Aditya Singh

Milk Powder Origin Classification Using Trace Element Fingerprinting

Overview

This project investigates whether trace elemental composition (ppm) can be used to accurately classify the regional origin of milk powder samples. Using ICP-style elemental data and a machine learning approach, a Random Forest classifier was developed and validated to determine whether geographic provenance can be reliably predicted.

The final model achieved ~96% classification accuracy under repeated cross-validation, demonstrating strong regional signals in elemental profiles.

Methodology:

Data Preprocessing:

The dataset contained trace element concentrations across multiple regions.

Values reported as <LOQ (below limit of quantitation) were replaced using:

Element-specific LOQ ÷ 2

This approach:

- Reduces downward bias compared to zero substitution
- Preserves distributional structure
- Is standard practice in trace chemical analysis

All elemental predictors were converted to numeric prior to modelling.

Random Forest Modelling:

A Random Forest classifier (1000 trees) was trained using all available elemental predictors.

Model selection was guided by:

- Mean Decrease Accuracy (MDA) — reduction in classification accuracy when a predictor is permuted (primary selection metric)
- Mean Decrease Gini (MDG) — contribution to split purity (supporting metric)

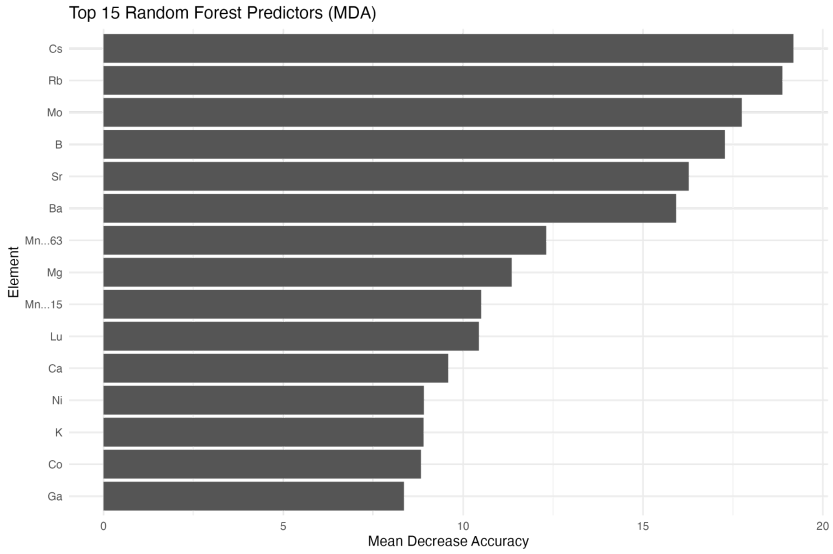

Figure 1: Random Forest variable importance ranked by Mean Decrease Accuracy (MDA).

Feature Selection:

Elemental predictors were ranked by MDA and the top 10 predictors were retained.

This reduced model:

- Lowered class error compared to the full model
- Reduced dimensionality
- Improved interpretability
- Removed low-signal noise variables

This confirms that regional discrimination is concentrated within a limited subset of elemental predictors.

Model Validation

To ensure robustness, model performance was evaluated using:

- Repeated 10-fold cross-validation (5 repeats)

This procedure reduces variance in performance estimates and guards against overfitting.

Key Performance Metrics

- Accuracy: ~96%
- Kappa: ~0.94

High Kappa indicates strong agreement beyond chance, suggesting robust regional discrimination.

Multivariate Structure

Principal Component Analysis (PCA) was performed using the selected predictors to visualise separation in reduced dimensional space.

