Mappatura del carbonio organico nel suolo in Toscana attraverso l'integrazione di osservazioni a terra, dati ausiliari e di telerilevamento

Montaghi Alessandro [1], Chiesi Marta [1], Maselli Fabio.[1] Gardin Lorenzo[1]
[1] CNR-IRET, alessandro.montaghi@cnr.it
[2] CNR IBE, Nome.Cognome@cnr.it

Abstract. Il carbonio organico immagazzinato nello strato superficiale del suolo è una componente essenziale del ciclo globale del carbonio che dovrebbe essere quantificato per diversi scopi. Il presente articolo propone un metodo per mappare la quantità di carbonio organico immagazzinato nei primi 0.3 m di suolo (SOC), basato sull'integrazione di numerose osservazioni con dati ausiliari e di telerilevamento. Il metodo di integrazione consiste nell'applicazione di un sistema scalabile *end-to-end* di *tree boosting* chiamato XGBoost. Grazie alle sue caratteristiche, XGBoost è considerato lo stato dell'arte in molte applicazioni di classificazione e regressione. Il framework di ottimizzazione di nuova generazione Optuna è stato utilizzato per identificare gli iper parametri. Questo approccio è applicato e sperimentato in Toscana, una regione dell'Italia centrale con caratteristiche ambientali estremamente diversificate ed eterogenee. Più di 4000 campioni di suolo sono stati raccolti e resi disponibili allo scopo, insieme ad un set di dati ausiliari e telerilevati. Queste informazioni sono state elaborate per produrre una mappa finale che descrive la distribuzione spaziale del SOC nella regione alla risoluzione spaziale di 100 m. La mappa riproduce bene la variabilità del SOC nella regione, mostrando valori più elevati per le foreste rispetto alle praterie e ai terreni coltivati, e picchi di SOC in torbiere e in suoli su substrato vulcanico.

Parole chiave: SOC; Apprendimento automatico, XGBoost; ottimizzazione.

1 Introduzione

La sostanza organica del suolo (SOM) è la porzione di suolo composta da tessuti vegetali o animali a vari stadi di decomposizione. Include diversi componenti come residui vegetali, biomassa microbica vivente, materia organica attiva (spesso chiamata detriti) e materia organica stabile, ovvero humus. L'elemento chimico primario del SOM è il carbonio, presente in vari composti organici e tipicamente indicato come carbonio organico del suolo (SOC), d'importanza rilevante nell'ambito del programma di rendicontazione dei gas serra dell'IPCC già da metà anni '90. Il SOC, infatti, è efficace nelle strategie di mitigazione del riscaldamento globale dovuto ai cambiamenti climatici e svolge un ruolo fondamentale nell'influenzare la disponibilità di nutrienti

