DEVELOPING PEDOTRANSFER FUNCTIONS AND SPATIAL ESTIMATES OF GROWING SEASON NITROGEN MINERALIZATION TO INFORM IMPROVED NITROGEN MANAGEMENT STRATEGIES IN PRINCE EDWARD ISLAND

by

Luke Laurence

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
May 2024

Dalhousie University is located in Mi'kma'ki, the ancestral and unceded territory of the Mi'kmaq. We are all Treaty people.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Decision support tools (DSTs) and a framework for incorporating estimates of growing season nitrogen (N) mineralization (GSN) in soil, using analysis of total soil N (TN) to quantify the stable N pool, and biological nitrogen availability (BNA) to quantify the labile N pool in a zero- plus first-order kinetic function, were developed to inform N fertilizer recommendations. Pedotransfer functions (PTFs) were trained for TN, BNA, and GSN (N response variables), which showed important predictor variables, intrinsic controls on N dynamics, and that high accuracy predictions could be obtained from surrogate soil data. Using machine learning (ML) and interpretation metrics, soil organic matter for the stable N pool and soil respiration for the labile N pool were shown to be both cost-effective and highly correlated predictors of GSN. Provincial scale digital soil maps (DSMs) were then developed with ML to create spatial estimates of N response variables where direct soil data is not available. Multi-year crop inventory covariates were generated to improve predictive strength and incorporate the influence of soil management on N response variables. Climate appeared to influence the stable N pool, whereas plant cover had a greater influence on the labile N pool. DSTs, both PTFs and DSMs, were then tested on six-separate producer fields that had been sampled to reflect infield variability. DSTs captured the differences between fields at a coarse-scale (30m) resolution. The higher quality soils used to train DST models provided a useful benchmark to identify poor quality fields, and inform N management decisions. To capture infield variability, infield maps (5m resolution) were developed and the results showed an increase in accuracy and decrease in the uncertainty of predictions. Overall, DSTs for biological processes as related to soil N dynamics were achieved using novel techniques and ML. From these results, a tiered approach to N fertilizer management was proposed. The DSM, as tier 1, provides estimates of GSN to producers who do not have point data; the PTF, as tier 2, is applied for producers who have surrogate data; and subsequent tiers are for producers with direct measures for N response variables.

LIST OF ABBREVIATIONS AND SYMBOLS USED

AC	Active carbon
AS	Aggregate stability
BC	British Columbia
BNA	Biological Nitrogen Availability
C	Carbon
CASH	Cornell's Comprehensive Assessment of Soil Health
CBFE	Cost-benefit feature elimination
CCC	Lin's concordance correlation coefficient
CEC	Cation exchange capacity
CI	Confidence interval
Cl	Clay
CLORPT	Climate (CL), organisms (O), relief (R), parent material (P), time (T)
CU	Cubist
DSM	Digital soil mapping
DST	Decision support tool
EON	Extractable organic nitrogen
Four (4)R	Right source, time, rate, and placement
GSN	Growing season nitrogen mineralization
k	Mineralization rate coefficient
MIR	Mid-infrared
ML	Machine learning/learner
MLR	Multiple linear regression

N	Nitrogen
N_0	Potentially mineralizable nitrogen
NDVI	Normalized difference vegetation index
NIR	Near-infrared
N_{min}	Cumulative nitrogen mineralization over a given time
OM	Soil organic matter
ON	Ontario
PEI	Prince Edward Island
PEIAL	Prince Edward Island Analytical Laboratory
Pool I	Gross nitrogen mineralization in the first two-week incubation period representing the labile N pool
POX_C	Permanganate oxidizable carbon
PTF	Pedotransfer function
QC	Quebec
R^2	Coefficient of determination
RF	Random forest
RFE	Recursive feature elimination
RMSE	Root mean square error
RSN	Residual soil nitrogen
Sa	Sand
scorpan	Soil (s), climate (c), organisms (o), relief (r), parent material (p), age (a), spatial position (n)
SD	Standard deviation
SGB	Stochastic gradient boosting

SHD	Soil health database
Si	Silt
SK	Saskatchewan
SOC	Soil organic carbon
SR	Soil respiration
SVM	Support vector machine
TC	Total soil organic carbon
TN	Total soil nitrogen
VIF	Variance inflation factor

ACKNOWLEDGEMENTS

Heartfelt thanks and gratitude to my supervisor, David, for the opportunity to join this extremely interesting project; and for all his help, wisdom, patience, and encouragement along the way. Also, to Brandon, a sincere and tremendous thank you for all his dedication, patience, and help in teaching me with this, my introduction to the world of machine learning and digital soil mapping. To Judith and Barry, thank you so much for all your insights, encouragement, and help from beginning to end! There were also many contributors, including Evan, Hardy, Jin, and Travis, who offered so much help and expertise at various stages along the way.

This research was supported by funding from the Natural Science and Engineering Research Council of Canada (NSERC) through a Discovery Grant (DB), the NSERC CREATE Climate Smart Soils Grant, and the Weston Family Foundation's Soil Health Imitative (DB). Thank you to the administration and faculty of Olds College of Agriculture and Technology for supporting this work, and "picking up the slack" through my absence.

Finally, I wish to thank my dear family. Patricia, Kathleen, Elizabeth, and Gabrielle, I love you.

CHAPTER 1: INTRODUCTION

There has been an ongoing effort by industry, researchers, and government agencies to improve Nitrogen (N) management in Eastern Canada (Bowles et al., 2018; He et al., 2012; Zebarth et al., 2009). This is particularly true in Prince Edward Island (PEI) where humid climatic conditions and permeable soil properties, combined with high N fertilizer rates associated with potato production, increase N loss potential to environmental receptors (Zebarth et al., 2015). A principal risk in PEI, and key driver of this thesis, was focused on minimizing the N remaining post-harvest, which is almost completely lost overwinter due to leaching or gaseous losses from denitrification (Nyiraneza et al., 2017; Sharifi et al., 2008; Zebarth et al., 2008). In alignment with the 4R nutrient stewardship (4R) initiative, where the "right" nutrient source, time, rate, and placement seek to improve fertilizer management (Bruulsema et al., 2016), this study focuses on the contributions from soil N mineralization (N_{min}) potential as an important source in order to inform right-rate N fertilizer recommendations.

The problems addressed in this thesis were twofold. The first relates to the existing N credit system in PEI, which estimates N contributions (N credits) from various sources (e.g., manure, legume crops, and soil N_{min}) against the proposed crop N requirement in order to obtain a recommended N fertilizer rate (Zebarth et al., 2008). The current system assigns a soil-based N credit (soil N_{min}) of 15 kg N/ha credit only for soils with a soil organic matter (OM) percentage above 3.5%. The average OM level in PEI is approximately 3% and therefore soil N credits are seldom assigned despite documented levels of N_{min} well in excess of this number in most soils (Nyiraneza et al., 2017; Zebarth

et al., 2008). Secondly, there is currently no soil N test in PEI wherein a crop producer can adjust N fertilizer rates based on direct measures of soil N_{min}. Due to the high degree of residual soil N loss over the fall to spring period, the Fall N test typically used in prairie Canada is not applicable and is therefore not offered by local laboratories (Zebarth et al., 2009).

Chapter 2 thus provides an introduction and background into the nature of this two-fold problem, and includes a literature review for the *quantification and prediction of soil N mineralization for supporting right-rate nutrient stewardship.* The thesis objectives, based on the literature review and outlined in Section 2.6, were to build from the predictive function presented by Dessureault-Rompre et al. (2015) in order to develop N decision support tools (DSTs) using machine learning. The DSTs, developed and tested in the following chapters, include pedotransfer functions and digital soil maps of N parameters for the purpose of enhancing the existing N credit system and providing an improved approach for informing N management strategies in PEI.

CHAPTER 2: LITERATURE REVIEW - QUANTIFICATION AND PREDICTION OF SOIL NITROGEN MINERALIZATION FOR SUPPORTING RIGHT-RATE NUTRIENT STEWARDSHIP

2.1 INTRODUCTION

To accurately quantify the cumulative nitrogen (N) mineralization of soil organic N over a given time (N_{min}), and prior to fertilization so that a producer's N rate can be adjusted accordingly, is one of the key objectives of N research (Bassanino et al., 2007; Fowler et al., 2013). In developing a standard measure of nitrogen mineralization potential (N_0), Stanford and Smith (1972) followed by Curtin and Campbell (2008) confirmed that substantial contributions of greater than 200 kg N/ha were indeed possible in many agricultural soils. However, quantifying N_{min} in soil at a particular location is only one aspect of the goal. As a biologically mediated process, there is more to consider when attempting to extrapolate these measures both temporally, and spatially. Being affected by climate, management practices, soil organic matter (OM) quality, etc., quantifying N_{min} temporally and spatially has required a multitude of studies to try and understand these relationships (Dessureault-Rompre et al., 2015; Goovaerts and Chiang, 1993; Heuvelink and Webster, 2001). As such, at the producer level, there has been no clear approach as to how an estimate of N_{min} could be applied throughout their fields and into their N rate recommendations. Further confusion arose with debates centered on soil carbon (C) storage and cycling, and how these might affect our understanding of N pool dynamics (Derrien et al., 2023; Janzen, 2019; Kleber and Lehmann, 2019; Olk et al., 2019). While much work has been done to allay these concerns, until a clear framework is proposed, a crop producer is more likely to forego the benefits of N_{min}, and leave these estimates out of their N rate recommendations altogether. Thus, with this hesitancy in

producers, decades of research can often be frustrated by a failure to provide accessible measures of soil N_{min}.

The thesis of this study is that the means for making reliable estimates of N_{min} currently exist. However, they require commentary, qualification, and development in order to be incorporated by practitioners into their N rate fertilizer recommendations. As such, and using an approach similar to Derrien et al. (2023) who addressed controversies related to soil carbon, the purpose of this chapter is to review the state of knowledge associated with the principal steps towards a practical system for incorporating N_{min} estimates into fertilizer recommendations, identify possible impediments, and suggest applications for the purpose of overcoming producer hesitancy. In order to explore these challenges, predictive functions (Section 2.2), N pool indices (Section 2.3), and field scale approaches (Section 2.4) will be considered. A chapter summary including research gaps is given in Section 2.5, and a statement of objectives and outline of the thesis are in Section 2.6.

2.2 PREDICTIVE FUNCTIONS

With N mineralization as a process occurring over time and under the influence of climate, one of the initial decisions the practitioner must make is the choice of predictive function. The standard function used in Curtin and Campbell (2008) is shown (Eq. 2.1) in order to demonstrate the principal components of a an estimate of N_{min} including, the predictive function itself (in this case, depicting a first-order kinetic model), potentially mineralizable N (N_0) here representing a single compartment N pool estimate, and the kinetic rate coefficient (k), reflecting the role of soil climate over a given time (t).

$$N_{min} = N_0\left[1 - e^{(-kt)}\right] \quad [2.1]$$

Benbi and Richter (2002) and Manzoni and Porporato (2009) outline multiple kinetic models with varying levels of complexity. These complexities are inherent with biologically driven processes in general, but are also due to the various aspects of temporal predictions. Each of the aspects of predictive functions including, the number of measurable N pools to consider, the means of explaining the breakdown relationship in the associated kinetic function (between a zero-, first-order, or mixed kinetic approach), the rate of breakdown understood in the rate coefficients (k), and the temporal scale (t) will be examined.

2.2.1 N Pool Divisions

An N pool is a theoretical or measurable division of the total and/or available organic-N in OM (Sharifi et al., 2007a; von Lützow et al., 2007). Characterized by the related breakdown kinetics (Section 2.2.2) and the method of extraction (Section 2.3), the decision on the number of N pools for estimating N_{min} is not only a function of what is biologically available, but in a practical sense, what is measurable. Generally, considering one kinetically uniform N pool to approximate N_{min} in soil (Eq. 2.1) was standard; but with the understanding that additional pools could be delineated based on incubation time and/or soil conditions (Curtin and Campbell, 2008; Stanford and Smith, 1972). Alternatively, Heumann et al. (2011) and Dessureault-Rompre et al. (2015) were among those who characterized organic N into two major divisions including a slowly (relatively stable) and quickly (relatively labile) mineralizing pool. While multiple theoretical divisions are possible, a total of three N pools were defined by Sharifi et al. (2007a) including Pool I (gross nitrogen mineralization in the first two-week incubation period representing the labile N pool), Pool II (gross nitrogen mineralization between the

two- and 24-week incubation period), and Pool III, which was derived based on curve-fitting as a theoretical pool extrapolated beyond the incubation period. The larger and more stable Pools II and III taken together, are used to estimate the variable N_0 (Sharifi et al., 2007a; Sharifi et al., 2007c). In some cases, additional pools emerged as statistical curve fitting approaches evolved. The goal was to increase the "goodness of fit", regardless of whether the pools could be measured or not. With divisions of one, two, three plus pools available in order to account for mineralizable N, the decision on the optimum relationship and their interactions provides practical complications.

Gillis and Price (2016) in comparing model scenarios for carbon mineralization from soil amended with organic residues found that a one-pool approach, with the addition of a logistic function to explain intrinsic competition dynamics, were significantly higher than two-pool models. Alternatively, and citing reasons of pool size and availability differences of organic-N content, a two-pool model wherein a larger and more slowly mineralizing stable pool, and a smaller and more quickly mineralizing labile pool was found superior to the one pool approach (Benbi and Richter, 2002; Cabrera and Kissel, 1988; Dessureault-Rompre et al., 2015; Heumann et al., 2013; Heumann et al., 2011). With respect to including a third pool (Pool III), its consideration as an extrapolated pool is important conceptually, but not practically.

The literature is relatively silent with applied examples of using a third N pool on its own in a model to quantify N_{min} in soil. As a theoretical pool, it provides an estimate of what could be potentially mineralizable beyond the 24-week incubation but was not measurable (Sharifi et al., 2008; Sharifi et al., 2007a). Practically speaking, the inclusion of pools and parameters that are measurable is of the utmost importance. Using a logistic

function or theoretical pool, not being measurable *per se*, renders it difficult to use in practice – especially if these aspects could be considered implicitly through other models or indices. Pool II and III, which considered together as the larger and more slowly mineralizing "stable" pool, has either been measured solely with Total N (TN) analysis, (Dessureault-Rompre et al., 2015; Mallory and Griffin, 2007), or with TN and additional parameters such as clay content (Heumann et al., 2013). Pool I as the more labile N pool, determined and derived from the first two weeks of aerobic incubation, is based on measurable indices relating to biological activity of the soil community, and is more feasible for practical use (Curtin and Campbell, 2008; Sharifi et al., 2008; Sharifi et al., 2007a).

Thus, using a model with parameters that are measurable and easy to parameterize is one of the main considerations in overcoming the obstacle of N pool divisions (Benbi and Richter, 2002). Also, while using two N pools is considered essential for estimating N_{min} in agricultural soils (Heumann et al., 2011), the data that is available to a practitioner can, and should, dictate the method of incorporation; namely, that if a bulk N measure of N_0 via TN is all that is available, than a one-pool approach should be attempted, rather than foregoing the advantage estimating N_{min} altogether. In practical circumstances, the selection of pools should be a decision based primarily on what is representative of the target field conditions (what is happening on the ground), and how a soils breakdown kinetics are understood.

2.2.2 N pool Kinetics

Estimating N_{min} for incorporation into N fertilizer recommendations requires that a kinetic relationship be applied to describe the mineralization process for the N pool(s)

selected. Kinetic models attempt to capture intrinsic and extrinsic soil dynamics affecting mineralization of N pools over time, which is influenced by such factors as soil properties, temperature, moisture, soil management, crop residues, organic amendments, etc. (Griffin et al., 2007). Adding to this complexity, the literature has provided many scenarios of single compartment (e.g., one model), dual-compartment (e.g., two models of the same order), or mixed-compartment (e.g., two models of a different order) scenarios (Benbi and Richter, 2002). Typically, the kinetic models used for estimating N_{min} are either a first-order model, zero-order model, or combination thereof. In first-order kinetics, the breakdown rate is linearly proportional to the concentration of the depleting substrate being degraded (i.e., OM), whereas in a zero-order model the breakdown relationship is independent of the substrate concentration. In higher order kinetic models, the relationship between rate and substrate concentration is non-linear. Applying a single, dual, or mixed-compartment arrangement, and deciding on the kinetic model(s) to include therein, is another point of confusion for the practitioner.

Stanford and Smith (1972) used a single compartment (first-order) model to describe net mineralization over a 30-week incubation period (Eq. 2.1). This approach has been adopted at various incubation intervals (e.g., 24-weeks) and in multiple field and laboratory studies as the optimum model for determining N_{min} (Curtin and Campbell, 2008; Heumann et al., 2013; Sharifi et al., 2007c). However, based on the understanding of soil organic carbon (SOC) cycling and the varying degrees of availability and recalcitrance (Lehmann and Kleber, 2015), the case for implementing two compartments, a quickly and a more slowly mineralizing N pool, is widely recognized over a wide range of cropping systems and seasonal variations (Benbi and Richter, 2002; Bonde and

Rosswall, 1987). The next question to emerge then is, if these pools should be described separately, do they breakdown according to similar or different kinetics? As such, a dual- or mixed-compartment kinetic description becomes necessary.

Two first-order equations (i.e., dual-compartment scenario assuming the fast and the slow pools both follow a first-order breakdown) were used under field conditions by Cabrera and Kissel (1988) and Heumann et al. (2013) with mixed results. Both studies saw deviations in predictive accuracy with some of the error being attributed to methodology (in particular, to wetting or sieving the soils for incubation). Their findings saw that fallow fields, where N mineralization is accelerated due to tillage practices, had lower average errors ranging from ~ 3.5 to 193% as compared with cropped plots where the average error ranged from 114 to 343% being overpredicted (Cabrera and Kissel, 1988). For both studies, it is conjectured that while first-order kinetics are representative of the smaller labile N pool, which is potentially consumed within a growing season, the larger slowly mineralizing pool, which may continue to mineralize over an entire growing season, is not being properly represented.

Dessureault-Rompre et al. (2013) showed that the slowly mineralizable stable organic N pool followed a zero-order, non-diminishing, relationship due to N cycling replenishment during the growing season. In this same study, multiple models (single-, dual- and mixed-compartment scenarios) including first-order alone, first- plus first- and zero- plus first-order combinations were examined. In that study, and in Dessureault-Rompre et al. (2015), results showed that a zero- plus first-order relationship, parameterized using TN for the stable, and Pool I (or N Flush via the 2-week aerobic incubation test) for the labile, respectively, best predicted the mineralizable N pools in

agricultural soils using regression analysis from soil properties (R^2 = 0.41 @ 15 cm). Since these parameters for the stable and labile fractions might be difficult to obtain in practice, when this optimal scenario is not achievable, a one pool model can generate usable estimations and is preferable to no estimate at all.

2.2.3 Mineralization Rate Coefficients

The mineralization rate coefficient (k) reflects the combined effects of the soil biological community and soil climate on the rate of N mineralization. Whether the rate coefficient depicts that the substrate is being diminished (-k) or is being replaced (k), it is an integral and potentially confounding aspect of estimating N_{min}. Estimating the rate of -kt as with N_0 (Eq. 2.1) is determined by regression technique and influenced by incubation time (t), temperature, and moisture (Curtin and Campbell, 2008). In addition, k is also associated with substrate (OM) quality and pool size, and must be calibrated (adjusted) to reflect local environmental factors (Benbi and Richter, 2002). For soil N mineralization applications, k is typically reported in units of per week (w^{-1}) or per day (d^{-1}) with increasing numbers depicting higher decay rates. Since deriving soil specific rates is not practical for the practitioner, reliance on published data is necessary, but should be qualified accordingly.

Sharifi et al. (2008) established mineralization rate coefficients using the first-order procedure described by Curtin and Campbell (2008) at 25°C for 24 weeks from field trials with various tillage practices. This was also completed across four geographic and climatic regions including British Columbia (BC), Saskatchewan (SK), Ontario (ON) and Quebec (QC). Comparing conventional tillage practices for consistency, k values ranged from 0.0093 d^{-1} (reported as 0.065 w^{-1} with N_0 = 170 kg N/ha) to a low of 0.0044

d^{-1} (reported as 0.031 w^{-1} with N_0 = 206 kg N/ha) which, highlights the variations between climatic regions (Sharifi et al., 2008). Heumann et al. (2013) also used first-order relationships, but implemented a dual-compartment approach (accounting for two pools) and generated specific k values for each pool with k_{fast} for the quickly (labile) and k_{slow} for the larger and more slowly (stable) mineralizable organic N pools. Dessureault-Rompre et al. (2015), in both dual- and mixed-compartment approaches, reported the differences between the specific k values for the labile (k_L) and stable (k_S) pools from a variety of fields and geographic locations across New Brunswick, Canada. Considering the first-order k value alone at a rate of 0.0084 d^{-1} (with N_0 = 124 mg N/kg), compared to the derived rate constants of k_S (0.492 d^{-1}) and k_L (0.074 d^{-1}), we see the need to account separately for the slow and fast pools, respectively. This difference in k value for both the labile and stable pools was also found by Bonde et al. (1988) and Benbi and Richter (2002) with the difference in range (from high to low) seemingly attributed to differences in soil management and location. As such, it is strongly encouraged by various authors to correct/calibrate k values for temperature, and or soil water content (Cabrera and Kissel, 1988; Dessureault-Rompre et al., 2015; Heumann et al., 2011). As an example, Dessureault-Rompre et al. (2015) corrected the k value using a biophysical water function on a daily time step (Dessureault-Rompre et al., 2011; Georgallas et al., 2012).

The variations observed in k values between climatic zones, management zones, and between single-, dual-, or mixed-compartment approaches is perhaps the most challenging aspect of predicting N_{min} from an applied perspective. Where possible, it is first recommended that k values derived from single compartment studies should only be

used if applied to make single-compartment predictions. For example, if a practitioner has a surrogate measure of N_0, the k value used in the single compartment (first-order) formula should have been derived from a single-compartment study. However, if the preferred approach of a mixed-compartment (zero- plus first order) relationship is put into practice, then a similarly derived k_L coefficient for the labile pool (e.g., aerobic incubation test) and a different k_S coefficient for the stable pool (e.g., total N) should be used, respectively. In addition, for both single-, dual-, or mixed-compartment functions, k values from similar climatic regions and/or management types should be sought. For example in humid climates and for both single- and multi-compartment scenarios, the results for k values taken from Dessureault-Rompre et al. (2015) are preferred since this is where, and how they were derived. Sharifi et al. (2008) has results from various locations across Canada, but for use in single compartment (zero-order) relationships; as such, they should only be used accordingly. Lastly, in applying these functions, a reasonable estimate of the temporal scale (t) is also recommended.

2.2.4 Considerations for the Temporal Scale

With respect to the temporal scale, the literature infers estimates of N_{min} can be made with any type compartment model, but with little reference to preferred durations. With k values per day or per week, the duration seems to be at the discretion of the practitioner and thus offers some points for consideration. An estimate of N per day or per week gives little help to a practitioner since N recommendations are usually made with the entire growing season in mind. Growing season length can be defined as the duration of time between the first and last occurrences of critical temperatures that occur when average soil temperatures (in the top 0.5m) are above biologic zero, which is 5°C

(Brinkmann, 1979). In terms of estimating N mineralization, biological activity and its duration is a key consideration. In Canada, the mean growing season in agriculturally prominent ecozones are approximately 130 d in the Atlantic Maritime, 110 d in the Boreal Shield, 120 d in the Prairies, and 100 d in the Boreal Plain (Gordon and Bootsma, 1993; Pedlar et al., 2015).

Using average growing season days specific to the practitioners ecozone, where the soil temperature is above biologic zero (5°C), is recommended for estimating N_{min} in order to account for biological activity in soil. Secondly, inclusion of the complete growing season accounts for the steady introduction of N into the agricultural system, and generates values more easily included into fertilizer recommendations (kg N/ha vs. mg N/kg). It should also be noted that considering a longer duration (such as a growing season), will result in larger estimates of N_{min} coming from the non-diminishing stable pool (k_S) since the average labile pool size (using k_L= 0.074 d^{-1}) can be consumed within the first 80 days (inferred from Soil Health Database in Chapter 3 using the average biological nitrogen availability (BNA) value of 41.5 mg N/kg). Finally, yearly fluctuations in growing season duration due to cold or wet springs for example, require careful consideration prior to incorporation. For example, if spring applications of N fertilizer must be delayed due to cold or wet conditions, modification of t may be altered accordingly. For potato crops, Dessureault-Rompre et al. (2013) noted that May to August delineates the most activity for both soil mineralization and plant N uptake. As such, a 90-day "effective" growing season might be considered or a 30% reduction from a 130-day estimate. Alteration of time (t), depending on producer application practices

(e.g., split N applications), may assist the practitioner in making final allotments of N_{min} estimates into their fertilizer recommendations.

2.3 COMMON INDICES FOR QUANTIFYING N POOLS

For the purpose of estimating N_{min} via the predictive function over a growing season, parameterization of the necessary rates of decomposition is required. If the preferred model is a mixed compartment function, (zero- plus first-order), measures of the stable pool (N_S) and labile pool (N_L) are appropriate. If a single-compartment function is used (first-order alone), then a bulk-measure of N_0 is required for considering both Pool II and III together. Besides the predictive function, a measured value for the kinetic pools is the next major point of consideration and of possible confusion for the practitioner. Much of the confusion arises as to the predictive power, and efficacy of various methods for estimating the size of plant available N pools. There is much divergence in the literature as to the best method, not to mention the appropriate pool-specific test among them. As we know, N mineralization itself is a catabolic reaction and conversion of microbial N (i.e., microbial decomposition of organic N) to inorganic ammonium – NH_4^+ (Paul, 2015). Being a microbially driven process, what any N test hopes to accomplish is to quantify the potential amount of organic N that can be converted into inorganic (mineral) N over a given period time. In turn, this will make soil N supply predictions possible and N management decisions more informed. The principal approaches, including biological and chemical methods, will be discussed along with their effectiveness in estimating respective pool sizes.

2.3.1 Biological Methods

It is important to reiterate that the N test is a measure of potential N mineralization, and not the actual N mineralized in the field during a particular growing season. While a lab-based measurement, at optimal conditions and using *ex-situ* soil can never precisely predict edaphic, climatic, or geographic conditions, or how much N will be mineralized or reach a crop at a given moment, the *Mineralizable Nitrogen* method is considered the industry standard for estimating N mineralization potential (St Luce et al., 2011). The aerobic incubation method (Stanford and Smith, 1972), the updated *Mineralizable Nitrogen* method (Curtin and Campbell, 2008) and its applications in practice (Dessureault-Rompre et al., 2015; Sharifi et al., 2007a; Sharifi et al., 2007c) involve a soil sample being mixed with sand and placed in leaching tubes where it is wetted, and the initial (inherent, or residual) leached mineral N is discarded prior to incubation. The leaching tubes with wetted soil are placed in an incubator at a set temperature (e.g., 25°C) where it is periodically aerated. The sample is leached with a $CaCl_2$ and N free nutrient solution every two weeks for the first 12 weeks, and then every following four weeks for the desired remaining incubation time > 20 weeks (e.g., 24-, 28-, or 32-weeks). The leachate, or mineral-N enriched leaching solution, is collected, and analyzed for concentrations of ammonium (NH_4^+) and nitrate (NO_3^-) using colorimetry techniques. With respect to limitations, aerobic incubation does not measure growing season mineralization, but mineralization potential, as this method is conducted under optimal conditions (e.g., no rainfall, controlled temperature, etc.). Also, being a laboratory method, *ex-situ* soil handling, drying/sieving, rewetting, and set climatic conditions are all factors that can influence the results.

As such, the experimental design is the criterion for delineating N pools, much like certain extraction methods define our understanding of humic substances (Janzen, 2019; Lehmann and Kleber, 2015). With aerobic incubation methods, important definitions and distinctions are made; namely, the N flush from the first two weeks of incubation is considered the labile organic N pool, and has been termed Pool I (Sharifi et al., 2007a). Further, Pool I is excluded from the estimation of N_0, which is defined as a bulk measure of potentially mineralizable N and includes Pool II (an intermediate pool) and the extrapolated more recalcitrant Pool III (Dessureault-Rompre et al., 2015; Dessureault-Rompre et al., 2013). In this understanding then, N_0 is representative of a comparatively slow mineralizing pool which is relatively stable and takes greater than 2-weeks to mineralize (Dessureault-Rompré et al., 2010). In humid temperate soils and in terms of estimating N_{min} over a growing season, this method provides the major N pool indices with the 2-week N flush (Pool I) as the preferred index for N_L to be included in the first-order relationship and N_0 (Pool II and II) for as the index for N_S to be included in the zero-order relationship.

However, while the biological method is the preferred approach, and with the goal of overcoming impediments with indices that are accessible, due to its completion time and involvement, it may be prohibitive at commercial scales and costly for the producer. Considering both the labile and stable parameters, the two-week N flush estimating Pool I (N_L) is more practical, but long-term incubations (>20 weeks) to capture N_0 or Pool II and III (N_S) requires surrogate measures. Total soil N (TN) as a surrogate, was most strongly correlated to N_0 ($R^2 = 0.40$), N_0 + Pool I ($R^2 = 0.39$), and Pool II alone ($R^2 = 0.40$) and is of practical value as it can be determined simultaneously with total soil organic Carbon

(TC) via dry combustion techniques (Dessureault-Rompré et al., 2010). Thus, within mixed-compartment functions, soil TN results have been used as the index for N_S in the zero-order compartment, and the 2-week (Pool I) N flush as the index for N_L in the first-order compartment (Dessureault-Rompre et al., 2015). Soil TN was also found by Heumann et al. (2011) to be significantly correlated to the stable pool N_S (0.54) using long-term incubations (~ 28 weeks) for validation. This correlation was also used successfully in pedotransfer functions using soil TN as a surrogate measure for N_S in estimating nitrate leaching in agricultural fields (Heumann et al., 2013). TN has shown poor correlation in field trials where organic N leaching was prevalent, but explained 91% of the variation in gross N mineralization in comparison to aerobic incubation methods (Wang et al., 2001).

Much of the literature cited deals with the comparison of various methodologies and the interpretation of their respective outcomes (i.e., incubation times, temperatures, etc., yielding less, or more inorganic N); consequently, a focus on these finer points may distract from the overarching conclusion that sizeable quantities of mineralized N are possible. For example, the “best methodology” may estimate 190 kg N/ha, while a “good methodology” from the same soil may estimate 110 kg N/ha. This may be a significant difference, but to the producer, the fact is that >100 kg N/ha is potentially available during a growing season. As such, while biological methods are the standard by which N_{min} potential can be measured, opportunities for alternative measures that will assist the practitioner should be pursued. Surrogate measures such as chemical methods (Section 2.3.2), or pedotransfer functions (Section 2.4.1) that translate data you have into data you

need, provide some of the tools to help utilize this potential, and inform N fertilizer recommendations.

2.3.2 Chemical Methods

The studies around which chemical method is the best surrogate measure for aerobic incubation, and under which circumstances, soil types, or environmental conditions, has itself filled tomes. Gerard Ros examined the subject at length regarding chemical extraction methods for estimating N_{min} and how methodologies influence concentration, for the purpose of informing and improving N fertilizer management (Ros, 2011). It was found that chemically extractable organic N (EON) from 20 available extraction techniques provided the significant ability to quantify N_{min} potential, but showed differences in pool size and composition due to variations in duration or temperature of extraction (Ros et al., 2011a; Ros et al., 2009; Ros et al., 2011b). Overall, these studies concluded that chemically EON was a reliable measure of the stable N pool (N_0), but attention to methodology, spatial, temporal, and cropping implications, was highly recommended (Ros, 2011; Ros, 2012). In order to provide clarity and resolution to practitioners, as well as assistance on a chemical index for predicting growing season N, we will focus on three fundamental questions, namely, which index will give a reasonable estimate of plant available N pools (N_S/N_0, or N_L) for usage in a chosen kinetic function, which index is available commercially, and what N pool does it represent.

The practicality of a surrogate for biological methods cannot be contested. Early on and in the context of their *Mineralizable Nitrogen* method, Curtin and Campbell (2008) noted the relative ease and speed with which chemical methods could estimate

N_{min} capacity. However, they also noted issues of an inability to factor in immobilization and, extract chemically what N would be considered biologically accessible. With aerobic incubation still required as the reference (St Luce et al., 2011), chemical methods attempt to estimate the component of OM that is biologically available by adding electrolyte solution to a given soil sample, water or acids, then incubated at a given temperature (e.g. 100°C) for a set time (e.g. 4 hours) in order to extract organic N from OM as NH_4-N (Ros, 2011).

The practitioner must pay close attention to which N pool is being identified via specific chemical methods. Recalling that N_0 is primarily made up of the larger and more slowly mineralizable stable pool (Pools II and III) or N_S, St Luce et al. (2011) reviewed multiple studies comparing extraction methods to estimate N_0. The best three methods included hot KCl extractable N (HKC; $R^2 = 0.78$), ultraviolet (UV) absorbance of $NaHCO_3$ extract at 260 nm ($R^2 = 0.74$), and direct steam distillation of phosphate-borate buffer (pH 11.2) extract ($R^2 = 0.73$) (St Luce et al., 2011). In another review, Ros et al. (2011b) found EON (comparable to N_0) was best predicted with hot $CaCl_2$, acid $KMnO_4$, acid $K_2Cr_2O_7$, hot water or hot KCl ($57\% < R^2 < 74\%$). In these manuscripts, the results show a conflict as to the optimum: is it hot KCL, or hot $CaCl_2$?

Looking at standalone studies, Schomberg et al. (2009) identified TN ($R^2 = 0.64$), hot KCl ($R^2 = 0.62$), and Hydrolysable N ($R^2 = 0.60$) among the best three predictors of N_0 in soils under different management systems in the southern United States. Nyiraneza et al. (2012), comparing both biological and chemical indices against field-based soil N supply (plant bioassay) as estimated by corn N uptake and corrected for starter N fertilizer, also found UV absorbance of $NaHCO_3$ extract at 205 nm ($r = 0.41$), hot KCl

extractable NH_4-N (r = 0.39), and Pool I plus the concentration of NO_3-N extracted using $CaCl_2$ solution prior to incubation (r = 0.44) as promising surrogates of soil N supply (interestingly and perhaps a testimony to the need for inclusion of Pool I, since N_0 alone was not as strongly correlated to soil N supply). With few studies focusing on surrogates for the more labile pool, Sharifi et al. (2007a) looked at soils across multiple different climatic conditions and management practices and concluded that $CaCl_2$ extractable N and hot KCL extractable (NH_4 and NO_3), while good measures of Pool II and III, were also highly correlated to Pool I. As is seen with all these studies, there is divergence, but also underlying agreement.

For the purpose of incorporating N_{min} estimates into fertilizer recommendations, what must be avoided is a tendency to allow the pursuit of the ideal method hinder its application. In particular, the study by Nyiraneza et al. (2012) confirms and proposes two main points that are important for moving forward amidst uncertainty. Firstly, that in order to correct for lab-based biases, using plant uptake in unfertilized crops is a suitable reference standard (Nyiraneza et al., 2012; Zebarth et al., 2005). This method would be ideal and is potentially reproducible by the practitioner at their own field level using test strips of zero, or reduced N applications for comparison. Secondly, and confirming Ros (2011), the measure of N_0 may not be an adequate sole estimator due to the poor agreement with soil N supply (Nyiraneza et al., 2012). Another takeaway, is that while there might be differences as to the "best" index, there is underlying agreement that all extraction methods perform well and, at the very least, provide some indication of mineralized N contribution.

From a practical perspective, the "best test" may be the one that is available. From an informal survey of agricultural soil labs, hot KCl appeared as the chemical method most commonly available. But unavailability of the "optimum" test should not hinder a practitioner from utilizing some estimation of N_{min} in their respective soil. The caveat, is that calibration is required to qualify chemical method results, and if this calibration is not performed by the lab itself, it is recommended that local field supply measurements (e.g., test strips of zero- or reduced-N), where possible, become standard practice.

2.4 TOOLS FOR INCORPORATING N_{MIN} ESTIMATES

Having addressed quantifying estimates of N pools within a soil sample, and using that measure to estimate N mineralization over time through predictive functions, the next step for the practitioner is scaling these specific point estimates across a landscape. As such, overcoming challenges and impediments to quantifying N_{min} may also involve the use of pedometric tools and support systems. A common finding in the literature for making estimates of N_{min}, has been the need to incorporate texture, moisture, soil management, climate, or other intrinsic and extrinsic factors that mediate soil N mineralization processes (Derrien et al., 2023; Dessureault-Rompre et al., 2015; Heumann et al., 2013; Nyiraneza et al., 2012; Zebarth et al., 2009). These factors step beyond the standard predictive function or laboratory measures, and attempt to account for the stochastic variability of N_{min} predictions, which can be confounding factors for mechanistic or empirical approaches. Besides this difficulty of accounting for outside controls to mineralization in the soil data, there is the added impediment of having no direct soil measures. From the perspective of practitioners attempting to incorporate N_{min} predictions on their farms, three main challenges exist, namely accounting for a lack of

data or for controls of N_{min}: at a given point (non-spatial), across their respective field (spatial), and utilizing those predictions to develop fertilizer recommendations. In response, this section will explore the potential for pedotransfer functions (PTFs), digital soil mapping (DSM), and the application of soil-based N credits.

2.4.1 Pedotransfer Functions for Estimating N_{min}

PTFs are a means of quantifying, via an empirical and/or qualitative function, what the relationship is between a parameter of interest (or response variable) and selected predictor variables. PTFs are derived by modelling the relationships between predictor and response variables using statistical techniques that range from simple regression to machine learning (McBratney et al., 2002; McBratney et al., 2000). Feature selection is typically performed by first addressing multicollinearity amongst predictor variables through variance inflation factor (VIF) analysis (Marquaridt, 1970), followed by recursive feature elimination (RFE), in order to identify the most important predictors for a simplified, parsimonious predictive function (Perreault et al., 2022; Román Dobarco et al., 2019b). What this means in practice, is that once the relationship between response and predictor variables are established, and obtained with the fewest and most relevant predictors, a practitioner can then use the *data they have to get the data they need* (Bouma, 1989). Thus, for the purpose of estimating growing season N_{min}, PTFs can then be used to establish the relationship between variables such as Pool I, N_0, etc., and available parameters in order to make predictions. As a tool, PTFs may be helpful to replace the need for lengthy incubation methods, understand what controls contribute to N_{min}, and/or overcome data gaps whenever direct measures are absent.

PTFs have been used for understanding the intrinsic controls in N mineralization for decades. Rasiah (1995) focused on the concept that N_0 was influenced by more than oxidized N released during incubation analysis, and compared PTFs using texture, SOC, TN, and cation exchange capacity (CEC) in order to predict variables in one and two-pool models. In his study, it was observed that the concentration of the labile pool increased with higher TN and pH and was reduced with higher clay contents (as opposed to the stable pool, which increased with higher clay content). As such, intrinsic properties such as the physical protection, afforded by the mineral component, were seen to have significant influence on N_0. Other studies have shown that PTFs can be derived successfully for measures of the stable pool (R^2 as high as 0.64) and the labile pool (R^2 as high as 0.42) using parameters such as clay content, temperature, humus class etc. (Glendining et al., 2011; Heumann et al., 2003; Heumann et al., 2011). Laurence et al. (2023) successfully derived PTFs for TN as an estimate of the stable pool (concordance = 0.80) with parameters such as texture, OM, soil respiration; BNA as an estimate of the labile pool (concordance = 0.78) using aggregate stability, active carbon, soil respiration, and TN; and an estimate of (130-day) growing season N_{min} (concordance = 0.82) using aggregate stability, active carbon, soil respiration, OM, and pH (Chapter 3).

Although PTFs provide the possibility of reliable predictions, some cautionary findings should be considered. Since PTFs are usually empirical in character, the intrinsic mechanisms influencing N_{min} can be inferred, but are not expressly evident; as such, with important drivers potentially not being accounted for, using PTFs outside their geomorphological origins can be problematic. McBratney et al. (2002) recommended that PTFs are best suited for locally derived applications, so for the practitioner, while a

PTF might be available, it should be used with caution if it was developed outside their source geographic region. In addition, PTFs derived with complicated sets of parameters should be avoided since their reproduction or usefulness is less likely. For example, PTFs designed for larger scale or global applications can include predictors such as latitude, mean C:N ratio of the soil classification group, etc., and may be difficult to populate (Glendining et al., 2011). It should also be noted that PTFs derived using certain machine learning models (e.g., cubist decision tree), sometimes referred to as "black box" models (Molnar et al., 2018), do not produce regression equations that are transferable outside the seed database. Where possible, researchers deriving PTFs using machine learning should include multiple-linear regression analysis in order to provide regression model coefficients for practical use applications (Chapter 3).

For the practitioner, the PTF is a useful tool for estimating N_{min} at a given point; however, extrapolation from that point, spatially across a field or region, is difficult since the controls may not be understood at each spatial location. Scale plays an important factor, and utilizing soil predictors (intrinsic properties) as did Dessureault-Rompre et al. (2015) may be sufficient to estimate N_{min} locally (proximal scale); however, extrapolation across a landscape may require more climate or geographic related inputs to increase accuracy (distal scales). Notably, PTFs that are derived from non-georeferenced data points will typically not include geospatial and/or climatic predictor's and, as a result, may reduce the potential to incorporate extrinsic controls. This is where DSM techniques are useful in order to fill in spatial gaps, extrapolate via continuous coverage of climatic/geographic information, and provide predictions where point data is not available (thus alleviating another impediment).

2.4.2 Digital Soil Maps for Estimating N_{min}

In order to adjust N fertilizer rates across a landscape, the prediction of N_{min} must be quantified spatially (Simard et al., 2001; Zebarth et al., 2009). DSM techniques are an appropriate tool to answer this need both regionally, and infield. DSM is a process by which soil sample points, obtained at a known location, are extrapolated spatially using predictive models. With these models, point data is used in conjunction with environmental covariate data layers to learn the relationships, make predictions in pixels or polygons across the landscape, and "fill in the gaps" where direct soil information is lacking (McBratney et al., 2003). Just as with PTF development, DSMs predict for a response variable of interest (e.g., Pool I, or N_0), but with the exception that the predictor variables are spatial layers that provide a means to take advantage of non-spatial (point) information over a broader range. The DSM process, including the use of machine learning, has been outlined extensively by McBratney et al. (2003), Minasny and McBratney (2016) and Heung et al. (2016). With respect to our goal of predicting growing season N mineralization across a landscape, a key element of DSMs is that environmental covariates spatially quantify the soil forming factors (Jenny, 1941) including climate (*c*), organisms (*o*), relief (*r*), parent material (*p*) and age/time (*a*), plus soil properties (*s*), and spatial position (*n*), in what is collectively known as the *scorpan* model (McBratney et al., 2003). By incorporating *scorpan*, and with spatial variation inherent in biologically driven processes such as N_{min}, DSM techniques lend themselves well to predicting N_{min} across a landscape.

Examples in the literature do include DSMs for Total N (Mponela et al., 2020; Uygur et al., 2010; Wang et al., 2013; Wang et al., 2018; Wang et al., 2017; Zhang et al.,

2019; Zhou et al., 2019; Zhou et al., 2020), but spatial predictions of biologically labile N (Pool I) are rare. Based on these regional studies, TN was best predicted with vegetation reflectance indices (e.g., normalized difference vegetation index, NDVI) from Sentinel 2 satellite data, especially band-3 – red (0.63 to 0.69 μm) as well as topographic variables, followed by climatic variables such as mean-annual precipitation and temperature (Wang et al., 2018). In addition to climatic variables, Zhou et al. (2020) found that separating for land-use types within climatic zones also had a strong prediction influence. With respect to maps based on aerobic incubation methods, the literature is relatively silent; however, infield mapping of N_0 has been done using predictors such as clay content and OM grouped according to yield, response curves, and residual soil N (Simard et al., 2001). While the infield map was useful for delineating spatial variations of N_0, the issue of extrapolation beyond field boundaries is impossible unless this data is available. Thus, if the aim of the DSM is to provide predictions for practitioners on a broader scale, regional studies should be investigated in the future.

The findings above are corroborated by Robertson and Groffman (2015) who note the differences between distal and proximal scales in addressing the controls of N mineralization. Distally, or regionally, climate, disturbance (soil management), soil type, and plant community structure are noted as the primary controls; whereas, proximally (field scale) the controls relate more to plant uptake (crop yield), moisture, structure, and CEC. Regionally, this is promising for mapping N_{min} because the covariates required are publicly available via satellite imagery; locally, this provides a practical insight for the practitioner in that proximal data collection across the field (e.g., drone imagery at a finer scale) is valuable for incorporating infield variability estimations.

2.4.3 Application of Soil Based N credits

The adoption, or implementation of N_{min} estimates is where research results can be applied in practice. With respect to the 4R nutrient stewardship system (right: source, rate, time, place), a practical approach to fertilizer management used by practitioners, estimates of N_{min} align with the determination of the right "rate" (Johnston and Bruulsema, 2014). Zebarth et al. (2009) discussed various strategies for predicting fertilizer N rates. These strategies attempt to consider the various N inputs that can be applied as credits to a standard fertilizer prescription in order to minimize overapplication, and potential losses to the environment. In order to determine a recommended fertilizer rate (F_N) as in Eq. 2.2, N credit systems first apply a general crop N requirement (R) based on field N response trials for standard crop types. Next, contributions from various N sources are subtracted from R which may include: a NH_4^+ credit from manure or compost (M_{AMM}), organic N credits from manure or compost (M_{ORG}), credits for a previous year's crop growth such as legumes (C), and finally soil based credits from OM content (S) via N mineralization (Zebarth et al., 2008).

$$F_N = R - M_{AMM} - M_{ORG} - C - S \quad [2.2]$$

N credit systems may take on different scales such as a global approach (Gu et al., 2021), but a standard feature is the need for reliable estimates of each input parameter. For the scope of this chapter, factors R and S will be considered.

Estimates of R are typically a tabularized base value of what a particular crop type, or varieties N requirement (output) would be for a growing season (kg N/ha). Being a generic value, there is not always reference to how this output requirement might differ

based on higher efficiency N sources, or differences in crop yield. Besides a generic value for R, there is a possibility to derive the crop requirement using a yield goal approach (Machet et al., 2017; Oglesby et al., 2023; Tamagno et al., 2022; Zebarth et al., 2009). With this approach, the producer chooses a target yield, and based on crop N uptake/removal estimates, calculates the total crop N requirement (R). The primary limitations of this approach relate to actual vs. predicted growing season, as well as setting a realistic target yield (Tamagno et al., 2022; Zebarth et al., 2009). Oglesby et al. (2023) found that basing calculations on unrealistic yield goals alone can often over-estimate N recommendations above what would be the agronomic optimum N rate (the maximum N rate to maximize yield). Studies suggest that considering the economic optimum N rate (i.e., the maximum N rate without monetary loss), coupled to the delta yield (i.e., optimum N rate minus the control yield without N fertilizer) provide more realistic yield estimates (Lory and Scharf, 2003). With the appropriate R value selected, N credits can then be applied.

With respect to the S credit, and the focus of this review, the inherent difficulty of these estimates may lead to an overly cautious approach in applying them. Prince Edward Islands (PEI) N credit system for example, applies a generic credit of 15 kg N/ha if the OM level is above 3.5% and a zero credit if OM is less than 3.5% (Zebarth et al., 2008). It is notable that in PEI, the mean OM levels ranged from 2.7% to 3.6% (Laurence et al., 2023; Nyiraneza et al., 2017). With most soils falling below the threshold value of 3.5%, soil-based credits (S) are effectively *nil*. As a result, soil based contributions, which have been found to average 106 kg N/ha across trials using a plant bioassay approach in potato crops (Nyiraneza et al., 2022), are being overlooked. Greater attention to the site-specific

estimates of S via N_{min} potential over a growing season are therefore highly recommended (Priesack et al., 2006).

The implementation of N_{min} estimates into recommended fertilizer rates (F_N), and standard practices for doing so, is an area in need of development. Overcoming the impediments and areas of confusion in applying N_{min} estimates opens an opportunity for decision support systems, or N credit systems, based on newly available technologies.

2.5 SUMMARY AND RESEARCH GAPS

The two N pool, zero- plus first-order mixed-compartment, predictive function was found to be the optimum device for recognizing the more slowly mineralizing, plus the more quickly mineralizing, compartments of organic N in OM. Mineralization rate coefficients (k), as an influential component of the predictive function, must be implemented within the same models from which they were derived (using first-order rate coefficients within first-order equations, for example), under similar climatic regions where possible, and applied for logical durations (t) such as a growing season. Conflicting opinion as to the optimum function(s) are warranted, as many show strong correlation and promise in making N_{min} predictions under various circumstances. While the above function is recommended, the practitioner should use prudential judgement, and opt to apply what is most suitable for their specific soil/climatic conditions.

Regarding indices for the desired N pool, it is suggested that aerobic-incubation, as a biological index, is preferred for determining the smaller, more labile, and quickly mineralizing N pool (Pool I). Hot KCl appeared as the most appropriate chemical extraction technique based on evidence in studies and consumer availability at commercial labs. In order to quantify what N pool is being represented via chemical

extraction techniques, calibration with biological indices or local field trials was recommended. Soil TN, a dry combustion technique, is recommended based on multiple studies for estimating the stable N pool.

In situations where direct soil data is unavailable for use, or in order to understand the controls related to N_{min}, the PTF was identified as a useful tool in overcoming these challenges. Applicable for predictions at a given soil sample location (i.e., non-spatial predictions), PTFs for soil TN (estimating the stable N pool) were more common in the literature showing a strong correlation with related properties such as OM. Studies that derived PTFs for estimating the labile pool (Pool I) are rare and are recommended. Also recommended are spatial applications (DSM techniques), which have the potential to incorporate climate, management, and other extrinsic controls. Digital maps have been produced primarily for TN but there was no evidence of a DSM for the aerobic incubation test (e.g., Pool I) or estimate of N_{min} over a growing season at a regional scale. DSMs are useful for identifying controls for N_{min}, and supplying predictions for N parameters in areas without direct soil information. Estimates from DSTs can then be incorporated into existing N credit systems.

Ultimately, and in order to overcome producer hesitancy with soil-based N credits, the field of N research must build on available predictive functions and indices, and transition these methods into applications via support tools such as PTFs and DSMs. In the following chapters of this thesis, a cohesive approach is developed wherein a practitioner can use indices for the stable and labile N pools in predictive functions, PTFs, and/or DSMs for informing N credit systems and soil-based estimates of growing season N mineralization.

2.6 THESIS OBJECTIVES AND OUTLINE

From research gaps identified in the literature review, the thesis objective is to build from the mixed-compartment (zero- plus first-order) predictive function from Dessureault-Rompre et al. (2015) to develop DSTs of N parameters for improving N management in PEI. Specific objectives included: 1) Developing a framework and PTFs for N parameters using surrogate data for circumstances where producers do not have direct soil measures of the stable (TN) or labile (Pool I) N pools for use in the predictive function; 2) the development of DSMs for spatial estimates of N parameters where direct soil information is not available; and 3) application and assessment of soil-based DSTs (PTFs and DSMs) of N parameters at the field-scale to build a system of use that can complement existing N management systems in PEI.

Chapter 3 addressed the problem of transforming the data a producer has, into the data a producer needs (Bouma, 1989), via a *cost-effective framework for estimating soil nitrogen pools using pedotransfer functions and machine learning.* Next, to address situations where field specific data does not exist, Chapter 4 considered *integrating multi-year crop inventories as a proxy for soil management practices within a digital soil mapping framework for predicting nitrogen indices* at the provincial scale. Resulting from PTF and DSM development, intrinsic and extrinsic factors influencing N dynamics were also explored. Chapter 5 focused on *applying provincially derived pedotransfer function and spatial estimates of nitrogen indices at the field scale* in order to examine how novel DSTs might be interpreted, and implemented. The conclusion (Chapter 6), synthesized the findings and proposed practical suggestions for amending the existing N credit system, and supplying estimates of growing season N_{min} to producers.

CHAPTER 3: TOWARDS A COST-EFFECTIVE FRAMEWORK FOR ESTIMATING SOIL NITROGEN POOLS USING PEDOTRANSFER FUNCTIONS AND MACHINE LEARNING [1]

3.1 ABSTRACT

Globally, the strategic use of nitrogen (N) is important in optimizing economic returns and reducing soil nitrogen losses to the environment. Incorporating reliable estimates of nitrogen (N) mineralized over a growing season (GSN) into N fertilizer rate recommendations is critical, but may often lack a direct measurement. For this purpose, Pedotransfer functions (PTFs) of total nitrogen (TN) – representing the stable pool from which N is mineralized, and biological nitrogen availability (BNA) – representing the labile pool of N mineralization, were used to estimate GSN. GSN was calculated based on TN and BNA results from a soil health database (SHD), which also includes a suite of related soil health parameters (n = 2,222). Using a process of recursive feature elimination (RFE) and cost-benefit feature elimination (CBFE), the best predictors of TN, BNA, and GSN were identified using a suite of machine learners (MLs) and regression analysis. For TN, RFE revealed that BNA, active carbon (AC), sand (Sa), and soil organic matter (OM) were the best predictors yielding a Lin's concordance correlation coefficient (CCC) of 0.80 and a reduction in theoretical cost of 41% compared to the control. CBFE resulted in AC, soil respiration (SR), clay, Sa, and OM as the most cost-

[1] Chapter 3 is a version of a manuscript that was submitted to Geoderma (open source) on May 31, 2023, revised on September 13th, 2023, accepted on October 16, 2023 and available online on November 20, 2023. This publication has multiple authors, in which the concept, design, data processing, and writing was done by the PhD Candidate with the assistance of all co-authors. The publication can be obtained using the following citation:

Laurence, L., Heung, B., Strom, H., Stiles, K. and Burton, D., 2023. Towards a cost-effective framework for estimating soil nitrogen pools using pedotransfer functions and machine learning. Geoderma, 440, p.116692. https://doi.org/10.1016/j.geoderma.2023.116692

effective predictors of TN with a CCC of 0.79 and a theoretical cost savings 49% below the cost of using all appropriate soil health parameters in the SHD. With respect to BNA, the best predictors from RFE were aggregate stability (AS), AC, SR, and TN with a CCC of 0.78 and a theoretical cost reduction of 23%. CBFE retained AC, SR, S, TN, OM, and pH as predictors of BNA with a CCC of 0.78 and reduction of 29% in theoretical cost. Finally, GSN results from RFE identified AS, AC, SR, OM, and pH as the best predictors with a 0.82 CCC and 17% reduction in theoretical cost. CBFE, on the other hand, identified AC, SR, sand, OM, and pH as the most cost-efficient predictors while maintaining a CCC of 0.82 and theoretical cost reduction of 29%. Of the MLs used for pattern recognition (i.e., cubist, random forest, support vector machine, and stochastic gradient boosting), cubist model outperformed the others for the majority of iterations of the RFE and CBFE processes. The cost-effective framework, and the N related PTFs developed in this study will greatly enhance our ability to predict soil N pool dynamics and the ability to incorporate GSN estimates into N fertilizer recommendations for producers worldwide. Improvements in predictive strength could be achieved by incorporating climate and soil management practices into PTF development. Another area for improvement and future study would include addition of spatial and landscape variability related to N measures via digital soil mapping applications.

3.2 INTRODUCTION

The global need to manage anthropogenic nitrogen (N) inputs to provide plant available N while minimizing nitrogen losses and residual soil nitrogen (RSN) remaining after harvest has reached a critical point (Chataut et al., 2023; Golden et al., 2023). With nitrogen use efficiencies often <50% in major crops (Tamagno et al., 2022), the transfer

of nitrate (NO_3^-) to ground- and surface-water systems or its transformation into ammonia (NH_3) or nitrous oxide (N_2O) emissions to the atmosphere necessitate the need for concrete solutions. With a view for optimizing the global N cycle and managing the N budget (Fowler et al., 2013; Heumann et al., 2013), practical strategies aimed at N management are emerging; including, sustainable nutrient policies, digital crop nutrition solutions, nutrient recovery and recycling, climate-smart fertilizers, and accelerated innovation (Dobermann et al., 2022).

The fertilizer industry is promoting 4R Nutrient Stewardship (4R) approach, wherein nutrients supplied from the right source, at the right rate, time and place to support more sustainable agricultural practices (Bruulsema et al., 2016). A crucial element of the 4R approach, as with earlier examples of conducting N balances (Bassanino et al., 2007; Stanford, 1973), is the quantification of all nutrient sources in determining 'right rate' fertilizer applications. This method is used to calculate fertilizer input rates (e.g., kg N/ha), and attempts to estimate the N demand for a given crop, relative to the N supply available from the soil (Frerichs et al., 2022; Morvan et al., 2022). The former is based on plant uptake requirement as a function of estimated crop yield, while the latter is the sum of all N sources including synthetic fertilizer, organic amendments, RSN, crop residues, and the N produced through soil N mineralization—a biologically mediated process. Of the variables required to complete an N balance correctly, an estimate for soil N mineralization (N_{min}), is perhaps the most crucial as it can account for approximately 50% of plant N uptake (Valenzuela, 2023). Zebarth et al. (2009) recommends a balance sheet or N budget approach as a solution to N management

and emphasized the need for reliable and practical methods to estimate N_{min} over the growing season.

In response to this need, and for the purpose of generating soil N_{min} estimates in a usable format for fertilizer recommendations (i.e., kg N/ha), here we develop a pedotransfer function (PTF) to support predictions of soil N_{min} over a growing season. Building on early aerobic incubation work from Stanford and Smith (1972), and using soils sampled from New Brunswick arable cropping systems, Dessureault-Rompre et al. (2015) found that a two pool (i.e., zero- plus first-order) regression equation best described soil N_{min} and resulted in R^2 values ranging from 0.41-0.49. Accounting for a stable, non-diminishing, zero-order N pool plus a labile, diminishing, first-order N pool, the regression equation formed the basis of a prediction function that can be used to estimate soil N_{min} supply (Dessureault-Rompre et al., 2012, 2015; Sharifi et al., 2007b). Applying this prediction function over a growing season requires a set timeframe (t) for prediction purposes (Zebarth et al., 2009). In Atlantic Canada, 130 days represents the duration of a standard growing season. In addition to a set timeframe, calculating the cumulative N concentration of soil N_{min} over a 130-day growing season also requires direct soil measures for both the stable and the labile N pools. In previous studies, the stable pool has been represented by total nitrogen (TN) analysis, and the labile pool via 2-week soil incubation methods, such as biological nitrogen availability (BNA) analysis (Dessureault-Rompre et al., 2015; Sharifi et al., 2007a; Heumann et al., 2003, 2011; Rasiah, 1995).

While the ability to make 130-day growing season N (GSN) predictions is possible with prediction functions, their implementation is hindered when direct measures for TN

(stable N pool) and BNA (labile N pool) are not available. Under current methodologies, TN is a relatively simple measurement to make and available in select databases; however, it is not always included in standard soil analytical suites at commercial laboratories. The BNA test (Pool I or N Flush test), determined by 2-week aerobic incubation and extraction of mineral-N generated, is based on the mineralizable nitrogen method (Curtin and Campbell, 2008). This measure of N mineralization potential can then be contextualized to reflect soil climate considerations (Dessureault-Rompre et al., 2015; Sharifi et al., 2007b) resulting in an estimate of GSN. BNA is a relatively novel test that is time consuming, costly, and, as such, absent from most datasets. As a result, when direct soil measures of TN and BNA are not feasible or unavailable, the opportunity to make GSN estimates is greatly reduced; and with it, the ability to use an estimate of GSN to inform N fertilizer recommendations, optimize N use, and minimize N losses.

In circumstances where data is difficult to obtain and/or non-existent, PTFs provide extremely useful approximations. In the absence of direct measures, PTFs are a means of obtaining an estimate for a soil parameter by quantifying the relationships between the parameter of interest (i.e., response or dependent variable) and other pedologically related parameters (i.e., predictor or independent variables; Bouma, 1989). In practice, this means using available data in order to predict costly, or unavailable data (McBratney et al., 2000, 2002). Once determined, the learned relationship can be used to make predictions in datasets where the response variable is absent. In the soil science literature, PTFs have been applied to mineralogical or soil hydraulic parameters (response variables), such as bulk density or available water holding capacity—estimates that are

time consuming to measure, and therefore difficult to obtain (Benites et al., 2007; Glendining et al., 2011; Román Dobarco et al., 2019b; Van Looy et al., 2017). Traditional approaches use linear regression, but machine learning (ML) techniques have been increasingly used (Amanabadi et al., 2019; Benke et al., 2020; Cisty and Cyprich, 2020; Heung et al., 2016; Khlosi et al., 2016; Schillaci et al., 2021; Wang et al., 2019; Xiao et al., 2022). There is good reason to believe that predicting biological properties like soil respiration and N mineralization will be highly correlated with fundamental soil properties. Soil respiration is the result of the soil biological community metabolizing soil organic matter, as influenced by the soil environment (e.g., aeration, water content, pH, clay content). The mineralization of N is one of the outcomes of soil biological activity and is influenced by the ratio of carbon to nitrogen in the soil organic matter. Aggregate stability is an outcome of the combined influence of soil biological activity and the mineral composition of the soil. Thus, it is reasonable to use ML to determine the extent these factors are correlated with these more difficult to measure parameters, and whether they can be used to develop a PTF.

