Data augmentation and interpolation improves machine learning-based pasture biomass estimation from Sentinel-2 imagery

B.N. Azubuike[ADE], A. Chlingaryan[BD], M. Correa-Luna[AD], C.E.F. Clark[CD], and S.C. Garcia[AD]

[A]Dairy Science Group, School of Life and Environmental Sciences, Faculty of Science, The University of Sydney, Camden, NSW 2570, Australia
[B]Livestock Production and Welfare Group, School of Life and Environmental Sciences, University of Sydney, Camden, NSW 2570, Australia
[C]Gulbali Institute, Charles Sturt University, Wagga Wagga, NSW 2650, Australia
[D]Dairy UP Program, Camden, NSW 2570, Australia
[E]Corresponding author. Email: blessing.azubuike@sydney.edu.au

Accurate pasture biomass (PB) estimation is critical for tactical grazing decisions, but conventional satellite-derived vegetation indices such as Normalised Difference Vegetation Index (NDVI) saturate once canopies exceed ≈3 t DM ha^{-1}, significantly limiting predictive accuracy. This study integrated raw Sentinel-2 (a satellite providing free multispectral imagery at 10–20 m resolution every 5 days) reflectance, rising plate meter (RPM) PB, climatic, and paddock predictors to enhance PB estimation in dairy systems. Data comprising 3,161 observations from 80 paddocks across 16 dairy farms in New South Wales (NSW), collected between November 2021 and July 2024, were utilised. Multiquadric interpolation of RPM measurements bridged temporal gaps between field measurements and cloud-free imagery, expanding the dataset and reducing prediction errors. Eight regression algorithms and four predictor sets were evaluated via five-fold, three-repeat cross-validation on an 80:20 farm/paddock-stratified train:test set split, complemented by independent validation sets and leave-farm-out tests. An XGBoost model, utilising full-band reflectance and concurrent weather data, achieved robust performance (R^2 = 0.63; mean absolute error (MAE) = 243 kg DM ha^{-1}) on the test set (20%) without interpolation. Augmenting the dataset with ~30% synthetic rows via multiquadric interpolation of the RPM measurements improved test set performance to (R^2 = 0.70 MAE = 216 kg DM ha^{-1}), with gains sustained on external validation sets (R^2 = 0.41/0.48; MAE = 267/235 kg DM ha^{-1}). Progressive training, continuously refreshing model parameters, reduced errors by ~30% compared to a sparse, weekly baseline, maintained accuracy even with withheld farms or seasons. This Sentinel-2 workflow, demonstrated comparable accuracy to Pasture.io, a commercial platform using satellite imagery with higher temporal and spatial resolution than Sentinel-2, with an R^2 = 0.66 and MAE of 240 kg DM ha^{-1}. This research provides a clear path towards automated paddock-level recalibration and near real-time farm-specific decision support for pasture-based dairy systems.

Keywords: data augmentation, machine learning, pasture biomass estimation, remote sensing, Sentinel-2