

Quantile-Based Soil Nutrient Assessment from Multispectral Data for Scalable Agronomic Decision Support

Abstract

Accurate characterization of soil nutrient status is foundational to effective agronomic decision-making, yet conventional soil testing remains constrained by high costs, sampling frequency, low resolution and methodological variability across laboratories. These limitations restrict the scalability and comparability of nutrient assessments at field and regional scales. This whitepaper presents a proprietary machine learning framework for measuring soil nutrient levels using multispectral satellite imaging. Rather than directly estimating absolute nutrient concentrations, the proposed approach predicts the relative position of a given field within empirically derived nutrient distributions, expressed as quantile-based classes. This formulation enables consistent interpretation across laboratory methods, geographic regions, and sampling protocols, while preserving practical agronomic relevance. Model performance was evaluated using independent laboratory-analyzed soil samples spanning diverse soil types and management conditions. Results demonstrate strong agreement between predicted and observed nutrient quantiles, with robust performance across multiple nutrients. These findings suggest that relative, distribution-based nutrient prediction offers a scalable and interpretable alternative to absolute concentration estimates, supporting precision agriculture applications scalability, cross-regional compatibility and timely decision support are critical.

1. Introduction

Effective soil nutrient management underpins crop productivity, environmental stewardship, and long-term soil health. Measurements of key nutrients such as nitrogen, phosphorus, and potassium inform fertilization strategies, yield optimization, and regulatory compliance. However, traditional soil testing approaches rely on physical sampling and laboratory analysis, processes that are often labor-intensive, costly, and spatially sparse. Moreover, absolute nutrient measurements can vary substantially across laboratories due to differences in methodology, calibration standards, and reporting conventions, complicating direct comparison even between samples of the same field analyzed by different laboratories.

Recent advances in data availability and predictive modeling have enabled alternative approaches for characterizing soil properties at scale. Despite their promise, many existing models focus on predicting absolute nutrient concentrations, which can limit interpretability

and transferability when applied across heterogeneous landscapes or laboratory methodologies.

In this work, we present a novel framework that reframes soil nutrient estimation as a quantile prediction problem. By predicting the quantile-based standing of a field within a broader nutrient distribution, the model produces outputs that are inherently comparable across measurement scales and laboratory techniques. This relative formulation aligns more closely with agronomic decision-making, which often depends on whether nutrient levels are low, adequate, or excessive relative to regional norms rather than on precise concentration values. The objective of this whitepaper is to describe the conceptual foundation, validation approach, and empirical performance of this modeling framework, demonstrating its applicability for scalable and interpretable soil nutrient assessments.

2. Data Sources and Preprocessing

2.1 Data Overview

Ground truth data were obtained from the ESDAC LUCAS soil sampling campaigns conducted across Europe in 2009, 2015, and 2018 (**Orgiazzi et al., 2018**). We also use soil samples from our partners across Europe to validate the model. Multispectral data was collected for every point in our ground truth dataset from the Sentinel-2 and Landsat-8 missions (for locations for which Sentinel-2 is unavailable) (**Claverie et al., 2018**). We've collected images from the 8 months prior to the date of sampling by the ESDAC lab for each point which gives us enough bare soil images as well as enough data on the lifecycle of the plant on top. We've then masked the cloudiest images. The multispectral data was harmonized, atmospherically, radiometrically, BRDF corrected and resampled to 10m resolution. In addition we also made use of a variety of auxiliary data including climate data collected from ERA5 global averages, topographic data derived from the Copernicus DEM, parent material from USGS rasters and variables derived from geological rasters.

2.2 Sampling Strategy

To minimize the risks of overfitting or the model learning to recognize the location based on the multispectral or other data we employ several data splitting techniques. For our train-val split we use a stratified spatial splitting. We divide the data into 250km blocks and leave a random subset of these blocks for validation. This evaluation design intentionally penalizes region-specific shortcuts and provides a conservative estimate of real-world performance (**Roberts et al., 2017**). The 250km size was chosen to prevent the model from learning to group nearby points into the same nutrient categories using regional similarities. This design encourages the model to learn features that generalize across regions and climatic conditions, rather than relying on localized spatial correlations. Samples of all locations within the validation blocks across different years were also excluded from the training set. This forces the model to learn patterns that are temporally and spatially invariant and minimises information leakage.