To develop a PTF, the primary considerations include both a decision on the response variable, and a decision on the predictors needed (Van Looy et al., 2017). Firstly, choosing the desired output parameter is decided both by need and by practicality. While the requirement for a PTF is predicated by need, and becomes somewhat self-evident, the concept of 'practicality' is more nuanced. Practicality includes factors, such as time, cost of analysis, and the principal that the effort in obtaining predictors should not exceed the effort in obtaining the response variable itself (McBratney et al., 2002). The second consideration, that of selecting soil parameters for

making predictions, takes into account the correlation, availability, time/cost of analysis, and the total number of predictors — all of which have implications on computational demands and model complexity (Pachepsky and Rawls, 2004). The underlying goal of a PTF is that it will be used; and as such, one that is accurate, achieved with the fewest, and most cost-effective set of predictors would be considered the optimum. Also, while a direct relationship may not exist between response and predictor variables, a complimentary aspect of PTFs is their ability to provide insight into intrinsic soil relationships (McBratney et al., 2002; Van Looy et al., 2017). With respect to TN, BNA, and GSN, wherein physical, chemical, and biological relationships are difficult to interpret, methods of developing PTFs with machine learning may provide a unique opportunity to increase our understanding of soil N pool dynamics and associated factors.

Practicality in a PTF is of high importance and relates primarily to achieving a parsimonious model with as few predictors as possible (McBratney et al., 2002). As an added benefit, a parsimonious model reduces the potential for model overfitting (Shahabi et al., 2022). While multiple parameters may be available to develop a PTF, the likelihood of it being used elsewhere, and for other datasets, is increased with fewer predictor variables. This process of feature selection, or eliminating predictor variables, is begun by addressing multicollinearity by means of variable inflation factor (VIF) analysis (Perreault et al., 2022; Román Dobarco et al., 2019b; Xiao et al., 2022). Secondly, to identify and remove irrelevant predictors, a process of recursive feature elimination (RFE) is conducted (Miranda et al., 2022; Xiao et al., 2022). Once the correlated and irrelevant predictors have been removed, the remaining variables, via the best performing pattern recognition method (e.g., ML), were included in the final PTF. It is notable that

the analytical cost and/or practicality of obtaining independent variables were not considered in VIF or RFE procedures. It may occur that some of the final predictors were, in whole or in part, more expensive or time consuming to produce than the predictor variable itself. PTFs must not only include *data we have* (Bouma, 1989), but also result in predictors that are available, practical, and cost effective.

The obstacles that hinder our ability to calculate and predict soil N_{min} over a growing season (GSN) are inextricably tied to the difficulty in obtaining TN and BNA as input parameters. The absence or difficulty of obtaining one, or both, of these required inputs impair the calculation of GSN. Thus, the development of a PTF becomes necessary to make these estimates accessible. Previously developed PTFs for stable, and labile N pool size can be found; Rasiah (1995), Glendining et al. (2011), and Heumann et al. (2003, 2011) for example, but '*commonsensically*' speaking, locally applied PTFs derived from geomorphically similar data are preferred (McBratney et al., 2002). While TN is now a relatively easy parameter to analyze, the fact that it is seldom available in datasets makes it a suitable candidate for a new PTF. BNA is also suited for PTF development since the analysis is time consuming, labor intensive, and rarely found in most datasets. Finally, in situations where neither TN or BNA data are available, a PTF for the calculated GSN output itself should be considered. PTFs are commonly developed for directly measurable soil parameters only; however, because of the pressing need for GSN estimates, the opportunity to test the possibility does exist.

In this study, to improve our understanding of soil N pool dynamics and for the purpose of making GSN estimates available to inform "right rate" N fertilizer recommendations, a framework for predicting TN, BNA and GSN using ML and

regression techniques is proposed. Specific objectives will include identifying what are the most important predictors of TN, BNA and GSN using (i) RFE, and (ii) identifying what are the most cost-effective predictors and arguments using a Cost-Benefit feature elimination (CBFE) approach. This study used data acquired from Prince Edward Island (PEI), Canada, to calibrate and validate the PTFs as a case study.

3.3 MATERIALS & METHODS

The methodological framework used in this study is shown and summarized in Figure 3.1. Summary statistics and modelling activities were performed with version 4.2.0 of the R statistical software (R Core Team, 2018).

3.3.1 Study Area

The study area includes soils collected throughout the province of PEI with a total land area of approximately 5,620 km^2, an undulating relief with the majority (~75%) of the land base between 45 m above mean sea level, and a maximum elevation of 139 m above mean sea level (MacDougall et al., 1988). With a cool, humid climate, the growing season extends between May and October with a frost-free period of 100 to 160 days, a mean temperature range of -7°C (January) to 19°C (July), and an annual precipitation, ranging from 900 to 1000 mm per year, including snowfall (MacDougall et al., 1988). Podzolic soils dominate the landscape followed by Luvisols, Brunisols and Gleysols, which were developed on medium to coarse textured glacial till derived predominantly from sandstone bedrock (MacDougall et al., 1988). Agricultural production consists mainly in potatoes, cereals, and legumes, with grain, pasture and forage production also supporting mixed operations, dairy and/or non-ruminant livestock production (MacDougall et al., 1988).

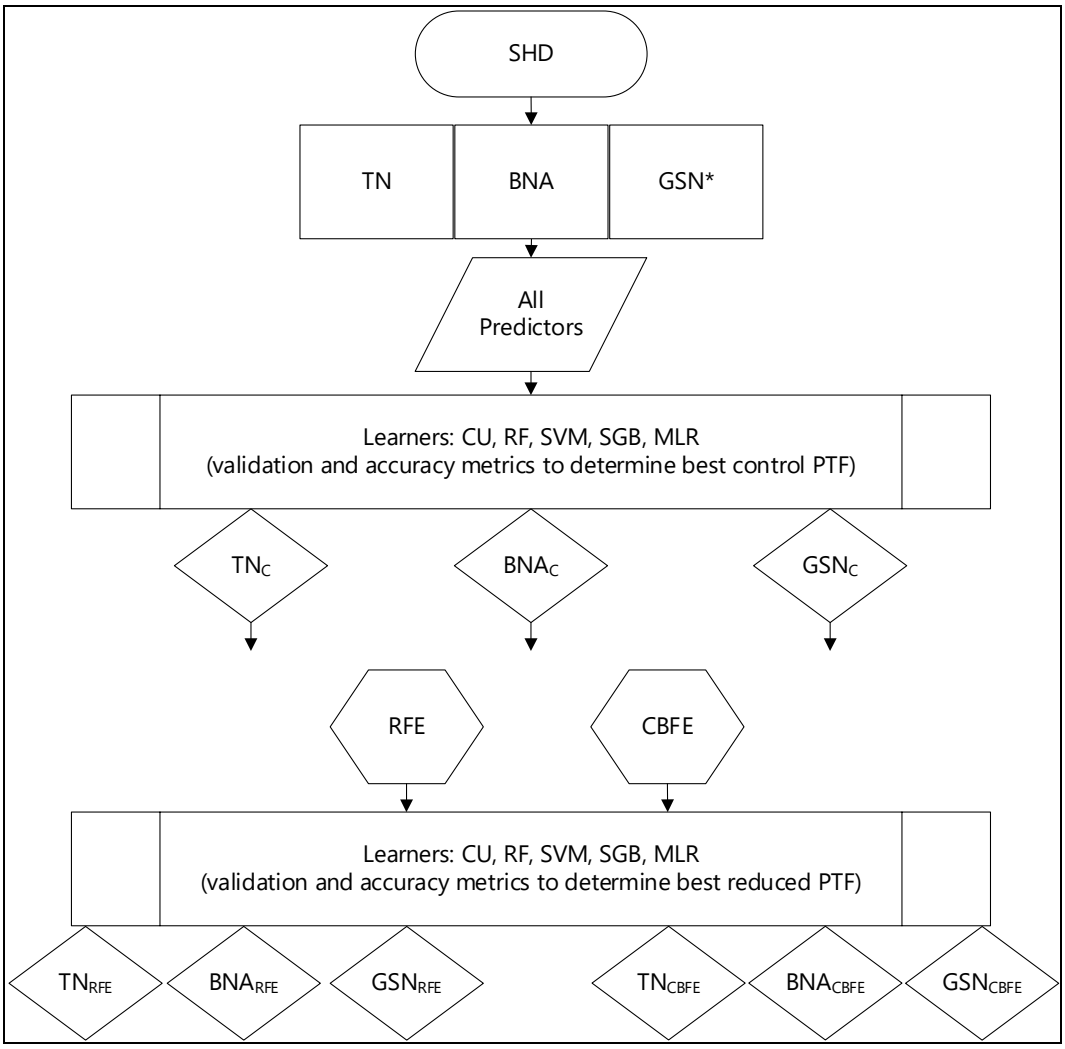


Figure 3.1 The methodological framework used in this study. From the soil health database (SHD), total nitrogen (TN), biological nitrogen availability (BNA), and growing season nitrogen (GSN*: calculated value) was selected as the dependent variables and then applied to predictor variables in respective conceptual models. Control PTFs for TN (TN_C), BNA (BNA_C) and GSN (GSN_C) were then developed by comparison of four machine learners (CU = cubist, RF = random forest, SVM = support vector machine, SGB = stochastic gradient boosting) and regression analysis (MLR = multiple linear regression) using validation techniques and accuracy metrics. Using recursive feature elimination (RFE), and cost benefit feature elimination (CBFE) incorporated with learner comparison and validation and accuracy metrics, reduced PTFs were developed for TN ($TN_{RFE\ \&\ CBFE}$), BNA ($BNA_{RFE\ \&\ CBFE}$) and GSN ($GSN_{RFE\ \&\ CBFE}$).

3.3.2 Soil Health Database

The Soil Health Database (SHD) used for this study consisted of agricultural soil samples (n = 2,222) collected throughout the province of Prince Edward Island (PEI).

PEI producers selected sample locations based on their respective need as part of a provincially subsidized soil testing program. As a voluntary program, sampling density varied with the density of agricultural land. Soil samples were collected between the years of 2019 and 2022 by landowners, or their designates, and submitted to PEI Analytical Laboratory (PEIAL) for analysis. Soil samples were collected either from 0 to 18 or 20 cm or presumed rooting depth using PEIALs Standard Operating Procedures for collection and preservation (PEI Analytical Laboratories, 2019).

The database includes soil parameters that were considered critical "soil health indicators" as described in Cornell's *Comprehensive Assessment of Soil Health* (CASH) framework (Moebius-Clune, 2016) and, as such, were appropriate and relevant for this study. PEIALs SHD, including results from their "Soil Health Test" analytical suite, is organized in terms of: biologically related parameters sensitive to management including aggregate stability, biological nitrogen availability, permanganate oxidizable carbon (or active carbon), soil respiration, total nitrogen, and soil organic matter calculated from a measure of total carbon; physical parameters that are not changed by management including particle size fractions (i.e., sand, silt, and clay percentages); and standard chemical parameters, such as pH, soil macro- and micro-nutrients. It should be noted that soil nutrients influenced by annual fertilization are unstable predictors and thus were not considered. Table 3.1 includes a summary of the parameters used in this study with the inclusion of the calculated output value for GSN (explained in Section 3.3.3 and based on TN and BNA results). PEIAL's procedures of soil analysis for the parameters in the SHD (Table 3.1) were taken from the CASH by Moebius-Clune (2016), and are also described in Marshall et al. (2021).

Table 3.1 Summary of soil health database parameters, sample size (*n*), and summary statistics including the minimum (Min) value, 1st (25%) quartile, Median, Mean, 3rd (75%) quartile and the maximum (Max) value.

	Units	*n* =	Min	1st	Median	Mean	3rd	Max
Aggregate Stability	%	2,219	5.10	25.1	37.1	41.5	53.5	96.5
Biological Nitrogen Availability	mg/kg	2,222	0.700	17.3	23.0	26.9	32.3	104
Permanganate Oxidizable Carbon	ug/g	2,158	121	347	410	421	482	799
Soil Respiration	mg/g	2,219	0.020	0.370	0.490	0.583	0.710	2.00
Total Nitrogen	%	2,180	0.0130	0.110	0.135	0.142	0.168	0.481
Organic Matter	%	2,038	0.500	2.20	2.60	2.71	3.10	6.50
Clay	%	2,182	1.40	10.3	11.7	11.8	13.3	35.7
Silt	%	2,182	1.00	26.8	29.8	30.7	33.8	76.0
Sand	%	2,182	2.90	53.5	58.5	57.4	62.3	81.3
pH	NA	2,037	4.40	5.80	6.00	6.04	6.30	7.40
Growing Season Nitrogen	Kg N/ha	2,180	42	102	125	136	157	448

Aggregate stability (AS) analysis followed the CASH framework using the method adapted from Schindelbeck et al. (2016). The AS method employs a rainfall simulator to assess the proportion of soil aggregates remaining after a simulated rainfall event. Soils were air-dried and aggregates were separated using 2 mm and 0.25 mm sieves. Dried soil aggregates were then placed on a 0.25 mm sieve and subject to rainfall simulation. AS is measured by percentage weight of stable soil aggregates (retained in the sieve after broken down aggregates and stones have been removed) divided by the initial (total) weight of the sample. The average AS reported in the SHD (*n* =2,219) showed that the mean amount of aggregate remaining was 42% (Table 3.1).

Biological nitrogen availability (BNA) as explained in Marshall et al. (2021) is performed using a two-week aerobic incubation adapted from Sharifi et al. (2007a). In this method, the mixed soil (50%) and inert Ottawa sand (50%) sample is prepared in a Buchner funnel and, after initial leaching of readily available mineral RSN (NH_4^+ and NO_3^-) with a 0.01 M $CaCl_2$ solution, the sand and soil mixture were incubated aerobically for 14 days and leached again using the same method. After the 14-day incubation, the

concentration of mineral N analyzed from the recovered leachate is measured as the BNA and used as an estimate of N mineralization potential from biological processes (Sharifi et al., 2007a). In the SHD, the mean BNA value (n = 2,222) was 26.9 mg N/kg soil (Table 3.1).

Permanganate oxidizable carbon (POX_C), utilizing colorimetric methods from the CASH framework, is referred to as active carbon (AC) by PEIAL and is adapted from Weil et al. (2003). In this process, soil is air-dried, sieved to 2 mm, mixed with a potassium permanganate solution, shaken, extracted, and then measured spectroscopically for POX_C (Marshall et al., 2021). POX_C is a measure of readily oxidizable organic matter (Moebius-Clune, 2016) and is an indicator of the labile soil N pool. In the SHD, POX_C analysis (n = 2,158) had a mean value of 420.9 ug C/g soil (Table 3.1).

Soil respiration (SR) measures CO_2 released after a four day aerobic incubation and is described in the CASH framework adapted from Zibilske (1994). In this procedure, a sample of soil is air-dried, sieved to 2 mm, weighed, and placed in a sealed jar along with a separate beaker filled with alkaline CO_2 trapping solution. The soil is slowly re-wetted and the jar is sealed and incubated for four days. Conductivity is measured against the original and known carbonate content trapping solution. The conductivity of the trapping solution declines in a linear relationship with increasing respiration (CO_2 adsorption) and is quantified by comparison with the conductivities of calibration solutions. The mean SR (n = 2,219) results were 0.6 mg CO_2/kg soil (Table 3.1).

Total nitrogen (TN) and soil organic matter (OM) are considered a multi-parameter analysis wherein both parameters are determined at the same time and with the same instrument. Analyses were done in accordance with the LECO Method Report: Plants and Soils 10cc Loop, 4/16/2019, CN 828 S/N:20014 combustion procedure. TN and total soil carbon (TC) were first determined by the combustion at 900 °C of previously oven dried and ground soil (Marshall et al., 2021). TN measures were recorded as-is and TC is converted to estimate OM, which was calculated by PEIAL using the conventional conversion formula of OM = TC * 1.72. While this conversion factor is of debatable origin and accuracy (Pribyl, 2010), OM is the reported value in the SHD and was therefore retained. The value average TN (n = 2,180) is 0.14% and the average OM (n = 2,038) is 2.7% (Table 3.1).

Soil texture analysis procedures (also a multi-parameter test) determining clay (Cl), silt (Si), and sand (Sa) percentages was taken from the CASH framework, which uses a "rapid texture" protocol adapted from Kettler et al. (2001). In this method soil is air-dried, sieved to 2 mm and, with the addition of a 3% sodium hexametaphosphate solution to act as a dispersant, is shaken for 2 hours and sieved again at 0.053 mm to remove Sa. The remaining soil in solution (Si and Cl) sits for a minimum of 2 hours and is decanted in order to remove Si so that the Sa and Si separates can be dried and weighed. The proportion of Cl is determined from the difference as calculated from total dry weight minus %Sa and %Si. In the SHD, the average (mean) proportions (n = 2,182) were 57% S, 31% Si, and 12% Cl (Table 3.1).

Soil pH (n = 2,037) used the 1:1 (soil:water) method found in PEIAL's Modified Laboratory Manual of Methods, Standards and Equipment (PEI Analytical Laboratories, 1996) and had a mean value of pH 6.

3.3.3 Growing Season N Mineralization Estimates

GSN estimates for PTF training were calculated values, based on direct soil data, using the two-pool regression equation (Eq. 3.1) cited in Dessureault-Rompre et al. (2015). Due to the relative climatic homogeneity of the study area, variation in soil climate was not considered.

$$N_{\min} = k_S t + N_L\left[1 - e^{(-k_L t)}\right] \quad [3.1]$$

In Eq. (3.1), cumulative N mineralization ($N_{\min}$) was estimated over a 130-day growing period (t) based on the sum of a non-depleting zero-order function describing the stable N pool and a first-order function describing the labile N pool. The stable fraction ($k_s t$) was calculated by estimating k_s (d^{-1}) as described in Dessureault-Rompré et al. (2015) using the following relationship (Eq. 3.2):

$$k_s = 0.123\ (TN) + 0.00312\ (BNA) + 0.0685 \quad [3.2]$$

where TN is in mg N/kg soil and BNA is in mg N/kg soil. The labile component $N_L(1 - e^{-k_L t})$ was estimated as a first order relationship where BNA was used as an estimate of N_L and a fixed value for k_L of 0.074 d^{-1} as suggested in Dessureault-Rompre et al. (2015) for sandy loam textured soils. As calculated from the TN and BNA results in the SHD, the mean GSN value (n = 2,180) was 136 kg N/ha.

3.3.4 Learner Approaches

Four ML approaches, including cubist, random forest, support vector machines with radial basis function, and stochastic gradient boosting were used in addition to multiple linear regression, and compared to determine the best learner for predicting TN, BNA, GSN. The *caret* package (Kuhn, 2020) and the *iml* package (Molnar et al., 2018) within the R statistical software (R Core Team, 2020) were used for all modeling procedures, and for interpretation, respectively.

Cubist (CU), a rule-based regression tree model (Kuhn and Johnson, 2013; Quinlan, 1992), is further described in Landré et al. (2018) and Deragon et al. (2023). The CU modelling approach, also called the model tree or M5, has been used for both PTF and digital soil mapping production (Deragon et al., 2023; Mello et al., 2022; Solomatine and Dulal, 2003; Xiao et al., 2022). CU consists of two hyperparameters that are required to optimize the model including the number of *committees* and *neighbors*. Committees are a boosting strategy to address over- and under-predictions; whereas, neighbors relate to the number of observations that are closest to the predicted values that will be averaged and compared (Deragon et al., 2023). For this study, a matrix of combinations was generated for the committees (i.e., 1, 10, 20, 30, 40, 50) and neighbors (i.e., 0, 2, 4, 6, 8) hyperparameters.

Random forest (RF) is a tree-based learner that utilizes, and trains a number of decision trees in order to generate predictions that have been compared and tested against an ensemble of uncorrelated trees (Breiman, 2001; Heung et al., 2016). RF has applications in PTF, predictive digital soil mapping, water resource, and spatial imaging applications (Deragon et al., 2023; Heung et al., 2016; Paul et al., 2022; Tyralis et al.,

2019). The main hyperparameter for optimizing the accuracy of predictions is m_{try} (Kuhn, 2020), which is the number of predictors that were randomly selected to be tested for each node-splitting rule. Given that m_{try} is dependent on n number of predictors available, the values tested for this hyperparameter were $m_{try} = 1, \ldots, n$.

Support vector machine (SVM), as discussed in Qin et al. (2022) is based on a statistical theory and method for reducing structural risk by optimizing the boundaries between classes, while enhancing generalization and prediction ability (Vapnik and Chapelle, 2000; Vapnik, 1999). The method has been used for PTF development and digital soil mapping applications relating to soil or land classification as well as soil moisture and related soil properties (Gill et al., 2006; Huang et al., 2002; Priori et al., 2014; Qin et al., 2022; Sedaghat et al., 2022). Further described in Boser et al. (1992); Hastie et al. (2009) and Heung et al. (2016), the kernel, or mathematical function, used in this study included the radial basis function (RBF) in the *caret* package using sigma (i.e., 0.0001, 0.001, 0.01, 0.1, 1) and cost (i.e., 0.1, 1, 10, 100, 1000) hyperparameters (Kovačević et al., 2010; Kuhn, 2020; Priori et al., 2014).

Stochastic gradient boosting (SGB), introduced by Freund and Schapire (1997) and modified by Friedman (2001, 2002), is an ensemble technique and adaptation of classification tree analysis. SGB is a mixture of blending and bagging procedures wherein the residuals from previous trees are built into smaller trees with each step of the boosting procedure (Lawrence et al., 2004; Rossel and Behrens, 2010). Perhaps due to its resilience to inaccurate training data, outliers, and a resistance to overfitting, SGB has been used in various PTF and digital soil mapping applications (Chen et al., 2018; Gebauer et al., 2020; Govil et al., 2022; Jalabert et al., 2010; Lawrence et al., 2004;

Szabó et al., 2019) For this study, parameter values included number of trees (i.e., n.trees = 500), interaction depth (i.e., 1, 3, 5, 7, 9), shrinkage (i.e., 0.1, 0.2, 0.3), and minimum terminal node size (i.e. 1, 5, 10, 15, 20).

Multiple linear regression (MLR), as described in Minasny et al. (1999) and Botula et al. (2015), is a regression technique that is common for analyzing relationships between dependent and independent variables for generating PTFs. In this multi-parameter method, a stepwise, forward, or backward selection technique removes, or adds variables in order to fit a straight line in dimensional space equal to the number (*n*) of predictors (Padarian et al., 2018). A stepwise approach to MLR was included, in addition to the suite of MLs, in order to obtain equation based PTFs that are more transferable than complex ML approaches. Based on preliminary examination of distributions of parameters in the SHD, using probability density functions, the data did not require transformation as in other related studies (Schillaci et al., 2021; Wösten et al., 1999).

3.3.5 Accuracy Assessment

Repeated 10-fold cross-validation, which measures the prediction error based on the predicted versus actual values within the dataset, was performed with 50 repeats to test each model's prediction accuracy (Ballabio et al., 2019). The repeats were used to ensure reliable and stable estimates of model accuracy.

Final predictions were assessed for accuracy using the coefficient of determination (R^2), Lin's concordance correlation coefficient (CCC), and the root mean square of error (RMSE) metrics (Donatelli et al., 2004; Román Dobarco et al., 2019b; Van Looy et al., 2017). CCC is a reproducibility index assessing both the closeness of

data to the line of best fit and the distance of the line of best fit from the 45° (1:1) line through the origin, and thereby accounting for systematic under- or over-prediction (Lin, 1989). With a range between 1 and -1, higher CCC values demonstrate higher correlation between the observed values and model prediction (Román Dobarco et al., 2019b).

To account for variance in prediction results, the 95% confidence interval (CI) was calculated using the standard deviation (SD) of the CCC. The 95% CI was selected in preference to a narrower (90%) CI to allow for a greater variance as is inherent with biological processes. The upper and lower bounds of the CI was taken as the prediction range, or margin, for feature elimination processes.

3.3.6 Feature Elimination

Feature (or parameter) elimination began by addressing multicollinearity and reviewing correlations (Figure 3.2) between predictors. The process for eliminating predictor variables was performed to obtain a parsimonious model while maintaining a similar model accuracy. Based on preliminary trials of model performance, PTFs tended to show the highest CCC's when all SHD were included as predictor variables. As such, the following approach was used to judge PTF success wherein a PTF including all SHD predictors was trialed with all MLs to identify the best scoring model (CCC) and thus the Control-PTF. Next, the 95% CI upper and lower CCC boundaries were determined for the Control-PTF, and finally, subsequent PTFs were considered successful if their respective CI's remained within the interval of the Control-PTF. The 95% CI upper and lower bounds of the Control-PTF was termed the Control-CI. Here, the upper and lower limits of the 95% CI were generated using the accuracy metrics from the 50 repeats of the 10-fold cross-validation procedure.

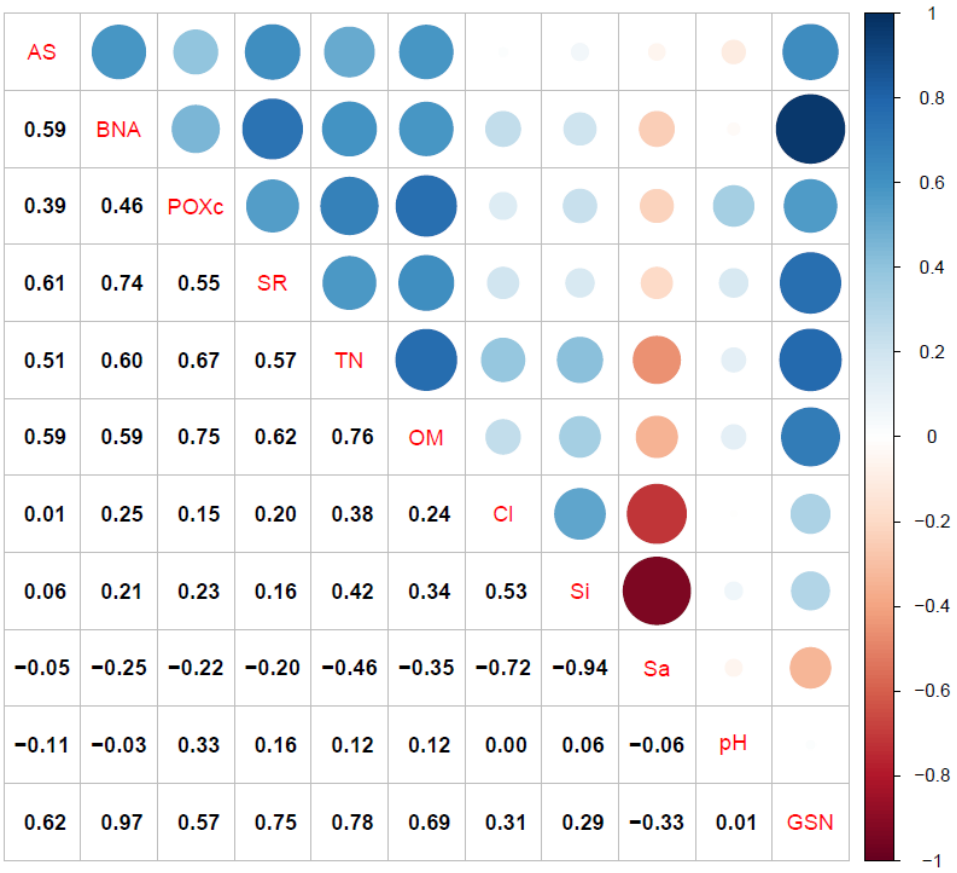


Figure 3.2 Correlations between soil health database variables (AS = aggregate stability, BNA = biological nitrogen availability, POX_C = permanganate oxidizable carbon, SR = soil respiration, TN = total nitrogen, OM = organic matter, Cl = clay, Si = silt, Sa = sand, pH) and the calculated growing season nitrogen (GSN) prediction (n = 2,222).

To incorporate cost into PTF selection, a total theoretical cost was determined for all control and final PTFs. Standard monetary rates for each individual parameter were obtained from PEIAL, ranked in order of cost, and secondarily, where one analysis renders multiple parameters (e.g., soil texture, TN/OM), in terms of processing time from high to low (Table 3.2). Standard rates for various analytical packages, or bulk pricing, was not considered in this study. Implementing these rationale, RFE and CBFE approaches were carried out. MLR results were reported for RFE and CBFE, regardless of performance, to provide model coefficients.

Table 3.2 Ranking and total cost (Canadian Dollars; CAD) of single parameter analysis and multi-parameter analysis in the soil health database.

Rank	Single Parameter Analysis	Multi-Parameter Analysis	Cost (CAD)
1	Aggregate Stability		$31.34
2	Biological Nitrogen Availability		$31.09
3	Permanganate Oxidizable Carbon		$23.32
4	Soil Respiration		$19.10
5		Clay	$18.26
6		Silt	
7		Sand	
8		Total Nitrogen	$9.50
9		Organic Matter	
10	pH		$6.00

3.3.6.1 Recursive feature elimination

The process of RFE, proposed by Guyon et al. (2002) is described in Xiao et al. (2022) as a backward feature elimination method for choosing the optimal and most relevant predictor variables. RFE works by: **Step 1**: including all predictors in a conceptual model (model) argument (Control PTF), **Step 2**: testing the model and determining the performance for each ML; **Step 3**: for each ML, variable importance metrics were generated and the least important predictor is removed; **Step 4:** Steps 2 and 3 were repeated iteratively until one predictor remains (Poggio et al., 2021; Xiao et al., 2022). The optimal PTF for each of the dependent variables (TN, BNA, GSN) was determined by comparing the results of all MLs tested, and choosing the best performing ML and model with the least predictors who's (95% CI) upper bound was within the Control-CI. This process was carried out in determining PTFs for TN, BNA, and GSN. It should be noted that TN and BNA were not included in the GSN model as these parameters were used in its derivation.

3.3.6.2 Cost-benefit feature elimination

Given the principle that PTFs should use 'easy' to get data to derive 'difficult' to get data, a cost-benefit approach (CBFE) was developed using sample cost as the metric for difficulty (i.e., effort of analysis). In theory, if costly soil parameters can be removed without sacrificing overall accuracy, then the opportunity exists to render an accurate PTF with an optimized theoretical cost. In terms of surrogate measures for N mineralization, the goal of CBFE would be obtaining a surrogate measure as cost-effectively as possible. As an example application, commercial laboratories may consider CBFE to determine which parameters may more easily predict difficult parameters such as aerobic incubation analysis (e.g., BNA), and package these into routine analytical suites. For this purpose, the following iterative and incremental CBFE process was developed and followed for each dependent variable (TN, BNA, and GSN): **Step 1**: To establish the Control-PTF, all available variables were ordered in sequence according to the ranking in Table 3.2, from high to low, where *x* is the dependent variable and n_1 to n_4 are the ranked independent variables:

$$\text{Model 1: } (x) = f(n_1, n_2, n_3, n_4, \dots)$$

For multi-parameter analyses, such as soil texture and TN/OM analysis, soil parameters were regarded individually for elimination purposes. **Step 2**: The Control model was tested with the MLs and the performance was determined based on CCC. **Step 3**: Taking the best CCC result of the Control Model, the upper and lower 95% CI was calculated. **Step 4**: Removing the most expensive parameter from Model 1 (i.e., AS), the simplified model (Model 2) was tested.

$$\text{Model 2: } (x) = f(n_2, n_3, n_4, \dots)$$

Step 5: Based on the accuracy metrics from Model 2, the best performing model was identified and observed to see if the bounds of the 95% CI were within the Control-CI. The following two scenarios would result from each iteration. In Scenario A, if the CI of the best model CCC dropped below the Control-CI, then parameter (n_1) was considered necessary and retained, and the next incremental parameter (n_2) was removed yielding an argument as follows:

$$\text{Scenario A: } (x) = f(n_3, n_4, \dots, n_1)$$

In Scenario B, if the CI of the best model CCC remained within the Control-CI, then parameter (n_1) was left out and the next incremental parameter (n_2) was removed yielding an argument as follows:

$$\text{Scenario B: } (x) = f(n_3, n_4, \dots).$$

Step 6: This process was continued until all parameters were tested. The resulting model was considered the optimum and most cost-effective PTF.

Using price per parameter from Table 3.2, the cost of the final PTF was calculated. For comparison purposes, costs were also calculated for Control-PTFs and RFE-PTFs. This process was carried out for TN, BNA and GSN.

3.4 RESULTS AND DISCUSSION

3.4.1 Control PTFs

The Control PTF results for TN, BNA and GSN (Figure 3.3) were not subject to RFE or CBFE as they were developed using all relevant parameters in the SHD. Notwithstanding, GSN is a calculated value derived from TN and BNA results and as such, these parameters were not included in the PTFs for GSN. In some cases, there are

concerns with multicollinearity among parameters (Menard, 1995); however, VIF analysis was conducted and no parameters were removed prior to RFE and CBFE procedures.

3.4.1.1 Total Nitrogen

The Control-PTF for TN (TN_C) yielded a range of 2% in CCC from 0.79 to 0.81 based on 1,984 observations and trial of five MLs (Figure 3.3, Table 3.3).

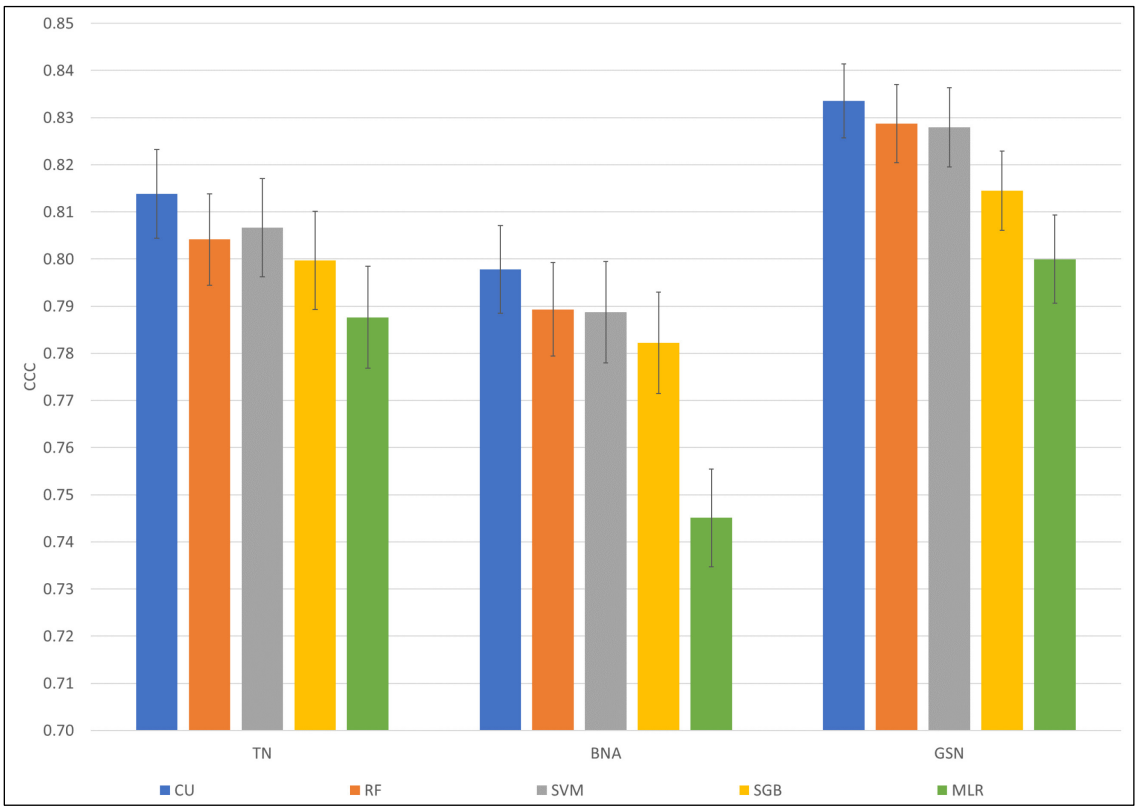


Figure 3.3 Chart of concordance (CCC) results of control pedotransfer functions for total nitrogen (TN), biological nitrogen availability (BNA), and growing season nitrogen (GSN) using all relevant predictor variables with cubist (CU), random forest (RF), support vector machine (SVM), stochastic gradient boosting (SGB), and multiple linear regression (MLR).

Table 3.3 Comparison of control, recursive feature elimination (RFE), and cost-benefit feature elimination (CBFE) results with the learner (cubist (CU); and multiple linear regression (MLR)), conceptual models (AS = aggregate stability, POX_C = permanganate oxidizable carbon, SR = soil respiration, OM = organic matter, Cl = clay, Si = silt, Sa = sand) with multi-parameter analysis shown in brackets, number of variables, sample number (*n*), coefficient of determination (R^2), root mean square error (RMSE), concordance (CCC), and the theoretical cost (Canadian dollars; CAD) for total nitrogen (TN), biological nitrogen availability (BNA) and growing season nitrogen (GSN).

Parameter	Method	Learner	Conceptual Models	Variables	*n* =	R^2	RMSE	CCC	SD of CCC	95% CI	Cost (CAD)
TN	Control	CU	AS + BNA + POX_C + SR + (Cl + Si + Sa) + (OM) + pH	9	1,984	0.686	0.0250	0.814	0.0339	0.00938	$139
	RFE	CU	BNA + POX_C + (Sa) + (OM)	4	1,986	0.665	0.0260	0.800	0.0361	0.0100	$82
	CBFE	CU	POX_C + SR + (Cl + Sa) + (OM)	5	1,985	0.655	0.0267	0.794	0.0402	0.0111	$70
	Control	MLR	AS + BNA + POX_C + SR + (Cl + Si + Sa) + (OM) + pH	9	1,984	0.659	0.0260	0.788	0.0369	0.0108	$139
	RFE	MLR	BNA + POX_C + (OM)	3	1,986	0.636	0.0270	0.771	0.0402	0.0111	$64
	CBFE	MLR	POX_C + (Sa) + (OM)	3	1,986	0.633	0.0275	0.769	0.0394	0.0109	$51
BNA	Control	CU	AS + POX_C + SR + (Cl + Si + Sa) + (TN + OM) + pH	9	1,984	0.657	7.98	0.798	0.0336	0.00930	$108
	RFE	CU	AS + POX_C + SR + (TN)	4	1,984	0.637	8.18	0.780	0.0330	0.00914	$83
	CBFE	CU	POX_C + SR + (Sa) + (TN + OM) + pH	6	1,984	0.639	8.14	0.781	0.0379	0.0105	$76
	Control	MLR	AS + POX_C + SR + (Cl + Si + Sa) + (TN + OM) + pH	9	1,984	0.602	8.51	0.745	0.0374	0.0104	$108
	RFE	MLR	AS + SR + (TN + OM)	4	1,996	0.581	8.79	0.727	0.0408	0.0113	$69
	CBFE	MLR	POX_C + SR + (TN) + pH	4	1,984	0.582	8.73	0.729	0.0398	0.0110	$58
GSN	Control	CU	AS + POX_C + SR + (Cl + Si + Sa) + (OM) + pH	8	1,984	0.715	24.9	0.834	0.0284	0.00788	$108
	RFE	CU	AS + POX_C + SR + (OM) + pH	5	1,984	0.692	25.9	0.818	0.0326	0.00903	$89
	CBFE	CU	POX_C + SR + (Sa) + (OM) + pH	5	1,984	0.690	26.0	0.817	0.0321	0.00890	$76
	Control	MLR	AS + POX_C + SR + (Cl + Si + Sa) + (OM) + pH	8	1,984	0.676	26.4	0.800	0.0337	0.00934	$108
	RFE	MLR	AS + POX_C + SR + (OM)	4	1,985	0.654	27.3	0.784	0.0325	0.00901	$83
	CBFE	MLR	SR + (Sa) + (OM) + pH	4	1,995	0.657	27.4	0.786	0.0325	0.00901	$53

Apart from the MLR and the CU models, there was no significant difference between ML performance. For TN_C, the top performing ML was the CU model with an R^2 of 0.69, RMSE of 0.025 and CCC of 0.81. For benchmarking purposes, the Control-CI was calculated from the SD of the CCC (Table 3.3). The upper and lower bounds of the Control-CI was 0.82 and 0.80 respectively. The CI for CU-TN_C was used to judge the success or failure of subsequent conceptual TN models that were trialed using both RFE and CBFE procedures. The theoretical cost of analysis for TN_C was $139 (Table 3.3). The MLR results were significantly different to CU-TN_C but within significant range of the remaining MLs (Figure 3.3). The CCC for MLR was 0.79, with a SD of CCC (0.039) and a 95% CI of 0.01 (Table 3.3). The CI was 0.80 (upper) and 0.78 (lower). This CI was exclusively used to evaluate subsequent MLR conceptual model results for RFE and CBFE. Coefficient results for the MLR method are presented in Table 3.4.

3.4.1.2 Biological Nitrogen Availability

For BNA, the Control-PTF (BNA_C) showed a drop in CCC compared to TN results (Figure 3.3, Table 3.3). The span in CCC between all MLs tested was 5%, ranging between 0.75 and 0.80 based on 1,984 observations. MLR performed significantly lower than all other MLs, while CU, RF, SVM and SGB showed no significant differences. The best performance, and selected as the control, was with the CU model. Accuracy metrics for CU-BNA_C were 0.66 for R^2, 8.0 for RMSE, and 0.80 for CCC. The Control-CI for CU-BNA_C was used for benchmarking throughout the RFE and CBFE processes. The theoretical cost of analysis was $108 (Table 3.3). MLR coefficient results for control parameters are given in Table 3.4. With a CCC of 0.75, the CI for the MLR-TN control was 0.010 (Table 3.3) and was used to evaluate RFE and CBFE for the MLR results only.

Table 3.4 Multiple linear regression coefficient results from control, recursive feature elimination (RFE), and cost-benefit feature elimination (CBFE) for total nitrogen (TN), biological nitrogen availability (BNA), and growing season nitrogen (GSN) and soil health database variables (AS = aggregate stability, POX_C = permanganate oxidizable carbon, SR = soil respiration, OM = organic matter, Cl = clay, Si = silt, Sa = sand) with multi-parameter analysis shown in brackets.

Parameter	Method	Conceptual Model	a	b	c	d	e	f	g	h	i	j	k
TN	Control	AS + BNA + POX_C + SR + (Cl + Si + Sa) + (OM) + pH	0.0205	0.000108	0.000517	0.000106	NCG	0.00160	NCG	-0.00058	- -	0.0260	NCG
	RFE	BNA + POX_C + (OM)	-0.00383	- -	0.000611	0.000100	- -	- -	- -	- -	- -	0.0312	- -
	CBFE	POX_C + (Sa) + (OM)	0.0545	- -	- -	0.000102	- -	- -	- -	-0.00087	- -	0.034031	- -
BNA	Control	AS + POX_C + SR + (Cl + Si + Sa) + (TN + OM) + pH	13.7	0.120	- -	-0.00920	23.8	0.377	NCG	NCG	55.7	1.29	-2.85
	RFE	AS + SR + (TN + OM)	0.169	0.146	- -	- -	21.9	- -	- -	- -	58.5	NCG	- -
	CBFE	POX_C + SR + (TN) + pH	27.5	- -	- -	-0.00449	28.5	- -	- -	- -	81.5	- -	-4.40
GSN	Control	AS + POX_C + SR + (Cl + Si + Sa) + (OM) + pH	76.5	0.416	- -	0.0254	70.3	1.79	NCG	-0.333	- -	16.05	-9.16
	RFE	AS + POX_C + SR + (OM)	19.6	0.349	- -	NCG	71.5	- -	- -	- -	- -	21.9	- -
	CBFE	SR + (Sa) + (OM) + pH	132	- -	- -	- -	82.4	- -	- -	-0.529	- -	22.8	-12.7

Notes: Formula for PTF estimating TN, BNA, or GSN = a + b (AS) + c (BNA) + d (POX_C) + e (SR) + f (Cl) + g (Si) + h (Sa) + i (TN) + j (OM) + k (pH); NCG = variable included in conceptual model, but no coefficient generated; ' - - ' = variable not included in conceptual model

The decrease in CCC observed between CU-TN_C and CU-BNA_C (Figure 3.3) was understandable based on their respective measures. TN, as a surrogate measure for the stable N pool, had a 2.9% higher R^2 and a 1.6% higher CCC when compared with BNA - the measure for the labile N pool. In addition, the variance between MLs increased with BNA trials (5%) versus TN trials (2%), which may reflect the biological variability inherent with BNA.

3.4.1.3 Growing Season Nitrogen

The Control-PTF for GSN (GSN_C) resulted in a 3% difference between MLs with the CU model (highest) at 0.83 and the MLR model (lowest) at 0.80 CCC (Figure 3.3, Table 3.3). As such, CU model results were selected for GSN_C, which had an R^2 of 0.72, an RMSE of 24.9, and a CCC of 0.83. The Control-CI was calculated from the SD of CCC (Table 3.3) and used as the principal metric for RFE and CBFE procedures. MLR and SGB underperformed CU comparatively, and there was no significant difference between CU-GSN_C and RF and SVM results (Figure 3.3). The theoretical analytical cost to obtain GSN_C parameters is $108. PTF coefficients from MLR results are shown in Table 3.4. The CI for MLR was used to assess MLR model results during RFE and CBFE for GSN (Figure 3.3, Table 3.3).

Interestingly, CU-GSN_C had the highest CCC as compared to CU-TN_C and CU-BNA_C (Figure 3.3). In fact, the GSN results for each ML outperformed their respective counterparts for TN and BNA, suggesting the higher CCC with the CU model was not an anomaly. Without a precedence for comparison in the literature, it is likely that the combination of the two parameters in the regression equation (Eq. 3.1) adds a predictive stability not found with the parameters taken individually. It may also be construed that

the high relative CCC of GSN lends a tacit endorsement of the relationship depicted in Eq. 3.1; in that, if the relationship were not sound, a consistent predictive pattern with related soil properties (independent variables) could not be learned.

3.4.2 Recursive Feature Elimination

3.4.2.1 Total Nitrogen

A reduced PTF with four predictors for TN (TN_{RFE}) was developed using RFE (Table 3.3). Over a series of nine iterations (Iter.), predictor variables were removed to identify a parsimonious conceptual model within the control range (Figure 3.4). The best PTF for TN_{RFE} was obtained with the CU model (Figure 3.4-Iter. 6) and the top predictors included BNA, POX_C, Sa, and OM (Table 3.3). Accumulated local effects (ALE) of TN_{RFE} (= y) for the CU model, showing how predictions change based on variance in the top predictors, is included in Figure 3.5. Calculated from 1,986 observations, the R^2 was 0.67, the RMSE was 0.026, the CCC was 0.80, and the theoretical cost was $82 (Table 3.3). PTF coefficients generated using MLR are listed in Table 3.4. The optimum MLR model result (Figure 3.4-Iter. 7) included BNA, POX_C, and OM as the most important predictors and a theoretical cost of $64. MLR accuracy metrics, based on 1,986 observations, included the R^2 at 0.64, RMSE at 0.027, and CCC at 0.77 (Table 3.3).

BNA, POX_C, Sa, and OM as the top predictors for CU-TN_{RFE} was consistent with correlation results (Figure 3.2), which showed BNA (0.60), OM (0.76) and POX_C (0.67) among the top correlated parameters. ALE of TN also shows positive correlations with BNA, POX_C, and OM (Figure 3.5). Figures 3.2 and 3.5 also show Sa negatively correlated with TN (-0.46) suggesting that TN increases with decreasing Sa content; a

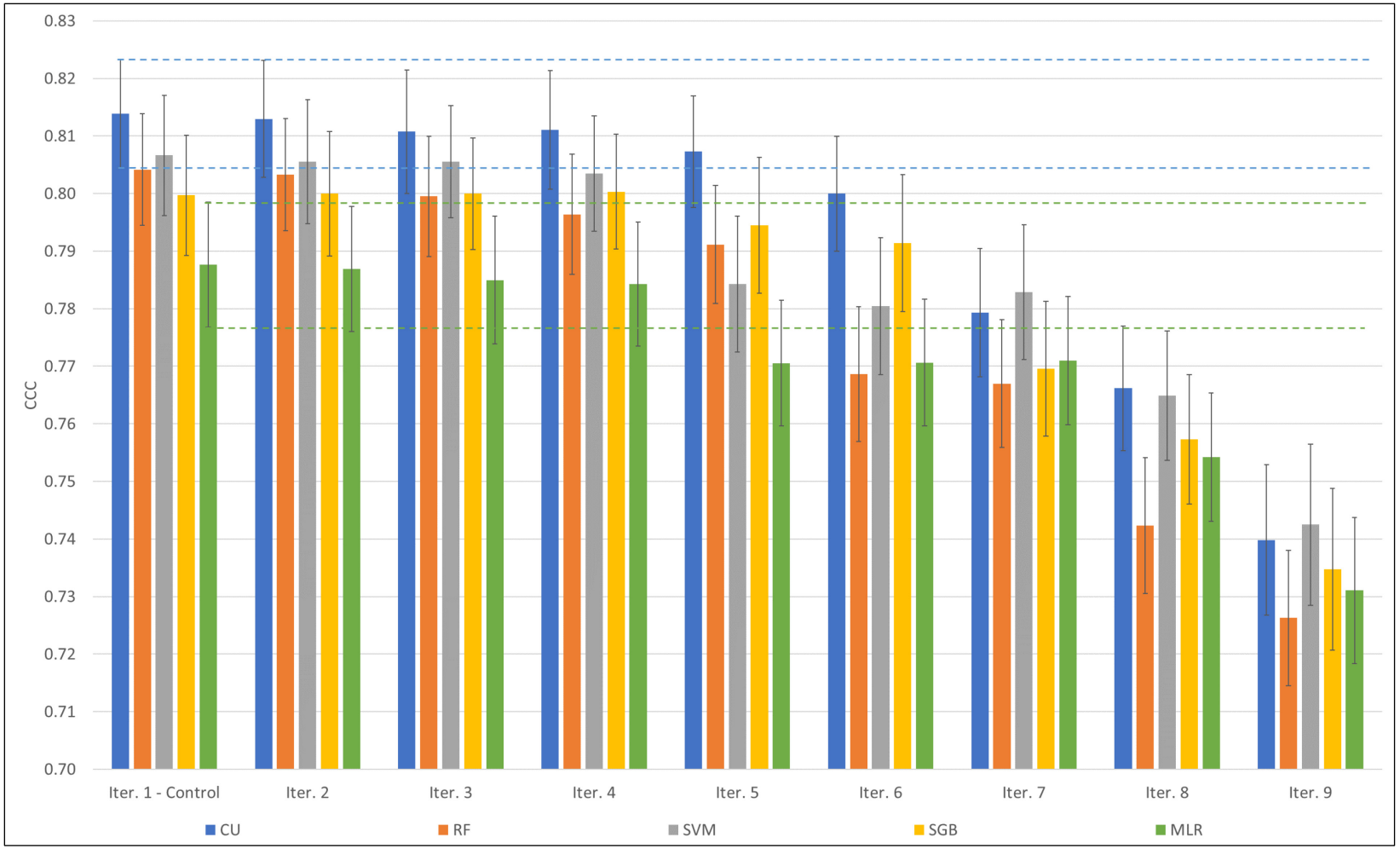


Figure 3.4 Chart of total nitrogen (TN) concordance (CCC) results from cubist (CU), random forest (RF), support vector machine (SVM), stochastic gradient boosting (SGB), and multiple linear regression (MLR) for each iteration (Iter.) of recursive feature elimination and showing the control interval from Iter. 1 for CU and MLR.

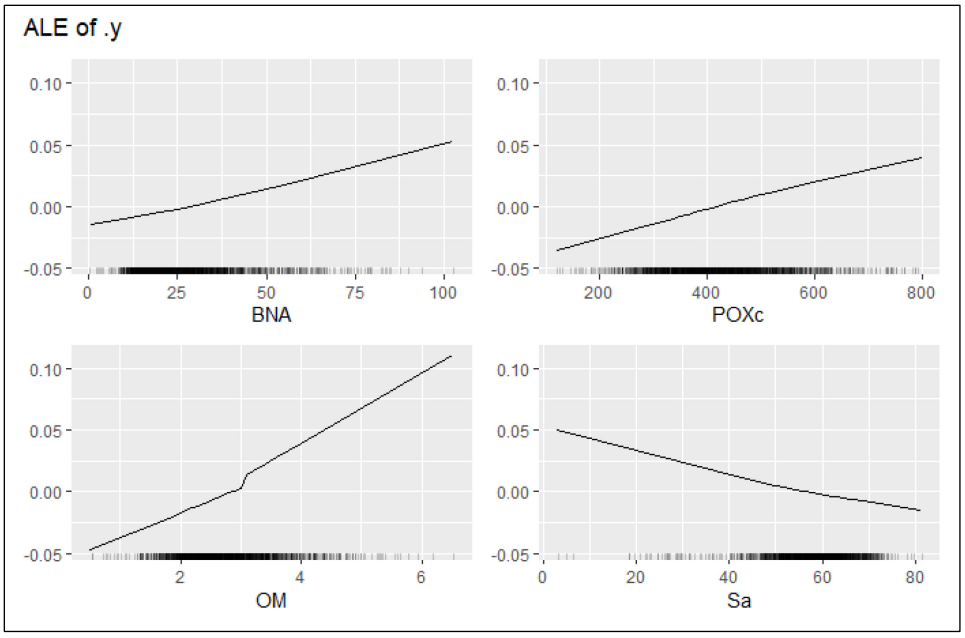


Figure 3.5 Accumulated local effects (ALE) of total nitrogen (TN = .y) for the cubist model depicting how biological nitrogen availability (BNA), permanganate oxidizable carbon (POX_C), organic matter (OM), and sand (Sa), the top predictors identified through the recursive feature elimination process, are influenced by variances in respective soil concentrations.

tenable conclusion based on Matus (2021) who found that Si and Cl content were the fundamental drivers of C and N storage in soil. The fact that neither Si nor Cl were retained as predictors, given their importance in accumulating TN, is likely a function of the study areas textural class, which is dominated by Sa (mean = 57%) versus Si (mean = 31%) and Cl (mean = 12%) percentages (Table 3.1). Comparing other PTFs for TN in the literature, soil organic carbon (SOC) was used as the sole predictor for Rashidi and Seilsepour (2009) and Mesele and Ajiboye (2020) who used regression analysis in order to yield R^2 values of 0.83. RMSE nor CCC were reported in their studies. With respect to SOC, the SHD reports SOC (or TC) converted to OM. OM as a single predictor for TN was tried in this study, but underperformed with the highest $R^2 = 0.59$ and a CCC of 0.74 from the SVM learner (Figure 3.4-Iter. 9). Similar R^2 results, as found in Rashidi and

Seilsepour (2009) and Mesele and Ajiboye (2020), were not achieved in this study where the highest R^2 for TN was 0.69 for CU-TN_C (Table 3.3). Using R^2 as the primary accuracy metric is problematic however, given the potential for systematic bias wherein predictions may adhere to the line of best fit but be skewed from the origin, or 1:1 line. Correcting for this bias, CCC as the preferred metric was comparatively strong at 0.80 and arguably a more reliable representation. CCC results were not reported in the aforementioned studies so could not be compared. Glendining et al. (2011), who developed a PTF for TN at a global scale, incorporated latitude as a predictor along with OC, the soil classification group's mean C:N ratios, and soil textural class identified based on Cl or Sa content. Due to the global scale for use, and the relative size and climatic homogeneity of the study area, this PTF was not tested for comparison.

In addition to yielding a simper model, RFE provided a more cost-efficient PTF and maintained a high level of accuracy. The theoretical cost was reduced by 41% without a significant drop in CCC (1.4%) when comparing CU-TN_C and CU-TN_{RFE} (Table 3.3). RFE's retention of BNA, as a relatively novel and costly two-week incubation analysis, lends a certain impracticality to CU-TN_{RFE} in terms of transferability with other users worldwide. In addition, the use of the CU learner for deriving TN_{RFE} complicates the ability to transfer model learnings; in that, a potential user would require a substantial database equipped with both dependent and independent variables for training purposes. MLR-TN_{RFE} results were statistically below CU-TN_C (Figure 3.4); however, with a CCC of 0.77, a 54% reduction in theoretical cost (Table 3.3), and the generation of model coefficients (Table 3.4), MLR-TN_{RFE} may be an appropriate PTF for a variety of circumstances such as *ad hoc* or small-scale applications.

3.4.2.2 Biological Nitrogen Availability

The simplified PTF for BNA using RFE (BNA_{RFE}) was obtained after six iterations with four predictors required to accurately predict BNA using the CU model (Figure 3.6, Table 3.3). The predictors selected, and their correlations (Figure 3.2), included AS (0.59), POX_C (0.46), SR (0.74), and TN (0.60). Accumulated local effects (ALE) of BNA_{RFE} (= y) for the CU model is included in Figure 3.7. CU-BNA_{RFE} had an R^2 of 0.64, RMSE of 8.2, and a CCC of 0.78 based on 1,984 observations. The theoretical cost to analyze the predictor variables was $83. Using MLR, four predictors were also required with the difference that OM (0.59) was selected in place of POX_C. For MLR-BNA_{RFE} (n = 1,996), the R^2 was 0.58, the RMSE was 8.8, the CCC was 0.73, and the theoretical cost was $69 (Table 3.3). Model coefficients from MLR are presented in Table 3.4.

Correlations (Figure 3.2) were consistent with the predictor variables selected by RFE. Interestingly, OM was not retained by the best model (CU) yet showed an equal or better correlation with AS and POX_C, respectively. AS analysis measures the strength of a soil's structure in relation to "destabilizing stressors" via simulated rainfall (Angers and Carter, 2020; Moebius-Clune, 2016). Soil structure is formed via biological breakdown processes where decomposed organics are transformed into agents for binding soil minerals (Rieke et al., 2022). Aggregate stability also influences the biological environment in which N mineralization occurs. In the CU interpretation, AS proved a stronger indicator of active biological processes than OM and was thus retained. However, the ALE plot of BNA (Figure 3.7) showed a relatively horizontal relationship with AS in a majority of samples which suggests that, while a useful predictor, it is not as

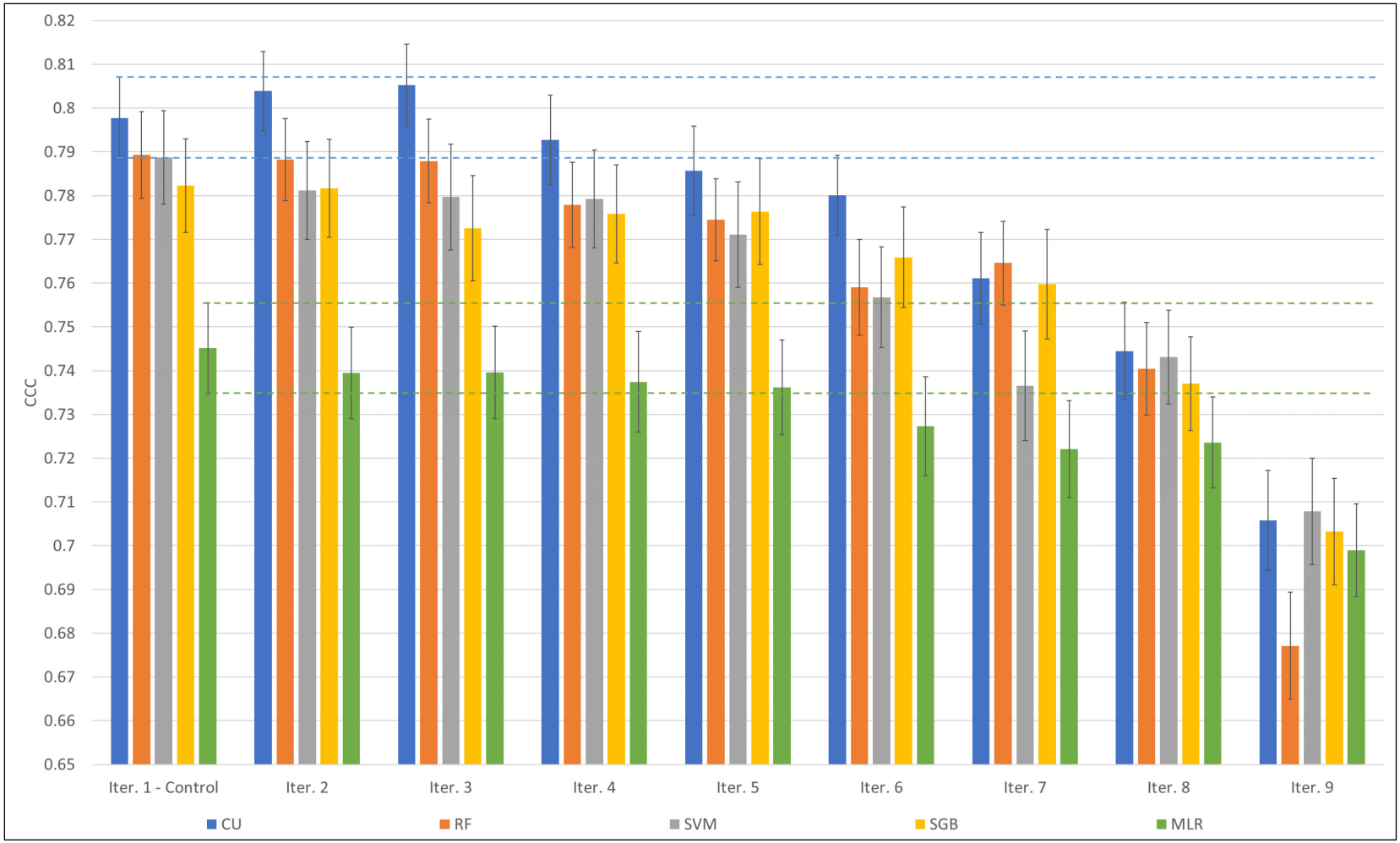


Figure 3.6 Chart of biological nitrogen availability concordance (CCC) results cubist (CU), random forest (RF), support vector machine (SVM), stochastic gradient boosting (SGB), and multiple linear regression (MLR) for each iteration (Iter.) of recursive feature elimination also showing the control interval from Iter. 1 for CU and MLR.

