\text{CUN} = \frac{\text{prélèvement avec N – prélèvement sans N}}{\text{dose N}} * 100 \qquad \text{(Équation 1.2)}$$

$$\text{Pr} = \frac{\text{prélèvement sans N}}{\text{prélèvement maximal avec N}} * 100 \qquad \text{(Équation 1.3)}$$

1.5.2 Les recommandations de P et de K

Les grilles de fertilisation sont élaborées à l'issue de deux étapes de modélisation : (i) le regroupement des sols en classes de fertilité suivi (ii) des modèles de recommandation par classe de fertilité. La démarche est la même pour le P et le K, mais est plus documentée pour le P.

1.5.2.1 Regroupement des sols en classes de fertilité

Les classes de fertilité du sol sont établies sur la base d'un indicateur de fertilité (agronomique ou environnemental) qui puisse être bien corrélé à la réponse de la culture aux engrais (Westermann et Davis, 1992; Beegle, 1995; Samson et al., 2008). Des analyses de sol (Mehlich-3) sont effectuées pour quantifier le P ou le K du sol disponible pour la plante (Horneck et al., 2011). Pour que les résultats soient utilisables à des fins de recommandations, le test est calibré en effectuant un grand nombre d'essais dans des conditions locales avec des sols représentatifs (allant des déficients aux suffisamment pourvus) pour chaque nutriment (Spargo, 2013). Cette calibration est requise puisque des doses optimales différentes peuvent être déterminées au sein d'une même classe de fertilité en raison d'autres facteurs influençant le rendement (Fitts, 1955).

Le seuil agronomique de fertilité est la valeur critique de l'analyse de sol à partir de laquelle une culture non fertilisée atteint 80 à 95 % du rendement maximal obtenu avec un apport supplémentaire du fertilisant (Black, 1993; Parent et al., 2010). Ce pourcentage est le rendement relatif (Équation 1.4) qu'il est préférable de corréler à l'indicateur de fertilité en utilisant le modèle de Mitscherlich si la relation est continue ou l'analyse de groupe de Cate-Nelson (Figure 1.3) si les réponses sont trop hétérogènes (Parent et al., 2010).

$$RR = 100 * \frac{\text{rendement témoin}}{\text{rendement maximum avec engrais}} \qquad \text{(Équation 1.4)}$$

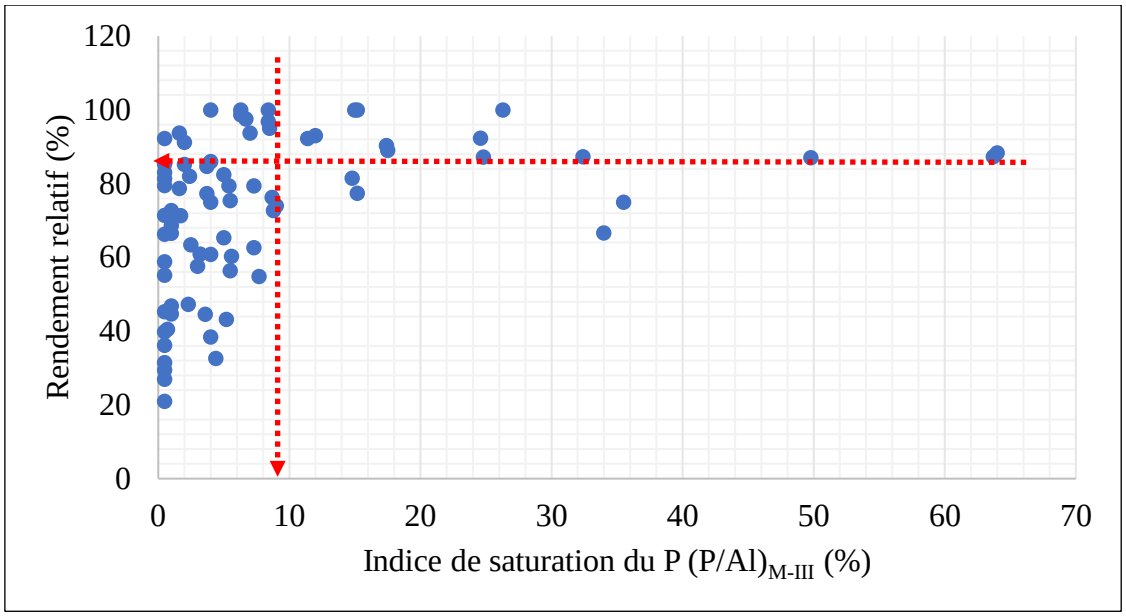


Figure 1.3 : Relation entre le rendement relatif de la culture et l'indice de fertilité du sol (ISP_1) en culture de pomme de terre au Québec, adapté de Khiari et al. (2000)

Le rendement relatif permet d'uniformiser la variabilité des réponses des cultures à la fertilisation entre les types de sol, les localités et les années (Whitney et al., 1985). La mise en commun du P et de l'aluminium (un des agents de fixation) sous forme de rapport décrit mieux la réponse de la culture aux engrais phosphatés (Khiari et al., 2000). Pour le K, c'est la teneur en K (Mehlich-3) qui est corrélé au rendement relatif.

À l'issue de cette étape, Cope et Rouse (1983) ont proposé une catégorisation en classes pouvant distinguer six classes de fertilité. À la valeur critique déterminée par la méthode itérative de Cate-Nelson (Nelson et Anderson, 1984) ou grâce au modèle optimal retenu, on attribue un indice de fertilité égal à 100. Pour les valeurs inférieures à la valeur

critique, trois groupes de fertilité sont délimités auxquels sont affectés les indices de 0 à 25, 25 à 50 et 50 à 100. Pour les valeurs supérieures à la valeur critique, trois autres groupes de fertilité peuvent être affectés avec des indices de fertilité de 100 à 200, 200 à 400 et 400 à 900. Pellerin (2010) a ainsi défini six classes de fertilité pour la culture de pomme de terre avec l'ISP_1 pour le P et sept classes pour le K (Tableau 1.1).

Tableau 1.1 : Classes de fertilité phosphatée et potassique pour la culture de pomme de terre au Québec (Pellerin, 2010)

Classes	**1**	**2**	**3**	**4**	**5**	**6**	**7**
P ISP_1 (%)	0 – 2,5	2,6 – 5,0	5,1 – 10,0	10,1 – 15,0	15,1 – 25,0	25,1 et +	
K_{M3} (kg/ha)	0 – 75	76 – 150	151 – 225	226 – 300	301 – 375	376 – 450	450 et +

1.5.2.2 Optimisation de la dose et modèle de recommandation

Cette étape consiste à déterminer la DO pour chaque site en mettant en relation les rendements vendables obtenus en fonction des doses appliquées (Samson et al., 2008). Au sein d'une même classe de fertilité, les DOs sont déterminées par site expérimental en ajustant des modèles statistiques aux données de rendement recueillies à partir des essais de fertilisation. Ces DOs sont classées en ordre croissant et celles correspondant aux espérances conditionnelles des percentiles 50 à 80 peuvent être recommandées (Khiari et al., 2000). La dose de P correspondant au percentile sélectionné dans une classe de fertilité donnée peut être jumelée à la valeur médiane de la classe de fertilité pour construire un modèle continu de recommandation (Khiari et al., 2000; Pellerin et al., 2006). Au-dessus de la valeur critique agronomique ou agro-environnementale, la recommandation de l'élément doit être limitée aux exportations par la culture (Khiari et al., 2000).

1.5.3 Les zones de gestion homogène

La gestion du champ par zones homogènes est une approche de l'agriculture de précision qui vise l'appariement des ressources et des pratiques agronomiques. La méthode tient compte de la variabilité des propriétés du sol et des exigences des cultures dans l'espace et le temps à l'échelle du champ (Whelan et McBratney, 2000). C'est une alternative à l'application uniforme de fertilisants sans tenir compte de cette variabilité. Le champ est divisé en zones de gestion qui sont supposées avoir des propriétés assez homogènes pour recevoir les mêmes doses de fertilisants.

La délimitation des zones de gestion nécessite l'analyse de la variabilité spatiale du rendement et des propriétés du sol. La méthode utilise l'analyse combinée de données issues de la télédétection, des cartes de rendements, des cartes altimétriques et des cartes de propriétés telles que la conductivité électrique apparente (Cambouris et al., 2014).

1.5.4 L'analyse de l'ionome à des fins diagnostiques

L'ionome est la composition minérale d'un organisme (Lahner et al., 2003). Son analyse à des fins diagnostiques repose sur l'hypothèse que des relations causales existent entre la croissance des plantes ou la productivité, et la teneur en éléments nutritifs des tissus (Marschner, 1995). Cet état nutritionnel se reflète mieux dans les feuilles qui de ce fait, sont préconisées pour le diagnostic (Marschner, 1995). En culture de pomme de terre, les échantillons foliaires ($3^{ème}$ à $5^{ème}$ feuille composée complète à partir du haut) sont prélevés 20 à 45 jours après l'émergence (Westermann et Davis, 1992; CRAAQ, 2003). Ce diagnostic sert d'une part, à déterminer si la recommandation du printemps était suffisante ou excessive (Giroux et Lemieux, 2006). En d'autres termes, l'analyse de l'ionome permet de détecter les carences, les excès ou les déséquilibres minéraux dans la plante à l'aide de valeurs critiques et d'indices de déséquilibre nutritif calculés (Khiari et Laga, 2010). D'autre part, l'ionome attribue une signature nutritionnelle spécifique au cultivar. Les rapports des logarithmes centrés (clr) et des logarithmes isométriques (ilr) des compositions foliaires ont été utilisés pour le regroupement de cultivars de pomme de terre (Parent et al., 1994; Hernandes et al., 2011), de variétés de manguiers (Parent et al., 2013b) et d'orangiers (Rozane et al., 2015). Ces ratios ont aussi servi à expliquer les concentrations saisonnières décroissantes de N, P et K dans les feuilles (Walworth et Sumner, 1987).

1.5.5 Les modèles statistiques courants

La régression est un outil statistique valable pour aider les chercheurs à comprendre l'interaction des facteurs présents en milieu naturel lorsqu'il s'agit d'utiliser des données brutes prises au champ (Salisbury et Ross, 1985). Supposons que l'on dispose d'un ensemble de données **D** de ***n*** observations, avec $\mathbf{D} = \{(x_i, y_i) \mid i = 1, \ldots, n\}$, où **x** est un vecteur (de covariables) d'entrées indépendantes de dimension *D* et **y** un vecteur scalaire de sorties ou cibles (variable dépendante, ou réponse). Les vecteurs colonnes servant de variables d'entrées pour les ***n*** observations sont agrégés dans une matrice

dénommée X, et les cibles sont contenues dans le vecteur **y**. Ainsi, **D = (X, y)**. En régression, les cibles sont des nombres réels. L'objectif est de faire des inférences sur la relation entre **X** et **y** c'est-à-dire la distribution conditionnelle des cibles étant donné les entrées (Rasmussen et Williams, 2006).

La méthode courante est de supposer que la relation est une fonction appartenant à une famille connue de fonctions. Par exemple, on peut supposer qu'elle est linéaire ou non linéaire simple (Barreto et Westerman, 1987; Cerrato et Blackmer, 1990; Dahnke, 1993; Bullock et Bullock, 1994; Belanger et al., 2000; Pellerin et al., 2006; Samson et al., 2008), ou linéaire ou non linéaire mixte, pouvant avoir un ou plusieurs niveaux de regroupement (Pinheiro et Bates, 2000) appelés dans ce cas des modèles hiérarchiques ou multiniveaux (Pinheiro et Bates, 2000; Parent et al., 2017). Ainsi, prédire sa valeur en n'importe quel point équivaut à déterminer les paramètres (intercept, coefficients) qui la caractérisent à partir des observations.

Qu'il s'agisse de l'N, du P ou du K, les doses optimales sont déterminées en ajustant des modèles statistiques aux données recueillies à partir d'essais de fertilisation (Belanger et al., 2000). En général, les agronomes sont intéressés à prédire les rendements des cultures (Acock et Acock, 1991). Ils donc des fonctions connues pour relier la ou les variables qui présentent un intérêt au rendement de la culture (Thornley et Johnson, 1990). Les modèles de régression linéaire, linéaire-plateau, quadratique et exponentiel sont les plus utilisés (Dahnke, 1993). Des variantes du modèle quadratique existent comme le modèle quadratique-plateau et le modèle racine-carrée. Le postulat de base considère que les niveaux de tous les facteurs autres que celui que l'on fait varier sont adéquats (Parent et al., 2010).

Le modèle linéaire est utilisé généralement pour des sols de fertilité faible (Samson et al., 2008). Le rendement augmente linéairement au fur et à mesure que la dose de l'élément augmente sans connaître un état stationnaire ni décroissance. La dose optimale correspond dans ce cas à la dose maximale appliquée.

Le modèle linéaire-plateau admet un état stationnaire. La réponse **y** est proportionnelle à la dose **X** jusqu'à une valeur limite (A) de **X** au-delà de laquelle aucune réponse n'est obtenue. Avec ce modèle, la DO se situe à l'intersection entre la section linéaire et le plateau si la pente de la section linéaire ascendante est supérieure au ratio du

prix unitaire d'achat du fertilisant ($ kg^{-1}) sur le prix unitaire de vente du produit ($ kg^{-1}), sinon est égale à zéro (Belanger et al., 2000; Elkhatib et al., 2004; Pellerin et al., 2006).

Avec le modèle quadratique, le rendement s'accroit de façon moins que proportionnelle avec l'ajout du fertilisant jusqu'à un maximum au-delà duquel la réponse diminue quand la dose augmente. Ce modèle permet de déterminer un rendement maximum à la valeur de X_{max} (dose maximale de l'élément apporté) qui annule la dérivée première de l'équation quadratique. Le rendement optimal est déterminé à la valeur de X_{op} (DO) qui égalise cette dérivée première au ratio du prix de l'engrais sur celui de la récolte (Barreto et Westerman, 1987; Cerrato et Blackmer, 1990; Belanger et al., 2000).

Le modèle exponentiel est la loi des accroissements moins que proportionnels de Mitscherlich utilisé par Parent et al. (2017) au premier niveau de modélisation pour la fertilisation azotée de la pomme de terre.

1.5.6 Le modèle multiniveau

Pendant que les modèles de niveau simple relient directement la réponse aux prédicteurs, les coefficients de régression de niveau 1 des modèles à plusieurs niveaux sont eux-mêmes estimés à partir des données (Pinheiro et Bates, 2000; Hall et Clutter, 2004; Gelman, 2006; Bressoux, 2010). Ces modèles ont une capacité élevée de généralisation et d'interpolation (Pinheiro et Bates, 2000). Le modèle de Parent et al. (2017), élaboré pour la gestion de l'azote dans la fertilisation de la pomme de terre, utilise l'équation de Mitscherlich $Y = A\left(1 - e^{-R(E+dose)}\right)$ au premier niveau. Au deuxième niveau de modélisation, les paramètres A (asymptote), E (environnement) et R (pente) sont estimés linéairement par : $A = \beta X + \mu Z + \varepsilon_A$ (le rendement potentiel), $E = \gamma X + \varepsilon_E$ (la dose équivalente d'engrais fournie par l'environnement) et $R = \delta X + \varepsilon_R$ (le taux qui est la pente de la courbe reliant Environnement à Asymptote) ; β, γ et δ étant les vecteurs des coefficients des effets fixes associés à A, E et R respectivement. La variable X est la matrice du modèle à effets fixes, μ est le vecteur des coefficients de l'effet aléatoire, Z est la matrice du modèle à effets aléatoires, ε_A, ε_E et ε_R sont les erreurs de modélisation associée respectivement à A, E et R.

1.5.7 Choix du modèle approprié

La sélection du modèle approprié n'est pas évidente puisque différents modèles ajustant un même ensemble de données peuvent donner des coefficients de détermination semblables, mais des doses optimales de fertilisants différentes (Bock et Sikora, 1990; Cerrato et Blackmer, 1990; Angus et al., 1993; Bullock et Bullock, 1994; Isfan et al., 1995; Belanger et al., 2000). La DO peut varier du simple au double selon le modèle sélectionné (Parent et al., 2010).

Considérant le modèle linéaire-plateau, Cerrato et Blackmer (1990) et Parent et al. (2010) affirment qu'il est douteux qu'un système comme la plante atteigne un plateau sans subir une atténuation progressive de sa croissance. Aussi, la trajectoire non-linéaire du modèle de Mitscherlich admet une asymptote au fur et à mesure que l'on se rapproche du rendement maximum (ou potentiel). À la limite, si on vise ce rendement potentiel, on aurait besoin d'un apport infini de fertilisant (Dahnke, 1993; Parent et al., 2010). Selon Dahnke et Olson (1990), ce modèle tend à recommander trop de fertilisant par rapport aux gains de rendement obtenus près du rendement maximum. Il faut donc viser un rendement inférieur (90, 95, 96 ou 97% du rendement potentiel) ou la DO. Belanger et al. (2000) ont trouvé que le modèle quadratique décrivait mieux la réponse de la pomme de terre pour estimer la DO d'azote. Bullock et Bullock (1994) ont préféré le modèle quadratique-plateau comparé au modèle quadratique pour prédire les besoins en azote du maïs.

1.6 Les limites des modèles actuels

En général, la dose économique optimale de fertilisant est déterminée pour un élément à la fois. Il faut un grand nombre de cas d'espèces représentant le plus grand nombre de facteurs locaux pour conduire une méta-analyse (Tremblay et al., 2012). Les zones de gestion uniforme exploitent les systèmes de positionnement global (GPS), les systèmes d'information géographique (SIG), les capteurs de rendement et les capteurs distants et proximaux pour les images satellites multi- et hyper-spectrales, et les mesures de conductivité électrique apparente du sol (Cambouris et al., 2014). Ces technologies sont parfois difficiles d'accès pour les producteurs. Les doses d'N requièrent un jugement additionnel d'expert (Zebarth et al., 2007). Les recommandations en P sont basées sur le test du P disponible du sol ou le rapport P/Al (Parent et Gagné, 2010). Les besoins en K

ne tiennent pas compte des interactions complexes entre la génétique, l'environnement et la gestion (Hatfield et Walthall, 2015) et tiennent peu compte des interactions entre cations impliquant le magnésium (Mg) et le calcium (Ca). De plus, les DOs dépendent du modèle utilisé, des prix d'achat des fertilisants et des prix de vente des tubercules (lorsqu'il faut déterminer une dose économique) qui varient selon les calibres, le poids spécifique et le marché visé (table, transformation, semence). Bien que les essais au champ puissent estimer des optimums nutritifs (Hofman et Salomez, 2000), ces derniers ne peuvent être généralisés à des conditions différentes de celles des expériences (Kyveryga et al., 2007b, 2007a) dû de la variabilité d'un site à l'autre et d'une année à l'autre. De ce fait, la fertilisation doit être raisonnée au cas par cas, chaque année, voire même être ajustée en cours de saison (Tremblay et Seydoux, 2016).

1.7 Les modèles d'apprentissage automatique

Le terme apprentissage automatique ou apprentissage machine - *machine learning* en anglais - (Blum, 2007), fait référence à la détection automatique de motifs significatifs dans des données ou encore l'extraction d'informations à partir de grands ensembles de données (Shalev-Shwartz et Ben-David, 2014). Les tâches typiques de l'autoapprentissage sont l'apprentissage de concepts, la modélisation prédictive (ou apprentissage de fonctions), le regroupement et la recherche de motifs prédictifs à travers des données issues d'expériences ou d'instructions (Willems, 2015). La technique consiste à attribuer soit une classe (à une variable dépendante qualitative) ou une valeur (à une variable dépendante quantitative) à une nouvelle observation, sur la base des données disponibles en utilisant des algorithmes d'apprentissage supervisé ou non supervisé (Jain et al., 1999; Kotsiantis et al., 2007).

En apprentissage supervisé, l'algorithme a accès à un ensemble d'apprentissage de la forme $\{(x_1, y_1) ; \dots ; (x_m, y_m)\}$ de m paires $(x;\ y)$ dont $y = f(x)$ pour une fonction f inconnue. Les valeurs x_i sont des vecteurs de la forme $(x_{i,1}, x_{i,2}, \dots, x_{i,n})$ dont les composantes sont discrètes ou réelles (Dietterich, 1997). Les valeurs y sont tirées d'un ensemble discret de classes $(1, \dots, K)$ dans le cas de la classification, ou de valeurs continues dans le cas de la régression, et elles sont connues. Ces classes ou ces valeurs connues sont utilisées pour apprendre les descriptions de classes qui, à leur tour, sont utilisées pour prédire de nouveaux ensembles (Jain et al., 1999). En apprentissage non supervisé, les valeurs y ne sont pas connues. En appliquant des algorithmes non-

supervisés, les chercheurs espèrent découvrir des structures inconnues mais utiles, uniquement à partir des données (Jain et al., 1999). Cette section présente sommairement les algorithmes utilisés dans le cadre de ce projet de doctorat.

1.7.1 Les *k* plus proches voisins

L'hypothèse de base qui sous-tend l'algorithme des k plus proches voisins est que chaque observation est similaire à ses voisins. Cet algorithme se fonde sur deux concepts fondamentaux : (1) **l'espace des variables** selon lequel chaque observation de l'ensemble d'apprentissage est représentée par un point dans un espace qui comporte autant de dimensions qu'il y a de variables prédictives, (2) il existe une **notion de distance (mesure de similarité)** dans cet espace, utilisée pour trouver les k points les plus proches du point à classer ou à estimer (Mucherino et al., 2009). Les choix courants pour la distance sont la distance euclidienne, la distance euclidienne pondérée ou la distance de Mahalanobis (McRoberts, 2009). L'algorithme ne construit pas de modèle, mais utilise l'ensemble d'apprentissage chaque fois que la sortie d'une nouvelle instance doit être prédite (Mucherino et al., 2009). Pour prédire la sortie pour un point x, un voisinage $N_k(x)$ correspondant à l'ensemble des k observations les plus proches de x est défini parmi les échantillons d'apprentissage (Crisci et al., 2012). La classe de x est celle qui est majoritaire parmi ses k plus proches voisins. Pour la régression, la prédiction est la sortie moyenne sur les k voisins les plus proches.

1.7.2 Les forêts aléatoires

Une forêt aléatoire est un ensemble d'arbres de décision formé chacun sur un sous-ensemble aléatoire des données d'apprentissage. Les prédictions des différents arbres sont ensuite agrégées pour fournir une prédiction unique (Breiman, 2001; Pal, 2017; Carvajal et al., 2018). Au lieu d'obtenir une méthode optimisée en une fois, il génère plusieurs prédicteurs pour mettre en commun leurs différentes prédictions.

Formellement, une forêt aléatoire est construite après la sélection du nombre d'arbres (T) dans la forêt, du nombre d'observations (n) et de variables (f) pour chaque arbre. Pour chaque arbre (t) dans T, l'algorithme sélectionne n observations avec remplacement dans toutes les observations et sélectionne f variables explicatives au hasard. L'arbre de décision est formé et enregistré à l'aide de l'ensemble de données de n observations avec f variables. Pour effectuer une prédiction à partir d'une nouvelle

observation (o) pour chaque arbre (t) du modèle, le résultat est prédit à l'aide de chaque arbre t appliqué à o. Si le modèle est un classificateur, la classe la plus fréquemment prédite est sélectionnée. S'il s'agit d'une régression, la moyenne de tous les résultats des arbres est calculée. C'est une méthode robuste quelle que soit la distribution des données, et est considérée comme une des techniques de prédiction les plus efficaces (Carvajal et al., 2018).

1.7.3 Les machines à vecteur de support

Les machines à vecteurs de support ou séparateurs à vaste marge (SVM) sont une approche de classification binaire exclusivement. Le principe de base consiste à trouver un hyperplan séparant deux groupes d'observations (Figure 1.4) de telles sorte que la marge entre les plus proches objets des deux ensembles et l'hyperplan soit maximale. Lorsque la séparation linéaire n'est pas évidente, les points de données sont projetés dans un espace de dimension supérieure facilitant la séparation linéaire en utilisant la technique des noyaux (Boser et al., 2003; Meyer, 2019).

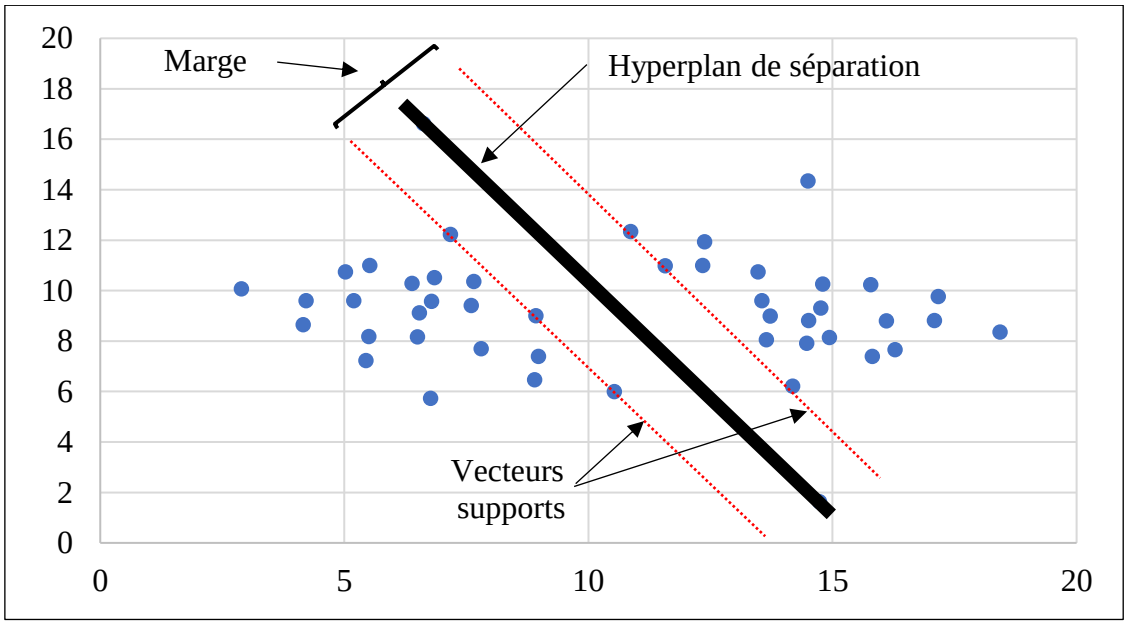


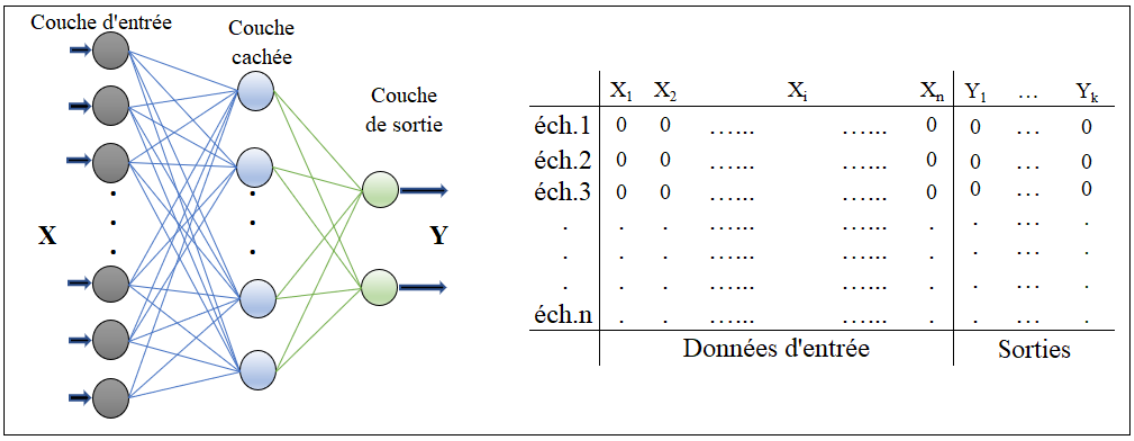
Figure 1.4 : Classification de cas linéairement séparable avec les séparateurs à vaste marge, adapté de Meyer (2019)

1.7.4 Les réseaux neuronaux

L'approche par réseaux de neurones (artificiels) se fonde sur le mode de fonctionnement du neurone biologique. Le réseau est caractérisé par son architecture,

l'algorithme d'entraînement et une ou plusieurs fonctions d'activation. L'architecture comprend une couche d'entrée définie par les variables indépendantes, une couche de sortie (la ou les réponses) et une ou plusieurs couches cachées entre les deux (Liakos et al., 2018) tel qu'illustré à la Figure 1.5. Les algorithmes sont conçus à partir des quatre fonctions de base de tous les neurones biologiques: (1) recevoir des entrées (signaux) d'autres neurones ou sources, (2) combiner leurs informations, (3) effectuer des opérations non linéaires sur ces résultats ayant différentes forces de connexion ou poids, (4) puis fournir un résultat (signal) de sortie (Klerfors et Huston, 1998). L'apprentissage ajuste le réseau aux résultats (Mucherino et al., 2009).

De manière formelle, chaque neurone de la couche d'entrée reçoit une variable d'entrée multipliée par un poids de connexion v_{ij} ($i = 1, \ldots, p$ variables; $j = 1, \ldots, n$ neurones) et diffuse ce signal vers les neurones de chaque couche cachée. Chaque neurone caché calcule sa fonction d'activation (f_j) et envoie le résultat ($z_1, \ldots, z_n$) aux neurones de la couche de sortie, ce qui produit la réponse finale du réseau (Y). Dans le cas le plus simple, ces produits sont additionnés et transformés via une ou plusieurs fonctions de transfert pour générer un résultat, puis une ou plusieurs réponses (Lek et Guégan, 1999; Chantre et al., 2012).



	X_1	X_2		X_i		X_n	Y_1	…	Y_k
éch.1	0	0	…..	…..		0	0	…	0
éch.2	0	0	…..	…..		0	0	…	0
éch.3	0	0	…..	…..		0	0	…	0
.	.	.	…..	…..		.	.	…	.
.	.	.	…..	…..		.	.	…	.
.	.	.	…..	…..		.	.	…	.
éch.n	.	.	…..	…..		.	.	…	.
	Données d'entrée						Sorties		

Figure 1.5 : Schéma d'un réseau neuronal à trois couches, avec une couche d'entrée, une couche cachée et une couche de sortie. La partie droite de la figure montre le jeu de données à utiliser. $X_1, \ldots, X_n$ sont les variables d'entrée, $Y_1, \ldots, Y_k$ sont les variables de sortie, éch.1, éch.2, …, éch.n sont les observations. Adapté de Lek et Guégan (1999).

Des réseaux de neurones plus complexes combinent plusieurs couches cachées, les couches suivantes ayant comme entrées les sorties des couches précédentes (Jain et

al., 1996). L'apprentissage avec un réseau de plus de trois couches est dit profond (*deep learning*).

1.7.5 Les processus gaussiens

La modélisation par processus gaussiens suppose que la variable dépendante peut être générée à partir de n'importe quelle fonction plutôt que de spécifier une fonction a priori (linéaire, quadratique ou exponentielle, …) pour apprendre les paramètres (Liutkus, 2012).

En supposant un ensemble de données $\boldsymbol{D} = \{\boldsymbol{X_i}, \boldsymbol{y_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{n}}$ où $\boldsymbol{X_i} \in \mathbb{R}^{\boldsymbol{d}}$ et $\{\boldsymbol{y_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{n}}$ sont les $\boldsymbol{n}$ sorties correspondantes à valeurs scalaires, cet ensemble de données peut être considéré comme un échantillonnage à partir d'une distribution conditionnelle $\boldsymbol{p(y|X)}$ dans une perspective probabiliste sous le postulat que les observations $\boldsymbol{y}$ sont indépendantes et identiquement distribuées. Ces sorties contiennent une composante systématique (f) et une autre aléatoire (ε) du type (Yuan et al., 2010) :

$\boldsymbol{y = f + \varepsilon}$ *(Équation 1.5)*

où $\boldsymbol{f}$ est une fonction latente quelconque ($\boldsymbol{f: X \rightarrow R}$) et $\boldsymbol{\varepsilon}$ un bruit gaussien additif : $\boldsymbol{p(\varepsilon)} = \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma_n^2})$. Les valeurs $\boldsymbol{f_i}$ sont traitées comme des variables aléatoires indexées par l'observation correspondante en entrée $\boldsymbol{X_i}$. Le processus gaussien (GP) est défini par une fonction de la moyenne, **m(x)**, et une autre appelée noyau générant la matrice de covariance $\boldsymbol{K(X_i, X_j)}$ entre des paires de sorties aléatoires (Rasmussen and Williams, 2006) :

$$\boldsymbol{GP} \sim \left(\boldsymbol{m(x)}, \boldsymbol{K(X_i, X_j)}\right) \quad \text{(Équation 1.6)}$$

En ajoutant la variation purement aléatoire au noyau, l'erreur du modèle s'exprime par :

$$\boldsymbol{p(y|x_1, x_2, \ldots, x_n)} = \mathbf{N}\left(\mathbf{0}, \boldsymbol{K(X_i, X_j)} + \boldsymbol{\sigma_n^2 I}\right) \quad \text{(Équation 1.7)}$$

Par convention, la fonction de moyenne retourne toujours un zéro (en s'assurant de centrer la variable réponse à zéro). Le noyau peut prendre différentes fonctions de covariance ou combinaisons de fonctions de covariance. La covariance entre les sorties

est fonction des paires de variables indépendantes correspondantes, et la fonction de base radiale est très couramment utilisée, définie par :

$$\boldsymbol{K(X_i,\ X_j) = \sigma_f^2\, exp\left(-\frac{1}{2}(X_i - X_j)^T M^{-1}(X_i - X_j)\right)} \qquad \text{(Équation 1.8)}$$

où $\boldsymbol{\sigma_f^2}$ la variance du signal commune à toutes les dimensions est généralement initialisée à 1. La matrice **M** présente sur sa diagonale le facteur d'échelle (*scaling factor*) de chaque dimension, $\boldsymbol{l = [l_1, \ldots, l_d]^T}$. Ce facteur est la distance à parcourir dans l'espace des variables prédictives avant que la valeur de la fonction puisse changer de manière importante. La covariance est proche de 1 entre des variables aux valeurs correspondantes très proches et diminue à mesure que leur distance augmente. Les valeurs attendues de la fonction devraient être similaires assurant une courbe lisse entre ces points (Rasmussen et Williams, 2006; Murphy, 2012).

En règle générale, le noyau doit pouvoir définir la hauteur ($\boldsymbol{\sigma_f}$) et la longueur ($\boldsymbol{l}$) de l'ondulation (*scaling factor*), et facultativement le bruit blanc ($\boldsymbol{\sigma_n}$) qui sont appris par le processus comme il le serait pour l'intercept et la pente dans une régression linéaire simple.

1.7.6 Limites des méthodes d'autoapprentissage

Bien que l'apprentissage machine remplace l'activité humaine par des techniques automatiques qui améliorent l'efficacité des analyses de données et la précision des résultats (Pat et Herbert, 1995), il est prétentieux d'attribuer des vertus trop grandes aux algorithmes d'apprentissage automatique (Marcus, 2018). La qualité des « décisions » prises par un algorithme dépend non seulement de la qualité (homogénéité, fiabilité, etc.) des données utilisées pour l'entrainement mais surtout de leur quantité. L'apprentissage automatique demande de grandes quantités de données pour fonctionner correctement ; données qui proviennent parfois de sources de collecte différentes dont il peut s'avérer difficile de contrôler l'intégrité. De plus, la représentativité des données est aussi nécessaire pour garantir la suffisance et l'intégrité (Yuan et al., 2010). La classification déséquilibrée opposant un petit nombre d'observations de classes ou groups minoritaires, à des centaines ou plus d'observations pour d'autres, pose un problème à la modélisation prédictive car la plupart des algorithmes sont conçus sous l'hypothèse d'un nombre égal d'observations pour chaque catégorie (Brownlee, 2020). Ce qui est rarement le cas en

pratique. Il en résulte des modèles qui ont de mauvaises performances de prédiction en particulier pour les classes minoritaires. La division aléatoire des données peut agréger les observations d'une même classe dans l'ensemble d'entrainement si bien que le modèle sera entrainé à ne prédire que cette classe, le rendant incapable à généraliser sur de nouvelles données (Kuhn et Johnson, 2013). D'après Marcus (2018), l'apprentissage machine ne distingue actuellement pas cause et corrélation de par sa construction mathématique, et est incapable d'aller au-delà du cadre imposé par les données. En d'autres termes, si on apprend à un algorithme de retourner le chiffre qu'on lui donne en l'entraînant uniquement avec les nombres 5 à 20, il sera incapable de correctement répondre à 25, donc incapable d'extrapoler. Il est donc recommandé d'avoir conscience du cadre de données que l'on a utilisé pour l'apprentissage lors de leur utilisation.