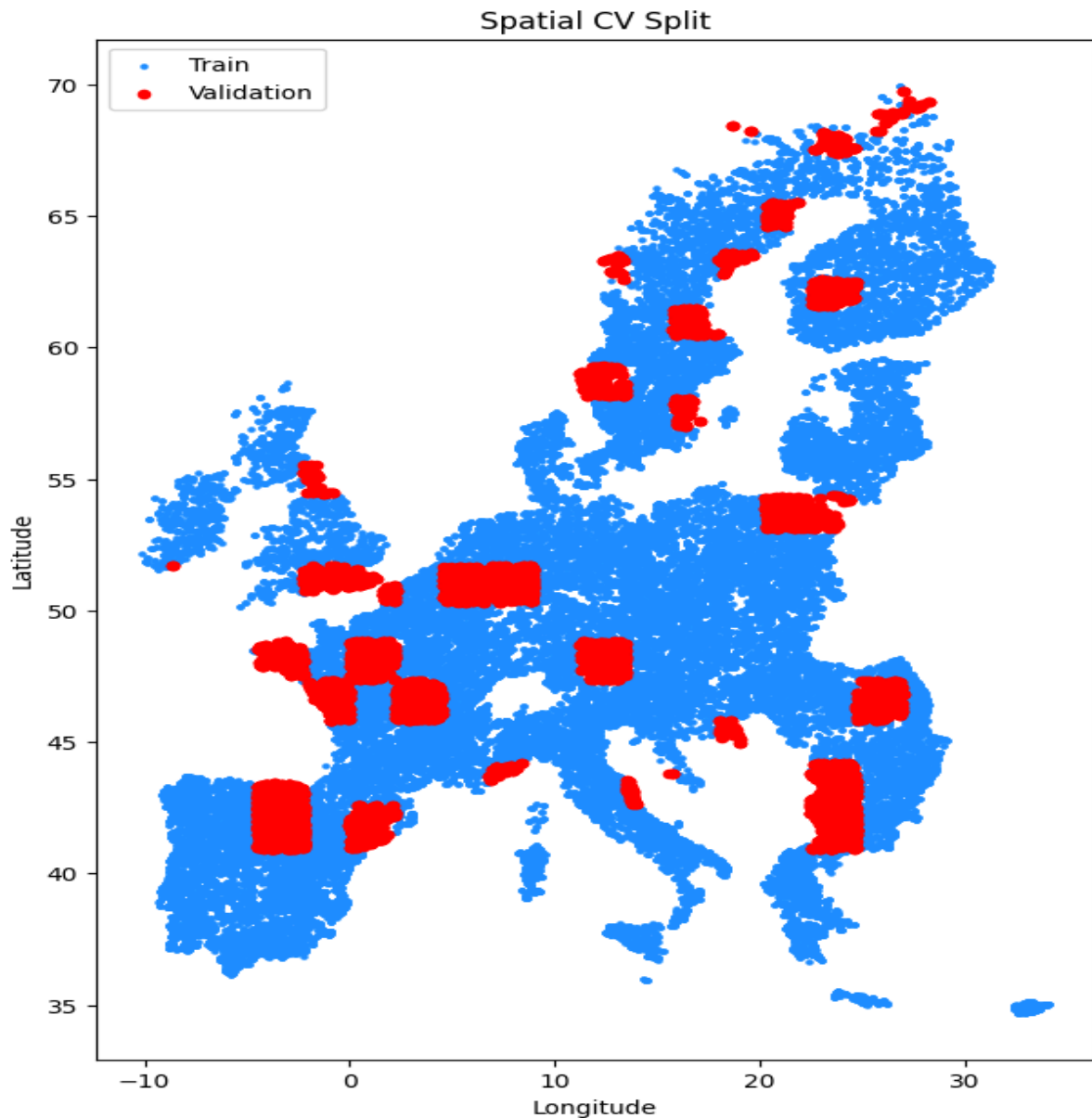


Fig 1. Spatial data split.