delle piante e nel regolare la ritenzione idrica, essendo in grado di migliorare la struttura del suolo e limitare l'erosione (Oldfield et al., 2019).
Per tutti questi motivi, il monitoraggio e la mappatura del SOC a diverse scale spaziali stanno diventando sempre più rilevanti per supportare il monitoraggio globale del carbonio nel suolo e le operazioni di agricoltura di precisione a livello locale (Post et al., 2001; Wetterlind et al., 2008; Lamichhane et al., 2019).
Il SOC è determinato da molti fattori ambientali, sia intrinseci alle caratteristiche del suolo e del terreno (es. topografia, litologia, tessitura, ecc.), sia correlati alle caratteristiche dell'ecosistema (es. copertura e produttività della vegetazione, pratiche di gestione del territorio, ecc.) (Schillaci et al., 2017; Calvo de Anta et al., 2020). Questi fattori hanno alta variabilità spaziale, che aumenta la discontinuità del SOC e la conseguente difficoltà nel mapparlo, soprattutto in aree frammentate come quelle del bacino del Mediterraneo (Adhikari et al., 2020).
Diversi approcci vengono utilizzati per la mappatura del SOC, a seconda della scala di indagine e della disponibilità di dati. Il metodo più semplice è ovviamente fornito dalle misurazioni a terra basato su campioni raccolti in campo ed analizzati in laboratorio con procedure standard. Questo approccio è molto accurato, ma anche costoso e dispendioso in termini di tempo; inoltre produce osservazioni puntuali che potrebbero non essere rappresentative di grandi aree (Chatterjee et al., 2009). Le osservazioni a terra sono quindi solitamente estese su vaste zone mediante tecniche di interpolazione e/o estrapolazione che possono sfruttare le informazioni di livelli di dati ausiliari (Lamichhane et al., 2019).
I dati telerilevati rappresentano una forma particolare di questi livelli e recentemente sono stati presi in considerazione per la mappatura del SOC. Questi dati, infatti, possono fornire informazioni sincrone e ripetitive su vaste aree geografiche, a volte inaccessibili per il rilievo in campo (Castaldi et al., 2016). Le immagini ottiche sono spesso utilizzate in questo contesto, grazie al loro valore informativo sulle caratteristiche dell'ecosistema e del suolo legate al SOC (Angelopoulou et al., 2019). Tali operazioni, tuttavia, sono limitate da diversi problemi. In primo luogo, le informazioni relative al SOC ritratte dall'immagine sono limitate ai terreni nudi, poiché la copertura vegetale ne ostacola la caratterizzazione. In secondo luogo, il segnale spettrale dei terreni nudi correlato all'OC è solitamente alterato da diversi fattori di disturbo, come la rugosità superficiale, l'umidità e il contenuto di minerali, ecc. (Castaldi et al., 2016). Ciò aumenta la complessità delle relazioni tra SOC ed informazioni spettrali, che possono essere teoricamente gestite da metodi avanzati di elaborazione dei dati (Lamichhane et al., 2019).
Il *gradient boosting* è una potente tecnica di apprendimento automatico introdotta da Friedman (2001). L'introduzione di questa tecnica è stata motivata dall'implementazione del metodo della discesa del gradiente nello spazio delle funzioni ammissibili del modello, in grado di adattarsi a modelli predittivi generici non parametrici. I modelli ad albero sono particolarmente adatti al gradient boosting quando applicati a modelli ad albero additivi (Hastie et al., 2009). Più recentemente, un nuovo metodo di potenziamento degli alberi è entrato in scena e ha rapidamente guadagnato popolarità. XGBoost, che è il nome abbreviato di eXtream Gradient Boosting, è un sistema di apprendimento automatico basato su un algoritmo di potenziamento del

gradiente proposto da Chen e Guestrin (2016). Questo approccio di apprendimento supervisionato applica un approccio di divisione binaria ricorsiva per scegliere la migliore divisione in ogni fase per ottenere il modello migliore (Chen e Guestrin, 2016). In modo simile a Random Forest, XGBoost non è sensibile ai valori anomali. Inoltre, XGBoost, come molti metodi di boosting, è robusto all'overfitting, utilizzando una formalizzazione del modello più regolarizzata, che facilita notevolmente la selezione del modello (Zhang et al., 2020). In particolare, XGBoost può eseguire le tre principali tecniche di potenziamento del gradiente, ovvero Gradient Boosting, Regularized Boosting e Stochastic Boosting. Infine, questo algoritmo è in grado di gestire tutti i tipi di distribuzione dei dati (Kankanamge et al., 2019). Grazie alla loro superiorità in termini di prestazioni e di tempo e memoria accessibili, i modelli XGBoost e le sue varianti sono ben noti per risolvere sia problemi di classificazione complessi (Liew et al., 2021, Zhang e Zhan, 2017), che problemi predittivi di regressione (Dong et al., 2020; Rusdah e Murfi, 2020; Ramraj et al., 2016).
Il presente articolo illustra i risultati preliminari di una metodologia che adotta queste tecniche per la mappatura del SOC a risoluzione spaziale di 100 m in Toscana (Italia centrale). Questo obiettivo è ottenuto integrando misure di SOC a 0.3 m di profondità del suolo ed altre informazioni ausiliarie (compresi i dati spettrali) utilizzando diversi algoritmi di apprendimento automatico parametrizzati mediante un framework avanzato di ottimizzazione, Optuna.

2 Area di studio

La Toscana si estende su una superficie di circa 23.000 km² (9°-12° di longitudine Est, 42°-44° di latitudine Nord) ed è caratterizzata da una ampia varietà di paesaggio, clima e copertura del suolo. La topografia della regione spazia dalle pianure costiere e dalle grandi valli fluviali alle aree collinari e montuose che si avvicinano alla catena appenninica. Il clima varia da mediterraneo vicino alla costa a temperato, caldo o freddo, a seconda dell'altitudine, della latitudine e della distanza dal mare.