Par ailleurs, les schémas de partitionnement des données en ensembles d'entrainement et d'évaluation varient selon les études. Fortin et al. (2010) ont utilisé 60 % à l'entrainement et 40 % pour l'évaluation. Parizeau (2006) préconise 50 % à l'entrainement, 20 % à la validation et 30 % pour l'évaluation. Crisci et al. (2012) ont utilisé une répartition de 75 % et 25 % tandis que Chantre et al. (2012) ont reparti en 82 % et 18 % pour l'entrainement et l'évaluation respectivement. Soman et Bobbie (2005) ont trouvé des temps d'apprentissage plus courts et des précisions plus élevées avec 70 % des données à l'entrainement et 30 % à l'évaluation de la précision des modèles. Il n'y aurait donc pas de schéma de répartition unanime des données à l'étude.

1.7.7 Quelques utilisations en agriculture

Bien que l'utilisation de l'apprentissage machine en agriculture soit un domaine de recherche relativement nouveau, ses progrès rapides fourniront des solutions économiques et complètes dans la prise de décision pour une meilleure estimation de l'état des cultures et de l'environnement (Mucherino et al., 2009; Chlingaryan et al., 2018; Kamilaris et Prenafeta-Boldu, 2018). Différentes problématiques du secteur agricole sont désormais étudiées grâce à ces techniques dont la gestion des cultures, du bétail, de l'eau, des sols et les prévisions d'indices climatiques. Pour ne citer que quelques-unes, Ramos et al. (2017) ont utilisé les machines à vecteur de support pour estimer la production de café. Pantazi et al. (2016) ont utilisé les réseaux neuronaux pour la prédiction du rendement du blé et Pantazi et al. (2017) pour diagnostiquer la carence azotée de la culture du blé. Fortin et al. (2011) ont simulé la croissance des tubercules de pomme de terre et

bien auparavant, Elizondo et al. (1994) avaient prédit les dates de floraison et de maturité du soja avec la même technique. Qin et al. (2018) ont utilisé quatre différents algorithmes de régression (linéaire, de crête, Lasso, arbres boostés par gradient) pour prédire la DO de N pour le maïs. Rajagopalan et Lall (1999) ont utilisé l'algorithme des *k* plus proches voisins pour simuler des variables météorologiques quotidiennes (précipitations, rayonnement solaire, températures minimum et maximum, vitesse moyenne du vent et température moyenne du point de rosée). Jagtap et al. (2004) ont estimé les paramètres de l'eau du sol en combinant les teneurs en argile, sable, limon et carbone organique.

1.8 Hypothèses et objectifs

Différents facteurs alliant les caractéristiques génétiques de l'espèce ou du cultivar, les facteurs environnementaux et les pratiques culturales locales influencent la réponse des cultures aux apports de fertilisants. Les modèles classiques ajustent une fonction connue *a priori* sur les données et décrivent des effets. Les techniques d'apprentissage automatique ont un intérêt croissant grâce à leur capacité de classification et de régression prédictives. Dans le cadre de ce projet, les hypothèses spécifiques suivantes ont été posées :

1. Les cultivars de pomme de terre peuvent être regroupés sur la base de l'ionome de la feuille diagnostique. Cette hypothèse est vérifiée au chapitre 2.

2. L'ionome du tissu diagnostique permet de détecter les déséquilibres nutritifs en culture de pomme de terre durant la saison. Cette hypothèse est vérifiée au chapitre 2.

3. Le vecteur de perturbation d'Aitchison peut rééquilibrer l'ionome foliaire spécifique d'un spécimen en une composition à potentiel de rendement élevé. Cette hypothèse est vérifiée au chapitre 2.

4. Les caractéristiques génétiques, environnementales et les pratiques de gestion locales sont les principaux facteurs influant les besoins en NPK de la pomme de terre. Cette hypothèse est vérifiée au chapitre 3.

5. La capacité prédictive des doses NPK optimales est équivalente (1) entre les techniques d'autoapprentissage, (2) et entre ces techniques et le modèle multiniveau de Mitscherlich. Cette hypothèse est vérifiée au chapitre 3.

L'objectif général de ce projet de doctorat est de développer un modèle d'autoapprentissage pour prédire les doses optimales à la fois de N, de P et de K selon le rendement, le calibre et le poids spécifique en culture de pomme de terre en combinant des facteurs spécifiques au sol, à la génétique et aux pratiques agronomiques locales. Les objectifs spécifiques sont de :

1. regrouper les cultivars de pomme de terre sur la base des similitudes dans la composition nutritionnelle de la feuille diagnostique. Dans la possibilité d'un nouveau regroupement des cultivars, cette variable de classe pourrait être incluse comme variable prédictive dans le modèle de fertilisation ;

2. évaluer la capacité des techniques d'autoapprentissage à diagnostiquer un déséquilibre nutritionnel en cours de saison en utilisant les données de composition foliaire ;

3. évaluer la capacité de la méthode du vecteur de perturbation à ajuster une composition foliaire déséquilibrée ;

4. déterminer des variables locales caractérisant l'environnement, la génétique et la gestion agronomique d'importance pour la prévision des besoins en NPK en culture de pomme de terre ;

5. développer et comparer des modèles d'autoapprentissage pour prédire la combinaison NPK optimale selon le rendement, le calibre et le poids spécifique en culture de pomme de terre.

1.9 Bibliographie

AAC. (2019). Revue d'information sur les marchés de la pomme de terre 2018-2019.

Acock, B., et Acock, M. C. (1991). Potential for using long-term field-research data to develop and validate crop simulators. *Agronomy Journal, 83*(1), 56-61. doi:10.2134/agronj1991.00021962008300010015x

Ahmadi, S. H., Agharezaee, M., Kamgar-Haghighi, A. A., et Sepaskhah, A. R. (2014). Effects of dynamic and static deficit and partial root zone drying irrigation strategies on yield, tuber sizes distribution, and water productivity of two field grown potato cultivars. *Agricultural Water Management, 134*, 126-136. doi:10.1016/j.agwat.2013.11.015

Ahmadi, S. H., Andersen, M. N., Plauborg, F., Poulsen, R. T., Jensen, C. R., Sepaskhah, A. R., et Hansen, S. (2010). Effects of irrigation strategies and soils on field grown potatoes: Yield and water productivity. *Agricultural Water Management, 97*(11), 1923-1930. doi:10.1016/j.agwat.2010.07.007

Allison, M. F., Fowler, J. H., et Allen, E. J. (2001). Responses of potato (Solanum tuberosum) to potassium fertilizers. *Journal of Agricultural Science, 136*, 407-426.

Angus, J. F., Bowden, J. W., et Keating, B. A. (1993). Modeling nutrient responses in the field. *Plant and Soil, 155*, 57-66. doi:10.1007/bf00024984

Baligar, V. C., Fageria, N. K., et He, Z. L. (2001). Nutrient use efficiency in plants. *Communications in Soil Science and Plant Analysis, 32*(7-8), 921-950.

Bansal, S. K., et Trehan, S. P. (2011). Effect of potassium on yield and processing quality attributes of potato. *Karnataka J. Agric. Sci., 24*(1), 48-54.

Barber, S. A. (1995). *Soil nutrient bioavailability: a mechanistic approach*: John Wiley & Sons.

Barreto, H. J., et Westerman, R. L. (1987). YIELDFIT: A computer program for determining economic fertilization rates. *J. Agron. Educ, 16*, 11-14.

Beegle, D. (1995). Interpretation of Soil Testing Results. Chapter 12. In *Recommended Soil Testing procedures for the Northeastern United States* (pp. 84-91): Northeastern Regional Publication No. 493.

Belanger, G., Walsh, J. R., Richards, J. E., Milburn, P. H., et Ziadi, N. (2000). Comparison of three statistical models describing potato yield response to nitrogen fertilizer. *Agronomy Journal, 92*(5), 902-908.

Black, C. A. (1993). Soil fertility evaluation and control. Lewis Publication, Boca Raton, FL. . *Ann Arbor, MI*, 746 pp.

Blum, A. (2007). Machine learning theory. *Carnegie Melon Universit, School of Computer Science*, 26.

Bock, B. R., et Sikora, F. J. (1990). Modified-quadratic/plateau model for describing plant-responses to fertilizer. *Soil Science Society of America Journal, 54*(6), 1784-1789.

Boiteau, G., Goyer, C., Rees, H. W., et Zebarth, B. J. (2014). Differentiation of potato ecosystems on the basis of relationships among physical, chemical and biological soil parameters. *Canadian Journal of Soil Science, 94*(4), 463-476. doi:10.4141/cjss2013-095

Bolinder, M. A., Katterer, T., Poeplau, C., Borjesson, G., et Parent, L. E. (2015). Net primary productivity and below-ground crop residue inputs for root crops: Potato (*Solanum tuberosum* L.) and sugar beet (*Beta vulgaris* L.). *Canadian Journal of Soil Science, 95*(2), 87-93. doi:10.4141/cjss-2014-091

Boser, B. E., Guyon, I. M., et Vapnik, V. N. (2003). *A training algorithm for optimal margin classifiers.* Paper presented at the Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Bressoux, P. (2010). Modéliser des données hiérarchisées : les modèles multiniveaux. Méthodes en sciences humaines. Chapitre 6. Dans P. Bressoux (Ed.), *Modélisation statistique appliquée aux sciences sociales* (pp. 271-338): De Boeck.

Brownlee, J. (2020). *Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning*.

Bucher, M., et Kossmann, J. (2007). Molecular Physiology of the Mineral Nutrition of the Potato. Dans D. Vreugdenhil, J. Bradshaw, C. Gebhardt, F. Govers, D. K. L. Mackerron, M. A. Taylor, et H. A. Ross (Eds.), *Potato Biology And Biotechnology* (First ed., pp. 311-329): Elsevier.

Bullock, D. G., et Bullock, D. S. (1994). Quadratic and quadratic-plus-plateau models for predicting optimal nitrogen rate of corn: A comparison. *Agronomy Journal, 86*(1), 191-195.

Busman, L., Lamb, J., Randall, G., Rehm, G., et Schmitt, M. (2002). Nutrient Management: The nature of phosphorus in soils. *Regents of the University of Minnesota*.

Cambouris, A. N., Nolin, M. C., Zebarth, B. J., et Laverdiere, M. R. (2006). Soil management zones delineated by electrical conductivity to characterize spatial and temporal variations in potato yield and in soil properties. *American Journal of Potato Research, 83*(5), 381-395.

Cambouris, A. N., Zebarth, B. J., Ziadi, N., et Perron, I. (2014). Precision Agriculture in Potato Production. *Potato Research, 57*(3-4), 249-262. doi:10.1007/s11540-014-9266-0

Camire, M. E., Kubow, S., et Donnelly, D. J. (2009). Potatoes and human health. *Critical Reviews in Food Science and Nutrition, 49*(10), 823-840. doi:10.1080/10408390903041996

Carvajal, G., Maucec, M., et Cullick, S. (2018). Components of artificial intelligence and data analytics. Dans G. Carvajal, M. Maucec, et S. Cullick (Eds.), *Intelligent Digital Oil and Gas Fields* (pp. 101-148). Boston: Gulf Professional Publishing.

Cerrato, M. E., et Blackmer, A. M. (1990). Comparison of models for describing corn yield response to nitrogen-fertilizer. *Agronomy Journal, 82*(1), 138-143.

CFIA. (2015). Potato plants characteristics, maturity. Canadian Food Inspection Agency. [En ligne], URL: http://www.inspection.gc.ca/plants/potatoes/characteristics/eng/1326490397702/1326490477981#mature

Chantre, G. R., Blanco, A. M., Lodovichi, M. V., Bandoni, A. J., Sabbatini, M. R., Lopez, R. L., Vigna, M. R., et Gigon, R. (2012). Modeling *Avena fatua* seedling emergence dynamics:

An artificial neural network approach. *Computers and Electronics in Agriculture, 88*, 95-102. doi:10.1016/j.compag.2012.07.005

Chlingaryan, A., Sukkarieh, S., et Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture, 151*, 61-69. doi:https://doi.org/10.1016/j.compag.2018.05.012

Condron, L. M. (2004). Phosphorus – surplus and deficiency. Chapter 5. Dans P. Schjønning, S. Elmholt, et B. T. Christensen (Eds.), *Managing Soil Quality, Challenges in Modern Agriculture* (pp. 69-84): Danish Institute of Agricultural Sciences, Research Centre Foulum, Tjele, Denmark.

Cope, J. T., et Rouse, R. D. (1983). Interpretation of soil test result. Dans L. M. Walsh and J. D. Beaton (Eds.), *Soil testing and plant analysis. Revised edition. SSSA, Madison, WI* (pp. 35–54).

CRAAQ. (2003). *Guide de référence en fertilisation* (1ère ed.): Centre de Référence en Agriculture et Agroalimentaire du Québec.

Crisci, C., Ghattas, B., et Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling, 240*, 113-122.

Dahnke, W. C. (1993). Soil test interpretation. *Communications in Soil Science and Plant Analysis, 24*(1-2), 11-27. doi:10.1080/00103629309368778

Dahnke, W. C., et Olson, R. A. (1990). Soil test correlation, calibration, and recommendation. Dans R. L. Westerman (Ed.), *Soil Testing and Plant Analysis* (3rd ed., pp. 45-71). Madison, WI: Soil Science Society of America.

Dampney, P., Wale, S., et Sinclair, A. (2011). Potash requirements of potatoes. Review. Project R443, Report 2011/4, Potato Council, Agric. Hortic. Dev. Board, Kenilworth, Warwickshire, UK.

Dietterich, T. G. (1997). Machine-learning research. *AI magazine, 18*(4), 97-136.

Diriba, S. G. (2017). Water-nutrients interaction: exploring the effects of water as a central role for availability & use efficiency of nutrients by shallow rooted vegetable crops – a review *J. Agric. Crops, 3*(10), 78-93.

Elizondo, D. A., McClendon, R. W., et Hoogenboom, G. (1994). Neural-network models for predicting flowering and physiological maturity of soybean. *Transactions of the ASAE, 37*(3), 981-988.

Elkhatib, H. A., Elkhatib, E. A., Allah, A. M. K., et El-Sharkawy, A. M. (2004). Yield response of salt-stressed potato to potassium fertilization: A preliminary mathematical model. *Journal of Plant Nutrition, 27*(1), 111-122. doi:10.1081/pln-120027550

Ellert, B. H., et Bettany, J. R. (1992). Temperature-dependence of net nitrogen and sulfur mineralization. *Soil Science Society of America Journal, 56*(4), 1133-1141. doi:10.2136/sssaj1992.03615995005600040021x

Eschemback, V., Kawakami, J., et Melo, P. E. d. (2017). Performance of modern and old, European and national potato cultivars in different environments. *Horticultura Brasileira, 35*, 377-384.

Fageria, N. K. (2010). Cassava and potato. Dans N. K. Fageria (Ed.), *Growth and Mineral Nutrition of Field Crops, Third Edition* (pp. 457-475): CRC Press.

Firman, D. (2013). Development of seed rate recommendations for new varieties. Final report R446, Potato Council (Warwickshire, UK).

Fitts, J. W. (1955). Using soil tests to predict a probable response from fertilizer application. *Better Crops with Plant Food, 39*(3), 17-20.

Fixen, P. E. (2014). Data: landfill or legacy? . *Better Crops, 2*, 32.

Fontes, P. C. R., Braun, H., Busato, C., et Cecon, P. R. (2010). Economic optimum nitrogen fertilization rates and nitrogen fertilization rate effects on tuber characteristics of potato cultivars. *Potato Research, 53*(3), 167-179. doi:10.1007/s11540-010-9160-3

Fortin, J. G., Anctil, F., Parent, L.-É., et Bolinder, M. A. (2010). A neural network experiment on the site-specific simulation of potato tuber growth in Eastern Canada. *Computers and Electronics in Agriculture, 73*(2), 126-132. doi:https://doi.org/10.1016/j.compag.2010.05.011

Fortin, J. G., Anctil, F., Parent, L. E., et Bolinder, M. A. (2011). Site-specific early season potato yield forecast by neural network in Eastern Canada. *Precision Agriculture, 12*(6), 905-923. doi:10.1007/s11119-011-9233-6

Fraser, N. (1998). La production biologique de la pomme de terre. *Centre d'agriculture biologique de la Pocatière*, 56p.

Gelman, A. (2006). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics, 48*(3), 432-435. doi:10.1198/004017005000000661

Giroux, M. (1982). Effet des doses, des sources et du mode d'apport de l'azote sur rendement et la maturité de la pomme de terre cultivée sur différents types de sols du Québec. *Canadian Journal of Soil Science, 62*(3), 503-517. doi:10.4141/cjss82-055

Giroux, M., et Lemieux, M. (2006). Comparaison de différentes méthodes d'évaluation de la fertilité azotée des sols et détermination de la dose N optimale du maïs ensilage. *Agrosol, 17*(1), 39-50.

Giroux, M., Morin , R., et Lemieux, M. (2000). Effet des doses d'engrais N, P et K sur les rendements, les prélèvements en éléments nutritifs et la teneur en huile du canola. *Agrosol, 11*(1), 4-14.

Griffin, T. S., Johnson, B. S., et Ritchie, J. T. (1993). *A simulation model for potato growth and development: Substor-potato Version 2.0*: Michigan State University, Department of Crop and Soil Sciences.

Hall, D. B., et Clutter, M. (2004). Multivariate multilevel nonlinear mixed effects models for timber yield predictions. *Biometrics, 60*(1), 16-24. doi:10.1111/j.0006-341X.2004.00163.x

Hatfield, J. L., et Walthall, C. L. (2015). Meeting global food needs: realizing the potential via genetics x environment x management interactions. *Agronomy Journal, 107*(4), 1215-1226. doi:10.2134/agronj15.0076

Hawkesford, M., Horst, W., Kichey, T., Lambers, H., Schjoerring, J., Møller, I. S., et White, P. (2012). Chapter 6 - Functions of Macronutrients. Dans *Marschner's mineral nutrition of higher plants* (Third ed., pp. 135-189). San Diego: Academic Press.

Hernandes, A., Parent, S.-É., Veillette, J.-P., Parent, P., Leblanc, M., Roy, G., Sylvestre, P., Samson, N., Natale, W., et Parent, L.-É. (2011). Compositional meta-analysis of the nutrient profile of potato cultivars.

Hofman, G., et Salomez, J. (2000). Management of nitrogen and water in potato production. Dans A. J. Haverkort and D. K. L. MacKerron (Eds.), *Pers, Wageningen* (pp. 121–135).

Hopkins, B. G., Horneck, D. A., et MacGuidwin, A. E. (2014). Improving phosphorus use efficiency through potato rhizosphere modification and extension. *American Journal of Potato Research, 91*(2), 161-174. doi:10.1007/s12230-014-9370-3

Horneck, D. A., Sullivan, D. M., Owen, J. S., et Hart, J. M. (2011). Soil Test Interpretation Guide. *Oregon State University Extension Service*, 12 p.

Hüwing, H. (2012). Düngung sichert ertrag und qualität. *Land & Fort*, 12, 22nd March 2012, 2036-2038.

Imas, P., et Bansal, S. K. (1999). *Potassium and Integrated Nutrient Management in Potato*. Paper presented at the Global Conference on Potato, New Delhi, INDIA.

IPC. (2008). International year of the potato. Potato world. [En ligne], URL : http://www.fao.org/potato-2008/en/world/index.html

IPI. (2018). Fertilizers and Potassium in the Environment. [En ligne], URL : http://www.ipipotash.org/en/k-center/potassium-and-environment (April, 3rd 2017)

Isfan, D., Zizka, J., Davignon, A., et Deschenes, M. (1995). Relationships between nitrogen rate, plant nitrogen concentration, yield, and residual soil nitrate-nitrogen in silage corn. *Communications in Soil Science and Plant Analysis, 26*(15-16), 2531-2557. doi:10.1080/00103629509369466

Iwama, K. (2008). Physiology of the potato: new insights into root system and repercussions for crop management. *Potato Research, 51*(3-4), 333-353. doi:10.1007/s11540-008-9120-3

Jagtap, S. S., Lall, U., Jones, J. W., Gijsman, A. J., et Ritchie, J. T. (2004). Dynamic nearest-neighbor method for estimating soil water parameters. *Transactions of the ASAE, 47*(5), 1437-1444.

Jain, A. K., Mao, J., et Mohiuddin, K. (1996). Artificial neural networks: A tutorial. *Computer*(3), 31-44.

Jain, A. K., Murty, M. N., et Flynn, P. J. (1999). Data clustering: A review. *Acm Computing Surveys, 31*(3), 264-323. doi:10.1145/331499.331504

Kamilaris, A., et Prenafeta-Boldu, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147*, 70-90. doi:10.1016/j.compag.2018.02.016

Kawakami, J., Iwama, K., Jitsuyama, Y., et Zheng, X. (2004). Effect of cultivar maturity period on the growth and yield of potato plants grown from microtubers and conventional seed tubers. *American Journal of Potato Research, 81*(5), 327-333.

Khiari, L., et Laga, H. (2010). L'analyse des plantes comme outil de gestion des fertilisants. Dans L. E. Parent et G. Gagné (Eds.), *Le Guide de référence en fertilisation, 2è éd.* (2è ed., pp. 219-249): Centre de référence en agriculture et agroalimentaire du Québec.

Khiari, L., Parent, L. E., Pellerin, A., Alimi, A. R. A., Tremblay, C., Simard, R. R., et Fortin, J. (2000). An agri-environmental phosphorus saturation index for acid coarse-textured soils. *Journal of Environmental Quality, 29*(5), 1561-1567.

Klerfors, D., et Huston, T. L. (1998). Artificial neural networks.

Kooman, P. L., et Haverkort, A. J. (1995). *Modelling development and growth of the potato crop influenced by temperature and daylength: LINTUL-POTATO* (Vol. 3).

Kotsiantis, S. B., Zaharakis, I., et Pintelas, P. (2007). Supervised machine learning: a review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 160*, 3-24.

Kuhn, M., et Johnson, K. (2013). *Applied predictive modeling*. http://dx.doi.org/10.1007/978-1-4614-6849-3

Kyveryga, P. M., Blackmer, A. M., et Morris, T. F. (2007a). Alternative benchmarks for economically optimal rates of nitrogen fertilization for corn. *Agronomy Journal, 99*(4), 1057-1065. doi:10.2134/agronj2006.0340

Kyveryga, P. M., Blackmer, A. M., et Morris, T. F. (2007b). Disaggregating model bias and variability when calculating economic optimum rates of nitrogen fertilization for corn. *Agronomy Journal, 99*(4), 1048-1056. doi:10.2134/agronj2006.0339

Lahner, B., Gong, J. M., Mahmoudian, M., Smith, E. L., Abid, K. B., Rogers, E. E., Guerinot, M. L., Harper, J. F., Ward, J. M., McIntyre, L., Schroeder, J. I., et Salt, D. E. (2003). Genomic scale profiling of nutrient and trace elements in Arabidopsis thaliana. *Nature Biotechnology, 21*(10), 1215-1221. doi:10.1038/nbt865

Leblanc, M. A. (2016). *Relations entre les caractéristiques pédologiques et les pratiques de fertilisation et de conservation des sols.* (PhD Thesis), Laval University, Quebec, CA.

Leblanc, M. A., Gagné, G., et Parent, L. E. (2016). Numerical clustering of soil series using morphological profile attributes for potato. Dans A. E. Hartemink et B. Minasny (Eds.), *Digital Soil Morphometrics*. New York, NY: Springer.

Lek, S., et Guégan, J.-F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling, 120*(2-3), 65-73.

Levallois, P., Theriault, M., Rouffignat, J., Tessier, S., Landry, R., Ayotte, P., Girard, M., Gingras, S., Gauvin, D., et Chiasson, C. (1998). Groundwater contamination by nitrates associated with intensive potato culture in Quebec. *Science of the Total Environment, 217*(1), 91-101.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., et Bochtis, D. (2018). Machine learning in agriculture: a review. *Sensors, 18*(8), 1-29. doi:10.3390/s18082674

Lindhauer, M. G. (1985). *The role of potassium in the plant with emphasis on stress conditions (water, temperature, salinity).* Paper presented at the Proceedings of the Potassium Symposium. Pretoria, October.

Liutkus, A. (2012). *Processus gaussiens pour la séparation de sources et le codage informé.* (Doctoral dissertation. 239 pages), Télécom ParisTech,

Lokers, R., Knapen, R., Janssen, S., van Randen, Y., et Jansen, J. (2016). Analysis of big data technologies for use in agro-environmental science. *Environmental Modelling & Software, 84*, 494-504. doi:10.1016/j.envsoft.2016.07.017

Marcus, G. (2018). Deep learning: A critical appraisal. *New York University, Departments of Psychology and Neural Science, arXiv preprint arXiv:1801.00631*, 1-27.

Marschner, H. (1995). Diagnosis of deficiency and toxicity of mineral nutrients. Dans H. Marschner (Ed.), *Mineral Nutrition of Higher Plants* (2e ed., pp. 461-479): Academic Press, London.

McRoberts, R. E. (2009). A two-step nearest neighbors algorithm using satellite imagery for predicting forest structure within species composition classes. *Remote Sensing of Environment, 113*(3), 532-545. doi:10.1016/j.rse.2008.10.001

Meyer, D. (2019). Support vector machines. The interface to libsvm in package e1071. 1 - 8.

Mori, K., Asano, K., Tamiya, S., Nakao, T., et Mori, M. (2015). Challenges of breeding potato cultivars to grow in various environments and to meet different demands. *Breeding Science, 65*(1), 3-16. doi:10.1270/jsbbs.65.3

Morissette, R., Jego, G., Belanger, G., Cambouris, A. N., Nyiraneza, J., et Zebarth, B. J. (2016). Simulating potato growth and nitrogen uptake in eastern Canada with the STICS model. *Agronomy Journal, 108*(5), 1853-1868. doi:10.2134/agronj2016.02.0112

Mucherino, A., Papajorgji, P., et Pardalos, P. M. (2009). A survey of data mining techniques applied to agriculture. *Operational Research, 9*(2), 121-140.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective* (1rst ed.): MIT press.

Nelson, L. A., et Anderson, R. L. (1984). Partitioning of soil test - crop response probability. Dans *Soil testing: Correlating and interpreting the analytical results. Special Publication 29* (pp. 19–38). ASA, Madison, WI: Stelly, M.

Nyiraneza, J., Bizimungu, B., Messiga, A. J., Fuller, K. D., Fillmore, S. A. E., et Jiang, Y. (2017). Potato yield and phosphorus use efficiency of two new potato cultivars in New Brunswick, Canada. *Canadian Journal of Plant Science, 97*(5), 784-795. doi:10.1139/cjps-2016-0330

O'Grady, M. J., et O'Hare, G. M. P. (2017). Modelling the smart farm. *Information Processing in Agriculture, 4*, 179-187.

Pal, R. (2017). Predictive modeling based on random forests. Dans R. Pal (Ed.), *Predictive Modeling of Drug Sensitivity* (pp. 149-188): Academic Press.

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., et Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture, 121*, 57-65. doi:10.1016/j.compag.2015.11.018

Pantazi, X. E., Moshou, D., Oberti, R., West, J., Mouazen, A. M., et Bochtis, D. (2017). Detection of biotic and abiotic stresses in crops by using hierarchical self organizing classifiers. *Precision Agriculture, 18*(3), 383-393. doi:10.1007/s11119-017-9507-8

Parent, L. E. (2014). Nouveaux outils de gestion de l'azote dans la production de la pomme de terre. *CRAAQ, Colloque sur la pomme de terre 2014*.

Parent, L. E., Cambouris, A. N., et Muhawenimana, A. (1994). Multivariate diagnosis of nutrient imbalance in potato crops. *Soil Science Society of America Journal, 58*(5), 1432-1438. doi:10.2136/sssaj1994.03615995005800050022x

Parent, L. E., Cambouris, A. N., Tremblay, G., Thibault, E., Grenier, M., Gagné, G., et Khiari, L. (2010). L'expérimentation à la ferme. Dans L. E. Parent et G. Gagné (Eds.), *Guide de référence en fertilisation* (2è ed., pp. 345-358).

Parent, L. E., et Gagné, G. (2010). *Guide de référence en fertilisation* (2e ed.): CRAAQ, Québec.

Parent, S. E., Leblanc, M., Parent, A. C., Coulibali, Z., et Parent, L. E. (2017). Site-specific multilevel modeling of potato response to nitrogen fertilization. *Front. Environ. Sci., 5*(81), 1-18. doi:10.3389/fenvs.2017.00081

Parent, S. E., Parent, L. E., Egozcue, J. J., Rozane, D. E., Hernandes, A., Lapointe, L., Hebert-Gentile, V., Naess, K., Marchand, S., Lafond, J., Mattos, D., Barlow, P., et Natale, W. (2013a). The plant ionome revisited by the nutrient balance concept. *Frontiers in plant science, 4*. doi:10.3389/fpls.2013.00039

Parent, S. E., Parent, L. E., Rozane, D. E., et Natale, W. (2013b). Plant ionome diagnosis using sound balances: case study with mango (Mangifera Indica). *Frontiers in plant science, 4*, 1-12.

Parizeau, M. (2006). Réseaux de neurones. *University Laval*, 27-51.

Pat, L., et Herbert, A. S. (1995). Applications of machine learning and rule induction. *Communications of the ACM, 38*(11), 54–64. doi:10.1145/219717.219768

Pellerin, A. (2010). Les grilles de références. Dans L. E. Parent et G. Gagné (Eds.), *Guide de référence en fertilisation* (2è ed., pp. 359-473).

Pellerin, A., Parent, L. E., Tremblay, C., Fortin, J., Tremblay, G., Landry, C. P., et Khiari, L. (2006). Agri-environmental models using Mehlich-III soil phosphorus saturation index for corn in Quebec. *Canadian Journal of Soil Science, 86*(5), 897-910.

Perrenoud, S. (1993). Fertilizing for high yield potato. *IPI Bulletin 8.2nd Edition. International Potash Institute, Basel, Switzerland*.

Pinheiro, J. C., et Bates, D. M. (2000). *Mixed effects models in S and S-Plus*: Springer Verlag NewYork.

Qin, Z. S., Myers, D. B., Ransom, C. J., Kitchen, N. R., Liang, S. Z., Camberato, J. J., Carter, P. R., Ferguson, R. B., Fernandez, F. G., Franzen, D. W., Laboski, C. A. M., Malone, B. D., Nafziger, E. D., Sawyer, J. E., et Shanahan, J. F. (2018). Application of machine learning methodologies for predicting corn economic optimal nitrogen rate. *Agronomy Journal, 110*(6), 2596-2607. doi:10.2134/agronj2018.03.0222

Rajagopalan, B., et Lall, U. (1999). A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resources Research, 35*(10), 3089-3101.

Ramos, P. J., Prieto, F. A., Montoya, E. C., et Oliveros, C. E. (2017). Automatic fruit count on coffee branches using computer vision. *Computers and Electronics in Agriculture, 137*, 9-22. doi:10.1016/j.compag.2017.03.010

Rasmussen, C. E., et Williams, C. K. I. (2006). Gaussian processes for machine learning. *The MIT Press, Cambridge, MA, USA, 38*, 715-719.

Raymundo, R., Asseng, S., Cammarano, D., et Quiroz, R. (2014). Potato, sweet potato, and yam models for climate change: a review. *Field Crops Research, 166*, 173-185. doi:10.1016/j.fcr.2014.06.017

Rosen, C. J., et Bierman, P. M. (2008). Potato yield and tuber set as affected by phosphorus fertilization. *American Journal of Potato Research, 85*(2), 110-120. doi:10.1007/s12230-008-9001-y

Rossignol, L., et Rousselle-Bourgeois, F. (1996). Botanique, morphologie et taxinomie. Dans P. Rousselle, Y. Robert, et J. C. Crosnier (Eds.), *La pomme de terre, production, amélioration, ennemis et maladies, utilisations* (pp. 49-69): INRA, Paris, France.

Rozane, D. E., Mattos, D., Parent, S. E., Natale, W., et Parent, L. E. (2015). Meta-analysis in the selection of groups in varieties of citrus. *Communications in Soil Science and Plant Analysis, 46*(15), 1948-1959. doi:10.1080/00103624.2015.1069307

Sabarina, K., et Priya, N. (2015). Lowering data dimensionality in big data for the benefit of precision agriculture. Dans S. Patnaik (Ed.), *International conference on computer, communication and convergence* (Vol. 48, pp. 548-554).

Salisbury, F. B., et Ross, C. W. (1985). *Plant physiology. Third edition, Wadsworth Publishing Company, Belmont, California. ISBN 0-534-04482-4.* .

Saluzzo, J. A., Echeverría, H. E., Andrade, F. H., et Huarte, M. (1999). Nitrogen nutrition of potato cultivars differing in maturity. *Journal of Agronomy and Crop Science, 183*(3), 157-165. doi:10.1046/j.1439-037x.1999.00323.x

Sample, E. C., Soper, R. J., et Racz, G. J. (1980). "Reactions of phosphate fertilizers in soils ", The role of phosphorus in agriculture. Dans F. E. Khasawneh et al. (Ed.), (pp. 263-310). Madison, Wisc.: Soil Science Society of America.

Samson, N., Parent, L. E., Pellerin, A., Khiari, L., et Landry, C. (2008). Fertilisation en phosphore de la pomme de terre – recommandations. *Centre de référence en agriculture et agroalimentaire du Québec*, 14 p.

Sandaña, P., et Kalazich, J. (2015). Ecophysiological determinants of tuber yield as affected by potato genotype and phosphorus availability. *Field Crops Research, 180*, 21-28. doi:https://doi.org/10.1016/j.fcr.2015.05.005

Sands, P. J., Hackett, C., et Nix, H. A. (1979). A model of the development and bulking of potatoes (Solanum Tuberosum L.) I. Derivation from well-managed field crops. *Field Crops Research, 2*, 309-331. doi:https://doi.org/10.1016/0378-4290(79)90031-5

Shahnazari, A., Liu, F. L., Andersen, M. N., Jacobsen, S. E., et Jensen, C. R. (2007). Effects of partial root-zone drying on yield, tuber size and water use efficiency in potato under field conditions. *Field Crops Research, 100*(1), 117-124. doi:10.1016/j.fcr.2006.05.010

Shalev-Shwartz, S., et Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*: Cambridge university press.

Soman, T., et Bobbie, P. O. (2005). Classification of arrhythmia using machine learning techniques. *WSEAS Transactions on computers, 4*(6), 548-552.

Spargo, J. (2013). Interpreting your soil test results. *University of Massachusetts, Soil and Plant Nutrient Testing Laboratory*, 4 p.

Spire, D., et Rousselle, P. (1996). Origine socio-historique. Dans P. Rousselle, Y. Robert, et J. C. Crosnier (Eds.), *La pomme de terre, production, amélioration, ennemis et maladies, utilisations* (pp. 1-48): INRA, Paris, France.