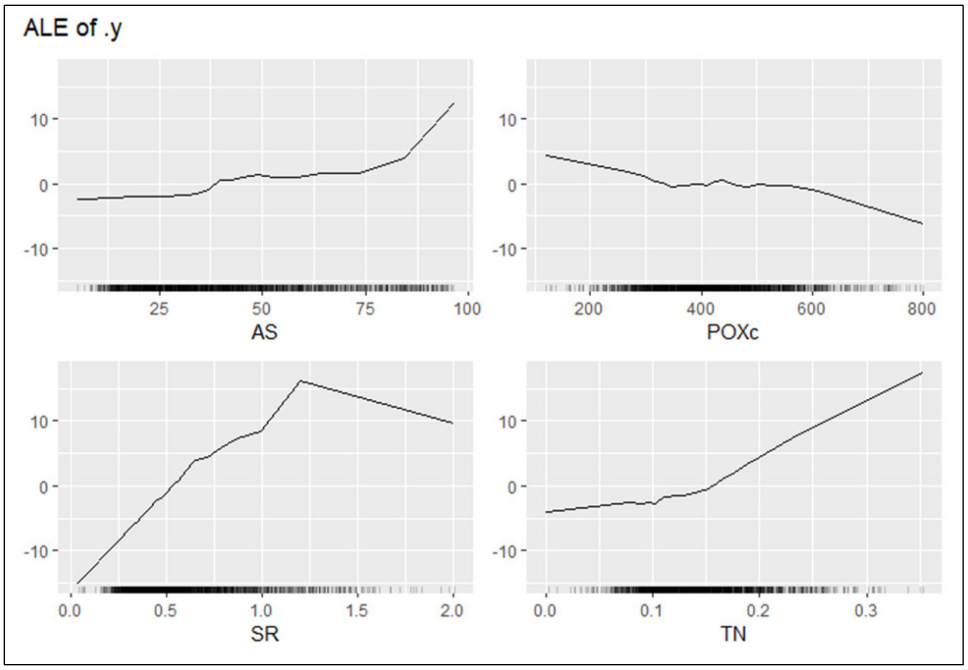


Figure 3.7 Accumulated local effects (ALE) of biological nitrogen availability (BNA = .y) for the cubist model depicting how aggregate stability (AS), permanganate oxidizable carbon (POX_C), soil respiration (SR), and total nitrogen (TN), the top predictors identified through the recursive feature elimination process, are influenced by variances in respective soil concentrations.

strongly associated to BNA as the other important predictors. RFE's selection of TN and POX_C is understood given that TN includes in its measure the total organic and inorganic N present in soil (including biologically available N), and given that POX_C measures the component of OM that is readily available for decomposition (Moebius-Clune, 2016). The most important predictor was SR, showing a strong correlation (Figure 3.2) that is positively correlated for most samples, and negatively correlated beyond the 3rd quartile of samples (Figure 3.7; Table 3.1). SR, like BNA, is an aerobic respiration test that measures metabolic activity in soils over a given time (Moebius-Clune, 2016). PTFs specifically using the BNA test could not be found in the literature; however, Rasiah (1995) developed PTFs for N_{min} parameters for both one and two N pool models using

the incubation method proposed by Stanford and Smith (1972). Representing the labile N pool (N_l), their PTF for potentially mineralizable N included TN, a residue removed factor (0 = yes, 1 = no), pH, and Cl content as the predictors with an R^2 of 0.94 (RMSE or CCC were not reported). A related study by Heumann et al. (2011) who used a similar incubation method, found that PTFs for the fast (labile) mineralizing pool included Cl and/or mean Cl content of textural class, humus class, and mean fall temperature were more successful when grouped by former land-use (R^2 = 0.34 to 0.42). In comparison with our study, related parameters were retained except for pH and Cl content. However, the inclusion of AS is perhaps a 'catch all' parameter in terms of the multiple processes (e.g., residue breakdown, biological activity) and structural factors (e.g., moisture holding capacity and aeration as a function of texture and matrix architecture) that are intrinsic to aggregation. Biological activity as influenced by cropping or soil management practices related to land-use were important factors for both Rasiah (1995) and Heumann et al. (2003) but were not present in the SHD so could not be considered as factors.

The theoretical cost of CU-BNA_{RFE} was 23% lower and had a minor drop in CCC of 1.8% when compared with the control (Table 3.3). RFE reduced the variables from nine to four and retained AS which is the highest cost analytical method in the SHD (Table 3.2). MLR-BNA_{RFE} resulted in a 46% reduction in theoretical cost as compared to the control and had a CCC that was satisfactory, yet significantly below CU-BNA_{RFE} (Figure 3.6, Table 3.3). The cost savings between the CU and MLR models is primarily due to the inclusion of the multiparameter analysis (TN and OM) with MLR in exchange for the need for POX_C in the CU model. This observation highlights the practicality of multiparameter analysis when factoring cost into PTF development.

3.4.2.3 Growing Season Nitrogen

Using the process of RFE, a simplified PTF for GSN (GSN_{RFE}) was successfully developed. Five predictors with the CU learner yielded the best results (Figure 3.8-Iter. 4) and the top predictors with their correlations included AS (0.62), POX_C (0.57), SR (0.75), OM (0.69), and pH (0.01) (Figure 3.2, Table 3.3). Accumulated local effects (ALE) of GSN_{RFE} (= y) for the CU model is included in Figure 3.9. Calculated from 1,984 observations, the R^2 was 0.69, the RMSE was 26, the CCC was 0.82, and the theoretical cost totaled $89. Using MLR, four predictors including AS, POX_C, SR, and OM were selected (Figure 3.8-Iter. 5, Table 3.3) and their respective coefficients are reported in Table 3.4. MLR-GSN_{RFE} was determined based on 1,985 observations with an R^2 of 0.65, a RMSE of 27, a CCC of 0.78, and a theoretical cost of $83 (Table 3.3).

GSN, as a calculated output from TN and BNA, retained similar predictors variables. ALE of GSN showed a relatively horizontal relationship with AS in most samples, which may indicate a weaker pedogenic association than the other GSN predictors (Figure 3.9). Recalling that GSN is an output of a two-pool regression calculation (Eq. 3.1), it is somewhat expected that the predictor variables would mirror both the stable N pool (TN) and the biologically active, labile N pool (BNA). GSN_{RFE} retained POX_C and OM which were required to predict TN (TN_{RFE}), as well as AS and SR, which were instrumental in BNA predictions (BNA_{RFE}). Uniquely, pH, which was not selected for TN_{RFE} or BNA_{RFE}, was selected as a predictor for GSN estimates. Fierer and Jackson (2006) described a similar phenomenon in their study that showed pH as a driver in differentiating microbial communities in soil at the continental scale. With respect to prediction accuracy, Figure 3.10 highlights a comparison of results and how

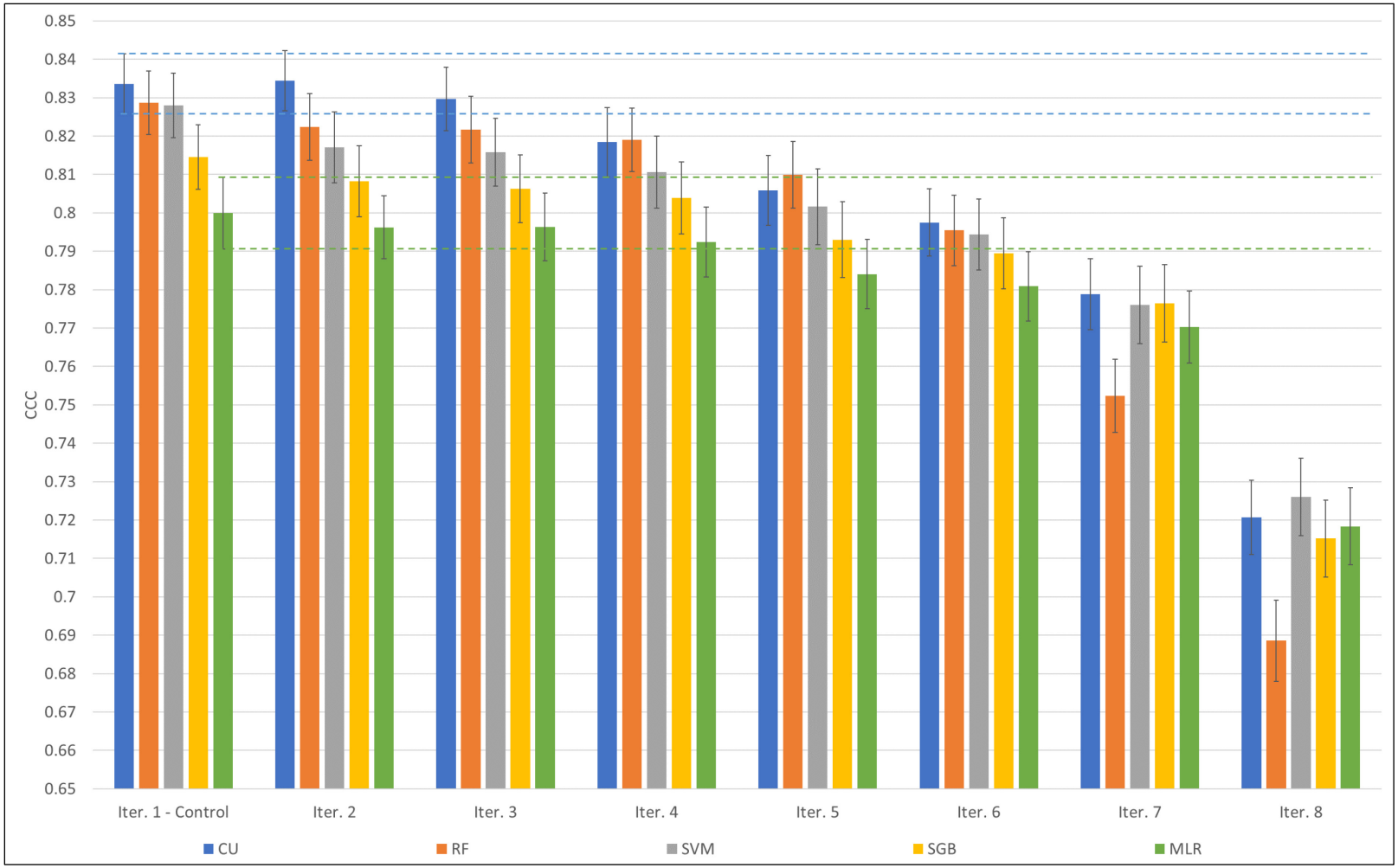


Figure 3.8 Chart of growing season nitrogen concordance (CCC) results from cubist (CU), random forest (RF), support vector machine (SVM), stochastic gradient boosting (SGB), and multiple linear regression (MLR) for each iteration (Iter.) of recursive feature elimination and showing the control interval from Iter. 1 for CU and MLR.

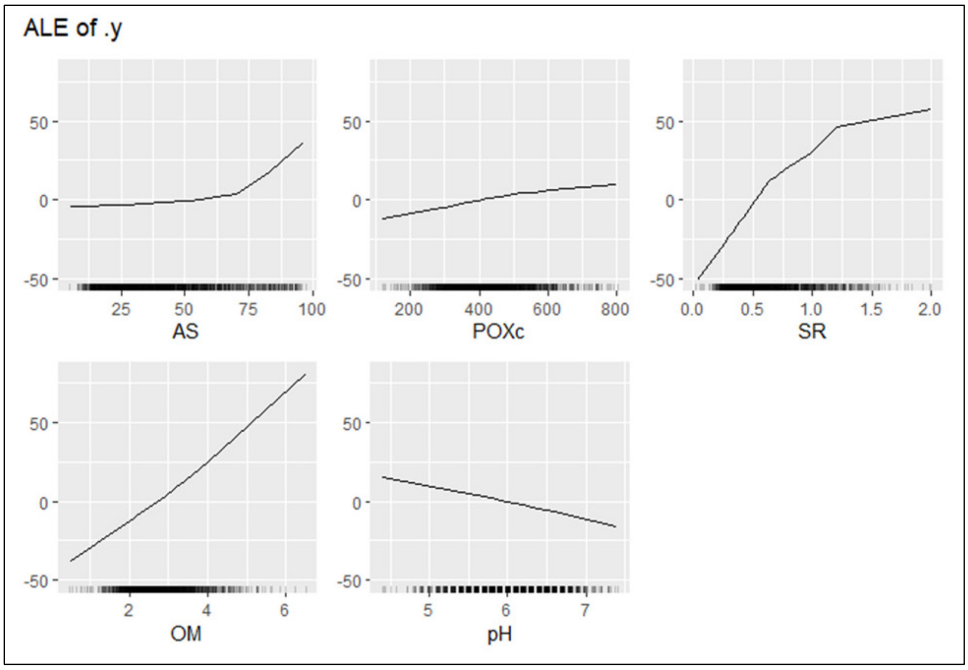


Figure 3.9 Accumulated local effects (ALE) of growing season nitrogen (BNA = .y) for the cubist model depicting how aggregate stability (AS), permanganate oxidizable carbon (POX_C), soil respiration (SR), organic matter (OM), and pH, the top predictors identified through the recursive feature elimination process, are influenced by variances in respective soil concentrations.

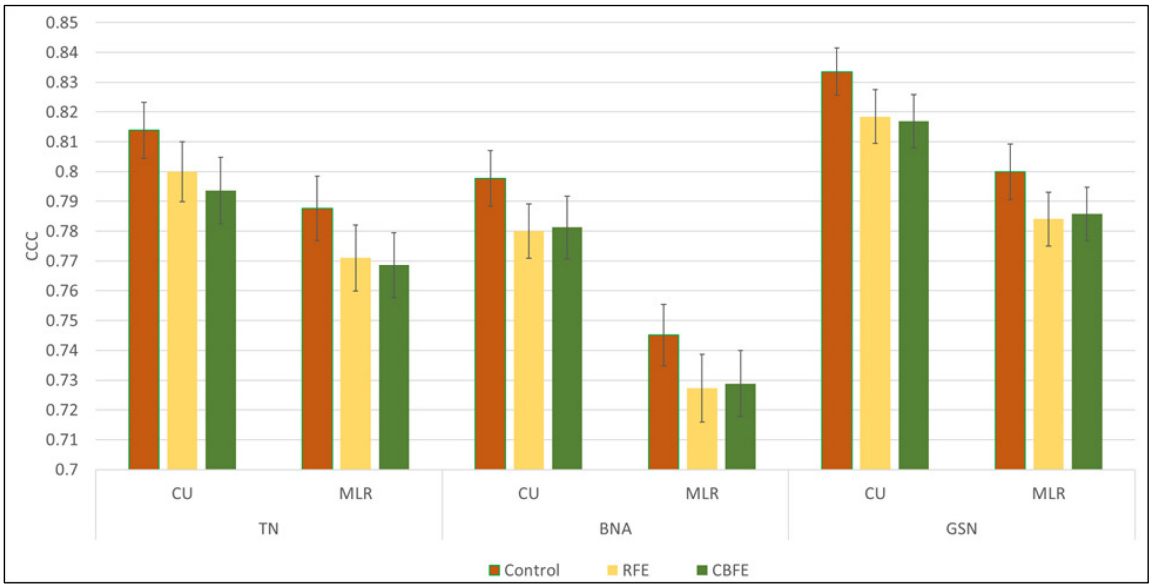


Figure 3.10 Comparison of cubist (CU) and multiple linear regression (MLR) concordance (CCC) results using recursive feature elimination (RFE) and cost-benefit feature elimination (CBFE) methods for total nitrogen (TN), biological nitrogen availability (BNA), and growing season nitrogen (GSN) pedotransfer functions.

CU-GSN_{RFE} showed a higher CCC (0.82) when compared with the respective TN (0.80) and BNA (0.78) counterparts. While this result is not significantly higher than TN, the CCC is significantly higher than BNA and shows that combining TN with BNA via Eq. 3.1 adds a predictive strength not found with these parameters alone.

The process of RFE was successful at reducing the theoretical cost by 17% as compared with the control (Table 3.3). In comparison, CU-TN_{RFE} achieved a 41% reduction, and CU-BNA_{RFE} had a 23% reduction, showing GSN_{RFE} to have the smallest theoretical cost savings. This higher overall cost may be explained in that CU-GSN_{RFE} is a more complex relationship requiring five predictors as opposed to CU-TN_{RFE} and CU-BNA_{RFE} (both with four predictors). MLR-GSN_{RFE} had a 23% cost reduction compared with the control and maintained a reasonable CCC at 0.78 (Table 3.3). Recalling that the GSN output value can be directly applied to fertilizer recommendation calculations, the theoretical savings of the MLR results combined with the ability to apply model coefficients to other datasets (or individual analysis), may render the MLR result more practical, albeit significantly lower in CCC than CU results.

3.4.3 Cost-Benefit Feature Elimination

3.4.3.1 Total Nitrogen

As an alternative to the RFE approach, CBFE was used to select parameters for predicting TN (TN_{CBFE}). Five predictors were required with the CU learner giving the best results (Figure 3.11-Iter. 10). The most cost-effective predictors for CU-TN_{CBFE} included POX_C, SR, Cl, Sa, and OM (Table 3.3). Calculated from 1,985 observations, the R^2 was 0.65, RMSE was 0.027, a CCC of 0.79 and a cost of $70. MLR-$TN_{CBFE}$ required three predictors (POX_C, Sa, OM) based on 1,986 observations (Figure 3.11-Iter. 10, Table

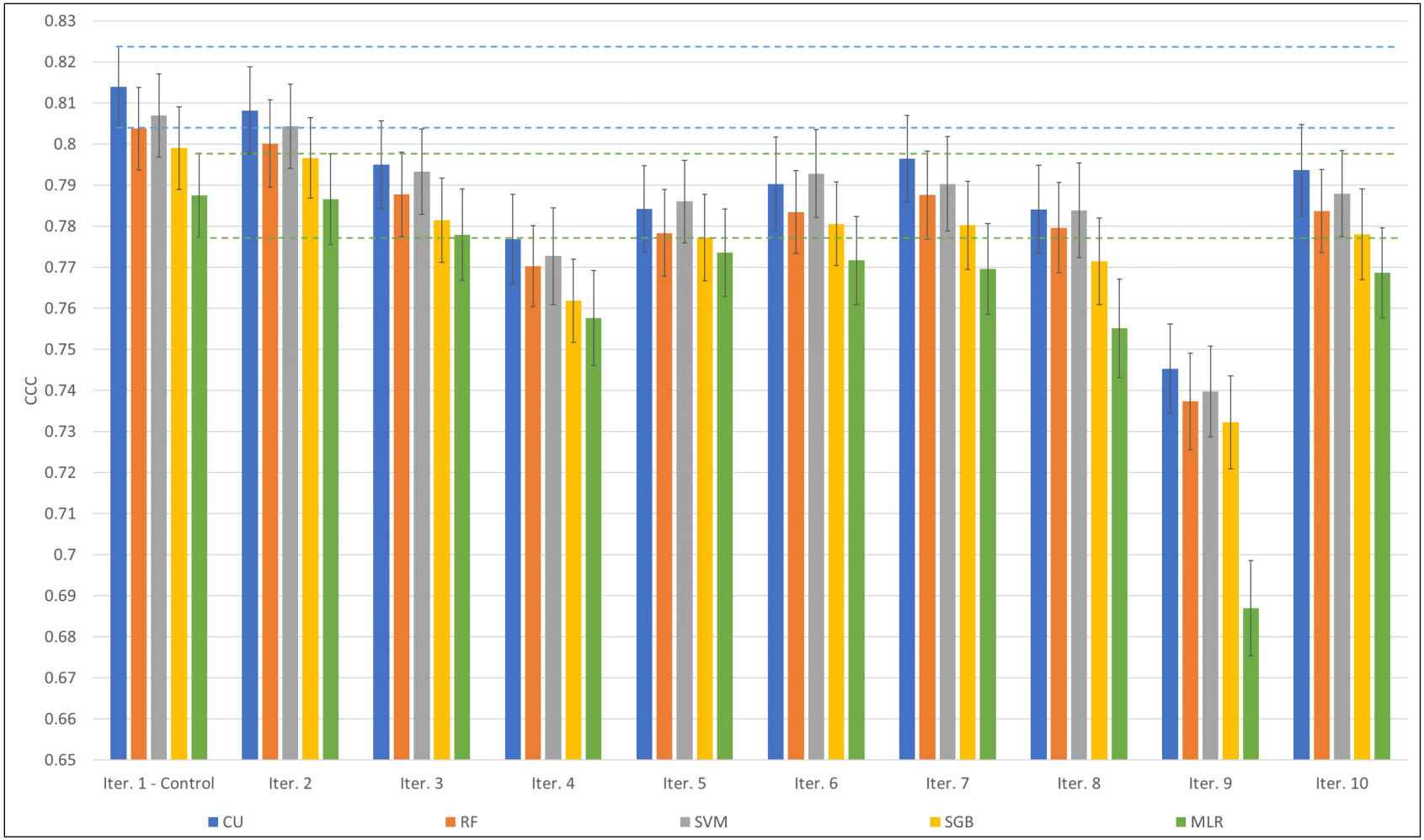


Figure 3.11 Chart of total nitrogen concordance (CCC) results from cubist (CU), random forest (RF), support vector machine (SVM), stochastic gradient boosting (SGB), and multiple linear regression (MLR) for each iteration (Iter.) of cost-benefit feature elimination and showing the control interval from Iter. 1 for CU and MLR.

3.3). The R^2 was 0.64, the RMSE was 0.027, the CCC was 0.77, and the cost was $51. MLR-$TN_{CBFE}$ model coefficients are in Table 3.4.

Using the CBFE approach, the theoretical cost savings were 49% when compared to the control. In this approach, BNA as a high-cost predictor was able to be removed (Figure 3.11-Iter. 3) without a significant drop in CCC. BNA was retained by CU-TN_{RFE}, which contributed to its higher theoretical cost. As such, the CBFE process enables a more cost-efficient set of predictors without losing accuracy. Another observation relates to parameters that were dropped by RFE (due to a lesser importance) but retained by CBFE because of a higher cost-benefit. The parameters SR vs. BNA are an example of this. Considering the conceptual model for CU-TN_{RFE} in Table 3.3, BNA was evidently retained because the inclusion of a labile-N parameter is important when predicting TN. However, CBFE showed that BNA could be replaced with another incubation test and surrogate measure of the labile N pool (i.e., SR) with similar predictive benefit, but at a reduced cost. MLR-TN_{CBFE} results, requiring only three parameters, were able to make a good prediction of TN (CCC = 0.77) without the inclusion of BNA or SR (Table 3.3). While MLR is significantly below CU results, the practicality of this PTF with its 63% cost savings compared to the control, and its fewer predictors, may render it useful for field-based applications.

3.4.3.2 Biological Nitrogen Availability

A reduced PTF for BNA with six predictors via the CU model was developed using CBFE (CU-BNA_{CBFE}). Figure 3.12-Iter. 11 and Table 3.3 display the results as calculated from 1,984 observations and accuracy metrics showing an R^2 of 0.64, an RMSE of 8.1, a CCC of 0.78 and a cost of $76. The final predictors of the CBFE process

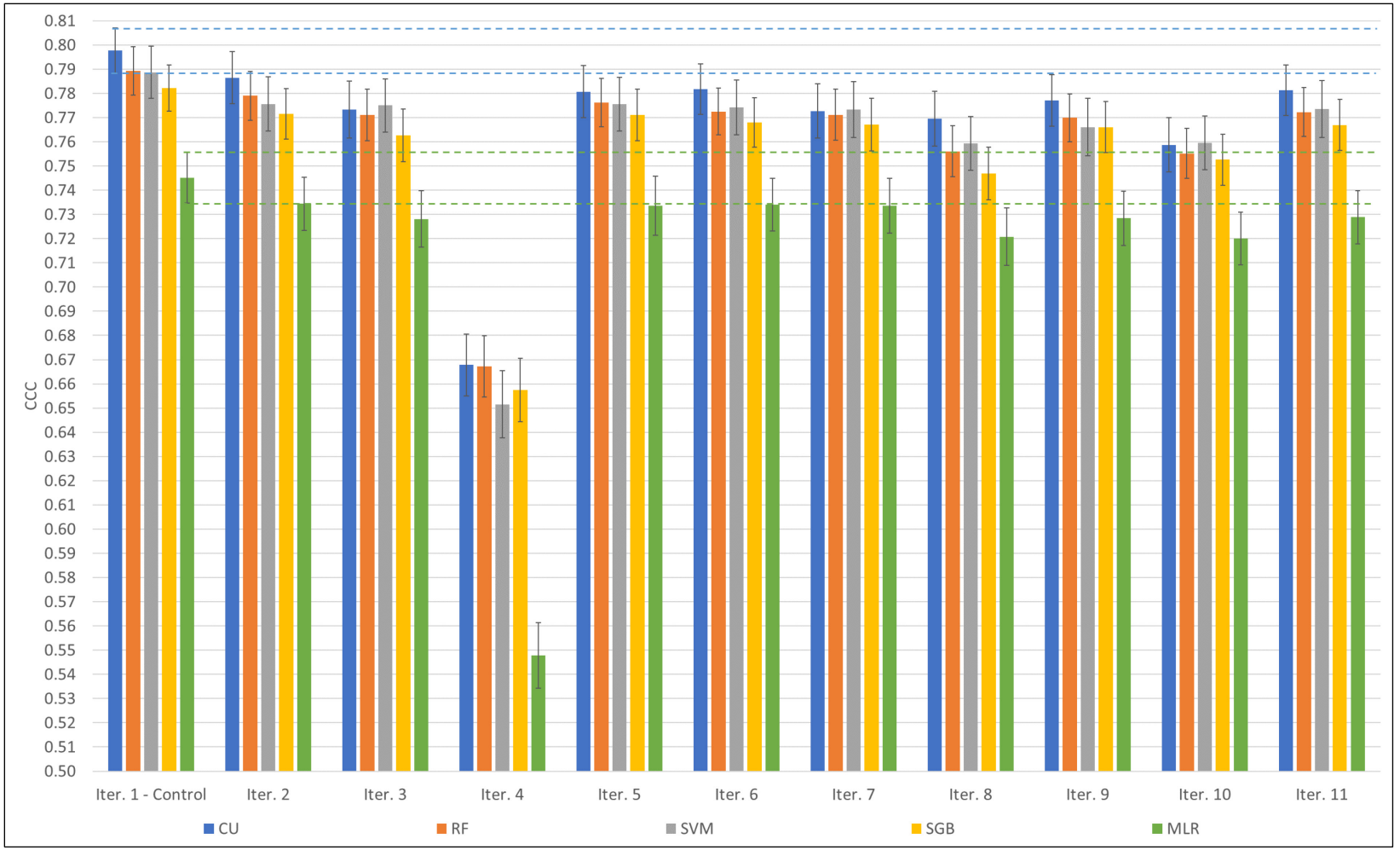


Figure 3.12 Chart of biological nitrogen availability concordance (CCC) results from cubist (CU), random forest (RF), support vector machine (SVM), stochastic gradient boosting (SGB), and multiple linear regression (MLR) for each iteration (Iter.) of cost-benefit feature elimination and showing the control interval from Iter. 1 for CU and MLR

included POX_C, SR, Sa, TN, OM, and pH for a 29% reduction in theoretical cost compared to the control. Comparing RFE (CU-BNA_{RFE}) vs CBFE (CU-BNA_{CBFE}) results in Table 3.3, the RFE process yielded a simpler PTF with four vs six predictors, but did so at a 6% higher cost. RFE retained AS as an important predictor (the highest priced predictor), while instead, CBFE retained Sa, OM, and pH (the lowest priced predictors) as necessary measures for aggregation. This substitution gives credence to the importance of soil structure when predicting BNA; a premise supported by Rasiah (1995) who found that decreasing structure, as with higher Sa content, allowed for higher N_l (labile N pool) biological activity due to accessibility of the substrate (e.g. aggregate). An example of an important predictor common to both the RFE and CBFE methods is SR, which is not surprising given its similarity to the BNA analysis. A clear impact of the SR parameter can be seen in Figure 3.12, with the 14% drop in CCC due its removal in Iter. 4, and its replacement in Iter. 5.

MLR-BNA_{CBFE} required four predictors (Figure 3.12-Iter. 11) including POX_C, SR, TN, and pH (Table 3.3). From 1,984 observations, the calculated R^2 was 0.58. the RMSE was 8.7, the CCC was 0.73, and the cost was $58. MLR model coefficients for BNA_{CBFE} are in Table 3.4. Comparing MLR-BNA results for both RFE and CBFE, the CBFE process yielded an 11% greater theoretical savings and retained the same CCC at 0.73 (Table 3.3). These results were significantly below CU-BNA results; however, given that the BNA analysis may be difficult to obtain, and yet given its practical importance for estimating the labile N pool, the MLR model may be a useful function under correct conditions.

3.4.3.3 Growing Season Nitrogen

Using the CBFE process, a simplified PTF for GSN was created using the CU model (CU-GSN_{CBFE}). The results showed that five predictors were necessary to achieve the best results (Figure 3.13-Iter. 10) and the top predictors included POX_C, SR, Sa, OM, and pH (Table 3.3). This result was calculated from 1,984 observations, where the R^2 was 0.69, the RMSE was 26, the CCC was 0.82 and the theoretical cost was $76. Like RFE results, GSN_{CBFE} had relatively higher CCC values compared to TN and BNA (Figure 3.10). The CBFE process, while maintaining a significant CCC, obtained a 29% reduction in theoretical cost compared to CU-GSN_C, and a 12% reduction over RFE. The main predictor eliminated by this process, but retained by RFE, was the AS parameter. In its place, Sa was selected with a similar benefit and in order to account for the influence of soil structure (Table 3.3). Similarly, with BNA_{RFE} (Figure 3.12-Iter. 4), Figure 3.13-Iter. 4 also showed a dramatic reduction in CCC following the elimination of SR; but with a smaller reduction (12%) as compared to BNA (14%). This smaller reduction can likely be attributed to the stability of GSN's inclusion of TN.

MLR-GSN_{CBFE} results identified four predictors including SR, Sa, OM, and pH (Table 3.3). Based on 1,995 observations, the calculated R^2 was 0.66 the RMSE was 27, the CCC was 0.79, and the cost was $53. Model coefficients generated from MLR are in Table 3.4. CBFE resulted in a 51% reduction in theoretical cost in comparison to the control (MLR-GSN_C), and a 28% reduction in comparison with RFE (MLR-GSN_{RFE}). AS and POX_C were eliminated in the CBFE process but retained in the RFE process. Instead, soil pH for its influence on biological activity (Fierer and Jackson, 2006), and Sa for its impact on aggregation and microbial access to substrate (Rasiah, 1995), was retained as

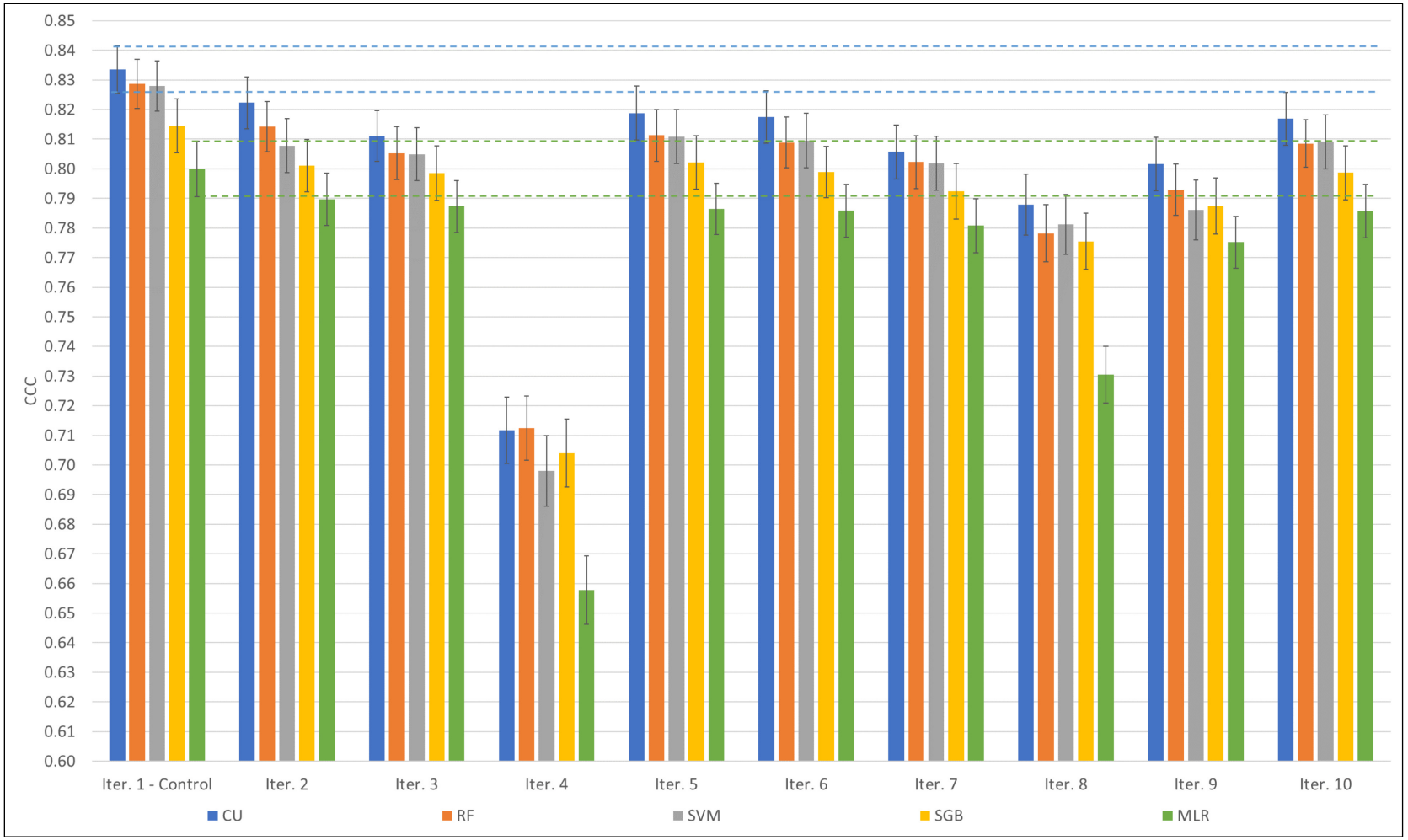


Figure 3.13 Chart of growing season nitrogen concordance (CCC) results from cubist (CU), random forest (RF), support vector machine (SVM), stochastic gradient boosting (SGB), and multiple linear regression (MLR) for each iteration (Iter.) of cost-benefit feature elimination and showing the control interval from Iter. 1 for CU and MLR.

important parameters. MLR results, while significantly different to CU results, produced a reliable PTF (CCC = 0.79) for GSN that requires fewer parameters than the CU model.

3.4.4 General Discussion

3.4.4.1 Pedotransfer function evaluation

Creation of a standard by which simplified PTFs could be evaluated was improved by establishing Control-PTFs for TN, BNA, and GSN. Control-PTFs provided a database specific, or internal benchmark for comparing accuracy metrics and a cut-off point for the feature elimination process. This process was able to augment sole reliance on comparing PTFs for N in the literature, which were sparse. The SHD used for this study was unique in that the parameters included in the database were specifically selected to capture not only standard soil quality parameters, but also biologically driven properties as recommended by Moebius-Clune (2016). While not all the desired parameters were included, available water holding capacity for example, the SHD provided the opportunity to develop a PTF with appropriate parameters in order to obtain a 'best case', benchmark, or control scenario. In general, obtaining a Control-PTF, and establishing the Control-CI for TN, BNA, and GSN were seen to be a valuable process and recommended when related parameters are available.

3.4.4.2 Comparison of machine learners

The use of multiple MLs for comparison purposes has become a relatively standard practice in recent years. While some studies still employ only one method for PTF development, comparison between two or three learners appeared more common in studies published after the year 2020. A cursory look at Figures 3.4 through 3.10 will show that while there is variation between the learners tested it is seldom a significant

variation. Except for MLR, the majority of MLs were within the same confidence range for each iteration. As such, even though the CU model was chosen consistently for the control model, it should not be considered a significantly better model than RF, SVM, or SGB for predicting TN, BNA, and/or GSN. MLR was significantly below the other MLs tested for most iterations; however, from a practical perspective, it was advantageous to include to provide parameter coefficients (Table 3.4). These coefficients are extremely useful for *ad hoc* applications such as supporting N fertilizer recommendations.

3.4.4.3 Feature elimination

Similar to testing multiple MLs, evaluation of different feature elimination approaches proved useful and is recommended for parameter selection. The process of RFE was used successfully and yielded final predictors with the highest variable importance, irrespective of cost. CBFE on the other hand appeared to achieve both a set of relevant predictors, and ones that were optimized for cost. With respect to CBFE, this typically meant that the reduced model may not have contained features with the highest variable importance, but with an equivalent importance at a lower theoretical cost. McBratney et al. (2002) noted that one of the principles for "*useful or efficient PTFs*" related to the quality, or cost of the information. And again, Odeh and McBratney (2005) mentioned cost when defining PTFs as the prediction of a soil property from other "*easily, routinely, or cheaply measured properties*". As an example, the CU-BNA_{RFE} relationship (Table 3.3) where $BNA = f(AS + AC + SR + TN)$. The cited principle suggests that one should not measure AS, POX_C, SR, and TN for the strict purpose of predicting BNA, since their combined cost is above the cost of BNA itself (Table 3.2). Cost, in this study, is used mainly as a surrogate measure for ease, or effort in obtaining a

soil measure. Based on the results, CBFE was able to obtain predictions of a similar strength to RFE, at a consistently lower theoretical cost. As such, the factoring of cost into feature elimination is useful, but since selected features may not have the highest variable importance, it is recommended as a complimentary process to RFE. Another recommendation with the cost-benefit approach is that, since variable importance isn't driving feature elimination, all control parameters (prior to feature elimination) must be pedogenically related to the dependent variable. And lastly, that reference to a Control-CI is imperative to maintain a standard accuracy of prediction.

3.4.4.4 Nitrogen pool dynamics

RFE and CBFE, coupled with ML model capabilities, provided notable insights for interpreting soil N pool dynamics. The importance of particle size fractions (Sa, Si, Cl) or AS in predicting TN, BNA, and GSN revealed the importance of soil architecture and microbial accessibility of the substrate to Cl and N storage (Matus, 2021; Rasiah, 1995). Furthermore, the predictive stability observed with PTFs for GSN suggests that purely mechanistic relationships with TN and BNA, taken in isolation, may not reflect the full diversity of factors at play when determining N pool dynamics. As such, directed ML approaches may grant a more robust approach for predicting complex relationships such as GSN.

3.5 CONCLUSION

The intent of this study was to improve our understanding of soil N pool dynamics and generate GSN estimates to support "right rate" N fertilizer recommendations in order to reduce RSN losses via leaching to groundwater and/or volatilization to atmosphere. The PTFs presented here could be incorporated into (two-

pool) regression equations that account for the labile N pool, and the stable N pool, to make duration-based N_{min} estimates. The framework, as demonstrated, identified important predictors for TN, BNA, and GSN using a variable-importance approach (RFE), and a cost-based approach (CBFE). Resultantly, a novel method for incorporating practicality, effort, or cost, was able to maintain high prediction standards through a suite of MLs and select important and cost-effective predictors. Testing multiple MLs was helpful to identify the best model; however, given the practical intent of this study, PTF regression coefficients may be the best suited for *ad hoc* predictions transforming the data a crop producer might have, into the data a crop producer might need. Unfortunately, and as a recommendation for further study, influences of climate and soil management practices could not be, but should be considered for N_{min} and N related PTFs. Fate and transport dynamics of N pools and N_{min} is greatly influenced by landscape, moisture, crop rotations, tillage, and residue management. These factors likely influenced some of the variance observed, and may greatly enhance further studies. Also, given climate and landscape influence on N_{min}, the findings of this study would be well suited for predictive soil mapping applications. Future maps of N_{min} parameters might allow more crop producers to access GSN predictions and provide greater support for informed "right-rate" N fertilizer applications worldwide.

CHAPTER 4: INTEGRATING MULTI-YEAR CROP INVENTORIES AS A PROXY FOR SOIL MANAGEMENT PRACTICES WITHIN A DIGITAL SOIL MAPPING FRAMEWORK FOR PREDICTING NITROGEN INDICES [2]

4.1 ABSTRACT

For the international digital soil mapping (DSM) community, adequate spatial estimates of nitrogen (N) mineralization have been difficult to generate. This is due, in part, to an inability to capture critical N controls at the regional and provincial scales. While the influence of climate, vegetation, and relief are accessible predictors in DSM, the effect of soil management, or crop rotation, is known for its important influence on N dynamics, but has hitherto been elusive for soil mappers. To help inform N fertilizer management, the intention of this study was to determine the importance of novel crop frequency/soil management layers as well as top predictors and controls through the development of provincial scale DSMs of total nitrogen (TN), biological nitrogen availability (BNA), representing the stable and labile N pools, respectively; and in addition, the calculated estimate of N mineralization over a growing season (GSN). DSMs were developed using a provincial soil quality monitoring database (SQMD), consisting of georeferenced sample points (n = 675) containing direct measures of TN and BNA for training data, and covariates at a 30 m spatial resolution representing climate, vegetation/organisms, and relief as predictor layers. In addition, novel crop and soil management layers were developed that estimate the number of times a particular crop type was planted over a 10-year period, thus capturing tillage intensity via cropping

[2] Chapter 4 is a version of a manuscript that was submitted to Geoderma (open source) on February 29, 2024, and conditionally accepted pending moderate revision on April 8, 2024. This was a multi-authored manuscript (Laurence, L., Heung, B., Zhang, J., Pennell, T., Nyiraneza, J., Strom, H., Stiles, K., and Burton, D.L.), in which the concept, design, data processing, and writing was done by the PhD Candidate with the assistance of all co-authors.

frequency. Results for TN were 27% higher with the use of novel soil management layers achieving a final CCC of 0.45 using SVM. BNA predictions increased by 24% using novel layers and had a CCC of 0.45 with the SGB learner. GSN showed the least improvement using novel layers (6%) but showed the highest CCC of the study (0.47) using novel soil management layers with SGB. Prediction maps, and related uncertainty maps, of TN, BNA, and GSN were developed showing higher uncertainty with N parameters in areas of intensive tillage, and increased erosion potential. The stable N pool, represented by TN, showed climate with the highest importance; whereas, the labile pool, based on BNA measures was best predicted and controlled by organisms, or plant cover. The successful inclusion of soil management layers into DSMs of N parameters indicated that the number of times in forages and potatoes over a 10-year period was of the greatest importance. As tillage intensity is most pronounced in potatoes, and as forages contribute to increased biomass and building OM levels, the results showed that increasing the number of years in forages had a positive correlation with GSN and the stable and labile N pools.

4.2 INTRODUCTION

The importance of reducing nitrogen (N) loss through informed N fertilizer management is well documented (Cassity-Duffey et al., 2020; Nyiraneza et al., 2010; Zebarth et al., 2009). Various studies have pointed to the need for quantifying both the stable and labile N pools in order to accurately predict N mineralization; and ultimately, to use this prediction for informing N fertilizer recommendations (Dessureault-Rompre et al., 2015; Heumann et al., 2013). The stable (i.e., slowly mineralizing) N pool is considered the larger pool consisting of mature, more recalcitrant, soil organic matter

(OM) stocks (Dessureault-Rompre et al., 2016). This pool has been observed to degrade according to a non-diminishing, zero-order kinetic relationship that can be parameterized using total soil nitrogen (TN) analysis (Dessureault-Rompre et al., 2013). The labile, or more quickly mineralizing, N pool is made up of more reactive portions of OM that can be parameterized using a two-week aerobic incubation method known as the N flush (Pool I), or commercially as the Biological Nitrogen Availability (BNA) test (Sharifi et al., 2008; Sharifi et al., 2007c). The labile N pool is small, relative to the stable pool and may be consumed within a growing season following a first-order kinetic relationship (Curtin and Campbell, 2008; Stanford and Smith, 1972). Direct soil measures of both TN (representing the stable N pool) and BNA (representing the labile N pool) may be used to parameterize a two-compartment kinetic model to predict the extent of soil N mineralization over a growing season (i.e., growing season N; GSN). Once complete, the point-prediction of mineralized N can be used to construct an N balance as a means of determining the need for supplemental N fertilizer addition, referred to as "right-rate" in 4R Nutrient Stewardship; an international program with the goal of improving N use efficiency (Johnston and Bruulsema, 2014).

Using indices of the stable and labile N pools for estimating N mineralization rates have various obstacles for producers including, the absence of data to make point predictions, spatially extrapolating these point-predictions throughout a landscape, and interpreting these relationships to assist soil management decisions. An absence of data is perhaps the most foundational since without it, a producer must rely on regional data, which may or may not be representative of site-specific soil conditions. With respect to spatially interpolating N mineralization within a landscape, due to the variability and

complexities of biological communities, extrapolating these point predictions can be a challenge (Zebarth et al., 2009). Biological communities are influenced by many soil factors (e.g., OM levels, texture, etc.), climate (e.g., temperature and moisture), vegetation or residues (e.g., crop selection, tillage, etc.), and topography (e.g., aspect, slope, etc.). As such, predictions of N mineralization between sample points requires an understanding of the relationships learned from temporal and spatial patterns in controlling variables. The spatial prediction of GSN mineralization represents an opportunity to apply these learnings toward more sustainable soil management practices. For the pedologist or agronomist, insight into whether distinct N pools (stable and labile) are more influenced by pedogenic processes, or by cropping decisions, could assist with crop selection and/or tillage practices that influence soil health over the long-term (Zebarth et al., 2009).

Soil management practices, including choice of crop rotation and tillage intensity, have a notable influence on N dynamics. Griffin (2008) stated that among the factors controlling N fraction sizes, management practices are ranked equally with other factors such as climate, biotic and abiotic factors (Dessureault-Rompre et al., 2015). Whittaker et al. (2023), studying the effects of forage mixes on yield, N cycling, and soil properties, found that the inclusion of forage legumes in crop rotations increased both soil quality and N supply but increased nitrate leaching compared to forage grasses. Forage legumes, in comparison to forage grasses, reduced the C:N ratio, which in turn increases both the quantity and quality of OM supply (Whittaker et al., 2023). Alternatively, the effects of tillage intensity inherent with potato production, have been found to reduce OM supply overall (Nyiraneza et al., 2017). As tillage practices increase in intensity, from zero-till,

no-till, minimum-till and conventional systems, there is a greater effect on mycorrhizae in soils, potentially mineralizable N, and soil structure (Kabir, 2005; Sharifi et al., 2008; Tivet et al., 2013). Therefore, it is important to quantitatively reflect the impact of soil management practices on N mineralization when making N management decisions.

Digital soil mapping (DSM), along with machine learning (ML) techniques, present an opportunity to address these challenges. DSM is a process by which direct soil data for a given soil attribute, or response variable of interest (e.g., TN, BNA, or GSN), are then coupled with environmental data (predictor variables) at the same spatial location, in order to recognize patterns and build numerical models (Heung et al., 2016; McBratney et al., 2003; Minasny and McBratney, 2016). Expanded from Jenny's five soil forming factors (Jenny, 1941), environmental predictor variables seek to quantify soil (s), climate (c), organisms (o), relief (r), parent material (p), age (a), and spatial position (n) plus autocorrelated residuals in what is collectively known as the *scorpan* factors (McBratney et al., 2003). Once the relationships between a soil attribute (e.g., BNA) and associated *scorpan* data (e.g., % slope, or mean average rainfall) have been modeled and learned, predictions can then be made in areas/pixels where direct soil data are absent. The DSM process begins by assembling covariates with complete coverage over the study area and that are representative of the *scorpan* factors. The number of covariate layers is simplified first with a process of variance inflation factor (VIF) analysis, which addresses multicollinearity among predictor variables (Craney and Surles, 2002; Deragon et al., 2024; Deragon et al., 2023; Paul et al., 2022; Saurette et al., 2023); and next, by a process of recursive feature elimination (RFE) where the most important predictors are identified in reference to the response variable (Deragon et al., 2023; Guyon et al., 2002;

Paul et al., 2022). Using model training and validation techniques to identify the best model, a prediction map can then be made of the study area. In some cases, where direct soil data is limited, pedotransfer functions (PTFs) can be used to fill in data gaps (Arbor et al., 2023; Laurence et al., 2023), thereby increasing the number of training data points, stabilize the DSM model, and potentially improve model accuracy and reduce uncertainty (Purushothaman et al., 2022; Reddy and Das, 2023). Another important aspect of the DSM and ML process, is the ability to interpret models in order to identify important predictors of response variables (Molnar et al., 2018). As such, with the tools unique to DSM procedures, the opportunity exists to make estimates of TN, BNA and GSN to inform N fertilizer recommendations, as well as provide insight into intrinsic or extrinsic controls for assisting soil management.

Zebarth et al. (2009) noted over a decade ago that spatial predictions of N parameters were indeed an opportunity for improving N fertilizer management. The literature provides multiple examples of spatial predictions of TN (representing the stable N pool) using DSM techniques (Mponela et al., 2020; Uygur et al., 2010; Zhou et al., 2019; Zhou et al., 2020); in particular, as relating to soil carbon coupled with TN to derive C:N ratios over a landscape (van der Westhuizen et al., 2023). On the other hand, seldom have there been DSMs of labile N pools such as Pool I (N flush) or BNA. This is likely due to a lack of direct soil information based on the relative expense of the 2-week aerobic incubation analysis. Also scarce in the literature, are DSMs that properly incorporate soil management practices into covariate layers for predicting spatial variation in N parameters.

While the influence of soil management on the supply of N is of great importance (Nyiraneza et al., 2022; Nyiraneza et al., 2012; Whittaker et al., 2023), its inclusion into a DSM framework, to the best of our knowledge, has yet to be accomplished. For mapping applications, covariate layers with complete coverage across the whole of the study area are necessary; and to date, a covariate layer depicting soil management in a definitive way has been elusive. Indicators of N cycling or plant N uptake, such as NDVI (Wang et al., 2018) or land use types per climatic zone (Zhou et al., 2020), have been attempted with predictive strength. Paul et al. (2022) used annual crop inventory (ACI) layers produced by Agriculture and Agri-Food Canada (AAFC) to generate a crop rotation map. In order to do this, ACI layers were grouped into four categories including annual crops, pasture-grassland-forage, perennial crops, and non-agricultural (Paul et al., 2022). While these layers showed importance, the use of broad categories (e.g., annual crops) did not provide a nuanced representation of soil management practices between categories. For example, tillage practices between potato when compared to cereal crops are markedly different in terms of tillage intensity but were included in the same category, both as annual crops.

The use of multi-year crop inventory (ACI) data has the potential to provide a more detailed assessment of the influence of cropping system by quantifying the frequency of a particular crop type over a particular span of years. The number of times a particular crop was planted at a given spatial location has both the potential to reflect the diversity of the crop rotation, and by assuming the typical soil management practices associated with each crop, the tillage intensity associated with the rotation. The ability to

incorporate soil management into a soil model or mapping framework represents a significant knowledge gap in the field of pedometrics.

For this study, DSM techniques were applied to the agricultural soils of PEI, Canada to understand the importance of soil and crop management in a mapping framework of soil N parameters and for the purpose of improving N fertilizer recommendations and soil management decisions. Included in this research, multi-year ACI crop frequency covariates were developed in order to predict N parameters, in particular growing season N mineralization and provide insight into cropping practices that promote sustainable soil management. Specific objectives included (i) determining the best *scorpan* factors for predicting N Pools (TN and BNA) and GSN while assessing the usefulness of novel crop frequency (ACI) management layers; (ii) developing DSMs of TN, BNA, and GSN; and (iii) interpreting the dominant controls for the stable and labile N pools to increase understanding of the impact of soil and crop management.

4.3 METHODOLOGY

4.3.1 Study Area

The study area focused on the 2,405 km^2 of soils classified as agricultural land use located throughout the 5,665 km^2 area of the province of PEI (PEI Department of Agriculture and Land, 2020). With respect to climate, the study area is classified as cool and humid with mean monthly temperatures ranging between -7°C and 19°C, and annual precipitation rates of 900 to 1000 mm (MacDougall et al., 1988). Vegetation is mixed with approximately 46% of the land base in forest/shrub and 51% in agricultural production with pasture/hay/grass, potatoes, and grains as the dominant crops. The remaining 3% of the study area consists of urban/bare soil and water (Jiang et al., 2015).

Landscape relief ranges from primarily gentle slopes (0 to 2%) dominating in the west, moderate slopes (2 to 4%) in the central to eastern regions, and hilly to hummocky surface expressions (4 to 8% and over) in the central and southeastern portions of the island (MacDougall et al., 1988). Parent materials consist of medium to coarse textured glacial till overlying sandstone bedrock. Dominant soils consist of the Podzolic, Luvisolic, Gleysolic, and Brunisolic soil orders in accordance with the Canadian System of Soil Classification (MacDougall et al., 1988; Nyiraneza et al., 2017; Soil Classification Working Group, 1998).

4.3.2 Soil Quality Monitoring Database

The PEI Soil Quality Monitoring database (SQMD), consisting of 675 georeferenced sample points at approximately 143 geographic regions (Figure 4.1A), was used for training and validation. The SQMD sample locations are based on a national forest inventory (4 km x 4 km) grid system with the potential for sample collection in a cluster pattern at the intersecting point on the grid (node) and the locations 100 meters north, south, east, and west of that point (Douglas et al., 2000). Soil samples were collected on agricultural fields only (n = 675). For reasons of access and land use, not all nodes were sampled at the centroid and four cardinal points but varied in sample intensity from one sample point to a maximum of five sample points collected per geographic cluster (Figure 4.1B). While the SQMD consisted of historical data from 1998 to present, the data used in this study were collected over a three-year cycle with the first set of samples collected in 2020 (n = 230), the second set in 2021 (n = 216), and the third set in 2022 (n = 229). At each location, samples were taken with an Edelman soil auger to a

maximum depth of 17 cm (Douglas et al., 2000; Nyiraneza et al., 2017) and submitted to the PEI Analytical Laboratory (PEIAL) for analysis.

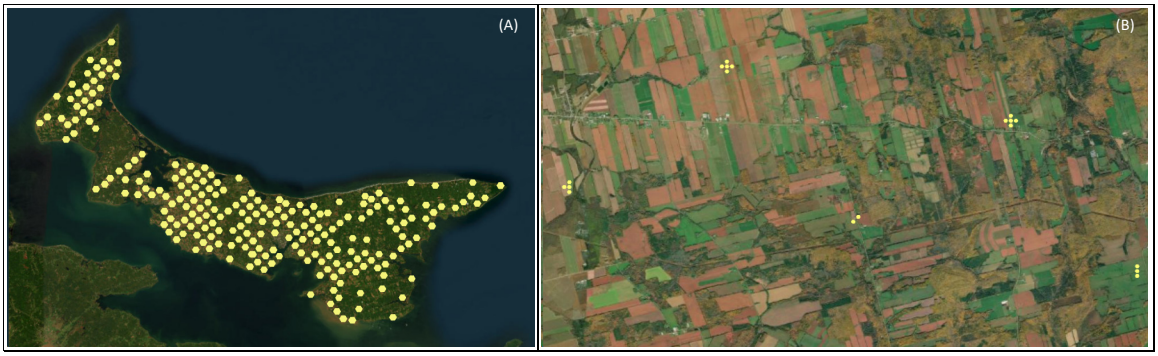


Figure 4.1 Maps of the study are showing clustered sample locations (A), and the variation in sample density within clusters (B).

Soil analytical parameters for this study included TN and BNA. TN (%) was determined using the LECO Method Report: Plants and Soils 10cc Loop, 4/16/2019, CN 828 S/N:20014 combustion procedure at 900°C (Marshall et al., 2021). BNA analysis, an aerobic incubation test adapted from Sharifi et al. (2007a) and detailed in Marshall et al. (2021), was conducted by first leaching initial concentrations of available ammonium (NH_4^+) and nitrate (NO_3^-) nitrogen with 200 mL of a 0.01 M $CaCl_2$ solution. Samples were then incubated for 14 days, and mineral N was leached again using 200 mL of a 0.01M $CaCl_2$ solution with results indicating the BNA (mg N/kg soil) and an estimate of N mineralization potential (Sharifi et al., 2007a). Summary statistics are presented in Table 4.1.

4.3.3 Growing Season N Mineralization Estimates

GSN was calculated using the regression equation (Eq. 4.1) from Dessureault-Rompre et al. (2015).

$$N_{\min} = k_S t + N_L\left[1 - e^{(-k_L t)}\right] \quad [4.1]$$

The convention of a 130-day growing season was used for the cumulative N mineralization ($N_{\min}$) at time (t) based on climate data for the Atlantic maritime ecozone (Gordon and Bootsma, 1993; Pedlar et al., 2015). The non-depleting (zero-order) stable fraction of the two-pool regression equation ($k_S t$) was calculated by estimating k_s (d^{-1}) from the relationship in Eq. 4.2 and SQMD results for TN (%) and BNA in mg N/kg soil (Dessureault-Rompre et al., 2015).

$$k_s = 0.123\ (TN) + 0.00312\ (BNA) + 0.0685 \quad [4.2]$$

The depleting (first-order) labile N pool, as $N_L(1 - e^{-k_L t})$ was calculated using BNA (N_L) and a constant value of 0.074 d^{-1} (k_L), as suggested by Dessureault-Rompre et al. (2015) for sandy loam textured soils. Summary statistics on GSN (kg N/ha) calculations are presented in Table 4.1.

Table 4.1 Summary of soil quality monitoring database parameters (n = 445) and summary statistics including the minimum (Min) value, 1st (25%) quartile, Median, Mean, 3rd (75%) quartile and the maximum (Max) value.

Parameter	Units	Min	1st	Median	Mean	3rd	Max
Total Nitrogen	%	0.052	0.11	0.14	0.14	0.17	0.32
Biological Nitrogen Availability	mg N/kg	0.00	27.7	35.6	38.3	45.9	120.2
Growing Season Nitrogen	Kg N/ha	54.2	130.2	157.6	167.0	193.7	440.3

4.3.4 Pedotransfer Functions

Since SQMD locations sampled in 2020 did not include TN or BNA analysis, PTFs were developed for the 2020 sample locations. OM and pH were selected as the predictor variables for development of PTFs for TN, BNA, and GSN, as these were the only appropriate parameters common between 2020 results, and the 2021 and 2022

SQMD datasets. In addition to using the SQMD 2021 and 2022 data (n = 445) for PTF development, to improve PTF model accuracy, the Soil Health Database (n = 2,222) from PEIAL was incorporated for a total of 2,667 samples available for model training (Chapter 3).

With two predictor variables available for PTF development, model simplification methods, such as recursive feature elimination were unnecessary. A suite of MLs were used to select the best model for predicting TN, BNA, and GSN. In order to select the best model, Lin's concordance correlation coefficient (CCC) was used as the primary accuracy metric (Lin, 1989). For PTFs of TN, support vector machines with radial basis function were selected (CCC = 0.76), while stochastic gradient boosting was selected for both BNA (CCC = 0.54) and GSN (CCC = 0.66) (Figure 4.2). The MLs are described in Section 4.3.7. PTF predictions for TN, BNA, and GSN were incorporated into the 2020 sample locations (n = 230) and used for model training purposes, but not validation due to the error associated with the PTFs.