For our test set we use the LOCO (Leave One Country Out) Meyer & Pebesma (2021) approach where we purposefully exclude data from Switzerland from the training set and use it only for testing **Meyer & Pebesma (2021)**. This is done to measure the model's ability to detect nutrients in unseen regions. We also have a 2nd test set consisting of soil samples collected by our partners across Europe for the years 2022, 2023 and 2024. These samples were analyzed by different regional laboratories to test the model's ability to work with data points in the future thus testing its temporal stability.

3. Methodology

The primary objective of our model is to provide consistent, reliable nutrient assessments across diverse regions, soil types on a European scale and to be easily interpretable and

comparable with a wide variety of laboratory methods and scales within that region. We faced two major challenges:

1. Modeling - mapping a sparse, noisy low resolution spectral signature to a given nutrient level
2. Interpretability - delivering results that farmers and agronomists across different countries within Europe can easily compare and interpret.

3.1 Modeling

The main challenges here are related to the nature of multispectral sensing itself namely: low resolution (1 pixel per 10-30 m), noise (from atmospheric, radiometric, topographic sources), moisture effects drastically affect soil reflectance, limited number of bands and low spectral resolution, lack of strong detectable features in the VIS-NIR range for P, label noise from differences in laboratory analysis methods and local calibration.

To address these challenges, we apply a combination of advanced image preprocessing methods and a model architecture incorporating chemistry-informed loss functions and custom spectral indices designed to enhance signal-to-noise ratios for weakly detectable nutrients such as phosphorus..

3.1.1 P detection

Soil P as already mentioned lacks strong detectable features in VIS-NIR which makes it very difficult to accurately detect using multispectral imaging on bare soil alone. To tackle this challenge we make use of several techniques:

1. Use of auxiliary data - we make use of climate, topographic, texture and geological data to determine the soil's ability to hold on to P. These parameters have been shown to be strong predictors of soil P (**Hengl, T., et al. (2017)**).
2. Detection through plant deficiency cues - rather than directly measuring P in soil we use cues from the crop that was planted on the patch in previous seasons. We generate a plant signature in which our model looks for disturbance in photosynthesis, purpling of leaves and delayed growth (**Osborne et al., 2002**). This approach improves sensitivity to low phosphorus availability by incorporating crop stress signals that are not directly observable from soil reflectance alone.
3. Chemistry informed loss (**Raissi et al. (2019)**) - we use a unique loss which aims to teach the model the relationship between P availability in soil and parameters such as pH, CEC, clay content, organic matter content and the availability of other nutrients such as Fe, Mg and Al. This helps the model learn not only the physical relationship between reflectance and P but also the chemical constraints that limit its availability in soil which cannot be detected from reflectance alone.