Stalham, M. A., Allen, E. J., et Herry, F. X. (2005). Effects of soil compaction on potato growth and its removal by cultivation. Research review. (R261), 1-60.

Stark, J., Westermann, D., et Hopkins, B. (2004). Nutrient management guidelines for Russet Burbank potatoes. *BUL 840. Unversity of Idaho. College of Agricultural and Life Sciences*, 1-12.

Stark, J. C., et Love, S. L. (2003). Tuber quality. *Potato production systems. Aberdeen: University of Idaho*, 329-343.

Struik, P. C. (2007). Responses of the potato plant to temperature. Dans D. Vreugdenhil (Ed.), *Potato biology and biotechnology: advances and perspectives* (pp. 366-396): Amsterdam: Elsevier, New York.

Tedeschi, L. O. (2006). Assessment of the adequacy of mathematical models. *Agricultural Systems, 89*(2-3), 225-247. doi:10.1016/j.agsy.2005.11.004

Tekalign, T., et Hammes, P. S. (2005). Growth and productivity of potato as influenced by cultivar and reproductive growth II. Growth analysis, tuber yield and quality. *Scientia Horticulturae, 105*(1), 29-44. doi:10.1016/j.scienta.2005.01.021

Thornley, J. H. M., et Johnson, I. R. (1990). *Plant and crop modelling - a mathematical approach to plant and crop physiology*: The Blackburn Press. Clarendon Oxford.

Thornton, M. K., Novy, R. G., et Stark, J. C. (2014). Improving phosphorus use efficiency in the future. *American Journal of Potato Research, 91*(2), 175-179. doi:10.1007/s12230-014-9369-9

Tremblay, G., et Breault, J. (2014). Fertilisation azotée du maïs-grain: démystification, planification et projet SCAN. [En ligne], URL : https://www.mapaq.gouv.qc.ca/SiteCollectionDocuments/Regions/LavalLanaudiere/JourneesagricolesLanaudiere2015/10h45GillesTremblayetJulie%20Breault.pdf (April, 3rd 2017)

Tremblay, N., Bouroubi, Y. M., Bélec, C., Mullen, R. W., Kitchen, N. R., Thomason, W. E., Ebelhar, S., Mengel, D. B., Raun, W. R., et Francis, D. D. (2012). Corn response to nitrogen is influenced by soil texture and weather. *Agronomy Journal, 104*(6), 1658-1671.

Tremblay, N., et Seydoux, S. (2016). *Viser la dose optimale d'azote pour concilier profits et environnement.* Centre de Référence en Agriculture et Agroalimentaire du Québec,

Truog, E. (1947). The liming of soils. *The yearbook of agriculture 1947: Science in farming*, 566-576.

Van Kempen, P., Le Corre, P., et Bedin, P. (1996). Phytotechnie. Dans P. Rousselle, Y. Robert, et J. C. Crosnier (Eds.), *La pomme de terre: production, amélioration, ennemis et maladies, utilisations* (pp. 363-414): INRA, Paris.

Vreugdenhil, D., Bradshaw, J., Gebhardt, C., Govers, F., Taylor, M. A., MacKerron, D. K. L., et Ross, H. A. (2011). *Potato biology and biotechnology: advances and perspectives*: Elsevier.

Walworth, J., et Sumner, M. (1987). The diagnosis and recommendation integrated system (DRIS). Dans *Advances in soil science* (pp. 149-188): Springer.

Westermann, D. T. (2005). Nutritional requirements of potatoes. *American Journal of Potato Research, 82*(4), 301-307.

Westermann, D. T., et Davis, J. R. (1992). Potato nutritional management changes and challenges into the next century. *American Potato Journal, 69*(11), 753-767. doi:10.1007/bf02853817

Whelan, B. M., et McBratney, A. B. (2000). The "null hypothesis" of precision agriculture management. *Precision Agriculture, 2*(3), 265-279.

White, P. J., Bradshaw, J. E., Dale, M. F. B., Ramsay, G., Hammond, J. P., et Broadley, M. R. (2009). Relationships between yield and mineral concentrations in potato tubers. *Hortscience, 44*(1), 6-11.

White, P. J., Wheatley, R. E., Hammond, J. P., et Zhang, K. (2011). *Potato biology and biotechnology: advances and perspectives* (D. Vreugdenhil, J. Bradshaw, C. Gebhardt, F. Govers, M. A. Taylor, D. K. L. MacKerron, and H. A. Ross Eds.): Elsevier.

Whitney, D. A., Cope, J., et Fred Welch, L. (1985). Prescribing soil and crop nutrient needs. Dans O. P. Engelstad (Ed.), *Fertilizer technology and use* (3rd ed., pp. 25-52). SSSA, Madison, WI: SSSA, Madison, WI.

Willems, K. (2015). Machine learning in R for beginners. [En ligne], URL : https://www.datacamp.com/community/tutorials/machine-learning-in-r

Wolfert, S., Ge, L., Verdouw, C., et Bogaardt, M. J. (2017). Big data in smart farming - a review. *Agricultural Systems, 153*, 69-80. doi:10.1016/j.agsy.2017.01.023

Xie, M., Tremblay, N., Tremblay, G. J., Bourgeois, G., Bouroubi, M. Y., et Wei, Z. (2013). Weather effects on corn response to in-season nitrogen rates. *Canadian Journal of Plant Science, 93*(3), 407-417.

Yuan, J., Liu, C. L., Li, Y. M., Zeng, Q. B., et Zha, X. F. (2010). Gaussian processes based bivariate control parameters optimization of variable-rate granular fertilizer applicator. *Computers and Electronics in Agriculture, 70*(1), 33-41. doi:10.1016/j.compag.2009.08.009

Zaheer, K., et Akhtar, M. H. (2016). Potato production, usage, and nutrition - a review. *Critical Reviews in Food Science and Nutrition, 56*(5), 711-721. doi:10.1080/10408398.2012.724479

Zebarth, B. J., Karemangingo, C., Scott, P., Savoie, D., et Moreau, G. (2007). Gestion de l'azote pour la pomme de terre : Recommandations générales. *Programme d'atténuation des gaz à effet de serre pour l'agriculture canadienne*, 4 pages.

Zebarth, B. J., Leclerc, Y., Moreau, G., et Botha, E. (2004a). Rate and timing of nitrogen fertilization of Russet Burbank potato: Yield and processing quality. *Canadian Journal of Plant Science, 84*(3), 855-863. doi:10.4141/p03-123

Zebarth, B. J., Tai, G., Tarn, R., de Jong, H., et Milburn, P. H. (2004b). Nitrogen use efficiency characteristics of commercial potato cultivars. *Canadian Journal of Plant Science, 84*(2), 589-598. doi:10.4141/p03-050

Ziadi, N., Zebarth, B. J., Bélanger, G., et Cambouris, A. N. (2012). Soil and plant tests to optimize fertilizer nitrogen management of potatoes. In *Sustainable potato production: Global case studies* (pp. 187-207): Springer.

Chapitre 2 : Cultivar-specific nutritional status of potato (*Solanum tuberosum* L.) crops

Zonlehoua Coulibali[1¶], Athyna N. Cambouris[2&] and Serge-Étienne Parent[1¶*]

1 Department of Soils and Agrifood Engineering, Université Laval, Québec City, Québec, Canada
E-mail: zonlehoua.coulibali.1@ulaval.ca

2 Quebec Research and Development Centre, Agriculture and Agri-Food Canada, Québec City, Québec, Canada
E-mail: athyna.cambouris@canada.ca

***** Corresponding author
Email: serge-etienne.parent.1@ulaval.ca

¶ These authors wrote the paper and ran all computations.

& This author reviewed the paper.

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

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

Competing interests: The authors have declared that no competing interests exist. All the funders (Natural Sciences and Engineering Council of Canada, Quebec Ministry of Agriculture, Fisheries and Food, Centre SEVE, Patate Dolbec Inc., Groupe Gosselin FG, Agriparmentier Inc., Patate Laurentienne, Ferme Bergeron-Niquet, and Patates

Lac-St-Jean) have declared that no competing interests exist. This does not alter our adherence to PLOS ONE policies on sharing data and materials**.**

2.1 Résumé

Les tests de diagnostic foliaire reposent sur l'hypothèse qu'une plante saine absorbe et assimile des quantités optimales de nutriments pour une production optimale dans des conditions environnementales favorables. Les caractéristiques spécifiques d'absorption et d'assimilation étant héritées des parents, les profils nutritionnels foliaires peuvent varier considérablement entre les cultivars des mêmes groupes de maturité. Notre objectif était d'évaluer la validité d'un regroupement des cultivars et de prédire les rendements des tubercules de pomme de terre à partir des ionomes foliaires. L'étude a rassemblé 3382 observations faites au Québec (Canada) de 1970 à 2017. Ces observations sont les analyses de composition de la feuille diagnostique en N, P, K, Ca et Mg. Ces concentrations ont été transformées en variables logarithmiques centrées (clr) à partir desquelles le regroupement a été testé en utilisant les centroïdes des cultivars. L'algorithme de regroupement non supervisé (dbscan) n'a pas permis de délimiter des groupes montrant la spécificité dans la composition ionomique des cultivars. Les ionomes prétraités ont également été utilisé pour évaluer leurs effets sur les classes de rendement des tubercules en utilisant les algorithmes de classification des k plus proches voisins, des forêts aléatoires et des machines à vecteurs de supports. Ces modèles ont renvoyé une précision moyenne pratiquement similaire de 70 % montrant un potentiel de diagnostic acceptable pour détecter le déséquilibre nutritionnel en cours de saison. De plus, le vecteur de perturbation de l'espace de composition d'Aitchison permet d'attribuer les régions ionomiques optimales des nouveaux cultivars à l'un des plus proches cultivars documentés.

2.2 Abstract

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

Key words:

Potato ionomics, *Solanum tuberosum* L., cultivar-specific fertilization, Aitchison perturbation, machine learning.

2.3 Introduction

Potato cultivars are commonly classified into maturity groups based on the number of days from planting to maturity (CFIA, 2015). Compared to other maturity groups, longer-maturity cultivars generally show similar to higher yield potential (Kawakami et al., 2004; Söğüt and Öztürk, 2011; Eschemback et al., 2017) due to higher genetic potential (Saric, 1983) related to higher foliar nitrogen status (Zebarth et al., 2011) and root acquisition rate (Sattelmacher et al., 1990). Hence, nutrient management of potato cultivars often consider the cultivar maturity group. However, nutrient profiles or ionomes (Lahner et al., 2003; Salt et al., 2008) may vary among potato cultivars of the same maturity groups because cultivars inherit from a diversity of parents specific traits for nutrient absorption and assimilation (Hernandes et al., 2011). Indeed, White et al. (2012) provided evidence of important ionome variations in angiosperm species and stated that plant families could be distinguished by their shoot ionomes. Successful classifications of plant species based on axis-reductions have been implemented on compositionally preprocessed plant ionomes (Parent et al., 2013a; Parent et al., 2013b). The potato cultivar may also be classified similarly, allowing newly introduced cultivars to benefit from the documented nutrient management of older cultivars. Hence, the foliar ionome, easily collected from field trials, could provide a tool for the fertilization of newly introduced cultivars.

Tissue ionome portrays plant nutritional status (Parent et al., 2013a) under the assumption of causal relationships between plant growth rate and nutrient concentration in a tissue (Jones et al., 1991; Marschner, 1995). In survey datasets, reference compositions are those that are nutritionally balanced (Parent et al., 2013b). Imbalanced ionomes could be rebalanced theoretically through a perturbation operation (Aitchison, 1986) *i.e.*, a change in tissue composition after nutrient stress has been applied. Any factor impacting yield response to nutrients can perturb leaf composition (Dumenil, 1958). Fertilization perturbs soil composition (McKenzie et al., 1992) by supplying readily available plant-nutrients (McKenzie, 1998).

Because nutrients interact in the plant, Baxter (2015) suggested that the ionome could be treated as a combination of elements rather than elements taken in isolation. Parent (Parent et al., 2013a) described ionomes as multivariate balance systems of isometric log-ratios (Aitchison, 1986). Isometric log-ratios maps vectors of

concentrations, which are strictly positive data constrained to the measurement unit that convey only relative information, to a real space of orthonormal coordinates (Pawlowsky-Glahn and Buccianti, 2011). Indeed, ionomes are intrinsically multivariate: each part cannot be interpreted without being related to the other parts of the whole (Tolosana-Delgado and Van Den Boogart, 2011). Parent and Dafir (1992) developed the compositional nutrient diagnosis in plants using row-centered log-ratios (clr). Thereafter, compositional data transformation has been used to preprocess combined nutrients traits of plant species and cultivars (Parent et al., 2013a; de Deus et al., 2018; Melo et al., 2018; Nicolas et al., 2019) as well as animal species (Prater et al., 2019), and human food (Leite, 2016; Leite and Prinelli, 2017).

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

2.4 Methodology

2.4.1 Data set

The data set is a collection of potato surveys, and nitrogen (N), phosphorus (P) and potassium (K) fertilizer trials conducted in the province of Québec (Canada) from 1970 to 2017 (S 2.1 Table) between the US border (45th parallel) and the Northern limit of cultivation (49th parallel). The data set was filtered to remove foliar samples collected too early or too late from the beginning (10%) of flowering, as reported by scouting teams, and where three or more of the five major elements (N, P, K, Ca and Mg) have not been quantified. The complete data set comprised 3382 observations of 151 field trials. Five maturity classes were represented and we made correspondence with the Canadian Food Inspection Agency classification (CFIA, 2015): early season (65 – 70 days), early mid-season (70 – 90 days), mid-season (90 – 110 days), mid-season late

(110 – 130) and late season (130 days and more) cultivars. The number of samples per cultivar and the corresponding maturity classes are reported in S 2.2 Table.

2.4.2 Diagnostic tissue composition

The potato diagnostic tissue is the first mature leaf (4th leaf from top) sampled at the beginning (10%) of the blooming stage (Jones et al., 1991; Westermann and Davis, 1992). Twenty to 30 leaves were collected at random in each plot, composited, dried at 65^{o}C, ground to pass through a 1 mm sieve, then analyzed for N, P, K, Ca and Mg concentrations after dissolution of combustion. Total N was determined by micro-Kjeldahl or Dumas combustion (Leco CNS-2000 analyzer, St. Joseph, MI, USA). After acid dissolution (Mills et al., 1996), K, Ca, and Mg concentrations were quantified by atomic absorption spectrometry or inductively coupled plasma spectroscopy (ICP), and P by colorimetry or ICP. We made no distinction between methodologies in the analysis of ionomes.

2.4.3 Processing nutrient composition to nutrient balances

The compositional space (Aitchison, 1986) of the leaf tissue comprised five nutrients (N, P, K, Mg, Ca) and undetermined components amalgamated into a filling value (Fv) computed by difference between the measurement unit and the sum of quantified nutrients. Tissue components were preprocessed using the row-centered log-ratio transformation, as follows (Parent and Dafir, 1992):

$$clr_i = ln\left(\frac{x_i}{g(x)}\right) \qquad \text{(Equation 2.1)}$$

where x_i is raw concentration of the ith component and $g(x)$ is the geometric mean across components including the filling value.

2.4.4 Clustering cultivars

Yield thresholds are useful for decision-making. Because tuber yield potential varies widely among cultivars, we processed by discretizing tuber yields into low- and high-productivity categories (Parent et al., 2013b) by ranking the marketable yield in ascending order within a given cultivar, and selecting the yield corresponding to the 65th percentile as cut-off between the two subgroups. Hence, each cultivar had its cut-

off with respect to its yield potential as shown in S2 Table. The high-yielding subpopulation ionomes were used to assess cultivars clustering ability. This subgroup comprised 1190 occurrences (after the exclusion of 144 outliers) across 151 trials and 47 cultivars. A density-based clustering method (Hahsler et al., 2017) was used to delineate cultivar groups of similar compositions using clr variables.

2.4.5 Ionome effect and yield prediction

Machine learning algorithms can either regress to predict continuous variables or classify to predict categories (James et al., 2013). Tuber yield categories were predicted using clr variables and information on ionomic groups of the full data set (high and low yielders *i.e.* 3382 rows). Three machine learning algorithms were compared: k-nearest neighbors, random forest and support vector machines.

We estimated the relative influence of variables in the model and their rank by examining how can prediction error increases when data for a variable is permuted while all others are left unchanged (Breiman, 2002; Liaw and Wiener, 2002). A variable can score a zero or too small value compared to others. Deleting such variable from the dataset should not impact on the results. The random forest model was used for feature selection to assess importance of each clr variable in predicting tuber yield, but none of the variable was removed.

The data were split into training (75%) and testing (remaining 25%) sets at cultivar level i.e., for each cultivar the samples were randomly separated according to these proportions. The performance of the classification models was assessed using accuracy computed with the testing set. Applied to the context, the four quadrants defined by Swets (1988) in binary system diagnosis to delineate the response classes are presented in the contingency table (Table 2.1). The accuracy is the proportion of correctly predicted instances:

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \qquad \text{(Equation 2.2)}$$

Table 2.1: Terms definitions used for the study.

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

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

2.4.6 Rebalancing a composition: the enchanting islands

A compositional perturbation is a translation in the compositional space (Aitchison and Ng, 2005; Monna et al., 2017). A perturbation vector expressed as clr values contains a series of deltas (differences). Once back-transformed into the compositional space, the perturbation vector alters a composition through a perturbation (⊕) operation as follows (Aitchison and Ng, 2005):

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

(Equation 2.3)

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

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

The high yielders of the training set correctly diagnosed as balanced (true negative specimens) by the most accurate model were used as the reference subpopulations. The clr values of these reference specimens were used as reference

nutritional status at high yield potential. A potato nutrient imbalance index was computed as a distance from the closest high-yielding specimen using the Aitchison distance, *i.e.* the Euclidean distance between compositions using clr-transformed concentrations (Egozcue and Pawlowsky-Glahn, 2006). For any misbalanced or new specimen of a given cultivar, the closest true negative (closest reference composition) was identified as the sample with the minimum Aitchison distance from the new composition. The nutrient clr differences defining the Aitchison distance may be considered as apparently excess or deficiency of the nutrient requiring correcting measures in a multivariate and compositional data perspective (Parent, 2020). Hence, the clr space of nutrient components (N, P, K, Ca, Mg) was described not as an ellipsoidal hyper-space (Hron, 2009) but as islands of high-yielding specimens dispersed in the hyper-space of differently yielding specimens. The closer is a specimen from the enchanting island, the higher its chance to become a high-yielder (Parent, 2020). The clr-difference was converted into a perturbation vector between two nutrient compositions expressed as familiar nutrient concentrations.

2.4.7 Statistical analysis

Statistical computations were performed in the R statistical environment version 3.6.1 (R Core Team, 2019). Compositional data analysis was conducted using the R *compositions* package version 1.40-2 (Van den Boogaart et al., 2014). Multivariate outliers were removed for robust multivariate analysis (Filzmoser and Hron, 2011) using the Mahalanobis distance at a 0.01 level of significance with the R *mvoutlier* package version 2.0.9 (Filzmoser and Gschwandtner, 2018). The clustering operation were performed using *dbscan* package version 1.1-3 (Hahsler et al., 2017). Linear discriminant analysis (LDA) was conducted using the R *ade4* package version 1.7-13 (Dray and Dufour, 2007) which allows computing linear combinations of clr coordinates that best discriminate cultivars ionomes centroids. Supervised analysis was conducted using the *caret* package version 6.0-84 (Kuhn et al., 2008). Computations performed for this paper are reproducible by using the R codes and data given as supplementary information and available online in a GitHub repository at https://git.io/Jvt2r.

2.5 Results

2.5.1 Cluster analysis

The data set used for clustering is described in S2 Table. The AC Chaleur cultivar showed the lowest tuber marketable yield cut-off (65th percentile) at 17.4 Mg ha^{-1} and Red-Maria, the highest at 64.6 Mg ha^{-1}. Average marketable yield was 40.5 Mg ha^{-1} for high yielders and 24.8 Mg ha^{-1} for low yielders. In comparison, average potato tuber yields in Canada and Québec were 31.2 Mg ha^{-1} and 28.8 Mg ha^{-1} respectively, in 2018 (Statistics Canada, 2017).

The *dbscan* clustering function looked for dense regions in the clr-space, and detected a single cluster of cultivars *i.e.*, cultivars were scattered without any particular dense region. A principle components analysis allowed to map cultivars and nutrients in the biplot shown in Figure 2.1. The principle components scores mapped on the distance biplot (Figure 2.1A) showed no particular pattern allowing groups partition. The clr correlation loadings (Figure 2.1B) showed a negative relationship between K and Mg, P and Ca, and positive relationship between N and P in agreement to concentration changes with time as the plant matures (Parent et al., 1994). Discrepancies between cultivars were driven mainly by Mg and K on the first axis, and by P and Ca on the second axis (right hand side plot).

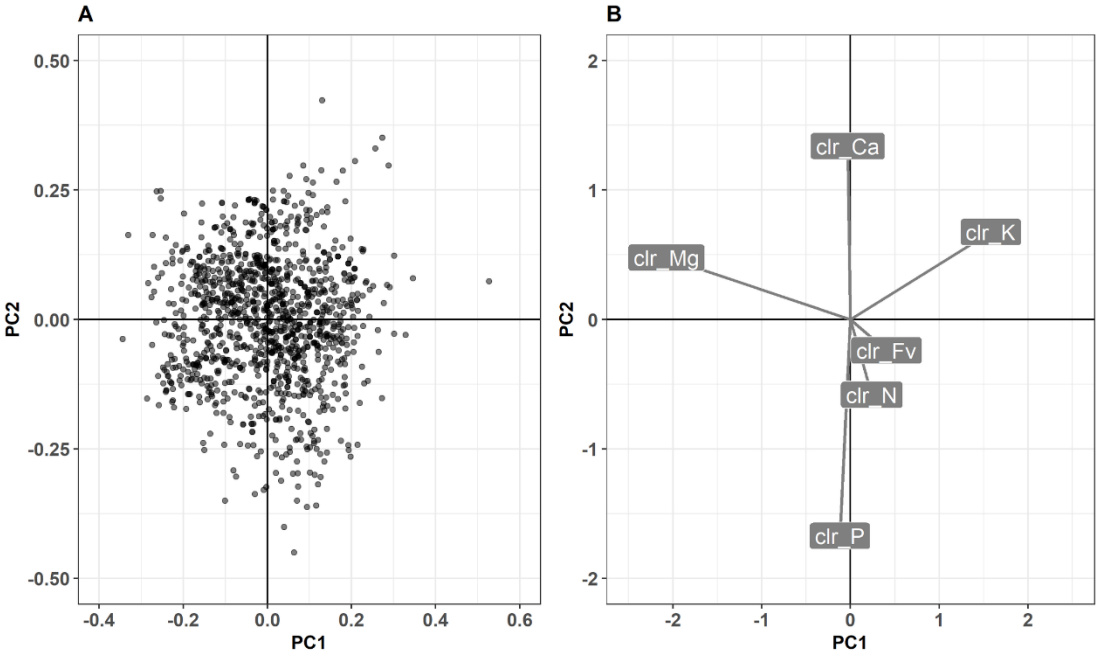


Figure 2.1: Principle components biplot of potato ionome showing (A) scores in distance scaling and (B) loadings in correlation scaling.

2.5.2 Predicting tuber yield

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

After splitting data into training (75%) and testing (25%) data sets, we used a ten-fold cross-validation process that sequentially splits the training data set into ten parts, using nine parts for calibration and the remainder for validation. The *k*-nearest neighbours, the random forest and the support vector machine models returned practically similar predictive accuracies (although slightly lower for the support vector machine algorithm), with a mean accuracy of 70% representing 591 successful and 254 unsuccessful cases classification with the testing set. The null hypothesis for a random classifier *i.e.*, non-informative classification consisting of an equal distribution of 50% successful and 50% unsuccessful cases was rejected after a χ^2 homogeneity test ($\chi^2 = 69.135$, $p < 2.2\ 10^{-16}$). Since all the models returned practically similar accuracy over the testing set, predictions with the *k*-nearest neighbors' model were used for interpreting. There was high variation in model fit by cultivar as shown in Figure 2.2. The accuracy at testing varied from 25% for Estima and Waneta, to 100% for Ambra, Carolina, Dark Red Chieftain, Harmony, Peribonka and Viking. All these cultivars had small sample sizes in the dataset as shown in the S2 Table.

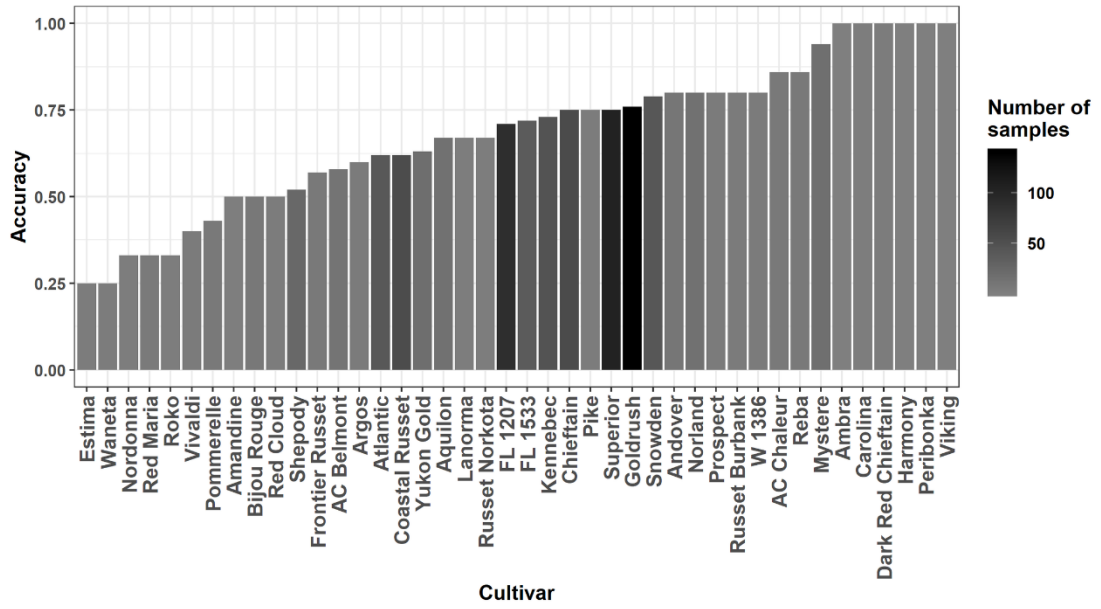


Figure 2.2: The k nearest neighbors model evaluation accuracies for cultivars.

2.5.3 Ionome perturbation

The true negative specimens (correctly diagnosed as balanced) comprising 783 occurrences in the training data set provided the clr reference values required to compute the Aitchison distance, which is equal to the Euclidean distance across clr-transformed compositions. The S 2.3 Table displays mean values for each cultivar. Using the Aitchison metric, the closest true negative specimen was set as the reference composition for each imbalanced specimen. In the clr-space, the difference between the reference and the imbalanced compositions returns a perturbation vector. Figure 2.3 shows the imbalanced sample with the highest Aitchison distance from its reference and the perturbation to apply as a translation to reach a balanced ionome. The most imbalanced observation nutrient composition was (0.0601, 0.0037, 0.0355, 0.0032, 0.0048. 0.8919), the nearest reference composition was (0.0561, 0.0036, 0.0603, 0.0052, 0.0184, 0.8565), the corresponding perturbation vector was (0.0919, 0.0965, 0.1696, 0.1629, 0.3832, 0.0959) for N, P, K, Mg, Ca and Fv respectively. The Aitchison distance computed between the observation and its associated true negative was 1.135.

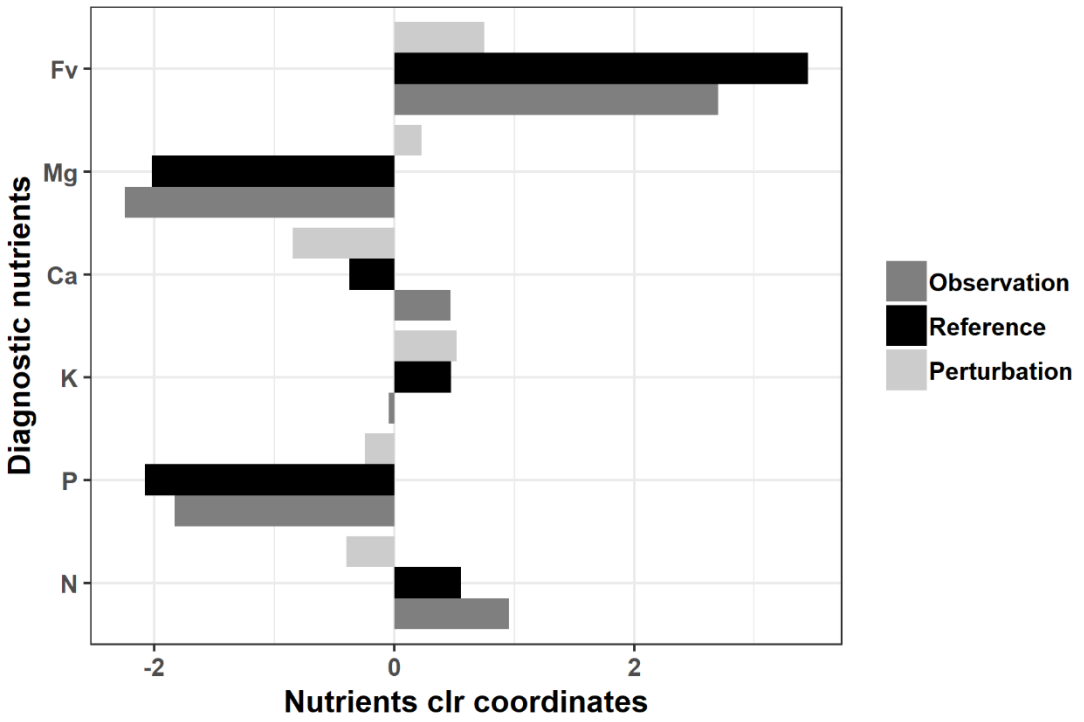


Figure 2.3: Perturbation vector example mapped using the most imbalanced sample.

2.6 Discussion

2.6.1 Clustering potato cultivars

The Canadian Food Inspection Agency classified potato cultivars broadly into maturity groups based on the time elapsed between planting and maturity (CFIA, 2015). However, nutrient requirements, especially nitrogen, vary widely between cultivars of the same maturity group. In New Brunswick (Canada), Zebarth et al. (2007) recommended 200-208 kg N ha^{-1} for Russet Norkota (early-season cultivar) and Russet Burbank (late-season cultivar), 190 kg N ha^{-1} for Superior (early-mid-season cultivar) and Goldrush (mid-season cultivar), 175 kg N ha^{-1} for Shepody (mid-season), 135 kg N ha^{-1} for early cultivars for the table market, 160-180 kg N ha^{-1} for other mid-season, 180-200 kg N ha^{-1} for other late, and 135-160 kg N ha^{-1} for low N requirement cultivars. Such large discrepancies within the same cultivar maturity group was attributed to differential foliar gene expression (Zebarth et al., 2011) and root development (Sattelmacher et al., 1990). Hence, information additional to maturity grouping is needed to assess nutrient requirements of potato cultivars. Huang and Salt (2016) reported that ionomics allows the discovery of genes controlling natural variation in the

plant ionome and for Salt et al. (2008), ionomics could capture information about the functional state of an organism driven by genetic and environmental factors. The content of plant tissue reflects what the plant can absorb from the soil and for each nutrient, there is a correlation between its concentration and yield. Since tissue analysis is carried out as well for observing the effect of fertilizer applications as for determining the in-season or next season nutrient requirement (Cottenie, 1980; Hochmuth et al., 2004), ionomes could be useful in discriminating potato cultivars. Indeed, using a small data set of eight potato cultivars, Hernandes et al. (2011) showed that foliar nutrient profiles varied widely among cultivars of the same maturity group. According to Parent et al. (2013b), variations in ionomes could be interpreted only partly as genotypic effect, and phenotypic plasticity can also be driven by nutrient supply capacity specific to agroecosystems while breeding programs are conducted under relatively luxurious environments to reach high productivity. The N, Mg and K clr values, that dominated principal components (Figure 2.1), could reflect the abilities of individual cultivars to acquire and use those nutrients more efficiently (White et al., 2007; Giletto and Echeverría, 2015). Natale et al. (2018) provided evidence that in general macronutrient contents differ among species and cultivars and within the same species for fruit trees. For N, K and Ca, this range is wider because of higher requirement of these elements by plants, and narrower for P, Mg and S, indicating smaller demand for the latter.

To cluster is to recognize that objects are sufficiently similar to be put in the same group, and to identify distinctions or separations between groups of objects (Legendre and Legendre, 2012; Borcard et al., 2018). Based on the assumptions of differential genotypic potential, root development, nutrients requirement, nutrients uptake and use efficiency, the goal was to discover interesting structure about the N, P, K, Mg and Ca contents of the diagnosis tissue to decipher dissimilarities between cultivars (James et al., 2013). However, the process failed to discriminate groups of cultivars along the clr coordinates. Hernandes et al. (2011) reached similar results with overlapping nutrient profiles between cultivar groups depending on isometric log ratio (ilr) coordinates. They found similar nutrient profiles between cultivars groups along some ilr coordinates and very different ones along others. While ionome dissimilarities are not numerically compelling, they could assist classifying new cultivars into appropriate ionomic group to benefit from costly fertilizer trials conducted on cultivars of the same group.

2.6.2 Tuber marketable yield prediction

The P content of the diagnostic leaf did not appear useful in predicting potato tuber yield classes. Other elements (N, K, Ca and Mg) showed important contribution of their clr values on the prediction quality metric, especially N which is directly related to photosynthesis (Andrews et al., 2013). Since the fertilization trials have been conducted over a time-span of 47 years (1970 – 2017), the question arises whether the different methodology of quantifying P (colorimetry/ICP) have contributed to depreciate this variable in predicting tuber yield classes. The ICP method is shown to be faster and to give higher results for total phosphorus content in 'soils' extracts in comparison to the colorimetric method. However, they are exceptions and controversial results (Sikora et al., 2005; Ivanov et al., 2010; Adesanwo et al., 2013). Ivanov et al. (2010) found that the two methods for total P determination in plant material were highly correlated, and the results were generally within 5 to 10% of each other. Moreover, Valkama et al. (2009) provided that although agricultural practices, soil conditions and analytical techniques have undergone substantial changes over time, the differences between old and recent experiments in yield responses to P application were not statistically important. For all these reasons, we consider the two analytical methods equally relevant to the analysis. The low importance of P clr variable in predicting tuber yield classes may comes from its correlation with Ca. Globally, the selection of relevant features is achieved, firstly, by checking the correlation between features and response to select the features which have correlation above a selected level (*e.g.*, 0.5). Then, the independent variables need to be uncorrelated with each other. If some features are correlated, only one is kept. The process selected the clr_Ca variable (alphabetical order) instead of clr_P since these features are correlated as shown in Figure 2.1B. In this study no element was discarded from the process relative to its importance.

