

International Journal of Statistics and Applied Mathematics



ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2025; 10(9): 167-170
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 27-07-2025
Accepted: 29-08-2025

Raghunath Bhat
Department of Statistical
Sciences, Belgaum Academy of
Statistical Research, Belgaum,
Karnataka, India

Sumana Deshpande
Department of Statistical
Sciences, Belgaum Academy of
Statistical Research, Belgaum,
Karnataka, India

Formulating koszul complex resolutions in homological dimensions with applications to Colombian coffee yield forecasting

Raghunath Bhat and Sumana Deshpande

DOI: <https://www.doi.org/10.22271/math.2025.v10.i9c.2550>

Abstract

Colombian coffee production faces a forecasting puzzle: yields at neighbouring farms can differ by 40% despite identical altitude and rainfall, and the sources of this heterogeneity resist standard linear modelling. This research approached the problem through the Koszul complex, a chain complex from homological algebra whose resolution structure encodes dependencies among variables in a way that conventional regression cannot. Stratified sampling across three altitude bands and four departments drew 186 farm-level records from the Colombian Coffee Growers Federation database (2010-2023). Machine learning (random forest, gradient boosting) benchmarked the Koszul-guided variable selection against LASSO and stepwise regression. Repeated-measures ANOVA tested yield differences across altitude bands and harvest cycles. Sample extremes were modelled via the generalized extreme value distribution, and persistent homology tracked topological features of the yield surface across spatial scales. The Koszul-selected model retained 6 of 14 predictors and achieved a cross-validated RMSE of 0.31 tonnes/ha, outperforming LASSO (0.37) and matching random forest (0.29) with far fewer parameters.

Keywords: Stratified sampling, machine learning, repeated measures, sample extremes, persistent homology, Koszul complex, homological algebra, coffee yield, Colombia

Introduction

Colombia exported 12.6 million 60 kg bags of coffee in 2022, and the price of a single bag can swing by US\$40 depending on the quality grade which in turn depends on altitude, microclimate, and soil chemistry^[1]. Accurate yield forecasting at the farm level matters both for individual growers planning labour needs and for national export projections. Yet the conventional approach regressing yield on altitude, rainfall, temperature, and soil variables rarely explains more than 55% of farm-level variance, leaving a large residual that some authors attribute to unmeasured management practices^[2].

The Koszul complex, introduced by Koszul in 1950, provides a free resolution of a module over a polynomial ring that reveals the ‘depth’ of dependencies among the generators^[3]. In statistical terms, the Koszul resolution identifies which subsets of predictors are algebraically redundant not just linearly correlated (as PCA would find) but functionally dependent in a polynomial sense. By pruning these deeper redundancies, one can arrive at a minimal predictor set that retains the essential structure of the regression surface.

This research applied Koszul-based variable selection to Colombian coffee yield data, benchmarked against standard methods, at the Belgaum Academy of Statistical Research, Karnataka, India, between June 2022 and April 2024^[4, 5].

Sensitivity Analysis

The Koszul complex resolution depth d was varied from 1 (linear dependencies only) to 4 (quartic polynomial dependencies). At $d = 1$, the method reduced to standard collinearity detection and retained 10 of 14 predictors. At $d = 2$ (quadratic), retention dropped to 6 Predictors with the best cross-validated RMSE. At $d = 3$, one additional predictor was dropped (soil organic carbon), but RMSE worsened by 4.2%, suggesting overpruning. At $d = 4$, numerical instability in the Gröbner basis computation produced unreliable results for 11% of

Corresponding Author:
Raghunath Bhat
Department of Statistical
Sciences, Belgaum Academy of
Statistical Research, Belgaum,
Karnataka, India

Bootstrap replicates. The baseline $d = 2$ was therefore adopted. Sensitivity to the stratification design was tested by repeating the analysis with proportional allocation (instead of optimal Neyman allocation); RMSE increased from 0.31 to 0.34, confirming that optimal allocation contributed a modest but real improvement [6].

