

Area: ANA

Classification of *Apis mellifera* and Stingless Bee Pollen Using NIR Spectroscopy Combined with Machine Learning

Caroline Simão (PG)^{1*}, Vanderlei Ap. Lima (PQ)³, Maria Lurdes Felsner (PQ)^{1,2}

carolinesimao1900@gmail.com; felsner@unicentro.br

¹Department of Chemistry, State University of Central-West (UNICENTRO) – Guarapuava, PR; ²Department of Chemistry, State University of Londrina – Londrina, PR; ³Department of Chemistry, Federal University of Technology of Paraná - Pato Branco, PR.
Keywords: Bee pollen; NIR spectroscopy; Machine learning; Classification; Green chemistry.

Highlights

Rapid and sustainable NIR-based strategy for classifying bee pollen.

Collector bee species discrimination achieved with high accuracy.

Random Forest model reached up to 95.3% external validation accuracy.

Abstract

Bee pollen is a complex natural product rich in proteins, amino acids, lipids, vitamins, and bioactive compounds, widely consumed as a functional food and applied in nutraceuticals and cosmetics. Beyond its nutritional relevance, pollen analysis is fundamental in apiculture and meliponiculture because its biochemical profile reflects both the floral sources visited and the species of bee that collected it. Therefore, classifying pollen according to the collector species is crucial not only for quality assurance and consumer safety, but also for nutritional evaluation and biodiversity research.

Near-infrared (NIR) spectroscopy provides comprehensive information on the chemical composition of bee pollen, making it a valuable tool for nutritional and botanical evaluation. In this study, NIR spectra of pollen loads from *Apis mellifera* and pollen pots from stingless bees were acquired using a Bruker MPA FT-NIR spectrometer (quartz cup, probe mode) over the range of 3600–12500 cm⁻¹ (~800–2778 nm). Spectra displayed distinctive absorption patterns across the range, reflecting variations in biochemical composition influenced by both bee species and floral sources. These characteristic spectral fingerprints enabled the discrimination of bee species based on the pollen they collected. To enhance classification efficiency, NIR data were modeled in Python using the Random Forest (RF) algorithm. The model was trained with a stratified division of the dataset (70% calibration, 30% external validation) and further validated through cross-validation. During training, RF achieved 100% accuracy, indicating excellent model fitting. In external validation, the model maintained robust performance with an overall accuracy of 95.3% and an average cross-validation accuracy of 97.9%. Class-specific metrics confirmed consistent results: *Scaptotrigona bipunctata* pollen reached 100% accuracy, while the lowest accuracy (91.4%) was observed for *Melipona quadrifasciata quadrifasciata*. Precision, sensitivity, and F1-scores ranged from 91% to 100%, confirming stable predictive capacity across all bee species. The Random Forest algorithm, based on ensemble decision trees, proved advantageous in handling the complexity and high dimensionality of NIR spectra. Its ability to mitigate overfitting through bootstrapping and random feature selection ensured generalizable and reliable predictions. These findings highlight the robustness of RF for multiclass classification tasks involving subtle spectral differences. This work demonstrates the potential of combining NIR spectroscopy with Random Forest as a sustainable, rapid, and non-destructive approach for bee pollen classification. Beyond providing reliable authentication, the method supports biodiversity monitoring, enhances traceability, and contributes to green analytical chemistry practices in apiculture and meliponiculture.

Acknowledgments

We gratefully acknowledge the financial support provided by CAPES and the Araucária Foundation.