In termini di uso del suolo, le pianure sono principalmente occupate da aree urbane ed attività agricole, mentre le colline e le montagne sono un mosaico di terreni agricoli e foreste. Le colture annuali (principalmente grano, mais ed ortaggi) coprono circa il 25% della regione, mentre le colture arboree come vigneti e uliveti rappresentano circa il 5%. I boschi coprono quasi la metà della superficie della Toscana.

3 Dati di studio

3.1 Misure a terra

Le osservazioni di SOC sono state derivate da analisi dettagliate di profili del suolo raccolti in Toscana dal 1990 ad oggi. Per l'elaborazione dei dati, sono stati considerati gli orizzonti minerali situati entro 0.3 metri dalla superficie per i suoli minerali e gli orizzonti organici per i suoli organici, in conformità con il Soil Survey Manual (Soil Survey Staff, 1993). I dati unitari per la sezione di suolo di 0-0.3 m sono stati calcolati

ponderando il contenuto di carbonio organico di ciascun orizzonte per il rispettivo spessore all'interno di quella profondità.

La determinazione del carbonio organico per tutti i campioni è stata effettuata in laboratorio seguendo gli standard ufficiali di analisi del suolo (MiPAAF, 1999).

3.2 Set di dati ausiliari e di telerilevamento

Alcuni strati informativi come la mappa di copertura del suolo, la litologia dei substrati e la mappa del suolo sono derivati dal data center della Regione Toscana (https://www.regione.toscana.it/-/geoscopio); ulteriori covariate sono state ottenute da parametri del suolo precedentemente spazializzati, come la granulometria del suolo ed il pH. Un modello digitale del terreno a 100 m di risoluzione spaziale viene impiegato per derivare indici morfometrici come pendenza, esposizione, potenziale di radiazione solare, ecc. I dati meteorologici, con una risoluzione spaziale di 250 metri, sono ricavati dal database regionale del Consorzio LaMMA ed elaborati in vari indici come la temperatura media annua, la precipitazione media annua ed indici di aridità (Indice di Aridità, indice di Fournier, Bagnouls e Gaussen).

Infine, dai dati di telerilevamento derivano le seguenti covariate: 1) valori medi di NDVI tratti dalle immagini MSI di Sentinel-2 di luglio 2022, ovvero il picco della stagione di crescita; ii) infrarosso medio a onde corte (SWIR) (i.e., banda 12 del dataset MSI di Sentinel-2) dei periodi estivi e primaverili 2022; iii) immagine media della produzione primaria lorda nel periodo 2000-2019 di tutti gli ecosistemi terrestri in Toscana ottenuta utilizzando un approccio Monteith (Maselli et al., 2009). Tutti i dati sono stati riportati in formato raster su un sistema di riferimento comune ed una risoluzione spaziale di 100 m.

4 Elaborazione dati

Le impostazioni predefinite degli iperparametri non possono garantire prestazioni ottimali delle tecniche di apprendimento automatico, ed è pertanto necessario porre attenzione all'impostazione ottimale degli iperparametri per ciascun modello e set di dati (Schratz et al., 2019).

Per quanto riguarda XGBoost, i principali parametri da regolare durante l'ottimizzazione degli iperparametri sono i seguenti. Il primo parametro è il booster che seleziona il tipo di modello da eseguire ad ogni iterazione in modo da massimizzarne le prestazioni. XGBoost implementa tre opzioni di booster: gbtree (XGBoost Tree Booster), gblinear e dart (Dropouts meet Multiple Additive Regression Trees). gbtree utilizza modelli basati su alberi per ogni iterazione di boosting e funziona bene su un'ampia gamma di set di dati. L'impostazione gblinear impiega modelli lineari, quindi è preferibile per set di dati in cui le relazioni tra le variabili sono ben approssimate linearmente. Infine, il sistema dart impedisce l'overfitting dovuto ad un approccio dropout durante l'allenamento.