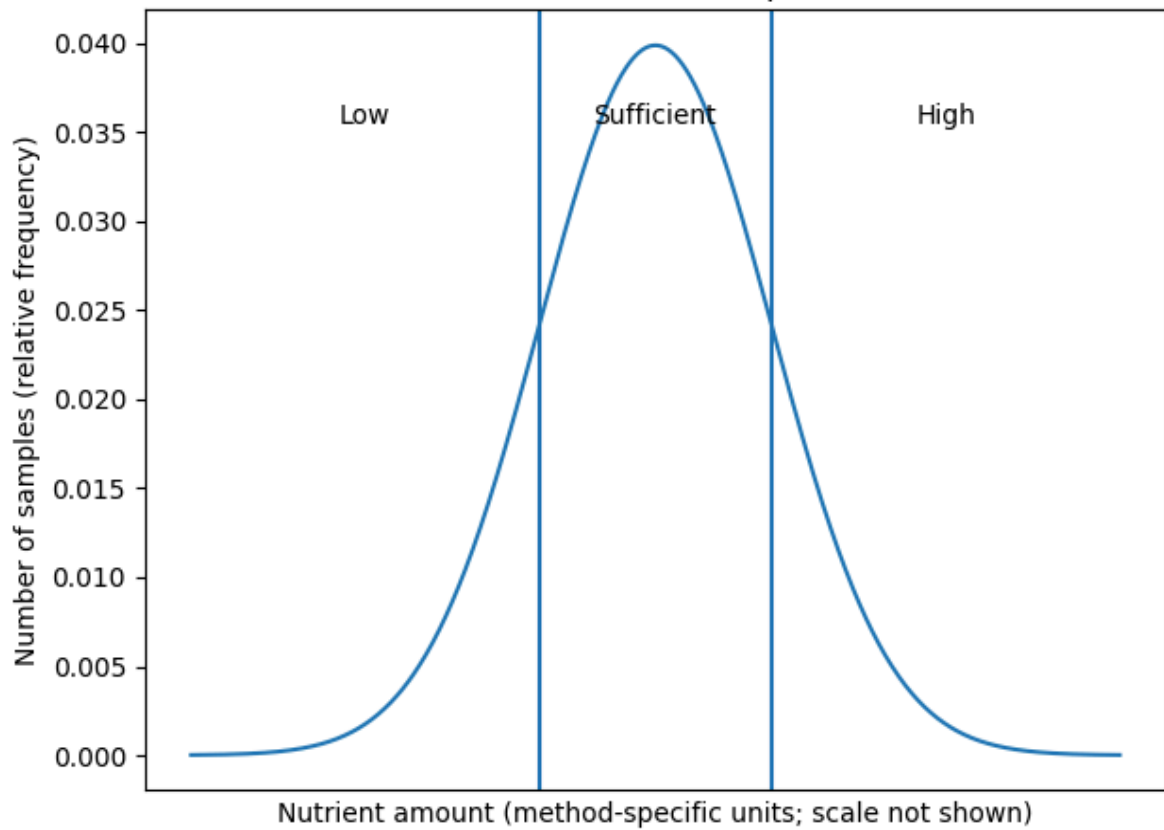
3.2 Interpretability

As already mentioned, comparing nutrient measurements from different labs even within the same region can be quite challenging due to the difference in scale, extraction methodology, calibration, measurement accuracy and even sampling techniques. All of these factors make

it difficult even to convert measurements from one laboratory to another. Since scales across labs and methods differ widely, building a model that predicts nutrient levels on a single scale makes the model difficult to interpret in a new country with different lab methods. Achieving interpretability and comparability across diverse laboratory methods and regions therefore presents a significant challenge.

To solve this task we decided to not use a single scale rather to build a model that predicts the relative quantile in which the nutrients of a given field fall within a wider nutrient distribution within the region (**Cade & Noon, 2003**). This approach relies on the assumption that, despite differences in reported absolute values, laboratories assign comparable agronomic status (e.g., deficient, sufficient, excessive) to fields occupying similar positions within their respective nutrient distributions (**Mallarino, 2003**). This is true because the way in which the agronomic categories have been derived for the given method is by collecting a large number of samples from different fields within a region and then ordering these fields by the magnitude of the measured nutrient. Then fields that fall on the lower end of the distribution are labelled as deficient for the given nutrient and the absolute values that this quantile corresponds to become the boundaries (in mg/kg, meq, ppm or other units) of the “low” nutrient availability category for the given method in this region. Other labs that use different methods will still follow this procedure to estimate their nutrient availability ranges. And even though the values and measurement units might be different, all labs will agree that a field with insufficient nitrogen levels should receive the “insufficient” label and not the “toxic” or “sufficient” one. Thus the universal language that different labs and measurement methods speak is that of agronomic labels.

Lab Method A: Distribution in Method-Specific Units (Illustrative)



Lab Method B: Distribution in mg/kg (Illustrative)