The tested algorithms (*k*-nearest neighbours, random forest and support vector machine) returned similar accuracies in the prediction of yield classes using clr variables as predictors and showed fair diagnostic potential to detect nutrient imbalance. The correctly predicted high and low yielders reached 70% in the testing data set. The models classified more accurately the yield categories compared to a random classifier (Hollander et al., 2013). Specimens classified as false negatives (*i.e.*, low yielders incorrectly classified as high yielders) are attributable to limiting

conditions other than N, P, K, Mg, and Ca nutrition: soil physical and chemical properties (Stalham et al., 2005; Boiteau et al., 2014), fertilization (Zebarth et al., 2004), management failures, diseases (Rich, 1983) or weather events (Herman et al., 2017) impacting plants growth and yield potential. False positive specimens (*i.e.*, high yielders incorrectly classified as low yielders) indicate luxury consumption when nutrient concentrations are higher (Parent et al., 2012; Parent et al., 2013b), or other particularly favorable growth conditions. The confusion matrix built for cultivars revealed poor predictive accuracy for certain cultivars (*i.e.*, 25% for Estima and Waneta) and conversely an accuracy of 100% for others (*i.e.*, Ambra, Peribonka) as shown in Figure 2.2. These cultivars involved mainly small sample sizes (only one, two or three high-yielders and five, six or a little more low-yielders). The problems of small-data in machine learning are numerous, but mainly revolve around over-fitting. The training and testing datasets division could only aggregate observations of one class in the training set so that the model would train to always predict this dominant class (Kuhn and Johnson, 2013). The model could either memorize labels, which is not ideal for generalizing on new data. Brownlee (2020) explained that imbalanced classifications (one or less examples in a minority class for hundreds or more examples in the other) pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class. This results in models that have poor predictive performance, specifically for the minority class. The controversial accuracy level for some cultivars (especially low level) could either come from other yield limiting factors specific to the experiments but not involved in this study, as for false positive specimens. Our model was not effective for these cultivars treated separately.

The differential nutrition of potato cultivars could be addressed objectively using mineral analysis of the diagnostic leaf. More data are needed for poorly documented cultivars. Moreover, dedicated models could be trained for cultivars for which sufficient data are available (e.g., Goldrush, Superior, FL 1207, Chieftain). Other algorithmic, sampling and quality measurement approaches could further be implemented to deal with the problems of small-data and unequal distribution of classes (Kuhn and Johnson, 2013; Brownlee, 2020). One could extend the predictors to the experimental conditions (soils, weather data), handling a site-and-cultivar-specific nutrients diagnosis model.

2.6.3 Perturbation vector for fertilizer recommendation

Rational fertilization requires information on the nutrients that are available in the soil, and the nutritional status of the plant (Marschner, 1995) as portrayed by the diagnostic tissue composition (Jones et al., 1991; Marschner, 1995). However, the diagnosis of deficiency and toxicity of mineral nutrients may be complicated in field-grown plants where more than one mineral nutrient is deficient or where there is a deficiency of one nutrient and simultaneously toxicity of another (Marschner, 1995). The scientific principle behind tissue analysis is that healthy plants contain predictable concentrations of analytical nutrients (Campbell, 2000). The values are compared to established norms for inadequate, adequate, and excess levels. However, Parent et al. (2013a) proved that this concept of growth-limiting nutrient concentrations supported by the "Law of minimum" and illustrated by Liebig's barrel, should be replaced by a concept of growth-limiting nutrient balances illustrated by a pan balance design, where groups of elements are balanced optimally against each other in weighing pans.

The difference between two equal-length compositional vectors can only be computed using tools of compositional data analysis. The perturbation vector concept applied to foliar tissue diagnosis returns a scaling operator (Pawlowsky-Glahn and Buccianti, 2011) that when applied to an imbalanced composition translates it (theoretically) into a balanced composition with high yield potential (*i.e.*, true negative). Although the closure of the simplex implies that a perturbation on the clr of a specific nutrient is methodologically not a change in proportion of a single nutrient, perturbations expressed in the clr space appear suitable for interpretation. Indeed, the difference measured between clr values of the diagnosed sample and reference (true negative) specimen can be ranked using the sign of that difference (Hernandes et al., 2011; Rozane et al., 2011; Rozane et al., 2015), hence indicating which components are at excessive or deficient levels. As provided by Parent (2020), K and Mg were apparently deficient while N, P and Ca were apparently in excess compared to the closest *reference* specimen (Figure 2.3). Using the same approach, ionomes of newly introduced cultivars with unknown nutrient requirements could be assigned to the cultivars of known nutrient requirements showing the closest ionomes.

A perturbation as the one shown in Figure 2.3 should not be interpreted as shifts of individual components, since the operation on a single component resonates on the

whole simplex (Parent, 2020). For instance, an offset in the simplex ***S*** (N, P, K, Ca, Mg, Fv) composition following the increase by 20% (theoretically) of N and P clr values is displayed on Figure 2.4. The K, Ca and Mg concentrations seemed more stable with respect to the others. Although P clr values have been increased, P proportion decreased globally for the new equilibrium of the simplex. The offset was higher for the selected components followed by the filling value (Fv).

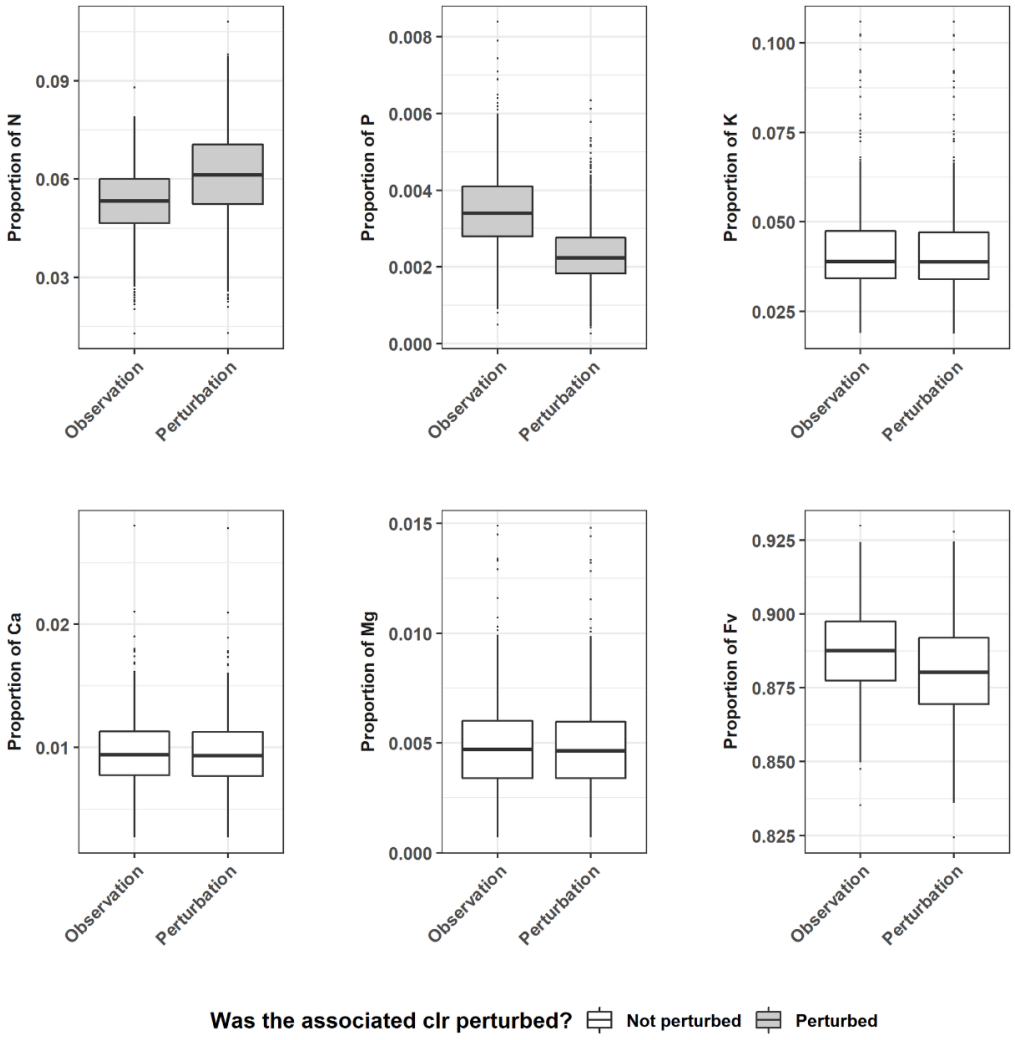


Figure 2.4. Effect of the perturbation of N and P clr coordinates on the other element proportions. 'Observation' stands for the element's original proportion, 'Perturbation' designates the new proportion after the 'Observed' vector's clr value was offset. Greyed boxplots plot distribution of perturbed elements of the simplex.

Perturbation (as defined in equation 2.3) is the measure of compositional change from one composition to another (Aitchison and Ng, 2005). Because foliar composition

belongs to compositional data family, the Figure 2.4 illustrates the principle that changing a proportion of such data affects at least another proportion of the simplex (Aitchison, 1986). The result displayed variable offsets for other elements, decreasing or increasing to reach another balance in the simplex.

2.7 Conclusion

Since the concept of compositional data analysis was applied to plant tissues, several studies classified plant species and cultivars using multivariate analysis of nutrients compositions. This study is, to our knowledge, the third (following Parent et al. (1994) and Hernandes et al. (2011)) to use statistical tools to address the differential nutrition of potato cultivars using combination of nutrient concentrations in the diagnostic leaf, and the first using tools of machine learning to predict tuber marketable yield. The potato ionomes showed some dissimilarities in principle components analysis, but not compelling to separate definite density-based clusters between cultivars on the basis of the clr values. However, the ionome showed a determinant effect on tubers yield. Used as predictors in machine learning tools, clr variables showed diagnostic potential to detect in-season nutrient imbalance to address objectively the differential response of cultivars to fertilization. The perturbation vector of the leaf compositional space could indicate cultivar sensitivity to fertilization and address specific problems of nutrient imbalance in new cultivars. Tissue testing remains an informative, diagnostic, and preventive tool with real-world applications for growers in evaluating the effectiveness of their nutrient management program. When using the right interpretation, this timely and right tissue testing helps diagnose suspected nutrient deficiency presence and magnitude. By using the compositional perturbation vector involving interactions among nutrients, our study provided a useful tool in potato precision fertilization in Quebec. The perturbation vector can help identify limiting nutrients requiring correcting measures as season progresses or for subsequent seasons. Moreover, our study implicitly provided robust multi-nutrient norms for potato crop gathering more cultivars of different maturity classes than the previous works. These norms are sets of true negative or nutritionally balanced compositions per cultivar (enchanting islands) with high-yield potential. More data are needed to fine tune the models, especially for poorly documented cultivars. New algorithms, other sampling methods and model quality measures could be tested to deal with the problem of small-

data and imbalanced classification. Further studies extending predictive features to site-specific conditions could improve the diagnosis with a site- and-cultivar specific nutrient diagnosis model.

2.8 Acknowledgements

The authors acknowledge the financial support of the Natural Sciences and Engineering Council of Canada (CRDPJ 385199-09 and DG-2254), the Quebec Ministry of Agriculture, Fisheries and Food (IA216581), Centre SEVE, Patate Dolbec Inc. (St-Ubalde, QC), Groupe Gosselin FG (St-Augustin-de-Desmaures, QC), Agriparmentier Inc. (Notre-Dame-du-Bon-Conseil, QC), Ferme Daniel Bolduc Inc. (Péribonka, QC), Patate Laurentienne (Notre-Dame-de-la-Paix, QC), Ferme Bergeron-Niquet (Péribonka, QC) and Patates Lac-St-Jean (Péribonka, QC).

2.9 References

Adesanwo, O. O., Ige, D. V., Thibault, L., Flaten, D., and Akinremi, W. (2013). Comparison of colorimetric and ICP methods of phosphorus determination in soil extracts. *Communications in Soil Science and Plant Analysis, 44*(21), 3061-3075. doi:10.1080/00103624.2013.832771

Aitchison, J. (1986). The statistical analysis of compositional data. *London: Chapman and Hall*.

Aitchison, J., and Ng, K. W. (2005). The role of perturbation in compositional data analysis. *Statistical Modelling, 5*(2), 173-185.

Andrews, M., Raven, J. A., and Lea, P. J. (2013). Do plants need nitrate? The mechanisms by which nitrogen form affects plants. *Annals of Applied Biology, 163*(2), 174-199. doi:10.1111/aab.12045

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*(8), 2127-2131. doi:10.1093/jxb/erv040

Boiteau, G., Goyer, C., Rees, H. W., and Zebarth, B. J. (2014). Differentiation of potato ecosystems on the basis of relationships among physical, chemical and biological soil parameters. *Canadian Journal of Soil Science, 94*(4), 463-476. doi:10.4141/cjss2013-095

Borcard, D., Gillet, F., and Legendre, P. (2018). *Numerical ecology with R*: Springer.

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA, 1*, 58.

Brownlee, J. (2020). *Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning*.

Campbell, C. R. (2000). *Reference sufficiency ranges for plant analysis in the southern region of the United States* (SAAESD Ed. Vol. Bulletin #394): SAAESD.

CFIA. (2015). Potato plants characteristics, maturity. Canadian Food Inspection Agency. Retrieved from http://www.inspection.gc.ca/plants/potatoes/characteristics/eng/1326490397702/1326490477981#mature

Cottenie, A. (1980). Soil and plant testing as a basis of fertilizer recommendations. *F.A.O. Soils Bulletin, 38*(2), 1-118.

de Deus, J. A. L., Neves, J. C. L., Correa, M. C. D., Parent, S. E., Natale, W., and Parent, L. E. (2018). Balance design for robust foliar nutrient diagnosis of "Prata" banana (*Musa* spp.). *Scientific Reports, 8*, 1-7. doi:10.1038/s41598-018-32328-y

Dray, S., and Dufour, A. B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software, 22*(4), 1-20.

Dumenil, L. C. (1958). *Relationship between the chemical composition of corn leaves and yield responses from nitrogen and phosphorus fertilizer* Iowa State University Capstones, Retrieved from https://lib.dr.iastate.edu/rtd/2277

Egozcue, J. J., and Pawlowsky-Glahn, V. (2006). Simplicial geometry for compositional data. In A. Buccianti, G. H. MateuFigueras, and V. GlahnPawlowsky (Eds.), *Compositional Data Analysis in the Geosciences: From Theory to Practice* (Vol. 264, pp. 145-159).

Eschemback, V., Kawakami, J., and Melo, P. E. d. (2017). Performance of modern and old, European and national potato cultivars in different environments. *Horticultura Brasileira, 35*, 377-384.

Filzmoser, P., and Gschwandtner, M. (2018). mvoutlier: Multivariate Outlier Detection Based on Robust Methods. R package version 2.0.9.

Filzmoser, P., and Hron, K. (2011). "Robust statistical analysis. Chapter 5. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional data analysis: Theory and applications* (pp. 59–72): John Wiley and Sons, New York, NY.

Giletto, C. M., and Echeverría, H. E. (2015). Critical nitrogen dilution curve in processing potato cultivars. *American Journal of Plant Sciences, 6*(19), 3144 - 3156.

Hahsler, M., Piekenbrock, M., Arya, S., and Mount, D. (2017). dbscan: Density based clustering of applications with noise (DBSCAN) and related algorithms. R package version 1.1-3.

Herman, D. J., Knowles, L. O., and Knowles, N. R. (2017). Heat stress affects carbohydrate metabolism during cold-induced sweetening of potato (*Solanum tuberosum* L.). *Planta, 245*(3), 563-582. doi:10.1007/s00425-016-2626-z

Hernandes, A., Parent, S.-É., Veillette, J.-P., Parent, P., Leblanc, M., Roy, G., Sylvestre, P., Samson, N., Natale, W., and Parent, L.-É. (2011). Compositional meta-analysis of the nutrient profile of potato cultivars.

Hochmuth, G. J., Maynard, D., Vavrina, C., Hanlon, E., and Simonne, E. (2004). Plant tissue analysis and interpretation for vegetable crops in Florida. In (pp. 1-48).

Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods* (Third ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Hron, K. (2009). Analytical representation of ellipses in the Aitchison geometry and its application. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica, 48*(1), 53-60.

Huang, X. Y., and Salt, D. E. (2016). Plant Ionomics: From Elemental Profiling to Environmental Adaptation. *Molecular Plant, 9*(6), 787-797. doi:10.1016/j.molp.2016.05.003

Ivanov, K., Zaprjanova, P., Angelova, V., Bekjarov, G., and Dospatliev, L. (2010). *ICP determination of phosphorous in soils and plants.* Paper presented at the 19th World Congress of Soil Science, Soil Solutions for a Changing World.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning-with applications in R. In: New York, NY: Springer.

Jones, J. J. B., Wolf, B., and Mills, H. A. (1991). *Plant analysis handbook. A practical sampling, preparation, analysis, and interpretation guide*: Micro-Macro Publishing, Inc.

Kawakami, J., Iwama, K., Jitsuyama, Y., and Zheng, X. (2004). Effect of cultivar maturity period on the growth and yield of potato plants grown from microtubers and conventional seed tubers. *American Journal of Potato Research, 81*(5), 327-333.

Kuhn, M., and Johnson, K. (2013). *Applied predictive modeling*. In (pp. 595). Retrieved from http://dx.doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., Wing, J., Weston, S., and Williams, A. (2008). Caret package: classification and regression training *Journal of statistical software, 28*(5), 1-26.

Lahner, B., Gong, J. M., Mahmoudian, M., Smith, E. L., Abid, K. B., Rogers, E. E., Guerinot, M. L., Harper, J. F., Ward, J. M., McIntyre, L., Schroeder, J. I., and Salt, D. E. (2003). Genomic scale profiling of nutrient and trace elements in Arabidopsis thaliana. *Nature Biotechnology, 21*(10), 1215-1221. doi:10.1038/nbt865

Legendre, P., and Legendre, L. (2012). Cluster analysis. In P. Legendre and L. Legendre (Eds.), *Developments in environmental modelling. Numerical ecology* (Vol. 24, pp. 337-424): Elsevier.

Leite, M. L. C. (2016). Applying compositional data methodology to nutritional epidemiology. *Statistical Methods in Medical Research, 25*(6), 3057-3065. doi:10.1177/0962280214560047

Leite, M. L. C., and Prinelli, F. (2017). A compositional data perspective on studying the associations between macronutrient balances and diseases. *European Journal of Clinical Nutrition, 71*(12), 1365-1369. doi:10.1038/ejcn.2017.126

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22.

Marschner, H. (1995). Diagnosis of deficiency and toxicity of mineral nutrients. In H. Marschner (Ed.), *Mineral Nutrition of Higher Plants* (2e ed., pp. 461-479): Academic Press, London.

McKenzie, R. (1998). Crop nutrition and fertilizer requirements. *Alberta Agriculture, Food and Rural Development. Lethbridge*, 1-7.

McKenzie, R. H., Stewart, J. W. B., Dormaar, J. F., and Schaalje, G. B. (1992). Long-term crop rotation and fertilizer effects on phosphorus transformations: I. In a Chernozemic soil. *Canadian Journal of Soil Science, 72*(4), 569-579.

Melo, G. W., Rozane, D. E., Brunetto, G., and Lattuada, D. S. (2018). Discriminant analysis in the selection of groups of peach cultivars. In T. Mimmo, Y. Pii, and F. Scandellari (Eds.), *Viii International Symposium on Mineral Nutrition of Fruit Crops* (Vol. 1217, pp. 335-342).

Mills, H. A. J. J., Walsh, L. M. B., James, D., CottenieE, A., Faithfull, N. T., Larrahondo, J. E., Palau, F. J., Ramirez, J., Lopez, A., Eepinosa, J., Vargas, A., and Malavolta, E. (1996). *Plant analysis handbook II: a practical preparation, analysis, and interpretation guide*: Potash and Phosphate Institute.

Monna, F., Marques, A. N., Guillon, R., Losno, R., Couette, S., Navarro, N., Dongarra, G., Tamburo, E., Varrica, D., Chateau, C., and Nepomuceno, F. O. (2017). Perturbation vectors to evaluate air quality using lichens and bromeliads: a Brazilian case study. *Environmental Monitoring and Assessment, 189*(11). doi:10.1007/s10661-017-6280-0

Natale, W., Lima Neto, A. J., Rozane, D. E., Parent, L.-É., and Corrêa, M. C. M. (2018). Mineral nutrition evolution in the formation of fruit tree rootstocks and seedlings. *Revista Brasileira de Fruticultura, 40*(6), (e-133).

Nicolas, O., Charles, M. T., Jennie, S., Toussaint, V., Parent, S. E., and Beaulieu, C. (2019). The ionomics of lettuce infected by Xanthomonas campestris pv. vitians. *Frontiers in plant science, 10*, 1-10. doi:10.3389/fpls.2019.00351

Parent, L. E., Cambouris, A. N., and Muhawenimana, A. (1994). Multivariate diagnosis of nutrient imbalance in potato crops. *Soil Science Society of America Journal, 58*(5), 1432-1438. doi:10.2136/sssaj1994.03615995005800050022x

Parent, L. E., and Dafir, M. (1992). A theoretical concept of compositional nutrient diagnosis. *Journal of the American Society for Horticultural Science, 117*(2), 239-242.

Parent, S. E. (2020). Why we should use balances and machine learning to diagnose ionomes. *Authorea*. https://www.authorea.com/users/23640/articles/281937-why-we-should-use-balances-and-machine-learning-to-diagnose-ionomes doi:https://doi.org/10.22541/au.157954751.17355951

Parent, S. E., Parent, L. E., Egozcue, J. J., Rozane, D. E., Hernandes, A., Lapointe, L., Hebert-Gentile, V., Naess, K., Marchand, S., Lafond, J., Mattos, D., Barlow, P., and Natale, W. (2013a). The plant ionome revisited by the nutrient balance concept. *Frontiers in plant science, 4*. doi:10.3389/fpls.2013.00039

Parent, S. E., Parent, L. E., Rozane, D. E., Hernandes, A., and Natale, W. (2012). Nutrient balance as paradigm of plant and soil chemometrics. Chapter 4. In R. N. Issaka (Ed.), *Soil Fertility* (pp. 83-114): Tech Publ, NY.

Parent, S. E., Parent, L. E., Rozane, D. E., and Natale, W. (2013b). Plant ionome diagnosis using sound balances: case study with mango (*Mangifera Indica*). *Frontiers in plant science, 4*, 1-12.

Pawlowsky-Glahn, V., and Buccianti, A. (2011). *Compositional data analysis. Theory and applications*: A John Wiley & Sons, Ltd, Publication.

Prater, C., Scott, D. E., Lance, S. L., Nunziata, S. O., Sherman, R., Tomczyk, N., Capps, K. A., and Jeyasingh, P. D. (2019). Understanding variation in salamander ionomes: A nutrient balance approach. *Freshwater Biology, 64*(2), 294-305. doi:10.1111/fwb.13216

R Core Team. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Rich, A. E. (1983). *Potato diseases*. New York: Academic Press.

Rozane, D. E., Mattos, D., Parent, S. E., Natale, W., and Parent, L. E. (2015). Meta-analysis in the selection of groups in varieties of citrus. *Communications in Soil Science and Plant Analysis, 46*(15), 1948-1959. doi:10.1080/00103624.2015.1069307

Rozane, D. E., Mattos Junior, D. d., Parent, S. E., Natale, W., and Parent, L. E. (2011). *Compositional meta-analysis of citrus varieties in the state of São Paulo, Brazil.* Paper presented at the 4th International Workshop on Compositional Data Analysis, Saint Feliu de Giuxols, Girona, Spain.

Salt, D. E., Baxter, I., and Lahner, B. (2008). Ionomics and the study of the plant ionome. *Annual Review of Plant Biology, 59*, 709-733. doi:10.1146/annurev.arplant.59.032607.092942

Saric, M. R. (1983). Theoretical and practical approaches to the genetic specificity of mineral-nutrition of plants. *Plant and Soil, 72*(2-3), 137-150. doi:10.1007/bf02181954

Sattelmacher, B., Klotz, F., and Marschner, H. (1990). Influence of the nitrogen level on root growth and morphology of two potato varieties differing in nitrogen acquisition. *Plant and Soil, 123*(2), 131-137.

Sikora, F. J., Howe, P. S., Hill, L. E., Reid, D. C., and Harover, D. E. (2005). Comparison of colorimetric and ICP determination of phosphorus in Mehlich3 soil extracts. *Communications in Soil Science and Plant Analysis, 36*(7-8), 875-887. doi:10.1081/css-200049468

Söğüt, T., and Öztürk, F. (2011). Effects of harvesting time on some yield and quality traits of different maturing potato cultivars. *African Journal of Biotechnology, 10*(38), 7349-7355.

Stalham, M. A., Allen, E. J., and Herry, F. X. (2005). Effects of soil compaction on potato growth and its removal by cultivation. Research review. (R261), 1-60.

Statistics Canada. (2017). Area, production and farm value of potatoes. Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3210035801&pickMembers%5B0%5D=1.6&request_locale=en

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293.

Tolosana-Delgado, R., and Van Den Boogart, K. G. (2011). Linear models with compositions in R. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional data analysis: Theory and applications* (pp. 356-371): (New York: John Wiley and Sons).

Valkama, E., Uusitalo, R., Ylivainio, K., Virkajarvi, P., and Turtola, E. (2009). Phosphorus fertilization: a meta-analysis of 80 years of research in Finland. *Agriculture Ecosystems & Environment, 130*(3-4), 75-85. doi:10.1016/j.agee.2008.12.004

Van den Boogaart, K. G., Raimon, T., and Bren, M. (2014). compositions: compositional data analysis. R package version 1.40-1.

Westermann, D. T., and Davis, J. R. (1992). Potato nutritional management changes and challenges into the next century. *American Potato Journal, 69*(11), 753-767. doi:10.1007/bf02853817

White, P. J., Broadley, M. R., Thompson, J. A., McNicol, J. W., Crawley, M. J., Poulton, P. R., and Johnston, A. E. (2012). Testing the distinctness of shoot ionomes of angiosperm families using the Rothamsted Park Grass Continuous Hay Experiment. *New Phytologist, 196*(1), 101-109.

White, P. J., Wheatley, R. E., Hammond, J. P., and Zhang, K. (2007). Minerals, soils and roots. In *Potato Biology and Biotechnology* (pp. 739-752): Elsevier.

Zebarth, B. J., Karemangingo, C., Scott, P., Savoie, D., and Moreau, G. (2007). Nitrogen management for potato: general fertilizer recommendations. *New-Brunswick Ministry of Agriculture, Fisheries and Aquaculture, Fredericton, NB, Canada*.

Zebarth, B. J., Leclerc, Y., Moreau, G., and Botha, E. (2004). Rate and timing of nitrogen fertilization of Russet Burbank potato: Yield and processing quality. *Canadian Journal of Plant Science, 84*(3), 855-863. doi:10.4141/p03-123

Zebarth, B. J., Tai, H. L., Luo, S. N., Millard, P., De Koeyer, D., Li, X. Q., and Xiong, X. Y. (2011). Differential gene expression as an indicator of nitrogen sufficiency in field-grown potato plants. *Plant and Soil, 345*(1-2), 387-400. doi:10.1007/s11104-011-0793-z

2.10 Supporting information

S 2.1 Table: Quebec potato leaves ionome data set. raw_leaf_df.csv file available online in data repository at https://git.io/Jvt2r.

S 2.2 Table. Description of potato data set used for cluster analysis.

Cultivar	Maturity class	Number obs.		Yield cut-off	Median clr values					
		High yielders	Total	Mg ha^{-1}	N	P	K	Ca	Mg	Fv
AC Belmont	early	21	60	29.87	0.70	-2.01	0.66	-1.11	-1.62	3.46
AC Chaleur	early	6	23	16.70	0.70	-1.97	0.35	-1.04	-1.54	3.51
Amandine	mid-season	2	6	35.42	0.52	-2.29	0.56	-0.70	-1.62	3.53
Ambra	mid-season	2	6	53.69	0.71	-1.98	0.28	-1.08	-1.39	3.46
Andover	early mid-season	11	30	39.97	0.87	-2.17	0.74	-0.81	-2.28	3.57
Aquilon	mid-season	23	66	24.60	0.82	-1.86	0.28	-0.91	-1.83	3.48
Argos	late	7	20	45.07	0.77	-1.77	0.57	-1.28	-1.70	3.43
Atlantic	mid-season	88	184	36.82	0.78	-1.89	0.17	-1.10	-1.50	3.57
Bijou Rouge	early	9	24	36.53	0.94	-1.93	0.56	-1.12	-1.93	3.61
Carolina	early	3	8	39.03	0.60	-2.19	0.13	-0.84	-1.34	3.63

Chieftain	mid-season	119	258	36.89	0.77	-1.92	0.40	-1.06	-1.91	3.60
Coastal Russet	mid-season	58	165	33.12	0.67	-1.86	0.39	-0.95	-1.84	3.51
Dark Red Chieftain	mid-season	8	22	36.81	0.96	-1.71	0.36	-1.37	-1.86	3.65
Estima	mid-season	12	35	54.85	0.93	-2.13	0.58	-0.98	-1.83	3.53
FL 1207	mid-season late	127	346	36.33	0.74	-2.09	0.28	-0.93	-1.51	3.45
FL 1533	mid-season	64	183	36.20	0.63	-1.80	0.22	-1.13	-1.35	3.47
Frontier Russet	mid-season	9	24	34.28	0.52	-2.20	0.62	-0.37	-2.06	3.58
Goldrush	mid-season	197	560	41.33	0.90	-1.97	0.56	-1.11	-1.95	3.57
Harmony	mid-season	2	5	33.98	0.88	-2.10	0.62	-1.01	-2.03	3.64
Kanona	mid-season	2	5	20.55	0.91	-1.99	0.60	-1.29	-1.87	3.63
Kennebec	mid-season	68	189	33.30	0.71	-1.88	0.58	-0.93	-2.12	3.64
Keuka Gold	mid-season late	2	5	31.26	0.88	-2.08	0.33	-1.26	-1.56	3.69
Krantz	mid-season	2	6	28.01	0.93	-1.88	0.42	-0.78	-2.18	3.50
Lamoka	late	2	5	29.10	0.84	-2.26	0.32	-1.14	-1.60	3.84
Lanorma	mid-season	3	9	36.82	1.01	-2.38	0.55	-1.29	-1.68	3.84
Mystere	late	24	68	33.64	0.63	-2.01	0.51	-0.80	-1.78	3.43

Nordonna	mid-season	4	10	31.64	0.79	-2.15	0.40	-0.99	-1.62	3.56
Norland	early mid-season	21	48	31.10	0.84	-1.94	-0.10	-1.05	-1.60	3.86
Peribonka	early mid-season	2	6	38.81	0.95	-1.87	0.49	-1.13	-1.98	3.54
Pike	mid-season	5	14	39.98	0.95	-1.95	0.58	-1.43	-1.89	3.71
Pommerelle	late	16	44	46.83	0.60	-2.22	0.62	-0.86	-1.51	3.35
Prospect	mid-season	9	24	37.34	0.90	-2.04	0.62	-1.11	-1.88	3.58
Reba	mid-season	9	24	42.92	0.89	-1.95	0.32	-0.97	-1.78	3.52
Red Cloud	mid-season	4	12	45.41	0.74	-1.61	0.14	-1.13	-1.58	3.46
Red Maria	late	5	15	64.59	0.85	-1.77	0.53	-1.28	-1.97	3.65
Roko	mid-season	6	18	42.51	0.67	-2.30	0.81	-1.02	-1.55	3.37
Russet Burbank	late	26	29	32.50	0.90	-2.10	0.24	-0.91	-1.70	3.55
Russet Norkota	early	6	16	49.05	0.97	-1.94	0.50	-1.38	-1.64	3.52
Shepody	mid-season	32	112	29.50	0.72	-2.12	0.59	-0.97	-1.81	3.58
Sifra	mid-season late	3	8	50.01	0.60	-2.40	0.25	-0.76	-1.07	3.40
Snowden	late	66	188	33.26	0.68	-1.93	0.49	-0.90	-2.06	3.67
Superior	early mid-season	199	367	32.60	0.73	-1.99	0.57	-0.98	-1.95	3.58

Viking	mid-season	2	5	26.02	0.80	-2.18	0.18	-0.96	-1.44	3.61
Vivaldi	early mid-season	11	24	59.07	0.53	-2.16	0.51	-0.47	-1.74	3.30
W 1386	mid-season late	4	12	26.65	0.91	-2.16	0.23	-0.76	-1.73	3.51
Waneta	late	6	16	31.50	0.77	-2.16	0.30	-0.98	-1.50	3.35
Yukon Gold	mid-season	27	78	30.66	0.68	-2.17	0.44	-0.89	-1.61	3.47
		1334	3382							

S 2.3 Table. True negatives mean clr values for cultivars.

Cultivar*	clr_N	clr_P	clr_K	clr_Mg	clr_Ca	clr_Fv
AC Belmont	0.703	-2.120	0.622	-1.623	-1.072	3.491
AC Chaleur	0.704	-1.932	0.360	-1.553	-1.074	3.496
Amandine	0.519	-2.290	0.557	-1.616	-0.696	3.526
Ambra	0.713	-1.978	0.275	-1.393	-1.082	3.464
Andover	0.904	-2.203	0.743	-2.194	-0.850	3.601
Aquilon	0.810	-1.939	0.404	-1.834	-0.894	3.453
Argos	0.819	-1.754	0.402	-1.698	-1.131	3.362
Atlantic	0.775	-1.901	0.132	-1.503	-1.069	3.568
Bijou Rouge	0.962	-2.059	0.598	-2.164	-1.012	3.676
Carolina	0.595	-2.179	0.114	-1.324	-0.833	3.627
Chieftain	0.790	-1.870	0.385	-1.879	-1.037	3.611
Coastal Russet	0.693	-1.838	0.413	-1.844	-0.973	3.550

Dark Red Chieftain	0.953	-1.748	0.307	-1.920	-1.351	3.759
Estima	0.973	-2.111	0.549	-1.959	-1.049	3.597
FL 1207	0.768	-2.060	0.252	-1.502	-0.920	3.461
FL 1533	0.625	-1.826	0.212	-1.380	-1.095	3.465
Frontier Russet	0.532	-2.484	0.575	-2.065	-0.164	3.606
Goldrush	0.886	-1.949	0.553	-1.943	-1.129	3.583
Harmony	0.941	-2.070	0.616	-2.106	-1.072	3.690
Kanona	0.913	-1.990	0.605	-1.871	-1.291	3.635
Kennebec	0.654	-2.051	0.693	-2.041	-0.873	3.617
Keuka Gold	0.916	-2.061	0.297	-1.591	-1.199	3.637
Krantz	0.930	-1.879	0.415	-2.178	-0.784	3.496
Lanorma	1.005	-2.377	0.472	-1.684	-1.260	3.843
Mystere	0.621	-2.029	0.515	-1.740	-0.791	3.424
Nordonna	0.800	-2.073	0.344	-1.599	-0.994	3.521
Norland	0.810	-2.029	0.030	-1.588	-1.003	3.781

Peribonka	0.938	-1.788	0.489	-1.989	-1.153	3.503
Pike	0.974	-1.983	0.590	-1.870	-1.460	3.749
Pommerelle	0.590	-2.231	0.588	-1.487	-0.830	3.369
Prospect	0.910	-2.011	0.530	-1.908	-1.113	3.591
Reba	0.868	-1.951	0.320	-1.774	-0.983	3.521
Red Cloud	0.765	-1.604	0.137	-1.587	-1.132	3.422
Red Maria	0.837	-1.739	0.539	-1.964	-1.320	3.648
Roko	0.666	-2.304	0.820	-1.536	-1.001	3.356
Russet Burbank	0.889	-2.076	0.259	-1.699	-0.967	3.595
Russet Norkota	0.955	-1.930	0.516	-1.673	-1.387	3.519
Shepody	0.737	-2.065	0.566	-1.831	-0.992	3.585
Snowden	0.680	-1.980	0.374	-1.883	-0.845	3.655
Superior	0.733	-1.968	0.572	-1.891	-1.038	3.593
Viking	0.836	-2.156	0.204	-1.446	-1.012	3.575
Vivaldi	0.537	-2.149	0.513	-1.724	-0.479	3.302

W 1386	0.921	-2.180	0.241	-1.747	-0.756	3.520
Waneta	0.758	-2.207	0.424	-1.485	-0.920	3.430
Yukon Gold	0.747	-2.155	0.395	-1.589	-0.907	3.509

** Cultivars Lamoka and Sifra had no true negative specimens.*

Chapitre 3 : Site-specific machine learning NPK predictive model for high potato yield and quality

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

[1] Department of Soils and Agrifood Engineering, Université Laval, Québec City, Quebec, Canada

E-mail: zonlehoua.coulibali.1@ulaval.ca

[2] Quebec Research and Development Centre, Agriculture and Agri-Food Canada, Québec City, Quebec, Canada

E-mail: athyna.cambouris@canada.ca

* Corresponding author

Email: serge-etienne.parent.1@ulaval.ca

[¶] These authors wrote the paper and ran all computations.