Per le ultime due opzioni (es. booster gbtree o dart), i parametri comunemente ottimizzati includono la velocità di apprendimento (velocità con cui l'algoritmo di boosting apprende da ogni iterazione, ed è l'iperparametro di potenziamento del

gradiente più importante), gamma (controlla la quantità minima di riduzione della perdita richiesta per effettuare un'ulteriore divisione su un nodo foglia dell'albero), grow_policy (controlla il modo in cui vengono aggiunti nuovi nodi all'albero), max_depth (determina la profondità alla quale ogni albero nel processo di boosting può crescere durante l'addestramento), n_estimators (specifica il numero di alberi, stimatori, da costruire nel modello), min_child_weight (determina la somma minima della matrice Hessiana delle derivate secondo, necessaria in un nodo per effettuare la sua divisone), tree_method (specifica l'algoritmo utilizzato per costruire gli alberi, ad esempio auto, esatto, approssimativo o storico), max_bin (determina il numero massimo di contenitori utilizzati per il raggruppamento di funzionalità continue durante il processo di costruzione dell'albero), sottocampione (la frazione di campioni utilizzati in ogni iterazione di boosting) e colsample_bytree (la frazione di elementi utilizzati in ogni albero). Oltre a questi, altri parametri comunemente ottimizzati includono sample_type (controlla il modo in cui vengono selezionati gli alberi eliminati durante il processo di addestramento del modello), normalize_type (determina il tipo di algoritmo di normalizzazione utilizzato per i pesi degli alberi eliminati e degli alberi appena aggiunti durante il processo di boosting), rate_drop (controlla il tasso di abbandono, che è la frazione di alberi che vengono eliminati casualmente a ogni iterazione di boosting), skip_drop (controlla la probabilità di saltare la procedura di esclusione durante un'iterazione di boosting). Questi parametri aiutano a controllare il compromesso bias-varianza e la capacità di generalizzazione del modello (Yang et al., 2020).

Infine, per gblinear booster, gli unici parametri ottimizzati sono updater (specifica l'algoritmo del modello lineare e le opzioni includono shotgun e coordinate descent) e feature_selector (determina l'algoritmo utilizzato per la selezione delle caratteristiche quando si adatta un modello lineare, come ciclico, shuffle, casuale, avido o parsimonioso).

Attualmente, gli iperparametri XGBoost sono stati ottimizzati utilizzando la strategia del framework Optuna (ver 4.0.0) all'interno di una procedura di cross validation nidificata k-folder (con k=5). Nel ciclo di cross validation esterno, abbiamo suddiviso in modo casuale i set di dati di 4863 punti in training (90%) e set di test (10%). Successivamente, abbiamo confrontato la capacità predittiva per ciascuna delle tre impostazioni di booster implementate in XGBoost. Per ogni booster di parametri, abbiamo eseguito 50000 prove, o 10.000 prove per ciascuna delle cinque strategie di ottimizzazione impiegate, ovvero lo stimatore TPE (Tree-structured Parzen Estimator) (Bergstra et al., 2011), lo stimatore NSGA-III (Non-dominated Sorting Genetic Algorithm III) (Deb and Jain, 2013), lo stimatore GP (Gaussian process-based Bayesian optimization) (Wilson et al., 2020), il campionatore multi-obiettivo utilizzando l'algoritmo MOTPE (Chiandussi et al., 2012) e CmaES (Covariance matrix adaptation evolution strategy) stimatore (Loshchilov e Hutter, 2016). Nel ciclo di cross validation interno di ogni percorso, abbiamo suddiviso in modo casuale ogni set di allenamento in 5 set di allenamento e 1 set di convalida ed Optuna ha selezionato un sottoinsieme di iperparametri mediante una strategia di ottimizzazione. In ogni iterazione, le prestazioni medie del set di convalida sono determinate per valutare le prestazioni di un'impostazione di iperparametro in base al minimo root mean square error (RMSE).

Dopo aver selezionato le migliori prestazioni iperparametriche dei 3 (gbtree, gblinear e dart) x 5 (TPE, NSGA-III, GP, MOTPE e CmaES) modelli, in base al RMSE, abbiamo addestrato ogni modello migliore con il set di addestramento e valutato le prestazioni utilizzando le seguenti metriche: errore assoluto medio (MAE), errore quadratico medio (MSE), RMSE, errore massimo e coefficiente di efficienza del modello di Nash-Sutcliffe. Infine, è stata ottenuta una mappa SOC con una risoluzione spaziale di 100 m utilizzando la migliore strategia di modellazione applicata all'intero set di dati.