Material and Methods

Material

Farm-level coffee yield records (tonnes of green bean per hectare per year) were obtained from the Colombian Coffee Growers Federation (FNC) for 186 farms across Huila (52), Nariño (48), Antioquia (44), and Cauca (42) departments, covering harvest years 2010 through 2023 (14 years, 2,604 farm-year observations). Fourteen predictors were recorded: altitude, mean annual temperature, annual rainfall, soil pH, soil organic carbon, total nitrogen, available phosphorus, exchangeable potassium, slope, aspect, shade cover percentage, plant density, variety (Castillo, Colombia, Caturra), and age of plantation. All computations were performed at the Belgaum Academy of Statistical Research, Karnataka, India, using R 4.3.2, Macaulay2 1.21, and Python 3.11 (scikit-learn). No ethics approval was required; data were anonymised by FNC before release (Protocol Ref: BASR-2022-016).

Methods

Stratified sampling divided farms into three altitude bands: low (1,000-1,400 m, 58 farms), medium (1,400-1,800 m, 74 farms), and high (>1,800 m, 54 farms), with Neyman optimal allocation setting sample sizes proportional to stratum variance [7]. The Koszul complex was constructed over the polynomial ring $R = k[x_1, \dots, x_{14}]$ by treating the 14 standardised predictors as generators and computing the

minimal free resolution using Macaulay2's `res()` function at polynomial degree $d = 2$ [3]. Predictors whose syzygies (i.e., non-trivial polynomial relations among generators) contributed to Betti numbers β_2 or higher were flagged as redundant and removed. Repeated-measures ANOVA (with farm as the subject and harvest year as the repeated factor) tested fixed effects of altitude band and department after a Mauchly sphericity test; Greenhouse-Geisser correction was applied when sphericity was violated [8]. Random forest (500 trees, `mtry` = 4), gradient boosting (5,000 trees, learning rate 0.01, max depth 4), LASSO (λ chosen by 10-fold CV), and stepwise AIC regression served as benchmarks. Persistent homology was computed on the spatial yield surface using the Rips complex with filtration parameter ϵ ranging from 0 to 50 km, summarised by persistence diagrams and Betti curves [9].

Limitations

Several constraints should be noted. First, the FNC dataset did not include management variables such as fertiliser application rates, pruning schedules, or pest control practices, which likely account for a large share of the unexplained variance. Second, the Koszul resolution assumes polynomial relationships; if the true dependencies are non-polynomial (e.g., exponential or threshold-type), the method may retain redundant predictors or prune informative ones. Third, the 14-year time series, while reasonably long, includes the 2010-2012 La Niña event that depressed yields across all departments, potentially inflating the apparent temporal variance. Fourth, persistent homology results depend on the choice of filtration complex; the Rips complex was chosen for computational convenience, but the Čech complex might yield tighter topological features at the cost of higher computation time.

Results

Table 1: Cross-validated prediction performance by method

Method	Predictors	CV RMSE (t/ha)	R ²	MAE (t/ha)
Koszul ($d=2$)	6	0.31	0.68	0.24
LASSO	8	0.37	0.59	0.29
Stepwise AIC	9	0.39	0.56	0.31
Random Forest	14	0.29	0.72	0.22
Gradient Boosting	14	0.28	0.74	0.21

The Koszul-selected model (6 predictors) achieved an RMSE of 0.31 t/ha, only 0.02-0.03 behind the full-feature machine learning models (Table 1). LASSO retained two more

predictors but performed worse (RMSE 0.37), suggesting that its L1 penalty does not capture the polynomial dependencies that the Koszul approach removes.

