A Machine Learning Framework for Automated Anomaly Detection in Precision Agriculture Geospatial Data

Zachary Komarnisky[1], Felippe H. S. Karp[1]

[1] Olds College of Agriculture & Technology, Olds, Alberta, Canada

Corresponding author e-mail: zachary.komarnisky@oldscollege.ca

A paper from the Proceedings of the

17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture

13–15 July 2026

Porto Alegre, Brazil

Abstract. Modern precision agriculture relies on the analysis of geospatial data generated by a wide range of equipment and sensors. While these datasets are foundational to data-driven management practices, they are often affected by inaccuracies arising from various sources. Existing filtering systems and software, that implement operational (e.g., abrupt changes in speed, speed limits, and removal of maneuvers), global statistical (e.g., observations that are inconsistent with the overall distribution of the data) and spatial outliers (e.g., observations that are inconsistent with their neighborhood) have been widely adopted in agriculture, but they require expert knowledge and parameter tuning, which limits their deployment at scale. This project aimed to build a rigorously annotated dataset and develop a machine learning framework capable of processing multiple geospatial data types to automatically detect and remove data anomalies. Apparent electrical conductivity (EC_a) and yield-monitoring datasets from the Olds College of Agriculture & Technology Smart Farm (Olds, Alberta, Canada) were manually analyzed, and each observation was annotated as valid or one of three anomaly classes: operational, global, and local. QGIS was used for spatial visualization, and custom Python scripts were developed to automate preprocessing and deploy machine learning models. Generalization was assessed with a leave-one-sensor-out scheme, in which an entire field was withheld for testing while the remaining fields were used for training, so that performance reflects transfer to an unseen sensor. Random Forest (RF) and XGBoost models were assessed, and both detected and classified anomalies across the held-out datasets. The annotated data enabled RF to reproduce the bulk of the manual annotation reliably, recovering clean observations (F1-score = 0.95) and partially transferring operational anomalies (F1-score = 0.52) across sensors, while performance on local spatial anomalies remained limited (F1-score = 0.30). This inconsistency is likely related to the non-spatial architecture of the tested models, which do not explicitly account for the spatial dependencies inherent in yield and EC_a data. Training spatially-explicit architectures (e.g., Graph Neural Networks) is a logical evolution for the framework. In conclusion, a high-quality dataset was created to support the initial development of a precision agriculture-focused machine-learning-based filter. Future research will evaluate the use of spatially-aware models and expand the manually annotated dataset.

Keywords. precision agriculture, anomaly detection, data cleaning, machine learning, apparent electrical conductivity, yield monitoring

Introduction

Precision agriculture depends on large volumes of geospatial data collected by a variety of field equipment and sensors, yet these data are frequently degraded in quality by operational errors, global statistical outliers, and localized false values. Rule-based filters such as Yield Editor (Sudduth and Drummond, 2007) and Map Filter (Maldaner et al., 2022) remove such errors using operational, global, and spatial-neighborhood criteria, but they rely on expert knowledge and manual parameter tuning, which limits their deployment at scale (Karp et al., 2022). The objectives of this study were (i) to build a rigorously annotated, multi-sensor geospatial dataset, and (ii) to develop a machine learning framework able to process multiple geospatial data types and automatically detect and remove data anomalies.

Materials and Methods

The data were collected from fields at the Olds College of Agriculture & Technology Smart Farm (Olds, Alberta, Canada). Apparent electrical conductivity (EC_a) was measured with an EM38 sensor (Geonics Limited, Mississauga, Ontario, Canada) and a Veris 3100 platform (Veris Technologies, Salina, Kansas, USA), and grain yield was recorded with the combine's GreenStar Harvest Monitor (Deere & Company, Moline, Illinois, USA). Each observation was manually annotated in QGIS version 3.28 (QGIS Development Team, 2022) as a valid observation or as one of three anomaly classes: operational (e.g., maneuvers and abrupt speed changes), global statistical (values inconsistent with the overall distribution of the data), or local (values inconsistent with the spatial neighborhood). Global and local z-scores (described below) were used to aid the manual annotation. The annotated dataset comprised 5,075 EM38 EC_a, 10,226 Veris EC_a, and 34,364 yield observations.

Custom Python 3.11 scripts were developed to implement the framework and automate preprocessing and model deployment. Preprocessing included calculating motion features (distance between consecutive points, heading, heading change, time interval, speed, and acceleration) and value features that were derived from global and spatial outlier detection principles established by Sudduth and Drummond (2007) and Vega et al. (2019), comprising a field-level z-score, a local neighborhood z-score computed at three fixed spatial radii (15, 30, and 45 m), and the ratio of each observation to its transect mean.