5 Risultati e Discussione

La Tabella 1 mostra tutte le statistiche di accuratezza ottenute applicando le diverse strategie di ottimizzazione. Le migliori prestazioni si ottengono utilizzando gli algoritmi gbtree e dart con la strategia di ottimizzazione CmaES; in entrambi i casi, tuttavia, le precisioni raggiunte sono relativamente scarse, come testimoniato dall'efficienza moderata del coefficiente di Nash-Sutcliffe. Una mappa di esempio ottenuta con il metodo dart CmaES è mostrata in Figura 1.

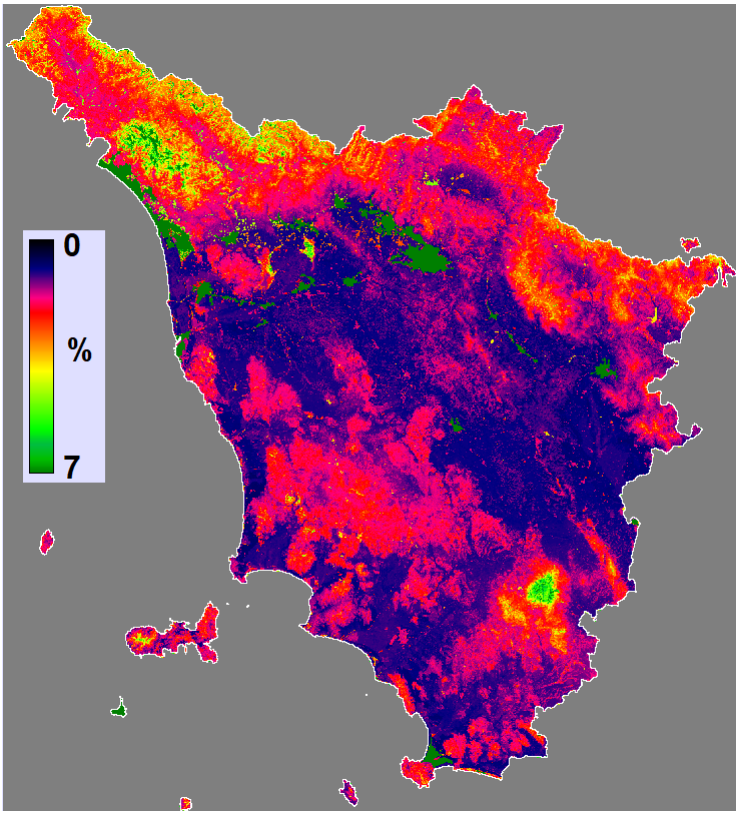

Figura 1 – Mappa di %SOC della Toscana ottenuta con il metodo dart CmaES.

Tabella 1: statistiche di accuratezza ottenute applicando le 3x5 opzioni di iperparametrizzazione al dataset di test.

Booster	Strategia di ottimizzazione	MAE	MSE	Errore massimo	RMSE	Nash-Sutcliffe
gbtree	TPE	0.705	1.589	10.157	1.261	-0.106
	NSGA-III	0.597	1.099	10.150	1.048	0.235
	GP	0.636	1.227	10.152	1.108	0.146
	MOTPE	0.745	1.598	10.204	1.264	-0.112
	CmaES	0.521	0.928	9.716	0.963	0.354
GBLinear	TPE	0.605	1.045	10.530	1.022	0.273
	NSGA-III	0.604	1.044	10.523	1.022	0.274
	GP	0.603	1.044	10.511	1.022	0.274
	MOTPE	0.605	1.044	10.528	1.022	0.273
	CmaES	0.605	1.045	10.529	1.022	0.273
dart	TPE	0.533	0.935	9.825	0.967	0.349
	NSGA-III	0.521	0.932	10.097	0.965	0.351
	GP	0.521	0.976	9.815	0.988	0.321
	MOTPE	0.645	1.366	10.152	1.169	0.050
	CmaES	0.521	0.928	9.716	0.963	0.354

Come precedentemente accennato, il presente studio è rivolto ad un primo sviluppo di un approccio metodologico per la mappatura del SOC in regioni mediterranee complesse dal punto di vista ambientale. La metodologia descritta, infatti, mostra come ottimizzare i parametri iniziali dei modelli statistici comunemente utilizzati per interpolare dataset spazialmente distribuiti. I primi risultati raggiunti evidenziano l'importanza di inizializzare correttamente gli algoritmi applicati, che possono essere eseguiti da Optuna, un framework di ottimizzazione degli iperparametri di nuova generazione. Questi risultati non possono ovviamente essere considerati conclusivi, ma possono servire come base per l'approfondimento degli aspetti metodologici più rilevanti sia dal punto di vista teorico che pratico.