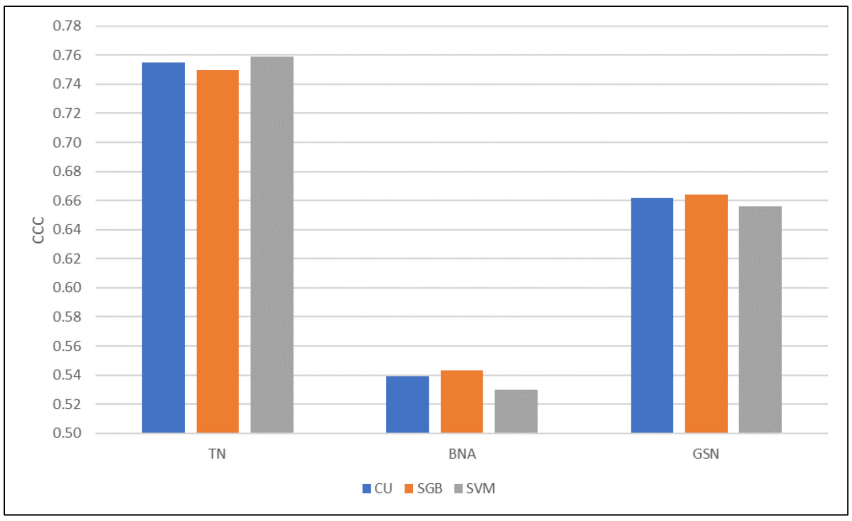


Figure 4.2 Concordance (CCC) results of pedotransfer functions (n = 2,667) for total nitrogen (TN), biological nitrogen availability (BNA), and growing season nitrogen (GSN) using the cubist (CU), stochastic gradient boosting (SGB), and support vector machine (SVM) models.

4.3.5 Environmental Covariates

Derived and processed from several sources, environmental covariates spanned the extent of the study area using a 30 m spatial resolution. A total of 85 covariates were considered for their potential applicability for predicting N pools and to reflect the appropriate *scorpan* factors including climate, organisms, and relief (Table 4.2). The *scorpan* layer for parent material (p) was not considered necessary for the study area due to its homogeneity across the study area. The spatial reference projection used for this study was the North American datum (NAD) 1983, Canadian spatial reference system (CSRS) universal transverse mercator (UTM) Zone 20N, European petroleum survey group (EPSG) spatial reference code 2961.

4.3.5.1 Climate variables

A set of 19 bioclimatic variables were obtained from the WorldClim dataset (Fick and Hijmans, 2017) in order to reflect bioclimatic processes related to soil N mineralization (Table 4.2). Each variable was derived from a 30-year average (1970-2000) representing major climatic factors such as temperature and precipitation trends, ranges, and seasonal extremes. Climate variables were resampled from the source spatial resolution of 30 arcseconds (approximately 643 m) to 30 m. Table 4.3 shows climate covariates retained after feature elimination procedures.

4.3.5.2 Organisms/vegetation variables

Four separate normalized difference vegetation index (NDVI) variables calculated from moderate resolution imaging spectroradiometer (MODIS; Didan, 2021) were included as a vegetation greenness indicator (Table 4.2). NDVI layers, obtained at a spatial resolution of 250 m and resampled to 30 m, consisted of median data over a span

of 10 years (2012-2021) collected throughout the growing season ranging from May to October. Covariate layers used for this study included NDVI maximum, mean, minimum, and range values over the 10-year period (Table 4.3).

Table 4.2 Environmental covariates considered for modelling N parameters and showing retained variables after variance inflation factor (VIF) analysis (threshold = 10).

Climate variables	VIF < 10
Annual mean temperature (Celsius)	
Annual precipitation	
Isothermality	✓
Max temp of warmest month	
Mean diurnal temperature range	
Mean temp of coldest quarter	
Mean temp of driest quarter	✓
Mean temp of warmest quarter	✓
Mean temp of wettest quarter	✓
Min temp of coldest month	✓
Precipitation of coldest quarter	✓
Precipitation of driest month	✓
Precipitation of driest quarter	
Precipitation of warmest quarter	✓
Precipitation of wettest month	
Precipitation of wettest quarter	
Precipitation seasonality (coefficient of variation)	✓
Temperature annual range	
Temperature seasonality (standard deviation x100)	
Organisms/vegetation variables	
Maximum NDVI value over a 10 year period (2012 - 2021)	✓
Mean NDVI value over a 10 year period (2012 - 2021)	
Minimum NDVI value over a 10 year period (2012 - 2021)	
Range of NDVI values over a 10 year period (2012 - 2021)	✓
Crop and soil management variables	
Berries	✓
Cereals	✓
Corn	✓
Fallow	✓
Grassland/Pasture/Forages	✓
N-fixing	✓
Oilseeds	✓
Potatoes	✓
Vegetables (other)	✓
Relief/topographic variables	
Catchment Area	✓
Convergence Index	✓
Dam Height	✓
Deviation from Mean Elevation, filter 3 m^2	✓
Deviation form Mean Elevation, filter 150 m^2	
Deviation from Mean Elevation, filter 800 m^2	
Deviation from Mean Elevation, filter 2000 m^2	

Relief/topographic variables (cont'd)	VIF < 10
Difference from Mean Elevation, filter 3 m	
Difference from Mean Elevation, filter 150 m^2	
Difference from Mean Elevation, filter 800 m^2	
Difference from Mean Elevation, filter 2000 m^2	
Eastness (Aspect)	✓
Elevation Percentile, filter 3 m^2	✓
Elevation Percentile, filter 150 m^2	✓
Elevation Percentile, filter 800 m^2	
Elevation Percentile, filter 2000 m^2	
General Curvature	✓
Maximum difference from mean elevation scaled, filter 150 m^2	✓
Maximum difference from mean elevation scaled, filter 800 m^2	✓
Maximum difference from mean elevation scaled, filter 2000 m^2	✓
Maximum difference from mean elevation, filter 150 m^2	✓
Maximum difference from mean elevation, filter 800 m^2	
Maximum difference from mean elevation, filter 2000 m^2	✓
Maximum Elevation Deviation scaled, filter 150 m^2	✓
Maximum Elevation Deviation scaled, filter 800 m^2	✓
Maximum Elevation Deviation scaled, filter 2000 m^2	✓
Maximum Elevation Deviation, filter 150 m^2	
Maximum Elevation Deviation, filter 800 m^2	✓
Maximum Elevation Deviation, filter 2000 m^2	
Mean Flooded Depth	✓
Mid-Slope Position	✓
Multiresolution Index of Ridge Top Flatness	✓
Multiresolution Index of Valley Bottom Flatness	✓
Multi-Scale Topographic Position Index	
Northness (Aspect)	✓
Plan Curvature	✓
Profile Curvature	✓
Sky View Factor	✓
Slope	✓
Slope Height	✓
Slope Length	✓
Slope Length Factor	✓
Standardized Height	
Stream Power Index	✓
Terrain Roughness Index	
Topographic Position Index	
Topographic Wetness Index	✓
Total Curvature	✓
Valley Depth	✓
Visibility	
Wetness Index	✓

Table 4.3 Accuracy metrics, selected covariates for *scorpan* factors, correlations for total nitrogen, biological nitrogen availability, and growing season nitrogen, and identifying top machine learners (CU = cubist; SGB = stochastic gradient boosting; SVM = support vector machines) and best model (*) for each response variable.

			Total Nitrogen (TN)				Biological Nitrogen Availability (BNA)				Growing Season Nitrogen (GSN)			
	Description	Abbreviation	Correlation	CU	SGB	SVM*	Correlation	CU	SGB*	SVM	Correlation	CU	SGB*	SVM
Metrics	Lin's concordance correlation coefficient	CCC		0.435	0.439	0.447		0.402	0.450	0.400		0.449	0.470	0.439
	Coefficient of determination	R^2		0.208	0.236	0.269		0.189	0.220	0.240		0.231	0.236	0.244
	Root mean square error	RMSE		0.035	0.034	0.033		13.7	14.6	12.9		43.9	44.0	43.0
Climate	Isothermality (mean diurnal range/temperature annual range) (X100)	C.iso	0.11	✓	✓		-0.04				-0.01			
	Mean temp of driest quarter	C.avgt.dq	0.07		✓		0.08	✓			0.08		✓	
	Mean temp of warmest quarter	C.avgt.waq	0.22	✓	✓	✓	-0.03				0.03			✓
	Mean temp of wettest quarter	C.avgt.weq	0.06	✓	✓	✓	0.17	✓	✓	✓	0.16	✓	✓	✓
	Min temp of coldest month	C.mint.cm	0.11	✓			0.17				0.17			✓
	Precipitation of coldest quarter	C.pcq	-0.01	✓	✓		0.12	✓			0.10			
	Precipitation of driest month	C.pdm	-0.21		✓		0.04				-0.02	✓		
	Precipitation of warmest quarter	C.pwq	-0.33	✓	✓	✓	-0.08	✓			-0.15	✓		
	Precipitation seasonality (coefficient of variation)	C.ps	-0.22	✓	✓	✓	0.03		✓	✓	-0.03	✓		✓
		Climate percentage of total covariates		39%	40%	50%		29%	17%	40%		31%	25%	31%
Organisms	Annual Crop Inventory - number of times in cereals from 2013 - 2022	O.ACI.ce	-0.21	✓			-0.20				-0.21			✓
	Annual Crop Inventory - number of times in grassland/pasture/forages from 2013 - 2022	O.ACI.gr	0.35	✓	✓	✓	0.40	✓	✓	✓	0.42	✓	✓	✓
	Annual Crop Inventory - number of times in potatoes from 2013 - 2022	O.ACI.po	-0.26	✓	✓	✓	-0.36	✓		✓	-0.36	✓		✓
	Maximum NDVI value from 2012 - 2021	O.NDVI.ma	0.12				0.25	✓	✓		0.24	✓		✓
	Range of NDVI values from 2012 - 2021	O.NDVI.ra	-0.09		✓		-0.22		✓	✓	-0.21		✓	✓
		Organisms percentage of total covariates		17%	15%	25%		21%	25%	60%		23%	25%	38%
Relief	Dam Height	R.dh	0.07	✓			0.02				0.03			
	Difference from Mean Elevation, filter 3 m^2	R.dme3	-0.02		✓		0.07				0.05			
	Deviation form Mean Elevation, filter 150 m^2	R.dme150	0.02	✓			0.16	✓			0.14	✓		
	Deviation form Mean Elevation, filter 2000 m^2	R.dme2000	0.10	✓	✓	✓	0.15	✓	✓		0.15	✓	✓	✓
	Eastness (Aspect)	R.aspe	0.09	✓			0.11		✓		0.12		✓	
	Elevation Percentile, filter 150 m^2	R.ep150	-0.03				0.14	✓			0.11			
	General Curvature	R.gcurv	-0.02		✓	✓	0.06				0.05			
	Maximum Elevation Deviation, filter 150 m^2	R.med150	-0.03	✓	✓		0.14	✓	✓		0.11	✓	✓	
	Maximum Elevation Deviation, filter 800 m^2	R.med800	0.01				0.14	✓			0.12	✓		
	Maximum Elevation Deviation, filter 2000 m^2	R.med2000	0.19				0.21	✓			0.22	✓		✓
	Mid-Slope Position	R.msp	-0.04		✓		-0.03				-0.03			
	Multiresolution Index of Valley Bottom Flatness	R.mivbf	-0.02	✓			-0.14				-0.12			
	Profile Curvature	R.pcurv	0.00		✓		0.11		✓		0.09		✓	
	Slope Height	R.sh	-0.04	✓			0.08				0.05			
	Slope Length Factor	R.slf	0.09		✓		0.08		✓		0.08			
	Topographic Wetness Index	R.twi	-0.08				-0.12		✓		-0.12			
	Total Curvature	R.tc	0.09		✓		0.03				0.05			✓
	Valley Depth	R.vd	0.14	✓	✓		0.05	✓	✓		0.07	✓		
	Wetness Index	R.wi	-0.08				-0.17				-0.16			✓
		Relief percentage of total covariates		44%	45%	25%		50%	58%	0%		46%	50%	31%
		Total coavariates required		18	20	8		14	12	5		13	8	13

4.3.5.3 Annual crop inventory and soil management variables

In order to capture soil management resulting from cropping practices, cropping history was characterized utilizing nine covariates extracted from the Annual Crop Inventory (ACI) produced by Agriculture and Agri-Food Canada (AAFC) at a 30 m spatial resolution (Table 4.2). The ACI data consisted of 72 possible classes including waterbodies, urban, and forested, as well as specific crop types at a target accuracy of 85% or greater (Fisette et al., 2014; Fisette et al., 2013). Using both optical and radar-based satellite imagery, ACI classification layers covering a 10-year span total (2013 to 2022) of the study area were grouped and re-classified into nine major cropping categories based on crop type and tillage intensity (Table 4.2). For example, cereal crops were considered separately from potato crops due to the differences in rooting biomass and tillage practices. All non-agricultural classifications (e.g., waterbodies or urban) were omitted because the SQMD includes only agricultural sample data. Once reclassified, raster layers representing the frequency for each cropping category for a 10-year period were generated for the study area. From this, a continuous covariate layer quantifying the frequency of each of the particular cropping categories was generated. ACI frequency covariates retained after feature elimination procedures are listed in Table 4.3. Figure 4.3 and 4.4 provide an example of novel ACI soil management covariates including the frequency in forages and potatoes, respectively.

4.3.5.4 Relief/topographic variables

Topographical metrics were generated using a 1 m spatial resolution digital elevation model (DEM) derived from light detection and ranging (LiDAR) data in 2020, which was provided by PEI's Department of Environment, Energy and Climate Change.

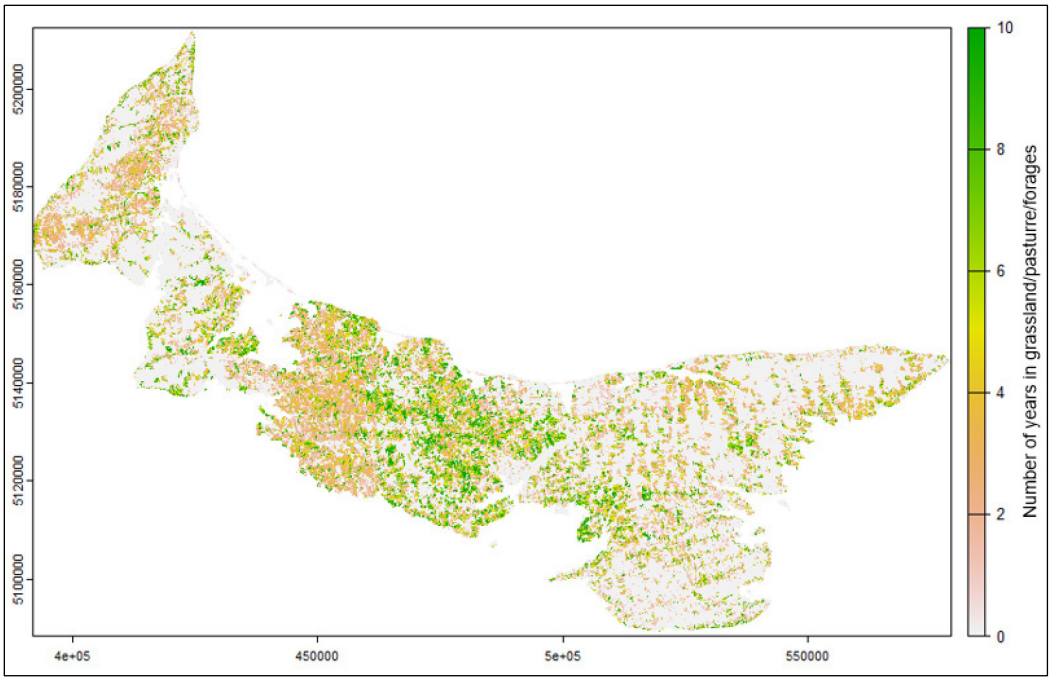


Figure 4.3 Multi-year annual crop inventory (ACI), frequency and soil management map showing the number of years that grassland/pasture/forages (O.ACI.gr) were recorded over a ten-year period (including 2013 to 2022).

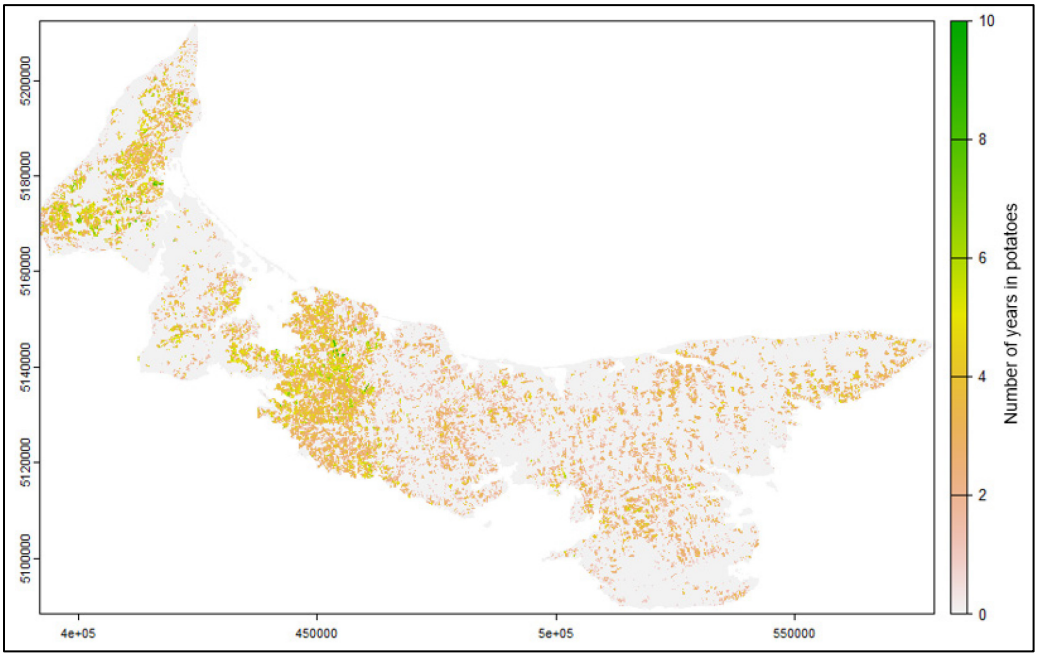


Figure 4.4 Multi-year annual crop inventory (ACI), frequency and soil management map showing the number of years that potatoes (O.ACI.po) were recorded over a ten-year period (including 2013 to 2022).

After smoothing with a mean filter window of 5 m x 5 m, and aggregating to a 30 m spatial resolution, the *RSAGA* (Brenning et al., 2018) and *whitebox* (Wu, 2021) packages were used within the R statistical software (R-CoreTeam, 2022) to generate 53 distinct covariates (Table 4.2). In addition to typical morphometry indicators such as slope or aspect etc., hydrological characteristics such as total curvature, valley depth, and topographic wetness indices were applied to the study area. Table 4.3 lists remaining relief covariates after feature elimination procedures.

4.3.6 Variance Inflation Factor Analysis

Feature elimination procedures began with consideration of multicollinearity among predictors; in particular, topographical covariates generated from a single DEM. Multicollinearity was addressed using variance inflation factor (VIF) analysis via the equation by Marquaridt (1970) and a value of 10 as the stopping threshold (James et al., 2013; O'brien, 2007). Using the *onsoilsurvey* package (Saurette, 2021) within the R statistical software (R-CoreTeam, 2022), ordinary least squares regression was performed starting with one predictor fitted against the other predictors without reference to dependent variables (e.g., TN or BNA). The VIF is calculated and the highest VIF score is removed, the VIF is recalculated, and the process goes on iteratively until all VIF < threshold. Those variables above the threshold value were removed. This process was performed until all variables were below VIF = 10 (Menard, 1995). Since no categorical covariates were included in this study, VIF was conducted for all predictors. After VIF analysis, 85 covariates were reduced to 54, with 31 being removed due to multicollinearity (Table 4.2).

4.3.7 Machine Learning

Initially, and to select the best three MLs for use in this study, five ML approaches were trialed including cubist (CU), random forest (RF), stochastic gradient boosting (SGB), support vector machines with radial basis function (SVM), and K-nearest neighbors (kNN). While many MLs can be used, these were selected for initial trials based on previous applications related to DSM and/or mapping of N based parameters (Deragon et al., 2023; Morellos et al., 2016; Parsaie et al., 2021; Shahbazi et al., 2019). All five MLs were used to model the relationships between each dependent variable (i.e., TN, BNA, and GSN) and the 54 remaining covariates (after VIF) for the 2021 and 2022 sample locations of the study area. The *caret* package (Kuhn, 2020) was used for all modeling procedures within the R statistical software (R Core Team, 2022). Based on the trial results, the top three MLs for all response variables were CU, SGB, and SVM.

The CU model has been used in various studies related to PTFs and DSM applications (Deragon et al., 2023; Mello et al., 2022; Paul et al., 2022). CU is a rule-based regression tree model that requires two hyperparameters for model optimization including the number of committees and neighbors. Further explanation of the CU model is found in Landré et al. (2018), Deragon et al. (2023), and in Chapter 3. In this study, a matrix of combinations was generated for both the committees (i.e., 1, 10, 50, 100) and neighbors (i.e., 0, 1, 5, 9) hyperparameters. The SGB model, described by Freund and Schapire (1997), Friedman (2002), and summarized in Chapter 3, is an ensemble technique based from classification tree analysis that has had multiple applications related to classification, PTFs, and DSM (Gebauer et al., 2020; Govil et al., 2022;

Hitziger and Ließ, 2014; Lamichhane et al., 2019). Parameter values used in this study included the interaction depth (i.e., 1, 3, 5, 7, 9), number of trees (i.e., 500), shrinkage (i.e., 0.1, 0.2, 0.3), and the minimum terminal node size (i.e., 10). Lastly, SVM works by optimizing the boundaries between data structures and variable classes in order to increase prediction abilities (Vapnik and Chapelle, 2000; Vapnik, 1999). The model is further described in Boser et al. (1992), Hastie et al. (2009), Heung et al. (2016) and in Chapter 3, with applications in both PTFs and DSMs (Gill et al., 2006; Lamichhane et al., 2019; Priori et al., 2014; Sedaghat et al., 2022). The radial basis function in the *caret* package was used (Kovačević et al., 2010; Kuhn, 2020) with sigma (i.e., 0.0001, 0.001, 0.01, 0.1, 1) and cost (i.e., 0.1, 1, 10, 100, 1000) as the hyperparameters. Selected MLs were used for PTF development and DSM modelling.

4.3.8 Mapping N pools and Growing Season Nitrogen

A methodological framework similar to Deragon et al. (2023) was used for this study. Summary statistics and modelling activities were performed with version 4.2.2 of the R statistical software (R Core Team, 2022).

4.3.8.1 Spatial cross-validation procedures

Due to the presence of clustered sample points in the SQMD (Figure 4.1), there was a likelihood of increased bias due to spatial auto-correlation (Pohjankukka et al., 2017). Preliminary trials confirmed this likelihood; thus, to generate realistic predictions, a leave-one-block-out (spatial) cross-validation (LOBOCV) procedure was performed (Deragon et al., 2023; Roberts et al., 2017). In addition, due to the inclusion of PTF predicted values for the 2020 locations, PTF predicted locations were not used for validation/testing but for model training purposes only (Román Dobarco et al., 2019a).

As such, training data included PTF and direct observations from 2020 to 2022 data (n = 675), while validation data included direct 2021 and 2022 sample results only (n = 445). For LOBOCV, the number of folds was determined by the number of clusters in the study area, which totaled 143. At each iteration, one block (i.e., cluster) was removed for testing from the validation data while the model was trained from the remaining training data (i.e., inner training loop) of 142 clusters. For the inner training loop, repeated 10-fold cross-validation was used with 20 repeats to optimize model hyperparameters (Ballabio et al., 2019). The model was then tested on the validation block, and prediction accuracy metrics were recorded (outer-loop). At the completion of each inner-loop, a new block was used for validation until each fold had been tested. Because the outer-loop generates an average CCC from the observed versus predicted values for each of the 143 folds, a standard-deviation of CCC could not be calculated as it was a pseudo-statistic. Spatial cross-validation was performed for TN, BNA, and GSN and for each ML.

4.3.8.2 Accuracy metrics

The primary model performance metric used for this study was Lin's concordance correlation coefficient (CCC). Preferred to the coefficient of determination (R^2) because it accounts for systematic under or over predictions, CCC can range between 1 and -1 with higher values showing better correlation between the observed values and the model predictions (Lin, 1989; Román Dobarco et al., 2019b). Secondarily, the root mean square error (RMSE) was used to demonstrate the average differences between the observed values and the model predictions.

4.3.8.3 Recursive feature elimination

Model simplification was done using recursive feature elimination (RFE), a backwards feature elimination process that is used to obtain a parsimonious model with the fewest, most relevant predictors, and with the highest CCC (Kuhn, 2020). Detailed in Chapter 3 and in Paul et al. (2022), RFE begins with all predictors, and at each iteration, removes the least important predictor based on variable importance analysis for each respective ML model (Guyon et al., 2002; Poggio et al., 2021). This process was conducted for each response variable (TN, BNA, GSN), using a model-agnostic process for calculating variable importance, approach to avoid model partiality, applied to each ML (CU, SGB, SVM) within the *caret* package (Kuhn, 2020).

Model training and validation accounting for spatial auto-correlation was performed according to the cross-validation procedures (LOBOCV) outlined in Section 4.3.8.1. Training data included the full dataset (n = 675), including 2020 PTF derived data points, while the validation data included only 2021 and 2022 direct soil observations (n = 445). The final model and ML were selected based on the highest CCC for each response variable. Once the final model and ML was selected, PTF data points were reintroduced to use the full dataset for both training and validation. To test for the possibility of increased or decreased accuracy from PTFs, RFE was conducted again for comparison with all response variables but without using PTFs for model training. Additionally, to test the effectiveness of novel ACI frequency layers, RFE was conducted a third time for comparison with all response variables but omitting ACI layers to observe the change in CCC. Once the best combination was determined (i.e., with or without

PTFs, and with or without ACI layers) the final model and ML were applied to the study area to obtain the final prediction map for TN, BNA, and GSN.

In order to identify the controls for the stable and labile N pools (TN and BNA, respectively) and GSN, interpretation of ML results was carried out using the *iml* package (Molnar et al., 2018) within the R statistical software (R-CoreTeam, 2022). Interpretation metrics included feature importance analysis, feature importance based on *scorpan* groups (i.e., climate, organisms, and relief), and accumulated local effects (ALE) plots to obtain correlations of each parameter with respect to the response variable (Molnar et al., 2018).

4.3.8.4 Uncertainty estimation

As observed by Deragon et al. (2023), uncertainty analysis is increasingly necessary when using DSMs in practical applications. As such, to estimate the uncertainty of N pool predictions for application in N fertilizer recommendations, a quantile regression (QR) approach first introduced by Koenker and Bassett (1978) and adapted by Kasraei et al. (2021) was performed. The required inputs for QR include the model residuals (i.e., the difference between the observed and predicted values) specific to each ML and response variable, and the associated predicted map. This feature adds flexibility, in that, by using the model residuals and predicted map generated from each ML, QR as a framework can be applied for each model, while maintaining model specific interpretations (Kasraei et al., 2021). In order to train the QR function, the *quantreg* package (Koenker, 2019) within the R statistical software (R-CoreTeam, 2022) was used with the model residuals and the predicted map for each ML (CU, SGB, SVM) and response variable (TN, BNA, GSN) as inputs. The 0.05 (5%, or lower prediction limit)

and 0.95 (95%, or upper prediction limit) quantiles were calculated, then subtracted from each other, and the result was applied to each cell of the predicted map in order to obtain the 90% prediction interval (PI). From this, three additional maps were generated including 5% lower limit, 95% upper limit, and from the difference, the 90% PI map of the study area. QR generated maps were then used to identify areas of higher or lower uncertainty, evidenced by areas of higher or lower PIs, respectively. In addition to uncertainty estimates, "eye-testing" was performed to assess the logic of spatial representations.

4.4 Results and Discussion

4.4.1 Feature Elimination and Model Performance

4.4.1.1 Variance inflation factor analysis

Beginning with a total of 85 covariate layers, 31 were removed due to multicollinearity (Table 4.2) using VIF analysis (63% decrease). The remaining 54 covariate layers below the stopping threshold (VIF = 10), included 9 climate layers reduced from 19 (53% reduction), 11 organism layers from an initial 13 (15% reduction), and 34 relief layers from 53 (36% reduction). The remaining covariates were used for RFE of each response variable and ML. The reduction in covariates compared to other studies that saw a 67% (Paul et al., 2022) or 58% (Deragon et al., 2023) reduction in covariates from VIF alone. Climate covariates calculated from single sources, such as temperature/precipitation climate models and relief covariates derived from a single DEM, are most susceptible to multicollinearity and were thus among the highest proportion removed (Mendonça-Santos et al., 2006).

4.4.1.2 Recursive feature elimination for Total Nitrogen

Comparing CU, SGB, and SVM to predict TN representing the stable N pool, the RFE process showed the best model to be SVM (CCC = 0.45) with a total of 8 covariates required for prediction (Figure 4.5, Table 4.3). The removal of irrelevant predictors using RFE resulted in a 20% increase in concordance when predicting TN with the inclusion of PTFs. To test for the possibility of confounding effects from PTFs, 2020 sample points were removed and the RFE process was repeated. The removal of PTFs resulted in an 8% reduction in CCC when compared to the best model (Figure 4.5). As such, PTFs were retained for use in final mapping procedures. In order to test the effectiveness of novel ACI frequency layers, these predictors were removed from the RFE process and the result was a 27% decrease in the CCC (Figure 4.5). As such, ACI frequency layers were retained for mapping procedures.

The relevance of predictor variables retained, and response of ACI layers as related to TN (representing the stable N pool) is discussed in Section 4.4.3.1. With respect to model performance, CCC was not calculated in other reported DSM studies of TN, so could not be directly compared. The RMSE for TN (0.033%, Table 4.3) showed favorable results when compared to similar studies (0.069%) in the literature (Hengl et al., 2015). Mapping TN, Mponela et al. (2020) had an R^2 of 0.14 using 3-fold cross validation, but 0.89 using out of the bag error, demonstrating the possible range based on validation techniques. Wang et al. (2013) obtained R^2 of 0.57 using geographically weighted regression and 0.68 using ordinary cokriging to predict TN. Results for R^2 have the potential for systematic bias, in that predictions may conform to the line of best fit, but deviate from the origin; as such, using CCC is preferred as it corrects for this by

assessing distance and closeness to the 1:1 line through the origin (Lin, 1989). Even with potential for bias, R^2 values for TN mapping were relatively modest, showing the inherent difficulty with predicting N parameters. Nevertheless, the CCC and the RMSE of this study showed comparable accuracy with the available literature (Uygur et al., 2010; Wang et al., 2017; Zhou et al., 2020).

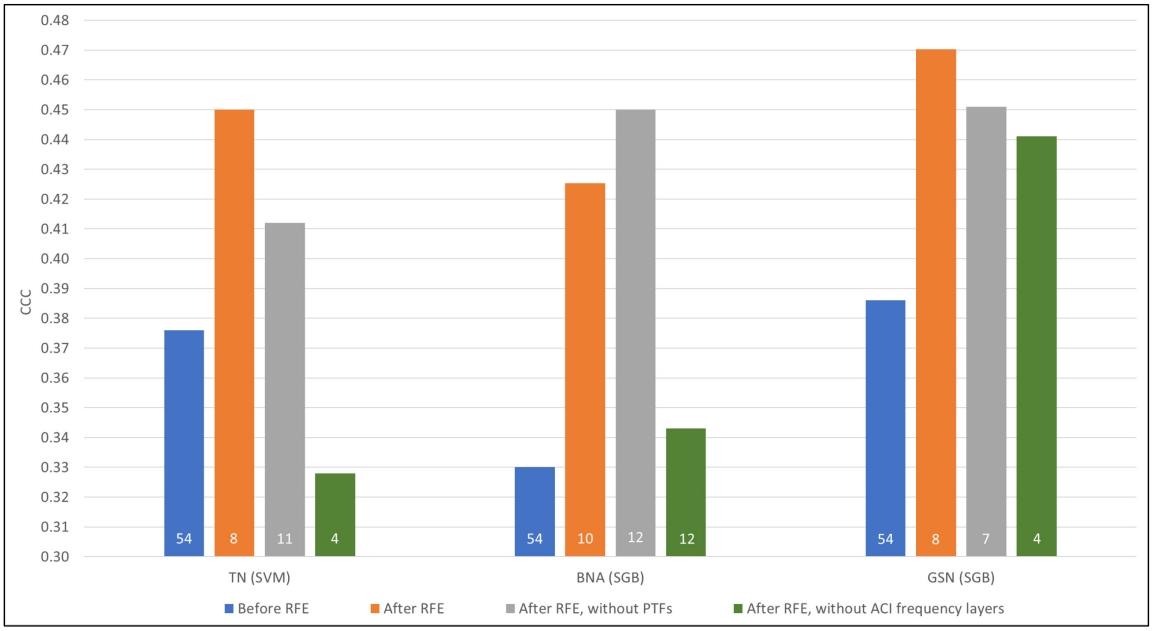


Figure 4.5 Feature elimination concordance (CCC) results for total nitrogen (TN) using the support vector machines (SVM) model, and biological nitrogen availability (BNA), and growing season nitrogen (GSN) using the stochastic gradient boosting (SGB) model with the number of predictors given for results before recursive feature elimination (RFE), after RFE with pedotransfer functions (PTFs), without PTFs, and without annual crop inventory (ACI) frequency layers.

4.4.1.3 Recursive feature elimination for Biological Nitrogen Availability

After RFE, the best ML for predicting BNA was SGB (CCC = 0.45) as compared with CU and SVM (Figure 4.5, Table 4.3). While RFE resulted in a strong increase in CCC using PTFs (29%), the increase was greater without the use of PTFs (36%), so PTFs were not used for final predictive mapping of BNA as representing the labile N pool

(Figure 4.5). The removal of ACI layers from the RFE process evidenced a 24% reduction as compared to the best model and were thus retained (see Section 4.4.3.2).

Regional DSMs using BNA as the soil attribute, as analyzed from 2-week aerobic incubation analysis (Pool I or N flush), could not be found in the literature and so results could not be directly compared. This absence of regional scale DSMs of BNA are likely due to the scarcity of geographically referenced databases containing this parameter. In comparison with results from the stable N pool, the same CCC of 0.45 for TN and BNA was achieved (Figure 4.5, Table 4.3) but required more predictors (8 vs. 12, respectively). The instability of BNA as representing the labile N pool, as the more active and quickly degrading N pool, likely contributes to the need of more covariates for prediction. Further, the variable nature of BNA is shown in that the assistance of PTFs for increasing model performance was not observed.

PTFs have been found to both increase and decrease model performance. Reddy and Das (2023), comparing DSMs derived with the inclusion, and exclusion of PTFs, found that PTFs reduced the total error. Román Dobarco et al. (2019a) on the other hand, saw an increase in the relative error (25% to 36%) with the inclusion of PTFs in soils of varying textural classes (very sandy or clayey soils). In this study area, with uniformly coarse textured soil (Laurence et al., 2023; Nyiraneza et al., 2017), the decrease in performance using PTFs is likely due to the relatively poor CCC (0.54) of the PTF itself (Figure 4.2). In comparison, the CCC for the PTF of TN was 0.76, and yielded more stable predictions for use in RFE. As such, the PTF for BNA with its lower CCC was not capable of yielding reliable predictions to assist RFE. This observation was also inferred by Román Dobarco et al. (2019a) who suggested updating results with more reliable

PTFs once training data became available. In terms of this study, the PTFs for all response variables (TN, BNA, and GSN) could only be derived using two predictor variables (OM and pH), thus as more parameters become available, PTF strength could be improved in the future. In the Chapter 3 study, CCC results ranging from 0.73 to 0.80 were possible with PTFs predicting BNA, using between four and nine predictors. It is assumed that using stronger PTFs for DSM procedures would contribute to better prediction potential of the BNA.

4.4.1.4 Recursive feature elimination for Growing Season Nitrogen

SGB showed the best CCC (0.47) after RFE procedures and was selected as the top model for GSN requiring 8 covariates (Figure 4.5, Table 4.3). CCC increased by 22% with the use of PTFs and 17% without PTFs; as a result, PTFs were used for final predictive mapping of GSN. ACI frequency layers, discussed in section 4.4.3.3, were also retained due to a drop in CCC of 6% with their removal.

Since DSMs of GSN or similar calculated N parameters were not available in the literature, comparison of RFE results could not be completed. Recalling that GSN is a calculated parameter (Eq. 4.1), derived using a two-compartment prediction function using TN and BNA, its higher CCC (0.47) than both TN and BNA (0.45) is notable. Interestingly, the combined effect of TN (representing the stable N pool) and BNA (representing the labile N pool) yielding a stronger CCC than the inputs alone show a predictive stability in the function (Eq. 4.1) itself. As observed in Chapter 3, PTFs of GSN consistently outperformed both TN and BNA showing that the inclusion of two parameters in the function added a model stability not achieved with the same parameters individually. The consistency of this finding, in PTFs using direct measures (Chapter 3),

and here with DSMs using indirect (e.g., remotely sensed data), increases confidence in the GSN output for practical use in N fertilizer recommendations (Laurence et al., 2023).

4.4.2 Spatial Representations of N pools and Growing Season Nitrogen

Spatial representations of the study area were conducted without masking non-agricultural lands (e.g., urban, water, forest, etc.) since these areas, having no soil samples collected, neither hindered nor helped overall accuracy and uncertainty (Nyiraneza et al., 2017).

4.4.2.1 Total Nitrogen

The final predictive map of TN, representing the stable N pool, and using SVM with 8 predictors (Table 4.3), had a range from 0.085 to 0.22% (Figure 4.6B), a mean of 0.14% and a standard deviation (SD) of 0.018% (Table 4.4). The lower prediction limit (5^{th} percentile) ranged from 0.072 to 0.11% TN (Figure 4.6A) and the upper prediction limit (95^{th} percentile) ranged from 0.14 to 0.28% TN (Figure 4.6C). The overall 90% prediction interval (PI) width ranged from 0.063 to 0.17% TN (Figure 4.6D) with a mean of 0.11% and a SD of 0.015% (Table 4.4).

Table 4.4 Descriptive statistics, including the minimum (Min), Mean, maximum (Max) values, and standard deviation (SD) for total nitrogen (TN), biological nitrogen availability (BNA), and growing season nitrogen (GSN) soil prediction and 90% prediction interval maps of the study area using support vector machines (SVM), and stochastic gradient boosting (SGB) learners.

Parameter	Units	Learner	Spatial Interpretation	Min	Mean	Max	SD
TN	%	SVM	Soil Prediction Map	0.085	0.14	0.22	0.018
			90% Prediction Interval Map	0.063	0.11	0.17	0.015
BNA	mg N/kg	SGB	Soil Prediction Map	-5.4	36.4	109.5	10.1
			90% Prediction Interval Map	19.5	41.0	78.5	5.2
GSN	kg N/ha	SGB	Soil Prediction Map	49.5	157.9	422.6	31.0
			90% Prediction Interval Map	43.1	126.0	328.3	23.7

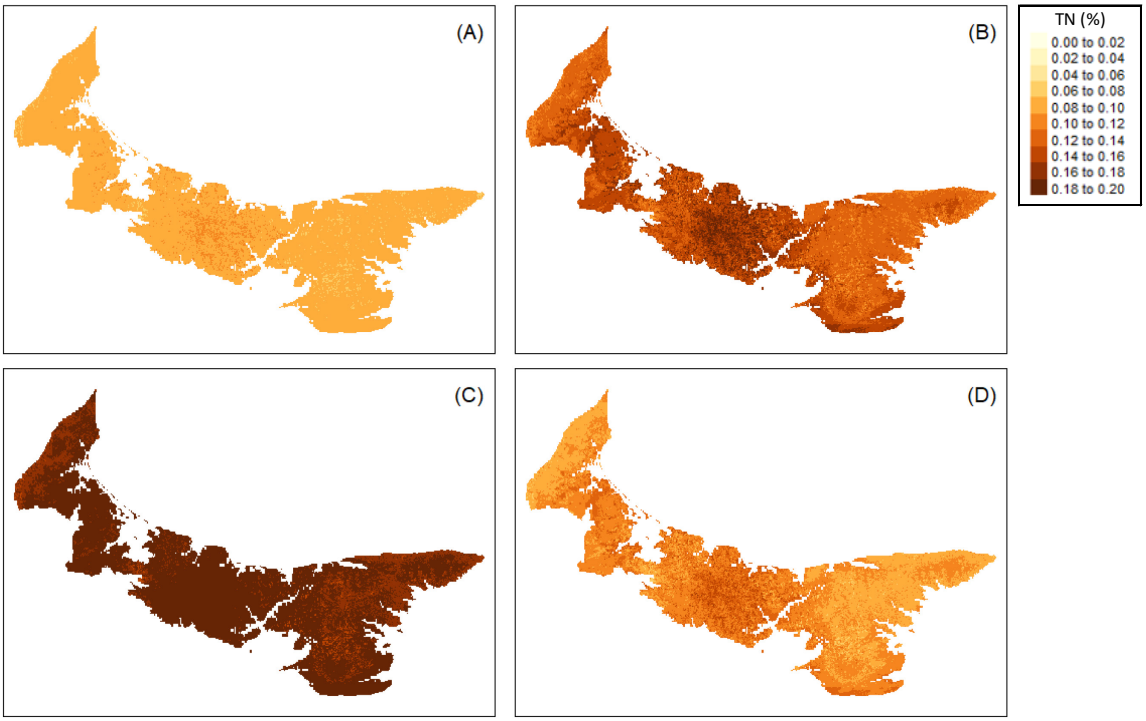


Figure 4.6 Soil total nitrogen (TN) maps (%) of the study area using the support vector machine (SVM) model and uncertainty maps using quantile regression. (A) Lower prediction limit map (5^{th} percentile), (B) prediction map, (C) upper prediction limit (95^{th} percentile), and (D) 90% prediction interval map.

Uncertainty estimation is relatively infrequent in DSMs of TN; however among the studies found, Zhou et al. (2020) assessed the uncertainty with results ranging from a SD of 0.05 to 0.14 g/kg with higher uncertainties of TN predictions in areas of low soil sampling density and complex surface structure. The issue of higher or lower sample density was not pronounced in this study due to the uniform grid system employed throughout the study area. Instead, higher uncertainty was evident in the central region of the study area (Figure 4.6D) where more intensive agricultural operations occur, as well as upper to upper-mid-slope locations where more erosion occurs. This is likely due to the atypical soil characteristics found with eroded soil conditions prevalent in these areas, as well as to the erosion related translocation of OM and associated nutrients from upper

to lower slope positions (Nyiraneza et al., 2017). In addition, it is also possible that these areas were proximate to livestock or confined animal production and therefore more land application of manure.

4.4.2.2 Biological Nitrogen Availability

Using SGB with a total of 12 predictors (Table 4.3), the final predictive map for BNA representing the labile N pool had a range from -5.4 to 109.5 mg N/kg (Figure 4.7B). The mean BNA value for the study area was 36.4 mg N/kg with a SD of 10.1 mg N/kg (Table 4.4). The lower prediction limit (Figure 4.7A) ranged from 76.4 to 165.4 mg N/kg, while the upper prediction limit (Figure 4.7C) had a range from 43.1 to 328.3 mg N/ha. The 90% PI map had a minimum interval of 19.5 mg N/kg, mean of 41 mg N/kg, maximum of 78.5 mg N/ha, and a SD of 5.2 mg N/ha (Figure 4.7D, Table 4.4).

Uncertainty estimates in regional representations of the labile N pool could not be found for comparison. However, comparing percent difference (max vs. min) of prediction intervals ranges (i.e., the greater the range, the greater the uncertainty) in PI maps between BNA and TN, BNA had a 12% greater interval range than TN (75% and 63% difference, respectively), but a proportional difference in SD (87% vs 86%, respectively). This indicates a similar model accuracy (Figure 4.5) but a greater uncertainty range in predicting the more labile BNA parameter. The labile N pool, as a smaller pool in comparison to the stable N pool (Dessureault-Rompre et al., 2013) and capable of being consumed within a growing season (Chapter 3), may contribute to the increased uncertainty observed in BNA. In addition, the labile fraction of OM is more sensitive to loss via cultivation, which may also increase prediction uncertainty (Six et al., 1999, 2002, 2020). Due to the coarse textured soils of the study area, erosion on PEI

is one of the major risks to OM and nutrient loss, including loss to N pools (Edwards et al., 1998; Nyiraneza et al., 2017).

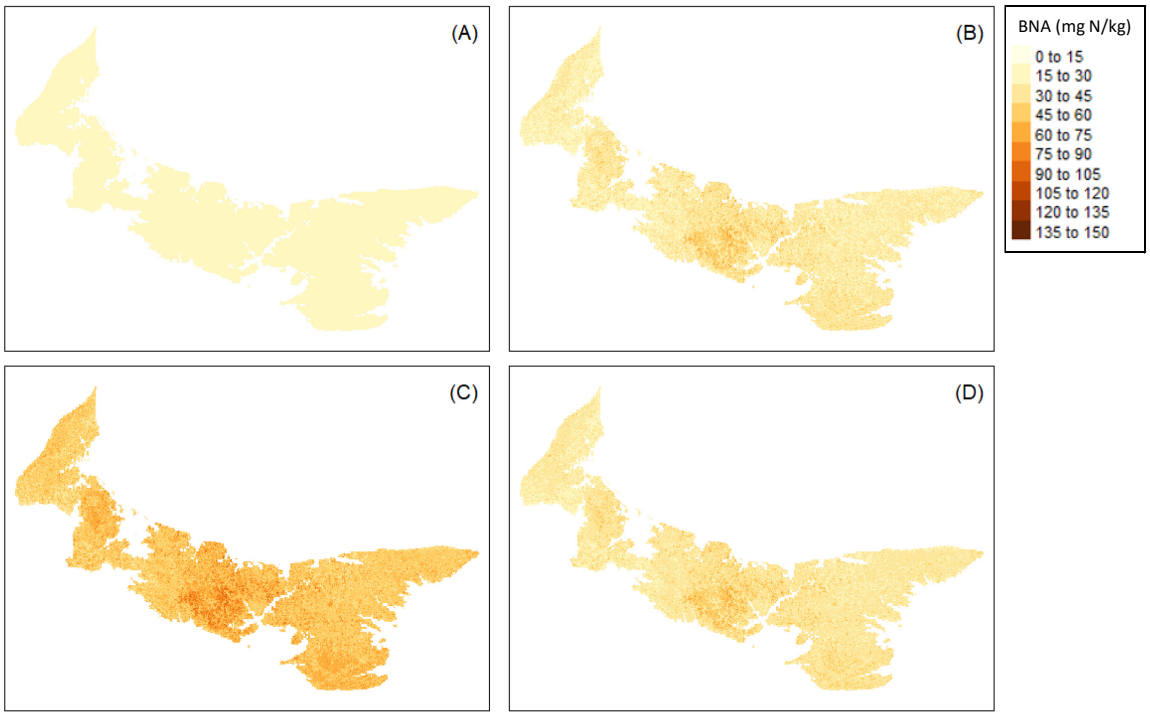


Figure 4.7 Biological nitrogen availability (BNA) maps (mg N/kg) of the study area using the stochastic gradient boosting (SGB) model and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

4.4.2.3 Growing Season Nitrogen

The predictive map of GSN, the calculated output based on TN and BNA (Eq. 4.1), was performed with 8 predictors using the SGB model (Figure 4.5, Table 4.3). The prediction range was between 49.5 to 422.6 kg N/ha (Figure 4.8B), with a mean of 157.9 kg N/ha and a SD of 31 kg N/ha (Table 4.4). Uncertainty maps showed the lower prediction limit (5th percentile) ranged from 76.4 to 165.4 kg N/ha GSN (Figure 4.8A) and the upper prediction limit (95th percentile) ranged from 119.5 to 493.7 kg N/ha GSN

(Figure 4.8C). The 90% PI ranged from 43.1 to 328.3 kg N/ha GSN (Figure 4.8D) and had a mean of 126 kg N/ha and a SD of 23.7 kg N/ha (Table 4.4).

Prediction uncertainty was greater with spatial estimates of GSN, as compared to both TN and BNA. The SD was proportionally less than TN and BNA, which was expected due to a higher prediction accuracy (Figure 4.5); yet, the PI range (between max and min) was proportionally greater (87% difference) as compared to TN (63% difference) and BNA (75% difference). Areas of greatest uncertainty appeared on upper slope regions which predominate the central and southeastern regions of the study area (Figure 4.8C). Upper slopes and knolls that experience high levels of erosion and translocation of OM and associated nutrients (Edwards et al., 1998; Six et al., 2002) likely contribute to the higher prediction uncertainty of GSN in these regions.

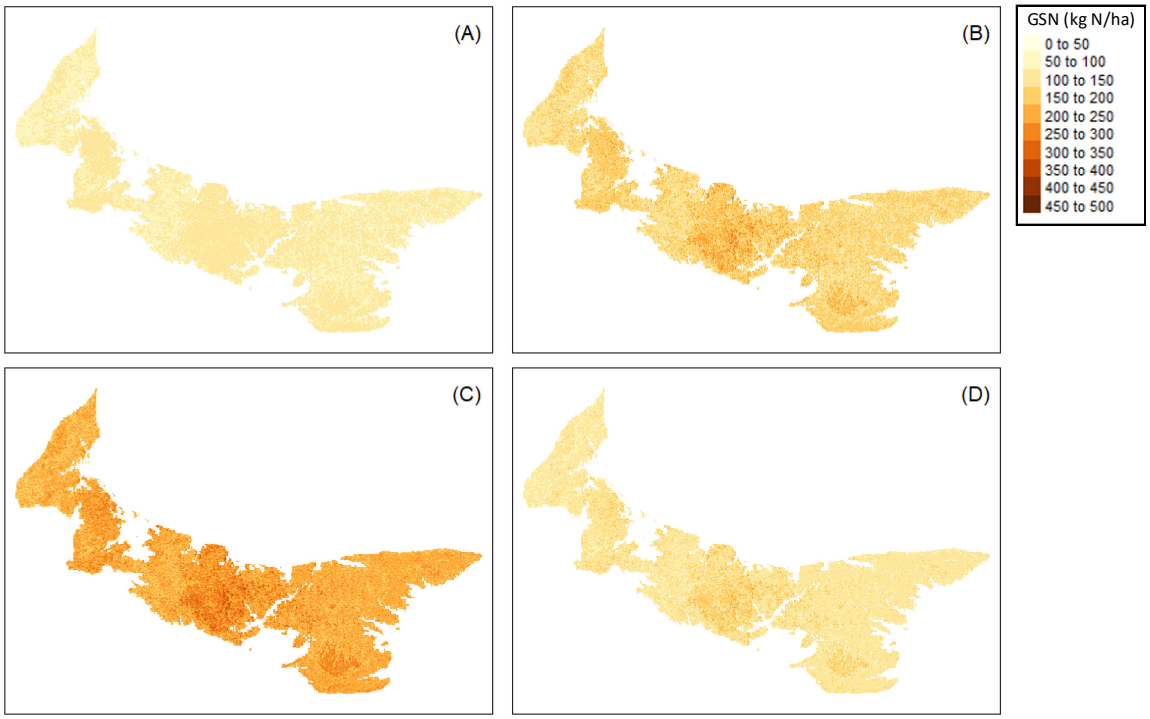


Figure 4.8 Estimated growing season nitrogen (GSN) maps (kg N/ha) of the study area using the stochastic gradient boosting (SGB) model and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

4.4.3 Interpretation of Predictors for N pools and Growing Season Nitrogen

4.4.3.1 Total Nitrogen

In order to assist the management of soil N stocks, the question of what factors control the stable N pool, estimated using TN measures, can be observed by first considering importance and ALE plots in Figure 4.9. Climate based parameters, including precipitation seasonality (C.ps) and mean temperature of the warmest quarter (C.avgt.waq), were the most important predictors selected by the SVM learner (Figure 4.9A). Consistent with studies related to the effect of climate on decomposition of OM and N pools (Dessureault-Rompré et al., 2010; Georgiou et al., 2022), the importance of climate for TN is also shown in Figure 4.9B, with climate variables collectively considered together as a group having the highest importance. Best correlation with climate variables (Table 4.3) and TN show precipitation of the warmest quarter (C.pwq) as the most important (-0.33). Interestingly, TN's negative correlation with precipitation was counter-intuitive as other DSM studies note the reverse, showing TN increase with increased precipitation (Wang et al., 2017, 2018). However, due to the predominantly coarse soil texture and relative absence of clay (mean = 12%) in the study area (Chapter 3), the decreasing TN with increasing precipitation can be understood. Amelung et al. (1998), as well as Georgiou et al. (2022), note the importance of clay for protection of carbon (C) and N stocks in soil and the resulting N loss in sandy soils with increased precipitation. Hence the negative correlation between TN and precipitation (Table 4.3) confirm the sensitivity of N stocks in humid climates with coarse textured soil, and the high potential for N leaching over winter and throughout the growing season (Nyiraneza et al., 2017; Sharifi et al., 2007c).

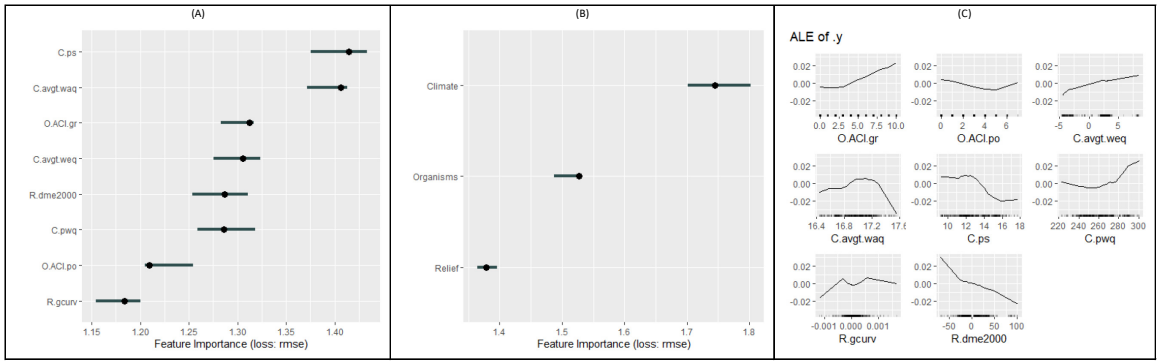


Figure 4.9 Variable importance plots (A), *scorpan* group importance plots (B), and accumulated local effects (ALE) of total nitrogen (TN = .y) with the support vector machines (SVM) model. Refer to Table 4.3 for a description of each covariate and abbreviation.

From Figure 4.9 and Table 4.3, after precipitation, temperature was the most important of the climate controls of TN followed by relief. This was consistent with Zhou et al. (2020) who observed that climate variables had the highest relative importance related to TN as compared to terrain, soil, and organism covariates. Temperature covariates selected by SVM (C.avg.waq and C.avg.weq) having a strongly positive correlation (Table 4.3), also displayed a reverse relationship to other spatial studies of TN that showed the expected decrease in TN with increasing temperature (Wang et al., 2017, 2018). The possitively correlated increase in TN with increasing temperature is likely a function of increased plant productivity increasing the C:N ratio and leading to increased immobilization of TN. Dessureault-Rompre et al. (2010), studying the relationship between climate and mineralizable N properties in soil, saw a possitive correlation with potential evapotransporation (a function of water content and temperature) and the larger theoretical Pool III, which suggests the more recalcitrant N pool may experience a net increase with increased temperature and plant biomass production. In addition, the positive correlation may highlight the greater role of biology, which increases with increased temperatures in soils, and retains TN via immobilization.

Organism/vegetation covariates important for modelling TN were second in importance to climate covariates (Table 4.3). Multi-year (ACI) crop frequency and soil management layers, indicating the number of times in grassland/pasture/forages (O.ACI.gr, Figure 4.3) showed high variable importance (Figure 4.7A). O.ACI.gr had the highest positive correlation (0.35) showing that TN increases with the number of times in grass species (Figure 4.9C, Table 4.3). Secondarily, the number of times in potatoes (O.ACI.po; -0.26) suggest that increasing frequency in potatoes yields a decrease in stable TN stocks (Figure 4.9C, Table 4.3). This result likely shows the result of tillage intensity, as potatoes require the highest rate of tillage of the crops considered, and forages require little to no tillage once the stand is established (Nyiraneza et al., 2017). Another factor contributing to decreased TN stocks is that potato production has a much lower crop residue input than forage systems. The importance of ACI frequency layers for predicting TN was also demonstrated by a 27% drop in CCC with their removal (Figure 4.5). The importance of organisms for predicting TN was also observed in other studies, but with relation to NDVI layers (Wang et al., 2018; Zhou et al., 2020). NDVI covariates had low relative correlations (Table 4.3) and were not selected by SVM for predicting TN. While ACI frequency covariates could not be compared directly, using landuse/land cover as a covariate has been used as a successful predictor with the highest relative importance for TN (Zhou et al., 2019). Of least importance in predicting TN as an estimate of the stable N pool, considered by group, were the relief covariates (Figure 4.9B, Table 4.3). Similar (Zhou et al., 2019), as well as conflicting results in the literature were observed (Wang et al., 2017). In the study by Wang et al. (2017), terrain indices were observed as the most important (elevation, and wetness index); however, it is

notable that the study measured TN to a depth of 100 cm in comparison to a maximum 17 cm depth with the current study area. TN predictions below the rooting-zone, with a lower influence on crop N availability in fertilizer recommendations, was not considered in this study.

4.4.3.2 Biological Nitrogen Availability

With respect to importance of predictors considered by *scorpan* group, BNA as an estimate of the labile N pool, exhibited a complete inversion of results as compared to TN (Figure 4.9B vs. Figure 4.10B) in showing relief as the most important, followed by organisms, and then climate. This alone is an important observation, which confirms the inherent differences and highlights varied controls of the stable and labile N pools. While most of the predictors for BNA were related to relief (58%), none of the relief variables selected by SGB had a particularly strong correlation (Table 4.3). The strongest correlations were observed from the organism *scorpan* group with the multi-year ACI frequency of forages (O.ACI.gr) layer (0.40), maximum NDVI (O.NDVI.ma) value (0.25) and range of NDVI (O.NDVI.ra) values (0.22; Table 4.3). NDVI layers were not selected by the top ML for TN, also highlighting the difference and importance of vegetation cover as a driver for the labile N fraction. Vegetative cover also reflects differences in the amount of crop and root residue input associated with various cropping systems, and as a result, would have a dramatic impact on BNA relative to TN. It is also notable that the top predictor for BNA (Figure 4.10A) was related to organisms' covariates (O.ACI.gr, Figure 4.3). This underscores the importance of crop frequency layers and soil management, as there was 24% decrease in CCC with the removal of ACI frequency layers as a whole (Figure 4.5). The labile N fraction, estimated with BNA or

the 2-week aerobic N flush (Pool I) analysis, has not been mapped regionally and therefore could not be compared; however, differences in controls between TN and BNA (known in the literature as the N Flush or Pool I) have been noted in various studies at the local (spatial) and discrete sample scales (Angst et al., 2022; Simard et al., 2001).

ALE plots (Figure 4.10C) and Table 4.3 show a similar trend with respect to BNA and the frequency of forages (O.ACI.gr) being positively correlated (0.40), and the frequency of potatoes being negatively correlated (-0.36). Based on this, it is evident that the effect of fibrous root crops adding biomass (Whittaker et al., 2023), and the relative absence of tillage with forages promote the growth of labile N stocks in similar fashion to the stable N pool (Kabir, 2005; Nyiraneza et al., 2010). Overall, while relief was a strong predictor, organisms and soil management were a key component in predicting BNA.

