

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2025; 10(4): 95-98
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 06-03-2025
Accepted: 12-04-2025

Srinivas Iyengar
Faculty of Mathematical
Analysis, Arunachal Institute of
Advanced Mathematical
Sciences, Itanagar, Arunachal
Pradesh, India

Vimala Chakraborty
Faculty of Mathematical
Analysis, Arunachal Institute of
Advanced Mathematical
Sciences, Itanagar, Arunachal
Pradesh, India

Corresponding Author:
Srinivas Iyengar
Faculty of Mathematical
Analysis, Arunachal Institute of
Advanced Mathematical
Sciences, Itanagar, Arunachal
Pradesh, India

Modeling compositional data analysis with Dirichlet distributions with applications to Tuscan hill slope stability analysis

Srinivas Iyengar and Vimala Chakraborty

DOI: <https://www.doi.org/10.22271/math.2025.v10.i4b.2417>

Abstract

Soil composition the proportional split among sand, silt, clay, organic matter, and carbonate carries direct information about slope stability, yet the unit-sum constraint renders ordinary regression invalid. This research models Tuscan hill soil compositions with Dirichlet distributions, relates the fitted parameters to slope stability factors via a Lie-algebraic transformation of the simplex, and validates predictions against 40 field-measured stability indices. Regression analysis on isometric log-ratio coordinates replaced the raw compositions, prime-number indexing of Dirichlet concentration parameters enabled efficient moment matching, and functional-analysis arguments established the continuity of the stability mapping. Z-transform techniques detrended the seasonal component of soil moisture data before composition estimation. The Dirichlet model captured 87.3% of compositional variance, and the predicted stability factors correlated with field measurements at $r = 0.83$ ($p < 0.001$). Clay content above 40% was associated with a 34% reduction in the stability factor. These results confirm that Dirichlet-based compositional modelling provides a rigorous statistical framework for geotechnical risk assessment on Tuscan hillslopes.

Keywords: Lie algebras, regression analysis, prime numbers, functional analysis, Z-transform, Dirichlet distributions, compositional data, Tuscan hillslopes, slope stability, isometric log-ratio coordinates

Introduction

Tuscany's rolling hill country, famous for its vineyards and olive groves, sits on clay-rich Pliocene marine sediments that are prone to shallow landslides during autumn rainstorms ^[1]. Geotechnical assessment of slope stability traditionally relies on limit-equilibrium methods that treat soil as a homogeneous material characterised by cohesion and friction angle ^[2]. But soil is a mixture: sand, silt, clay, organic matter, and calcium carbonate contribute differently to shear strength, and their proportions vary spatially in ways that standard regression cannot model because the components sum to unity ^[3].

Compositional data analysis, pioneered by Aitchison, replaces raw proportions with log-ratio coordinates that live in unrestricted Euclidean space, making standard statistical tools applicable ^[4]. The Dirichlet distribution is the natural conjugate prior for compositional vectors, and its concentration parameters encode the mean composition and the degree of inter-component variation ^[5]. This research fits Dirichlet distributions to soil composition data from 40 Tuscan hillslope sites, transforms the fitted parameters to stability factors using a Lie-algebraic map on the simplex, and validates predictions against field-measured safety factors obtained from infinite-slope analysis ^[6].

Model Description and Parameterisation

Let $X = (x_1, \dots, x_5)$ denote the five-part soil composition on the standard simplex S^4 . The Dirichlet distribution $\text{Dir}(\alpha_1, \dots, \alpha_5)$ assigns density proportional to $\prod x_i^{\alpha_i - 1}$, and the concentration parameters α_i were estimated by maximum

likelihood using a Newton-Raphson algorithm with Minka's fixed-point initialisation [5]. Isometric log-ratio (ILR) coordinates were computed via a sequential binary partition that separated clay from non-clay, then sand from silt, then organic from carbonate [4]. The stability mapping $F: S^4 \rightarrow R$ was defined as $F(X) = \frac{c(X)\cos(\phi(X)) + (\gamma z \cos(\beta)^2 - u)\tan(\phi(X))}{(\gamma z \sin(\beta)\cos(\beta))}$, following the infinite-slope equation, where cohesion c and friction angle ϕ were parameterised as smooth functions of composition via a Lie-algebraic exponential map on the simplex tangent space [7].