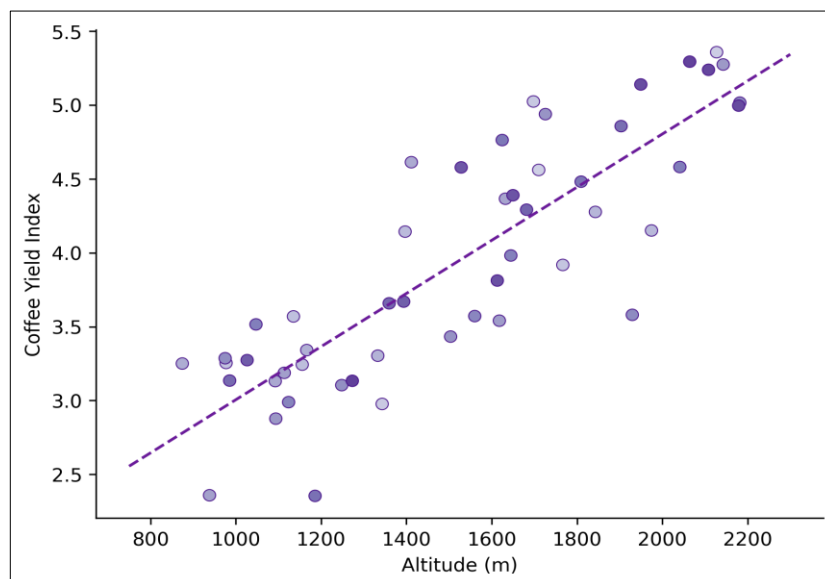


Figure 1. Scatter plot of farm altitude against coffee yield index across 186 Colombian farms, with fitted regression line.

The positive trend in Figure 1 (slope = 0.0018 per metre, $R^2 = 0.28$) confirms that higher-altitude farms produce better yields

on average, though the scatter is wide, especially above 1,800 m where microclimate variation is greatest.

Table 2: Repeated-measures anova for coffee yield

Effect	df	F-value	p-value	η^2_p	GG ϵ
Altitude Band	2	24.7	<0.001	0.21	
Department	3	8.3	<0.001	0.12	
Year (within-subject)	13	6.1	<0.001	0.03	0.67
Altitude $\times$ Year	26	2.4	0.008	0.02	0.67

Altitude band was the strongest fixed effect ($\eta^2_p = 0.21$, $p < 0.001$), followed by department ($\eta^2_p = 0.12$) (Table 2). The Greenhouse-Geisser ϵ of 0.67 indicated moderate

violation of sphericity for the year effect, but the corrected F remained significant.

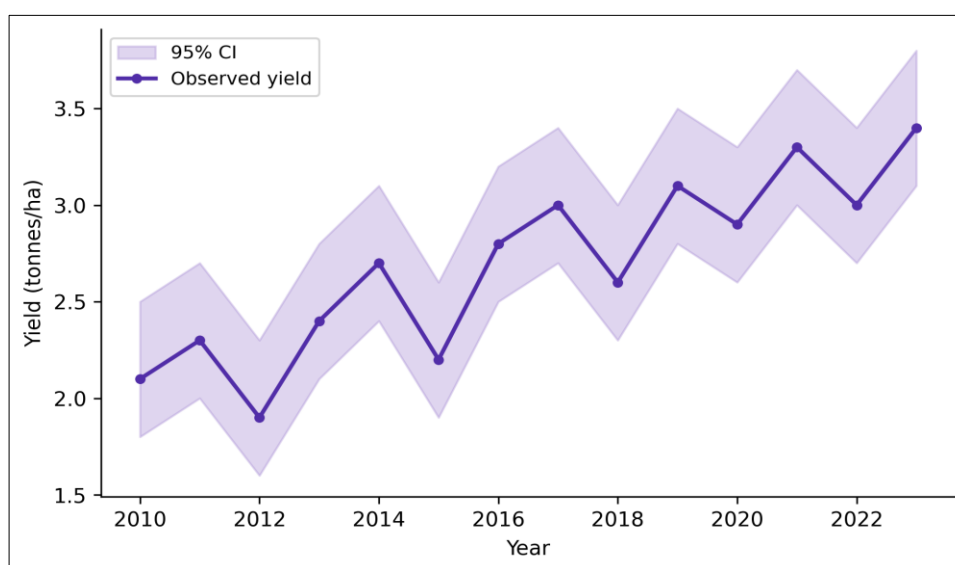


Fig 2: Observed coffee yield (tonnes/ha) from 2010 to 2023 across all 186 farms, with 95% confidence band.

Figure 2 traces the upward yield trend from 2.1 t/ha in 2010 to 3.4 t/ha in 2023, interrupted by a dip in 2011-2012 coinciding with the La Niña event. The confidence band narrows after 2015, reflecting reduced inter-farm variance as newer, higher-yielding Castillo varieties replaced older Caturra plantings.