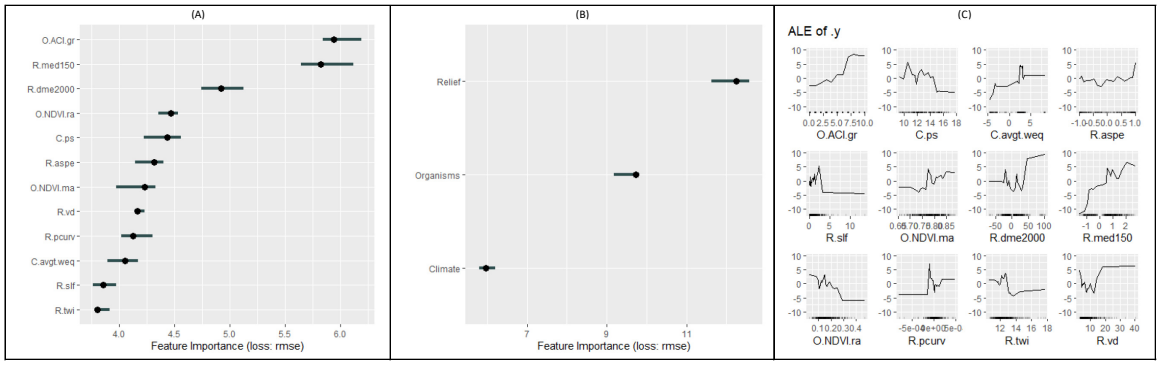


Figure 4.10 Variable importance plots (A), *scorpan* group importance plots (B), and accumulated local effects (ALE) of biological nitrogen availability (BNA = .y) with the stochastic gradient boosting (SGB) model. Refer to Table 4.3 for a description of each covariate and abbreviation.

4.4.3.3 Growing Season Nitrogen

The nature of the GSN output, being calculated from, and including the combined effects of both TN and BNA, presents a novel awareness of N supply in general. Based on the results from the SGB learner, GSN appears to mirror BNA more closely in regards

to variable specific and grouped importance plots (Figure 4.11A and B). The top predictor of GSN (from SGB), was the multi-year ACI frequency layer related to number of times in forages (O.ACI.gr, Figure 4.3), which also had a correlation of 0.42 - the strongest of all correlations across all parameters and learners (Figure 4.11A, Table 4.3). Similarly with BNA, GSN was most influenced by the relief *scorpan* group, followed by organisms, and then climate (Figure 4.10B vs. 11B). Also of similarity, relief variables were the most abundant of the SGB predictors (50%) but held the weakest correlations. The impact of relief on GSN, as well as BNA, demonstrates its foundational importance to N supply in general, but not necessarily the importance of one attribute in particular. The overarching importance of terrain has been observed for mapping N parameters in various studies (Zhou et al., 2019, 2020). However, as mapping scale changes from landscape to infield applications, it may become evident that specific relief attributes appear more important.

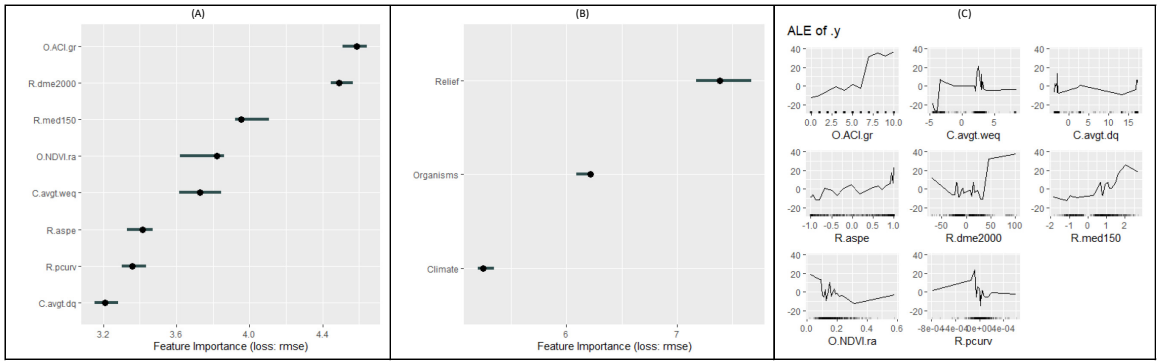


Figure 4.11 Variable importance plots (A), *scorpan* group importance plots (B), and accumulated local effects (ALE) of growing season nitrogen (GSN = .y) with the stochastic gradient boosting (SGB) model. Refer to Table 4.3 for a description of each covariate and abbreviation.

Organism covariates on the other hand have a more pronounced impact on GSN as shown by correlation of predictors (Table 4.3), and feature importance (Figure 4.11).

Unlike both TN and BNA, GSN showed the least reduction in CCC with the removal of ACI frequency layers (Figure 4.5) with a 6% reduction (versus 27% for TN, and 24% for BNA). This resilience with the removal of crop frequency layers is somewhat expected due to the collective strength of the GSN output in comparison to TN and BNA observed in Figure 4.5. Prior to RFE, with the confounding effects of using all predictors, GSN had the highest CCC compared to both TN and BNA. Also in the same figure, it is interesting that the highest predictive strength (CCC = 0.47) is derived from the GSN value. Lastly, the most parsimonious model also belongs to GSN, shown by the highest CCC with the fewest variables as compared to TN and BNA (Figure 4.5 and Table 4.3).

4.4.3.4 Soil management

Based on correlation, feature importance, and ALE plots, it is strongly suggested that TN (representing the stable N pool) is driven by climate; whereas, BNA (representing the labile N pool) is functionally driven by organisms. The best correlations for TN, and the most selected parameters across all learners shown in Table 4.3 (CU, SGB, SVM) relate to climate. On the other hand, the best correlations, and the most selected parameters for BNA, belong to the organisms *scorpan* factor. With a mainly pedogenically driven stable N pool, and a management driven labile N pool, the key to managing N stocks points towards strategies that promote OM development. OM is known to control soil climate via buffering heat exchange via the mineral component, and promote retention of moisture levels (Leirós et al., 1999; Sieber et al., 2022). Coarse texture soils, which predominate the study area and are prone to N loss via leaching, can be improved (in terms of protecting N pools) by increasing OM via increasing vegetative biomass and ground cover (Yang et al., 2020).

This understanding of the effect of forages, or inversely from the effect of tillage intensity, was seen via ACI frequency layers (Figure 4.3 and 4.4, respectively) as a predictor of N parameters. The positive correlation (Figures 4.9 to 4.11, Table 4.3) and the importance of including O.ACI.gr (Figure 4.3) in predictions speaks to the inclusion of forages (pasture or forage crops) in a crop-rotation cycle as one aspect of increasing biomass and building OM reserves (Nyiraneza et al., 2010; Whittaker et al., 2023). Also, observed was the negative correlation (Figures 4.9 to 4.11, Table 4.3) with frequency of potato crops and tillage intensity (Figure 4.4). As a root crop, potatoes contribute minimally to increasing soil biomass in comparison to fibrous roots indicative of forages (Stark and Porter, 2005). Furthermore, tillage, which decreases OM by aerating and heating the soil, has the increased potential to reduce N pools and soil health overall (Kabir, 2005; Roscoe and Buurman, 2003; Sharifi et al., 2008). In terms of soil management, the practitioner would do well to include forage crops, particularity legume-based forages, into rotations as well as minimizing tillage where possible (Whittaker et al., 2023).

4.5 CONCLUSION

Soil management covariates using multi-year, ACI crop frequency layers, were instrumental in mapping soil N parameters. Prediction accuracy of TN and BNA increased substantially with the inclusion of ACI frequency layers, and confirmed the importance of cropping rotation on mapping N pool dynamics. In addition, this study was conducted in order to model and map surrogate parameters for the stable N pool via TN, the labile N pool via BNA, and the calculated estimate of GSN in order to inform N fertilizer and soil health management decisions. Climate factors including precipitation

and temperature covariates were among the best predictors of TN followed by organisms and relief. Alternatively, BNA and GSN were best predicted with relief and organism covariates. In particular, multi-year and novel ACI frequency covariates showed high importance in predicting N parameters, especially the number of times in forages and the number of times in potatoes over a 10-year period. As indicative of biomass introduction and tillage intensity, this finding provides the opportunity to include soil management into a mapping framework, as well as insight into soil management decisions. In addition, highly erodible soil conditions in the study area were apparent from increased uncertainty of predictions on upper slope positions. For the purpose of informing N fertilizer recommendations, novel DSMs of TN, BNA, and GSN can be used by practitioners across the study area. While this study focused at the regional scale with a 30 m spatial resolution, future studies may consider infield N dynamics as the controls at finer scales are likely to be influenced by different soil forming factors. Furthermore, additional studies using multi-year crop frequency layers as a method for incorporating soil management in a soil mapping framework, is highly recommended due to their prediction importance.

CHAPTER 5: APPLYING PROVINCIALLY DERIVED PEDOTRANSFER FUNCTION AND SPATIAL ESTIMATES OF NITROGEN INDICES AT THE FIELD SCALE [3]

5.1 INTRODUCTION

In previous chapters, pedotransfer functions (PTFs) and digital soil maps (DSMs) were developed from province wide datasets in order to be used as decision support tools (DSTs) for informing nitrogen (N) fertilizer recommendations at a field scale. Informing N recommendations, via estimates of organic-N mineralized from soil over a growing season, is a fundamental component of completing N balances for quantifying N sources, versus N demand from the crop (Frerichs et al., 2022; Morvan et al., 2022; Zebarth et al., 2009) and is integral to a 4R Nutrient Stewardship (Johnston and Bruulsema, 2014). Growing Season Nitrogen (GSN) mineralization is estimated from a mixed-compartment kinetic model as described in Chapter 2 and in Dessureault-Rompre et al. (2015). The prediction function captures contributions from both the slowly mineralizing stable N pool, estimated from total nitrogen (TN) analysis, and the quickly mineralizing labile N pool, estimated from the 2-week aerobic incubation test commercially known as biological nitrogen availability (BNA) analysis. Frequently, due to perceived risk of yield loss, or from a lack of information on mineralized N sources, N fertilizers can be over applied, and lead to nitrous oxide (N_2O) losses through the growing season and/or high concentrations of residual soil nitrogen (RSN) remaining in the soil after harvest. As a consequence, high RSN levels can increase the potential for nitrate leaching or N_2O

[3] Chapter 5 was prepared as a thesis chapter intended for future publication. As a manuscript, this chapter would be considered as multi-authored (Laurence, L., Heung, B., MacDonald, E., Ramsay, M., and Burton, D.L.), in which the concept, design, data processing, and writing was done by the PhD Candidate with the assistance of all co-authors.

emissions (Zebarth et al., 2015). The implementation of DSTs (PTFs and DSMs) provides the opportunity to improve N management to sustain yield and minimize the potential for environmental impact. In this chapter we explore the potential for DSTs to be used in developing estimates of N parameters within a field, and supporting N fertilizer management.

In Chapter 3, PTFs for TN, BNA, and GSN were developed for the purpose of establishing a framework, determining the optimum predictor variables, and making use of existing data to estimate the contribution of mineralizable N in soil. While the study showed that accurate predictions were achievable, from a practical point of view, it also revealed that the best predictors were not among those historically and commonly available to producers. The standard soil analytical suite common to agricultural producers is called the S3-package and includes a limited set of parameters applicable for PTF development, including soil organic matter (OM), pH, and cation exchange capacity (CEC). As such, there is an opportunity to use the framework established in Chapter 3 to develop novel PTFs that could make use of existing datasets held by producers. In Chapter 4, provincial scale DSMs of N indices were made to assist N fertilizer recommendations in situations where producers had no historical data. At a 30 m spatial resolution, the provincial scale DSM was developed successfully using the PEI soil quality monitoring database (SQMD) to predict TN, BNA, and GSN. However, based on the effects of scale (landscape vs. field), and the disparity that may exist between provincial (SQMD) training data and on-farm soil quality (Malone et al., 2017), there is a need to observe how spatial predictions might perform at the field level. For example, slope position will likely have a greater impact on field-scale soil variability, and tillage-

intensive operations may have a lower soil quality as compared to the provincial (SQMD) benchmark (Lin et al., 2005). As such, potential issues between provincial scale predictions, infield soil results, and how they might be interpreted, is required for successful incorporation into N fertilizer recommendations.

The aspect of how a producer mechanically applies their fertilizer is also essential from a DST application and adoption perspective. Broadly speaking, a producer may apply N fertilizers as a single rate application (SRA) or as a variable rate application (VRA). The SRA, sometimes known as a blanket-rate application, occurs where a producer creates one N fertilizer rate and applies it to the whole field regardless of infield variations. While not ideal, this application approach is typically chosen when VRAs are not feasible to the producer. VRAs are performed when infield soil or nutrient variations have been delineated into zones, and on-farm technology (e.g., applicator equipment) is available to modify fertilizer rates during application passes (Zebarth et al., 2009). As such, if PTFs and DSMs are to be put into practice, they should meet producer needs by informing both SRA or VRA scenarios.

In order to address these issues, six producer fields were chosen that reflected the PEI standard 3-year potato rotation practices as described in Nyiraneza et al. (2017). Each was sampled to capture variability throughout the field, and analyzed for applicable soil quality and soil health parameters. With this infield data, there was an opportunity to compare observed N indices (i.e., sample results of TN, BNA, and GSN) versus provincial scale N predictions at specific sample locations. In addition, novel PTFs could be generated from a limited set of predictor variables (e.g., the S3-package) so that producers can make use of historical and widely available soil data. Lastly, in order to

meet the needs of producers who practice VRA, infield mapping of N indices provided the opportunity to apply a rich dataset of infield variation in N mineralization to inform N rate decisions. While there is some precedent for infield mapping of N parameters (Simard et al., 2001), infield mapping of N mineralization potential and predicted N mineralization over a growing season could not be found. Ultimately, DSTs that address producer needs, and provide a variety, or hierarchy of options to support N fertilizer management is imperative.

The specific objectives of this study included (i) a comparison of directly measured (observed) parameters collected from field specific data versus, predictions derived from a provincial scale database; (ii) development and assessment of potential for PTFs based on limited predictor variables; (iii) infield mapping of TN, BNA, and GSN and assessment of the dominant predictors; and (iv) providing a logical framework for incorporating regional and locally derived estimates of GSN to support N fertilizer recommendations.

5.2 METHODOLOGY

The methodological framework used in this study is shown in Figure 5.1. Modelling activities and summary statistics were performed with version 4.2.2 of the R statistical software (R Core Team, 2022).

5.2.1 Study Area

The study area consisted of six agricultural fields throughout PEI that were also part of a separate study entitled “Satellite-Derived Bare Soil Mapping in Prince Edward Island – A Potential Tool for Site Specific Field Management” (MacDonald, 2023). Field sizes ranged from 14.0 to 39.2 hectares (ha) with an average area of 23.1 ha and a total

area of 138.8 ha (Table 5.1). Study area locations were selected based on the following criteria: to identify fields where the anticipated soil N mineralization would be below average, on the basis of management practices, and by geographic location. With respect to management practices, select producers were asked to identify poorly performing fields that generally followed a 3-year rotation with potato occurring once in the 3-year rotation. F2 provides an exception with prolonged forages and cereals in rotation prior to potatoes (Table 5.1).

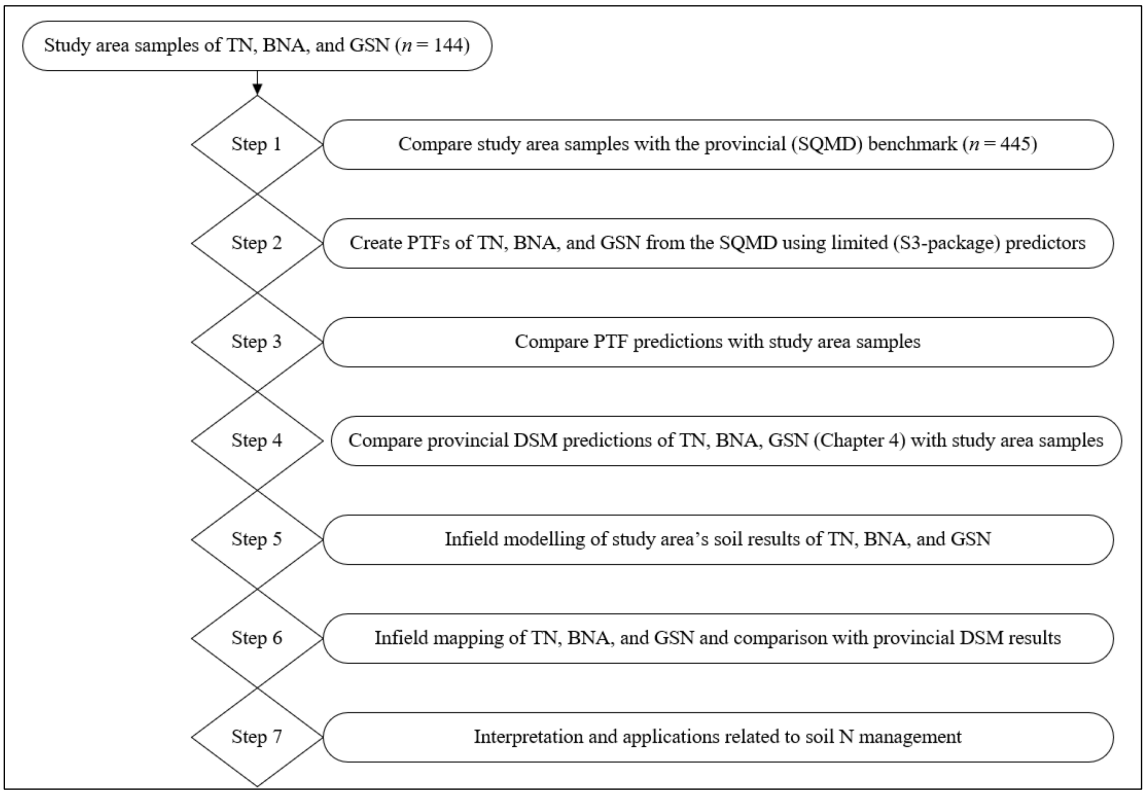


Figure 5.1 Methodological flow chart for field-scale application of pedotransfer functions (PTFs) and digital soil maps (DSMs) of total nitrogen (TN), biological nitrogen availability (BNA), and growing season nitrogen (GSN) that were derived from the provincial soil quality monitoring database (SQMD) benchmark.

The climate of PEI is classified as cool and humid with mean temperatures ranging between -7 and 19°C, and precipitation rates ranging between 900 and 1000 mm

annually (MacDougall et al., 1988). Agricultural crop rotations consist mainly of potato, grain, corn, and legume or forage crops on a 3-year rotation (Nyiraneza et al., 2017; Whittaker et al., 2023). With most of the agricultural production concentrated in the central region of PEI, four of the fields were located accordingly, while the remaining two fields were selected to include the western and eastern areas of the province (Figure 5.2, Table 5.1).

Table 5.1 Overview of the study area's six fields (F1 through F6) based on climatic location in Prince Edward Island, total area, and management considerations with rotations in order of progression ending with the most recent.

Field ID:	Location	Area (ha)	Irrigated	General three year rotation
F1	West	39.2	Yes	cereal - forage - potato
F2	Central	14.8	No	cereal - forage - forage
F3	Central	22.5	No	corn - potato - corn
F4	Central	14.0	No	cereal - forage - potato
F5	Central	17.9	No	cereal - forage - potato
F6	East	30.4	No	potato - cereal - forage

5.2.2 Soil Data

Samples were collected in May of 2023 prior to fertilization and planting using a soil auger to a maximum depth of 20 cm. The spatial sample design consisted of dividing each field into three zones based on soil colour and landscape position (upper, mid, lower) and collecting eight discrete georeferenced sample points per zone for a total of 24 samples per field. Once collected, samples were delivered to the Prince Edward Island Analytical Laboratory (PEIAL) for analysis using their standard soil fertility (S3) package and the soil health package.

Figure 5.2 Map of the study area showing the geographic location of fields F1 through F6.

TN and BNA were used as the principal soil analytical parameters for this study. TN was determined by combustion at 900°C using the LECO Method Report: Plants and Soils 10cc Loop, 4/16/2019, CN 828 S/N:20014 procedure. The average TN value, based on all soil analytical results (n = 144) was 0.12 % N (Table 5.2). BNA analysis are performed with a two-week aerobic incubation procedure adapted from Sharifi et al. (2007a). Detailed in Marshall et al. (2021), the procedure included premixing the soil sample (1:1) with inert Ottawa sand, placing in a Buchner funnel, and leaching available mineral N (NH_4^+ and NO_3^-) with 200 mL of a 0.01 M solution of $CaCl_2$. The soil:sand mixture was incubated for 14 days at 25ºC and leached again to obtain and estimate of potentially mineralizable N (mg N/kg soil) resulting from biological processes (Sharifi et al., 2007a). Based on the analytical results from this study (n = 144), the average BNA value was 18.12 mg N/kg (Table 5.2).

In addition to measured parameters of soil samples collected from the study area, georeferenced sample points of TN and BNA analysis collected in the spring of 2021 and 2022 from the provincial soil quality monitoring database (SQMD) were included (n = 445), and are summarized in Table 5.2. The SQMD was used in Chapter 4 to develop regional DSMs of TN, BNA, and GSN. The SQMD was applied in this study for comparison purposes, and to develop novel PTFs for testing on the study areas dataset.

Table 5.2 Summary of soil analytical data from the study area, considering all fields (AF) together and individually (F1 through F6), and the provincial benchmark soil quality monitoring database (SQMD), showing sample size (n) for total nitrogen (TN), biological nitrogen availability (BNA), and growing season nitrogen (GSN) with summary statistics including the minimum (Min) value, 1st (25%) quartile, Mean, 3rd (75%) quartile and the maximum (Max) value.

Area	n =	Parameter	Units	Min	1st	Mean	3rd	Max
AF	144	TN	%	0.05	0.10	0.12	0.14	0.20
		BNA	mg N/kg	5.1	14.2	18.1	21.1	59.8
		GSN	Kg N/ha	53.1	89.2	104.1	119.9	217.9
F1	24	TN	%	0.07	0.10	0.11	0.11	0.17
		BNA	mg N/kg	8.8	15.7	18.3	18.8	36.7
		GSN	Kg N/ha	69.5	90.0	100.6	102.3	171.8
F2	24	TN	%	0.14	0.16	0.17	0.18	0.20
		BNA	mg N/kg	16.6	20.3	22.2	24.0	27.3
		GSN	Kg N/ha	115.7	124.0	131.2	137.1	146.9
F3	24	TN	%	0.05	0.08	0.10	0.11	0.17
		BNA	mg N/kg	7.3	11.4	13.6	14.8	25.1
		GSN	Kg N/ha	53.1	74.5	84.9	91.9	139.3
F4	24	TN	%	0.05	0.09	0.11	0.12	0.14
		BNA	mg N/kg	5.8	12.9	15.3	18.5	22.7
		GSN	Kg N/ha	58.3	81.5	93.2	106.8	122.5
F5	24	TN	%	0.08	0.12	0.13	0.14	0.15
		BNA	mg N/kg	5.1	14.7	15.9	17.4	23.1
		GSN	Kg N/ha	73.7	94.3	100.0	106.2	126.7
F6	24	TN	%	0.08	0.10	0.11	0.11	0.12
		BNA	mg N/kg	13.1	18.7	23.4	24.9	59.8
		GSN	Kg N/ha	77.6	100.5	114.7	120.9	217.9
SQMD	445	TN	%	0.05	0.11	0.14	0.17	0.32
		BNA	mg N/kg	0.0	27.7	38.3	45.9	120.2
		GSN	Kg N/ha	54.2	130.2	167.0	193.7	440.3

5.2.3 Data Organization

With the objective of assessing, developing, and implementing practical tools for producer use in both SRAs and VRAs scenarios, and to obtain optimum modelling results, soil data from the study area was considered in a variety of ways. The "all-field" (AF) consideration (Table 5.2), combining the data from all six fields (n = 144), approached the study area as a whole, despite differences in field location. With this approach, modelling AF data together helped determine if predictions would be more accurate based on the larger (study area) dataset, and with similar soil management practices, as compared with models trained solely from field-specific (FS) data (n = 24). Models trained using AF data, and FS data modelling could then be applied to the six producer fields (F1 through F6) to map infield variation for farming practices that use VRA. Lastly, to consider farms without application equipment capable of applying variable rates, the "mean value per field" (MF) was considered to look at SRA approaches. MF modelling was performed by obtaining the mean value for all 24 samples (to mimic a composite sampling strategy), which provides one measurement to represent the entire field (effectively n = 1). For comparing MF soil results with provincial DSM predictions (Section 5.3.3), the arithmetic mean of all pixels within the field boundary was calculated.

5.2.4 Growing Season N Mineralization Estimates

A two-pool, first-order and second-order, regression equation (Eq. 5.1) was used to estimate GSN (Dessureault-Rompre et al., 2015).

$$N_{min} = k_S t + N_L\left[1 - e^{(-k_L t)}\right] \quad [5.1]$$

Cumulative N mineralization potential (N_{min}) over the growing season (t, d^{-1}) , estimated at 130 days in the Atlantic maritime ecozone (Gordon and Bootsma, 1993; Pedlar et al., 2015), was calculated using soil analytical results for TN (%) and BNA (mg N/kg). The zero-order regression equation ($k_S t$), representing the stable N pool, was calculated from the relationship in Eq. 5.2 described in Dessureault-Rompre et al. (2015):

$$k_s = 0.123\ (TN) + 0.00312\ (BNA) + 0.0685 \quad [5.2]$$

The labile N pool, estimated using the first-order regression equation $N_L(1 - e^{-k_L t})$, included BNA measures for N_L and a value of 0.074 d^{-1} (k_L) as suggested by Dessureault-Rompre et al. (2015) for sandy loam textured soils. The average GSN value, calculated from 144 sample results for TN and BNA, was 104.1 Kg N/ha (Table 5.2).

5.2.5 Pedotransfer Functions

Novel PTFs, different from those developed in Chapter 3 (Laurence et al., 2023), were required to reflect the standard agricultural-soil analytical package (S3-package) most commonly requested and historically available to PEI producers. The S3-package consisted of three soil-quality parameters conducive for PTF development, including OM, pH, and CEC. In addition to these parameters, there was an opportunity to include ACI frequency layers that were developed in Chapter 4 to increase predictive capabilities where possible. ACI frequency layers (explained in Section 5.2.6.5) were relevant only for AF and MF modelling scenarios since infield variation of crops is applicable for mixed-cropping (or multi-cropping) practices only, and was not practiced in the study area. After variance inflation factor analysis on ACI frequency layers (Section 5.2.7), preliminary trials were conducted to select ACI frequency layers with the highest variable

importance. As a result, the number of years in potatoes, cereals, and forages were identified as the best predictors. A limited process of recursive feature elimination (RFE), explained in Section 5.2.9.2, was then conducted. For RFE, the S3-package was considered as the minimum suite of variables due to its high variable importance. In addition to the S3-package, the frequency of potatoes, cereals, and grasses/forages over a 10-year period were included at the first iteration of RFE. Based on variable importance the least important ACI frequency layer was removed at each subsequent iteration, until the S3-package remained.

The SQMD (n = 444), having both response and predictor variables, was used with a suite of machine learners (Section 5.2.8), plus multiple linear regression (MLR) to obtain equation coefficients. For MLR, transformation of the data was not required due to the approximately normal distribution observed in probability density plots, and due to the comparable performance with machine learners (Section 5.2.8). Predictor variable summary statistics for the SQMD and the study area are given in Table 5.3.

Table 5.3 Summary of the soil quality monitoring database (SQMD) and study area response variable parameters with organic matter (OM), pH, cation exchange capacity (CEC), frequency of potatoes (ACIp), cereals (ACIc), and grasses (ACIg) in years per decade (yrs/dec), sample size (n), and summary statistics including the minimum (Min) value, 1st (25%) quartile, Mean, 3rd (75%) quartile, and the maximum (Max) value.

Source	n =	Parameter	Units	Min	1st	Mean	3rd	Max
SQMD	444	OM	%	1.2	2.5	2.9	3.3	7.0
		pH	NA	4.7	5.7	6.0	6.3	7.3
		CEC	Meq/100 g	3.0	8.0	9.7	11.3	21.0
		ACIp	yrs/dec	0	0	2	3	7
		ACIc	yrs/dec	0	1	2	3	6
		ACIg	yrs/dec	0	2	4	5	10
Study Area	124	OM	%	1.1	1.9	2.3	2.7	3.8
		pH	NA	5.2	5.9	6.2	6.5	7.3
		CEC	Meq/100 g	5.0	8.0	9.7	11.0	22.0
		ACIp	yrs/dec	2	3	4	4	7
		ACIc	yrs/dec	1	2	3	4	5
		ACIg	yrs/dec	0	1	2	3	6

5.2.6 Environmental Covariates

Covariates used for model training, and based on applicable *scorpan* factors (McBratney et al., 2003), were generated on a 5 m spatial resolution for field boundary extents of F1 to F6 in the study area. All covariate layers were transformed to the same projection (European petroleum survey group spatial reference code 2961). A total of 90 environmental covariates were considered for the study area (Table 5.4).

Table 5.4 Environmental covariates considered for modelling study area data points (n = 144) and showing retained variables after variance inflation factor (VIF) analysis (threshold = 10).

Soil variables	VIF < 10	**Relief/topographic variables (cont'd)**	VIF < 10
Soil color normalized / bare soil index	✓	Convergence Index	✓
Biological Nitrogen Availability	✓	Dam Height	
Growing Season Nitrogen	✓	Deviation from Mean Elevation, filter 3 m^2	✓
Total Nitrogen	✓	Deviation form Mean Elevation, filter 50 m^2	
		Deviation from Mean Elevation, filter 150 m^2	
Climate variables		Deviation from Mean Elevation, filter 500 m^2	
Annual mean temperature (Celsius)		Difference from Mean Elevation, filter 3 m	✓
Annual precipitation		Difference from Mean Elevation, filter 50 m^2	
Isothermality		Difference from Mean Elevation, filter 150. m^2	
Max temp of warmest month		Difference from Mean Elevation, filter 500 m^2	
Mean diurnal temperature range		Eastness (Aspect)	✓
Mean temp of coldest quarter		Elevation Percentile, filter 3 m^2	
Mean temp of driest quarter		Elevation Percentile, filter 50 m^2	✓
Mean temp of warmest quarter		Elevation Percentile, filter 150 m^2	
Mean temp of wettest quarter		Elevation Percentile, filter 500 m^2	
Min temp of coldest month	✓	General Curvature	
Precipitation of coldest quarter		Maximum difference from mean elevation scaled, filter 50 m^2	✓
Precipitation of driest month		Maximum difference from mean elevation scaled, filter 150 m^2	✓
Precipitation of driest quarter		Maximum difference from mean elevation scaled, filter 500m^2	✓
Precipitation of warmest quarter		Maximum difference from mean elevation, filter 50 m^2	
Precipitation of wettest month		Maximum difference from mean elevation, filter 150 m^2	
Precipitation of wettest quarter		Maximum difference from mean elevation, filter 500 m^2	
Precipitation seasonality (coefficient of variation)		Maximum Elevation Deviation scaled, filter 50 m^2	✓
Temperature annual range		Maximum Elevation Deviation scaled, filter 150 m^2	✓
Temperature seasonality (standard deviation x100)		Maximum Elevation Deviation scaled, filter 500 m^2	✓
		Maximum Elevation Deviation, filter 50 m^2	
Organisms/vegetation variables		Maximum Elevation Deviation, filter 150 m^2	
Growing season peak NDVI value: 2019	✓	Maximum Elevation Deviation, filter 500 m^2	
Growing season peak NDVI value: 2020	✓	Mean Flooded Depth	
Growing season peak NDVI value: 2021	✓	Mid-Slope Position	✓
Growing season peak NDVI value: 2022	✓	Multiresolution Index of Ridge Top Flatness	✓
Maximum NDVI value over a 10 year period (2012 - 2021)	✓	Multiresolution Index of Valley Bottom Flatness	✓
Mean NDVI value over a 10 year period (2012 - 2021)		Multi-Scale Topographic Position Index	
Minimum NDVI value over a 10 year period (2012 - 2021)		Northness (Aspect)	✓
Range of NDVI values over a 10 year period (2012 - 2021)	✓	Plan Curvature	
		Profile Curvature	✓
Crop and soil management variables		Sky View Factor	
Berries		Slope	✓
Cereals	✓	Slope Height	
Corn		Slope Length	✓
Fallow		Standardized Height	✓
Grassland/Pasture/Forages	✓	Stream Power Index	✓
N-fixing	✓	Terrain Roughness Index	
Oilseeds		Topographic Position Index (normalized)	✓
Potatoes		Topographic Wetness Index	✓
Vegetables (other)	✓	Total Curvature	✓
		Valley Depth	✓
Relief/topographic variables		Visibility	
Catchment Area		SAGA Wetness Index	✓

5.2.6.1 Bare-soil Index

To incorporate soil-based variables into modelling procedures, average soil colour values from a 5 m buffer around each sample location, and normalized on a scale from 0 (darker coloured) to 1 (lighter coloured), were generated from Plant Labs application program interface data at an original source resolution of 3 m. Resampled to 5 m, the soil colour normalized (S.SCN) layer was an average over four years (2018 to 2021) to account for seasonal anomalies. Soil colour was included due to the high correlation with soil organic carbon content (darker coloured) and conversely, depleted or eroded (lighter coloured) soils (Bartholomeus et al., 2011; Paul et al., 2020).

5.2.6.2 Provincial nitrogen prediction variables

Provincial scale predictive DSMs of TN (S.TNp), BNA (S.BNAp), and GSN (S.GSNp), generated at a 30 m spatial resolution were taken from Chapter 4 results. Provincial maps were tried for modelling to assess if provincial scale maps of TN, BNA, and GSN could be used to improve predictive strength at the field scale. Provincial DSMs were resampled to a 5 m spatial resolution using the nearest-neighbour method to avoid new values being generated through the interpolation process (Deragon et al., 2024).

5.2.6.3 Climate variables

Bioclimatic variables, including trends in temperature and precipitation, were only used in modelling of AF data together so that differences in climate across the province could be reflected. A total of 19 variables (Table 5.4) from the World-Clim 2.1 dataset were obtained based on a 30-year (1970-2000) average (Fick and Hijmans, 2017). The original resolution of 30 arcseconds (approximately 643 m) was resampled to 5 m using the nearest-neighbour approach. Infield variability of climate, as measured by the

World Clim dataset, was not appropriate for use infield and was thus not used for FS modelling.

5.2.6.4 Normalized difference vegetation index variables

As a vegetation greenness indicator, a total of eight normalized difference vegetation index (NDVI) variables were calculated. Four NDVI layers were derived from median data over a 10-year period (2012-2021) during the months of May to October in order to represent a growing season; including, the maximum, mean, minimum, and range. These images were taken from moderate resolution imaging spectroradiometer (MODIS) and resampled from an original spatial resolution of 250 m to 5 m for each individual field boundary (Didan, 2021).

An additional four NDVI layers were calculated based on the peak NDVI values of the previous four growing seasons from Sentinel-2 satellite imagery at an original spatial resolution of 10 m (Table 5.4). NDVI covariates for 2019, 2020, 2021 and 2022 were generated using the average index value from a 5 m buffer around each sampling point. Resampling to a 5 m spatial resolution was done using the nearest-neighbour resampling method (Deragon et al., 2024).

5.2.6.5 Annual crop inventory and soil management variables

Annual crop inventory (ACI) data, from optical and radar-based satellite imagery provided by Agriculture and Agri-Food Canada (AAFC), was used to generate 9 crop frequency covariates as described in Chapter 4. ACI cropping data, at a 30 m spatial resolution and spanning 10 years from 2013 to 2022, was reclassified into nine categories based on crop type (e.g., cereals vs. pulses) and tillage intensity (e.g., forages, vs. cereals, vs. root crops). A “value raster” layer was created for each crop layer, which assigned a

number value to the crop categories. Each value raster layer was applied to each year of the ACI cropping data layers to quantify the frequency/number of times a particular category was planted over the 10-year period (Table 5.4). Crop frequency layers were used in AF modelling only in order to account for crop management practices between fields; as such, these layers were not included for FS modelling.

5.2.6.6 Topographic Variables

Topographical variables were aggregated to 5 m from a 1 m spatial resolution light detection and ranging (LiDAR) digital elevation model (DEM) provided by the PEI Department of Environment, Energy and Climate Change. The DEM was smoothed with a mean filter window of 5 m x 5 m, and the *RSAGA* (Brenning et al., 2018) and *whitebox* (Wu, 2021) packages within the R statistical software (R-CoreTeam, 2022) were used to generate 50 topographic covariates. Two additional topographic covariates, including slope and topographic position index, were generated to reflect landscape features surrounding sample locations by averaging values within a 5 m buffer at each point (Table 5.4).

5.2.7 Variance Inflation Factor Analysis

Prior to modelling, the assembled 90 covariates were assessed for multicollinearity using variance inflation factor (VIF) analysis with a stopping threshold of 10 (James et al., 2013; Marquaridt, 1970; O'brien, 2007). Using the R statistical software (R-CoreTeam, 2022) in addition to the *onsoilsurvey* package (Saurette, 2021), ordinary least squares regression was performed iteratively between all variables without reference to TN, BNA, or GSN. After VIF analysis, a total of 40 covariates were retained (Table 5.4).

5.2.8 Modelling Approaches

Modelling the relationships between TN, BNA, and GSN and predictor variables was done by comparing three machine learners (MLs) commonly used for both PTFs, and regional or infield DSM approaches (Deragon et al., 2024). For all modelling procedures, the *caret* package (Kuhn, 2020) within the R statistical software (R Core Team, 2022) was used. Cubist (CU), random forest (RF), and support vector machines with radial basis function (SVM) were compared for each response variable and the top performing model was selected for final predictions.

The CU model, using a rule and tree-based structure (Kuhn and Johnson, 2013; Quinlan, 1992), is applicable for interpreting linear and non-linear relationships as is common with N parameters (Clingensmith and Grunwald, 2022; Kim et al., 2013). Hyperparameters in the CU model include the number of committees (1, 10, 50, 100) that will indicate the number of trees to be aggregated, and the number of neighbours (0, 1, 5, 9) to identify and assist prediction in relation to "neighbour" tree nodes in the training dataset (Deragon et al., 2023; Landré et al., 2018; Mello et al., 2022). Similarly, RF is a tree-based model that uses a non-parametric ensemble technique wherein predictions from decision trees are compared and tested against other uncorrelated trees to obtain optimum (least biased) predictions (Breiman, 2001; Heung et al., 2016). Accuracy of predictions is optimized with the m_{try} hyperparameter and the number of trees which was set at 500 (Kuhn, 2020). The hyperparameter m_{try} was set based on the number of predictors in sequence (by one's) from 1 to *n* predictors. The SVM learner is a method that attempts to optimize boundaries (vectors) between set structures and classes of training data to increase prediction capabilities (Qin et al., 2022). Based on Kovačević et

al. (2010) and Priori et al. (2014), the hyperparameters of sigma (i.e., 0.0001, 0.001, 0.01, 0.1, 1) and cost (i.e., 0.1, 1, 10, 100, 1000) were used with the radial basis function in the *caret* package (Kuhn, 2020).

For PTF applications, and in addition to the MLs described, MLR was included in order to obtain model coefficients (i.e., equation based PTFs) that are capable of being used by producers without the source dataset. MLR is a common regression technique for quantifying the relationships between a response variable and multiple predictor variables (Padarian et al., 2018; Tabachnick et al., 2013). Transformation of the data via squaring, log-transforming or rooted predictors is often considered in non-normal distributions (Schillaci et al., 2021; Wösten et al., 1999); however, the training data had an approximately normal distribution and displayed comparable performance with machine learners. As such, data transformation was not required for MLR application.

5.2.9 Modelling N pools and Growing Season Nitrogen

Summary statistics and modelling activities were conducted using version 4.2.2 of the R statistical software (R-CoreTeam, 2022). As in Chapter's 3 and 4, model evaluation included the concordance correlation coefficient (CCC) from Lin (1989) as the primary accuracy metric, followed by the coefficient of determination (R^2), and the root mean square error (RMSE). Modelling of N pools and GSN was performed for both novel PTF and DSM development.

5.2.9.1 Cross-validation procedures

For sample point-specific (non-spatial) modelling associated with PTF development, a repeated 10-fold cross-validation procedure was performed with 50 repeats (Ballabio et al., 2019; Laurence et al., 2023). For spatial applications, two

separate methods of spatial cross-validation were required for understanding if field predictions were best conducted based on training from a larger management-specific dataset (i.e., AF) or a smaller field specific (i.e., FS) dataset (see Section 5.2.3).

For model training based on AF data, which implicitly includes a cluster of 24 samples per field, a leave-one-field-out cross-validation (LFOCV) procedure was conducted in similarity to the leave-one-block-out spatial cross-validation in Chapter 4 (Deragon et al., 2023; Roberts et al., 2017). In this method, and to overcome spatial auto-correlation (Pohjankukka et al., 2017), six folds (one per field) were created, and at each iteration: a field was removed, the model was trained from the 5 remaining fields, and the model validated with observations from the one field left out. The model training (with 5 fields) at each iteration, called the "inner-loop", was done using repeated 10-fold cross validation with 20 repeats (Ballabio et al., 2019). After each iteration of the inner-loop, model predictions were validated against the training field (in the "outer loop") and accuracy metrics were recorded. After each outer loop, a new field was put aside as the validation fold and the model was trained again with the remaining five fields. The LFOCV process, conducted for TN, BNA, and GSN as the response variables, was repeated until all six fields had been used for validation.

For modelling based on FS data, and with a comparatively low number of sampling points per field (n = 24), a leave-one-out cross-validation (LOOCV) procedure was selected. LOOCV was chosen in preference to the repeated (k-fold) cross-validation method due to the low sample size per field (Deragon et al., 2023). In this method, one observation per iteration is removed during model training and used for validation. The number of iterations is equal to the number of sample points and the procedure is carried

out until all sample points have been used for testing/validation. As such, the LOOCV method negates the need for multiple repetitions.

5.2.9.2 Feature elimination

For PTF and DSM procedures, a process of RFE was used to select the fewest, most important predictors for each response variable (TN, BNA, and GSN). The RFE process was performed using a model-agnostic approach via the *caret* package (Kuhn, 2020) allowing each ML to select important predictors specific to the learner. In general, the RFE process is considered a "backwards" elimination process in that all covariates are included in the first iteration, and after identification of the least important variable, it is removed and modelled again until the fewest variables are achieved with the highest CCC (Paul et al., 2022). RFE was performed for mapping using the LFOCV procedure for AF modelling of response variables, and with the LOOCV method, using the *caretFuncs* function, for FS modelling of response variables. After selection of the most important variables, model interpretation was conducted using the *iml* package (Molnar et al., 2018) within the R statistical software (R-CoreTeam, 2022). Interpretation metrics were done with feature importance analysis, considered by individual feature (covariate) and based on *scorpan* groups (i.e., climate, organisms, and relief), and accumulated local effects (ALE) plots in order to obtain correlations with respect to the response variable (Molnar et al., 2018).

5.2.9.3 Infield mapping and uncertainty estimates

For each of the six fields of the study area, and for each response variable (TN, BNA, and GSN), a prediction map trained from both AF data and FS data was generated for a total of 36 N prediction maps. Uncertainty estimates were also generated for each

prediction map using model residuals with the *quantreg* package (Koenker, 2019) within the R statistical software (R-CoreTeam, 2022). Uncertainty was performed using the quantile regression (QR) approach (Kasraei et al., 2021; Koenker and Bassett, 1978) wherein the lower (5% quantile) and upper (95% quantile) prediction limits were calculated and subtracted to provide a 90% prediction interval (PI) for each cell (pixel) of the predicted map. As a result, three maps (5% lower limit, 95% upper limit, and 90% PI) were produced in connection to each prediction map for a total of 108 uncertainty maps. In addition to identifying situations of higher uncertainty for each field, "eye-testing" was performed to check conformance of spatial representations to known field conditions.

5.3 RESULTS AND DISCUSSION

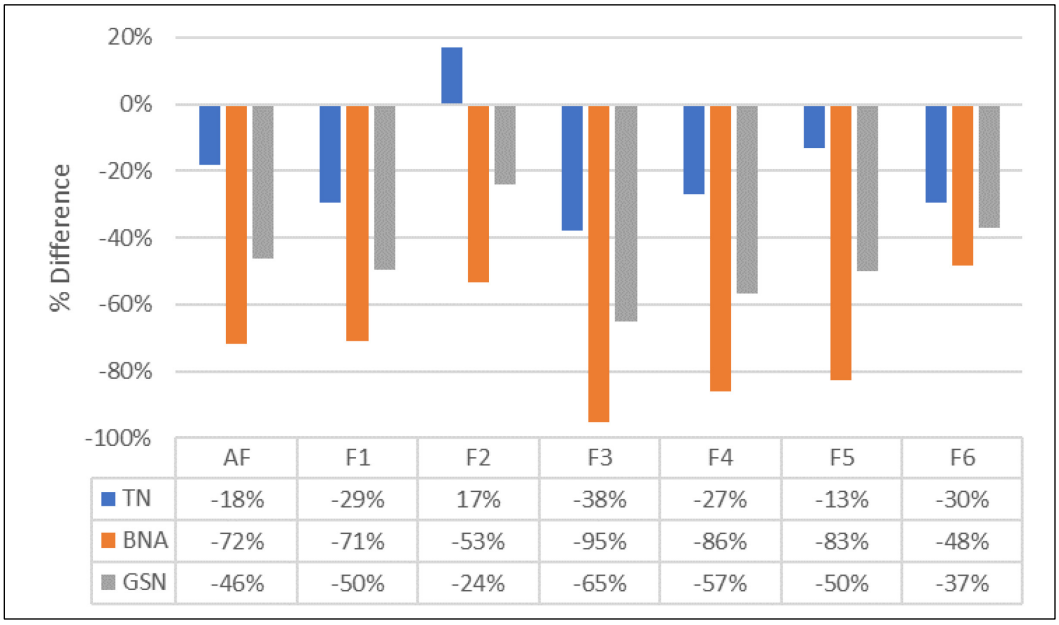
5.3.1 Soil Quality Monitoring Database Comparison with the Study Area

Study area fields, identified by respective producers as having poor yield performance and a lower expected N mineralization potential, required qualification to understand how provincially derived DSTs (at the distal scale) might be implemented at the field level (at the proximal scale) to support N fertilizer recommendations (Figure 5.1, Step 1). As a benchmark, soil samples collected from the study area were compared with soil results of SQMD parameters (Figure 5.3), including TN (representing the stable N pool), BNA (representing the labile N pool), and GSN (the calculated output of Eq. 5.1 and 5.2 including both TN and BNA values).

5.3.1.1 Total Nitrogen

Considering the mean TN of all six fields (AF) considered together (n = 144), TN was 18% below the provincial SQMD (n = 445) average (Figure 5.3, Table 5.2). With only one field above the provincial average (F2) at 17% difference, the remaining fields

were below average to a maximum of 38% (F3) difference, and a range of 55% (Figure 5.3, Table 5.2).



	AF	F1	F2	F3	F4	F5	F6
TN	-18%	-29%	17%	-38%	-27%	-13%	-30%
BNA	-72%	-71%	-53%	-95%	-86%	-83%	-48%
GSN	-46%	-50%	-24%	-65%	-57%	-50%	-37%

Figure 5.3 Percent differences between mean total nitrogen (TN), biological nitrogen availability (BNA) and growing season nitrogen (GSN) soil results from the provincial soil quality monitoring database versus the mean soil data results of the six fields in the study area considered together (AF), and individually (F1 through F6) with negative results depicting soil observations below the provincial average.

Since both the SQMD and study area samples were collected in the spring, differences in TN levels cannot be attributed to differences in seasonal variation. The lower mean TN observed in the study area clearly showed that fields were below the provincial benchmark, as chosen to provide a useful case study for implementing DSTs. The difference in TN likely displays the effect of management, as the SQMD includes soils collected from a variety of land uses, including exclusively forage based operations (PEI Department of Agriculture and Land, 2023). In contrast, the study area fields mainly implement potato-based rotations on the 3-yr cycle regulated since 2008 (Nyiraneza et al., 2017). F2 was an exception with a longer frequency in forages and cereals prior to

potatoes (Table 5.1), and was evidenced by higher relative TN levels (Figure 5.3). This observation is consistent with literature results, showing that an increase in forage rotations can improve soil organic matter and associated N pools (Whittaker et al., 2023).

5.3.1.2 Biological Nitrogen Availability

All soil BNA results collected from the study area were below the mean SQMD results (Figure 5.3, Table 5.2). The AF average difference was 72% below the SQMD with a range in fields from a minimum of 48% in F6 to a maximum of 95% below the SQMD in F3, with a range of 47% (Figure 5.3). As might be expected with this labile parameter, the mean AF BNA had a percentage difference that was 54% greater than TN showing more variability with BNA relative to TN. The large disparity between BNA values in the study area, as compared to the SQMD, was confirmed and also highlights the variable nature of this biologically mediated N pool.

5.3.1.3 Growing Season Nitrogen

GSN values for the study area, as calculated using Eq. 5.1 and 5.2 from TN and BNA soil results, were also below the provincial SQMD average (Figure 5.3, Table 5.2). The AF results for GSN had a mean difference that was 46% below the SQMD, a minimum of 24% (F2) to a maximum of 65% (F3) below the SQMD, and a range of 41% (Figure 5.3). GSN results for F2, with a longer rotation in forages and cereals (Table 5.1), was the field with the closest conformance to the SQMD benchmark (Figure 5.3). As has been observed in Chapter 3 and Chapter 4, GSN as a calculated value from TN and BNA, are more stable than the BNA parameter. Likewise, here, the mean AF results (Figure 5.3) for GSN (46% below the SQMD) falls between the extremes of TN (18% below the SQMD) and BNA (72% below the SQMD).

5.3.1.4 Soil data considerations

The soil data collected from the study area confirmed that the selected fields were of lesser soil quality than the SQMD provincial mean. The question of local vs provincial level disparity is a necessary consideration if provincial tools (i.e., PTFs or DSM derived from provincial scale data) are to be used at the field level. In addition, the variability observed between fields is notable (Figure 5.3, Table 5.2) and gives credence to the need for DSTs to help account for this variability when making N fertilizer recommendations.

5.3.2 Pedotransfer Function Development and Comparison with the Study Area

Following the framework offered in Chapter 3 (Laurence et al., 2023), and in alignment with PTF theory from McBratney et al. (2002) recommending greater accessibility for end-users, novel PTFs were developed from the provincial SQMD for response variables based on PEIAL's standard analytical suite (the "S3-package"). As noted (Section 5.2.5), the purpose of producing novel PTFs was to determine if a reliable prediction could be made from a limited, yet widely available suite of soil parameters (OM, pH, CEC) and ACI soil management layers (Figure 5.1, Step 2). Once trained from the provincial dataset, PTFs were used to make predictions of study area data points (n = 144) for comparison, and external validation with observed TN, BNA, and GSN soil results from the study area (Figure 5.2, Step 3).

5.3.2.1 Predictor variables

VIF analysis was conducted on S3-package parameters and applicable ACI soil management layers with no parameters being removed due to multi-collinearity. For comparison purposes, Figure 5.4 displays correlation plots of response and predictor variables for both the SQMD (Figure 5.4A) and the study area (Figure 5.4B). While

limited, the available predictor variables in the S3-package and ACI layers showed good correlations to response variables. The differences in correlation between the PTF training data (SQMD, Figure 5.4A) and the study area data (Figure 5.4B) were generally similar and provide confidence in the reliability of PTF use at the field scale.

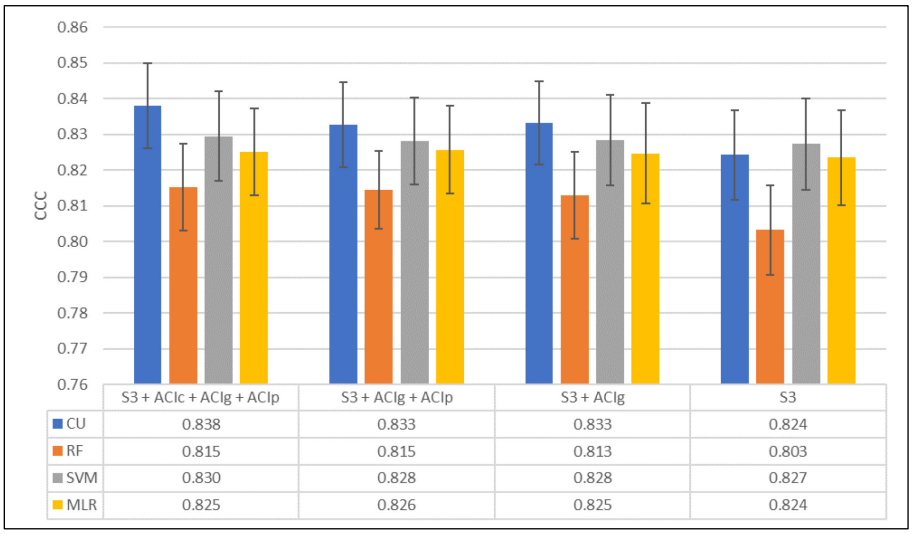
Figure 5.4 Correlations between (A) the soil quality monitoring database and (B) the study area results for total nitrogen (TN), biological nitrogen availability (BNA), growing season nitrogen (GSN), soil organic matter (SOM), pH, cation exchange capacity (CEC), ACI crop frequency of potatoes (ACIp), cereals (ACIc), and forages (ACIg).

With respect to the correlation between pH and TN, in the SQMD (Figure 5.4A) there was a 0.06 correlation, while in the study area there is a 0.49 correlation (Figure 5.4B). The correlation in Chapter 3 between TN and pH based on PEI's soil health database (SHD, n = 2,222) was 0.12 and more akin to the provincial SQMD (Laurence et al., 2023). This difference could be attributed to liming applications, or the higher proportion of forages and lower proportion of potatoes in rotation at the provincial scale compared to the study area (Table 5.3). With a higher frequency of potatoes, the reduction in OM, decreased buffering capacity, and decreased rooting biomass, the pH

parameter may exhibit a greater importance in depleted soils. Also, limestone is applied less frequently to potato fields in PEI, as producers prefer an acidic pH as a means of controlling potato scab (Waterer, 2002). Another notable difference was the change between a negative correlation of all response variables and the frequency in cereal crops over a 10-year span (ACIc) in Figure 5.4A to a positive correlation in Figure 5.4B. The suggestion here is that at the local scale, inclusion of cereal crops, that require fewer tillage passes and contribute to increased crop residue as opposed to potatoes, can have a positive correlation with N parameters.

5.3.2.2 Total Nitrogen

PTFs for TN, trained and internally validated with the SQMD (n = 445), were reduced over four iterations using the S3-package and removing the least important crop frequency layer at each iteration. The highest concordance (CCC = 0.84) was achieved with the CU learner and the S3-package and with all soil management layers (Figure 5.5).



	S3 + ACIc + ACIg + ACIp	S3 + ACIg + ACIp	S3 + ACIg	S3
CU	0.838	0.833	0.833	0.824
RF	0.815	0.815	0.813	0.803
SVM	0.830	0.828	0.828	0.827
MLR	0.825	0.826	0.825	0.824

Figure 5.5 Chart of total nitrogen (TN) concordance (CCC) results of recursive feature elimination using the S3-package (S3), the frequency of cereals (ACIc), grasses (ACIg), and potatoes (ACIp) over a 10-yr period and modeled using cubist (CU), random forest (RF), support vector machine (SVM), and multiple linear regression (MLR).

However, after RFE and the removal of soil management layers, there was no significant difference with SVM and the S3-package alone (Figure 5.5, CCC = 0.83). As such, and opting for the most parsimonious model, TN with predictor variables OM, pH, CEC (the S3-package) and the SVM model (SVM-TN) was used for applying the PTF to the study area.

Table 5.5 External validation results of PTFs derived using the support vector machine (SVM) learner and derived from the soil quality monitoring database for total nitrogen (TN), biological nitrogen availability (BNA), growing season nitrogen (GSN) and applied to the study area's six fields including all fields considered together (AF, *n* = 144), fields considered individually (F1 through to F6, *n* = 24/field) and the mean value of each field (*n* = 6) using organic matter (OM), pH, cation exchange capacity (CEC) and/or the frequency of grasses (ACIg) and potatoes (ACIp) in a 10 year period as conceptual models showing the concordance (CCC), coefficient of determination (R^2), and the root mean square error (RMSE).

Parameter	Observations	Conceptual Model	CCC	R^2	RMSE
TN	AF	OM + pH + CEC	0.91	0.84	0.013
	F1	OM + pH + CEC	0.86	0.84	0.011
	F2	OM + pH + CEC	0.63	0.62	0.015
	F3	OM + pH + CEC	0.88	0.90	0.012
	F4	OM + pH + CEC	0.85	0.80	0.010
	F5	OM + pH + CEC	0.76	0.68	0.012
	F6	OM + pH + CEC	0.12	0.02	0.015
	MF	OM + pH + CEC	0.80	0.95	0.006
BNA	AF	OM + pH + CEC	0.13	0.12	15.9
	F1	OM + pH + CEC	0.32	0.82	10.9
	F2	OM + pH + CEC	0.01	0.01	24.0
	F3	OM + pH + CEC	0.16	0.70	15.3
	F4	OM + pH + CEC	0.12	0.76	12.2
	F5	OM + pH + CEC	0.02	0.05	18.7
	F6	OM + pH + CEC	0.03	0.03	9.5
	MF	OM + pH + CEC	0.08	0.14	14.4
GSN	AF	OM + pH + CEC + ACIg + ACIp	0.33	0.39	43.2
	F1	OM + pH + CEC	0.50	0.88	28.8
	F2	OM + pH + CEC	0.03	0.15	60.0
	F3	OM + pH + CEC	0.30	0.81	45.8
	F4	OM + pH + CEC	0.28	0.84	32.8
	F5	OM + pH + CEC	0.08	0.27	53.3
	F6	OM + pH + CEC	0.05	0.03	27.2
	MF	OM + pH + CEC + ACIg + ACIp	0.25	0.51	37.8

The SVM-TN PTF was then used to make predictions on the study area's independent data set for external validation. The CCC results, seen in Table 5.5, Figure 5.6A and 5.6B, show that observed TN with AF samples had a CCC of 0.91 in relation to predicted TN in the study area. This result was promising considering the SQMD, from which SVM-TN was derived, had 18% higher TN values than study area results (Figure 5.3). The scatterplot of AF results (Figure 5.6B), shows strong conformance to the 1:1 line (predicted vs. observed TN) with relatively few outliers. F6 results, which had the lowest CCC (0.12), were mainly clustered above the 1:1 line, which means that predicted values tended to be less than observed TN results (Figure 5.6A, Table 5.5).

The comparatively poor results in F6 may be indicative of the role climate plays in controlling TN (Chapter 4), as F6 is on the eastern and milder extreme of the province (Table 5.1). With respect to how PTFs might be used in practice, the MF result (Figure 5.6A, Table 5.5) was of comparable strength (CCC = 0.80) to AF and the bulk of individual fields; as such, this shows that reliable estimates of TN could be made for whole-field N fertilizer applications and are more conducive to SRA scenarios.

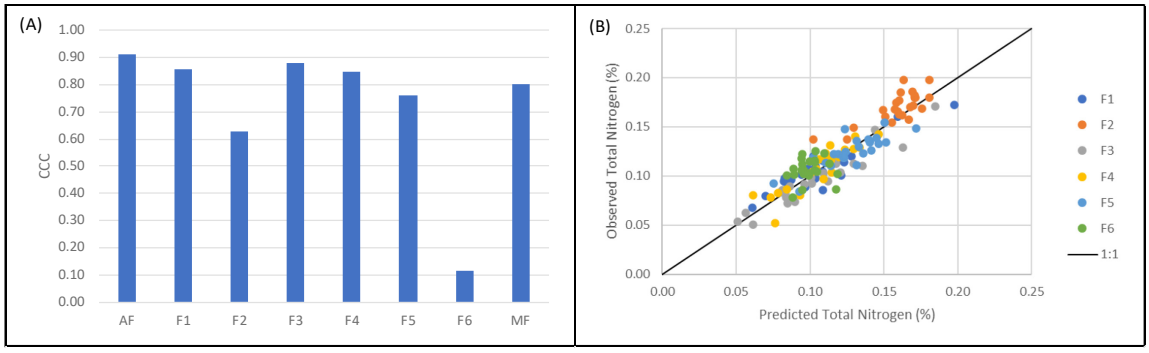
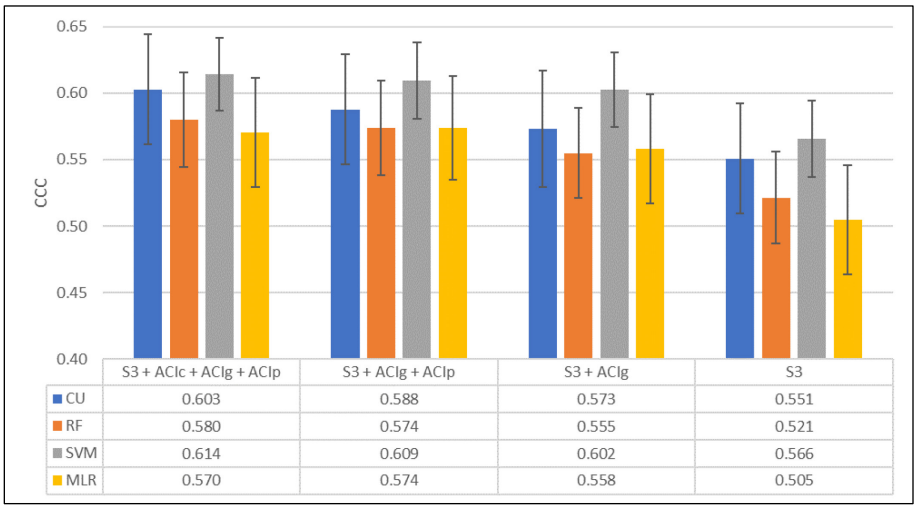


Figure 5.6 (A) Concordance (CCC) results of predicting total nitrogen (TN) on the study area with all field data together (AF), each field specifically (F1 through F6) and the mean of each field (MF) and (B) a scatter plot of the predicted vs. observed TN (%) with the 1:1 line for all results from a pedotransfer function modelled with support vector machine (SVM) and trained with the S3-package (organic matter, pH, and cation exchange capacity) from the soil quality monitoring database (SQMD).