[&] This author reviewed the paper.

3.1 Résumé

La modélisation statistique est couramment utilisée pour relier les performances de la pomme de terre (*Solanum tuberosum* L.) et les besoins en fertilisants au lieu de ne servir qu'à décrire des effets. La prescription des doses optimales de nutriments est difficile en raison de l'implication de nombreuses variables, notamment de la météorologie, du sol, de la gestion, la génétique, les ravageurs, les maladies, etc. Lorsque des données suffisantes sont disponibles, les algorithmes d'apprentissage automatique peuvent être utilisés pour prédire les performances des cultures. L'objectif de cette étude était de prédire le rendement et la qualité des tubercules (taille et poids spécifique) en fonction de la fertilisation azotée, phosphorée et potassique ainsi que des variables météorologiques, des sols et de la gestion du champ. Cette étude a utilisé les données de 273 essais menées au champ de 1979 à 2017 au Québec (Canada). Les modèles de Mitscherlich et les algorithmes des k plus proches voisins, des forêts aléatoires, des réseaux de neurones et de processus gaussiens ont été ajustés aux données divisées en un ensemble d'entraînement et de test, puis comparés pour la qualité d'ajustement et les prédictions sur les données test. Les modèles basés sur les algorithmes d'apprentissage automatique se sont ajustés aux données test avec des coefficients de détermination (R^2) variant entre 0,49 et 0,59 pour la prédiction du rendement, plus élevés que celui du modèle de Mitscherlich (0,37). Bien que les différences soient faibles, les R^2 des modèles de prédiction de la balance des tubercules de taille moyenne (R^2 = 0,60 – 0,69) et du poids spécifique (R^2 = 0,58 – 0,67) étaient plus élevés comparés aux R^2 de la balance des tubercules de grande taille (R^2 = 0,55 – 0,64) et du rendement vendable. Les surfaces de réponse du modèle de Mitscherlich, des réseaux de neurones et des processus gaussiens se sont montrées plus adéquates pour faires des inférences comparées aux courbes discontinues dérivées des modèles des k plus proches voisins et des forêts aléatoires. Enfin, des divergences apparaissent entre les modèles lorsqu'utilisés pour déterminer des doses économiques ou agronomiques optimales pour des conditions météorologiques, de sols et de gestion constantes. Les processus gaussiens sont apparus les plus appropriés en raison de leur capacité d'élaborer des recommandations probabilistes.

3.2 Abstract

Statistical modeling is commonly used to relate the performance of potato (*Solanum tuberosum* L.) to fertilizer requirements. Prescribing optimal nutrient doses is challenging because of the involvement of many variables including weather, soils, land management, genotypes, and severity of pests and diseases. Where sufficient data are available, machine learning algorithms can be used to predict crop performance. The objective of this study was to determine an optimal model predicting nitrogen, phosphorus and potassium requirements for high tuber yield and quality (size and specific gravity) as impacted by weather, soils, and land management variables. We exploited a data set of 273 field experiments conducted from 1979 to 2017 in Quebec (Canada). We developed, evaluated, and compared predictions from a hierarchical Mitscherlich model, *k*-nearest neighbors, random forest, neural networks, and Gaussian processes. Machine learning models returned R^2 values of 0.49–0.59 for tuber marketable yield prediction, which were higher than the Mitscherlich model R^2 (0.37). The models were more likely to predict medium-size tubers (R^2 = 0.60–0.69) and tuber specific gravity (R^2 = 0.58–0.67) than large-size tubers (R^2 = 0.55–0.64) and marketable yield. Response surfaces from the Mitscherlich model, neural networks and Gaussian processes returned smooth responses that agreed more with actual evidence than discontinuous curves derived from *k*-nearest neighbors and random forest models. When marginalized to obtain optimal dosages from dose-response surfaces given constant weather, soil and land management conditions, some disagreements occurred between models. Due to their built-in ability to develop recommendations within a probabilistic risk-assessment framework, Gaussian processes stood out as the most promising algorithm to support decisions that minimize economic or agronomic risks.

Keywords:

Precision fertilization, Mitscherlich model, *k*-nearest neighbors, random forest, neural networks, Gaussian process, economic optimal dose, agronomic optimal dose, *Solanum tuberosum* L.

3.3 Introduction

Modeling provides a quantitative understanding of how crop systems operate (Sinclair and Seligman, 2000). Site-specific simulations of fertilizer requirements to obtain high local potato yield and quality rely on models' ability to detect subtle variations in factors affecting plant growth and environment and to learn from the past to make predictions (Di Paola et al., 2016). Several crop models have been developed with different degrees of sophistication, scale, and representativeness (Di Paola et al., 2016). Mechanistic models have been published for potato cropping systems (Marshall, 2007; Raymundo et al., 2014). Semi-mechanistic growth models could be used to downscale tuber yield assessment from regional to field levels (MacKerron, 2007; Fortin et al., 2010). Multilevel modeling can assist in selecting a set of relevant parameters that impact tuber yield and fertilizer requirements, but can hardly predict site-specific nutrient requirements (Parent et al., 2017).

Several variables can impact fertilization at optimum tuber yield: soil type and quality (Stalham et al., 2005; Boiteau et al., 2014), organic fertilizers (Neeteson and Zwetsloot, 1989; Firman and Allen, 2007), preceding crops (Li et al., 1999; Sincik et al., 2008; Sharifi et al., 2009; Zebarth et al., 2009a; Zebarth et al., 2009b), weather conditions (Sands et al., 1979), irrigation (Cambouris et al., 2016), timing, location and chemical form of the fertilizer applied (Zebarth et al., 2004), pests and diseases (Raman and Radcliffe, 1992) and genetic factors such as cultivar longevity and growth rate (Gregory and Simmonds, 1992; Kooman et al., 1996). Air temperature, photoperiod, day length, intercepted radiation, water abundance, precipitations, root development and crop management were reported to be the driving variables for potato growth and development (Stalham et al., 2005; Fortin et al., 2008; Dessureault-Rompre et al., 2012; Boiteau et al., 2014; Dessureault-Rompre et al., 2015; Haverkort and Struik, 2015). While the nitrogen (N) requirement of potato crops compares with other high N-demanding crops, phosphorus (P) uptake depends largely on close contact between roots and soil particles that, in turn, depends on soil texture, buffering capacity and moisture content (Barber, 1995; White et al., 2007). Due to a shallow system of fine roots and small biomass (Bolinder et al., 2015), especially in compacted soils (Stalham et al., 2005; Boiteau et al., 2014), potato is sensitive to nutrient and water stresses (Diriba, 2017).

The N, P and K (potassium) requirements are thought to be cultivar- and market-specific (Gianquinto and Bona, 2000; Dampney et al., 2011; Hüwing, 2012). Specific gravity (SG) is of particular concern for North-American processors (Kirkman, 2007; Bohl and Johnson, 2010). Other characteristics, such as tuber size and grade are also valued (Bohl and Johnson, 2010). No model has yet addressed K requirements accounting for interactions between genetics, environment and management (Hatfield and Walthall, 2015).

Growers tend to over-fertilize because of the potential economic loss from under-fertilizing (Rajsic and Weersink, 2008; Parent, 2014). While N can cause nitrate contamination (Peralta and Stockle, 2002; Jiang et al., 2011; Zebarth et al., 2015a; Zebarth et al., 2015b) and P the eutrophication of surface waters (Khiari et al., 2000; Pellerin et al., 2006a; Pellerin et al., 2006b), K has no known deleterious effect on the quality of natural and drinking water. Attempts have been made to synthesize the results of fertilizer experiments using meta-analysis to derive N optima for specific soil texture and pH groups (Valkama et al., 2013) or multilevel modeling combining soil, climate indices and management variables (Parent et al., 2017). Even where field trials could identify nutrient optima (Hofman and Salomez, 2000), such optima cannot be generalized to conditions different from those of particular experiments (Kyveryga et al., 2007b, 2007a).

Although experimental data grow continuously in size and quality, it is still beyond researchers' ability to integrate, analyze and make the best-informed decisions. Machine learning is an emerging technology that can aid in the discovery of rules and patterns in large sets of data (Zhang and Jeffery, 2007). The technology bypasses intermediate processes otherwise explicitly explained by a mechanistic modeling system and makes predictions directly based on input data (Qin et al., 2018). Machine learning methods can combine fertilizer dosage, genetics, environmental and land management variables to predict tuber yield and quality. Classical models such as Mitscherlich are limited to plant-nutrient relationships (Dahnke and Olson, 1990).

We hypothesized that (1) genetics, environment and local land management practices are the main drivers of fertilizer requirements, (2) *k*-nearest neighbors, random forest, neural networks and Gaussian processes are more accurate in predicting marketable yield than classical Mitscherlich predictive models, and (3) the machine

learning algorithms are equally able to predict economic optimal or agronomic optimal fertilizer doses. The objective of this study was to develop, evaluate and compare the performance of machine learning models in predicting N, P and K requirements for potato.

3.4 Methodology

3.4.1 Data set

The Quebec (Canada) potato data set is a collection of field fertilizer trials conducted from 1979 to 2015 between the US border (45th parallel) and the Northern limit of cultivation (49th parallel). We added 17 trials conducted in 2016 and 2017. Figure 3.1 shows the location of experimental sites.

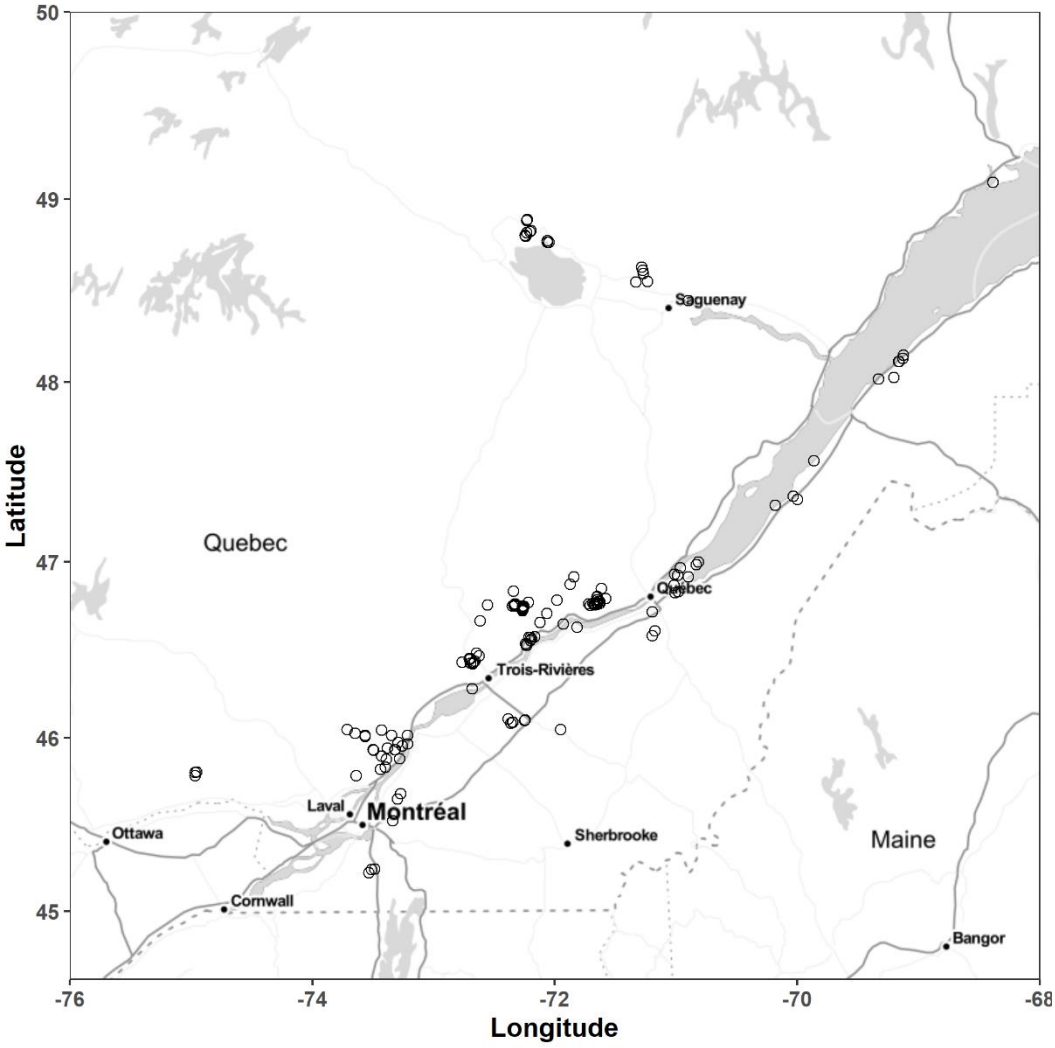


Figure 3.1. Location of experimental sites (Kahle and Wickham, 2013)

The trials with maximum yield less than 28 Mg ha^{-1} were discarded to avoid extreme cases of diseases, management failures or catastrophic weather events. The data set contains 4254 – 5913 observations from 208 – 273 field trials, depending on the number of missing values found in the target variable. Most experiments have been carried out from 1991 (Table 3.1). The number of trials, the number of samples, minimum and maximum number of blocks and treatments are given in S 3.1 and S 3.2 Tables according to the study year and the fertilizer tested.

Table 3.1. Global structure of the machine learning modeling data sets

A. Marketable yield model dataset

Period	Number of trials	Number of samples	Percentage (%)
1979-1980	2	20	0.3
1981-1990	4	38	0.6
1991-2000	68	1768	29.9
2001-2010	113	2386	40.4
2011-2020	86	1701	28.8
Total	273	5913	100.0

B. Tuber-size balances model dataset

Period	Number of trials	Number of samples	Percentage (%)
1971-1980	0	0	0.0
1981-1990	0	0	0.0
1991-2000	44	1196	26.2
2001-2010	81	1703	37.4
2011-2020	83	1658	36.4
Total	208	4557	100.0

C. Specific gravity model dataset

Period	Number of trials	Number of samples	Percentage (%)
1971-1980	0	0	0.0
1981-1990	0	0	0.0

1991-2000	61	1474	34.6
2001-2010	70	1144	26.9
2011-2020	83	1636	38.5
Total	214	4254	100.0

There were 48 cultivars classified as early (65 – 70 days), early mid-season (70 – 90 days), mid-season (90 – 110 days), mid-season late (110 – 130) or late maturity (130 days and more) as suggested on the website of the Canadian Food Inspection Agency (CFIA, 2015), with 4%, 13%, 62%, 12% and 9% of the samples respectively. The growing season lengths were provided by scouting teams covering the period from seeding to harvest. The names of the cultivar maturity classes consigned in the data set do not strictly match those of the Canadian Food Inspection Agency (CFIA, 2015). The preceding crop was categorized as in Parent et al. (2017) as grasslands, legumes, cereals, low-residue crops and high-residue crops (S 3.3 Table). The data set also includes fertilizers other than N, P or K (classified as NA), fertilizer dosage and application method, seeding density and date, harvest date, tuber marketable yield (excluding tubers < 2.5 cm in diameter), tuber size distribution (small, medium, large) and specific gravity.

3.4.2 Experimental procedures

The experiments included four to six treatments arranged mostly in a randomized complete block design with a minimum of three replications of each treatment (S 3.1 Table). One trial conducted in 1987 had two replications and 8% to 10% of the experiments were arranged as factorial design combining N, P and K fertilizers. We also retained one trial were N, P and K were fixed at their grower-optimum level (S 3.2 Table). Each experimental unit consisted of four or six rows measuring 6 or 8 m in length, with an average row spacing of 0.915 m and within-row spacing varying with cultivar. The potato seeds were planted in May (excepting June in the Outaouais region) then harvested in September. Median plant density was 36000 plants ha^{-1} in N trials, 33100 plants ha^{-1} in P trials, 36400 plants ha^{-1} in K trials, and 43700 plants ha^{-1} in factorial NPK trials. The N doses varied from 0 to 260 kg N ha^{-1} with varying steps, and P was applied at a dosage of 0 to 130 kg P ha^{-1} with varying steps. The K was applied at a dosage of 0 to 350 kg K ha^{-1} with varying steps. The P and K fertilizers could be converted to P_2O_5 and K_2O by multiplying P by 2.291 and K

by 1.205. Nitrogen fertilizers were either entirely applied at planting or split-applied between planting and hilling. Phosphorus fertilizers were banded at planting. Potassium fertilizers were band-applied or split-applied before planting and at planting. No animal manure or compost had been applied in the spring and the preceding fall. Other practices were managed uniformly by the grower.

At harvest, 3-m-long ridges in the middle two rows of each plot were dug and hand harvested. Tubers were divided into four categories as follows: culls, small (S), medium (M) or large (L), depending on the smallest diameter size measured with a ruler. The size cut-offs varied with cultivars and market. The marketable yield was calculated as total yield minus culls (tubers < 25 mm in size). Tubers with external defects such as secondary growth and soft rot were discarded. A representative sample of 20 medium-size tubers from each plot was used to determine tuber specific gravity.

3.4.3 Soil characteristics

3.4.3.1 Basic soil composition

Composite soil samples from the 0–20 cm layer were collected in the spring of the study year before planting to determine the initial soil physicochemical characteristics. Particle size distributions were measured as % clay (0 – 0.002 mm), % silt (0.002 – 0.05 mm), and % sand (0.05 – 2 mm) by sedimentation (Gee and Bauder, 1986) or laser diffraction (Yang et al., 2015). Where soil textural classes were not recorded, central values computed for sand, silt, and clay percentages (S 3.4 Table) using the Quebec soil data set (Tabi et al., 1990) were assigned as proxies.

Soil carbon concentration was determined using the Walkley-Black method (Nelson and Sommers, 1982) or Dumas combustion (Leco Instrument, Saint-Louis, MO). The two methods are closely related as in equation 3.1 (Grewal et al., 1991):

$$Dumas\ C(\%) = 0.126 + 1.25 * Walkley\text{-}Black\ C(\%) \qquad \text{(Equation 3.1)}$$

Because soil particle-size distribution and organic matter content are compositional data, they were transformed into isometric log-ratios (ilr) to avoid self-redundancy, non-normal distribution and scale dependency (Egozcue and Pawlowsky-Glahn, 2005). The ilr transformation consists in log ratios of the geometric means of hierarchically-arranged components and groups of components, and can be interpreted

as balances (Morton et al., 2017). The hierarchical arrangement of components follows a balance scheme where balances split groups of components sequentially until each group contains a single part. Each balance is computed as in equation 3.2:

$$ilr_j = \sqrt{\frac{r_j s_j}{r_j + s_j}} ln\left(\frac{g\left(c_j^+\right)}{g\left(c_j^-\right)}\right) \qquad \text{(Equation 3.2)}$$

where for the j^{th} balance in [1, ..., D-1] (D is the length of the compositional vector), r_j is the number of parts on the left-hand side, s_j is the number of parts on the right-hand side, c_j^- is the compositional vector at the left-hand side, c_j^+ is the compositional vector at the right-hand side, and g() is the geometric mean function. Hence, the textural components and carbon content were balanced as [Sand, Silt, Clay | C], [Clay | Sand, Silt] and [Silt | Sand]. We followed the [denominator parts | numerator parts] notation (Parent et al., 2013).

3.4.3.2 Soil pH

Soil pH was measured in water (1:1, v/v) or in a 0.01 M $CaCl_2$ solution (1:1 v/v) (Hendershot et al., 1993). The pH_{CaCl2} was converted into pH_{H2O} where required, as in equation 3.3 (Cescas, 1978):

$$pH_{water} = 0.27 + 1.03\ pH_{CaCl2} \qquad \text{(Equation 3.3)}$$

3.4.3.3 Soil Mehlich-3 extractable P, K, Al, Mg and Ca

Soil P was extracted using the Mehlich-3 method (Tran and Simard, 1993) or Bray-2 converted to P Mehlich-3 values using the Khiari et al. (2000) equation as in equation 3.4:

$$P_{Mehlich\text{-}3}\ (mg\ kg^{-1}) = -34.6 + 0.86 * P_{Bray\text{-}2}\ (mg\ kg^{-1}) \qquad \text{(Equation 3.4)}$$

Soil Al was extracted using the Mehlich-3 method or, where not available, from the typical Al-Mehlich-3 value of soil series as reported by Tabi et al. (1990). Soil K, Ca and Mg were extracted using the ammonium acetate method or its closely-related Mehlich-3 extractant (Michaelson et al., 1987). The P concentration was determined colorimetrically (Murphy and Riley, 1962) or by inductively coupled plasma (ICP). The

K concentration was determined by flame emission or ICP, and Ca, Mg, and Al concentrations were quantified by atomic absorption spectrometry or ICP.

Soil chemical compositions were partitioned into two simplexes *S*(P, Al) and *S*(K, Ca, Mg). The ilr variables were [Fv | Al, P], [Al | P] on the one hand and [Fv, Mg, Ca | K], [Fv | Mg, Ca], [Mg | Ca] on the other.

3.4.3.4 Soil profiles

The soils in our data set were classified according to the Canadian Soil Classification Working Group (1998) and ordered along a gleyzation-podzolization gradient using tools of pedometrics (Leblanc et al., 2016). Soil profile reflects the influence of subsoil on crop growth, in particular its impact in regulating the availability of water (Piikki et al., 2015). The continuous expressions for Quebec potato soil types defined by Leblanc et al. (2016) and used by Parent et al. (2017) *i.e.*, poorly-drained loam, poorly-drained sand and well-drained sand, were balanced as [Gleyed | Podzolized] and [Loamy gleyed | Sandy gleyed].

3.4.4 Weather data

Weather data were collected from the Environment Canada information system (Hutchinson et al., 2009) using geographical coordinates for each site. The selected weather indexes were the cumulative precipitation – PPT, the Shannon Diversity Index for rainfall distribution – SDI (Tremblay et al., 2012), the mean temperature, and the number of growing degree days – GDD.

The cumulative precipitation was computed as the sum of daily rainfall from planting to harvest. The Shannon Diversity Index is the precipitation evenness or the fraction of daily rainfall relative to the total rainfall in a given time period (in days). A SDI = 1 implies complete evenness *i.e.*, equal amounts of rainfall in each day of the period while a SDI = 0 implies complete unevenness *i.e.*, all rain in 1 day (Tremblay et al., 2012). The mean temperature was computed from the planting date to harvest date. The growing degree days index was computed using daily mean temperatures and using 5°C as baseline temperature (*i.e.*, sum of daily mean temperatures equal or superior to 5°C only). Weather variables were computed as in (Table 3.2) for the period between

planting and harvest dates using the historical weather data of the past 5 years (from the corresponding study year) at each site.

Table 3.2. Equations to compute climatic indices

Index	Description	Unit	Formula
PPT	Cumulative precipitations	mm	$PPT = \sum_{i=1}^{n} Rd_i$
SDI	Shannon Diversity Index for rainfall	Unitless	$SDI = \frac{-\sum_{i=1}^{n}[P_i\ ln(P_i)]}{ln(n)}$ $P_i = \frac{Rd_i}{PPT}$
T	Mean temperature	°C	$\frac{1}{n}\sum_{i=1}^{n} Tm_i$
GDD	Growing degree-days	°C	$GDD = \sum_{i=1}^{n} Tm_i\ with\ Tm_i \geq 5$

Rd is daily rainfall; n is the number of days and Tm is daily mean temperature.

3.4.5 Selection of features

3.4.5.1 Predictive features

The study focused on potato yield-impacting factors reported by Parent et al. (2017). Candidate variables were soil Mehlich-3 P, K, Mg, Ca, Al and Fe composition, soil pH, and soil profile classes expressed as balances across soil textural gradients and across gleization-podzolization processes as in Leblanc et al. (2016). The length of the growing season, the preceding crop categories, seeding density and N, P and K fertilizer dosages were used as land management variables. The average 5-yr temperature (T), PTT, GDD and SDI were used as weather features.

The importance of features can be assessed by assigning them a score based on how useful they are at predicting a target variable. We assessed features importance using *ExtraTreesRegressor* function from the scikit-learn Python package (Pedregosa et al., 2011) on the training set of each target variable.

3.4.5.2 Target variables

The data set is a collection of several experiments with specific objectives. Target variables were total yield, yield fractions, and SG. We separated marketable yield fractions with respect to tuber size as follows: large (L), medium (M) or small (S) size. Because these three fractions must add up to 100% of the marketable yield, they were treated as compositions. These compositional variables were transformed into isometric log-ratios of large-size tubers divided by the geometric mean of small- and medium-size tubers [M, S | L], and medium-size tubers divided by small-size tubers [S | M]. Since analysis of compositional data based on log-ratios of parts is not suitable when zeros are present in a data set (Martin-Fernandez et al., 2003), we proceeded by firstly imputing zero observations (Palarea-Albaladejo and Martin-Fernandez, 2015), reported mostly for large-size tubers. The detection limit was fixed at 65%. Table 3.3 summarizes the variables used for modeling. Tuber SG was determined by the weight-in-air to weight-in-water method (Young et al., 1964) as in equation 3.5:

$$SG = \frac{Weight\ in\ air}{Weight\ in\ air\ minus\ Weight\ in\ water} \qquad \text{(Equation 3.5)}$$

Table 3.3. Variables used for modeling

A. Predictive variables

Variable	Type	Description
N, P, K doses	Numeric (kg ha^{-1})	Fertilizer doses used during the experiments
Planting density	Numeric (plants ha^{-1})	The number of plants within 1 ha of area
Preceding crops	Categorical	Crop existing on the experimental site along the previous season categorized as small grain, high-residue crop, legume, grassland and low-residue crop (S 3.5 Table)
Growing season length	Numeric (days)	Number of days between planting and harvest
Temperature	Numeric (°C)	Average daily mean temperature from planting to harvest (Table 3.2) computed with temperature data through the five seasons preceding the season of the study

Precipitations	Numeric (mm)	Sum of daily rainfall from planting to harvest (Table 3.2) computed with data through the five seasons preceding the season of the study
Shannon diversity index	Numeric (unitless)	Precipitations evenness from planting to harvest **(**Table 3.2) computed with data through the five seasons preceding the season of the study
Number of growing degree days	Numeric (°C)	Sum of daily mean temperature from planting to harvest (Table 3.2) computed with temperature data through the five seasons preceding the season of the study (5°C as baseline)
Soil texture (0–20 cm) and carbon	Numeric (unitless)	Ilr coordinates: [Sand, Silt, Clay | C], [Clay | Sand, Silt] and [Silt | Sand]
Soil types	Numeric (unitless)	Ilr coordinates representing drainage capacity: [Gleyed | Podzolized] and [Loamy gleyed | Sandy gleyed]
Soil pH	Numeric (unitless)	Soil pH measured in water or expressed as pH in water
Soil chemical composition	Numeric (unitless)	P and its fixation agent Al, ilr coordinates: [Fv | Al, P], [Al | P]
	Numeric (unitless)	K, Ca and Mg, ilr coordinates: [Fv, Mg, Ca | K], [Fv | Mg, Ca], [Mg | Ca]

B. Target variables

Variable	Type	Description
Marketable yield	Numeric (Mg ha^{-1}), 1Mg = 1000 kg	Sum of small-, medium- and large-size tuber weight
Yield ratios	Numeric (unitless)	Ilr coordinates of large-size against small- and medium-size tuber weight [M, S | L], medium-size tubers against small-size tuber weight [S | M]
Tuber specific gravity	Numeric (unitless)	Ratio of tuber weight-in-air to weight-in-water (equation 3.5)

3.4.6 Data preprocessing

The data were partially preprocessed in the R 3.6.2 statistical computing environment (R Core Team, 2019). The tidyverse 1.3.0 package (Wickham, 2017) was used for general data handling and visualization. The compositions 1.40-3 package (Van den Boogaart et al., 2014) functions helped to transform compositional data into isometric log-ratios, and the robCompositions 2.2.0 package (Templ et al., 2011) helped to robustly impute missing values. The replacement of zeros in tuber sizes was performed using zCompositions 1.3.3-1 package (Palarea-Albaladejo and Martin-Fernandez, 2015).

The data preprocessing continued in Python 3.8.1 software (Van Rossum and Drake Jr, 1995). The data set used to model tuber SG was cleaned of outliers using the Python SciPy package version 1.4.1 (Virtanen et al., 2019). We used a z-score *i.e.*, a signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being measured on the multivariate data set. The threshold of the score value was set at 3. The data were handled in Python using NumPy version 1.17.5 (Van Der Walt et al., 2011) and pandas 1.0.0 (McKinney, 2010) libraries. The matplotlib 3.1.3 package (Hunter, 2007) was used for data visualization.

All the quantitative variables were scaled and centered to obtain zero mean and unit variance. The categorical variables were encoded by declining their factors in binary columns, each of which was denoted by 1 to specify the membership of the group of the column, and 0 otherwise.

3.4.7 Training and testing data sets

Schemes for partitioning data into training and testing sets vary between studies. Fortin et al. (2010) used 60% for training and 40% for testing. Parizeau (2006) suggested 50%, 20% and 30% for training, validation and testing, respectively. Crisci et al. (2012) used a 75%–25% split while Chantre et al. (2012) used a 82%–18% partition for training and testing, respectively. In this paper, the corresponding total input/output data pairs were divided into 70% for training and 30% for testing and model accuracy assessment. Soman and Bobbie (2005) found shorter learning times and highest accuracies with such split proportions. Moreover, self-contained and representative data collection is an important step to ensure the sufficiency and integrity

of the training data (Yuan et al., 2010). Thereby, we partitioned the data set according to whether the tested element was N, P K, factorial design, or another element (Mg, Ca). Thereafter, data were split at block level to avoid testing models on blocks comprising training samples.

3.4.8 Training models

3.4.8.1 Machine learning algorithms

Four machine learning models were trained to derive an optimal model: *k*-nearest neighbors (KNN), random forest (RF), neural networks (NN) and Gaussian processes (GP). Model parameters were tuned using the random search with cross-validation method (*RandomSearchCV*) of the scikit-learn library version v0.22.1 (Pedregosa et al., 2011).

3.4.8.2 Mitscherlich model

We used a Mitscherlich-related 3D response surface for three variables inspired by Dodds et al. (1995) in the multilevel modeling scheme of Parent et al. (2017). The Mitscherlich-related multilevel response surface was used as a predictive model for comparison with machine learning algorithms. The model was trained using the following equation:

$$Y = A\ x\left(1 - e^{-R_N \times (E_N + dose_N)}\right) \times \left(1 - e^{-R_P \times (E_P + dose_P)}\right) \times \left(1 - e^{-R_K \times (E_K + dose_K)}\right)$$

(Equation 3.6)

where Y is the target variable *i.e.*, marketable yield, A (for *Asymptote*) is the value of the target variable toward which the curve converges at increasing dosage, E (for *Environment*) describes the fertilizer-equivalent N (E_N), P (E_P) and K (E_K) doses from the environment, and R *(Rate)* is the steepness of the curve relating each fertilizer equivalent environmental supply to *Asymptote*. The first-level parameters (A, E and R) were modeled as linear combinations of the predictors with random effect added to the intercept of the *Asymptote*. To make comparison with preceding models, the model performances were computed without any random effect (*level = 0*). The Mitscherlich multilevel model was fitted in R 3.6.2.

3.4.9 Evaluation of model performance

In all cases, the goodness-of-fit measure or predictive capacity of the developed models was based on the coefficient of determination (R^2), the mean absolute error (MAE) and the root-mean-square error (RMSE). The R^2 evaluates the proportion of variance in the target variable explained by the model as in equation 3.7:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \qquad \text{(Equation 3.7)}$$

where y_i is the observed target variable value, $\hat{y}_i$ is the predicted target variable value, and $\bar{y}_i$ is the mean of observed target variable. The best possible score of R^2 is 1 (or 100%), but the score may also be negative when the model is arbitrarily worse. Higher R^2 values indicate less error variance. A constant model that always predicts the expected value of y disregarding the input features would yield a R^2 score of 0 (Pedregosa et al., 2011). Typically, values greater than 0.5 are considered acceptable (Moriasi et al., 2007). The MAE is the average of the absolute differences between predictions and observations as in equation 3.8:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad \text{(Equation 3.8).}$$

The MAE attributes equal weight to individual errors and is less sensitive than R^2 or RMSE to large prediction errors. The RMSE is the square root of the average of squared differences between predictions and observations computed as in equation 3.9:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \text{(Equation 3.9).}$$

The RMSE attributes high weight to large errors due to squaring. Both MAE and RMSE indicate prediction errors in the units of variable of interest. Zero values indicate a perfect fit. Values less than half of the standard deviation of measured data were considered low (Singh et al., 2005). The trained models were used to predict optimal N, P and K doses using some left-out experimental sites data.

3.4.10 Economic or agronomic optimal doses

The optimal nutrient input is the one returning yield of high-quality tubers (Hüwing, 2012), where profitability is maximized and the environmental footprint

minimized (Inman et al., 2005; Fortin et al., 2015). To compute the optimal economic N, P, K doses at a given site, all the predictive features, but not N, P and K doses, were held constant (fixed input data). The row of fixed input variables is stacked (reproduced) 1000 times to obtain a table with 1000 identical rows. We generated 1000 random N-P-K combinations of doses from uniform distributions of plausible doses varying between zero and 250 kg ha^{-1} for N, 110 kg ha^{-1} for P, 208 kg ha^{-1} for K. The table was altered in such a way that only N-P-K dosage changed following the random combinations.

A fertilizer cost was computed for each N-P-K triplet. Unit fertilizer costs were set at \$1.20 CDN kg^{-1} for N, \$1.10 CDN kg^{-1} for P and \$0.90 CDN kg^{-1} for K. Tuber price was set at \$250 CDN Mg^{-1} (1 Mg = 1000 kg) as in Parent et al. (2017). No environmental footprint effect was used because of a lack of reliable sources, although they could have been implemented as an increase in the cost of unit dosage. The difference between fertilizer cost and tuber revenue provided the marginal benefit from fertilizing. Economic optimal N-P-K dosage was reached where the net return was maximum. For tuber size and SG, an agronomic optimal N-P-K fertilizer dosage was deducted where the target variable reached a maximum.

Our results are reproducible by using the codes, data and package requirements provided in a GitHub repository at https://git.io/JvYxd.

3.4.11 Model interpretation data

We randomly selected four trials in the testing set for model interpretation (Table 3.4). The trials showed soil pH levels ranging between the adequate limits of 5.2 to 6.2 for potato crops according to the *Centre de Référence en Agriculture et Agroalimentaire du Québec* (CRAAQ, 2010). The phosphorus saturation environmental index $(P/Al)_{Mehlich3}$ classified the sites at extremely low environmental risk for P trials (1.4% to 1.6%), medium risk for N trial (11.1%) and very high risk for K trial (28.7%). Soil potassium levels showed extremely low (71.5 mg kg^{-1}) and very low (83.1 mg kg^{-1}) levels for P trials, medium level for K trial and high level for N trial (Pellerin, 2010).