Labeled observations from each field were partitioned into four spatially contiguous subfields using K-means clustering applied to geographic coordinates (k=4). One subfield per field was held out as the spatial validation set during training. The remaining three subfields were used for model fitting. Generalization was assessed using a Leave-One-Sensor-Out (LOSO) cross-validation scheme in which one entire dataset from that sensor was withheld as the test set across three folds, with the model trained on the remaining sensors in each fold. Random Forest (Breiman, 2001) was implemented using scikit-learn (Pedregosa et al., 2011; v1.9.0), and gradient boosting using the XGBoost library (Chen and Guestrin, 2016; v3.2.0) both using a final set of 13 motion and value features. Both models were trained with class-balanced weighting. Model performance was evaluated using per-class precision, recall, macro- and weighted-average F1, and a confusion matrix.

Results and Discussion

Both classifiers transferred across sensors under the LOSO scheme, sharing the same overall pattern (Table 1). RF reached a macro-averaged accuracy of 0.90 across the three held-out folds (EM38 0.89, Yield 0.89, Veris 0.94) and an overall F1 of 0.59, while XGBoost reached 0.83 (EM38 0.73, Yield 0.88, Veris 0.89) and an overall F1 of 0.55, with EM38 the weakest fold for XGBoost specifically. Clean observations were reliably recovered across all held-out sensors (RF F1 = 0.95, XGBoost F1 = 0.90). Operational anomalies (i.e., maneuvers, headland turns, and abrupt speed changes) were the best-transferred anomaly class (RF F1 = 0.52, XGBoost

F1 = 0.40), as their distinctive kinematic signature is captured directly by the motion features and is largely sensor-independent. Local spatial anomalies were transferred poorly (RF F1 = 0.30, XGBoost F1 = 0.34); most undetected local outliers were labeled clean, reflecting the limited sensitivity of models that treat each observation independently and do not represent spatial neighborhood structure. The global class was too rare to evaluate, affecting the model training and predictions.

Despite this limitation, the framework reproduces the bulk of the manual annotation across heterogeneous sensors without per-field parameter tuning. Spatially explicit architectures, such as Graph Neural Networks that propagate information between neighboring observations, are a logical next step, and expanding the annotated dataset is expected to further stabilize performance.

Table 1. Per-class classification performance, macro-averaged across three leave-one-sensor-out folds. The global class is excluded from per-class reporting (14 examples across all folds).

Class	RF Prec.	RF Rec.	RF F1	XGB Prec.	XGB Rec.	XGB F1
Clean	0.96	0.94	0.95	0.96	0.85	0.90
Operational	0.46	0.63	0.52	0.30	0.65	0.40
Local	0.44	0.24	0.30	0.32	0.38	0.34

Prec.: Precision; Rec.: Recall; F1: F1-score

Conclusion

A high-quality, manually annotated dataset spanning multiple geospatial data types was created to support the initial development of a precision agriculture machine learning filter. Under a LOSO evaluation, tree-ensemble models transferred clean and operational anomalies across sensors but generalized poorly to local spatial anomalies, underscoring the limits of non-spatial architectures. Future research will evaluate spatially-aware models, such as Graph Neural Networks, and expand the manually annotated dataset.

Acknowledgments

This work was supported by an NSERC CCI Mobilize Grant (CCMOB-2024-00038). The first author gratefully acknowledges the Olds College Office of Research Services and the Werklund School of Agriculture Technology for funding toward conference participation.

The authors used Claude (Anthropic, PBC) to assist with manuscript drafting, reference verification, and iterative text editing. All scientific content, interpretations, and conclusions are the authors' own. The authors take full responsibility for the accuracy and integrity of the work as submitted.

References

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

Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

Karp, F. H. S., Adamchuk, V., Melnitchouck, A., and Dutilleul, P. (2022). Optimization of batch processing of high-density anisotropic distributed proximal soil sensing data for precision agriculture purposes. In Proceedings of the 15th International Conference on Precision Agriculture.

Maldaner, L. F., Molin, J. P., and Spekken, M. (2022). Methodology to filter out outliers in high spatial density data to improve maps reliability. Scientia Agricola, 79(1), e20200178. https://doi.org/10.1590/1678-992X-2020-0178

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

QGIS Development Team. (2022). QGIS Geographic Information System (version 3.28). Open Source Geospatial Foundation Project. https://qgis.org

Sudduth, K. A., and Drummond, S. T. (2007). Yield Editor: software for removing errors from crop yield maps. Agronomy Journal, 99(6), 1471–1482. https://doi.org/10.2134/agronj2006.0326

Vega, A., Córdoba, M., Castro-Franco, M., and Balzarini, M. (2019). Protocol for automating error removal from yield maps. *Precision Agriculture*, 20, 1030–1044. https://doi.org/10.1007/s11119-018-09632-8