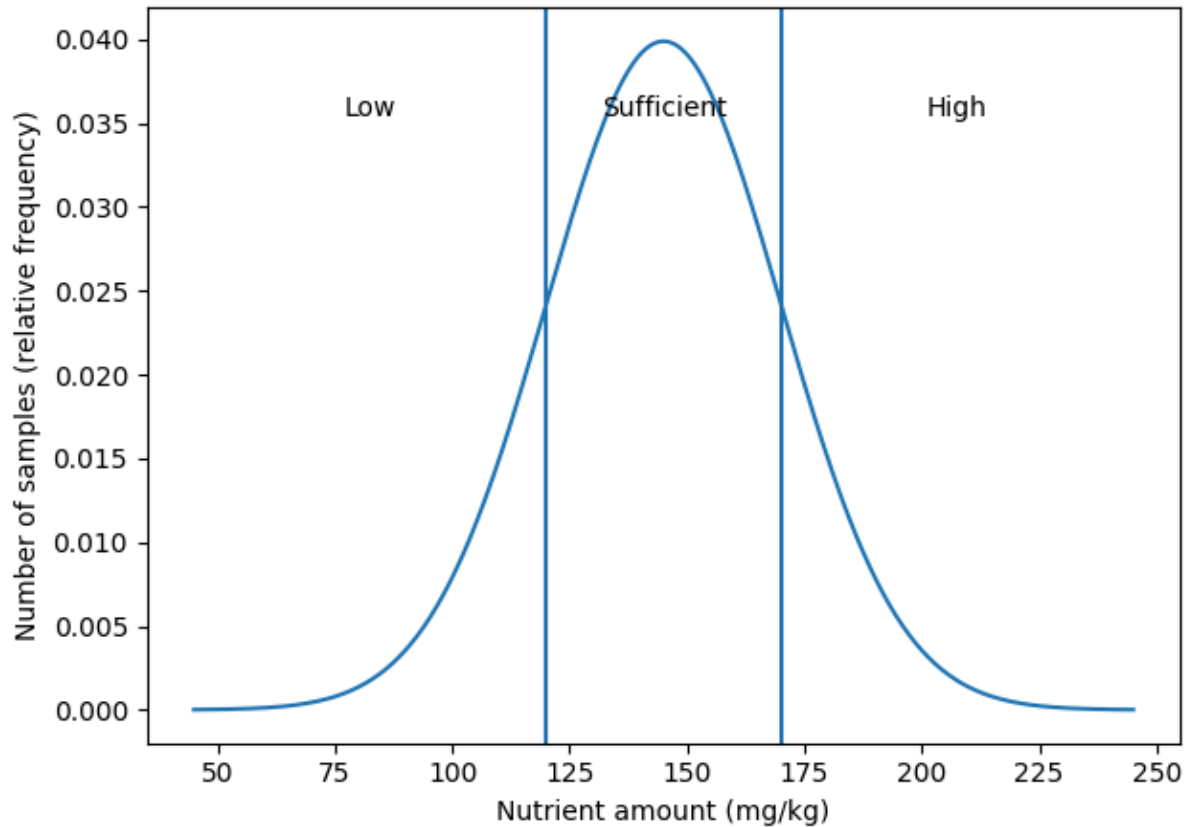


Fig 2. Schematic representation of nutrient distributions under different laboratory measurement scales

To make use of this our model predicts a “nutrient index” ranging from 0 to 100 which corresponds to the percentile of fields that exhibit the same spectral signature within this region for the given nutrient. Conversion to lab units (mg, ppm etc) is done by mapping the estimated percentile to the nearest quantile (agronomic category) of the reference lab method and calculating the distance between the model’s estimated percentile and the upper and lower bounds of the lab method’s ranges.

4. Evaluation Metrics

Because our model estimates the percentile in which the observed field falls with respect to the wider distribution of the given nutrient in the region, our main evaluation metric is the Spearman rank. To measure the performance in laboratory units we convert the predicted percentiles to lab units via the method described in 3.2 and then compare the converted values and the lab values using the following metrics: Correlation Concordance Coefficient (CCC), R² and MAE. These are secondary metrics because the model is optimized for rank consistency rather than absolute magnitude, scale-sensitive metrics such as R² are expected to underestimate practical utility. Furthermore these metrics are also susceptible to the percentile transformer’s calibration error. If the percentile transformer for a given model is ill-defined or not properly calibrated it can introduce a bias error which can skew these metrics.

5. Results

The results provided below are obtained by converting the percentile measurements that the model produces to lab units using a non-calibrated standard percentile transformer. In practice the percentile conversion itself can introduce additional error especially for scale sensitive metrics such as R², CCC and MAE. In practice we tune the percentile transformer for each region and the results may differ.

Parameter	Spearman	CCC	R ²	MAE
N	0.613	0.544	0.374	0.3198
P	0.385	0.445	0.167	18.16
K	0.840	0.57	0.54	24.9
OC	0.713	0.689	0.486	0.468
pH	0.904	0.904	0.822	0.626

Table 1. Per parameter metrics on the LUCAS validation set. MAE values are reported in mg/kg for P and K, g/kg for N and OC, and pH units for pH.