Table 3.4. Description of trials used for model analysis

	Trial 194*	**Trial 8804**	**Trial 412**	**Trial 320**
Nutrient tested	P	N	P	K
Cultivar	Superior	FL 1533	Goldrush	Krantz
Maturity class	Early mid-season	Mid-season	Mid-season	Mid-season
Growing season length (days)	102	131	108	112
Planting density (seeds ha^{-1})	36430	43716	36433	31226
Mean temperature (5 years) T°C	16	18	16	18
Total rainfall (5 years) mm	378	359	363	448
Soil pH	5.5	5.5	5.8	6.1
Soil P (Mehlich 3) mg kg^{-1}	23	175	46	349
Soil K (Mehlich 3) mg kg^{-1}	83	265	72	200
Soil Al (Mehlich 3) mg kg^{-1}	1580	1570	2839	1216
ISP$_1$ (environmental index %)	1.4	11.1	1.6	28.7
Texture	Sandy loam	Fine sand	Sandy loam	Loam
Minimum dose (kg ha^{-1})	0	0	0	0
Maximum dose (kg ha^{-1})	300	200	200	300

* Trial n° 194 used for economic optimal and agronomic optimal doses computation. Al: aluminium, ISP: phosphorus saturation index

3.5 Results

3.5.1 Feature importance

The feature importance, computed using the *ExtratreesRegressor* function, revealed that the N fertilizer dose was by far the most informative feature in the marketable yield prediction models, followed by soil type, air temperature, length of growing season and soil texture. To predict large-size tuber yield ([M, S | L] balance), the N dose remained the most informative feature, followed by soil type and texture. Tuber planting density exceeded other features for medium-size tubers ([S | M] balance), followed by N dose, soil elements (P and Al Mehlich-3) and soil type. For tuber SG, weather indices, *i.e.*, Shannon diversity index, total rainfall and temperature, returned the highest scores (Figure 3.2). Preceding crops were not informative across target variables and were deleted before modeling.

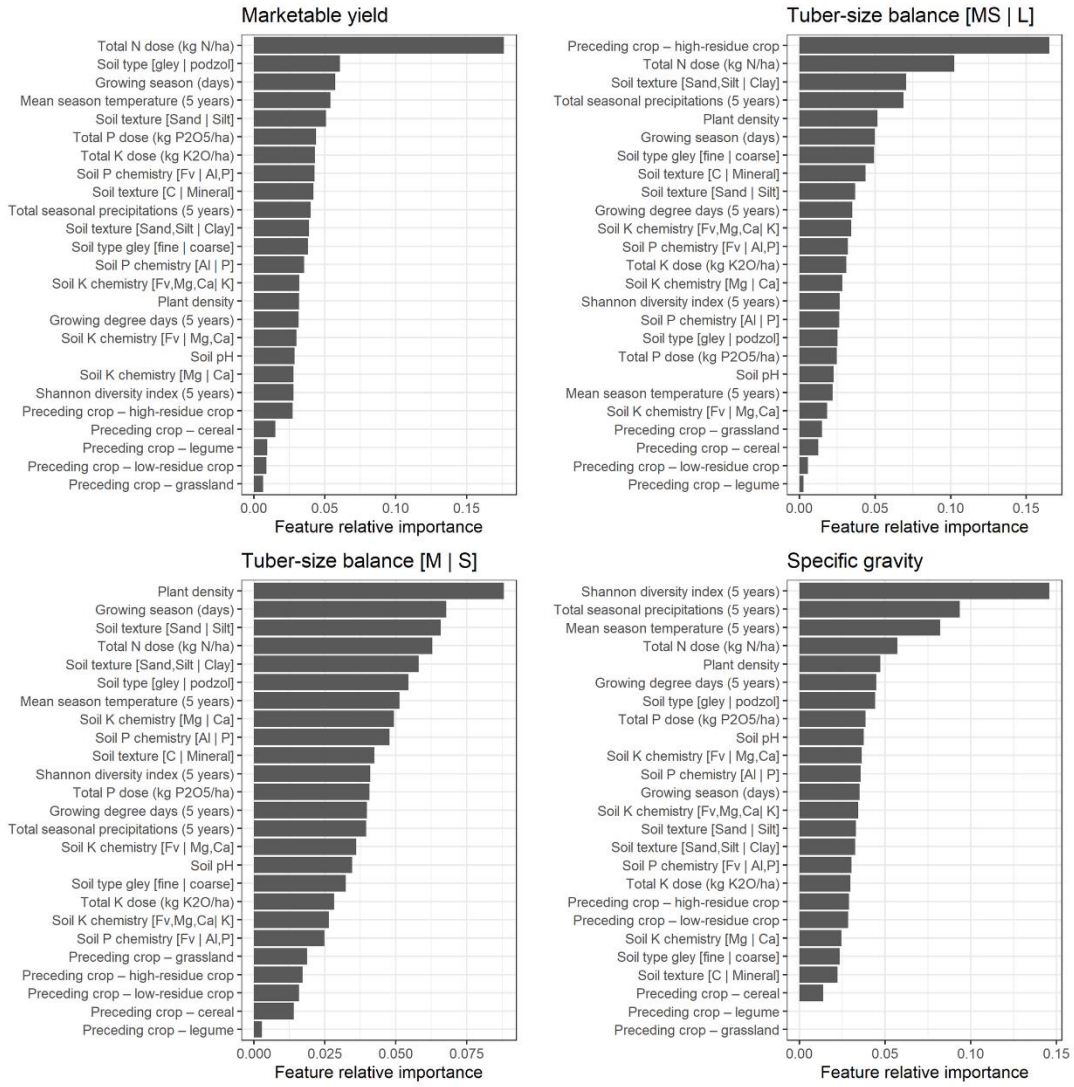


Figure 3.2. Predictive features importance for modeling

3.5.2 Model tuning parameters

The tuning parameters varied within the models depending on target variables (Table 3.5). The parameters were tuned during modeling using python random search method with 5-fold cross-validation. For each target variable the corresponding training set was used.

The basic assumption in the KNN algorithm is that similar samples should return similar output (class or value) (Mucherino et al., 2009). The two parameters to tune are the distance function which determines the similarity, and the optimal number of neighbors (similar known observations, *k*) to use for assigning the unknown output. The regressions were run with 19 nearest neighbors ($k = 19$) for yield, tuber size [M, S

| L] balance and SG prediction models. For the [S | M] balance prediction model, k was set at 18 neighbors. With uniform weights, all the points in each neighborhood are weighted equally while with an inverse distance weight, closer neighbors have a greater influence than neighbors which are further away.

Table 3.5. Tuned model parameters

A. Data used to tune parameters

Target variable	**Yield**	**[M, S | L]**	**[S | M]**	**SG**
Number of samples	5913	4557	4557	3180*
Training set	4139	3203	3203	2242
Testing set	1774	1354	1354	938

B. k-nearest neighbors

Target variable	**Yield**	**[M, S | L]**	**[S | M]**	**SG**
k	19	19	17	19
Distance	Euclidean	Euclidean	Euclidean	Euclidean
Weight	Inverse distance	Uniform	Inverse distance	Uniform

C. Random forest

Target variable	**Yield**	**[M, S | L]**	**[S | M]**	**SG**
Number of trees	92	12	17	19
Number of features	'auto'	'auto'	'auto'	'auto'

D. Neural networks

Target variable	**Yield**	**[M, S | L]**	**[S | M]**	**SG**
Input layer size	20	20	20	20
Hidden layers size	100	200	100	200

E. Gaussian process

Target variable	Yield	[M, S | L]	[S | M]	SG
Kernel	Matern	Matern	Matern	Matern
Noise level (alpha)	0.195	0.136	0.031	0.932

* The total number of samples (3180) differs from that of this target variable in Table 3.1 because 1074 outliers have been excluded during the process.

The parameters of a RF include mainly the number of decision trees in the forest and the number of features considered by each tree when splitting a node. The optimization procedure set the number of trees in the forests to 92, 12, 17 and 19 for yield, tuber size [M, S | L] balance, tuber size [S | M] balance and SG prediction models, respectively. The number of features considered for splitting at each leaf node were selected automatically.

A NN is characterized by its architecture, the training algorithm and the activation function. We used a multilayer perceptron in which neurons are organized in layers: an input layer where data are fed into the system, one or more hidden layers where the learning takes place, and an output layer where the decision/prediction is given (Liakos et al., 2018). We tuned the number of neurons for one hidden layer, and the activation function. A hyperbolic tangent activation function was selected for all the target variables prediction models. The tuned numbers of the hidden layer neurons were 100, 200, 100 and 200 for yield, tuber size [M, S | L] and [S | M] balances, and tuber SG respectively.

GPs are defined by a mean function m(x), a kernel or covariance function generating the covariance matrix $k(x_i, x_j)$ between pairs of random outputs. A white noise (σ^2) can optionally be added to the kernel (Rasmussen and Williams, 2006). The Matern kernel without white noise returned the lowest error for each target variables. Different noise levels were found to be optimal: 0.195 for marketable yield prediction model, 0.136 for tuber size [M, S | L] balance, 0.031 for [S | M] balance, and 0.932 for tuber SG. Because all the target variables were scaled and centered, mean functions m(x) were null.

3.5.3 Comparison between models

Model performance to predict marketable yield, tuber-size balances and tuber SG was assessed using R^2, MAE, RMSE, response curves shapes and economic optimal N-P-K dosage predictions for each model. For all the models, the predictive accuracy level was not affected after discarding the preceding crop classes.

3.5.3.1 Goodness of fit

The model scores at training and testing for the different target variables are presented in Figure 3.3. There was a large gap between training and testing scores. The difference was lower for the Mitscherlich model, which also showed the lowest coefficient of determination and the highest MAE and RMSE. Its R^2 values were 0.35 and 0.37 at training and testing, respectively. The R^2 values of machine learning algorithm-based models ranged between 0.78 (NN) and 0.92 (KNN) at training, and between 0.49 (NN) and 0.59 (RF) at testing in predicting marketable yield. With the large-size tuber yield balance [M, S | L], the R^2 values ranged between 0.72 (KNN) and 0.87 (RF) at training, and between 0.55 (KNN) and 0.64 (GP) at testing. The medium- versus small-size tuber [S | M] balance and SG prediction models were the most informative, as shown by the highest R^2 values at both training and testing. The R^2 values ranged between 0.83 (NN) and 0.93 (KNN) at training and between 0.62 (RF) and 0.69 (KNN) at testing in predicting small-size tuber balance, while for SG, they ranged between 0.72 (KNN) and 0.94 (RF), then between 0.58 (KNN) and 0.67 (RF) at training and testing, respectively. In general, model MAE and RMSE were slightly higher when R^2 values were low. The practically-similar magnitudes between RMSE and MAE meant that all the individual differences between predictions and observations had equal weight.

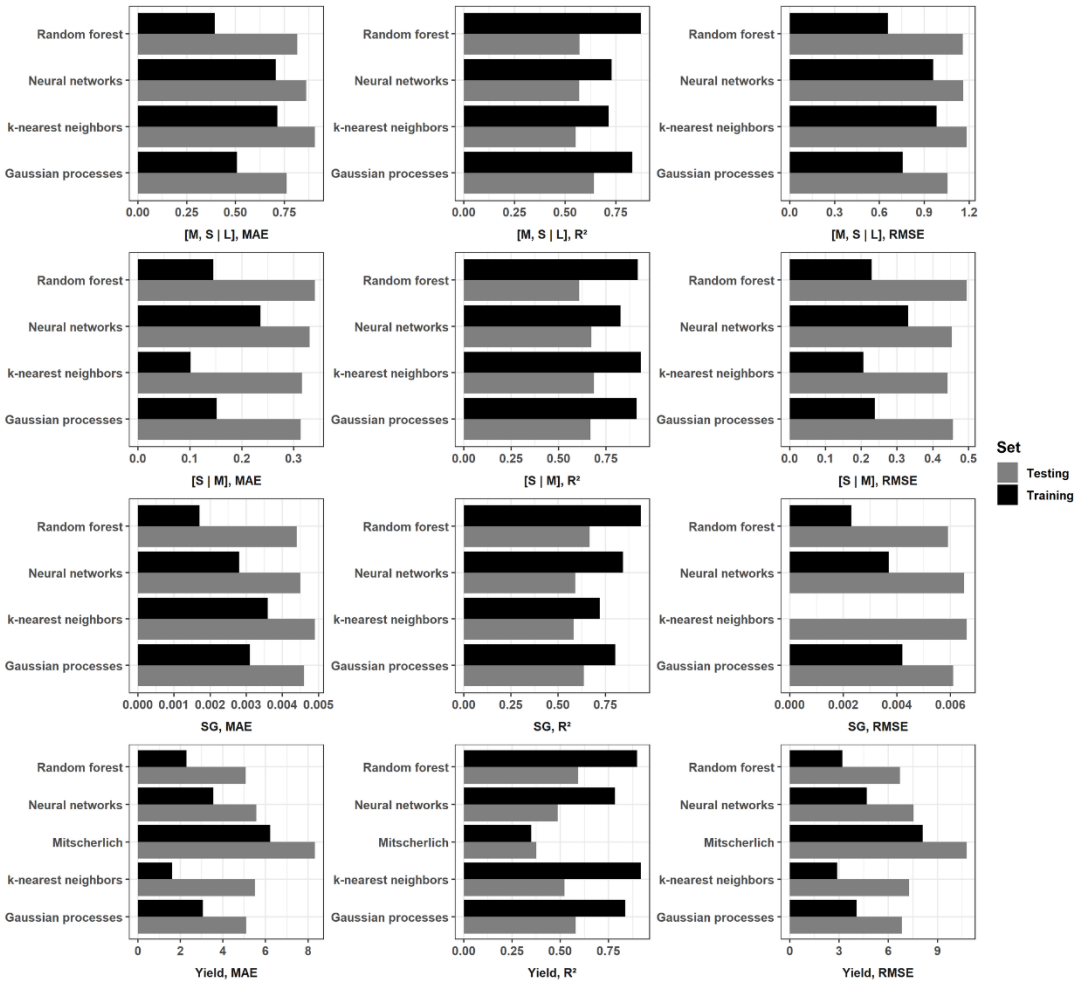


Figure 3.3. Comparison of models goodness of fit using R^2, MAE and RMSE

3.5.3.2 Response curves

The marketable yield response curves are plotted in Figure 3.4 for each model with respect to the tested nutrient. There were disagreements between models. The Mitscherlich, NN and GP models generated smooth response curves, while the KNN and RF models generated stepped curves. The marketable yield was non-responsive to P application in the RF model. There was also no effect of K fertilization on the yield shown by the Mitscherlich and RF models. All models for the P trial somewhat underestimated marketable yield while response curves followed data for N.

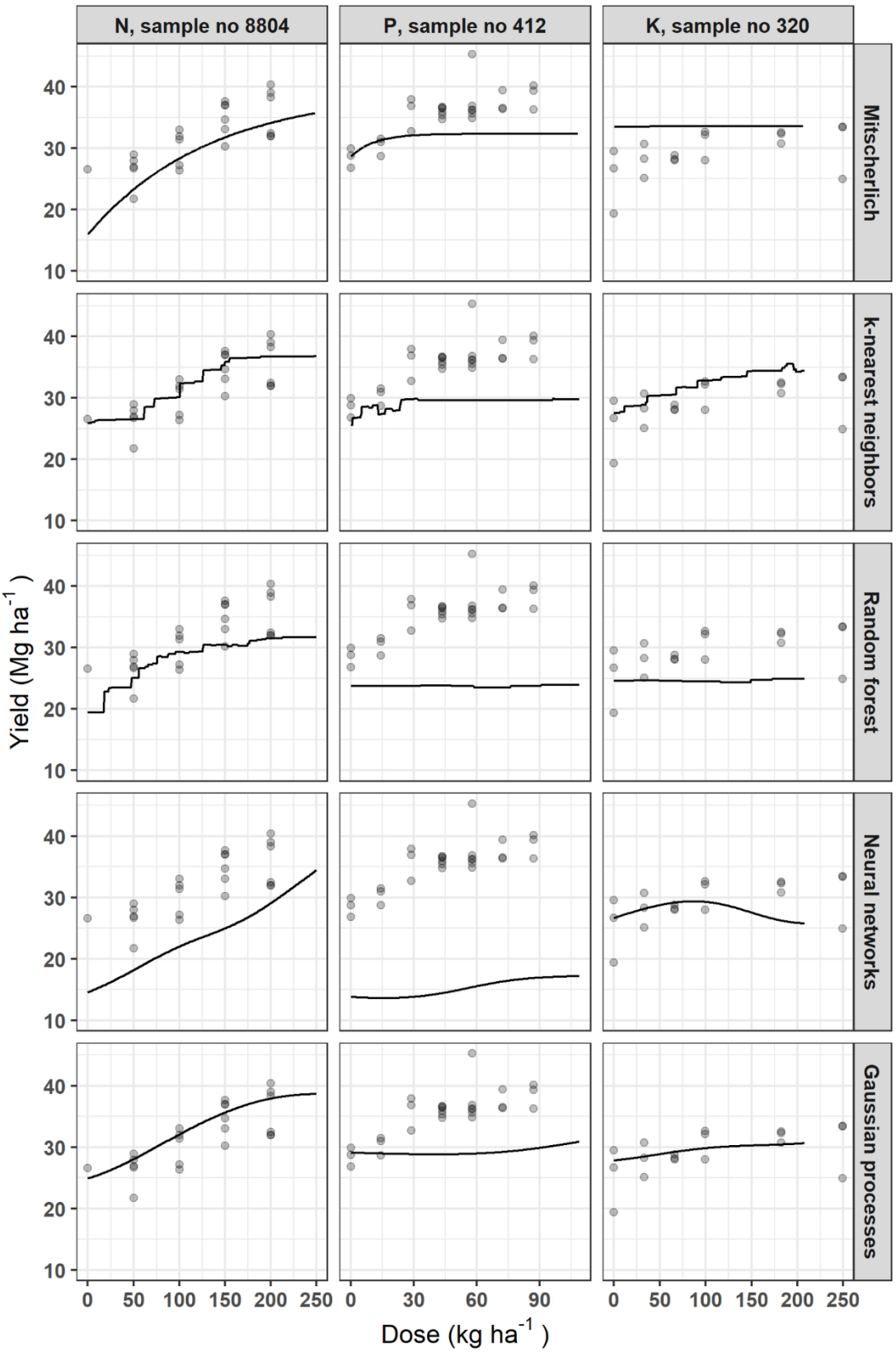


Figure 3.4. Examples of potato yield response to N, P or K fertilization using different models

The Mitscherlich model was excluded for the analysis of other target variables. Figures 3.5 – 3.7 show how each model fits responses of tuber size balances ([M, S | L] and [S | M]), and SG, respectively, with respect to N, P or K dosage. The NN and GP models generated smooth curves, while the KNN and RF models generated stepped curves. The [M, S | L] balance (Figure 3.5) showed increasing response to N fertilization across models, while response was globally poor for P and K. For the [S | M] balance, responses increased with increasing fertilizer doses, except for P and K trials data fitted with GP model (Figure 3.6). There was also poor response for K trial with SG (Figure 3.7). The SG response decreased from zero K levels and increased then decreased as P dosage increased. For N trials, SG slightly increased then decreased as N dose increased in the RF model but was non-responsive with the other models.

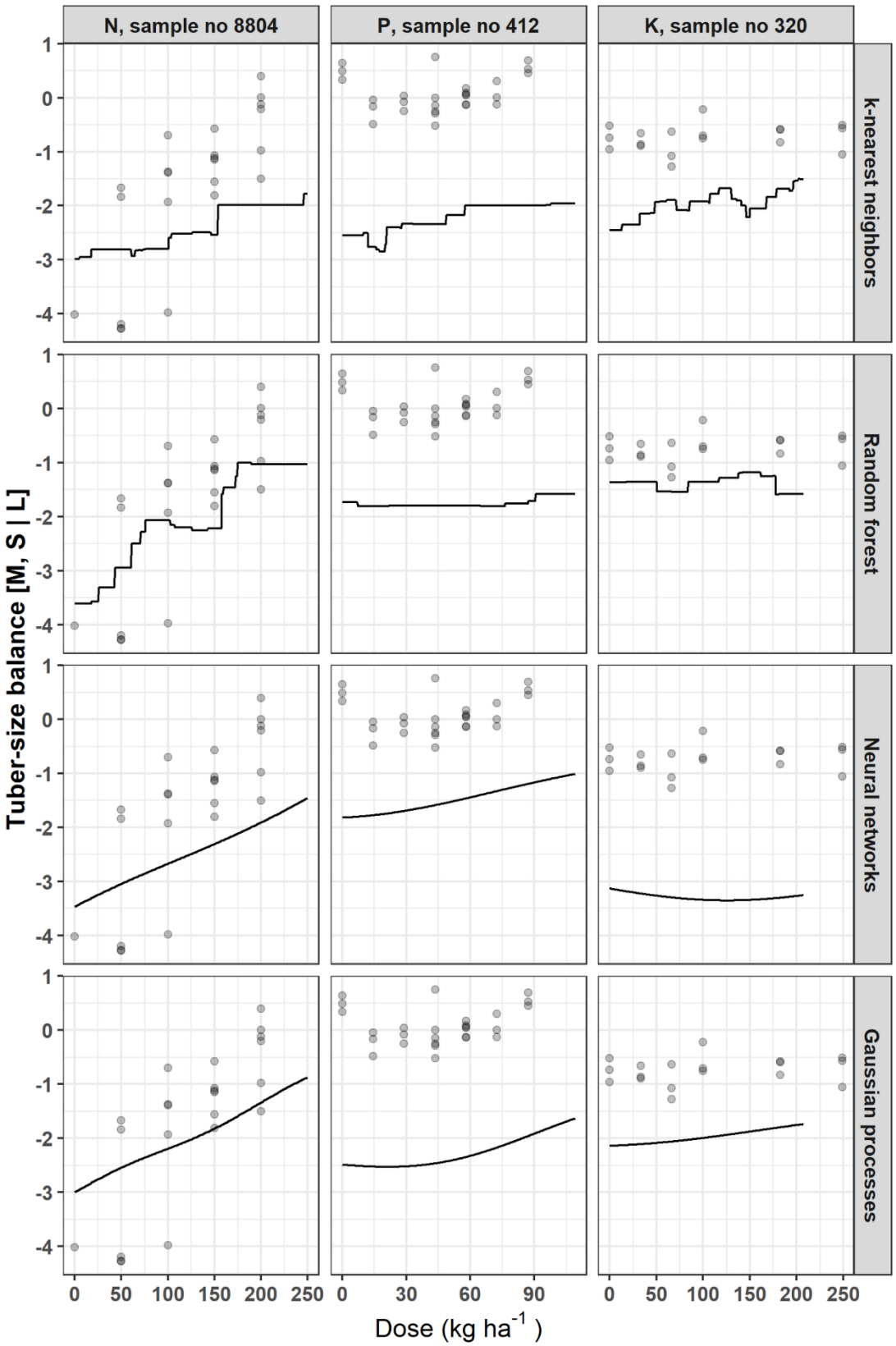


Figure 3.5. Examples of potato tuber size [M, S | L] balance response to N, P or K fertilization using different models

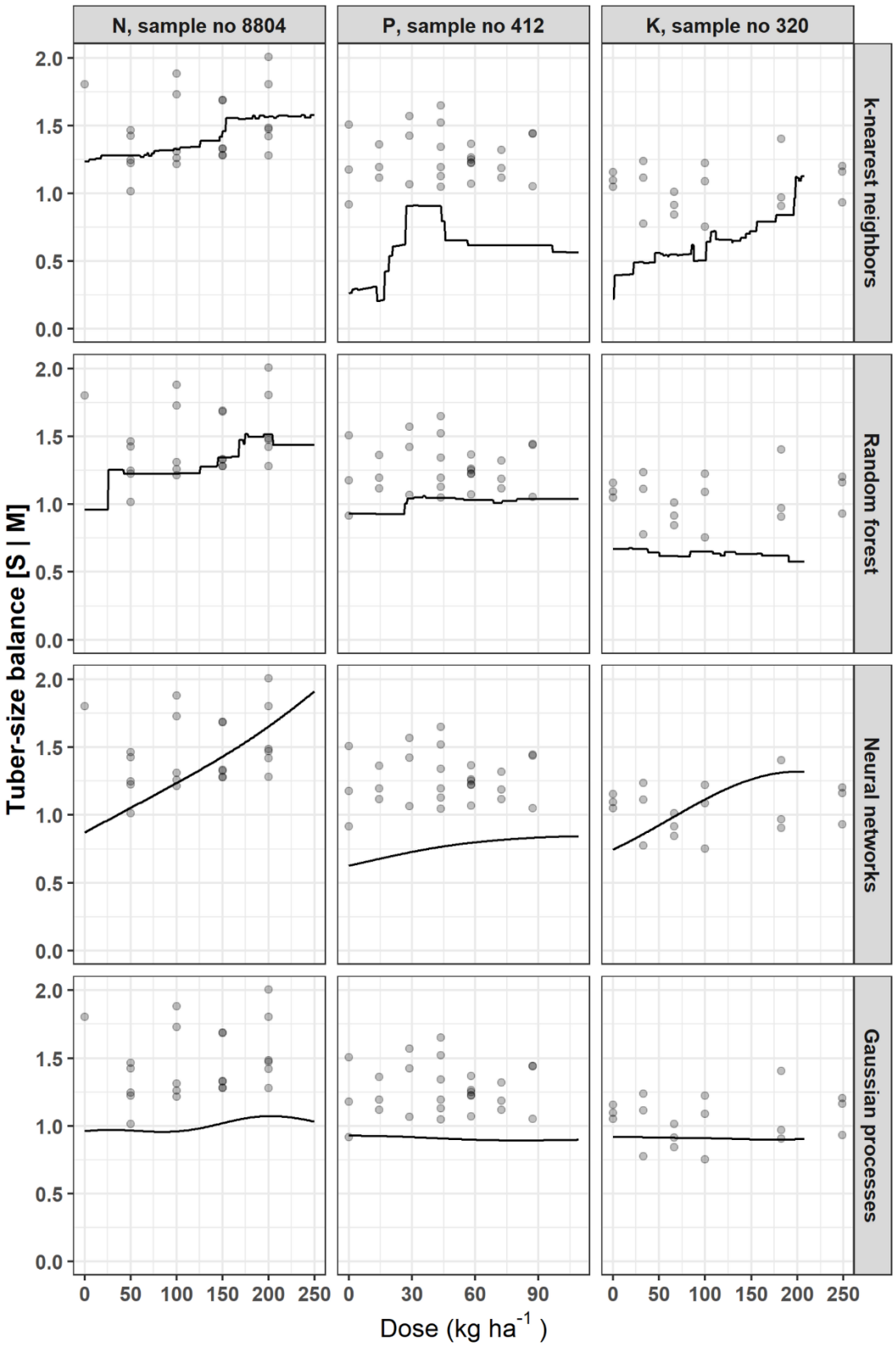


Figure 3.6. Examples of potato tuber size [S | M] balance response to N, P or K fertilization using different models

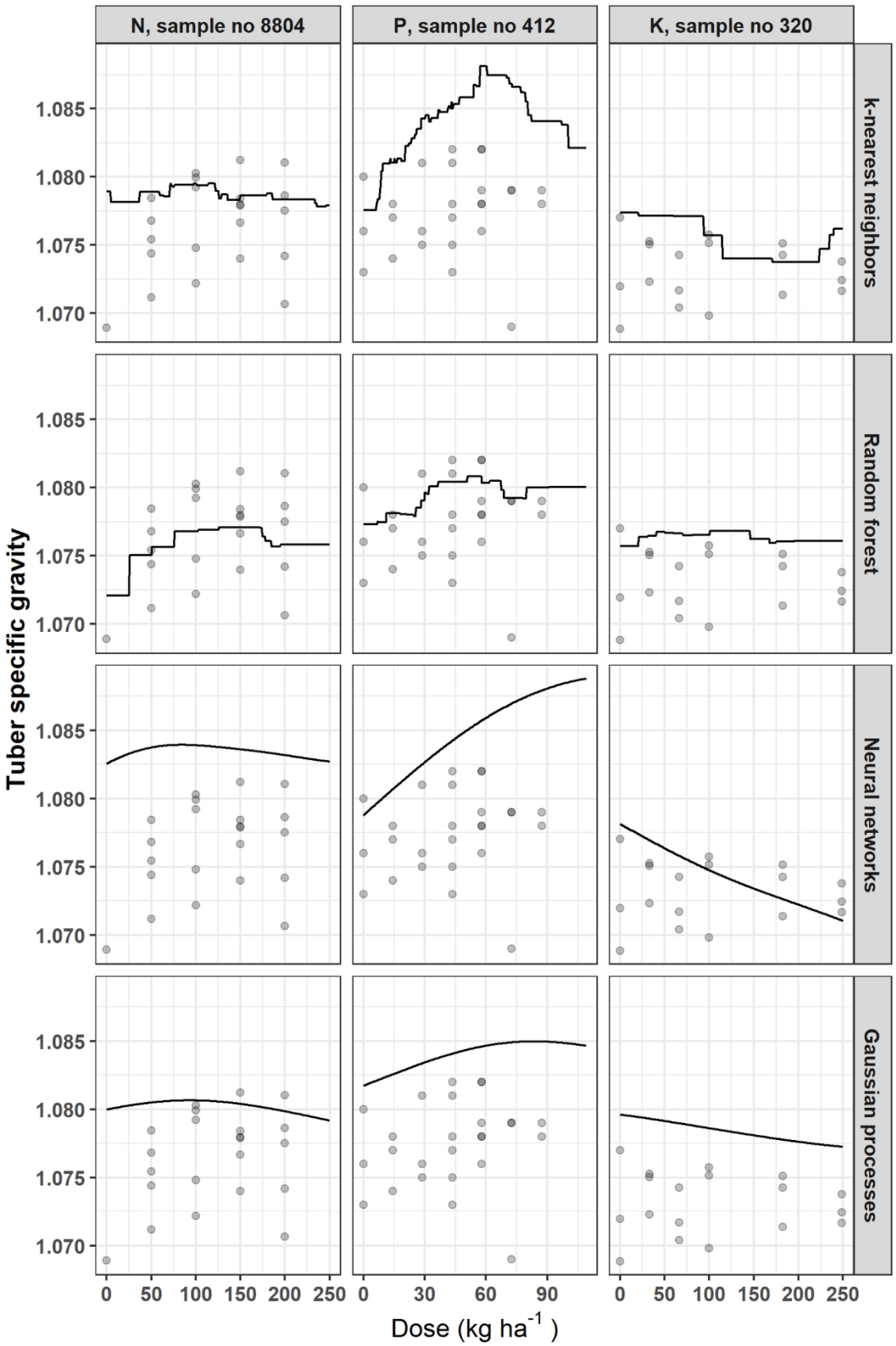


Figure 3.7. Examples of potato tuber SG response to N, P or K fertilization using different models

3.5.3.3 Predictions

The fertilizer recommendations and output predictions varied with the model and the target (Figure 3.8). The Mitscherlich and NN models predicted negligible economic optimal K doses (11 and 12 kg ha^{-1} respectively) in marketable yield prediction models, while the site Mehlich-3 K level was classified as very low (83.1 mg kg^{-1}) according to local standards (CRAAQ, 2010). The RF model suggested the highest cumulative agronomic optimum fertilizer doses, although its outputs were not the highest. With the tuber size [M, S | L] balance prediction model, practicable doses were recommended only by the GP model for P (107 kg ha^{-1}) and the RF model for K (185 kg ha^{-1}), a scheme that is almost similar to the [S | M] balance prediction models. For this output, the GP model recommended only 17 kg P ha^{-1}, while N and K were impracticable (1 kg ha^{-1} and 4 kg ha^{-1}, respectively). Despite the extremely low environmental risk for P and the low level of soil K, some models predicted negligible doses of P and K mainly for tuber size balances.

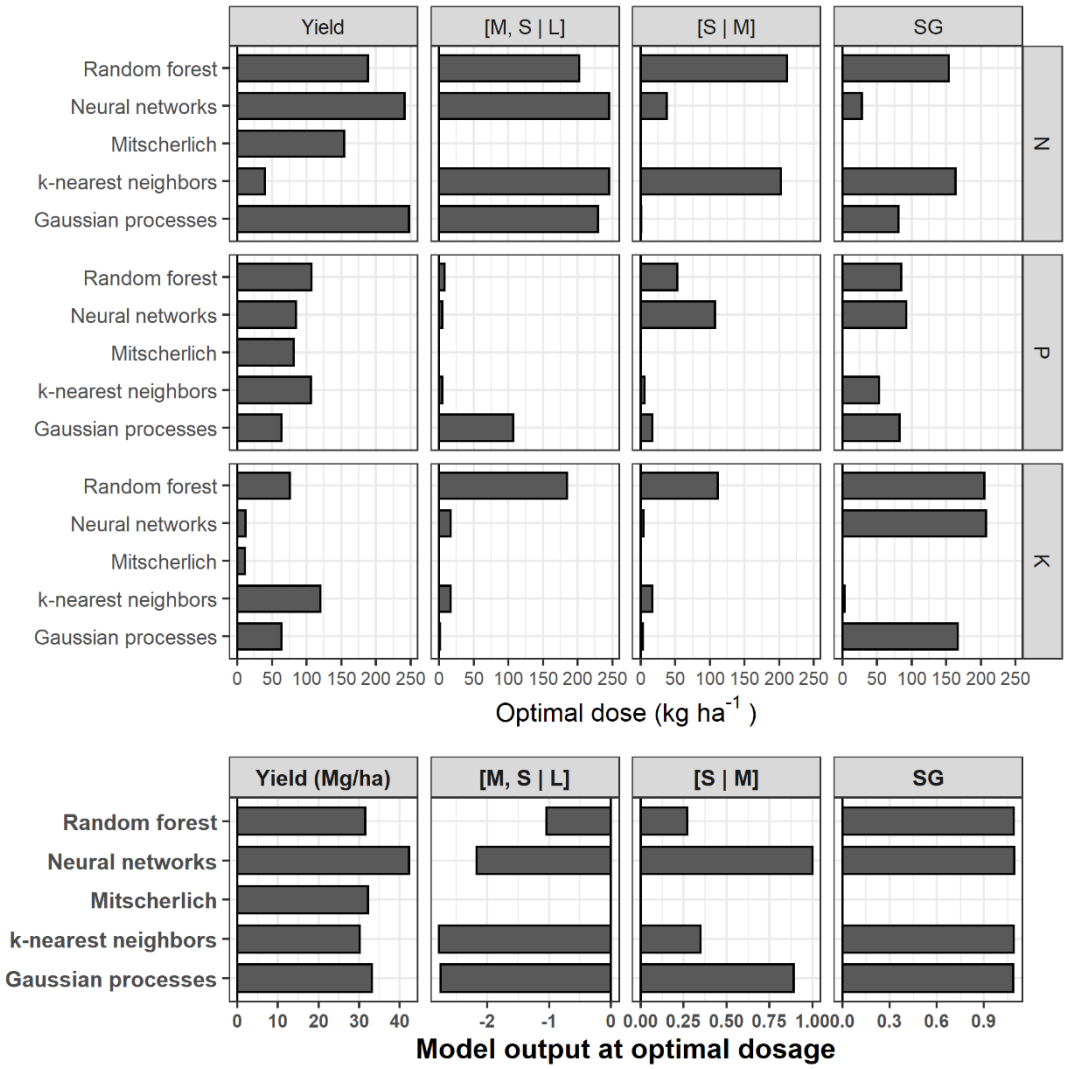


Figure 3.8. Economic or agronomic optimal doses and output predictions at optimal dosages for each model with a random selected test trial (N° 194)

3.5.4 Probabilistic predictions

In addition to point estimates shown by each model, the GP model can return posterior samples. Each sample is a function from which we can compute an economic optimal (marketable yield) or agronomic optimal (size balances or SG) fertilizer dose. Figures 3.9 – 3.12 present the results of 1000 generated samples for each target variable for the selected N, P and K trials. The average GP curve is shown as a black line, with its optimal dosage as a black dot. Five sampled GP curves are plotted as grey lines, with their optimal doses as grey dots. The probability distributions of the 1000 optimal doses are shown under the respective response curves. The figures show that predicted means

of optimal dosage (black dot) did not always correspond to the most likely dosage (highest histogram bar) computed after running the sampling process. With yield prediction models (Figure 3.9), the mean economic optimal dose corresponded to the probabilistic prediction only for the N trial (250 kg N ha^{-1}). For the tuber size [M, S | L] balance (Figure 3.10), the probabilistic prediction was equal to the mean GP prediction for P trial i.e., 87 kg P ha^{-1}, while N and K trials returned equal predictions with the [S | M] balance prediction models with 0.0 kg ha^{-1} and 0.70 kg ha^{-1}, respectively (Figure 3.11). For tuber SG prediction models, none of the probabilistic recommendation matched the mean GP optimal dosage (Figure 3.12).