5.3.2.3 Biological Nitrogen Availability

Using RFE to select the optimum predictor variables for BNA, and the SQMD (*n* = 445) to train and internally validate models using MLs and MLR, the highest concordance (CCC = 0.61) was obtained with all predictors and the SVM model (Figure 5.7). The optimum PTF, having the fewest predictors without a significant difference, was SVM and the S3-package (CCC = 0.57, Figure 5.7) including OM, pH, and CEC as the predictor variables (SVM-BNA). Comparing these results to TN (CCC = 0.83), the lower CCC in predicting BNA (CCC = 0.57) was somewhat expected given the multiple influences, variable, and labile nature of BNA. However, the substantial reduction in CCC between predicting TN vs. BNA was not observed in Chapter 3 (Laurence et al., 2023) and was likely due to the limited suite of predictor variables available (OM, pH, CEC), and their respective correlations with BNA (Figure 5.4). For example, the correlation between BNA, OM, and pH was higher in the SQMD (0.58 and 0.10, respectively) than in the study area (0.40 and 0.04, respectively).



	S3 + ACIc + ACIg + ACIp	S3 + ACIg + ACIp	S3 + ACIg	S3
CU	0.603	0.588	0.573	0.551
RF	0.580	0.574	0.555	0.521
SVM	0.614	0.609	0.602	0.566
MLR	0.570	0.574	0.558	0.505

Figure 5.7 Chart of biological nitrogen availability (BNA) concordance (CCC) results of recursive feature elimination using the S3-package (S3), the frequency of cereals (ACIc), grasses (ACIg), and potatoes (ACIp) over a 10-yr period and modeled using cubist (CU), random forest (RF), support vector machine (SVM), and multiple linear regression (MLR).

The PTF of choice (SVM-BNA), was then used to make predictions of BNA in the study area, and compared with observed values. The CCC results for BNA were markedly different from TN, and showed a CCC of 0.13 for AF samples (Table 5.5, Figure 5.8A and 5.8B). Compared with TN results (CCC = 0.91), the PTF for BNA showed lower predictive strength. Observing the scatterplot (Figure 5.8B), PTF predictions were typically higher than observations as indicated with most of the scatterplot points below the 1:1 line. Recalling that the SQMD had a mean BNA of 38.3 mg/kg and the study area had a mean BNA of 18.3 mg/kg (Table 5.2), a difference of 72%, the higher predicted vs. observed results can be understood. In addition, the selection of predictor variables, solely including OM, pH, and CEC, shows that these predictors do not fully capture the intrinsic controls for BNA. In the study by Laurence et al. (2023) in Chapter 3, the optimum predictors for BNA included aggregate stability, permanganate oxidizable carbon, soil respiration, and TN; none of which are available in the S3-package. Like TN, the PTF for BNA was successful in identifying field performance with respect to provincial (SQMD) averages.

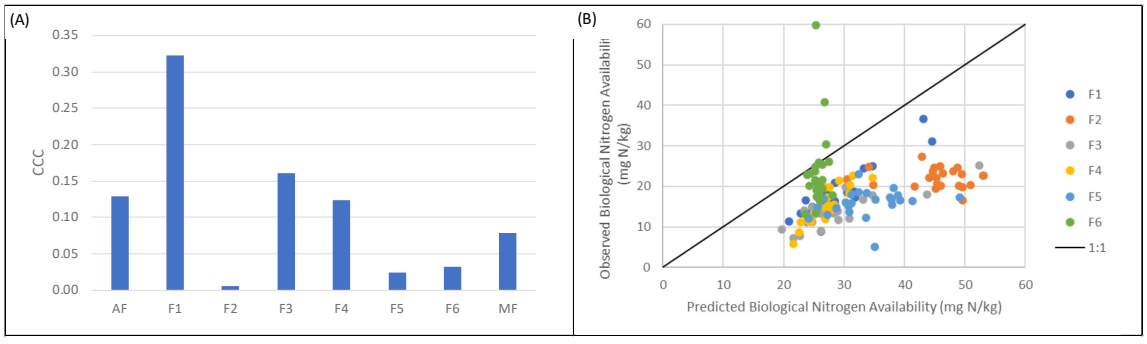


Figure 5.8 (A) Concordance (CCC) results of predicting biological nitrogen availability (BNA) on the study area with all field data together (AF), each field specifically (F1 through F6) and the mean of each field (MF) and (B) a scatter plot of the predicted vs. observed TN (%) with the 1:1 line for all results from a pedotransfer function modelled with support vector machine (SVM) and trained with the S3-package (organic matter, pH, and cation exchange capacity) from the soil quality monitoring database (SQMD).

5.3.2.4 Growing Season Nitrogen

Using the SQMD (n = 445) for training and validation, and after RFE, the optimum predictor variables for GSN were OM, pH, CEC from the S3-package plus crop management layers including the number of times in forages (ACIg) and potatoes (ACIp) over a 10-year period (Figure 5.9). The best CCC result (0.70) using these predictor variables was with the SVM model (SVM-GSN_{S3+}), which was significantly different to other models that used the S3-package alone. In comparison to PTFs for TN (SVM-TN, CCC = 0.83) and BNA (SVM-BNA, CCC = 0.57), GSN (SVM-GSN_{S3+}, CCC = 0.70) had a stronger predictive strength than BNA alone; likely due to the presence of TN in the GSN calculation (Eq. 5.1 and 5.2). It was observed in Chapter 3 that approximately half of the GSN came from the more stable fraction driven by TN; as such, it was understandable that GSN would be better predicted from the S3-package (containing OM) than would be the labile components that dominate BNA fractions. As a limitation, the inclusion of soil management (ACI crop frequency) layers was only useful for training AF or MF functions since cropping differences infield are only applicable in mixed-cropping scenarios. As such, for infield usage in the study area, the SVM model and the S3-package alone (SVM-GSN_{S3}, CCC = 0.66) was the appropriate PTF (Figure 5.9).

Applied to the study area for external validation, the SVM-GSM_{S3+} and SVM-GSN_{S3} PTFs were used to predict GSN and accuracy metrics were recorded (Figure 5.10A and 5.10B, Table 5.5). Overall, for AF and MF results, the GSN predictions were promising with greater CCC results than BNA, but less than TN; a trend that was expected based on previous results. As observed in the scatterplot (Figure 5.10B), the

model predictions of GSN were higher than observed, which was evidenced by the points dominant below the 1:1 line. Like TN and BNA, the GSN results in the SQMD were above those in the study area; and in like manner, PTFs yielded predictions from the model that were reflective of field conditions (Figure 5.3). With fields considered individually (Figure 5.10), the same variability in GSN exists as observed with TN (Figure 5.6) and BNA (Figure 5.8) with ranges in CCC from 0.50 in F1 to 0.03 in F2 (Table 5.5).

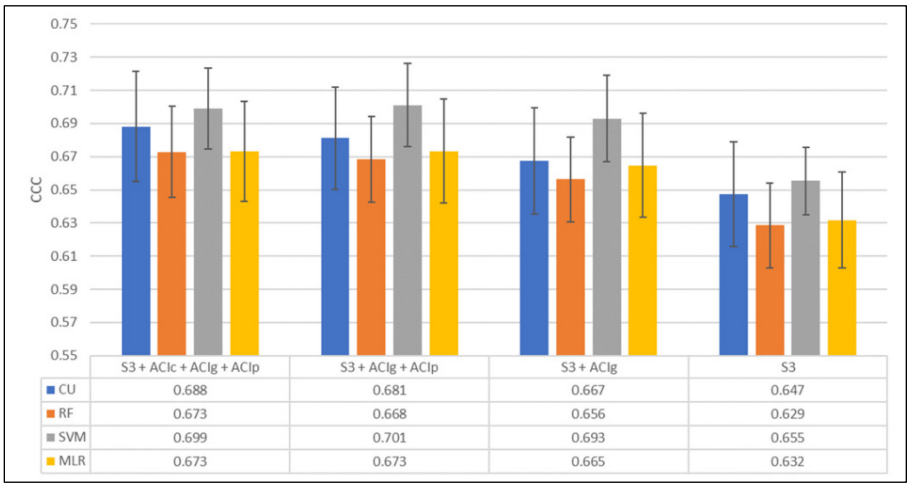


Figure 5.9 Chart of growing season nitrogen (GSN) concordance (CCC) results of recursive feature elimination using the S3-package (S3), the frequency of cereals (ACIc), grasses (ACIg), and potatoes (ACIp) over a 10-yr period and modeled using cubist (CU), random forest (RF), support vector machine (SVM), and multiple linear regression (MLR).

The accuracy of MF predictions (CCC = 0.25) was comparatively strong and suggests that applying "whole-field" PTFs of GSN may yield useful estimates for informing SRAs of N fertilizer. In general, estimates of GSN, as well as TN and BNA, from provincially derived PTFs were very successful in placing these fields in context with reference to provincial (SQMD) benchmark data. This aspect of the PTF is important for producers to identify fields that provide lower than average GSN supply in

order to adjust N fertilizer rates accordingly. For example, with predictions of GSN higher than observed, a reduction in soil N credits might be considered.

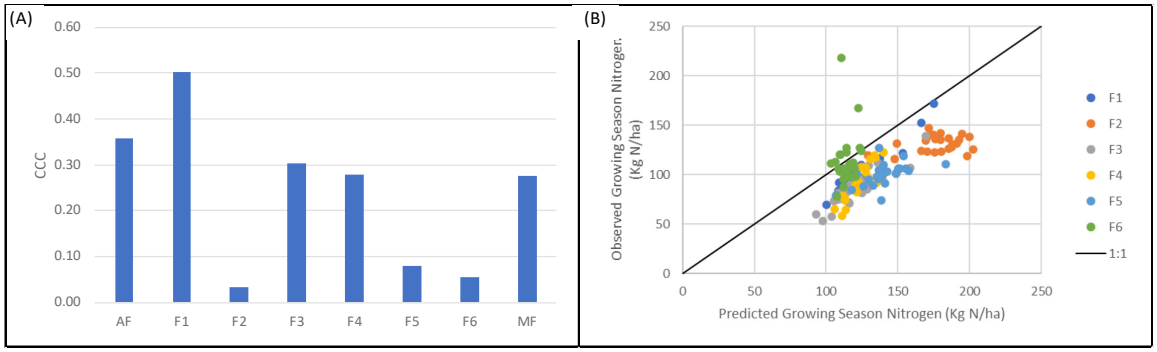


Figure 5.10 (A) Concordance (CCC) results of predicting growing season nitrogen (GSN) on the study area with all field data together (AF), each field specifically (F1 through F6) and the mean of each field (MF) and (B) a scatter plot of the predicted vs. observed TN (%) with the 1:1 line for all results from a pedotransfer function modelled with support vector machine (SVM) and trained with the S3-package (organic matter, pH, and cation exchange capacity), and the frequency of grasses (ACIg), and potatoes (ACIp) over a 10-yr period from the soil quality monitoring database (SQMD).

5.3.2.5 Pedotransfer function coefficients

MLR modelling (Figure 5.5, 5.7, and 5.9) showed reliable results for TN (MLR-TN, CCC = 0.82), BNA (MLR-BNA, CCC = 0.51) and GSN (MLR-GSN, CCC = 0.63) using the S3-package. PTF coefficients are provided in Table 5.6. From the producer

Table 5.6 Pedotransfer function accuracy metrics including Lin's concordance correlation coefficient (CCC), coefficient of determination (R^2), and root mean square error (RMSE) and coefficients from multiple linear regression results for total nitrogen (TN), biological nitrogen availability (BNA), and growing season nitrogen (GSN) using S3-package parameters including organic matter (OM), pH, and cation exchange capacity (CEC).

Parameter	CCC	R^2	RMSE	a	b	c	d
TN	0.82	0.74	0.020	-0.0178	0.0475	0.0038	NCG
BNA	0.51	0.40	12.8	-16.7	13.9	3.4	-0.597
GSN	0.63	0.52	37.4	-33.9	52.9	10.3	-1.57

Note: 1. PTF formula for estimating TN, BNA, or GSN = a + b (OM) + c (pH) + d (CEC)
2. NCG = variable included in conceptual model, but no coefficient generated

perspective, and as evidenced by the variability observed within fields, the optimum application of PTF coefficients would be at the whole-field (e.g., MF) scale in preference to single point predictions.

5.3.3 Provincial Digital Soil Map Comparison with the Study Area

Having first considered the results of direct soil observations, and PTF point predictions (Sections 5.3.1 and 5.3.2, respectively), spatial applications of N parameters are now considered. For the purpose of understanding how spatial representations might be implemented by producers, provincial scale DSM predictions (i.e., maps completed in Chapter 4) of TN, BNA, and GSN were compared with soil observations from the study area at each geographic location (Figure 5.1, Step 4).

5.3.3.1 Total Nitrogen

Predictions of TN from the provincial DSM (Section 4.4.2.1), developed using the SVM model with the SQMD (n = 445), had a CCC of 0.38 for AF, 0.40 for MF, and a range from -0.10 (F2 and F6) to 0.06 (F1) when compared with observed values at the same geographic location in the study area (Table 5.7).

While there was relatively no CCC between observed and predicted values for fields individually (Figure 5.11), there was improved CCC for broader scale considerations such as AF and MF (Table 5.7). This result suggests that regional map predictions of TN are better suited for broad scale applications such as whole field nutrient management practices (e.g., SRA) but less so for infield approaches (e.g., VRT). However, while there was a lack of CCC at the infield scale, there remained a high percentage of infield observations within the DSMs uncertainty PI (90% prediction

interval), especially AF and MF results, showing good reliability of uncertainty estimates (Table 5.7).

Table 5.7 External validation results from provincial digital soil map predictions of total nitrogen (TN), biological nitrogen availability (BNA), and growing season nitrogen (GSN) derived from the soil quality monitoring database (SQMD) and applied to the study area's six fields including all fields considered together (AF, *n* = 144), fields considered individually (F1 through to F6, *n* = 24/field) and the mean value of each field (MF, *n* = 6) with the concordance (CCC), coefficient of determination (R^2), and the root mean square error (RMSE).

Parameter	Observations	CCC	R^2	RMSE	w/in 90% PI
TN	AF	0.38	0.34	0.028	90%
	F1	0.06	0.05	0.026	92%
	F2	-0.10	0.26	0.028	100%
	F3	0.00	0.00	0.036	71%
	F4	0.01	0.00	0.035	77%
	F5	-0.03	0.02	0.030	90%
	F6	-0.10	0.04	0.015	95%
	MF	0.40	0.52	0.021	100%
BNA	AF	0.02	0.01	16.3	49%
	F1	-0.03	0.02	15.4	42%
	F2	-0.01	0.00	16.7	70%
	F3	-0.13	0.38	15.8	29%
	F4	0.03	0.08	22.2	15%
	F5	-0.03	0.18	15.7	30%
	F6	-0.01	0.00	12.7	76%
	MF	0.05	0.26	15.0	33%
GSN	AF	0.19	0.25	48.6	63%
	F1	-0.06	0.03	46.7	54%
	F2	-0.01	0.01	56.5	100%
	F3	-0.01	0.01	46.6	36%
	F4	0.10	0.20	60.2	31%
	F5	0.05	0.10	38.6	50%
	F6	0.01	0.00	37.4	76%
	MF	0.19	0.78	43.5	50%

The scatter plot (Figure 5.11) shows a higher proportion below the 1:1 line, which indicates that predictions of TN are higher than observed values. Fields, such as F2 were almost exclusively above the 1:1 line (Figure 5.11), showing higher TN observations vs.

predictions. This relationship was in keeping with Figure 5.3 that showed F2 as the only field with TN results above the provincial benchmark. Similarly, single point (PTF) predictions from F2 and reported in Section 5.3.2.2 (Figure 5.6B), were above the 1:1 line. As such, the results grant validity to provincial DSTs as a soil quality benchmark to successfully qualify how specific fields compare with the provincial average.

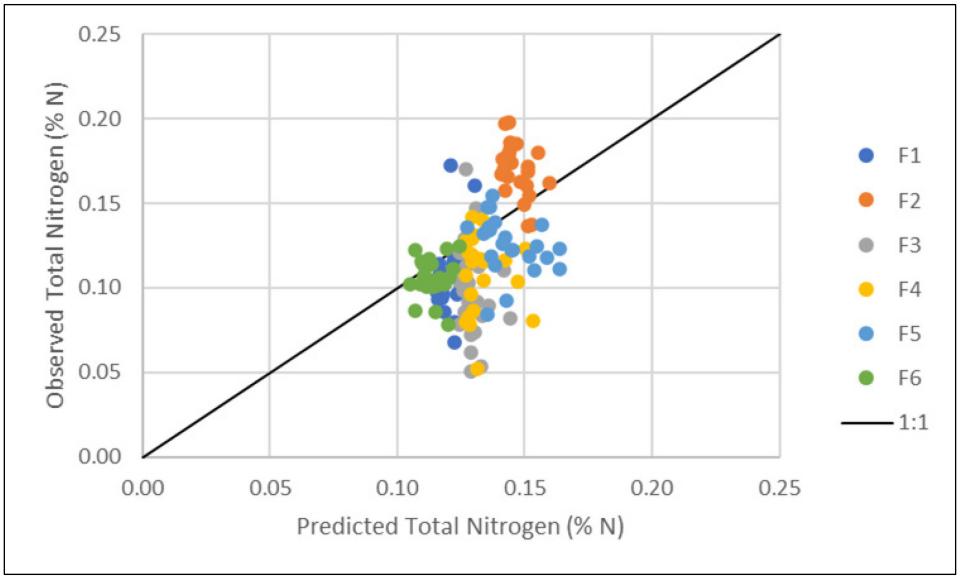


Figure 5.11 Scatter plot of observed Total Nitrogen (TN) from direct soil data in the study area versus predictions of TN from the provincial digital soil map (DSM) derived from the soil quality monitoring database (SQMD).

5.3.3.2 Biological Nitrogen Availability

BNA had limited CCC in the AF, MF, or individual field results (Table 5.2). However, AF results showed 49% of the soil observations within the DSMs uncertainty PI (Table 5.2). In addition, soil observations ranged from 15% (F4) to 70% (F2) being within the PI, which shows reliability with provincial DSM predictions of BNA. As observed in Figure 5.12, observations for BNA in the study area were almost exclusively below the regional predictions (i.e., below the 1:1 line), showing that the DSM was

reflective of the provincial benchmark with all BNA field samples below SQMD averages (Figure 5.3). Representing the labile N pool, the data also demonstrates that BNA was a more sensitive indicator, and responds more quickly than TN to management practices that negatively impact soil health.

The clustering observed in Figure 5.12 shows a moderate level of precision vs. accuracy, which may indicate the negative bias introduced via poor soil quality inherent in the study area, then from issues introduced by the DSM itself. However, while the provincial scale DSM showed lower accuracy in point-specific predictions, the percentage of results within the uncertainty limits showed that 49% (AF) and 33% (MF) of the results were within the prediction interval. As such, the provincial DSM appears more suited to landscape or whole-field scale estimates of BNA.

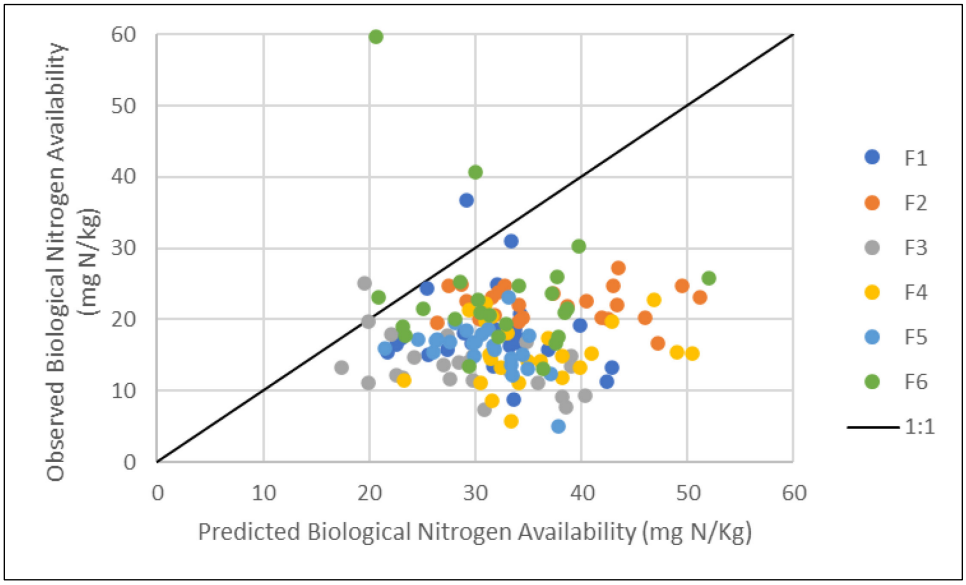


Figure 5.12 Scatter plot of observed Biological Nitrogen Availability (BNA) from direct soil data in the study area versus predictions of BNA from the provincial digital soil map (DSM) derived from the soil quality monitoring database (SQMD).

5.3.3.3 Growing Season Nitrogen

Considering AF and MF, soil observations in the study area had a CCC of 0.19 with provincial DSM predictions of GSN (Table 5.7). Considering individual fields however, infield GSN observations showed minimal CCC with DSM predictions as was similar with TN and BNA. The scatter plot in Figure 5.13 shows that provincial DSM predictions were higher in comparison to actual observed GSN values. Regarding uncertainty of DSM predictions, 63% of AF observations were within the PI, 50% of MF observations were within the PI, and a range of 100% to 31% of infield specific observations were within the PI (Table 5.7).

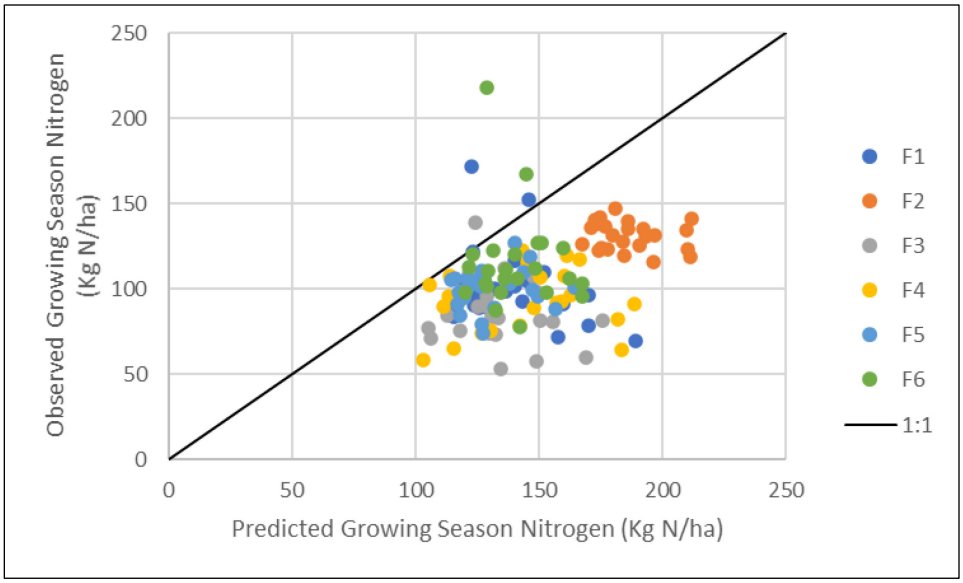


Figure 5.13 Scatter plot of observed Growing Season Nitrogen (GSN) from direct soil data in the study area versus predictions of GSN from the provincial digital soil map (DSM) derived from the soil quality monitoring database (SQMD).

From an applied perspective, there was a good level of consistency in predictions, both as being reflective of the provincial benchmark (Figure 5.3), and as remaining close within the 90% PI. The higher CCC and number of observations within AF and MF

uncertainty limits, both of which are broader scale considerations, coupled with the lower CCC with infield applications, show that provincial scale DSM predictions of GSN are better applied to whole-field or landscape applications than fine-scale or zone-based approaches. As with PTF predictions, the provincially derived DSMs of GSN, in addition to TN and BNA, could identify variability between fields and place these fields in context with respect to the provincial SQMD benchmark. As such, the DSM provides a useful tool for producers when site specific data are not available and reasonable estimates of N parameters are required.

5.3.4 Modelling Infield Nitrogen Indices

Provincial scale predictions of N indices from DSMs were able to qualify how specific fields compared to provincial averages; and in addition, provide reasonable estimates for landscape or whole-field approaches to N fertilizer management (e.g., SRA). To assist producers in VRA scenarios, infield modelling was conducted to identify the best infield model and predictors of TN, BNA, and GSN - thereby identifying N controls at the proximal scale (Figure 5.1, Step 5). Data from the SQMD was not used for infield modelling procedures.

5.3.4.1 Feature elimination

Covariate layer values at AF data point locations were extracted and tested for multi-collinearity. Of the 90 covariate layers, 41 layers remained below the stopping threshold (VIF = 10) with a total reduction of 54% (Table 5.4).

5.3.4.2 Total Nitrogen

Representing the stable N pool, and to identify the optimum training approach, TN was modelled using all study area (AF) samples, and then with each field individually

using field-specific (FS) samples. For AF modelling of TN with LFOCV, the CU model obtained a CCC of 0.39 with two covariates required for prediction (Table 5.8). This CCC was comparable to the provincial DSM of TN in Chapter 4, which had a CCC of 0.45 using eight predictors (Chapter 4).

Training the model with FS data yielded the strongest results with a range from 0.47 to 0.83 (Figure 5.14, Table 5.8), which was an improvement to using AF data for modelling (CCC = 0.39, Figure 5.14), and the provincial DSM of TN (Chapter 4). The suggestion from the results was that as the scale narrows to capture field level variability, field specific data was preferred for infield mapping and model training. The relevance of predictor variables retained for TN are discussed in Section 5.3.6.1.

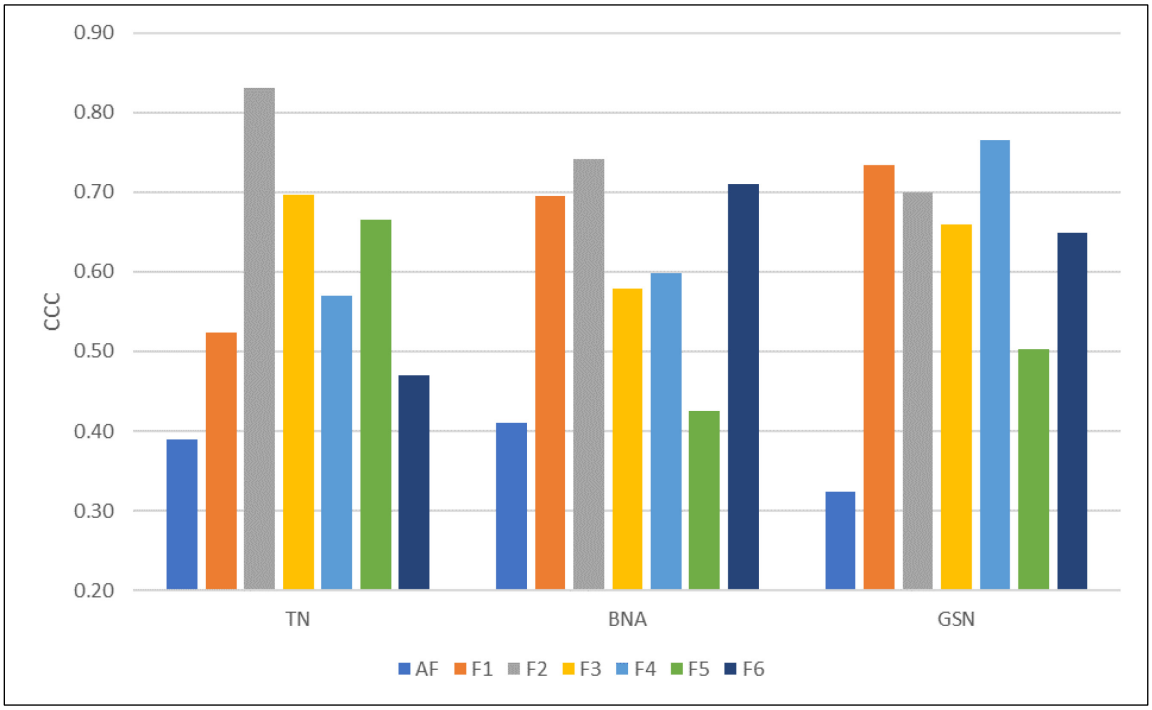


Figure 5.14 Comparison of concordance (CCC) results for infield modelling of total nitrogen (TN) using all fields considered together (AF) and individually (F1 to F6).

Table 5.8 Accuracy metrics, selected covariates by *scorpan* factor, and correlations for total nitrogen (TN) of all-field samples together (AF) and individually (F1 through F6) and identifying optimum models (CU = cubist, SVM = support vector machines) and validation procedures (LFOCV = leave field out cross validation, LOOCV = leave one out cross validation).

Category	Description	Abbrev.	Correlation	AF CU LFOCV	F1 SVM LOOCV	F2 SVM LOOCV	F3 CU LOOCV	F4 SVM LOOCV	F5 SVM LOOCV	F6 SVM LOOCV
Accuracy metrics	Lin's concordance correlation coefficient	CCC		0.39	0.52	0.83	0.70	0.57	0.67	0.47
	Coefficient of determination	R^2		0.04	0.02	0.01	0.02	0.02	0.01	0.02
	Root mean square error	RMSE		0.17	0.31	0.76	0.58	0.38	0.49	0.35
Soil variables	Soil color normalized / bare soil index	S.SCN	-0.18				✓			
	Biological Nitrogen Availability, provincial DSM prediction	S.BNAp	0.16				✓			
Climate variables	Min temp of coldest month	C.mint.cm	0.10	✓						
Organism variables	Growing season peak NDVI value: 2020	O.NDVI20	0.17		✓				✓	
	Growing season peak NDVI value: 2021	O.NDVI21	-0.04		✓					
	Growing season peak NDVI value: 2022	O.NDVI22	0.44						✓	
Relief variables	Deviation from Mean Elevation, filter 3 m^2	R.dme3	-0.14							✓
	Eastness (Aspect)	R.E.asp	-0.53	✓		✓				
	Elevation Percentile, filter 50 m^2	R.ep50	-0.20			✓	✓			
	Mid-Slope Position	R.msp	-0.05			✓	✓			
	Multiresolution Index of Ridge Top Flatness	R.rtf	-0.30			✓	✓		✓	
	Northness (Aspect)	R.N.asp	0.02			✓				
	Slope Height	R.sh	-0.22					✓		
	Stream Power Index	R.spi	0.08				✓			
	Topographic Position Index (normalized)	R.TPIn	-0.19						✓	✓
	Total Curvature	R.tc	-0.06						✓	
	Valley Depth	R.vd	0.07			✓				
		Total covariates required		2	2	6	6	1	5	2
		n =		144	24	24	24	13	21	23

5.3.4.3 Biological Nitrogen Availability

Representing the labile N pool, BNA modelled with AF samples had a CCC of 0.41 and required five covariates for prediction using the SGB learner (Figure 5.14, Table 5.9). This result was also like the provincial scale DSM for BNA in Chapter 4, which obtained a CCC of 0.45 using the SGB learner and the SQMD.

However, the best CCC results were obtained using FS samples to model respective fields (F1 to F6). The maximum CCC for BNA ranged from a maximum of 0.74 (F2) using seven predictors, to a minimum of 0.42 (F5) using one predictor and the SVM learner (Table 5.9). The 49% difference between the FS CCC (0.74) and the provincial DSM CCC (0.45) for BNA, in addition to the comparatively low performance of AF modelling (0.41), confirms what has been observed thus far, that predictions of N indices were much improved if obtained from localized data.

5.3.4.4 Growing Season Nitrogen

GSN, the combined result of TN and BNA (Eq. 5.1 and 5.2), and as modelled using AF samples, had a CCC of 0.32 (Figure 5.14, Table 5.10). Again, FS modelling outperformed AF modelling and ranged from a maximum CCC of 0.77 (F4) and two predictors with CU learner, to a minimum of 0.50 (F5) and two predictors with the SVM learner (Table 5.10).

In comparison with provincial scale DSM performance in Chapter 4, the best CCC for GSN was 0.47 and required eight predictors with the SGB learner. With FS modelling, there was an 83% difference in CCC (0.77 vs. 0.32, respectively) from the AF sampling (Figure 5.14, Table 5.10), and a 48% difference in CCC from the provincial

Table 5.9 Accuracy metrics, selected covariates for *scorpan* factors, and correlations for biological nitrogen availability (BNA) of all-field samples together (AF) and individually (F1 through F6) and identifying optimum models (SGB = stochastic gradient boosting, CU = cubist, SVM = support vector machines, and RF = random forest) and validation procedures (LFOCV = leave field out cross validation, LOOCV = leave one out cross validation).

Category	Description	Abbrev.	Correlation	AF SGB LFOCV	F1 CU LOOCV	F2 SVM LOOCV	F3 RF LOOCV	F4 RF LOOCV	F5 SVM LOOCV	F6 SVM LOOCV
Accuracy metrics	Lin's concordance correlation coefficient	CCC		0.41	0.69	0.74	0.58	0.60	0.42	0.71
	Coefficient of determination	R^2		0.21	0.54	0.60	0.49	0.44	0.21	0.58
	Root mean square error	RMSE		5.1	4.0	1.7	3.0	3.2	3.4	3.6
Soil variables	Soil color normalized / bare soil index	S.SCN	-0.34	✓	✓			✓		✓
	Biological Nitrogen Availability, provincial DSM prediction	S.BNAp	0.07						✓	
Organism variables	Growing season peak NDVI value: 2019	O.NDVI19	0.26	✓						✓
	Growing season peak NDVI value: 2020	O.NDVI20	-0.01		✓					
	Growing season peak NDVI value: 2021	O.NDVI21	0.27	✓						
	Growing season peak NDVI value: 2022	O.NDVI22	0.13	✓						✓
Relief variables	Convergence Index	R.conv	-0.04			✓				
	Deviation from Mean Elevation, filter 3 m^2	R.dme3	-0.14					✓		✓
	Eastness (Aspect)	R.E.asp	-0.25	✓						
	Mid-Slope Position	R.msp	0.12			✓				
	Multiresolution Index of Ridge Top Flatness	R.rtf	-0.14				✓			
	Northness (Aspect)	R.N.asp	0.14				✓			✓
	Slope Average	R.sa	-0.22			✓				
	Slope Length	R.sl	0.10							✓
	Stream Power Index	R.spi	0.08			✓				
	Topographic Position Index (normalized)	R.tpi	0.01			✓				✓
	Valley Depth	R.vd	-0.07			✓				
	Wetness Index (SAGA)	R.wi	0.20			✓				
			Total covariates required	5	2	7	2	2	1	7
			n =	105	24	24	14	24	20	21

Table 5.10 Accuracy metrics, selected covariates by *scorpan* factor, and correlations for growing season nitrogen (GSN) of all-field samples together (AF) and individually (F1 through F6) and identifying optimum models (SGB = stochastic gradient boosting, CU = cubist, SVM = support vector machines) and validation procedures (LFOCV = leave field out cross validation, LOOCV = leave one out cross validation).

Category	Description	Abbrev.	Correlation	AF SGB LFOCV	F1 SVM LOOCV	F2 CU LOOCV	F3 CU LOOCV	F4 CU LOOCV	F5 SVM LOOCV	F6 SVM LOOCV
Accuracy metrics	Lin's concordance correlation coefficient	CCC		0.32	0.73	0.70	0.66	0.77	0.50	0.65
	Coefficient of determination	R^2		0.14	0.60	0.54	0.60	0.65	0.28	0.51
	Root mean square error	RMSE		22.8	15.5	5.9	14.0	11.3	10.4	12.1
Soil variables	Soil color normalized / bare soil index	S.SCN	-0.31	✓	✓		✓	✓		✓
	Growing Season Nitrogen, provincial DSM prediction	S.GSNp	0.46	✓			✓			
Organism variables	Growing season peak NDVI value: 2019	O.NDVI19	0.32	✓						✓
	Growing season peak NDVI value: 2020	O.NDVI20	0.07		✓				✓	
	Growing season peak NDVI value: 2021	O.NDVI21	0.17	✓	✓		✓			
	Growing season peak NDVI value: 2022	O.NDVI22	0.28							✓
	Maximum NDVI value over a 10 year period (2012 - 2021)	O.NDVI.ma	0.20				✓			
	Range of NDVI value over a 10 year period (2012 - 2021)	O.NDVI.ra	-0.26					✓		
Relief variables	Convergence Index	R.conv	-0.02			✓				
	Deviation from Mean Elevation, filter 3 m^2	R.dme3	-0.15							✓
	Eastness (Aspect)	R.E.asp	-0.41	✓		✓	✓			
	Mid-Slope Position	R.msp	0.06			✓				
	Multiresolution Index of Ridge Top Flatness	R.rtf	-0.23				✓			
	Northness (Aspect)	R.N.asp	0.10				✓			
	Slope Length	R.sl	0.10							✓
	Stream Power Index	R.spi	0.08				✓			
	Topographic Position Index (normalized)	R.tpi	-0.07						✓	✓
	Wetness Index (SAGA)	R.wi	0.31	✓						
			Total covariates required	6	3	3	8	2	2	6
			n =	105	24	24	22	24	21	21

DSM result (0.77 vs. 0.47, respectively). As such, there appears consistency that while modelling GSN with samples from different regions yield reasonable results (e.g., AF data or SQMD data), the best option for modelling N indices was with FS sample results.

5.3.5 Infield Mapping of N parameters

Based on models built from AF and FS data (Section 5.3.4), and to assist producers in VRA scenarios, infield maps of TN, BNA and GSN were produced for each field (F1 through F6) in the study area. Also, for comparison purposes, provincial scale DSM predictions and uncertainty estimates from Chapter 4 were extracted for the total area of each field (Figure 5.1, Step 6).

5.3.5.1 Total Nitrogen

Including both prediction and uncertainty maps (12 maps and 36 maps, respectively), a total of 48 infield maps of TN were generated for the study area consisting of fields F1 through F6. Descriptive statistics of TN predictions, the 90% (uncertainty) PI's for maps modelled from AF and FS data for each field in the study area (F1 through F6), and the provincial scale DSM, are given in Table 5.11.

Comparing the model accuracies for TN in fields F1 through F6, Section 5.3.4.2 (Figure 5.14, Table 5.8), the highest CCC (0.83) was obtained from F2. The predictive map for F2, trained with SVM and six predictors using FS soil data, had a range from 0.13 to 0.22% TN (Figure 5.15B), a mean of 0.17% and a standard deviation (SD) of 0.016% TN (Table 5.11). The lower prediction limit (5^{th} percentile) ranged from 0.13 to 0.22% TN (Figure 5.15A) and the upper prediction limit (95^{th} percentile) ranged from

0.13 to 0.23% TN (Figure 5.15C). The overall 90% PI width ranged from 0.003 to 0.01% TN (Figure 5.15D) with a mean of 0.01% and a SD of 0.001% (Table 5.11). The highest CCC results for remaining fields, F1 and F3 through F6, were obtained using FS soil data and are available in the Appendix (Figure A.1 to A.5).

Table 5.11 Descriptive statistics (minimum (Min), Mean, maximum (Max) and standard deviation (SD) values) for total nitrogen (%) soil prediction and 90% prediction interval maps for the study area fields (F1 through F6) including extracted results from the provincial scale digital soil map (DSM) and novel maps trained with all-field samples (AF) and field-specific (FS) samples and including machine learners (SVM = support vector machines; CU = cubist)

Field	Data Source	Learner	Spatial Interpretation	Min	Mean	Max	SD
F1	Provincial DSM	SVM	Soil Prediction Map	0.11	0.12	0.14	0.004
			90% Prediction Interval Map	0.08	0.09	0.11	0.003
	AF Data	CU	Soil Prediction Map	0.10	0.11	0.14	0.007
			90% Prediction Interval Map	0.002	0.04	0.05	0.008
	FS Data	SVM	Soil Prediction Map	0.09	0.11	0.18	0.018
			90% Prediction Interval Map	0.04	0.06	0.10	0.010
F2	Provincial DSM	SVM	Soil Prediction Map	0.14	0.15	0.16	0.005
			90% Prediction Interval Map	0.11	0.11	0.12	0.004
	AF Data	CU	Soil Prediction Map	0.12	0.17	0.17	0.013
			90% Prediction Interval Map	0.02	0.02	0.05	0.007
	FS Data	SVM	Soil Prediction Map	0.13	0.17	0.22	0.016
			90% Prediction Interval Map	0.003	0.01	0.01	0.001
F3	Provincial DSM	SVM	Soil Prediction Map	0.12	0.13	0.15	0.004
			90% Prediction Interval Map	0.09	0.10	0.11	0.003
	AF Data	CU	Soil Prediction Map	0.07	0.11	0.17	0.020
			90% Prediction Interval Map	0.05	0.05	0.06	0.003
	FS Data	CU	Soil Prediction Map	0.03	0.11	0.19	0.019
			90% Prediction Interval Map	0.02	0.02	0.03	0.001
F4	Provincial DSM	SVM	Soil Prediction Map	0.13	0.13	0.15	0.007
			90% Prediction Interval Map	0.10	0.10	0.12	0.006
	AF Data	CU	Soil Prediction Map	0.07	0.10	0.12	0.012
			90% Prediction Interval Map	0.04	0.05	0.09	0.013
	FS Data	SVM	Soil Prediction Map	0.08	0.11	0.14	0.019
			90% Prediction Interval Map	0.004	0.04	0.07	0.025
F5	Provincial DSM	SVM	Soil Prediction Map	0.13	0.14	0.17	0.009
			90% Prediction Interval Map	0.10	0.11	0.13	0.007
	AF Data	CU	Soil Prediction Map	0.11	0.13	0.16	0.010
			90% Prediction Interval Map	0.004	0.09	0.12	0.024
	FS Data	SVM	Soil Prediction Map	0.03	0.12	0.19	0.018
			90% Prediction Interval Map	0.02	0.03	0.05	0.003
F6	Provincial DSM	SVM	Soil Prediction Map	0.13	0.14	0.17	0.009
			90% Prediction Interval Map	0.10	0.11	0.13	0.007
	AF Data	CU	Soil Prediction Map	0.11	0.13	0.16	0.010
			90% Prediction Interval Map	0.004	0.09	0.12	0.024
	FS Data	SVM	Soil Prediction Map	0.03	0.12	0.19	0.018
			90% Prediction Interval Map	0.02	0.03	0.05	0.003

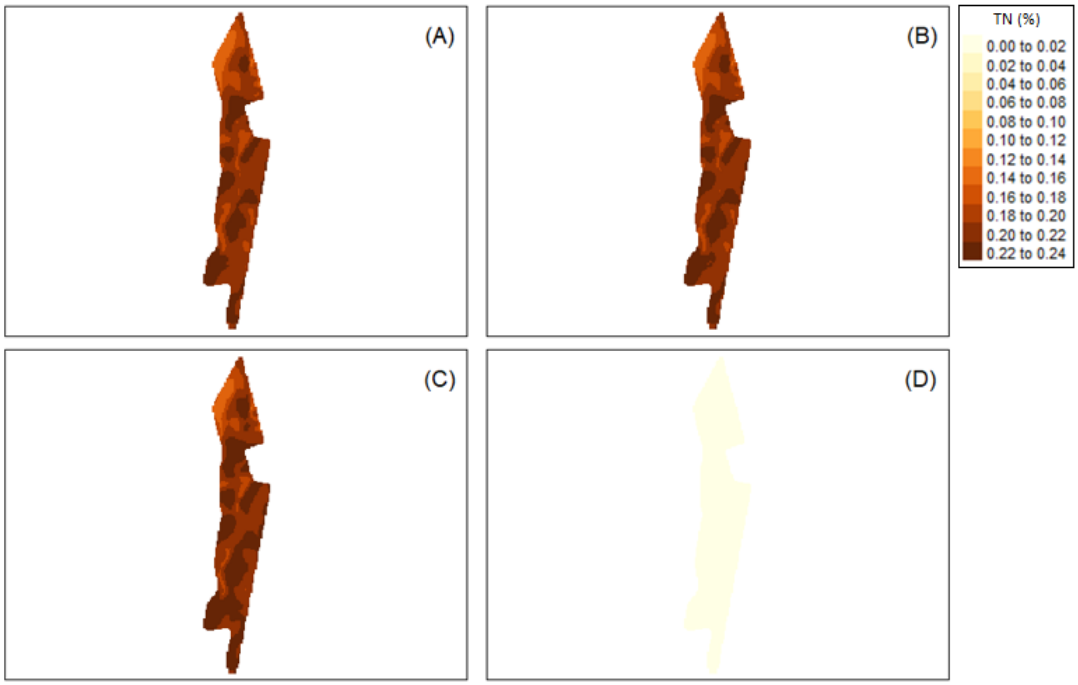


Figure 5.15 Soil total nitrogen (TN) maps (%) of best concordance field (F2) in the study area using the support vector machine learner and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

While in most cases, provincial DSM estimates were above TN for each field of the study area (Figure 5.11), F2 was the only field in which observed TN values were higher than the provincial DSM prediction (Figure 5.11, Table 5.11). As such, F2 appears to have had the strongest N capability among the fields in the study area, and was more similar in quality to the soils observed in the provincial SQMD (Figure 5.3). Except for F1, maps trained using FS data had the lowest 90% uncertainty PIs in comparison with AF trained maps (Figure 5.16, Table 5.11). Further, uncertainty was the greatest in field predictions from the provincial DSM of TN. It appears that for TN, prediction uncertainty increases with lower map resolution and with training data used from outside the geographic area.

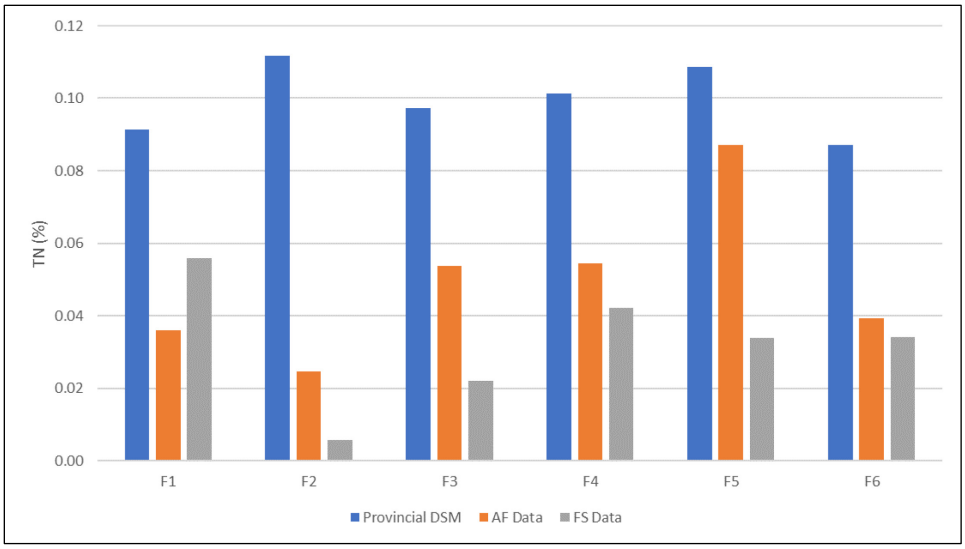


Figure 5.16 Comparison of mean total nitrogen (TN) 90% prediction interval uncertainty estimates from the provincial digital soil map (DSM), and novel maps modelled from all-fields (AF) and field-specific (FS) data for each field (F1 through F6) in the study area.

5.3.5.2 Biological Nitrogen Availability

Field scale maps of BNA modelled from AF and FS data were produced for fields F1 through F6 of the study area. A total of 48 infield maps, including 12 prediction and 36 uncertainty maps, were generated, and are summarized in Table 5.12. The best CCC results for BNA (Figure 5.14 and Table 5.9) were from F2 of the study area (0.74) using the SVM learner with FS data and seven predictor variables. The final BNA predictive map of F2 (Figure 5.17B) had a range from 15.8 to 38.3 mg N/kg BNA, a mean of 21.6 mg N/kg and a SD of 2.4 mg N/kg BNA (Table 5.12). The lower prediction limit (5^{th} percentile) ranged from 15.6 to 38.1 mg N/kg BNA (Figure 5.17A) and the upper prediction limit (95^{th} percentile) ranged from 15.4 to 42.4 mg N/kg BNA (Figure 5.17C). The overall 90% PI width ranged from -0.2 to 4.3 mg N/kg (Figure 5.17D) with a mean of 1.0 and a SD of 0.5 mg N/kg BNA (Table 5.12). For the remaining fields, F1 and F3

through F6, the best CCC results were also obtained using FS data, and are shown in the Appendix (Figure A.6 to A.10).

Table 5.12 Descriptive statistics (minimum (Min), Mean, maximum (Max) and standard deviation (SD) values) for biological nitrogen availability (BNA, mg N/kg) soil prediction and 90% prediction interval maps for the study area fields (F1 through F6) including extracted results from the provincial scale digital soil map (DSM) and novel maps trained with all-field samples (AF) and field-specific (FS) samples and including machine learners (SGB = stochastic gradient boosting; CU = cubist; SVM = support vector machine; RF = random forest).

	Map source	Learner	Spatial Interpretation	Min	Mean	Max	SD
F1	Provincial DSM	SGB	Soil Prediction Map	17.2	31.2	48.5	6.0
			90% Prediction Interval Map	31.1	38.3	47.2	3.1
	AF Data	SGB	Soil Prediction Map	9.3	17.8	29.2	2.7
			90% Prediction Interval Map	15.5	15.7	15.9	0.1
	FS Data	CU	Soil Prediction Map	8.3	18.1	38.0	4.3
			90% Prediction Interval Map	-0.1	1.1	1.8	0.3
F2	Provincial DSM	SGB	Soil Prediction Map	25.8	36.2	54.7	6.6
			90% Prediction Interval Map	35.5	40.8	50.3	3.4
	AF Data	SGB	Soil Prediction Map	13.5	21.5	29.5	1.9
			90% Prediction Interval Map	7.2	8.2	9.2	0.2
	FS Data	SVM	Soil Prediction Map	15.8	21.6	38.3	2.4
			90% Prediction Interval Map	-0.2	1.0	4.3	0.5
F3	Provincial DSM	SGB	Soil Prediction Map	13.8	26.8	40.7	5.7
			90% Prediction Interval Map	29.3	36.0	43.1	2.9
	AF Data	SGB	Soil Prediction Map	8.6	14.4	24.5	2.1
			90% Prediction Interval Map	1.6	8.2	19.5	2.4
	FS Data	RF	Soil Prediction Map	10.8	14.0	20.7	2.5
			90% Prediction Interval Map	3.8	4.3	5.2	0.4
F4	Provincial DSM	SGB	Soil Prediction Map	22.4	35.7	50.4	6.2
			90% Prediction Interval Map	33.7	40.6	48.1	3.2
	AF Data	SGB	Soil Prediction Map	7.0	15.2	24.2	3.0
			90% Prediction Interval Map	9.8	10.7	11.8	0.3
	FS Data	RF	Soil Prediction Map	9.9	15.5	20.8	2.3
			90% Prediction Interval Map	6.0	6.6	7.1	0.2
F5	Provincial DSM	SGB	Soil Prediction Map	19.7	30.7	40.9	4.4
			90% Prediction Interval Map	32.4	38.0	43.2	2.3
	AF Data	SGB	Soil Prediction Map	9.7	16.3	24.8	2.2
			90% Prediction Interval Map	-3.3	3.3	8.4	1.7
	FS Data	SVM	Soil Prediction Map	-5.8	15.3	17.2	2.4
			90% Prediction Interval Map	2.3	10.9	106.1	10.9
F6	Provincial DSM	SGB	Soil Prediction Map	18.3	30.3	52.0	5.1
			90% Prediction Interval Map	31.7	37.8	49.0	2.6
	AF Data	SGB	Soil Prediction Map	8.9	15.2	22.0	2.3
			90% Prediction Interval Map	-2.6	15.8	35.5	6.6
	FS Data	SVM	Soil Prediction Map	10.6	21.3	102.8	4.5
			90% Prediction Interval Map	-29.4	8.8	13.9	2.1

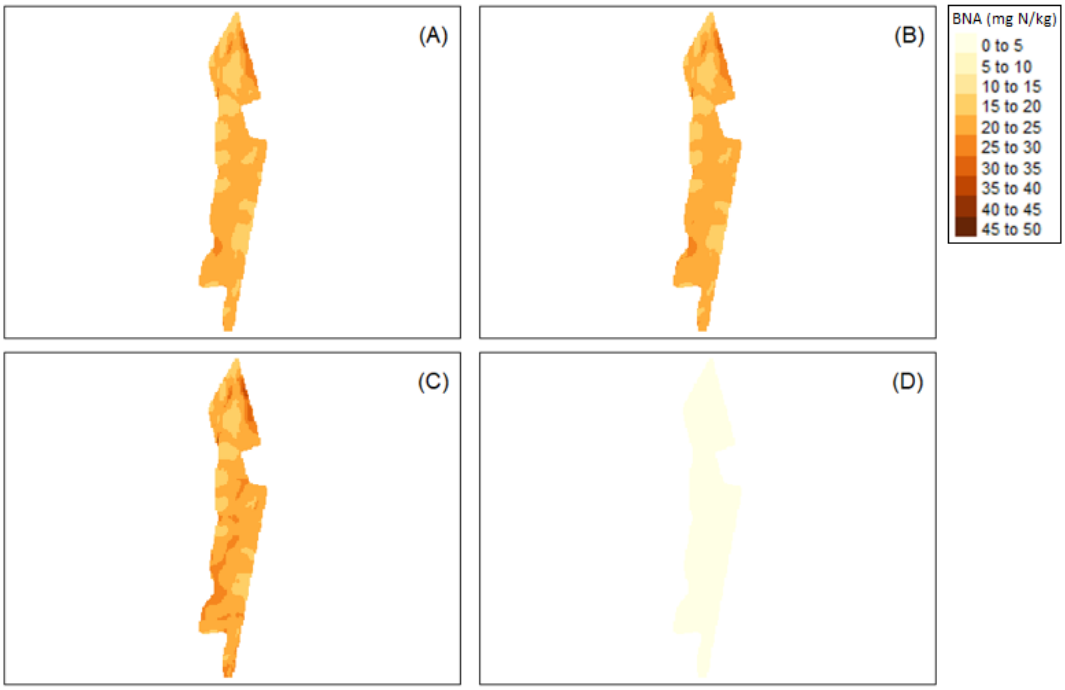


Figure 5.17 Biological nitrogen availability (BNA) maps (mg N/kg) of the best concordance field (F2) in the study area using the support vector machine (SVM) learner and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

Overall, maps trained using FS data had the lowest 90% uncertainty PIs, with the exception of F5 (Figure 5.18). The largest difference was observed in F2 between the provincial scale DSM mean uncertainty PI (40.8 mg N/kg) and the FS mean uncertainty PI (0.10 mg N/kg). Fields trained with AF data also showed a significant drop in uncertainty in BNA predictions from the provincial DSM, but were still of higher uncertainty than FS data. The improvement of AF predictions over provincial DSM predictions (trained from the SQMD) was likely due to the similarly managed fields present in the study area.

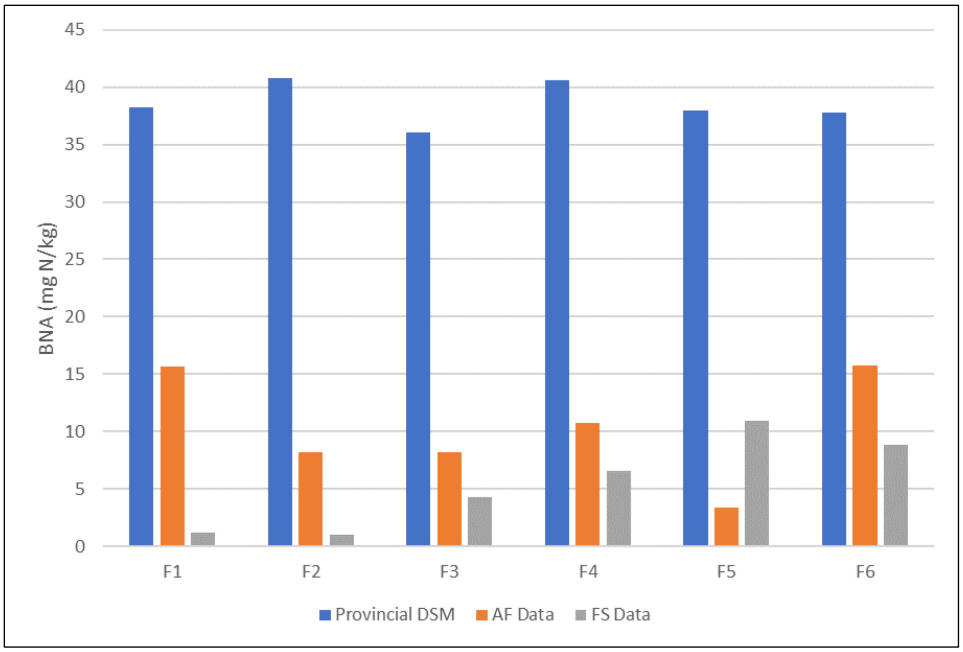


Figure 5.18 Comparison of mean biological nitrogen availability (BNA) 90% prediction interval uncertainty estimates from the provincial digital soil map (DSM), and novel maps modelled from all-fields (AF) and field-specific (FS) data for each field (F1 through F6) in the study area.

5.3.5.3 Growing Season Nitrogen

A total of 48 infield maps of GSN (12 prediction, and 36 uncertainty maps), modelled using both AF and FS data, were generated for the study area (fields F1 through F6). Descriptive statistics, including predictions from the provincial scale DSM of GSN, are summarized in Table 5.13.

The best model results, comparing both AF and FS results, was from field F4 (CCC = 0.77) using the CU model with FS data and two predictors (Figure 5.14, Table 5.10). The F4 prediction map of GSN (Figure 5.19B) ranged from 56.6 to 134.0 kg N/ha GSN, a mean of 94.2 kg N/ha GSN and a SD of 13.2 kg N/ha GSN (Table 5.13). The lower prediction limit (5th percentile) ranged from 37.2 to 38.1 kg N/ha GSN (Figure 5.19A) and the upper prediction limit (95th percentile) ranged from 74.6 to 142.2 kg N/ha

GSN (Figure 5.19C). The overall 90% PI width ranged from 20.9 to 37.4 kg N/ha GSN (Figure 5.19D) with a mean of 29.4 and a SD of 2.8 kg N/ha GSN (Table 5.13). The best results for the remaining fields (F1 to F3, F5 and F6) were achieved with FS data (Appendix, Figures A.11 to A.15).

Table 5.13 Descriptive statistics (minimum (Min), Mean, maximum (Max) and standard deviation (SD) values) for growing season nitrogen (GSN, kg N/ha) soil prediction and 90% prediction interval maps for the study area fields (F1 through F6) including extracted results from the provincial scale digital soil map (DSM) and novel maps trained with all-field samples (AF) and field-specific (FS) samples and including machine learners (SGB = stochastic gradient boosting; SVM = support vector machine; CU = cubist).