Parameter	Spearman	CCC	R2	MAE
N	0.540	0.459	0.303	0.359
P	0.210	0.14	0.17	20.88
K	0.493	0.442	0.309	93.28
OC	0.495	0.452	0.470	0.480
pH	0.43	0.42	0.24	0.899

Table 2. Per parameter metrics on the LOCO (Swiss) test set. MAE values are reported in mg/kg for P and K, g/kg for N and OC, and pH units for pH.

Parameter	Spearman	CCC	R2	MAE
N	NA	NA	Na	NA
P	0.39	0.27	0.39	20.91
K	0.58	0.49	0.49	57.4
OC	0.49	0.47	0.49	0.46
pH	0.5	0.47	0.67	0.65

Table 3. Per parameter metrics on the partner test set. MAE values are reported in mg/kg for P and K, g/kg for N and OC, and pH units for pH.

6. Discussion

Across all evaluation settings, the proposed framework demonstrates that relative nutrient position can be inferred from multispectral data with moderate to strong rank consistency, even under strict spatial and temporal generalization constraints. Among the evaluated parameters, pH exhibits the strongest and most consistent performance, maintaining high Spearman and R^2 scores across all three test sets. This behavior suggests strong generalization and is expected given that the primary drivers of soil pH—such as ferric oxidation, organic matter content, and calcium carbonate—exhibit pronounced and stable reflectance features in the VIS–NIR spectrum **Gholizadeh et al. (2018)**.

Organic carbon (OC) and nitrogen (N) show stable and sufficient performance across all datasets. This is consistent with the design of modern multispectral satellite sensors, particularly Sentinel-2, which are optimized for detecting vegetation and soil organic matter

signals. In addition, the strong biogeochemical coupling between OC and N enables the model to exploit shared spectral and contextual features, contributing to their comparatively robust generalization.

Potassium (K) demonstrates strong performance on the LUCAS validation set and moderate performance on the Swiss and partner test sets. The observed degradation on the LOCO Swiss dataset is expected, as this evaluation explicitly tests the model's ability to generalize to an unseen region with potentially different soil mineralogy and management practices. Nonetheless, the retention of moderate rank-based performance indicates that the model captures transferable signals related to potassium availability rather than region-specific calibration artifacts.

Phosphorus (P) exhibits the weakest overall performance, which is consistent with the lack of strong direct VIS–NIR spectral signatures for soil phosphorus, particularly at moderate to high availability levels. Plants generally do not display distinct spectral deficiency cues when phosphorus is abundant, which leads the model to occasionally misclassify high-P soils as sufficient. From an agronomic perspective, this behavior is acceptable provided that phosphorus-deficient soils are not misclassified as sufficient or excessive. To assess this risk, categorical classification accuracy was evaluated, yielding a value of 0.729 on the LUCAS validation set. This indicates that the model correctly assigns phosphorus sufficiency status in approximately 73% of cases. While this performance is expected to decline in fully unseen regions, it remains sufficient for guiding fertilization decisions where the primary objective is the identification of deficient soils. The reduced performance observed in the Swiss dataset further highlights the challenges of phosphorus generalization and underscores the need for region-specific contextual calibration.

Despite relatively modest R^2 values when predictions are expressed in laboratory concentration units, the consistently higher Spearman correlations demonstrate that the model reliably preserves the relative ordering of fields within nutrient distributions. From an agronomic standpoint, this property is often more actionable than precise concentration estimates, as fertilizer recommendations are typically based on categorical sufficiency classes rather than exact nutrient values. In practice, the proposed nutrient index can be used to prioritize field scouting, guide variable-rate fertilization, and identify zones at risk of nutrient deficiency, even in regions where laboratory calibration data are sparse or unavailable.

Finally, it should be noted that the Swiss and partner test evaluations were conducted using an uncalibrated percentile transformer with heuristically reconstructed nutrient distributions corresponding to the respective laboratory methods. In operational deployments, the percentile transformer would be empirically fitted using region- and laboratory-specific reference data. Such calibration is expected to substantially reduce scale-dependent errors and further improve R^2 and concordance-based performance metrics without altering the underlying rank-based predictions.