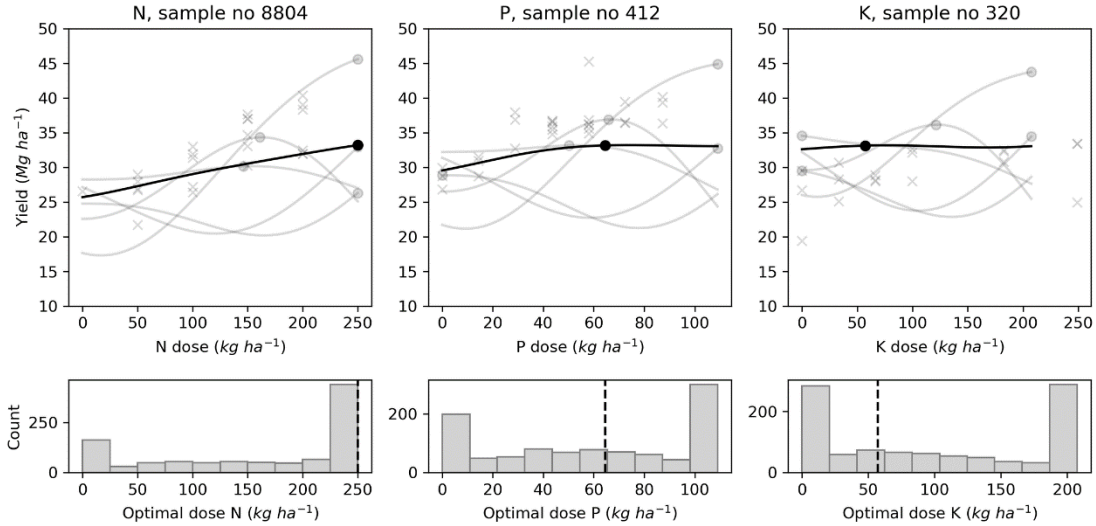


Figure 3.9. Examples of optimal economic N, P, K doses distribution with Gaussian processes using marketable yield for selected trials

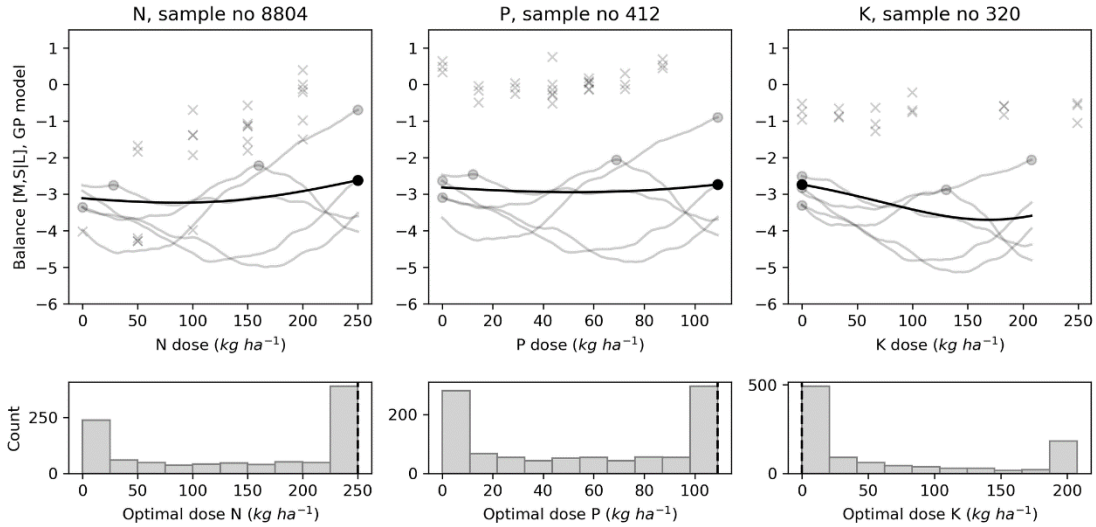


Figure 3.10. Examples of agronomic optimal N, P, K doses distribution with Gaussian processes using tuber size [M, S | L] balance for selected trials

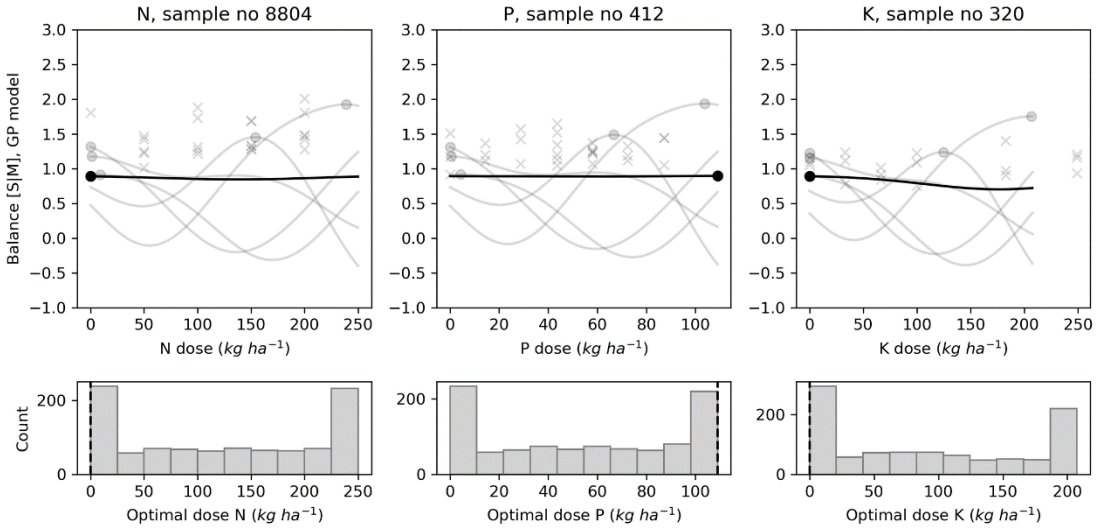


Figure 3.11. Examples of agronomic optimal N, P, K doses distribution with Gaussian processes using tuber size [S | M] balance for selected trials

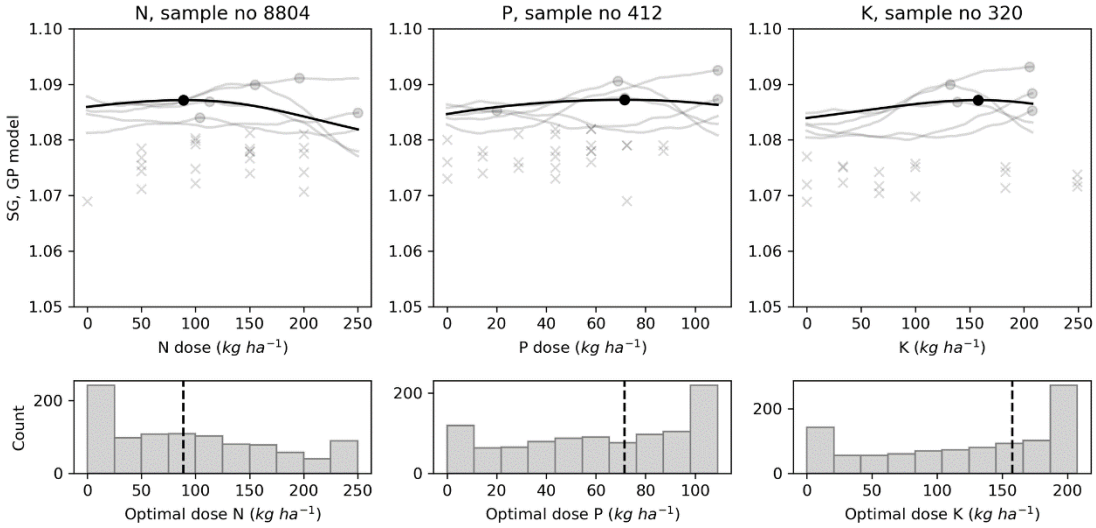


Figure 3.12. Examples of agronomic optimal N, P, K doses distribution with Gaussian processes using tuber SG for selected trials

3.6 Discussion

3.6.1 Selection of features

Fertilization trials were conducted over a time span of four decades (1979 – 2017). Although agricultural practices, soil conditions and analytical techniques have undergone substantial changes over time, Valkama et al. (2009) has shown that the differences between old and recent experiments in yield responses are not statistically important. Moreover, where the analytical techniques for the same element differed, correlation equations were available to converting to one technique before data analysis. It is the case for soil carbon converted from Walkley-Black to Leco CNS (equation 3.1), soil pH processed with $CaCl_2$ converted to pH water (equation 3.3), and P-Bray-2 converted to P-Mehlich-3 (equation 3.4). Since there were similarities in experimental procedures and ability to uniformly convert measurement methods, we found that the data set could be used for machine learning.

The feature selection function selects a subset of variables for a learning algorithm to focus attention on the subset, especially when dealing with a large number of explanatory variables. The model-based approach incorporates the correlation structure between predictors and provides scores that indicate how useful or valuable each feature is in model building. Features with low or no importance could be removed without affecting model performance (Pedregosa et al., 2011). The preceding crops categories *i.e.*, grassland, small grains, legumes, low-residue crops and high-residue crops, as categorized by Parent et al. (2017), returned zero (for tuber SG) or faintest scores (for other target variables) and were thus removed despite a substantial body of literature on the advantages of crop rotation to the next crop. Nonetheless, Zebarth et al. (2007) stated that the amount of nitrogen mineralized from organic matter during the growing season cannot be predicted accurately. Torma et al. (2018) found that the N supplied by soil and crop residues (maize, potato, silage maize, soybean, sunflower, winter rape, winter wheat) ranged from 20 to 132 kg ha^{-1}, while the phosphorus ranged from 2 to 24 kg ha^{-1} and potassium from 13 to 218 kg ha^{-1}. Rangarajan (2009) stated that nutrient availability to the next crop depends on whether the entire plant or only the root system is left in the field, and on how environmental conditions govern the rate of organic matter decomposition.

For marketable yield and tuber size balances prediction models, the N dose was the most informative feature, probably because of its close relation to photosynthesis (Andrews et al., 2013). Applied in excess, it delays tuber maturity, stimulates foliage production, increases plant susceptibility to diseases and reduces tuber SG (Hawkesford et al., 2012). Crop yield is also determined by environmental conditions driving the physical, chemical and biological reactions (Feddes, 1971) that are important in empirical or mechanistic models (Griffin et al., 1993; Stalham et al., 2005; Boiteau et al., 2014; Raymundo et al., 2014; Parent et al., 2017).

The selection process retained soil profile characteristics and weather events as major features. Levy and Veilleux (2007) reported the effects of air and soil temperatures on potato growth mechanisms and tuber yield. Leblanc et al. (2016) pointed out soil drainage conditions for loamy-gleyed profiles (poorly-drained loam), sandy-gleyed profiles (poorly-drained sand) and sandy-podzolized profiles (well-drained). Soil compaction has a negative impact on root extension and water movement *i.e.*, the reduction of nutrient uptake potential leading to a severe reduction of tuber yield (Stalham et al., 2005). Xu et al. (2017) developed pedotransfer functions for potato grown on light-textured soils that could be useful in future models.

Dry matter production of potato crops is determined by the length of the growth cycle (Struik, 2007), which turned out to be a valuable feature. Camire et al. (2009) stated that long growing season favors high-yielding late-season cultivars. Rex (1991) found a close relationship between delayed harvest date and total yield, main-size marketable tubers and SG.

Seeding density was the most informative feature of the medium- to small-size tubers balance. Seeding density differentiates the number of tubers harvested, the weight of the tubers and the size distribution; higher plant densities promote higher yields in small and medium sizes (Rex, 1991; Ellissèche, 1996; Bussan et al., 2007).

The feature selection algorithm showed the impact of weather indices on tuber SG. The Shannon diversity index, total rainfall and temperature yielded the highest scores in a decreasing order. Al Soboh et al. (2002) reviewed the factors affecting SG loss in crops of crisping potato and stressed that irrigation during early growth stages increases tuber dry matter content. Specific gravity could be reduced substantially if

heavy rain occurred at the end of the season before harvest. They stated that potatoes grown during a period of increasing day length, temperature and light intensity produce tubers of high SG. In this study, GDD considered only daily mean temperatures higher than or equal to 5 °C as used by Parent et al. (2017). Moulin et al. (2012) used a baseline of 7 °C and 30 °C as upper limit. Moreover, the general trend of SG response curves with respect to fertilization supported the results of Belanger et al. (2002), Zebarth et al. (2004) and Laboski and Kelling (2007). Excessive application doses of N and K along with high soil levels of either nutrient may reduce SG. Phosphorous application may increase tuber solids when soil test P levels are low. Specific gravity was not influenced by the relatively high levels of N and P used by Dubetz and Bole (1975), while Maier et al. (1994) found contrasted effects between trials.

The relative importance of a variable in a model is related to its effect on the output through its gradient in the data set. Hence the predominance of N doses, and P and K doses to some extent, could have been caused by the origin of the data set, which is a collection of fertilizer trials, where large gradients of doses are found by design. This study did not address fertilizer source and timing of application. While Marouani et al. (2015) found equivalency of ammonium nitrate (33.5% N), urea (46% N), NP fertilizer (33% N – 14% P_2O_5) and NPK fertilizer (27%N – 5% P_2O_5 – 5% K_2O), Petropoulos et al. (2020) found that the form of the fertilizer (ammonium sulfate, ammonium sulfate + zeolite, manure, slow release N fertilizer with urease inhibitor) and the cultivar (Kennebec and Spunta) may affect yield and chemical composition of potato tubers, affecting the end use of the product. Flis (2019) reported that the peculiarities of potato cultivar, plant root structure, and timing of nutrient uptake impact on the selection of a site-specific fertilization regime. Trehan et al. (2001) showed that some cultivars exhibit strong symptoms of N, P and K deficiencies compared to others. Potato cultivars may sustain leaf development and nutrient uptake while maintaining maximum tuber growth rates to reach higher final tuber yields with contrasting nutrient requirements (Kleinkopf et al., 1981). Differential effects of cultivar and fertilizer on tuber yield have also been reported by Daoui et al. (2014). In a previous study, Coulibali et al. (2020) found that genetic traits were not compelling to set apart clusters of cultivar based on N, P, K, Mg and Ca compositions of diagnostic leaves. The cultivar effect was thus excluded from the present study to keep models parsimonious. In our analysis, we focused on the gradients of N, P and K doses while keeping the other site-specific

factors constant. Nevertheless, predictive features such as biotic factors (length of growing season, preceding crop, and seeding density), could also be predicted and optimized by the models with respect to tuber yield and quality.

3.6.2 Comparison of models

The performance of a predictive model is evaluated at testing or with unseen data set. The goodness of fit refers to how closely the model-predicted values match the true or observed values. Overfitting occurs where models perform well at training and badly at testing, while underfitting characterizes a model performing badly in both training and testing. Except for the Mitscherlich model, the model scores at testing showed discrepancies with training, reflecting problems of overfitting. The differences between R^2 values were highest for the marketable yield prediction models (Figure 3.3), reaching 0.40 with KNN. Based on those gaps, one could argue that our models did not generalize well from training to testing data. However, we used a robust approach by comparing different algorithms, tuning the hyperparameters and tuning the models using 5-fold cross-validation. The R^2 values at testing varied with respect to target variables but were practically similar between models. The models estimated the proportions of medium- and small-size tubers ([S | M] balance) more accurately than those of large-size tubers ([M, S | L] balance), probably because of the high number of zero weight values among large-size tubers (21%) compared to tubers of small (0.06%) and medium (0.4%) size, at the early stage of our analysis. Imputing zeros to deal with measures where the large size was completely absent (Martin-Fernandez et al., 2003) improved the prediction quality of this fraction. Except for the Mitscherlich model in predicting yield, the R^2 values at testing were greater than 0.50 and could be considered acceptable according to Moriasi et al. (2007) for complex systems.

The Mitscherlich model returned a lower coefficient of determination in tuber yield prediction and was discarded for quality analysis (tuber size balances and SG). The KNN, RF, NN and GP algorithms more accurately approximated the unknown functions explaining tuber yield given the predictive features. However, it was difficult to select the best model since scores were practically similar. Cerrato and Blackmer (1990) and several others (Bock and Sikora, 1990; Angus et al., 1993; Bullock and Bullock, 1994; Isfan et al., 1995; Belanger et al., 2000) described similar ambiguities using classical statistical models.

Figures 3.4 –3.7 indicated that the calibration and generalization procedures returned smooth response curves for the Mitscherlich, NN and GP models for all the target variables. Except for the low R^2 value of the former, the NN and GP models appeared more suitable for making inferences.

The prediction of optimum fertilizer doses and optimum or maximum outputs showed some disagreements for the case presented (Figure 3.8). There should be a single economic optimal dose or agronomic optimal dose at each site each year. Some models were more consistent than others in deriving optimal doses depending on the target variable. At extremely low predicted N, P or K doses, it could be challenging to manage the fertilization program at low economic risk for producers, who generally consider that the cost of over-fertilization is low compared to the cost of under-fertilization (Rajsic and Weersink, 2008; Parent, 2014). The probabilistic prediction capability of Gaussian processes may help to determine credible dosage.

3.6.3 Probabilistic predictions

Sampling from a Gaussian process looks like rolling a die, returning a different function each time. Figures 3.9 – 3.12 showed only five possible functions for each target variable. By sampling the process numerous times, we generated a distribution of economic or agronomic optimal fertilizer doses as those shown by the histograms of the figures. The distributions often show frequent optima at the edges to the NPK grid, *i.e.*, at dose of 0 or 250 kg ha^{-1}. This phenomenon emerges from sampling continuously increasing or decreasing GP samples, which are more frequent when the sample is close to patterns in data where the response to fertilizer is flat. A zero-fertilizer recommendation could be interpreted as a soil sufficiently fertile to supply the crop, or a soil poorly responsive due to other constraints (Vanlauwe et al., 2011) such as pests and diseases (Rich, 1983; Raman and Radcliffe, 1992) or weed damage (Mondani et al., 2011). Nevertheless, we covered a wide range of factors that may impact potato crop growth and yield without falling into mechanistic modeling. Fertilizer doses more than 250 kg ha^{-1} may be excessive, since the maximum limits according to local standards are 175 kg ha^{-1} for N, 87 kg ha^{-1} for P and 199 kg ha^{-1} for K (CRAAQ, 2010).

To face predictions falling at the edges, the optimal fertilizer dosage could be selected within a range of conditional expectation as processed by Khiari et al. (2000)

when defining P optimal dose for acid coarse-textured soils. The x^{th} conditional expectation dose is the optimal dose that produces optimal yield *x%* of the time. For example, the 60th percentile would be the sampled optimal dose that produces optimal yield 60% of the time for a given site. Khiari et al. (2000) assessed the 50th and 80th percentiles. The mean (50%), the median or any other percentile dose could be computed to support decision-making. For example, the mean GP and the probability distribution processes returned the upper bound of the simulation dosage (*i.e.*, 250 kg N ha^{-1}) as the economic optimal dose for the N trial with the marketable yield prediction model (Figure 3.9). The conditional expectation percentiles showed that a lower dose (*i.e.*, 223 kg N ha^{-1}) could be recommended, producing optimal yield 55% of the time. At the 60th percentile or more, the full dose *i.e.*, 250 kg N ha^{-1} must be applied.

3.7 Conclusion

This study assessed machine learning techniques as an alternative for potato fertilizer recommendations at local scale usually handled by statistical models or meta-analysis at regional scale. A large collection of field trial data provided information to fit machine learning models with specific traits of cultivars, soil properties, weather indexes, and N, P and K fertilizers dosage used as predictive features. Five models, Mitscherlich, KNN, RF, NN and GP, were evaluated against optimal economic N, P and K doses derived from yield, or against optimal agronomic N, P and K doses derived from tuber size and SG. The models trained using machine learning algorithms outperformed the Mitscherlich tri-variate response predictive model. The marketable yield prediction coefficient (R^2) varied between 0.49 and 0.59, while the Mitscherlich model returned 0.37. The large-size tuber balance was predicted with a coefficient varying between 0.55 and 0.64. The R^2 varied between 0.60 and 0.69 in predicting medium-size tuber balance, and between 0.58 and 0.67 for SG. The N, P and K optimal doses could be recommended with respect to marketable yield, tuber size or SG using the NN and GP models, which appeared to be the most suitable for making inferences. Response surfaces were obtained by marginalizing the models using N-P-K doses generated from uniform distributions under constant weather conditions, soil properties and land management factors. The GP model stood up by its probabilistic framework in risk estimation for potato fertilizer recommendation in Quebec conditions.

As large amounts of data are being assembled into observational data sets, machine learning models may surrogate statistical models in making fertilizer recommendations in the context of precision agriculture. To assess model performance under real-world situations, it was an effective strategy to combine historical weather data since accurate future weather data covering the growing season are unavailable. We also focused on using easily-available features collected from routine analyses as predictors instead of mechanistic processes models. Any biotic factor other than fertilizer, e.g., length of growing season or planting density, could be optimized with our model. Improvement will require more data from many more diverse environments and management scenarios. With more experiment data, the training and testing division could be performed at trial level to improve the model predictive ability. Moreover, since the data for this analysis were collected from small research plots, validation at production-scale fields is needed for decision making.

3.8 Acknowledgements

The authors acknowledge the financial support of the Natural Sciences and Engineering Council of Canada (CRDPJ 385199-09 and DG-2254), the Quebec Ministry of Agriculture, Fisheries and Food (IA216581), Centre SEVE, Patate Dolbec Inc. (St-Ubalde, QC), Groupe Gosselin FG (St-Augustin-de-Desmaures, QC), Agriparmentier Inc. (Notre-Dame-du-Bon-Conseil, QC), Ferme Daniel Bolduc Inc. (Péribonka, QC), Patate Laurentienne (Notre-Dame-de-la-Paix, QC), Ferme Bergeron-Niquet (Péribonka, QC) and Patates Lac-St-Jean (Péribonka, QC).

3.9 References

Al Soboh, G., Sully, R., and Andreata, S. (2002). *Factors affecting specific gravity loss in crisping potato crops in Koo Wee Rup, Victoria*. Retrieved from https://trove.nla.gov.au/version/20463664

Andrews, M., Raven, J. A., and Lea, P. J. (2013). Do plants need nitrate? The mechanisms by which nitrogen form affects plants. *Annals of Applied Biology, 163*(2), 174-199. doi:10.1111/aab.12045

Angus, J. F., Bowden, J. W., and Keating, B. A. (1993). Modeling nutrient responses in the field. *Plant and Soil, 155*, 57-66. doi:10.1007/bf00024984

Barber, S. A. (1995). *Soil nutrient bioavailability: a mechanistic approach*: John Wiley & Sons.

Belanger, G., Walsh, J. R., Richards, J. E., Milburn, P. H., and Ziadi, N. (2000). Comparison of three statistical models describing potato yield response to nitrogen fertilizer. *Agronomy Journal, 92*(5), 902-908.

Belanger, G., Walsh, J. R., Richards, J. E., Milburn, P. H., and Ziadi, N. (2002). Nitrogen fertilization and irrigation affects tuber characteristics of two potato cultivars. *American Journal of Potato Research, 79*(4), 269-279. doi:10.1007/bf02986360

Bock, B. R., and Sikora, F. J. (1990). Modified-quadratic/plateau model for describing plant-responses to fertilizer. *Soil Science Society of America Journal, 54*(6), 1784-1789.

Bohl, W. H., and Johnson, S. B. (2010). *Commercial potato production in North America* (W. H. Bohl and S. B. Johnson Eds. 2nd ed.). Ann Arbor, USA: The Potato Association of America Handbook.

Boiteau, G., Goyer, C., Rees, H. W., and Zebarth, B. J. (2014). Differentiation of potato ecosystems on the basis of relationships among physical, chemical and biological soil parameters. *Canadian Journal of Soil Science, 94*(4), 463-476. doi:10.4141/cjss2013-095

Bolinder, M. A., Katterer, T., Poeplau, C., Borjesson, G., and Parent, L. E. (2015). Net primary productivity and below-ground crop residue inputs for root crops: Potato (*Solanum tuberosum* L.) and sugar beet (*Beta vulgaris* L.). *Canadian Journal of Soil Science, 95*(2), 87-93. doi:10.4141/cjss-2014-091

Bullock, D. G., and Bullock, D. S. (1994). Quadratic and quadratic-plus-plateau models for predicting optimal nitrogen rate of corn: A comparison. *Agronomy Journal, 86*(1), 191-195.

Bussan, A. J., Mitchell, P. D., Copas, M. E., and Drilias, M. J. (2007). Evaluation of the effect of density on potato yield and tuber size distribution. *Crop Science, 47*(6), 2462-2472.

Cambouris, A. N., St Luce, M., Zebarth, B. J., Ziadi, N., Grant, C. A., and Perron, I. (2016). Potato response to nitrogen sources and rates in an irrigated sandy soil. *Agronomy Journal, 108*(1), 391-401. doi:10.2134/agronj2015.0351

Camire, M. E., Kubow, S., and Donnelly, D. J. (2009). Potatoes and human health. *Critical Reviews in Food Science and Nutrition, 49*(10), 823-840. doi:10.1080/10408390903041996

Cerrato, M. E., and Blackmer, A. M. (1990). Comparison of models for describing corn yield response to nitrogen-fertilizer. *Agronomy Journal, 82*(1), 138-143.

Cescas, M. P. (1978). Table interprétative de la mesure du pH des sols du Québec par quatre méthodes différentes. *Naturaliste canadien, 105*, 259-263.

CFIA. (2015). Potato plants characteristics, maturity. Canadian Food Inspection Agency. Retrieved from http://www.inspection.gc.ca/plants/potatoes/characteristics/eng/1326490397702/1326490477981#mature

Chantre, G. R., Blanco, A. M., Lodovichi, M. V., Bandoni, A. J., Sabbatini, M. R., Lopez, R. L., Vigna, M. R., and Gigon, R. (2012). Modeling Avena fatua seedling emergence dynamics: An artificial neural network approach. *Computers and Electronics in Agriculture, 88*, 95-102. doi:10.1016/j.compag.2012.07.005

Coulibali, Z., Cambouris, A. N., and Parent, S. E. (2020). Cultivar-specific nutritional status of potato (*Solanum tuberosum* L.) crops. *Plos One, 15*(3), 1-15. doi:10.1371/journal.pone.0230458

CRAAQ. (2010). *Guide de référence en fertilisation* (2ème ed.): Centre de Référence en Agriculture et Agroalimentaire du Québec.

Crisci, C., Ghattas, B., and Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling, 240*, 113-122.

Dahnke, W. C., and Olson, R. A. (1990). Soil test correlation, calibration, and recommendation In R. L. Westerman (Ed.), *Soil Testing and Plant Analysis* (3rd ed., pp. 45-71). Madison, WI: Soil Science Society of America.

Dampney, P., Wale, S., and Sinclair, A. (2011). Potash requirements of potatoes. Review. Project R443, Report 2011/4, Potato Council, Agric. Hortic. Dev. Board, Kenilworth, Warwickshire, UK.

Daoui, K., Mrabet, R., Benbouaza, A., and Achbani, E. H. (2014). Responsiveness of different potato (*Solanum tuberosum*) varieties to phosphorus fertilizer. *Procedia Engineering, 83*, 344-347.

Dessureault-Rompre, J., Zebarth, B. J., Burton, D. L., and Georgallas, A. (2015). Predicting soil nitrogen supply from soil properties. *Canadian Journal of Soil Science, 95*(1), 63-75. doi:10.4141/cjss-2014-057

Dessureault-Rompre, J., Zebarth, B. J., Burton, D. L., Georgallas, A., Sharifi, M., Porter, G. A., Moreau, G., Leclerc, Y., Arsenault, W. J., Chow, T. L., and Grant, C. A. (2012). Prediction of soil nitrogen supply in potato fields using soil temperature and water content information. *Soil Science Society of America Journal, 76*(3), 936-949. doi:10.2136/sssaj2011.0377

Di Paola, A., Valentini, R., and Santini, M. (2016). An overview of available crop growth and yield models for studies and assessments in agriculture. *Journal of the Science of Food and Agriculture, 96*(3), 709-714. doi:10.1002/jsfa.7359

Diriba, S. G. (2017). Water-nutrients interaction: exploring the effects of water as a central role for availability & use efficiency of nutrients by shallow rooted vegetable crops – a review *J. Agric. Crops, 3*(10), 78-93.

Dodds, K. G., Sinclair, A. G., and Morrison, J. D. (1995). A bivariate response surface for growth data. *Fertilizer Research, 45*(2), 117-122. doi:10.1007/bf00790661

Dubetz, S., and Bole, J. B. (1975). Effect of nitrogen, phosphorus, and potassium fertilizers on yield components and specific gravity of potatoes. *American Potato Journal, 52*(12), 399-405. doi:10.1007/bf02852794

Egozcue, J. J., and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology, 37*(7), 795-828. doi:10.1007/s11004-005-7381-9

Ellissèche, D. (1996). Aspects physiologiques de la croissance et du développement. In P. Rousselle, Y. Robert, and J. C. Crosnier (Eds.), *La pomme de terre: production, amélioration, ennemis et maladies, utilisations* (pp. 71-124). PARIS: INRA.

Feddes, R. A. (1971). *Water, heat and crop growth.* Veenman, Wageningen. Retrieved from http://edepot.wur.nl/193068

Firman, D. M., and Allen, E. J. (2007). Agronomic practices. In D. Vreugdenhil (Ed.), *Potato Biology and Biotechnology* (pp. 719-738): Elsevier.

Flis, S. (2019). 4R practices for fertilizer management in potatoes. *Crops & Soils, 52*(2), 8-10.

Fortin, J. G., Anctil, F., Parent, L.-É., and Bolinder, M. A. (2010). A neural network experiment on the site-specific simulation of potato tuber growth in Eastern Canada. *Computers and Electronics in Agriculture, 73*(2), 126-132. doi:https://doi.org/10.1016/j.compag.2010.05.011

Fortin, J. G., Anctil, F., Parent, L. E., and Bolinder, M. A. (2008). Comparison of empirical daily surface incoming solar radiation models. *Agricultural and Forest Meteorology, 148*(8-9), 1332-1340. doi:10.1016/j.agrformet.2008.03.012

Fortin, J. G., Morais, A., Anctil, F., and Parent, L. E. (2015). SVMLEACH - NK POTATO: A simple software tool to simulate nitrate and potassium co-leaching under potato crop. *Computers and Electronics in Agriculture, 110*, 259-266. doi:10.1016/j.compag.2014.11.025

Gee, G. W., and Bauder, J. W. (1986). Particle-size analysis. In A. Klute (Ed.), *Methods of soil analysis: Part 1 - Physical and mineralogical methods (Agronomy M.)* (pp. 383–411): Soil Science Society of America, Madison, Wisconsin.

Gianquinto, G., and Bona, S. (2000). The significance of trends in concentrations of total nitrogen and nitrogenous compounds. In H. A. J. and M. D. K. L. (Eds.), *Management of nitrogen and water in potato production* (pp. 35-54). Wageningen.

Gregory, P. J., and Simmonds, L. P. (1992). Water relations and growth of potatoes. In *The potato crop* (pp. 214-246): Springer.

Grewal, K. S., Buchan, G. D., and Sherlock, R. R. (1991). A comparison of three methods of organic carbon determination in some New Zealand soils. *Journal of Soil Science, 42*(2), 251-257.

Griffin, T. S., Johnson, B. S., and Ritchie, J. T. (1993). *A simulation model for potato growth and development: Substor-potato Version 2.0*: Michigan State University, Department of Crop and Soil Sciences.

Hatfield, J. L., and Walthall, C. L. (2015). Meeting global food needs: realizing the potential via genetics x environment x management interactions. *Agronomy Journal, 107*(4), 1215-1226. doi:10.2134/agronj15.0076

Haverkort, A. J., and Struik, P. C. (2015). Yield levels of potato crops: Recent achievements and future prospects. *Field Crops Research, 182*, 76-85. doi:10.1016/j.fcr.2015.06.002

Hawkesford, M., Horst, W., Kichey, T., Lambers, H., Schjoerring, J., Møller, I. S., and White, P. (2012). Chapter 6 - Functions of Macronutrients. In *Marschner's mineral nutrition of higher plants* (Third ed., pp. 135-189). San Diego: Academic Press.

Hendershot, W. H., Lalande, H., and Duquette, M. (1993). Soil reaction and exchangeable acidity. In M. R. Carter and E. G. Gregorich (Eds.), *Soil sampling and methods of analysis* (2nd ed., Vol. 2, pp. 201-206).

Hofman, G., and Salomez, J. (2000). Management of nitrogen and water in potato production. In A. J. Haverkort and D. K. L. MacKerron (Eds.), *Pers, Wageningen* (pp. 121–135).

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing In Science Engineering, 9*(3), 90-95.

Hutchinson, M. F., McKenney, D. W., Lawrence, K., Pedlar, J. H., Hopkinson, R. F., Milewska, E., and Papadopol, P. (2009). Development and testing of Canada-wide interpolated spatial models of daily minimum-maximum temperature and precipitation for 1961-2003. *Journal of Applied Meteorology and Climatology, 48*(4), 725-741. doi:10.1175/2008jamc1979.1

Hüwing, H. (2012). Düngung sichert ertrag und qualität. *Land & Fort*, 12, 22nd March 2012, 2036-2038.

Inman, D., Khosla, R., Westfall, D. G., and Reich, R. (2005). Nitrogen uptake across site specific management zones in irrigated corn production systems. *Agronomy Journal, 97*(1), 169-176.

Isfan, D., Zizka, J., Davignon, A., and Deschenes, M. (1995). Relationships between nitrogen rate, plant nitrogen concentration, yield, and residual soil nitrate-nitrogen in silage corn. *Communications in Soil Science and Plant Analysis, 26*(15-16), 2531-2557. doi:10.1080/00103629509369466

Jiang, Y. F., Zebarth, B., and Love, J. (2011). Long-term simulations of nitrate leaching from potato production systems in Prince Edward Island, Canada. *Nutrient Cycling in Agroecosystems, 91*(3), 307-325. doi:10.1007/s10705-011-9463-z

Kahle, D., and Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *The R Journal, 5*(5), 144-161.

Khiari, L., Parent, L. E., Pellerin, A., Alimi, A. R. A., Tremblay, C., Simard, R. R., and Fortin, J. (2000). An agri-environmental phosphorus saturation index for acid coarse-textured soils. *Journal of Environmental Quality, 29*(5), 1561-1567.

Kirkman, M. A. (2007). Global markets for processed potato products. In *Potato Biology and Biotechnology, advances and perspectives* (pp. 27-44): Elsevier.

Kleinkopf, G. E., Westermann, D. T., and Dwelle, R. B. (1981). Dry matter production and nitrogen utilization by six potato cultivars. *Agronomy Journal, 73*(5), 799-802.