	Map source	Learner	Spatial Interpretation	Min	Mean	Max	SD
F1	Provincial DSM	SGB	Soil Prediction Map	97.5	133.9	189.3	16.9
			90% Prediction Interval Map	79.8	107.6	150.0	12.9
	AF Data	SGB	Soil Prediction Map	60.8	101.1	147.9	12.5
			90% Prediction Interval Map	0.1	43.7	94.4	13.5
	FS Data	SVM	Soil Prediction Map	13.3	100.1	172.8	21.9
			90% Prediction Interval Map	-44.7	27.4	87.6	18.1
F2	Provincial DSM	SGB	Soil Prediction Map	137.1	185.2	221.3	16.5
			90% Prediction Interval Map	110.1	146.8	174.5	12.6
	AF Data	SGB	Soil Prediction Map	83.7	127.4	157.2	11.1
			90% Prediction Interval Map	14.0	37.4	53.3	5.9
	FS Data	CU	Soil Prediction Map	15.8	21.6	38.3	2.4
			90% Prediction Interval Map	-0.2	1.0	4.3	0.5
F3	Provincial DSM	SGB	Soil Prediction Map	104.6	130.1	172.9	12.6
			90% Prediction Interval Map	85.2	104.7	137.5	9.6
	AF Data	SGB	Soil Prediction Map	54.0	91.1	138.9	15.1
			90% Prediction Interval Map	35.6	50.6	62.2	4.7
	FS Data	CU	Soil Prediction Map	61.0	87.6	138.4	15.6
			90% Prediction Interval Map	28.9	30.2	32.6	0.8
F4	Provincial DSM	SGB	Soil Prediction Map	103.1	144.3	208.7	26.5
			90% Prediction Interval Map	84.1	115.6	164.8	20.2
	AF Data	SGB	Soil Prediction Map	59.5	92.2	135.5	11.9
			90% Prediction Interval Map	27.4	54.4	74.8	7.4
	FS Data	CU	Soil Prediction Map	56.6	94.2	134.0	13.2
			90% Prediction Interval Map	20.9	29.4	37.4	2.8
F5	Provincial DSM	SGB	Soil Prediction Map	111.3	135.4	168.2	14.1
			90% Prediction Interval Map	90.3	108.8	133.9	10.8
	AF Data	SGB	Soil Prediction Map	55.5	95.8	133.4	11.9
			90% Prediction Interval Map	0.9	23.9	48.6	7.3
	FS Data	SVM	Soil Prediction Map	42.4	101.0	122.0	7.9
			90% Prediction Interval Map	3.1	38.6	51.3	4.8
F6	Provincial DSM	SGB	Soil Prediction Map	104.7	137.7	198.1	13.4
			90% Prediction Interval Map	85.3	110.6	156.7	10.3
	AF Data	SGB	Soil Prediction Map	54.6	87.7	127.0	10.8
			90% Prediction Interval Map	43.8	68.0	96.7	7.9
	FS Data	SVM	Soil Prediction Map	73.2	108.6	169.4	11.0
			90% Prediction Interval Map	-2.1	30.4	49.3	5.9

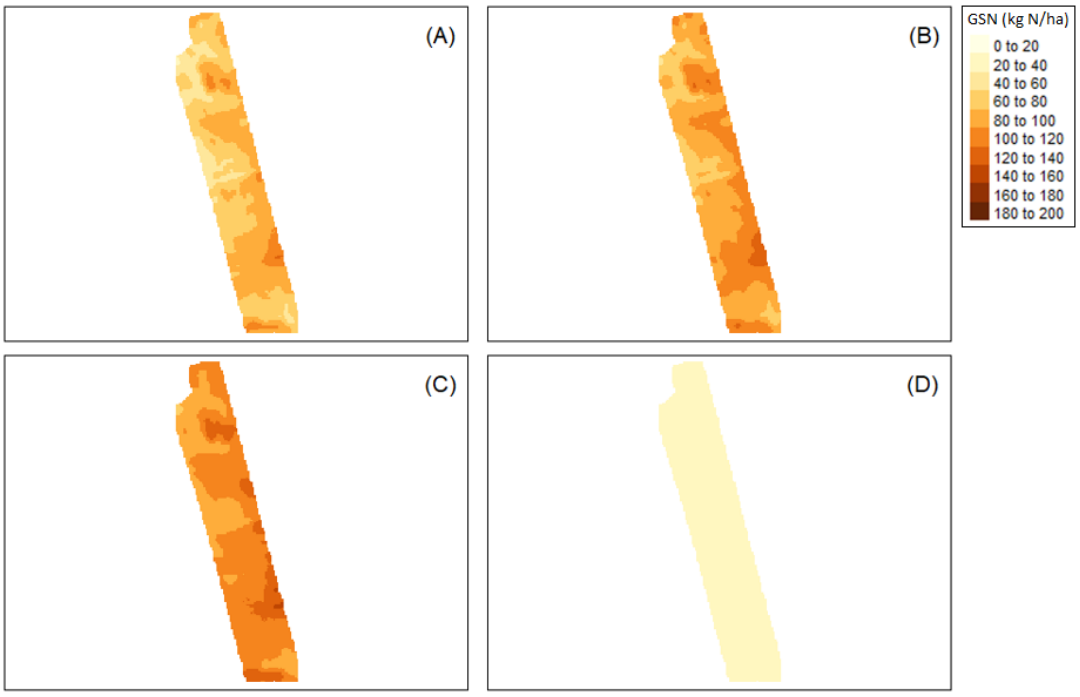


Figure 5.19 Growing season nitrogen (GSN) maps (kg N/ha) of the best concordance field (F4) in the study area using the cubist (CU) learner and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

In similarity to TN and BNA, the uncertainty of map predictions was reduced in most fields by using FS data alone (Figure 5.20). The relatively higher uncertainty of provincial scale DSMs applied at the local field level, as seen in Figure 5.20, highlights the need for finer resolution maps and the value of FS soil data to improve estimates of GSN for use in VRA N fertilizer recommendations. There was a substantial reduction in uncertainty using AF data from the study area; however, except for F5, FS data showed the best overall reduction in prediction uncertainty (Figure 5.20). The most drastic reduction in uncertainty was observed with the 90% uncertainty PI of F2 in the provincial DSM at 147 kg N/ha GSN reduced to 1.0 kg N/ha GSN in the FS data map (Figure 5.20).

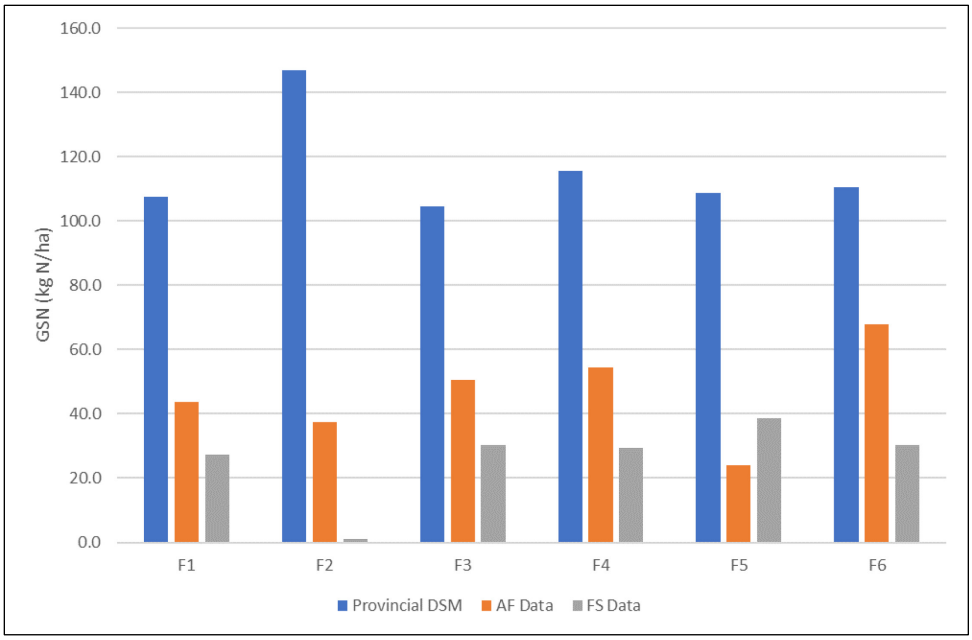


Figure 5.20 Comparison of mean growing season nitrogen (GSN) 90% prediction interval uncertainty estimates from the provincial digital soil map (DSM), and novel maps modelled from all-fields (AF) and field-specific (FS) data for each field (F1 through F6) in the study area.

5.3.6 Infield Predictors for N Indices

In order to improve understanding of N dynamics at the field scale, predictors and dominant *scorpan* factors selected from infield mapping of TN, BNA, and GSN were considered (Figure 5.1, Step 7). Interpretation of predictors was conducted using variable importance, *scorpan* group importance, and ALE plots (Molnar et al., 2018), based on modelling results outlined in Section 5.3.4.

5.3.6.1 Total Nitrogen

The stable N pool, which was estimated using TN, was best predicted using FS data (Figure 5.14) and relief variables (Figure 5.21, Table 5.8). Apart from F1, relief had the highest variable and group importance (Figure 5.21A and 5.21B) in the majority of fields in the study area. Specific relief variables showing importance (Figure 5.21A), as

well as high correlations (Table 5.8), included the eastness aspect (R.E.asp) with the highest correlation (-0.53) followed by multiresolution ridge top flatness (R.rtf, -0.30), slope height (R.sh, -0.22), and topographic position index (R.TPIn, -0.19). Interpreted from ALE plots (Figure 5.21C), results suggest that there was increased TN with increasing shade and water collecting scenarios related to slope and landscape flatness (Gallant and Dowling, 2003).

In comparison to TN predictions at the provincial scale, relief had the least importance relative to climate and organism *scorpan* groups (Chapter 4). With the fact that micro-climates were a function of relief (Liu et al., 2023), and given that typical climate indices were not applicable at the fine scales, the importance of micro-climate (relief) at the infield scale was thus in agreement with provincial findings. Relief, as a *scorpan* factor, was perhaps best expressed at a field scale (5 m resolution) since DSM estimates in Chapter 4 averaged relief over a larger area (30 m resolution) and therefore dilute the effects of relief locally. In the literature, terrain has also been found to be an important control of spatial distributions of TN, in addition to organisms (vegetative cover) via NDVI layers (Wang and Zhao, 2017; Zhou et al., 2020). In this study, organisms, primarily plants as represented by crop cover (O.NDVI20 and O.NDVI22) were also observed as important (Figure 5.21A and 5.21B) with correlations of 0.17 and 0.44, respectively (Table 5.8). With positive correlations between TN and vegetative growth at the infield (proximal) scale, there was also agreement at the provincial DSM (distal scale) and with related studies (Parsaie et al., 2021). Wang et al. (2018) and Zhou et al. (2020), who mapped TN at the regional scale, also found NDVI to be of importance. Lastly, soil color (S.SCN) was observed as important for predicting infield

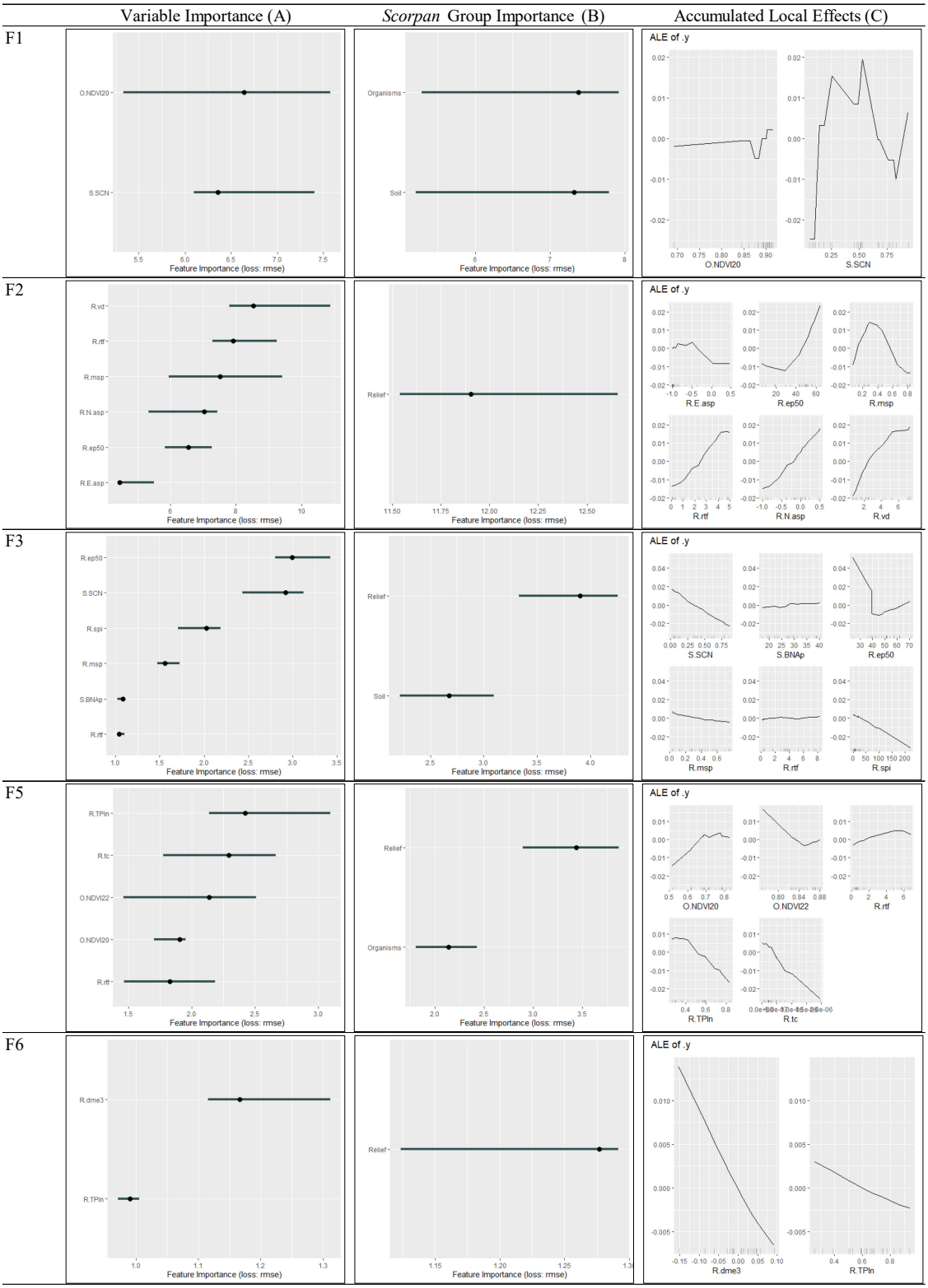


Figure 5.21 Variable importance, *scorpan* group importance, and accumulated local effects (ALE) plots of total nitrogen (TN = .y) field-specific (FS) modelling for each field (F1 through F6) of the study area (plots for F4 not included as only one relief variable was required for prediction). Refer to Table 5.8 for covariate descriptions and abbreviations.

TN in F1 and F3 (Figure 5.21A). With a negative correlation between TN and S.SCN (Table 5.8, -0.18), there was a logical connection with N dynamics in that as soil brightness increases (e.g., from less OM via erosion), the percentage TN decreases.

5.3.6.2 Biological Nitrogen Availability

Infield predictors of BNA, as a measure of the labile N pool, showed high importance related to relief and soil *scorpan* factors. As noted with TN, local relief as a driver of micro-climatic conditions also had a strong impact on BNA (Figure 5.22); in particular, eastness aspect (R.E.asp) and northness aspect (R.N.asp) with correlations of -0.25 and 0.14, respectively (Table 5.9). In addition to aspect, slope factors also showed importance in a similar way to TN and the stable N pool.

Provincial scale DSM predictions of BNA also showed relief as important, but with an especially strong relationship to ACI soil management layers from Chapter 4. Infield variation of crop frequency was only applicable in fields where mixed-cropping was practiced; therefore, ACI soil management infield covariate layers cannot be compared with regional predictions. Interestingly however, the inclusion of soil variables did show importance at the field scale, specifically with the soil color (S.SCN) layer (Figure 5.22, Table 5.9). Soil color, having the highest correlation with BNA (Table 5.9, -0.34), suggesting that BNA decreases with lighter colored soils, shows the impact of reduced OM levels in soil. Soil color, especially in PEI, was related to eroded soil conditions wherein lighter colored areas of a field indicate areas of increased erosion (Conforti et al., 2013). Erosion, as a negative influence on N indices, was inferred in Chapter 4 and in focused studies of the effects that erosion has on OM and nutrient loss (Edwards et al., 1998; Nyiraneza et al., 2017).

Figure 5.22 Variable importance (A), *scorpan* group importance (B), and accumulated local effects (ALE) plots (C) of biological nitrogen availability (BNA = .y) field-specific (FS) modelling for each field (F1 through F6) of the study area (plots for F5 are not included as only one soil variable was required for prediction). Refer to Table 5.9 for covariate descriptions and abbreviations.

5.3.6.3 Growing Season Nitrogen

GSN, as calculated from TN and BNA (Eq. 5.1 and 5.2), showed the strongest correlations with soil *scorpan* variables overall (Figure 5.23, Table 5.10). Using the provincial scale DSM as a covariate layer (S.GSNp), this covariate had the highest correlation (0.46) with calculated GSN values in the study area (Table 5.10). Soil color (S.SCN), as the other soil variable used for prediction, and with a correlation of -0.31, had high importance with infield predictions of GSN (Figure 5.23, Table 5.10). As the top predictor in fields F4 and F6 (Figure 5.23), soil color demonstrates the connection between OM, or its lack via eroded soils, and GSN dynamics.

At the provincial scale, GSN was best predicted with relief and organism *scorpan* groups (Chapter 4). This study showed similar findings in the majority of fields with relief as the best predictor group in four of the six fields (F2, F3, F5, and F6) and organisms showing relative importance in all fields but F2 (Figure 5.23). With respect to correlations as a group (Table 5.10), organism variables (i.e., plant cover) showed strong correlation with multiple NDVI layers, and relief layers especially eastness aspect (-0.41) and wetness index (0.31). In terms of GSN, there was similarity with TN and BNA, and the connection between increased GSN with increased shade and reduced slopes.

5.3.6.4 Infield versus provincial scale

Overall, relief appears to have the strongest connection and control of N indices, both stable and labile, at the field scale. Relief, comprising variations of slope and aspect and with its effect on moisture regimes and erosion potential, strongly influences soil climate at the proximal scale. At the provincial scale, TN was mainly driven by climate and BNA was mainly driven by organisms and relief (Chapter 4). With the connection

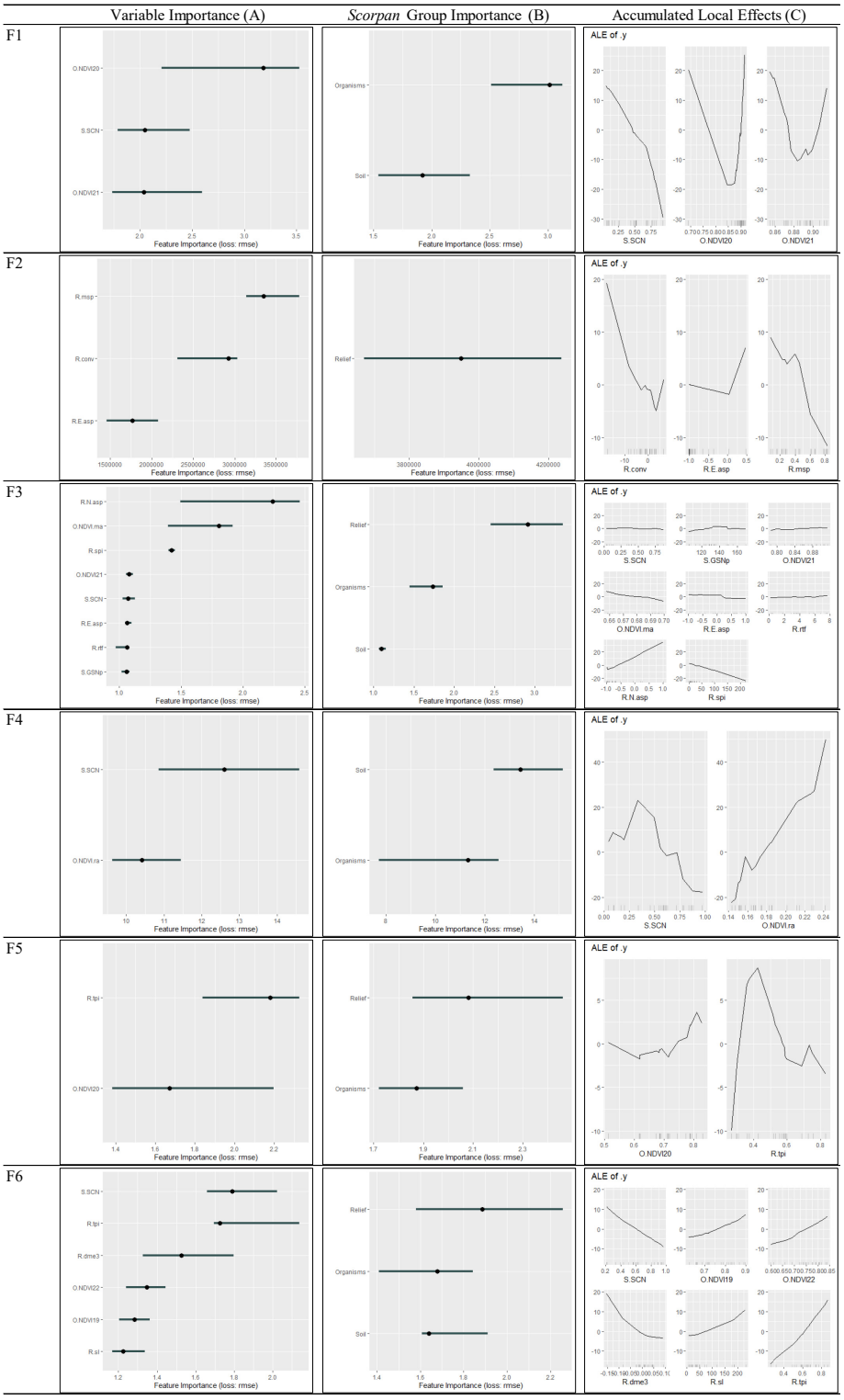


Figure 5.23 Variable importance (A), *scorpan* group importance (B), and accumulated local effects (ALE) plots (C) of biological nitrogen availability (BNA = .y) infield (IF) modelling for each field (F1 through F6) of the study area (plots for F5 are not included as only one soil variable was required for prediction). Refer to Table 5.10 for covariate descriptions and abbreviations.

between relief and local soil climate, there appears to be no contradiction as to important influences on N dynamics. Vegetation cover, shown by NDVI importance, and OM levels or the degree of erosion shown by the importance of soil colour, speak to the influence of soil management at both the provincial and infield scales.

The variability of relief within a landscape, as well as the connection between relief and erosion potential in PEI, identifies the infield variability of N dynamics that could not be captured at the provincial (30 m resolution) scale. As such, infield mapping of N parameters including TN, BNA, and GSN in order to capture the influence of relief and micro-climate variation in VRA scenarios, is strongly encouraged.

5.3.7 Application of Decision Support Tools

DSTs, including both PTFs and DSMs, were developed for the intention of making information on N pools and GSN mineralization accessible to producers to inform N fertilizer management. Based on results, there appears a logical framework, or hierarchy of options, available to the producer related to non-spatial and spatial applications.

5.3.7.1 Point prediction applications

The two N pool predictive function (Chapter 2), which offers estimates of GSN by accounting for the stable and labile N pools via surrogate measures of TN and BNA, respectively, was the basis of (non-spatial) sample point-predictions (Dessureault-Rompre et al., 2015), and the optimum device for informing N fertilizer recommendations. From a practical perspective, the output units of the GSN calculation, being in Kg N/ha, provide easily relatable and transferable results to assist N fertilizer recommendations. Also, being based on direct soil observations of TN and BNA, and

providing estimates in conformity with plant bioassay approaches in similar soils (Nyiraneza et al., 2022), the zero- plus first-order regression equation (Eq. 5.1 and 5.2) should be considered as optimal over estimates based on surrogate measures.

Where direct measures of TN and BNA do not exist, non-spatial (point sample) estimates from surrogate measures can be obtained via novel PTFs developed in Chapter 3 or Section 5.3.2. Based on the variability observed in BNA, the PTFs for GSN should be preferred as opposed to using the PTFs for TN, and BNA separately for input into Eq. 5.1 and 5.2. PTFs for GSN showed a predictive stability, not seen with BNA alone, and are thus recommended. As such, the most accessible PTF due to data availability and transferability of model coefficients, was the MLR-GSN PTF (CCC = 0.63), which included coefficients for OM, pH, and CEC (available in the standard S3-package; Section 5.3.2.5, Table 5.6). However, due to the potential of specific fields having a poorer soil quality than the provincial scale (SQMD) benchmark, the producer should first qualify how their specific fields compare to published provincial averages and adjust accordingly.

5.3.7.2 Spatial applications

Considering spatial estimates of GSN for inclusion in N fertilizer recommendations, FS data achieved the best prediction results for infield mapping (Section 5.3.5.3), as compared to provincial DSM predictions (Chapter 4), which were optimal at the regional scale (Section 5.3.3). Where possible, georeferenced infield sampling points, and analysis of TN and BNA parameters in order to generate spatial predictions of GSN is recommended in VRA scenarios. Secondarily, where direct soil measures of TN, BNA, or the S3-package are not available, provincial scale DSM

predictions are recommended for GSN estimations at a landscape scale (i.e., SRA scenarios).

It was observed in Section 5.3.3 and 5.3.5 that provincial scale DSM predictions of GSN (30 m resolution) were not able to capture infield dynamics for VRAs to the same degree as infield mapping (5 m resolution) with FS data. From a coarse-scale/landscape perspective, the provincial DSM provided a reasonable "ball-park" estimate of GSN predictions and is ideal for situations where a producer has no point data available in order to use novel PTFs. Benchmarking, with reference to the quality of a producer's own field in comparison with published soil averages and/or provincially derived DSMs, is also recommended to achieve optimized N fertilizer applications.

5.4 CONCLUSIONS

Provincially derived DSTs, including PTFs and DSMs, provided an extremely useful and effective benchmark to assess how well local fields compare to provincial "average" conditions. DSTs were also able to offer estimates of N parameters in situations where a producer has "the wrong data", or no data at all. DSTs offered predictions of TN, BNA, and GSN that reflected the differences in field specific soil observations. The results therefore provide confirmation of the sensitivity of N parameters, and the value of provincial averages and provincially derived DSTs for guiding N management decisions. Fields F1 through F6 were clearly below the provincial average, and with the assistance of novel DSTs, there was an important benefit available for these producers to be able to confirm this, quantify it, and inform sustainable soil N management. Novel PTFs, trained with a reduced suite of predictors more accessible to PEI producers (i.e., the S3-package), proved capable of providing reasonable estimates

for informing N fertilizer recommendations. Spatial estimates of N parameters based on field-specific training data, outperformed results with all field samples considered together, and provincial DSM predictions. Infield dynamics were more successfully captured based on localized data at a finer resolution, and map uncertainty was also greatly reduced in comparison to provincial scales. Relief *scorpan* factors, followed by soil and organism (i.e., plant cover) variables showed the best predictive strength for TN, BNA, and GSN. Soil colour, a surrogate for OM loss via erosion processes, and the provincial scale DSM for GSN, had the best correlations for predicting GSN infield.

Based on results of this study, a practical framework for incorporating provincially derived DSTs can be inferred; namely, a hierarchal or tiered approach to N management. From a producer's perspective, seeking insight into estimates of GSN for informing N fertilizer recommendations, the first tier may consist of the provincial DSM. These maps offered coarse-scale estimates and would be recommended where a producer has no baseline data available other than the location of their field. For improved estimates, the second tier would include the use of PTFs for estimates of GSN. PTFs, from surrogate measures, based on direct soil data from a producer's field, provide a more relevant point-estimate then the generic DSM prediction. Lastly, the third tier might include direct soil data with measures of TN and BNA are available for use in the GSN predictive function (Eq. 5.1 and 5.2). This third tier could be applied for both SRAs, or more preferably, for VRAs based on infield mapping with field-specific point data. Future studies should consider applying this tiered approach to N management in a variety of N trials, and cropping situations for further optimization. Also, in terms of applying DSTs, partitioning of provincial datasets, increasing training data, and model

training based on soil management practice, might alleviate producer risks associated with DSM predictions on poor quality fields. The provincial database, and associated DSTs, were useful and effective in documenting that study area fields were below standard with respect to soil health, as reflected in poorer soil N supply. As an affirmation of the value of the provincial benchmark values, and the sensitivity of N parameters to soil management, DSTs provided both an effective means for identifying variability between and within producer fields, and an ability to quantify and inform N fertilizer recommendations.

CHAPTER 6: CONCLUSIONS

6.1 SYNTHESIS

The predictive function - parameterized with total nitrogen (TN) for the stable N pool, and biological nitrogen availability (BNA) for the labile N pool, to estimate 130-day growing season nitrogen (GSN) as described in Chapter 2 - was applied to develop decision support tools (DSTs) in Prince Edward Island (PEI) to increase the efficiency of N use, and reduce N losses to the environment. The main issues facing PEI producers are an inadequate N credit system and the absence of an N test to inform N fertilizer recommendations. DSTs, in the form of Pedotransfer functions (PTFs) and digital soil maps (DSMs) developed using machine learning (ML), provided a means for non-spatial and spatial estimates of TN, BNA, and GSN (N response variables) to support N fertilizer management decisions.

A framework and PTFs were developed using ML and multiple-linear regression (MLR), to make sample point predictions of N response variables and gain insight on N pool dynamics from correlated predictor variables (Chapter 3). In addition to standard recursive feature elimination (RFE) practices, a novel feature elimination method was used to obtain optimum and cost-efficient predictor variables (CBFE) without a significant reduction in predictive accuracy. For the stable and labile N pools, a Lin's concordance correlation coefficient (CCC) of 0.80 and 0.78 was achieved for predictions of TN and BNA, respectively. The highest CCC of 0.82, for GSN, was obtained using aggregate stability, active carbon, soil respiration, organic matter, and pH as predictor variables. The cubist ML outperformed other learners in predicting N response variables,

and a maximum cost reduction of 49% was achieved with CBFE. Using this framework, novel PTFs generated reliable predictions of N response variables for possible use in N fertilizer recommendations. Based on the findings, the soil health suite of parameters used for prediction was able to capture intrinsic soil N controls at sample point (non-spatial) locations. The framework and PTFs developed may best be suited for laboratory use to identify "best predictors" for N response variables, and to incorporate into producer laboratory reports for supporting N management decisions.

Spatial estimates at the provincial scale, modelled from the province wide soil quality monitoring database (SQMD), were obtained through DSMs of N response variables (Chapter 4). Novel soil management layers, using multi-year crop frequency covariates, were instrumental in capturing important controls on N dynamics and achieved a maximum increase of 27% in CCC. TN, representing the stable N pool, was primarily controlled by climate variables (CCC = 0.45); whereas, BNA, representing the labile N pool, was best predicted with cropping and topographic variables (CCC = 0.45). GSN had the highest CCC (0.47) using the stochastic gradient boosting ML with multi-year crop frequency covariates as the most important predictors. The ability to achieve a provincial scale DSM of N response variables, with prediction uncertainty estimates, provides the producer with a strong and useful predictive model of GSN when direct soil measures are not available. In addition, the insight gained on potential drivers of the stable and labile N pools can help the management of soil N stocks for reducing N losses.

PTFs and DSMs derived from provincial scale soils data were then applied to six low productivity potato fields to qualify how DSTs might be used in practice, and on fields of lower soil quality (Chapter 5). Based on the results, predicted soil N

mineralization (N_{min}) potential from PTFs and DSMs were successful at qualifying field performance (i.e., as a benchmark comparison of N response variables) at the field scale. To increase the accessibility of N estimates, PTFs were developed from the limited suite of parameters commonly used in PEI (the "S3-package"). Using ML, the PTFs had a CCC of 0.83 for TN, 0.57 for BNA, and 0.70 for GSN. DSTs were capable of identifying fields with lower soil N_{min} potential, and gave useful approximations of N response variables at the landscape or whole-field scales (e.g., for single rate N applications). To assist producers using variable rate application of N fertilizer, infield maps were generated using ML and field-specific data. Infield maps showed the greatest prediction accuracy as well as the lowest uncertainty of predictions. Using field specific data, infield maps of TN and BNA achieved CCC results as high as 0.83 and 0.74, respectively, with the support vector machine learner. GSN results, with a CCC of 0.77 with the cubist learner, showed that soil-based covariates (such as soil colour and the provincial scale map of GSN) were the most important and highly correlated predictors. From the testing of DSTs at the field scale (Chapter 5), it was concluded that a tiered implementation approach could maximize the strengths for each DST in order to assist with N fertilizer recommendations.

6.2 APPLICATIONS

Sections 5.3.7 and 5.4 suggested a tiered hierarchy for implementing predictive DSTs. The specific objectives, outlined in Chapter 1 and Chapter 2 (Section 2.6), addressed the impediments that producers face when making N fertilizer recommendations; namely, no direct soil data, insufficient data, and the need for an updated system of incorporation. The following provides an example application, which

was informed from cumulative thesis results (Chapters 2 through 5). The study area from Chapter 5, including data from six producer fields (Fields F1 through F6), were used for demonstration.

6.2.1 Tier 1

This initial step, of a proposed tiered approach to N management, was based on the provincial scale DSM and is considered as an entry level DST for producers who do not have existing soil measurements. Table 6.1, Line 1 is populated with the average predictions of GSN from within the field boundary at the geographic location for fields F1 through F6. The DSM (Chapter 5) best performed at the whole-field/landscape scale and provided acceptable uncertainty estimations with a mean 90% prediction interval (PI) of 126 kg N/ha, or approximately +/- 63 kg N/ha of the predicted value. The degree of uncertainty in the GSN estimate is a function of limited field measurements, and demonstrate a need for further direct soil measures (training data) of TN and BNA as it becomes available. To translate the predicted GSN value from the provincial DSM into an N credit (Chapter 3 to 5), Line 2 (Table 6.1) applies a 60% "Tier 1 risk factor" to reflect the lower limit of the uncertainty PI (Section 4.4.2.3), and obtain a conservative estimation in response to this uncertainty. It is anticipated that the risk factor would become smaller (move closer to 1.0) over time and with increased training data. The active growing season adjustment (Line 3) acknowledges that plant uptake does not occur for the complete 130-day estimate of the growing season, but for approximately 70% of the 130-day growing season between May to August for potatoes (Section 2.2.4). Preferably, the predictive function could be adjusted to calculate either a 90-day uptake, or an acceptable duration for the period leading up to the most rapid plant uptake. Lastly,

a factor that considers the nitrogen use efficiency of GSN, estimated at approximately 50%, is based on the and is in conformity to other N fertilizer sources (Chapter 2 and Section 3.2). This factor could also be adjusted to reflect more efficient on-farm N management practices.

Table 6.1 Summary of Tier 1 to Tier 3 nitrogen (N)-credit options for fields in the study area (F1 through F6) based on predictions of the provincial digital soil map (DSM), the pedotransfer functions (PTFs) from the S3 soil analytical package (including soil organic matter, OM; pH; and cation exchange capacity, CEC), and soil health (SH) analytical package (including total nitrogen, TN; and biological nitrogen availability, BNA).

Line #	Options	Source	Description	F1	F2	F3	F4	F5	F6
1	Tier 1	DSM provincial scale	Tier 1 GSN, field average - kg N/ha	134	185	130	144	135	138
2			*Tier 1 risk factor*	*0.6*	*0.6*	*0.6*	*0.6*	*0.6*	*0.6*
3			*Active growing season adjustment*	*0.7*	*0.7*	*0.7*	*0.7*	*0.7*	*0.7*
4			*Nitrogen use efficiency*	*0.5*	*0.5*	*0.5*	*0.5*	*0.5*	*0.5*
5			**Adjusted Tier 1: N-Credit (kg N/ha)**	**28**	**39**	**27**	**30**	**28**	**29**
6	Tier 2	PTF from S3-package	Average OM (%)	2.1	3.2	2.1	2.1	2.5	2.0
7			Average pH	6.5	6.7	6.1	5.9	6.4	5.6
8			Average CEC	8.0	9.2	10.1	10.4	10.9	9.8
9			Tier 2 GSN - kg N/ha	132	190	124	122	147	114
10			*Tier 2 risk factor*	*0.7*	*0.7*	*0.7*	*0.7*	*0.7*	*0.7*
11			*Active growing season adjustment*	*0.7*	*0.7*	*0.7*	*0.7*	*0.7*	*0.7*
12			*Nitrogen use efficiency*	*0.5*	*0.5*	*0.5*	*0.5*	*0.5*	*0.5*
13			**Adjusted Tier 2: N-Credit (kg N/ha)**	**32**	**47**	**30**	**30**	**36**	**28**
14	Tier 3	SH-package	Average TN (%)	0.11	0.17	0.10	0.11	0.13	0.11
15			Average BNA (mg N/ha)	18.3	22.2	13.6	15.3	15.9	23.1
16			Tier 3 GSN - kg N/ha	102	131	86	94	101	115
17			*Tier 3 risk factor*	*0.8*	*0.8*	*0.8*	*0.8*	*0.8*	*0.8*
18			*Active growing season adjustment*	*0.7*	*0.7*	*0.7*	*0.7*	*0.7*	*0.7*
19			*Nitrogen use efficiency*	*0.5*	*0.5*	*0.5*	*0.5*	*0.5*	*0.5*
20			**Adjusted Tier 3 N-Credit (kg N/ha)**	**29**	**46**	**30**	**33**	**36**	**40**

After applying the correction factors, the adjusted Tier 1 N credit (Table 6.1, Line 5) would be used by producers for fields F1 through F6. This N credit, in kg N/ha, could be directly applied to offset alternative N fertilizer sources for single-rate application (SRA) scenarios. Due to the difference in the factors that were dominant in predicting GSN at regional vs. infield scales, the provincial DSM is not recommended as a sole data-source for infield variable rate applications; however, the DSM is ideal for baseline estimates or as a covariate layer, for creating N management zones.

6.2.2 Tier 2

The DST for the second-tier could account for situations where producers have direct soil data from their fields, but where analysis does not include TN or BNA. Based on the framework established in Chapter 3, PTFs were created in Chapter 5 from the standard soil testing suite offered at the PEI Analytical Laboratory, and called the S3-package. Organic matter (OM), pH, and cation-exchange capacity (CEC) are all included in the S3-package and were used as predictor variables. Lines 6-8 (Table 6.1) allow producers to directly input soil results for OM, pH, and CEC, respectively from each field. It is notable, that with the N credit system currently in practice (Chapter 1 and 2), zero credits would have been applied to these fields since all OM results (Table 6.1, Line 6) were below the 3.5% requirement. The raw output (Line 9) is the predicted GSN as derived from multiple-linear regression (MLR-GSN) PTF (CCC = 0.63) and trained with the provincial SQMD (Chapter 5.3.2.5, Table 5.6). The 70% "Tier 2 risk factor" (Table 6.1, Line 10) allows for a higher N credit compared to Tier 1 (Line 2). This is due to the increased accuracy (CCC) associated with PTF predictions (being from direct soil measures) versus remotely sensed DSM predictions (Section 5.3.2). The active growing season adjustment (Line 11) and nitrogen use efficiency allowance (Line 12) were applied in similarity to Tier 1.

The adjusted Tier 2 N credit (Table 6.1, Line 13) allows for potentially higher N credits due to the increased confidence of using field specific soils data. The Tier 2 (S3-package) PTF could be used in a variety of applications including discrete soil results, composite soil results, or management-zone based soil results.

6.2.3 Tier 3

The third tier is proposed for use when a producer has direct soil measures of TN and BNA (Table 6.1, Lines 14 and 15, respectively) for input into the prediction function outlined in Chapter 3.3.3 (Eq. 3.1 and 3.2). Based on direct soil measures, the prediction function's output (Line 16) will offer estimates with higher confidence as compared to PTFs or DSMs. As such, the Tier 3 risk factor (Line 17), containing the lowest risk of the previous tiers, retains the highest percentage of the raw N credit prediction. The Tier 3 risk factor was estimated based on the accuracy associated with the prediction function itself (Section 2.2). After applying the active growing season adjustment (Line 18) and nitrogen use efficiency allowance (Line 19), as with Tiers 1 and 2, the adjusted Tier 3 N credit (Line 20) can be applied directly for fields F1 through F6 to inform SRA of N fertilizer recommendations.

6.2.4 Tier 4

For producers who have the capability for variable-rate N applications (i.e., proper equipment/implements), a fourth tier is proposed. In the previous tiers, the adjusted N credit outputs (Table 6.1, Lines 5, 13, and 20), are most conducive for SRA scenarios; however, where georeferenced field-specific (discrete) samples of TN and BNA plus applicable covariates (Section 5.3.4.4) are available, infield mapping is a viable option. Infield mapping, which accounts for infield variation in soil N_{min} potential, showed the highest accuracy and lowest uncertainty of predictions. Tier 4 mapping would therefore be considered the best option for minimizing RSN losses (Section 5.3.5.3).

6.3 LIMITATIONS AND RECOMMENDATIONS

The methods and DSTs developed in this thesis focused specifically on the soil-based N credit portion of PEIs N credit system (Chapter 2). As such, contributions from manure, compost, or previous legume crops, and their impact on the stable and labile N pools, were outside of this scope but are recommended for further study. Also, the effect of crop rotation or soil management, demonstrated by multi-year crop management covariates in Chapter 4, showed substantial importance in predicting N response variables and warrants further investigation. For example, reclassification of crop frequency layers could enhance predictive accuracies and provide further insight into N pool dynamics. Another aspect for consideration relates to increasing sample density of the SQMD, or creating management-specific DSMs. It is conjectured that with increased training data, or focused data from conventional potato operations, there may be a reduction in uncertainty estimates. However, perhaps the most critical aspect stemming from this research, is its implementation. In retrospect, with the disparity that exists between the current N credit system now in use, and the adjusted (tiered) N credits proposed, there is an opportunity to study how producers might receive this credit score, how would they respond to it, their willingness in adopting it, and what the outcome of that adoption might be. Research from the producer's perspective into understanding their risks and hesitancy, and further consideration of approaches to make producers more confident with the proposed system, would greatly compliment the thesis findings. It is promising that the PEI Federation of Agriculture has begun including this framework into their N balance decision support system as part of the delivery of the On Farm Climate Action Fund programming.

REFERENCES

Amanabadi, S., Vazirinia, M., Vereecken, H., Vakilian, K.A., Mohammadi, M., 2019. Comparative study of statistical, numerical and machine learning-based pedotransfer functions of water retention curve with particle size distribution data. Eurasian Soil Science 52(12), 1555-1571.

Amelung, W., Zech, W., Zhang, X., Follett, R.F., Tiessen, H., Knox, E., Flach, K.W., 1998. Carbon, nitrogen, and sulfur pools in particle-size fractions as influenced by climate. Soil Sci. Soc. Am. J. 62(1), 172-181.

Angers, D., Carter, M., 2020. Aggregation and organic matter storage in cool, humid agricultural soils, Structure and organic matter storage in agricultural soils. CRC Press, pp. 193-211.

Angst, G., Lichner, L., Csecserits, A., Emsens, W.-J., van Diggelen, R., Veselá, H., Cajthaml, T., Frouz, J., 2022. Controls on labile and stabilized soil organic matter during long-term ecosystem development. Geoderma 426, 116090.

Arbor, A., Schmidt, M., Saurette, D., Zhang, J., Bulmer, C., Filatow, D., Kasraei, B., Smukler, S., Heung, B., 2023. A framework for recalibrating pedotransfer functions using nonlinear least squares and estimating uncertainty using quantile regression. Geoderma 439, 116674.

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L., Panagos, P., 2019. Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. Geoderma 355, 113912.

Bartholomeus, H., Kooistra, L., Stevens, A., van Leeuwen, M., van Wesemael, B., Ben-Dor, E., Tychon, B., 2011. Soil organic carbon mapping of partially vegetated agricultural fields with imaging spectroscopy. International Journal of Applied Earth Observation and Geoinformation 13(1), 81-88.

Bassanino, M., Grignani, C., Sacco, D., Allisiardi, E., 2007. Nitrogen balances at the crop and farm-gate scale in livestock farms in Italy. Agriculture, ecosystems & environment 122(3), 282-294.

Benbi, D.K., Richter, J., 2002. A critical review of some approaches to modelling nitrogen mineralization. Biol. Fertil. Soils 35(3), 168-183.

Benites, V.M., Machado, P., Fidalgo, E.C.C., Coelho, M.R., Madari, B.E., 2007. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil. Geoderma 139(1-2), 90-97.

Benke, K.K., Norng, S., Robinson, N.J., Chia, K., Rees, D.B., Hopley, J., 2020. Development of pedotransfer functions by machine learning for prediction of soil electrical conductivity and organic carbon content. Geoderma 366, 114210.

Bonde, T.A., Rosswall, T., 1987. Seasonal variation of potentially mineralizable nitrogen in four cropping systems. Soil Sci. Soc. Am. J. 51(6), 1508-1514.

Bonde, T.A., Schnürer, J., Rosswall, T., 1988. Microbial biomass as a fraction of potentially mineralizable nitrogen in soils from long-term field experiments. Soil Biology and Biochemistry 20(4), 447-452.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers, Proceedings of the fifth annual workshop on Computational learning theory, pp. 144-152.

Botula, Y.-D., Nemes, A., Van Ranst, E., Mafuka, P., De Pue, J., Cornelis, W.M., 2015. Hierarchical pedotransfer functions to predict bulk density of highly weathered soils in Central Africa. Soil Sci. Soc. Am. J. 79(2), 476-486.

Bouma, J., 1989. Using soil survey data for quantitative land evaluation. Advances in soil sciences 9, 177-213.

Bowles, T.M., Atallah, S.S., Campbell, E.E., Gaudin, A.C., Wieder, W.R., Grandy, A.S., 2018. Addressing agricultural nitrogen losses in a changing climate. Nature Sustainability 1(8), 399-408.

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

Brenning, A., Bangs, D., Becker, M., Schratz, P., Polakowski, F., 2018. RSAGA: SAGA geoprocessing and terrain analysis. R package version 1.3. 0. Obtenido de https://CRAN. R-project. org/package= RSAGA (citado en pág. 23).

Brinkmann, W., 1979. Growing season length as an indicator of climatic variations? Climatic Change 2(2), 127-138.

Bruulsema, T., Fixen, P., Sulewski, G., 2016. 4R plant nutrition manual: A manual for improving the management of plant nutrition, North American version (revised). Norcross, GA: International Plant Nutrition Institute.

Cabrera, M., Kissel, D., 1988. Evaluation of a method to predict nitrogen mineralized from soil organic matter under field conditions. Soil Sci. Soc. Am. J. 52(4), 1027-1031.

Cassity-Duffey, K., Cabrera, M., Gaskin, J., Franklin, D., Kissel, D., Saha, U., 2020. Nitrogen mineralization from organic materials and fertilizers: Predicting N release. Soil Sci. Soc. Am. J. 84(2), 522-533.

Chataut, G., Bhatta, B., Joshi, D., Subedi, K., Kafle, K., 2023. Greenhouse gases emission from agricultural soil: A review. Journal of Agriculture and Food Research, 100533.

Chen, S., Richer-de-Forges, A.C., Saby, N.P.A., Martin, M.P., Walter, C., Arrouays, D., 2018. Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area. Geoderma 312, 52-63.

Cisty, M., Cyprich, F., 2020. Evaluation of Linear and Machine Learning Models for Determining Pedotransfer Functions, IOP Conference Series: Earth and Environmental Science. IOP Publishing, pp. 012083.

Clingensmith, C.M., Grunwald, S., 2022. Predicting soil properties and interpreting vis-NIR models from across Continental United States. Sensors 22(9), 3187.

Conforti, M., Buttafuoco, G., Leone, A.P., Aucelli, P.P., Robustelli, G., Scarciglia, F., 2013. Studying the relationship between water-induced soil erosion and soil organic matter using Vis–NIR spectroscopy and geomorphological analysis: A case study in southern Italy. Catena 110, 44-58.

Craney, T.A., Surles, J.G., 2002. Model-dependent variance inflation factor cutoff values. Quality engineering 14(3), 391-403.

Curtin, D., Campbell, C., 2008. Mineralizable nitrogen. Soil sampling and methods of analysis 2, 599-606.

Deragon, R., Heung, B., Lefebvre, N., John, K., Cambouris, A.N., Caron, J., 2024. Improving a regional peat thickness map using soil apparent electrical conductivity measurements at the field-scale. Frontiers in Soil Science 3, 1305105.

Deragon, R., Saurette, D.D., Heung, B., Caron, J., 2023. Mapping the maximum peat thickness of cultivated Organic soils in the southwest plain of Montreal. Can. J. Soil Sci.

Derrien, D., Barré, P., Basile-Doelsch, I., Cécillon, L., Chabbi, A., Crème, A., Fontaine, S., Henneron, L., Janot, N., Lashermes, G., 2023. Current controversies on mechanisms controlling soil carbon storage: implications for interactions with practitioners and policy-makers. A review. Agronomy for Sustainable Development 43(1), 21.

Dessureault-Rompre, J., Zebarth, B.J., Burton, D.L., Georgallas, A., 2015. Predicting soil nitrogen supply from soil properties. Can. J. Soil Sci. 95(1), 63-75.

Dessureault-Rompre, J., Zebarth, B.J., Burton, D.L., Grant, C.A., 2016. Depth distribution of mineralizable nitrogen pools in contrasting soils in a semi-arid climate. Can. J. Soil Sci. 96(1), 1-11.

Dessureault-Rompre, J., Zebarth, B.J., Burton, D.L., Gregorich, E.G., Goyer, C., Georgallas, A., Grant, C.A., 2013. Are Soil Mineralizable Nitrogen Pools Replenished during the Growing Season in Agricultural Soils? Soil Sci. Soc. Am. J. 77(2), 512-524.

Dessureault-Rompré, J., Zebarth, B.J., Burton, D.L., Sharifi, M., Cooper, J., Grant, C.A., Drury, C.F., 2010. Relationships among mineralizable soil nitrogen, soil properties, and climatic indices. Soil Sci. Soc. Am. J. 74(4), 1218-1227.

Dessureault-Rompre, J., Zebarth, B.J., Georgallas, A., Burton, D.L., Grant, C.A., 2011. A biophysical water function to predict the response of soil nitrogen mineralization to soil water content. Geoderma 167-68, 214-227.

Dessureault-Rompre, J., Zebarth, B.J., Georgallas, A., Burton, D.L., Grant, C.A., Drury, C.F., 2010. Temperature dependence of soil nitrogen mineralization rate: Comparison of mathematical models, reference temperatures and origin of the soils. Geoderma 157(3-4), 97-108.

Didan, K., 2021. MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V061 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. Accessed 2023-08-14 from https://doi.org/10.5067/MODIS/MOD13Q1.061

Dobermann, A., Bruulsema, T., Cakmak, I., Gerard, B., Majumdar, K., McLaughlin, M., Reidsma, P., Vanlauwe, B., Wollenberg, L., Zhang, F., Zhang, X., 2022. Responsible plant nutrition: A new paradigm to support food system transformation. Global Food Security 33, 100636.

Donatelli, M., Wösten, J., Belocchi, G., 2004. Methods to evaluate pedotransfer functions. Developments in soil science 30, 357-411.

Douglas, B.W., MacLeod, J.A., Mellish, T.M., Glen, W.M., Thompson, B.L., DeHaan, K.R., Sturz, A.V., Carter, M.R., Brimacombe, M.B., 2000. A method for measuring Prince Edward Island soil quality. Commun. Soil Sci. Plant Anal. 31(11-14), 1837-1845.

Edwards, L., Richter, G., Bernsdorf, B., Schmidt, R.-G., Burney, J., 1998. Measurement of rill erosion by snowmelt on potato fields under rotation in Prince Edward Island (Canada). Can. J. Soil Sci. 78(3), 449-458.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International journal of climatology 37(12), 4302-4315.

Fierer, N., Jackson, R.B., 2006. The diversity and biogeography of soil bacterial communities. Proc. Natl. Acad. Sci. U. S. A. 103(3), 626-631.

Fisette, T., Davidson, A., Daneshfar, B., Rollin, P., Aly, Z., Campbell, L., 2014. Annual space-based crop inventory for Canada: 2009–2014, 2014 IEEE Geoscience and Remote Sensing Symposium. IEEE, pp. 5095-5098.

Fisette, T., Rollin, P., Aly, Z., Campbell, L., Daneshfar, B., Filyer, P., Smith, A., Davidson, A., Shang, J., Jarvis, I., 2013. AAFC annual crop inventory, 2013 Second International Conference on Agro-Geoinformatics (Agro-Geoinformatics). IEEE, pp. 270-274.

Fowler, D., Coyle, M., Skiba, U., Sutton, M.A., Cape, J.N., Reis, S., Sheppard, L.J., Jenkins, A., Grizzetti, B., Galloway, J.N., Vitousek, P., Leach, A., Bouwman, A.F., Butterbach-Bahl, K., Dentener, F., Stevenson, D., Amann, M., Voss, M., 2013. The global nitrogen cycle in the twenty-first century. Philos. Trans. R. Soc. B-Biol. Sci. 368(1621), 13.

Frerichs, C., Glied-Olsen, S., De Neve, S., Broll, G., Daum, D., 2022. Crop Residue Management Strategies to Reduce Nitrogen Losses during the Winter Leaching Period after Autumn Spinach Harvest. Agronomy 12(3), 653.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 55(1), 119-139.

Friedman, J.H., 2002. Stochastic gradient boosting. Computational statistics & data analysis 38(4), 367-378.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water resources research 39(12).

Gebauer, A., Ellinger, M., Brito Gomez, V.M., Ließ, M., 2020. Development of pedotransfer functions for water retention in tropical mountain soil landscapes: spotlight on parameter tuning in machine learning. Soil 6(1), 215-229.

Georgallas, A., Dessureault-Rompre, J., Zebarth, B.J., Burton, D.L., Drury, C.F., Grant, C.A., 2012. Modification of the biophysical water function to predict the change in soil mineral nitrogen concentration resulting from concurrent mineralization and denitrification. Can. J. Soil Sci. 92(5), 695-710.

Georgiou, K., Jackson, R.B., Vindušková, O., Abramoff, R.Z., Ahlström, A., Feng, W., Harden, J.W., Pellegrini, A.F., Polley, H.W., Soong, J.L., 2022. Global stocks and capacity of mineral-associated soil organic carbon. Nature communications 13(1), 3797.

Gill, M.K., Asefa, T., Kemblowski, M.W., McKee, M., 2006. Soil moisture prediction using support vector machines 1. JAWRA Journal of the American Water Resources Association 42(4), 1033-1046.

Gillis, J.D., Price, G., 2016. Linking short-term soil carbon and nitrogen dynamics: environmental and stoichiometric controls on fresh organic matter decomposition in agroecosystems. Geoderma 274, 35-44.

Glendining, M.J., Dailey, A.G., Powlson, D.S., Richter, G.M., Catt, J.A., Whitmore, A.P., 2011. Pedotransfer functions for estimating total soil nitrogen up to the global scale. European Journal of Soil Science 62(1), 13-22.

Golden, H.E., Evenson, G.R., Christensen, J.R., Lane, C.R., 2023. Advancing Watershed Legacy Nitrogen Modeling to Improve Global Water Quality. Environmental Science & Technology.

Goovaerts, P., Chiang, C.N., 1993. Temporal persistence of spatial patterns for mineralizable nitrogen and selected soil properties. Soil Sci. Soc. Am. J. 57(2), 372-381.

Gordon, R., Bootsma, A., 1993. Analyses of growing degree-days for agriculture in Atlantic Canada. Climate Research, 169-176.

Govil, S., Lee, A.J., MacQueen, A.C., Pricope, N.G., Minei, A., Chen, C., 2022. Using Hyperspatial LiDAR and Multispectral Imaging to Identify Coastal Wetlands Using Gradient Boosting Methods. Remote Sensing 14(23), 6002.

Griffin, T., Honeycutt, C., Albrecht, S., Sistani, K., Torbert, H., Wienhold, B., Woodbury, B., Hubbard, R., Powell, J., 2007. Nationally coordinated evaluation of soil nitrogen mineralization rate using a standardized aerobic incubation protocol. Commun. Soil Sci. Plant Anal. 39(1-2), 257-268.

Griffin, T.S., 2008. Nitrogen availability. Nitrogen in agricultural systems 49, 613-646.

Gu, B., van Grinsven, H.J., Lam, S.K., Oenema, O., Sutton, M.A., Mosier, A., Chen, D., 2021. A credit system to solve agricultural nitrogen pollution. The Innovation 2(1).

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine learning 46(1), 389-422.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

He, Z., Larkin, R., Honeycutt, W., 2012. Sustainable potato production: global case studies. Springer Science & Business Media.

Hengl, T., Heuvelink, G.B., Kempen, B., Leenaars, J.G., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PloS one 10(6), e0125814.

Heumann, S., Bottcher, J., Springob, G., 2003. Pedotransfer functions for the pool size of slowly mineralizable organic N in sandy arable soils. Journal of Plant Nutrition and Soil Science 166(3), 308-318.

Heumann, S., Fier, A., Hassdenteufel, M., Hoeper, H., Schaefer, W., Eiler, T., Boettcher, J., 2013. Minimizing nitrate leaching while maintaining crop yields: insights by simulating net N mineralization. Nutrient Cycling in Agroecosystems 95(3), 395-408.

Heumann, S., Ringe, H., Boettcher, J., 2011. Field-specific simulations of net N mineralization based on digitally available soil and weather data: II. Pedotransfer functions for the pool sizes. Nutrient Cycling in Agroecosystems 91(3), 339-350.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62-77.

Heuvelink, G., Webster, R., 2001. Modelling soil variation: past, present, and future. Geoderma 100(3-4), 269-301.

Hitziger, M., Ließ, M., 2014. Comparison of three supervised learning methods for digital soil mapping: Application to a complex terrain in the Ecuadorian Andes. Applied and Environmental Soil Science 2014.

Huang, C., Davis, L., Townshend, J., 2002. An assessment of support vector machines for land cover classification. International Journal of remote sensing 23(4), 725-749.

Jalabert, S.S.M., Martin, M.P., Renaud, J.-P., Boulonne, L., Jolivet, C., Montanarella, L., Arrouays, D., 2010. Estimating forest soil bulk density using boosted regression modelling. Soil Use and Management 26(4), 516-528.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning, 112. Springer.

Janzen, H., 2019. The Future of Humic Substances Research: Preface to a Debate. Journal of Environmental Quality 48(2), 205-206.

Jenny, H., 1941. Factors of soil formation: A system of quantitative pedology. Mineola, NY: Dover.

Jiang, Y., Nishimura, P., van den Heuvel, M.R., MacQuarrie, K.T.B., Crane, C.S., Xing, Z., Raymond, B.G., Thompson, B.L., 2015. Modeling land-based nitrogen loads from groundwater-dominated agricultural watersheds to estuaries to inform nutrient reduction planning. Journal of Hydrology 529, 213-230.

Johnston, A.M., Bruulsema, T.W., 2014. 4R Nutrient Stewardship for Improved Nutrient Use Efficiency. Procedia Engineering 83, 365-370.

Kabir, Z., 2005. Tillage or no-tillage: impact on mycorrhizae. Canadian Journal of Plant Science 85(1), 23-29.

Kasraei, B., Heung, B., Saurette, D.D., Schmidt, M.G., Bulmer, C.E., Bethel, W., 2021. Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. Environmental Modelling & Software 144, 105139.

Kettler, T.A., Doran, J.W., Gilbert, T.L., 2001. Simplified method for soil particle-size determination to accompany soil-quality analyses. Soil Sci. Soc. Am. J. 65(3), 849-852.

Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., Cornelis, W., 2016. Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. European Journal of Soil Science 67(3), 276-284.

Kim, D.-G., Hernandez-Ramirez, G., Giltrap, D., 2013. Linear and nonlinear dependency of direct nitrous oxide emissions on fertilizer nitrogen input: A meta-analysis. Agriculture, Ecosystems & Environment 168, 53-65.

Kleber, M., Lehmann, J., 2019. Humic Substances Extracted by Alkali Are Invalid Proxies for the Dynamics and Functions of Organic Matter in Terrestrial and Aquatic Ecosystems. Journal of Environmental Quality 48(2), 207-216.

Koenker, R., 2019. Quantreg: Quantile regression. http://CRAN. R-project. org/package= quantreg.

Koenker, R., Bassett, G., 1978. Regression Quantiles. Econometrica 46(1), 33-50.

Kovačević, M., Bajat, B., Gajić, B., 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma 154(3), 340-347.

Kuhn, M., 2020. Classification and Regression Training. R package version 6.0-86.

Kuhn, M., Johnson, K., 2013. Applied predictive modeling, 26. Springer.

Lamichhane, S., Kumar, L., Wilson, B., 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. Geoderma 352, 395-413.

Landré, A., Saby, N.P.A., Barthès, B.G., Ratié, C., Guerin, A., Etayo, A., Minasny, B., Bardy, M., Meunier, J.D., Cornu, S., 2018. Prediction of total silicon concentrations in French soils using pedotransfer functions from mid-infrared spectrum and pedological attributes. Geoderma 331, 70-80.