7. Limitations and Future Work

While the proposed framework demonstrates robust performance across multiple nutrients and evaluation settings, several limitations should be acknowledged. First, the reliance on multispectral satellite data inherently constrains the model's sensitivity to nutrients that lack strong or direct spectral signatures in the narrow band ranges that these satellites offer. This limitation is most evident for phosphorus, where prediction performance depends heavily on indirect cues derived from auxiliary data and plant response rather than direct soil reflectance. As a result, model performance for phosphorus is more sensitive to regional soil chemistry, crop composition, and management practices, which may limit generalization in fully unseen regions.

Second, the framework estimates relative nutrient position within empirically derived distributions rather than absolute concentrations. While this formulation aligns well with agronomic decision-making and enables cross-laboratory comparability, it also means that scale-dependent metrics such as R^2 and MAE are influenced by the quality of the percentile-to-concentration transformation. In this study, evaluations on the Swiss and partner datasets relied on heuristically reconstructed nutrient distributions due to limited access to region-specific laboratory calibration data. Consequently, reported scale-based errors should be interpreted as conservative estimates of operational performance.

Third, the model does not explicitly account for short-term management interventions such as recent fertilization, tillage, or irrigation events, which can transiently affect soil nutrient availability without producing detectable spectral or contextual signatures. This may introduce uncertainty when predictions are generated shortly after such interventions, particularly for mobile nutrients such as nitrogen.

In addition, the spatial resolution of available multispectral imagery (10–30 m) imposes limits on the detection of fine-scale soil heterogeneity, especially in small or highly fragmented fields. Mixed pixels containing vegetation, residue, or non-agricultural surfaces may further dilute soil-specific signals despite preprocessing and masking efforts.

Future work will focus on addressing these limitations through several avenues. Incorporating hyperspectral data, where available, may improve sensitivity to weakly detectable nutrients by leveraging finer spectral resolution. Spectral embeddings generated from multi-year multispectral or hyperspectral images can help improve the accuracy and reliability of less variable soil parameters such as pH by helping separate the transient management induced effects from the mineralogical processes which these embeddings are well suited to capture.

From an operational perspective, future deployments will prioritize empirical calibration of percentile transformers using region- and laboratory-specific reference datasets. This calibration is expected to substantially improve scale-dependent evaluation metrics while preserving the rank-based consistency that underpins the framework's interpretability. Finally, extending the approach to jointly model nutrient interactions and uncertainty estimates would further enhance its utility for risk-aware agronomic decision support.

References

- Cade, B. S., & Noon, B. R. (2003).** A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8), 412-420.
- Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J. C., ... & Justice, C. (2018).** The Harmonized Landsat and Sentinel-2 (HLS) product. *Remote Sensing of Environment*, 219, 145-161.
- Gholizadeh, A., Žižala, V., Saberioon, M., & Borůvka, L. (2018).** Soil organic carbon and texture retrieving and mapping using Sentinel-2 data and machine learning. *Geoderma*, 330, 115-126.
- Hengl, T., et al. (2017).** *SoilGrids250m: Global gridded soil information based on machine learning*.
- Mallarino, A. P. (2003).** Field calibration for corn of the Melich-3 soil phosphorus test. *Soil Science Society of America Journal*, 67(6), 1928-1934.
- Meyer, H., & Pebesma, E. (2021).** Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9), 1620-1633.
- Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., & Fernández-Ugalde, O. (2018).** LUCAS Soil, the largest expandable soil dataset for Europe: a review. *European Journal of Soil Science*, 69(1), 140-153.
- Osborne, S. L., Schepers, J. S., Francis, D. D., & Schlemmer, M. R. (2002).** Detection of phosphorus and nitrogen deficiencies in corn using spectral radiance measurements. *Agronomy Journal*, 94(6), 1215-1221.
- Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019).** *Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations*. Journal of Computational Physics.
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Aroita, G., ... & Dormann, C. F. (2017).** Cross-validation strategies for spatio-temporal machine learning models. *Ecography*, 40(8), 913-929.
- Vaudour, E., Gilliot, J. M., Bel, L., Lefevre, J., & Chehata, N. (2019).** Regional-scale mapping of organic carbon in surface soils using Sentinel-2 data. *Remote Sensing*, 11(18), 2148.