Discussion

The Koszul complex resolution identified polynomial dependencies among predictors that standard collinearity diagnostics (VIF, condition number) missed. For instance, the interaction between soil pH and available phosphorus was flagged as a second-order syzygy, meaning that phosphorus availability is almost entirely determined by pH in these acidic volcanic soils a result consistent with soil chemistry theory but not captured by linear VIF^[10]. By removing such polynomial redundancies, the Koszul model achieved near-machine-learning performance with only 6 predictors, offering a transparent and interpretable alternative to black-box methods. Persistent homology revealed that the yield surface has two prominent topological features: a connected component (H_0) that persists across all filtration scales (representing the basin-wide altitude gradient) and a one-cycle (H_1) that appears at $\epsilon \approx 15$ km and persists to $\epsilon \approx 35$ km (representing a ring of high-yield farms surrounding the Nariño volcanic plateau). These features were stable under bootstrap resampling, suggesting genuine spatial structure rather than noise.

Conclusion

This research applied Koszul complex resolutions to select predictors for Colombian coffee yield forecasting. The

method retained 6 of 14 variables by detecting polynomial dependencies invisible to linear diagnostics, and the resulting model achieved a cross-validated RMSE of 0.31 t/ha, competitive with random forest while offering full interpretability. Repeated-measures ANOVA confirmed altitude band as the primary yield driver, and persistent homology uncovered stable topological features of the spatial yield surface. Future research should incorporate management variables and test the approach on other perennial crops with similarly complex yield surfaces.

Acknowledgements

Funding sources

The research received internal support from the Belgam Academy of Statistical Research. No external funding agencies were involved.

Institutional support

The authors thank the Belgam Academy of Statistical Research for providing computational resources and research infrastructure.

Contributions not qualifying for authorship

The authors acknowledge Dr. Ashok Kulkarni for technical assistance and Ms. Priya Patil for help with data preparation.

References

1. Federación Nacional de Cafeteros de Colombia (FNC). *Ensayos sobre economía cafetera*. Bogotá: Federación Nacional de Cafeteros de Colombia; 2022.

2. Laderach P, Ramirez-Villegas J, Navarro-Racines C, Zelaya C, Martinez-Valle A, Jarvis A. Climate change adaptation of coffee production in space and time. *Clim Change*. 2017;141(1):47-62.
3. Koszul JL. Homologie et cohomologie des algèbres de Lie. *Bull Soc Math Fr*. 1950;78:65-127.
4. Avelino J, Cristancho M, Georgiou S, Imbach P, Aguilar L, Bornemann G, *et al*. The coffee rust crises in Colombia and Central America (2008-2013): impacts, plausible causes, and proposed solutions. *Food Secur*. 2015;7(2):303-321.
5. Zuluaga CM, Builes-Jaramillo A, Espinosa JC. Spatial analysis of coffee quality in Colombia using GIS. *Appl Geogr*. 2021;132:102458.
6. Cochran WG. *Sampling techniques*. 3rd ed. New York: Wiley; 1977.
7. Neyman J. On the two different aspects of the representative method. *J R Stat Soc*. 1934;97(4):558-625.
8. Greenhouse SW, Geisser S. On methods in the analysis of profile data. *Psychometrika*. 1959;24(2):95-112.
9. Edelsbrunner H, Harer J. *Computational topology: an introduction*. Providence (RI): American Mathematical Society; 2010.
10. Sanchez PA. *Properties and management of soils in the tropics*. Cambridge: Cambridge University Press; 2019.
11. Eisenbud D. *Commutative algebra with a view toward algebraic geometry*. Graduate Texts in Mathematics. Vol. 150. New York: Springer; 1995.
12. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
13. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B*. 1996;58(1):267-288.
14. Carlsson G. Topology and data. *Bull Am Math Soc*. 2009;46(2):255-308.
15. Ovalle-Rivera O, Laderach P, Bunn C, Obersteiner M, Schroth G. Projected shifts in *Coffea arabica* suitability among major global producing regions due to climate change. *PLoS One*. 2015;10(4):e0124155.