Laurence, L., Heung, B., Strom, H., Stiles, K., Burton, D., 2023. Towards a cost-effective framework for estimating soil nitrogen pools using pedotransfer functions and machine learning. Geoderma 440, 116692.

Lawrence, R., Bunn, A., Powell, S., Zambon, M., 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote sensing of environment 90(3), 331-336.

Lehmann, J., Kleber, M., 2015. The contentious nature of soil organic matter. Nature 528(7580), 60-68.

Leirós, M., Trasar-Cepeda, C., Seoane, S., Gil-Sotres, F., 1999. Dependence of mineralization of soil organic matter on temperature and moisture. Soil Biology and Biochemistry 31(3), 327-335.

Lin, H., Wheeler, D., Bell, J., Wilding, L., 2005. Assessment of soil spatial variability at multiple scales. Ecological Modelling 182(3-4), 271-290.

Lin, L.I., 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 45(1), 255-268.

Liu, J., Qiu, L., Chen, C., Wang, M., Wei, X., Kong, W., Cheng, J., 2023. Response of soil N dynamics to land-use change and topographic position in a semiarid hilly grassland. Land Degradation & Development 34(14), 4157-4167.

Lory, J., Scharf, P., 2003. Yield goal versus delta yield for predicting fertilizer nitrogen need in corn. Agronomy Journal 95(4), 994-999.

MacDonald, E., 2023. Thesis chapter: Satellite-Derived Bare Soil Mapping in Prince Edward Island – A Potential Tool for Site Specific Field Management.

MacDougall, J.I., Wilson, F., Veer, C., 1988. Soils of Prince Edward Island : Prince Edward Island soil survey. Land Resource Research Centre contribution ; no. 83-54. Research Branch, Agriculture Canada, Charlottetown, P.E.I.

Machet, J.M., Dubrulle, P., Damay, N., Duval, R., Julien, J.L., Recous, S., 2017. A Dynamic Decision-Making Tool for Calculating the Optimal Rates of N Application for 40 Annual Crops While Minimising the Residual Level of Mineral N at Harvest. Agronomy-Basel 7(4).

Mallory, E., Griffin, T., 2007. Impacts of soil amendment history on nitrogen availability from manure and fertilizer. Soil Sci. Soc. Am. J. 71(3), 964-973.

Malone, B.P., Styc, Q., Minasny, B., McBratney, A.B., 2017. Digital soil mapping of soil carbon at the farm scale: A spatial downscaling approach in consideration of measured and uncertain data. Geoderma 290, 91-99.

Manzoni, S., Porporato, A., 2009. Soil carbon and nitrogen mineralization: Theory and models across scales. Soil Biology & Biochemistry 41(7), 1355-1379.

Marquaridt, D.W., 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics 12(3), 591-612.

Marshall, C.B., Burton, D.L., Heung, B., Lynch, D.H., 2021. Influence of cropping system and soil type on soil health. Can. J. Soil Sci. 101(4), 626-640.

Matus, F.J., 2021. Fine silt and clay content is the main factor defining maximal C and N accumulations in soils: a meta-analysis. Scientific Reports 11(1), 6438.

McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109(1-2), 41-73.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97(3-4), 293-327.

McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117(1-2), 3-52.

Mello, F.A., Demattê, J.A., Rizzo, R., de Mello, D.C., Poppiel, R.R., Silvero, N.E., Safanelli, J.L., Bellinaso, H., Bonfatti, B.R., Gomez, A.M., 2022. Complex hydrological knowledge to support digital soil mapping. Geoderma 409, 115638.

Menard, S.W., 1995. Applied logistic regression analysis. Sage Publications, Thousand Oaks, Calif. :.

Mendonça-Santos, M.L., McBratney, A.B., Minasny, B., 2006. Chapter 21 Soil Prediction with Spatially Decomposed Environmental Factors. In: P. Lagacherie, A.B. McBratney, M. Voltz (Eds.), Developments in Soil Science. Elsevier, pp. 269-278.

Mesele, S., Ajiboye, G., 2020. Pedo-Transfer Functions for Predicting Total Soil Nitrogen in Different Land Use Types under Some Tropical Environments. Ghana Journal of Science 61(2), 45-56.

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301-311.

Minasny, B., McBratney, A.B., Bristow, K.L., 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma 93(3), 225-253.

Miranda, R., Nobrega, R., Silva, E., Silva, J., Araújo Filho, J., Moura, M., Barros, A., Souza, A., Verhoef, A., Yang, W., 2022. Hybrid machine learning for digital soil mapping across a longitudinal gradient of contrasting topography, climate and vegetation.

Moebius-Clune, B.N., D.J. Moebius-Clune, B.K. Gugino, O.J. Idowu, R.R. Schindelbeck, A.J. Ristow, H.M. van Es, J.E. Thies, H.A. Shayler, M.B. McBride, K.S.M Kurtz, D.W. Wolfe, and G.S. Abawi,, 2016. Comprehensive Assessment of Soil Health – The Cornell Framework, Edition 3.2. In: C. University (Ed.), Geneva, NY.

Molnar, C., Casalicchio, G., Bischl, B., 2018. iml: An R package for interpretable machine learning. Journal of Open Source Software 3(26), 786.

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., Mouazen, A.M., 2016. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosystems Engineering 152, 104-116.

Morvan, T., Beff, L., Lambert, Y., Mary, B., Germain, P., Louis, B., Beaudoin, N., 2022. An Original Experimental Design to Quantify and Model Net Mineralization of Organic Nitrogen in the Field. Nitrogen 3(2), 197-212.

Mponela, P., Snapp, S., Villamor, G.B., Tamene, L., Le, Q.B., Borgemeister, C., 2020. Digital soil mapping of nitrogen, phosphorus, potassium, organic carbon and their crop response thresholds in smallholder managed escarpments of Malawi. Applied Geography 124, 102299.

Nyiraneza, J., Chantigny, M., N'dayegamiye, A., Laverdière, M., 2010. Long-term manure application and forages reduce nitrogen fertilizer requirements of silage corn–cereal cropping systems. Agronomy journal 102(4), 1244-1251.

Nyiraneza, J., Murnaghan, D., Mills, A., Jiang, Y., 2022. Using a plant bioassay approach to estimate soil nitrogen contribution to potato crop, Agriculture and Agri-Food Canada, https://peipotatoagronomy.com/wp-content/uploads/2023/01/Soil_Derived_N_AAFC_Nyiraneza2023.pdf.

Nyiraneza, J., Thompson, B., Geng, X.Y., He, J.X., Jiang, Y.F., Fillmore, S., Stiles, K., 2017. Changes in soil organic matter over 18 yr in Prince Edward Island, Canada. Can. J. Soil Sci. 97(4), 745-756.

Nyiraneza, J., Ziadi, N., Zebarth, B.J., Sharifi, M., Burton, D.L., Drury, C.F., Bittman, S., Grant, C.A., 2012. Prediction of Soil Nitrogen Supply in Corn Production using Soil Chemical and Biological Indices. Soil Sci. Soc. Am. J. 76(3), 925-935.

O'brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. Quality & quantity 41, 673-690.

Odeh, I., McBratney, A., 2005. Pedometrics. Encyclopedia of Soils in the Environment. Elsevier.

Oglesby, C., Dhillon, J., Fox, A., Singh, G., Ferguson, C., Li, X., Kumar, R., Dew, J., Varco, J., 2023. Discrepancy between the crop yield goal rate and the optimum nitrogen rates for maize production in Mississippi. Agronomy Journal 115(1), 340-350.

Olk, D.C., Bloom, P.R., Perdue, E.M., McKnight, D.M., Chen, Y., Farenhorst, A., Senesi, N., Chin, Y.P., Schmitt-Kopplin, P., Hertkorn, N., Harir, M., 2019. Environmental and Agricultural Relevance of Humic Fractions Extracted by Alkali from Soils and Natural Waters. Journal of Environmental Quality 48(2), 217-232.

Pachepsky, Y., Rawls, W.J., 2004. Development of pedotransfer functions in soil hydrology, 30. Elsevier.

Padarian, J., Morris, J., Minasny, B., McBratney, A.B., 2018. Pedotransfer functions and soil inference systems. Pedometrics, 195-220.

Parsaie, F., Farrokhian Firouzi, A., Mousavi, S.R., Rahmani, A., Sedri, M.H., Homaee, M., 2021. Large-scale digital mapping of topsoil total nitrogen using machine learning models and associated uncertainty map. Environmental Monitoring and Assessment 193, 1-15.

Paul, S., Dowell, L., Coops, N., Johnson, M., Krzic, M., Geesing, D., Smukler, S., 2020. Tracking changes in soil organic carbon across the heterogeneous agricultural landscape of the Lower Fraser Valley of British Columbia. Science of The Total Environment 732, 138994.

Paul, S.S., Heung, B., Lynch, D.H., 2022. Modeling of total and active organic carbon dynamics in agricultural soil using digital soil mapping: a case study from Central Nova Scotia. Can. J. Soil Sci. (ja).

Pedlar, J.H., McKenney, D.W., Lawrence, K., Papadopol, P., Hutchinson, M.F., Price, D., 2015. A comparison of two approaches for generating spatial models of growing-season variables for Canada. Journal of Applied Meteorology and Climatology 54(2), 506-518.

PEI Analytical Laboratories, P., 1996. Water pH and SMP Buffer pH in Soil pH by pH Meter.

PEI Analytical Laboratories, P., 2019. Soil Health Test Sampling Instructions.

PEI Department of Agriculture and Land, P., 2020. Agriculture on PEI.

PEI Department of Agriculture and Land, P., 2023. Prince Edward Island Soil Quality Monitoring Project: Observed soil nutrient trends on PEI over 23 years (1998-2021). In: P.D.o.A.a. Land (Ed.).

Perreault, S., El Alem, A., Chokmani, K., Cambouris, A.N., 2022. Development of Pedotransfer Functions to Predict Soil Physical Properties in Southern Quebec (Canada). Agronomy 12(2), 526.

Poggio, L., De Sousa, L.M., Batjes, N.H., Heuvelink, G., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7(1), 217-240.

Pohjankukka, J., Pahikkala, T., Nevalainen, P., Heikkonen, J., 2017. Estimating the prediction performance of spatial models via spatial k-fold cross validation. International Journal of Geographical Information Science 31(10), 2001-2019.

Pribyl, D.W., 2010. A critical review of the conventional SOC to SOM conversion factor. Geoderma 156(3), 75-83.

Priesack, E., Gayler, S., Hartmann, H.P., 2006. The impact of crop growth sub-model choice on simulated water and nitrogen balances. Nutrient Cycling in Agroecosystems 75(1-3), 1-13.

Priori, S., Bianconi, N., Costantini, E.A.C., 2014. Can γ-radiometrics predict soil textural data and stoniness in different parent materials? A comparison of two machine-learning methods. Geoderma 226-227, 354-364.

Purushothaman, N.K., Reddy, N.N., Das, B.S., 2022. National-scale maps for soil aggregate size distribution parameters using pedotransfer functions and digital soil mapping data products. Geoderma 424, 116006.

Qin, W., Fan, G., Hongxing, L., 2022. Estimation and predicting of soil water characteristic curve using the support vector machine method. Earth Science Informatics.

Quinlan, J.R., 1992. Learning with continuous classes, 5th Australian joint conference on artificial intelligence. World Scientific, pp. 343-348.

R-CoreTeam, 2022. R: A language and Environment for Statistical Computing, Viennna, Austria.

Rashidi, M., Seilsepour, M., 2009. Total nitrogen pedotransfer function for calcareous soils of Varamin region. Int. J. Agric. Biol 11, 89-92.

Rasiah, V., 1995. COMPARISON OF PEDOTRANSFER FUNCTIONS TO PREDICT NITROGEN-MINERALIZATION PARAMETERS OF 1-POOL AND 2-POOL MODELS. Commun. Soil Sci. Plant Anal. 26(11-12), 1873-1884.

Reddy, N.N., Das, B.S., 2023. Digital soil mapping of key secondary soil properties using pedotransfer functions and Indian legacy soil data. Geoderma 429, 116265.

Rieke, E.L., Bagnall, D.K., Morgan, C.L.S., Flynn, K.D., Howe, J.A., Greub, K.L.H., Mac Bean, G., Cappellazzi, S.B., Cope, M., Liptzin, D., Norris, C.E., Tracy, P.W., Aberle, E., Ashworth, A., Bañuelos Tavarez, O., Bary, A.I., Baumhardt, R.L., Borbón Gracia, A., Brainard, D.C., Brennan, J.R., Briones Reyes, D., Bruhjell, D., Carlyle, C.N., Crawford, J.J.W., Creech, C.F., Culman, S.W., Deen, B., Dell, C.J., Derner, J.D., Ducey, T.F., Duiker, S.W., Dyck, M.F., Ellert, B.H., Entz, M.H., Espinosa Solorio, A., Fonte, S.J., Fonteyne, S., Fortuna, A.-M., Foster, J.L., Fultz, L.M., Gamble, A.V., Geddes, C.M., Griffin-LaHue, D., Grove, J.H., Hamilton, S.K., Hao, X., Hayden, Z.D., Honsdorf, N., Ippolito, J.A., Johnson, G.A., Kautz, M.A., Kitchen, N.R., Kumar, S., Kurtz, K.S.M., Larney, F.J., Lewis, K.L., Liebman, M., Lopez Ramirez, A., Machado, S., Maharjan, B., Martinez Gamiño, M.A., May, W.E., McClaran, M.P., McDaniel, M.D., Millar, N., Mitchell, J.P., Moore, A.D., Moore, P.A., Mora Gutiérrez, M., Nelson, K.A., Omondi, E.C., Osborne, S.L., Osorio Alcalá, L., Owens, P., Pena-Yewtukhiw, E.M., Poffenbarger, H.J., Ponce Lira, B., Reeve, J.R., Reinbott, T.M., Reiter, M.S., Ritchey, E.L., Roozeboom, K.L., Rui, Y., Sadeghpour, A., Sainju, U.M., Sanford, G.R., Schillinger, W.F., Schindelbeck, R.R., Schipanski, M.E., Schlegel, A.J., Scow, K.M., Sherrod, L.A., Shober, A.L., Sidhu, S.S., Solís Moya, E., St. Luce, M., Strock, J.S., Suyker, A.E., Sykes, V.R., Tao, H., Trujillo Campos, A., Van Eerd, L.L., van Es, H.M., Verhulst, N., Vyn, T.J., Wang, Y., Watts, D.B., Wright, D.L., Zhang, T., Honeycutt, C.W., 2022. Evaluation of aggregate stability methods for soil health. Geoderma 428, 116156.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8), 913-929.

Robertson, G.P., Groffman, P.M., 2015. Chapter 14 - Nitrogen Transformations. In: E.A. Paul (Ed.), Soil Microbiology, Ecology and Biochemistry (Fourth Edition). Academic Press, Boston, pp. 421-446.

Román Dobarco, Bourennane, H., Arrouays, D., Saby, N.P.A., Cousin, I., Martin, M.P., 2019a. Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. Geoderma 344, 14-30.

Román Dobarco, M., Cousin, I., Le Bas, C., Martin, M.P., 2019b. Pedotransfer functions for predicting available water capacity in French soils, their applicability domain and associated uncertainty. Geoderma 336, 81-95.

Ros, G.H., 2011. Predicting soil nitrogen supply: relevance of extractable soil organic matter fractions. Wageningen University and Research.

Ros, G.H., 2012. Predicting soil N mineralization using organic matter fractions and soil properties: A re-analysis of literature data. Soil Biology & Biochemistry 45, 132-135.

Ros, G.H., Hanegraaf, M.C., Hoffland, E., van Riemsdijk, W.H., 2011a. Predicting soil N mineralization: Relevance of organic matter fractions and soil properties. Soil Biology & Biochemistry 43(8), 1714-1722.

Ros, G.H., Hoffland, E., van Kessel, C., Temminghoff, E.J.M., 2009. Extractable and dissolved soil organic nitrogen - A quantitative assessment. Soil Biology & Biochemistry 41(6), 1029-1039.

Ros, G.H., Temminghoff, E.J.M., Hoffland, E., 2011b. Nitrogen mineralization: a review and meta-analysis of the predictive value of soil tests. European Journal of Soil Science 62(1), 162-173.

Roscoe, R., Buurman, P., 2003. Tillage effects on soil organic matter in density fractions of a Cerrado Oxisol. Soil & Tillage Research 70(2), 107-119.

Rossel, R.V., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158(1-2), 46-54.

Saurette, D., 2021. onsoilsurvey: making PDSM in Ontario better. R package version 0.0. 0.9000.

Saurette, D.D., Heck, R.J., Gillespie, A.W., Berg, A.A., Biswas, A., 2023. Divergence metrics for determining optimal training sample size in digital soil mapping. Geoderma 436, 116553.

Schillaci, C., Perego, A., Valkama, E., Märker, M., Saia, S., Veronesi, F., Lipani, A., Lombardo, L., Tadiello, T., Gamper, H.A., 2021. New pedotransfer approaches to predict soil bulk density using WoSIS soil data and environmental covariates in Mediterranean agro-ecosystems. Science of the total environment 780, 146609.

Schindelbeck, R., Moebius-Clune, B., Moebius-Clune, D., Kurtz, K., van Es, H., 2016. Cornell University comprehensive assessment of soil health laboratory standard operating procedures. Cornell Univ., Ithaca, NY.

Schomberg, H.H., Wietholter, S., Griffin, T.S., Reeves, D.W., Cabrera, M.L., Fisher, D.S., Endale, D.M., Novak, J.M., Balkcom, K.S., Raper, R.L., 2009. Assessing indices for predicting potential nitrogen mineralization in soils under different management systems. Soil Sci. Soc. Am. J. 73(5), 1575-1586.

Sedaghat, A., Shahrestani, M.S., Noroozi, A.A., Fallah Nosratabad, A., Bayat, H., 2022. Developing pedotransfer functions using Sentinel-2 satellite spectral indices and Machine learning for estimating the surface soil moisture. Journal of Hydrology 606, 127423.

Shahabi, A., Nabiollahi, K., Davari, M., Zeraatpisheh, M., Heung, B., Scholten, T., Taghizadeh-Mehrjardi, R., 2022. Spatial prediction of soil properties through hybridized random forest model and combination of reflectance spectroscopy and environmental covariates. Geocarto International, 1-24.

Shahbazi, F., Hughes, P., McBratney, A.B., Minasny, B., Malone, B.P., 2019. Evaluating the spatial and vertical distribution of agriculturally important nutrients—nitrogen, phosphorous and boron—in North West Iran. Catena 173, 71-82.

Sharifi, M., Zebarth, B.J., Burton, D.L., Grant, C.A., Bittman, S., Drury, C.F., McConkey, B.G., Ziadi, N., 2008. Response of potentially mineralizable soil nitrogen and indices of nitrogen availability to tillage system. Soil Sci. Soc. Am. J. 72(4), 1124-1131.

Sharifi, M., Zebarth, B.J., Burton, D.L., Grant, C.A., Cooper, J.M., 2007a. Evaluation of some indices of potentially mineralizable nitrogen in soil. Soil Sci. Soc. Am. J. 71(4), 1233-1239.

Sharifi, M., Zebarth, B.J., Burton, D.L., Grant, C.A., Porter, G.A., Cooper, J.M., Leclerc, Y., Moreau, G., Arsenault, W.J., 2007b. Evaluation of laboratory-based measures of soil mineral nitrogen and potentially mineralizable nitrogen as predictors of field-based indices of soil nitrogen supply in potato production. Plant Soil 301(1-2), 203-214.

Sharifi, M., Zebarth, B.J., Burton, D.L., Grant, C.A., Porter, G.A., Cooper, J.M., Leclerc, Y., Moreau, G., Arsenault, W.J., 2007c. Evaluation of laboratory-based measures of soil mineral nitrogen and potentially mineralizable nitrogen as predictors of field-based indices of soil nitrogen supply in potato production. Plant Soil 301(1), 203-214.

Sieber, P., Ericsson, N., Hammar, T., Hansson, P.-A., 2022. Albedo impacts of current agricultural land use: Crop-specific albedo from MODIS data and inclusion in LCA of crop production. Science of the Total Environment 835, 155455.

Simard, R., Ziadi, N., Nolin, M., Cambouris, A., 2001. Prediction of nitrogen responses of corn by soil nitrogen mineralization indicators. TheScientificWorldJOURNAL 1, 135-141.

Six, J., Conant, R.T., Paul, E.A., Paustian, K., 2002. Stabilization mechanisms of soil organic matter: implications for C-saturation of soils. Plant Soil 241, 155-176.

Six, J., Elliott, E., Paustian, K., 1999. Aggregate and soil organic matter dynamics under conventional and no-tillage systems. Soil Sci. Soc. Am. J. 63(5), 1350-1358.

Six, J., Paustian, K., Elliott, E.T., Combrink, C., 2000. Soil structure and organic matter: I. Distribution of aggregate-size classes and aggregate-associated carbon. Soil Sci. Soc. Am. J. 64(2), 681-689.

Soil Classification Working Group, A.a.A.-F.C.A., 1998. The Canadian system of soil classification, 3rd ed. NRC Research Press.

Solomatine, D.P., Dulal, K.N., 2003. Model trees as an alternative to neural networks in rainfall—runoff modelling. Hydrological Sciences Journal 48(3), 399-411.

St Luce, M., Whalen, J.K., Ziadi, N., Zebarth, B.J., 2011. Nitrogen Dynamics and Indices to Predict Soil Nitrogen Supply in Humid Temperate Soils. In: D.L. Sparks (Ed.), Advances in Agronomy, Vol 112. Advances in Agronomy. Elsevier Academic Press Inc, San Diego, pp. 55-102.

Stanford, G., 1973. Rationale for optimum nitrogen fertilization in corn production. Journal of Environmental Quality 2(2), 159-166.

Stanford, G., Smith, S.J., 1972. Nitrogen mineralization potential of soils. Soil Sci. Soc. of Am. Proceedings 36(3), 465-&.

Stark, J., Porter, G., 2005. Potato nutrient management in sustainable cropping systems. American Journal of Potato Research 82, 329-338.

Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K., Pásztor, L., 2019. Mapping soil hydraulic properties using random-forest-based pedotransfer functions and geostatistics. Hydrol. Earth Syst. Sci. 23(6), 2615-2635.

Tabachnick, B.G., Fidell, L.S., Ullman, J.B., 2013. Using multivariate statistics, 6. pearson Boston, MA.

Tamagno, S., Eagle, A.J., McLellan, E.L., van Kessel, C., Linquist, B.A., Ladha, J.K., Pittelkow, C.M., 2022. Quantifying N leaching losses as a function of N balance: A path to sustainable food supply chains. Agriculture, Ecosystems & Environment 324, 107714.

Tivet, F., de Moraes Sa, J.C., Lal, R., Briedis, C., Borszowskei, P.R., dos Santos, J.B., Farias, A., Eurich, G., Hartman, D.d.C., Nadolny Junior, M., Bouzinac, S., Seguy, L., 2013. Aggregate C depletion by plowing and its restoration by diverse biomass-C inputs under no-till in sub-tropical and tropical regions of Brazil. Soil & Tillage Research 126, 203-218.

Tyralis, H., Papacharalampous, G., Langousis, A., 2019. A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. Water 11(5), 910.

Uygur, V., Irvem, A., Karanlik, S., Akis, R., 2010. Mapping of total nitrogen, available phosphorous and potassium in Amik Plain, Turkey. Environmental Earth Sciences 59(5), 1129-1138.

Valenzuela, H., 2023. Ecological Management of the Nitrogen Cycle in Organic Farms. Nitrogen 4(1), 58-84.

van der Westhuizen, S., Heuvelink, G.B.M., Hofmeyr, D.P., 2023. Multivariate random forest for digital soil mapping. Geoderma 431, 116365.

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y.A., Padarian, J., Schaap, M.G., Toth, B., Verhoef, A., Vanderborght, J., van der Ploeg, M.J., Weihermuller, L., Zacharias, S., Zhang, Y.G., Vereecken, H., 2017. Pedotransfer Functions in Earth System Science: Challenges and Perspectives. Reviews of Geophysics 55(4), 1199-1256.

Vapnik, V., Chapelle, O., 2000. Bounds on error expectation for support vector machines. Neural computation 12(9), 2013-2036.

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE transactions on neural networks 10(5), 988-999.

von Lützow, M., Kögel-Knabner, I., Ekschmitt, K., Flessa, H., Guggenberger, G., Matzner, E., Marschner, B., 2007. SOM fractionation methods: relevance to functional pools and to stabilization mechanisms. Soil Biology and Biochemistry 39(9), 2183-2207.

Wang, H., Wellmann, F., Zhang, T.Q., Schaaf, A., Kanig, R.M., Verweij, E., von Hebel, C., van der Kruk, J., 2019. Pattern Extraction of Topsoil and Subsoil Heterogeneity and Soil-Crop Interaction Using Unsupervised Bayesian Machine Learning: An Application to Satellite-Derived NDVI Time Series and Electromagnetic Induction Measurements. Journal of Geophysical Research-Biogeosciences 124(6), 1524-1544.

Wang, K., Zhang, C., Li, W., 2013. Predictive mapping of soil total nitrogen at a regional scale: A comparison between geographically weighted regression and cokriging. Applied Geography 42, 73-85.

Wang, S., Jin, X., Adhikari, K., Li, W., Yu, M., Bian, Z., Wang, Q., 2018. Mapping total soil nitrogen from a site in northeastern China. Catena 166, 134-146.

Wang, S., Zhuang, Q., Wang, Q., Jin, X., Han, C., 2017. Mapping stocks of soil organic carbon and soil total nitrogen in Liaoning Province of China. Geoderma 305, 250-263.

Wang, W., Smith, C.J., Chalk, P.M., Chen, D., 2001. Evaluating chemical and physical indices of nitrogen mineralization capacity with an unequivocal reference. Soil Sci. Soc. Am. J. 65(2), 368-376.

Wang, Y., Zhao, T., 2017. Statistical interpretation of soil property profiles from sparse data using Bayesian compressive sampling. Geotechnique 67(6), 523-536.

Waterer, D., 2002. Impact of high soil pH on potato yields and grade losses to common scab. Canadian journal of plant science 82(3), 583-586.

Weil, R.R., Islam, K.R., Stine, M.A., Gruver, J.B., Samson-Liebig, S.E., 2003. Estimating active carbon for soil quality assessment: A simplified method for laboratory and field use. American Journal of Alternative Agriculture 18(1), 3-17.

Whittaker, J., Nyiraneza, J., Zebarth, B.J., Jiang, Y., Burton, D.L., 2023. The effects of forage grasses and legumes on subsequent potato yield, nitrogen cycling, and soil properties. Field Crops Research 290, 108747.

Wösten, J., Lilly, A., Nemes, A., Le Bas, C., 1999. Development and use of a database of hydraulic properties of European soils. Geoderma 90(3-4), 169-185.

Wu, Q., 2021. whitebox:"WhiteboxTools" R Frontend. R package version 1.4. 0.

Xiao, Y., Xue, J., Zhang, X., Wang, N., Hong, Y., Jiang, Y., Zhou, Y., Teng, H., Hu, B., Lugato, E., Richer-de-Forges, A.C., Arrouays, D., Shi, Z., Chen, S., 2022. Improving pedotransfer functions for predicting soil mineral associated organic carbon by ensemble machine learning. Geoderma, 116208.

Yang, T., Siddique, K.H., Liu, K., 2020. Cropping systems in agriculture and their impact on soil health-A review. Global Ecology and Conservation 23, e01118.

Zebarth, B., Karemangingo, C., Savoie, D., Scott, P., Moreau, G., 2008. Nitrogen management for potatoes: General fertilizer recommendations. GHG Taking Charge Team Factsheet.

Zebarth, B., Leclerc, Y., Moreau, G., Sanderson, J., Arsenault, W., Botha, E., Wang-Pruski, G., 2005. Estimation of soil nitrogen supply in potato fields using a plant bioassay approach. Can. J. Soil Sci. 85(3), 377-386.

Zebarth, B.J., Danielescu, S., Nyiraneza, J., Ryan, M.C., Jiang, Y., Grimmett, M., Burton, D.L., 2015. Controls on nitrate loading and implications for BMPs under intensive potato production systems in Prince Edward Island, Canada. Groundwater Monitoring & Remediation 35(1), 30-42.

Zebarth, B.J., Drury, C.F., Tremblay, N., Cambouris, A.N., 2009. Opportunities for improved fertilizer nitrogen management in production of arable crops in eastern Canada: A review. Can. J. Soil Sci. 89(2), 113-132.

Zhang, Y., Sui, B., Shen, H., Ouyang, L., 2019. Mapping stocks of soil total nitrogen using remote sensing data: A comparison of random forest models with different predictors. Computers and Electronics in Agriculture 160, 23-30.

Zhou, T., Geng, Y.J., Chen, J., Sun, C.L., Haase, D., Lausch, A., 2019. Mapping of Soil Total Nitrogen Content in the Middle Reaches of the Heihe River Basin in China Using Multi-Source Remote Sensing-Derived Variables. Remote Sensing 11(24).

Zhou, Y., Xue, J., Chen, S.C., Zhou, Y., Liang, Z.Z., Wang, N., Shi, Z., 2020. Fine-Resolution Mapping of Soil Total Nitrogen across China Based on Weighted Model Averaging. Remote Sensing 12(1).

Zibilske, L.M., 1994. Carbon mineralization. Methods of Soil Analysis: Part 2 Microbiological and Biochemical Properties 5, 835-863.

APPENDIX

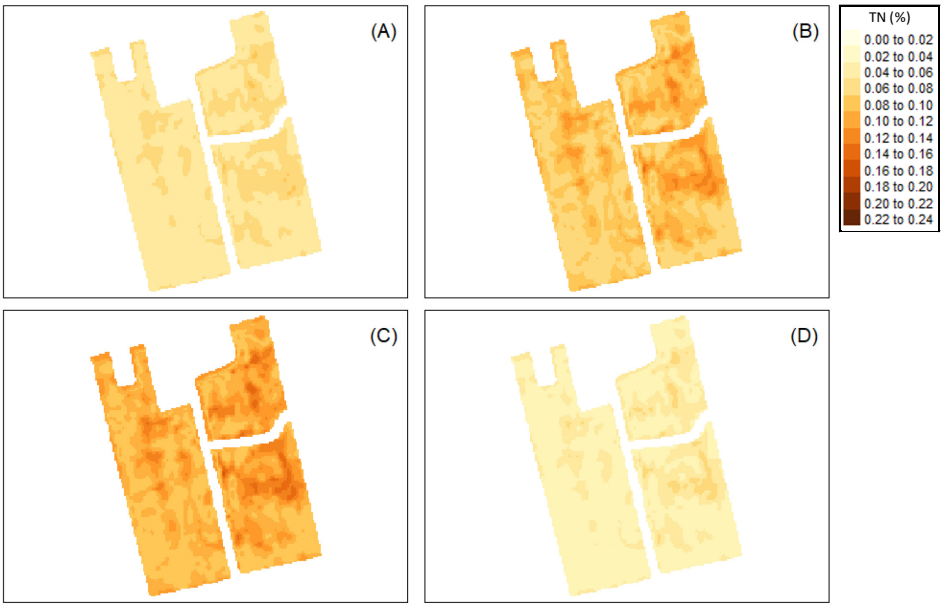


Figure A.1 Soil total nitrogen (TN) maps (%) of Field 1 (F1) in the study area modeled with field-specific (FS) data using the support vector machine learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

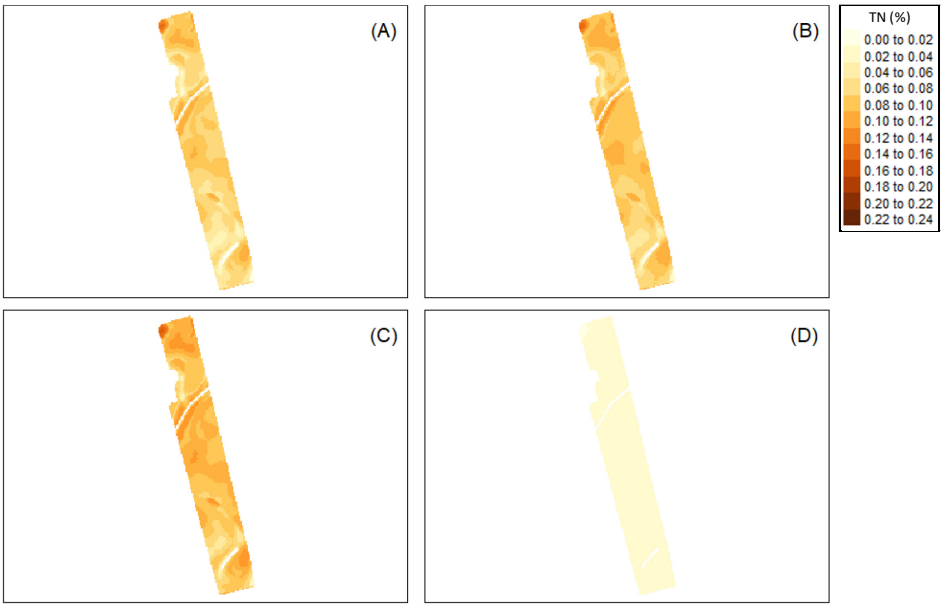


Figure A.2 Soil total nitrogen (TN) maps (%) of Field 3 (F3) in the study area modeled with field-specific (FS) data using the cubist learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

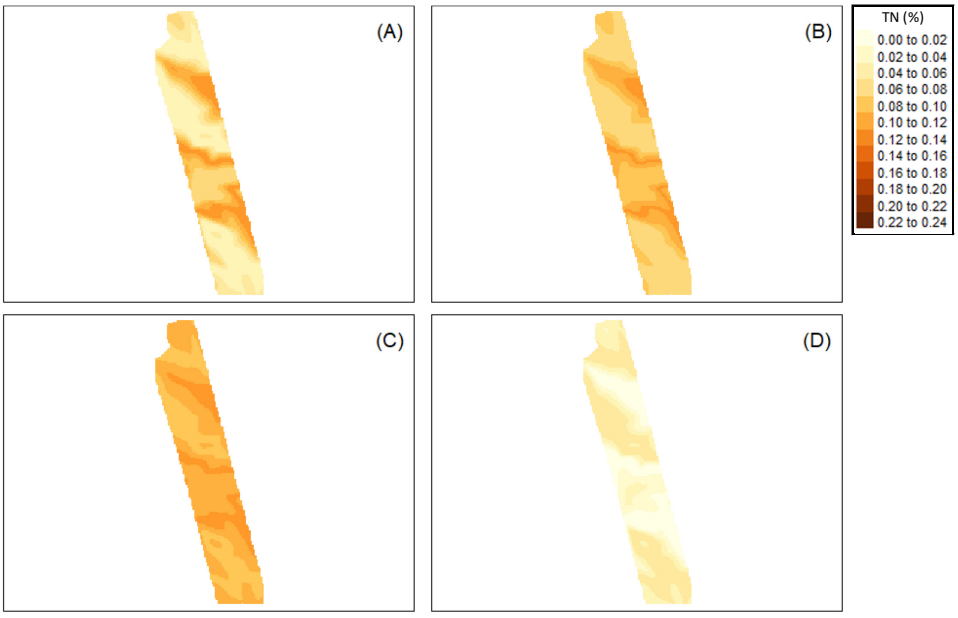


Figure A.3 Soil total nitrogen (TN) maps (%) of Field 4 (F4) in the study area modeled with field-specific (FS) data using the support vector machine learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limitpper prediction limit (95th percentile), and (D) 90% prediction interval map.

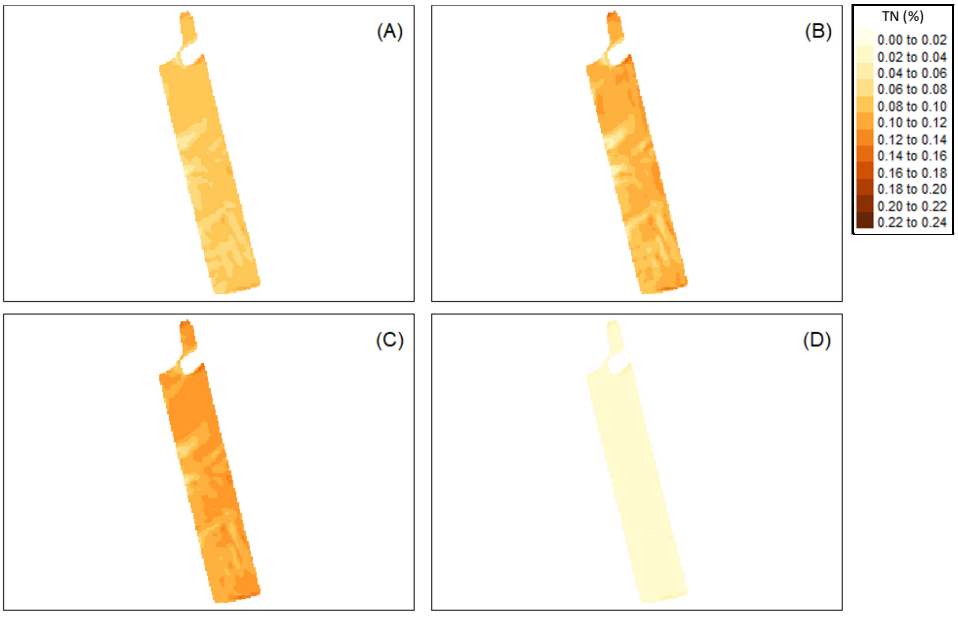


Figure A.4 Soil total nitrogen (TN) maps (%) of Field 5 (F5) in the study area modeled with field-specific (FS) data using the support vector machine learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

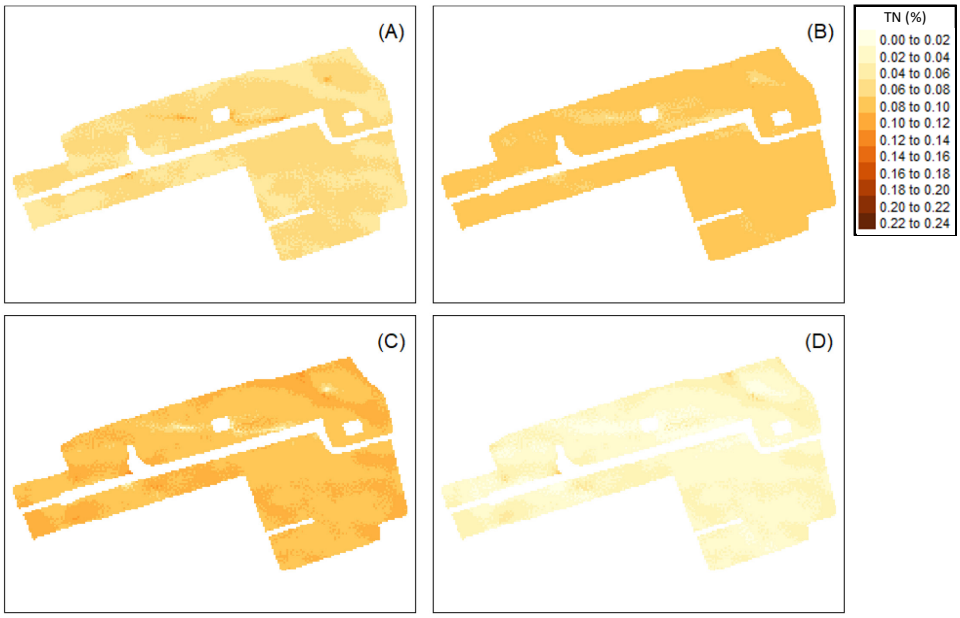


Figure A.5 Soil total nitrogen (TN) maps (%) of Field 6 (F6) in the study area modeled with field-specific (FS) data using the support vector machine learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

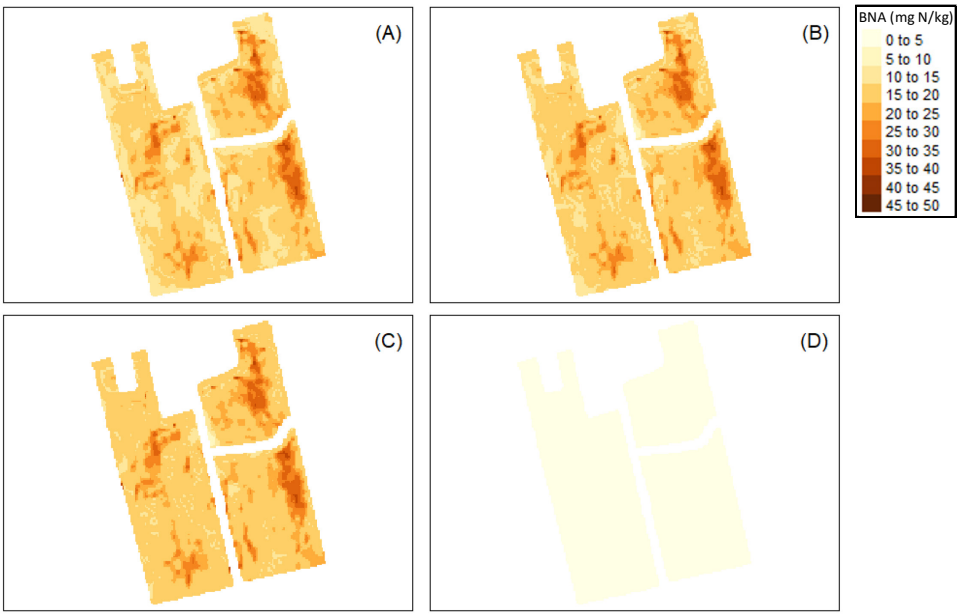


Figure A.6 Biological nitrogen availability (BNA) maps (mg N/kg) of Field 1 (F1) in the study area modeled with field-specific (FS) data using the cubist learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

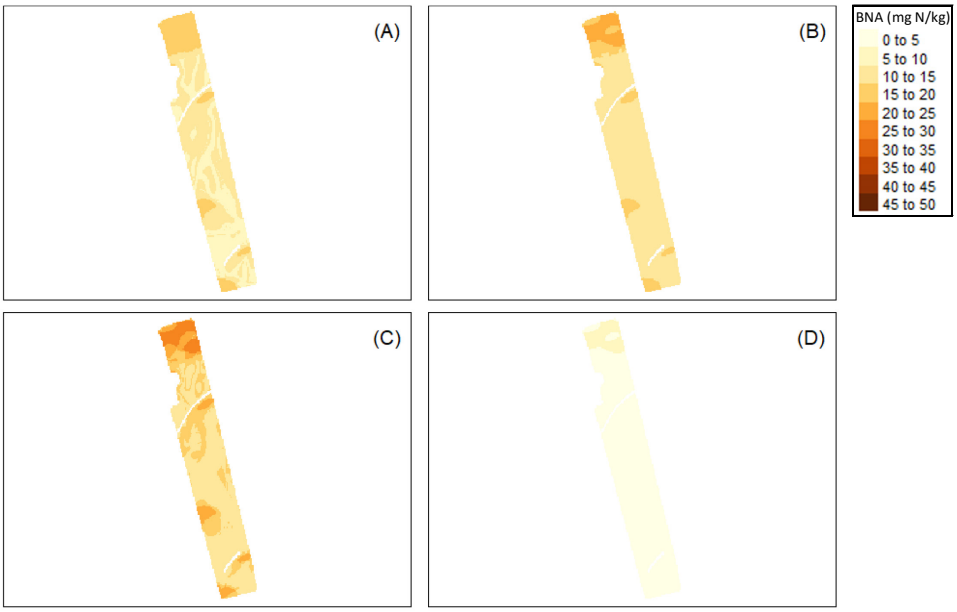


Figure A.7 Biological nitrogen availability (BNA) maps (mg N/kg) of Field 3 (F3) in the study area modeled with field-specific (FS) data using the random forest learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

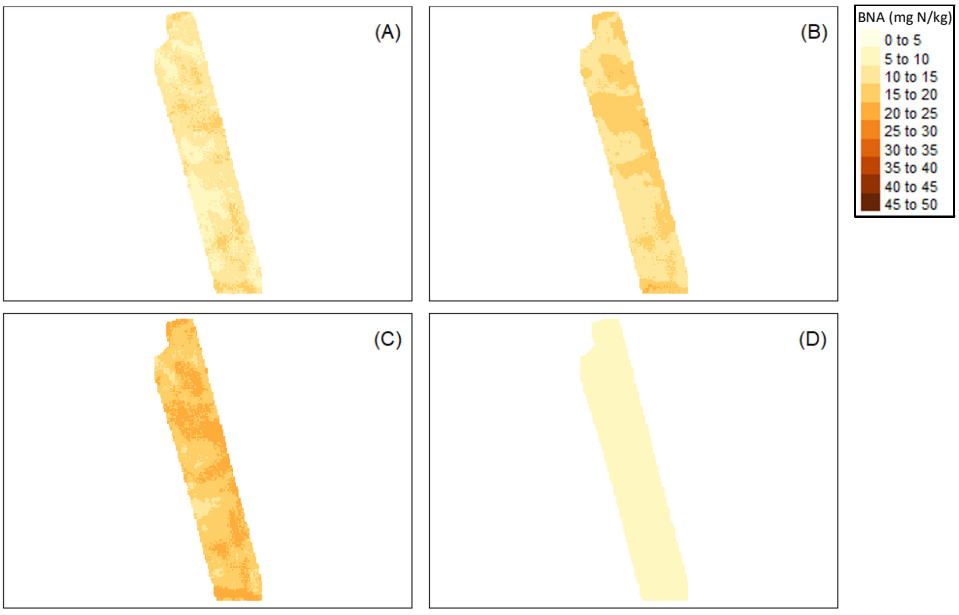


Figure A.8 Biological nitrogen availability (BNA) maps (mg N/kg) of Field 4 (F4) in the study area modeled with field-specific (FS) data using the random forest learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

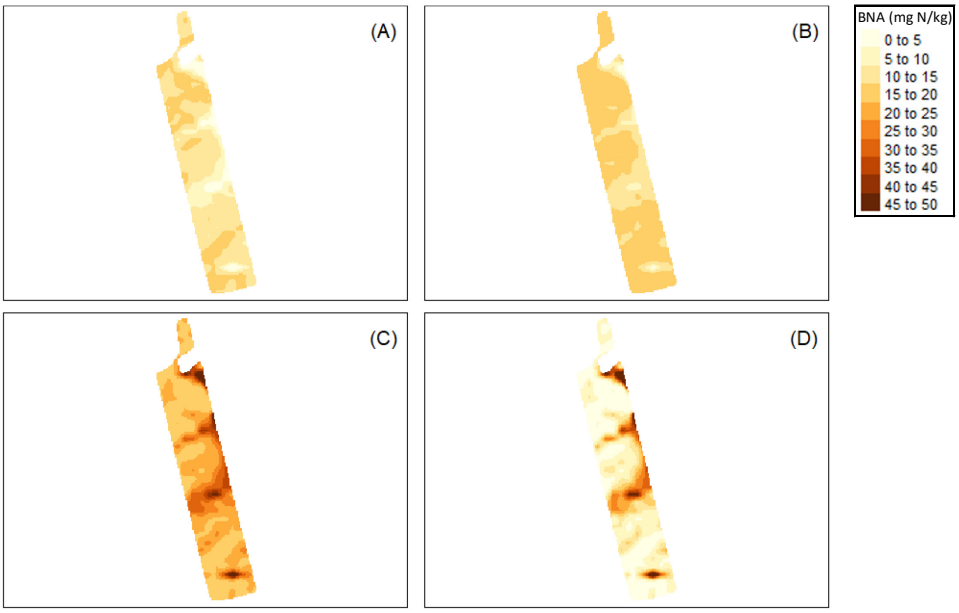


Figure A.9 Biological nitrogen availability (BNA) maps (mg N/kg) of Field 5 (F5) in the study area modeled with field-specific (FS) data using the support vector machine learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

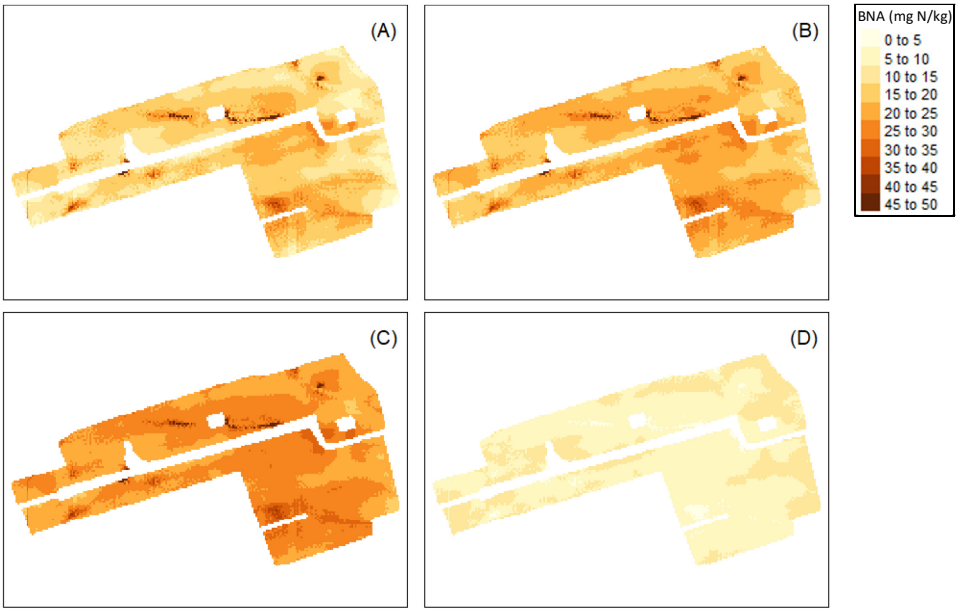


Figure A.10 Biological nitrogen availability (BNA) maps (mg N/kg) of Field 6 (F6) in the study area modeled with field-specific (FS) data using the support vector machine learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

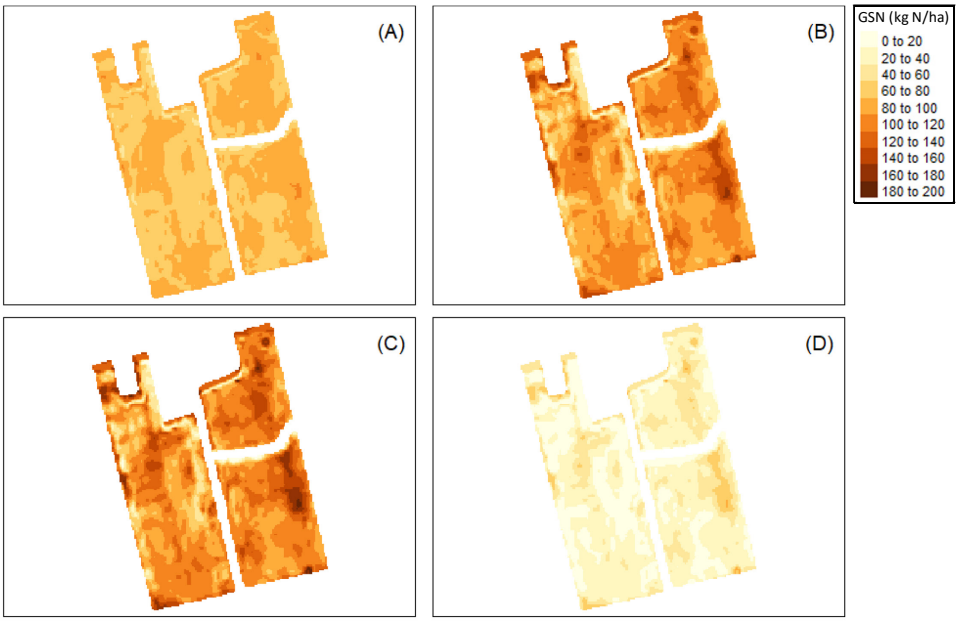


Figure A.11 Growing season nitrogen (GSN) maps (kg N/ha) of Field 1 (F1) in the study area modeled with field-specific (FS) data using the support vector machine learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

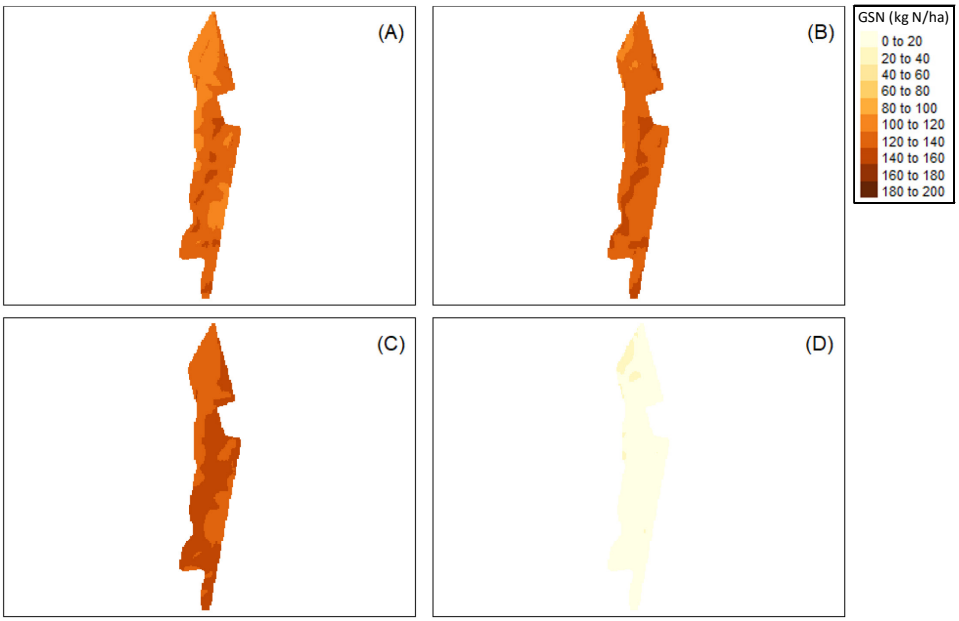


Figure A.12 Growing season nitrogen (GSN) maps (kg N/ha) of Field 2 (F2) in the study area modeled with field-specific (FS) data using the cubist learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

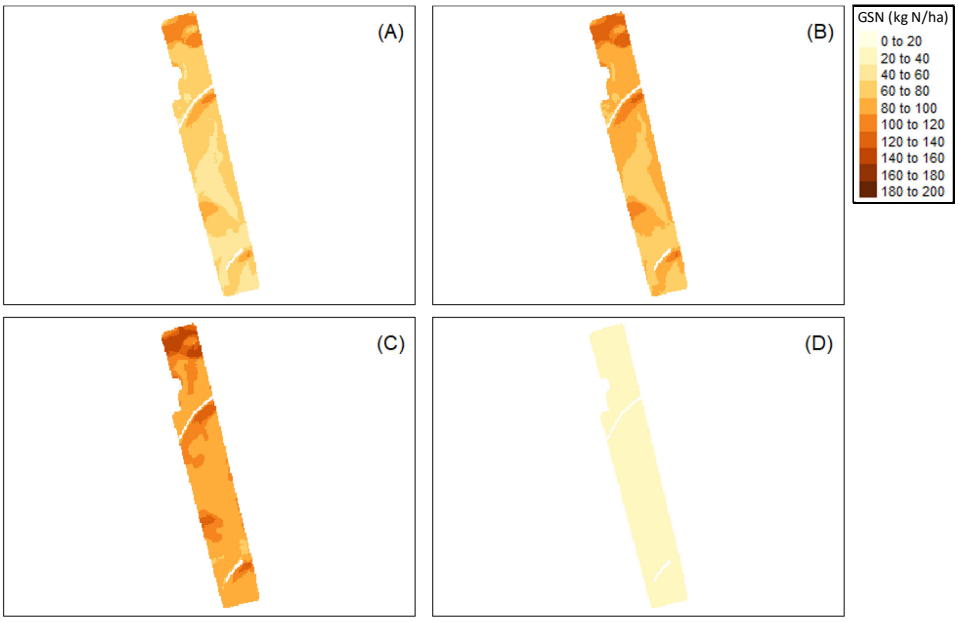


Figure A.13 Growing season nitrogen (GSN) maps (kg N/ha) of Field 3 (F3) in the study area modeled with field-specific (FS) data using the cubist learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

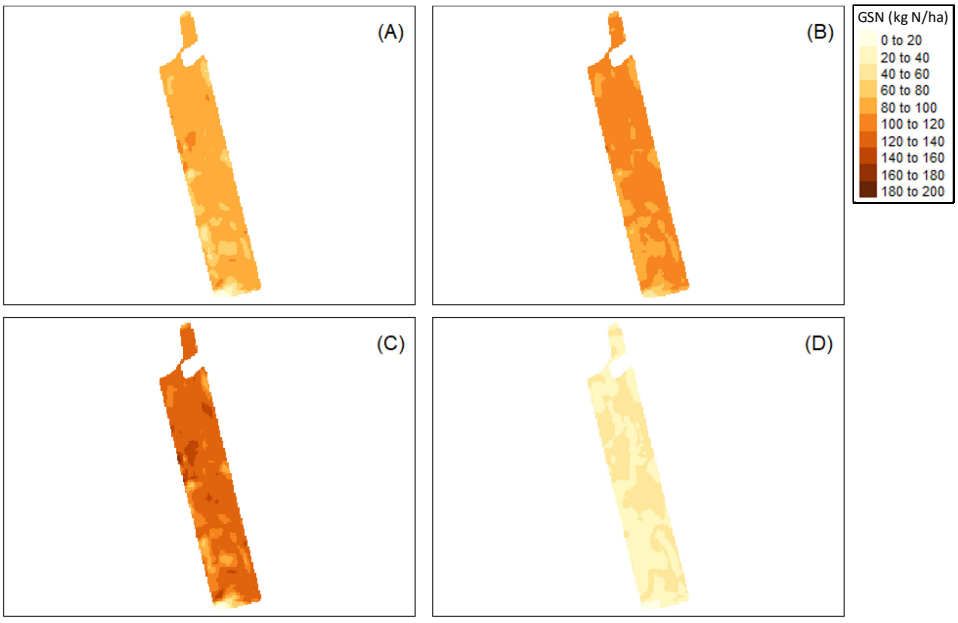


Figure A.14 Growing season nitrogen (GSN) maps (kg N/ha) of Field 5 (F5) in the study area modeled with field-specific (FS) data using the support vector machine learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5th percentile), (B) prediction map, (C) upper prediction limit (95th percentile), and (D) 90% prediction interval map.

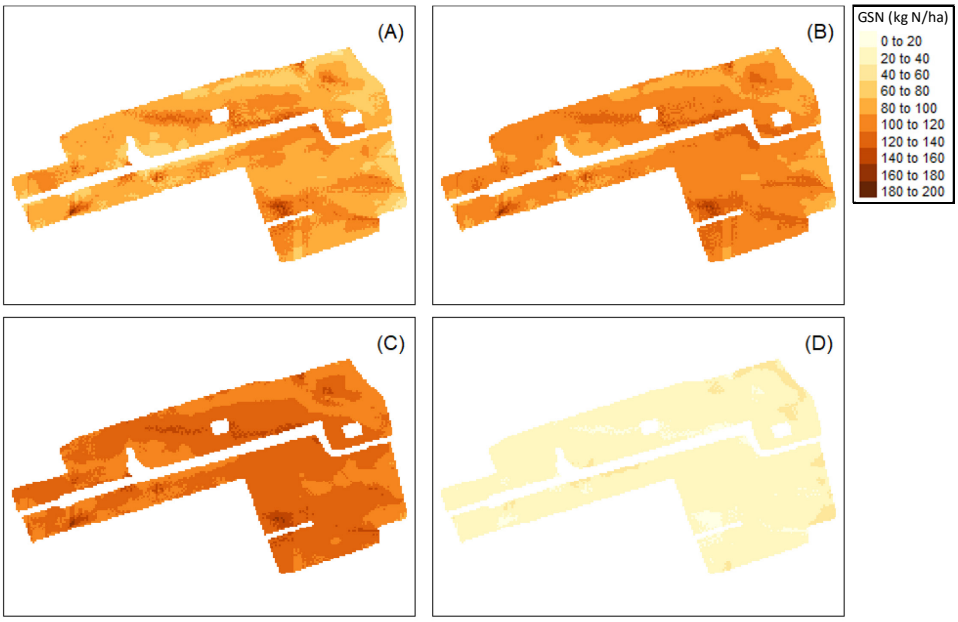


Figure A.15 Growing season nitrogen (GSN) maps (kg N/ha) of Field 6 (F6) in the study area modeled with field-specific (FS) data using the support vector machine learner, and uncertainty maps using quantile regression. (A) Lower prediction limit map (5^{th} percentile), (B) prediction map, (C) upper prediction limit (95^{th} percentile), and (D) 90% prediction interval map.