Riferimenti bibliografici

1. Adhikari K., Mishra U., Owens P.R., Libohova Z., Wills S.A., Riley W.J., Hoffman F.M., Smith D.R. Importance and strength of environmental controllers of soil organic carbon changes with scale. Geoderma, 375, 114472 (2020).
2. Angelopoulou T., Tziolas N., Balafoutis A., Zalidis G., Bochtis D. Remote sensing techniques for soil organic carbon estimation: a review. Remote Sensing, 11, 676 (2019).
3. Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., & Cox, D. D. Hyperopt: a python library for model selection and hyperparameter optimization. Computational Science & Discovery, 8 (1), 014008 (2005).

4. Calvo de Anta R., Luisa E., Febrero-Bande M., Galinanes J., Macias F., Ortiz R., Casas F. Soil organic carbon in peninsular Spain: Influence of environmental factors and spatial distribution. Geoderma, 370, 114365 (2020).
5. Castaldi F., Palombo A., Santini F., Pascucci S., Pignatti S., Casa R. Evaluation of the potential of the current and forthcoming multispectral and hyperspectral images to estimate soil texture and organic carbon. Remote Sensing of Environment, 179: 54-65 (2016).
6. Chatterjee A., Lal R., Wielopolski L., Martin M. Z., Ebinger M. H. Evaluation of different soil carbon determination methods. Critical Reviews in Plant Science, 28(3), 164-178 (2009).
7. Chiandussi, G., Codegone, M., Ferrero, S., Varesio, F. E. Comparison of multi-objective optimization methodologies for engineering applications. *Computers & Mathematics with Applications*, *63*(5), 912-942 (2012).
8. 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.
9. Loshchilov, I., Hutter, F., 2016. CMA-ES for hyperparameter optimization of deep neural networks. *arXiv preprint arXiv:1604.07269*.
10. Maselli, F., Papale, D., Puletti, N., Chirici, G., Corona, P., 2009. Combining remote sensing and ancillary data to monitor the gross productivity of water-limited forest ecosystems. *Remote Sens. Environ*. 113 (3), 657–667.
11. MiPAAF (1999). Official methods of soil chemical analysis. Gazzetta Ufficiale Supplemento Ordinario 248, Istituto Poligrafico e Zecca dello Stato, Rome, Italy.
12. Oldfield E.E., Bradford M.A., Wood S.A. (2019). Global meta-analysis of the relationship between soil organic matter and crop yields. Soil, 5: 15-32.
13. Post W.M., Izaurralde R.C., Mann L.K., Bliss N. (2001). Monitoring and verifying changes of organic carbon in soil. Climate Change, 51, 73-99.
14. Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., Brenning, A., 2019. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109-120.
15. Schillaci C., Acutis M., Lombardo L., Lipani A., Fantappie M., Marker M., Saia S. (2017). Spatio-temporal topsoil organic carbon mapping of a semi-arid Mediterranean region: the role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling. Science of the Total Environment, 601–602, 821–832.
16. Soil Survey Staff (1993). Soil Survey Manual, USDA-SCS, Handbook, vol. 18. U.S. Govt. Print. Off., Washington, DC
17. Yang, Z., Yu, Y., You, C., Steinhardt, J., Ma, Y., 2020. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, 10767-10777.
18. Wilson, J., Borovitskiy, V., Terenin, A., Mostowsky, P., Deisenroth, M., 2020. Efficiently sampling functions from Gaussian process posteriors. In *International Conference on Machine Learning*, pp. 10292-10302. PMLR.
19. Wetterlind, J.; Stenberg, B. & Soderstrom, M. (2008). The use of near infrared (NIR) spectroscopy to improve soil mapping at the farm scale. Precision Agriculture, 9, 57-69.