Kooman, P. L., Fahem, M., Tegera, P., and Haverkort, A. J. (1996). Effects of climate on different potato genotypes. 2. Dry matter allocation and duration of the growth cycle. *European Journal of Agronomy, 5*(3-4), 207-217. doi:10.1016/s1161-0301(96)02032-1

Kyveryga, P. M., Blackmer, A. M., and Morris, T. F. (2007a). Alternative benchmarks for economically optimal rates of nitrogen fertilization for corn. *Agronomy Journal, 99*(4), 1057-1065. doi:10.2134/agronj2006.0340

Kyveryga, P. M., Blackmer, A. M., and Morris, T. F. (2007b). Disaggregating model bias and variability when calculating economic optimum rates of nitrogen fertilization for corn. *Agronomy Journal, 99*(4), 1048-1056. doi:10.2134/agronj2006.0339

Laboski, C. A. M., and Kelling, K. A. (2007). Influence of fertilizer management and soil fertility on tuber specific gravity: a review. *American Journal of Potato Research, 84*(4), 283-290.

Leblanc, M. A., Gagné, G., and Parent, L. E. (2016). Numerical clustering of soil series using morphological profile attributes for potato. In A. E. Hartemink and B. Minasny (Eds.), *Digital Soil Morphometrics*. New York, NY: Springer.

Levy, D., and Veilleux, R. E. (2007). Adaptation of potato to high temperatures and salinity-a review. *American Journal of Potato Research, 84*(6), 487-506.

Li, H., Parent, L. E., Tremblay, G., and Karam, A. (1999). Potato response to crop sequence and nitrogen fertilization following sod breakup in a Gleyed Humo-Ferric Podzol. *Canadian Journal of Plant Science, 79*(3), 439-446. doi:10.4141/p98-042

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. (2018). Machine learning in agriculture: a review. *Sensors, 18*(8), 1-29. doi:10.3390/s18082674

MacKerron, D. K. L. (2007). Mathematical models of plant growth and development. In D. Vreugdenhil (Ed.), *Potato Biology and Biotechnology* (pp. 753-776): Elsevier.

Maier, N. A., Dahlenburg, A. P., and Williams, C. M. J. (1994). Effects of nitrogen, phosphorus, and potassium on yield, specific gravity, crisp colour, and tuber chemical composition of potato (*Solanum tuberosum* L.) cv. Kennebec. *Australian Journal of Experimental Agriculture, 34*(6), 813-824. doi:10.1071/ea9940813

Marouani, A., Behi, O., Ben Ammar, H., Sahli, A., and Ben Jeddi, F. (2015). Effect of various sources of nitrogen fertilizer on yield and tubers nitrogen accumulation of Spunta potato cultivar (*Solanum tuberosum* L.). *J. of New Sciences, Agriculture and Biotechnology, 13*(1), 399-404.

Marshall, B. (2007). Decision support systems in potato production. In D. Vreugdenhil (Ed.), *Potato Biology and Biotechnology* (pp. 777-800): Elsevier.

Martin-Fernandez, J. A., Barcelo-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology, 35*(3), 253-278. doi:10.1023/a:1023866030544

McKinney, W. (2010). *Data structures for statistical computing in python.* Paper presented at the Proceedings of the 9th Python in Science Conference.

Michaelson, G. J., Ping, C. L., and Mitchell, G. A. (1987). Correlation of Mehlich 3, Bray 1, and ammonium acetate extractable P, K, Ca, and Mg for Alaska agricultural soils. *Communications in Soil Science and Plant Analysis, 18*(9), 1003-1015. doi:10.1080/00103628709367877

Mondani, F., Golzardi, F., Ahmadvand, G., Ghorbani, R., and Moradi, R. (2011). Influence of weed competition on potato growth, production and radiation use efficiency. *Notulae Scientia Biologicae, 3*, 42-52. doi:10.15835/nsb.3.3.6125

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the Asabe, 50*(3), 885-900.

Morton, J. T., Sanders, J., Quinn, R. A., McDonald, D., Gonzalez, A., Vázquez-Baeza, Y., Navas-Molina, J. A., Song, S. J., Metcalf, J. L., Hyde, E. R., Lladser, M., Dorrestein, P. C., and Knight, R. (2017). Balance trees reveal microbial niche differentiation. *mSystems, 2*(1), e00162-00116. doi:10.1128/mSystems.00162-16

Moulin, A. P., Cohen, Y., Alchanatis, V., Tremblay, N., and Volkmar, K. (2012). Yield response of potatoes to variable nitrogen management by landform element and in relation to petiole nitrogen - A case study. *Canadian Journal of Plant Science, 92*(4), 771-781. doi:10.4141/cjps2011-005

Mucherino, A., Papajorgji, P., and Pardalos, P. M. (2009). A survey of data mining techniques applied to agriculture. *Operational Research, 9*(2), 121-140.

Murphy, J., and Riley, J. P. (1962). A modified single solution method for the determination of phosphate in natural waters. *Analytica chimica acta, 27*, 31-36.

Neeteson, J. J., and Zwetsloot, H. J. C. (1989). An analysis of the response of sugar beet and potatoes to fertilizer nitrogen and mineral soil mineral nitrogen. *Netherlands Journal of Agricultural Science, 37*(2), 129-141.

Nelson, D. W., and Sommers, L. E. (1982). Total carbon, organic carbon, and organic matter. In *Methods of soil analysis. Part 2. Chemical and microbiological properties* (pp. 539-579).

Palarea-Albaladejo, J., and Martin-Fernandez, J. A. (2015). zCompositions - R Package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems, 143*, 85-96. doi:10.1016/j.chemolab.2015.02.019

Parent, L. E. (2014). Nouveaux outils de gestion de l'azote dans la production de la pomme de terre. *CRAAQ, Colloque sur la pomme de terre 2014*.

Parent, S. E., Leblanc, M., Parent, A. C., Coulibali, Z., and Parent, L. E. (2017). Site-specific multilevel modeling of potato response to nitrogen fertilization. *Front. Environ. Sci., 5*(81), 1-18. doi:10.3389/fenvs.2017.00081

Parent, S. E., Parent, L. E., Rozane, D. E., and Natale, W. (2013). Plant ionome diagnosis using sound balances: case study with mango (*Mangifera Indica*). *Frontiers in plant science, 4*, 1-12.

Parizeau, M. (2006). Réseaux de neurones. *University Laval*, 27-51.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *12(Oct)*, 2825−2830.

Pellerin, A. (2010). Les grilles de références. In L. E. Parent and G. Gagné (Eds.), *Guide de référence en fertilisation* (2è ed., pp. 359-473).

Pellerin, A., Parent, L. E., Fortin, J., Tremblay, C., Khiari, L., and Giroux, M. (2006a). Environmental Mehlich-III soil phosphorus saturation indices for Quebec acid to near neutral mineral soils varying in texture and genesis. *Canadian Journal of Soil Science, 86*(4), 711-723. doi:10.4141/s05-070

Pellerin, A., Parent, L. E., Tremblay, C., Fortin, J., Tremblay, G., Landry, C. P., and Khiari, L. (2006b). Agri-environmental models using Mehlich-III soil phosphorus saturation index for corn in Quebec. *Canadian Journal of Soil Science, 86*(5), 897-910.

Peralta, J. M., and Stockle, C. O. (2002). Dynamics of nitrate leaching under irrigated potato rotation in Washington State: a long-term simulation study. *Agriculture, ecosystems & environment, 88*(1), 23-34.

Petropoulos, S. A., Fernandes, Â., Polyzos, N., Antoniadis, V., Barros, L., and CFR Ferreira, I. (2020). The impact of fertilization regime on the crop performance and chemical composition of potato (*Solanum tuberosum* L.) cultivated in central Greece. *Agronomy, 10*(4), 474-491.

Piikki, K., Wetterlind, J., Soderstrom, M., and Stenberg, B. (2015). Three-dimensional digital soil mapping of agricultural fields by integration of multiple proximal sensor data obtained from different sensing methods. *Precision Agriculture, 16*(1), 29-45. doi:10.1007/s11119-014-9381-6

Qin, Z. S., Myers, D. B., Ransom, C. J., Kitchen, N. R., Liang, S. Z., Camberato, J. J., Carter, P. R., Ferguson, R. B., Fernandez, F. G., Franzen, D. W., Laboski, C. A. M., Malone, B. D., Nafziger, E. D., Sawyer, J. E., and Shanahan, J. F. (2018). Application of machine learning methodologies for predicting corn economic optimal nitrogen rate. *Agronomy Journal, 110*(6), 2596-2607. doi:10.2134/agronj2018.03.0222

R Core Team. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Rajsic, P., and Weersink, A. (2008). Do farmers waste fertilizer? A comparison of ex post optimal nitrogen rates and ex ante recommendations by model, site and year. *Agricultural Systems, 97*(1-2), 56-67. doi:10.1016/j.agsy.2007.12.001

Raman, K. V., and Radcliffe, E. B. (1992). *The potato crop: the scientific basis for improvement* (P. M. Harris Ed. 2nd ed. Vol. 2). London: Chapman and Hall.

Rangarajan, A. (2009). Crop rotation effects on soil fertility and plant nutrition. In C. L. Mohler and S. E. Johnson (Eds.), *Crop Rotation on Organic Farms*. University of Maryland: NRAES.

Rasmussen, C. E., and Williams, C. K. I. (2006). Gaussian processes for machine learning. *The MIT Press, Cambridge, MA, USA, 38*, 715-719.

Raymundo, R., Asseng, S., Cammarano, D., and Quiroz, R. (2014). Potato, sweet potato, and yam models for climate change: a review. *Field Crops Research, 166*, 173-185. doi:10.1016/j.fcr.2014.06.017

Rex, B. L. (1991). The effect of in-row seed piece spacing and harvest date of the tuber yield and processing quality of Conestoga potatoes in southern Manitoba. *Canadian Journal of Plant Science, 71*(1), 289-296. doi:10.4141/cjps91-039

Rich, A. E. (1983). *Potato diseases*. New York: Academic Press.

Sands, P. J., Hackett, C., and Nix, H. A. (1979). A model of the development and bulking of potatoes (*Solanum Tuberosum* L.) I. Derivation from well-managed field crops. *Field Crops Research, 2*, 309-331. doi:https://doi.org/10.1016/0378-4290(79)90031-5

Sharifi, M., Zebarth, B. J., Porter, G. A., Burton, D. L., and Grant, C. A. (2009). Soil mineralizable nitrogen and soil nitrogen supply under two-year potato rotations. *Plant and Soil, 320*(1-2), 267-279. doi:10.1007/s11104-009-9892-5

Sincik, M., Turan, Z. M., and Goksoy, A. T. (2008). Responses of potato (*Solanum tuberosum* L.) to green manure cover crops and nitrogen fertilization rates. *American Journal of Potato Research, 85*(2), 150-158. doi:10.1007/s12230-008-9011-9

Sinclair, T. R., and Seligman, N. (2000). Criteria for publishing papers on crop modeling. *Field Crops Research, 68*(3), 165-172. doi:10.1016/s0378-4290(00)00105-2

Singh, J., Knapp, H. V., Arnold, J. G., and Demissie, M. (2005). Hydrological modeling of the Iroquois river watershed using HSPF and SWAT 1. *Journal of the American Water Resources Association, 41*(2), 343-360.

Soil Classification Working Group. (1998). Canadian system of soil classification, 3rd Ed. *Canadian system of soil classification*, 188.

Soman, T., and Bobbie, P. O. (2005). Classification of arrhythmia using machine learning techniques. *WSEAS Transactions on computers, 4*(6), 548-552.

Stalham, M. A., Allen, E. J., and Herry, F. X. (2005). Effects of soil compaction on potato growth and its removal by cultivation. Research review. (R261), 1-60.

Struik, P. C. (2007). Above-ground and below-ground plant development. In D. Vreugdenhil (Ed.), *Potato biology and biotechnology: advances and perspectives* (pp. 219–236): Amsterdam: Elsevier, New York.

Tabi, M., Tardif, L., Carrier, D., Laflamme, G., and Rompré, M. (1990). Inventaire des problèmes de dégradation des sols agricoles du Québec: rapport synthèse. *Entente auxiliaire Canada-Québec sur le développement agro-alimentaire Québec. Service de recherche en sols*.

Templ, M., Hron, K., and Filzmoser, P. (2011). robCompositions: an R-package for robust statistical analysis of compositional data. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis. Theory and Applications* (pp. 341-355). Chichester (UK): John Wiley & Sons.

Torma, S., Vilcek, J., Losak, T., Kuzel, S., and Martensson, A. (2018). Residual plant nutrients in crop residues - an important resource. *Acta Agriculturae Scandinavica Section B-Soil and Plant Science, 68*(4), 358-366. doi:10.1080/09064710.2017.1406134

Tran, T. S., and Simard, R. R. (1993). Mehlich-III extractable elements. In M. R. Carter (Ed.), *Soil sampling and methods of analysis* (pp. 43-49): Canadian Society of Soil Science, CRC Press, Boca Raton, FL.

Trehan, S. P., Roy, S. K., and Sharma, R. C. (2001). Potato variety differences in nutrient deficiency symptoms and responses to NPK. *Better Crops International. Potash and Phosphate Institute of Canada (PPIC), 15*, 18-21.

Tremblay, N., Bouroubi, Y. M., Bélec, C., Mullen, R. W., Kitchen, N. R., Thomason, W. E., Ebelhar, S., Mengel, D. B., Raun, W. R., and Francis, D. D. (2012). Corn response to nitrogen is influenced by soil texture and weather. *Agronomy Journal, 104*(6), 1658-1671.

Valkama, E., Salo, T., Esala, M., and Turtola, E. (2013). Nitrogen balances and yields of spring cereals as affected by nitrogen fertilization in northern conditions: A meta-analysis. *Agriculture Ecosystems & Environment, 164*, 1-13. doi:10.1016/j.agee.2012.09.010

Valkama, E., Uusitalo, R., Ylivainio, K., Virkajarvi, P., and Turtola, E. (2009). Phosphorus fertilization: a meta-analysis of 80 years of research in Finland. *Agriculture Ecosystems & Environment, 130*(3-4), 75-85. doi:10.1016/j.agee.2008.12.004

Van den Boogaart, K. G., Raimon, T., and Bren, M. (2014). compositions: compositional data analysis. R package version 1.40-1.

Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering, 13*(2), 22-30.

Van Rossum, G., and Drake Jr, F. L. (1995). *Python tutorial, technical report CS R9526*: Centrum voor Wiskunde en Informatica (CWI) Amsterdam.

Vanlauwe, B., Kihara, J., Chivenge, P., Pypers, P., Coe, R., and Six, J. (2011). Agronomic use efficiency of N fertilizer in maize-based systems in sub-Saharan Africa within the context of integrated soil fertility management. *Plant and Soil, 339*(1-2), 35-50. doi:10.1007/s11104-010-0462-7

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., and Bright, J. (2019). SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python. *arXiv preprint arXiv:1907.10121*.

White, P. J., Wheatley, R. E., Hammond, J. P., and Zhang, K. (2007). Minerals, soils and roots. In *Potato Biology and Biotechnology* (pp. 739-752): Elsevier.

Wickham, H. (2017). Tidyverse: easily install and load the 'Tidyverse'. R package version 1.2.1. In.

Xu, Y., Jimenez, M. A., Parent, S. E., Leblanc, M., Ziadi, N., and Parent, L. E. (2017). Compaction of coarse-textured soils: balance models across mineral and organic compositions. *Frontiers in Ecology and Evolution, 5*. doi:10.3389/fevo.2017.00083

Yang, X. L., Zhang, Q. Y., Li, X. Z., Jia, X. X., Wei, X. R., and Shao, M. A. (2015). Determination of soil texture by laser diffraction method. *Soil Science Society of America Journal, 79*(6), 1556-1566. doi:10.2136/sssaj2015.04.0164

Young, D. A., Voisey, P. W., and Dixon, N. (1964). A specific gravity calculator for potatoes. *American Journal of Potato Research, 41*(12), 401-405.

Yuan, J., Liu, C. L., Li, Y. M., Zeng, Q. B., and Zha, X. F. (2010). Gaussian processes based bivariate control parameters optimization of variable-rate granular fertilizer applicator. *Computers and Electronics in Agriculture, 70*(1), 33-41. doi:10.1016/j.compag.2009.08.009

Zebarth, B. J., Arsenault, W. J., Moorehead, S., Kunelius, H. T., and Sharifi, M. (2009a). Italian ryegrass management effects on nitrogen supply to a subsequent potato crop. *Agronomy Journal, 101*(6), 1573-1580. doi:10.2134/agronj2009.0184

Zebarth, B. J., Danielescu, S., Nyiraneza, J., Ryan, M. C., Jiang, Y. F., Grimmett, M., and Burton, D. L. (2015a). Controls on nitrate loading and implications for BMPs under intensive potato production systems in Prince Edward Island, Canada. *Ground Water Monitoring and Remediation, 35*(1), 30-42. doi:10.1111/gwmr.12088

Zebarth, B. J., Karemangingo, C., Scott, P., Savoie, D., and Moreau, G. (2007). Nitrogen management for potato: general fertilizer recommendations. *New-Brunswick Ministry of Agriculture, Fisheries and Aquaculture, Fredericton, NB, Canada*.

Zebarth, B. J., Leclerc, Y., Moreau, G., and Botha, E. (2004). Rate and timing of nitrogen fertilization of Russet Burbank potato: Yield and processing quality. *Canadian Journal of Plant Science, 84*(3), 855-863. doi:10.4141/p03-123

Zebarth, B. J., Ryan, M. C., Graham, G., Forge, T. A., and Neilsen, D. (2015b). Groundwater monitoring to support development of BMPs for groundwater protection: the Abbotsford-Sumas aquifer case study. *Ground Water Monitoring and Remediation, 35*(1), 82-96. doi:10.1111/gwmr.12092

Zebarth, B. J., Scott, P., and Sharifi, M. (2009b). Effect of straw and fertilizer nitrogen management for spring barley on soil nitrogen supply to a subsequent potato crop. *American Journal of Potato Research, 86*(3), 209-217. doi:10.1007/s12230-009-9074-2

Zhang, D., and Jeffery, J. P. T. (2007). *Advances in machine learning applications in software engineering*: IGI Global.

3.10 Supporting Information

S 3.1 Table. Description of the marketable yield modeling data set

Study year	Number of trials	Number of samples	Percentage (%)	Minimum number of blocks	Maximum number of blocks	Minimum number of treatments	Maximum number of treatments
1979	1	10	0.2	3	3	4	4
1980	1	10	0.2	3	3	4	4
1981	3	30	0.5	6	6	4	4
1987	1	8	0.1	2	2	4	4
1993	6	144	2.4	3	3	27	27
1994	3	84	1.4	3	3	10	10
1995	3	81	1.4	3	3	9	9
1996	8	258	4.4	6	6	11	11
1997	8	306	5.2	6	6	22	22
1998	14	280	4.7	6	6	5	5
1999	18	431	7.3	6	6	14	14
2000	8	184	3.1	6	6	5	5
2001	8	152	2.6	4	4	5	5
2002	8	183	3.1	4	4	11	11
2003	21	348	5.9	4	4	10	10
2004	11	267	4.5	3	3	11	11
2005	20	574	9.7	4	4	11	11
2006	10	186	3.1	3	3	10	10
2007	15	297	5.0	3	3	12	12
2008	6	107	1.8	3	3	7	7
2009	6	118	2.0	5	5	8	8
2010	8	154	2.6	4	4	8	8
2011	13	261	4.4	4	4	8	8
2012	13	268	4.5	4	4	8	8
2013	16	344	5.8	4	4	8	8
2014	25	535	9.0	4	4	10	10
2015	2	33	0.6	3	3	6	6
2016	5	48	0.8	3	3	6	6
2017	12	212	3.6	3	3	8	8
Total	273	5913	100.0				

Similar tables are reproducible for the tuber-size balances and specific gravity models by using the codes (file 1.3) available at https://git.io/JvYxd.

S 3.2 Table. Description of the data sets used for modeling per trial type

A. Marketable yield

Trial type	Number of samples	Number of trials	Percentage	Minimum N dosage	Maximum N dosage	Minimum P_2O_5 dosage	Maximum P_2O_5 dosage	Minimum K_2O dosage	Maximum K_2O dosage
			(%)	(kg ha^{-1})					
K	936	45	15.8	80	260	75	240	0	420
N	3,068	151	51.9	0	250	88	250	57	270
NPK	591	16	10.0	0	225	0	300	0	300
P	1,300	60	22.0	0	260	0	300	0	270
NA	18	1	0.3	218	218	110	110	55	55
Total	**5,913**	**273**	**100.0**						

B. Tuber size balances

Trial type	Number of samples	Number of trials	Percentage	Minimum N dosage	Maximum N dosage	Minimum P_2O_5 dosage	Maximum P_2O_5 dosage	Minimum K_2O dosage	Maximum K_2O dosage
			(%)	(kg ha^{-1})					
K	901	43	19.8	80	220	75	200	0	300
N	2,378	122	52.2	0	250	100	216	57	270
NPK	363	9	8.0	0	225	0	300	0	300
P	897	33	19.7	110	210	0	300	0	270
NA	18	1	0.4	218	218	110	110	55	55
Total	**4,557**	**208**	**100.0**						

C. Tuber specific gravity

Trial type	Number of samples	Number of trials	Percentage	Minimum N dosage	Maximum N dosage	Minimum P_2O_5 dosage	Maximum P_2O_5 dosage	Minimum K_2O dosage	Maximum K_2O dosage
			(%)	(kg ha^{-1})					
K	880	42	20.7	80	260	75	240	0	420
N	1,956	117	46.0	0	250	88	215	57	270
NPK	410	16	9.6	0	225	0	300	0	300
P	990	38	23.3	110	260	0	300	0	270
NA	18	1	0.4	218	218	110	110	55	55
Total	**4,254**	**214**	**100.0**						

S 3.3 Table. Classification of preceding crops (Parent et al., 2017)

Metal-level	Preceding crop
Small grain	Oat, oat and mustard, wheat, canola, cereal, mustard, white mustard, barley, barley and mustard, sunola, and cereal
High-residue crops	Grain corn, pearl millet, rye
Legume	Perennial legumes: birdsfoot trefoil, alfalfa, legume mix, clover, ryegrass, and clover
Grassland	Orchard grass, ryegrass, and cereal
Low-residue crops	Annual low-residue crops: broccoli, cabbage, Chinese cabbage, beans, strawberry, silage corn, onion, pea, potato, potato followed by green manure, buckwheat, soybean

S 3.4 Table. Centroids of soil textural classes derived from the Quebec soils data set (Tabi et al., 1990)

Textural class	Sand (%)	Silt (%)	Clay (%)
Coarse sand	93.8	3.3	2.9
Sand	92.7	4.4	2.9
Fine sand	92.3	4.6	3.1
Very fine sand	92.9	4.4	2.7
Coarse loamy sand	82.5	11.3	6.2
Loamy sand	82.5	12.1	5.4
Fine loamy sand	81.4	14.4	4.2
Very fine loamy sand	80.7	11.2	8.1
Coarse sandy loam	66.2	20.4	13.4
Sandy loam	64.5	25.0	10.5
Fine sandy loam	60.4	31.6	8.0
Very fine sandy loam	60.2	32.8	7.0
Loam	42.4	41.1	16.5
Silty loam	18.2	67.0	14.8
Silt	6.0	88.1	5.9
Sandy clay loam	55.7	19.9	24.4

Clay loam	29.5	38.4	32.1
Silty clay loam	7.8	58.1	34.1
Sandy clay	50.2	11.2	38.6
Silty clay	2.7	48.1	49.2
Clay	11.9	34.9	53.2
Heavy clay	1.3	23.9	74.8

S 3.5 Table. Quebec potato data set used for modeling. ‘Potato_df.csv’ file available in ‘data’ repository at https://git.io/JvYxd.

Chapitre 4 : Discussion générale

L'analyse de l'ionome, objet du chapitre 2, a montré qu'il est hasardeux de regrouper les cultivars en lien avec les données de compositions foliaires disponibles et l'algorithme de classification non supervisée (*dbscan*) utilisé. Cependant, l'étude a montré qu'avec ces données, les algorithmes des k plus proches voisins, des forêts aléatoires et des machines à vecteur de supports ont un potentiel de diagnostic acceptable pour détecter un déséquilibre nutritionnel en cours de saison avec une précision de 70 %. De plus, en considérant une sous-population de tête regroupant les observations ayant un rendement en tubercules vendables supérieur ou égal à celui du $65^{ème}$ percentile pour chaque cultivar, le vecteur de perturbation de l'espace de composition d'Aitchison (Aitchison and Ng, 2005) pourrait illustrer la sensibilité des nouveaux cultivars à la fertilisation. Des études ionomiques antérieures montrent des précisions de 75 à 85 % (Hernandes et al., 2011; Parent et al., 2013a; Parent et al., 2013b), mais leurs précisions sont rapportées sur des données d'entraînement sans évaluation en prédiction. Les qualités de prédiction rapportées dans cette étude ont été évaluées sur des données test.

La réponse des cultures à la fertilisation est conditionnée par les indices météorologiques, les caractéristiques pédologiques de la couche de surface et du sous-sol et les pratiques culturales locales (Leblanc, 2016; Parent et al., 2017). Au chapitre 3, l'analyse de l'importance des variables dans les modèles prédictifs les classe dans un ordre d'importance variant en fonction de la variable dépendante. Seul le précédent cultural pouvait être éliminé sans influence majeure sur les performances des modèles. Pour les modèles de prédiction du rendement des tubercules vendables et le ratio caractérisant les tubercules de gros calibre, l'azote s'est révélé être la variable la plus informative. L'azote est impliqué dans les phases sensibles de la production agricole et ses effets de carences ou d'excès sont largement étudiés. En visant des tubercules de calibre moyen comme objectif de production, la densité de plantation joue un rôle plus important dans les modèles parce que des densités plus élevées favorisent des rendements plus élevés de tubercules de petits et moyens calibres. Le poids spécifique des tubercules dépend en grande partie de la répartition des pluies sur la saison culturale (indice de Shannon) puis de leur abondance. L'irrigation (ou la disponibilité régulière de l'eau) durant les premiers stades de croissance tend à augmenter la teneur en matière

sèche des tubercules qui peut être considérablement réduite en cas de fortes pluies en fin de saison avant la récolte (Al Soboh et al., 2002).

Le rendement vendable, les balances caractérisant les calibres et le poids spécifique des tubercules ont été modélisés en fonction (1) des ratios logarithmiques isométriques calculés à partir de la composition du sol de surface en P, K, Ca, Mg et Al, (2) du pH du sol, (3) des ratios logarithmiques isométriques calculés à partir des propriétés du sous-sol, (4) des précipitations cumulées, (5) de l'indice de diversité de Shannon, (6) du cumul des températures moyennes, (7) des degrés-jours de croissance, (8) de la durée de la saison, (9) la densité de plantation et (10) des doses de N, P et K appliquées. Ces variables ou les données à partir desquelles elles sont calculées sont accessibles à tous les conseillers et producteurs agricoles au Québec.

Les algorithmes des k plus proches voisins (KNN), des forêts aléatoires (RF), les réseaux neuronaux (NN) et les processus gaussiens (GP) ajustés à ces conditions expérimentales ont des coefficients de détermination pratiquement similaires et supérieurs à celui du modèle de Mitscherlich pour le rendement vendable. Les R^2 étaient de 0,52, 0,59, 0,49 et 0,58 respectivement pour les modèles d'autoapprentissage (KNN, RF, NN, GP) et 0,37 pour le modèle de Mitscherlich. Les erreurs absolues moyennes (MAE) et quadratiques moyennes (RMSE) sont demeurées également similaires entre les premiers modèles (variant entre 5,1 et 5,6 Mg ha^{-1} pour MAE et entre 6,7 et 7,5 Mg ha^{-1} pour RMSE), mais supérieurs pour les modèles de Mitscherlich (8.3 Mg ha^{-1} et 10.8 Mg ha^{-1} pour MAE et RMSE respectivement). Bien que les différences soient faibles, le modèle du log-ratio isométrique entre les rendements des tubercules de calibres moyen et petit ont présenté les plus grandes valeurs de R^2 (0.60 – 0.69) suivi du poids spécifique (0.58 – 0.67), puis du log-ratio isométrique caractérisant la proportion de tubercules de gros calibre (0.55 – 0.64). Les dissemblances sont apparues entre les modèles dans les courbes de réponse et la détermination des doses économiques optimales de fertilisation N, P et K. Des courbes de réponse continues sont obtenues pour les modèles NN et GP les rendant plus propices pour faire des inférences. Les prédictions de doses économiques optimales ou agronomiques optimales sont spécifiques au site et varient également entre les modèles. Cependant, il est possible avec les modèles GP d'élaborer des recommandations stratégiques dans un cadre probabiliste et éventuellement d'estimer un risque

économique. Ces modèles peuvent échantillonner de nombreuses fonctions plausibles pour en déduire une distribution de doses optimales. Le conseiller ou le producteur peut apprécier les avantages en termes de rendement et de revenus d'appliquer une dose optimale plutôt qu’une autre. Le modèle par processus gaussien semble ainsi être un outil approprié pour les recommandations d'engrais en culture de pomme de terre au Québec.

Conclusion générale et perspectives

L'objectif de ce projet de doctorat était de proposer un modèle de recommandation des doses optimales à la fois pour l'azote, le phosphore et le potassium selon le rendement, le calibre et le poids spécifique en culture de pomme de terre. Pour tenir compte de la complexité des facteurs d'impact en production agricole, il a été préalablement établi de combiner aux fertilisants appliqués des indices spécifiques au sol, à la génétique et aux pratiques agronomiques locales. Ce travail a été divisé en deux parties. La première a testé la possibilité de structurer les cultivars sur la base de la composition foliaire en N, P, K, Ca et Mg en vue de produire une variable recouvrant de l'information génétique. La deuxième partie devait combiner cette variable aux autres facteurs d'impact afin d'élaborer le modèle de prédiction et ce, avec les données de 273 essais de fertilisation menés au Québec depuis 1970.

Depuis que le concept d'analyse des données de composition a été appliqué aux tissus végétaux, plusieurs études ont classé des espèces de plantes et des cultivars en utilisant une analyse multivariée de la composition tissulaire en éléments nutritifs (Parent et al., 2013a; Parent et al., 2013b). Cette étude est la première à utiliser des outils statistiques pour traiter la nutrition différentielle des cultivars de pomme de terre en combinant les concentrations foliaires en nutriments et des outils d'apprentissage automatique. L'échec du regroupement laisse la réflexion ouverte pour l'implémentation d'autres procédés. La discrétisation d'une variable continue en catégories, ici en deux catégories (haut rendement et faible rendement), permet une bonne précision, mais prédire des valeurs continues est une avenue intéressante à explorer bien que les R^2 des régressions utilisant l'ionome sont reconnus pour être faibles. Étant donné que les essais de fertilisation ont été menés sur une période d'environ 50 ans (1970 – 2017), l'ajout d'une variable temporelle pourrait être pertinente puisque les pratiques agricoles et les conditions du sol ont subi des modifications substantielles (Valkama et al., 2009). Ce qui peut se refléter dans les compositions tissulaires. Les méthodes d'analyse ont également évolué avec le temps. Les analyses de N total des feuilles de patates étaient faites avec la méthode Kjeldahl et les plus récentes par combustion Dumas (Leco). Plusieurs auteurs ont proposé des modèles de conversion des données Kjeldahl en Dumas-N (Simonne et al., 1994; Simonne et al., 1997; Jung et al., 2003; Oftedal et al., 2014). De telles uniformisations

pourraient favoriser une structuration des cultivars pour leurs caractéristiques génétiques et permettre de tirer un avantage pour les modèles de fertilisation.

Pour le modèle de fertilisation, de nombreux algorithmes ont été testé pour ne présenter que ceux-ci, mais les techniques d'analyses s'améliorent avec le temps. Les quatre algorithmes d'autoapprentissage évalués sont tous à même de faire des recommandations fiables de fertilisants comparés au modèle de Mitscherlich, mais le modèle par processus gaussiens s'avère plus propice par sa capacité à ajouter une crédibilité à la recommandation. L'on pourrait travailler à son amélioration notamment au niveau des paramètres du noyau gaussien. Il pourrait également être informatif de générer un processus gaussien avec des sorties multiples en simultané afin de favoriser à la fois (par exemple) un rendement optimal de tubercules avec une forte proportion de calibre moyen et de poids spécifique élevé. Une variable régionale pourrait être introduite également pour apprécier son impact sur les performances du modèle. Il est recommandé de continuer l'approvisionnement de la BD pour l'affinage progressif du modèle. Les doses extrêmes parfois prédites à la suite de l'échantillonnage gaussien incitent à tenter des modèles en ne retenant que les essais pour lesquels il y a eu une réponse à la fertilisation. Enfin, l'ajout de données d'essais régionalisés et de données de champs collectées avec les producteurs pourraient aider à améliorer la version en cours. L'application pratique du modèle consisterait à le déployer sur plateforme web où les producteurs et conseillers pourront simuler leurs besoins en fertilisants pour compléter leur jugement professionnel.

Bibliographie

Aitchison, J., and Ng, K. W. (2005). The role of perturbation in compositional data analysis. *Statistical Modelling, 5*(2), 173-185.

Al Soboh, G., Sully, R., and Andreata, S. (2002). *Factors affecting specific gravity loss in crisping potato crops in Koo Wee Rup, Victoria*. Retrieved from https://trove.nla.gov.au/version/20463664

Hernandes, A., Parent, S.-É., Veillette, J.-P., Parent, P., Leblanc, M., Roy, G., Sylvestre, P., Samson, N., Natale, W., and Parent, L.-É. (2011). Compositional meta-analysis of the nutrient profile of potato cultivars.

Jung, S., Rickert, D. A., Deak, N. A., Aldin, E. D., Recknor, J., Johnson, L. A., and Murphy, P. A. (2003). Comparison of Kjeldahl and Dumas methods for determining protein contents of soybean products. *Journal of the American Oil Chemists Society, 80*(12), 1169-1173. doi:10.1007/s11746-003-0837-3

Leblanc, M. A. (2016). *Relations entre les caractéristiques pédologiques et les pratiques de fertilisation et de conservation des sols.* (PhD Thesis), Laval University, Quebec, CA.

Oftedal, O. T., Eisert, R., and Barrell, G. K. (2014). Comparison of analytical and predictive methods for water, protein, fat, sugar, and gross energy in marine mammal milk. *Journal of Dairy Science, 97*(8), 4713-4732. doi:10.3168/jds.2014-7895

Parent, S. E., Leblanc, M., Parent, A. C., Coulibali, Z., and Parent, L. E. (2017). Site-specific multilevel modeling of potato response to nitrogen fertilization. *Front. Environ. Sci., 5*(81), 1-18. doi:10.3389/fenvs.2017.00081

Parent, S. E., Parent, L. E., Egozcue, J. J., Rozane, D. E., Hernandes, A., Lapointe, L., Hebert-Gentile, V., Naess, K., Marchand, S., Lafond, J., Mattos, D., Barlow, P., and Natale, W. (2013a). The plant ionome revisited by the nutrient balance concept. *Frontiers in plant science, 4*. doi:10.3389/fpls.2013.00039

Parent, S. E., Parent, L. E., Rozane, D. E., and Natale, W. (2013b). Plant ionome diagnosis using sound balances: case study with mango (Mangifera Indica). *Frontiers in plant science, 4*, 1-12.

Simonne, A. H., Simonne, E. H., Eitenmiller, R. R., Mills, H. A., and Cresman, C. P. (1997). Could the dumas method replace the Kjeldahl digestion for nitrogen and crude protein determinations in foods? *Journal of the Science of Food and Agriculture, 73*(1), 39-45.

Simonne, E. H., Mills, H. A., Jones, J. B., Smittle, D. A., and Hussey, C. G. (1994). A comparison of analytical methods for nitrogen analysis in plant tissues. *Communications in Soil Science and Plant Analysis, 25*(7-8), 943-954. doi:10.1080/00103629409369090

Valkama, E., Uusitalo, R., Ylivainio, K., Virkajarvi, P., and Turtola, E. (2009). Phosphorus fertilization: a meta-analysis of 80 years of research in Finland. *Agriculture Ecosystems & Environment, 130*(3-4), 75-85. doi:10.1016/j.agee.2008.12.004