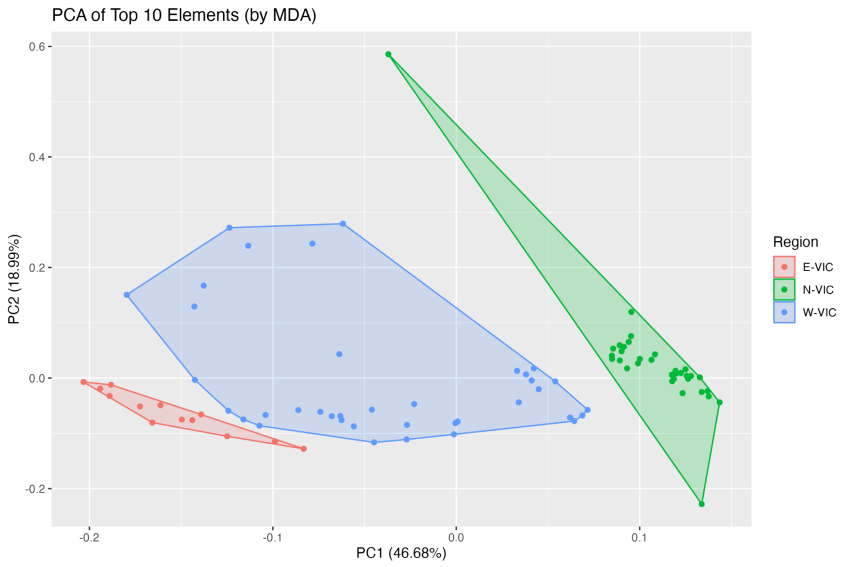

Figure 2: PCA of the top 10 elemental predictors showing regional clustering

Distinct clustering in PCA space supports the Random Forest classification results and indicates structured compositional differences across regions.

Element-Level Differences

To interpret which predictors drive classification, boxplots were generated for the selected elements.

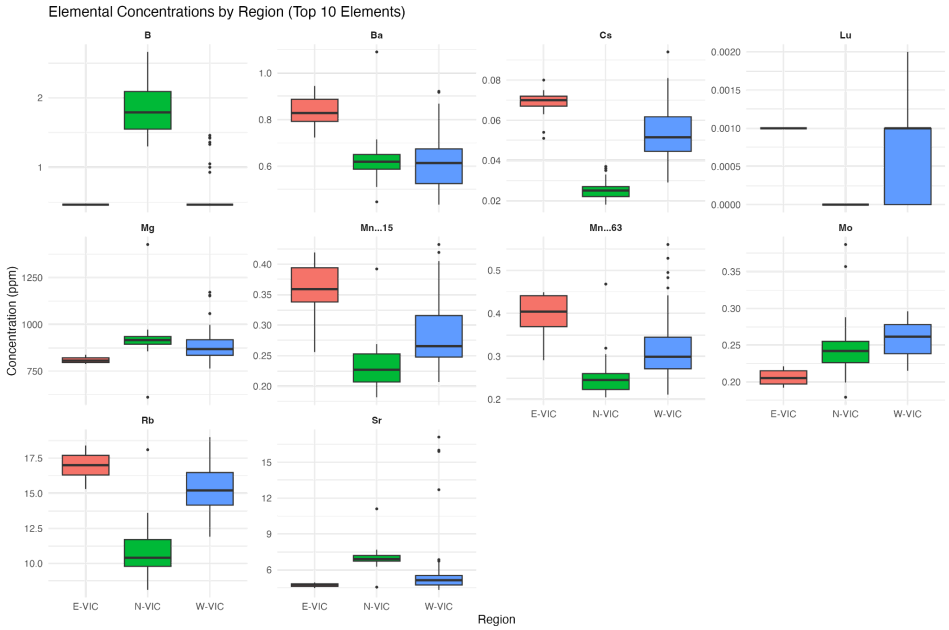

Figure 3: Distribution of selected elemental concentrations across regions

Several predictors exhibit clear shifts in median concentration between regions, supporting their high importance scores in the Random Forest model.

CONCLUSION

This project demonstrates that trace elemental composition contains strong geographic signals suitable for provenance classification. A reduced subset of predictors identified via Random Forest importance ranking achieved high and stable cross-validated accuracy.

The workflow is reproducible, modular, and transferable to other food authentication or traceability applications.

Works Cited

Dolor, L.I. *Lorem ipsum dolor sit amet, consectetuer adipiscing elit*, 1998. Print.

Dolor, L.I. *Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh. New York*: Columbia UP, 1998. Print.

Doe, R. John. *Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh,* 1998. Print.