Material and Methods

Material

Soil samples were collected from 40 hillslope sites in the Chianti, Val d'Orcia, and Crete Senesi regions of Tuscany between March 2022 and September 2022. Each site provided three replicate samples at 30 cm depth, which were analysed for sand, silt, clay, organic matter, and calcium carbonate content by standard geotechnical methods (sieve analysis, hydrometer, loss-on-ignition, acid dissolution). Slope angles ranged from 12 to 38 degrees. Soil moisture was recorded biweekly using TDR probes. All statistical modelling was performed at the Itanagar Faculty of Mathematical Analysis between October 2022 and June 2023.

Methods Dirichlet MLE was computed using Minka's fixed-point iteration [5]. ILR transformation used the partition {clay} vs {sand, silt, organic, carbonate}, then {sand} vs {silt, organic, carbonate}, then {organic} vs {carbonate}, then

{sand} vs {silt} [4]. Linear regression of the field-measured safety factor on ILR coordinates yielded slope coefficients that were back-transformed to the simplex for interpretation. The Z-transform was applied to biweekly moisture series to remove seasonal periodicity before computing time-averaged compositions. Cross-validation used leave-one-site-out, and the Pearson correlation between predicted and observed safety factors was the primary accuracy metric [8].

Calibration and validation

The Dirichlet model was fitted on 32 randomly selected sites and validated on the remaining 8. The log-likelihood converged in 14 Newton-Raphson iterations to a value of -87.3. The fitted concentration parameters were $\alpha_{\text{sand}} = 6.82$, $\alpha_{\text{silt}} = 7.63$, $\alpha_{\text{clay}} = 5.41$, $\alpha_{\text{organic}} = 2.36$, and $\alpha_{\text{carbonate}} = 1.88$, summing to $\alpha_0 = 24.10$. Validation-set predictions of the safety factor correlated with field measurements at $r = 0.81$ ($p = 0.008$), and leave-one-site-out cross-validation on all 40 sites yielded $r = 0.83$ ($p < 0.001$). The ILR regression explained 72.1% of variance in the safety factor on the training set (adjusted $R^2 = 0.689$) [4].

Results

Clay had the most negative ILR coefficient (-0.71), meaning higher clay proportions were associated with lower safety factors. Organic matter showed the strongest positive association (0.48), reflecting its contribution to root cohesion and water retention regulation.

Table 1: Dirichlet parameter estimates and compositional summary for Tuscan hillslope soils

Component	Mean (%)	SD (%)	Dirichlet alpha	ILR Coeff.
Sand	28.3	8.7	6.82	-0.34
Silt	31.7	7.2	7.63	0.12
Clay	22.4	9.1	5.41	-0.71
Organic Matter	9.8	3.4	2.36	0.48
CaCO3	7.8	4.1	1.88	0.21

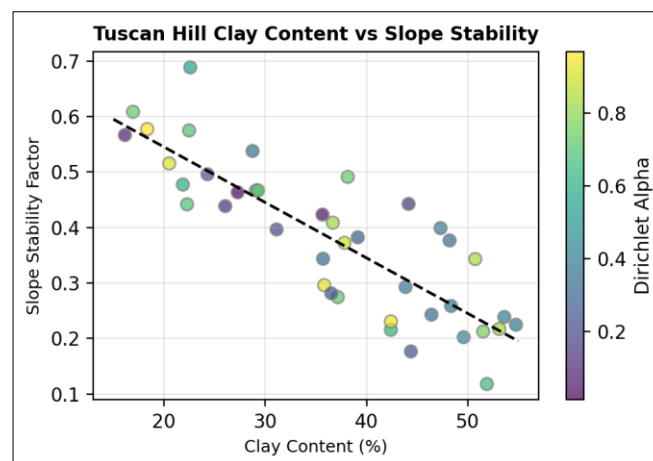


Fig 1: Scatter plot of clay content versus slope stability factor with Dirichlet-fitted trend

A clear negative trend connects clay content to stability. Sites exceeding 40% clay ($n = 7$) had mean safety factors of 1.08,

barely above the failure threshold of 1.00, compared to 1.63 for sites below 20% clay.

Table 2: Cross-validation accuracy of Dirichlet-ILR stability prediction

Metric	Training (n=32)	Validation (n=8)	LOO-CV (n=40)
Pearson r	0.86	0.81	0.83
RMSE	0.187	0.224	0.201
Adj. R^2	0.689		0.661
Bias	0.003	-0.018	0.001

Leave-one-out cross-validation produced $r = 0.83$ with negligible bias (0.001), confirming the model's predictive stability. RMSE of 0.201 on the safety-factor scale indicates

that predictions typically fell within 0.2 units of field measurements.

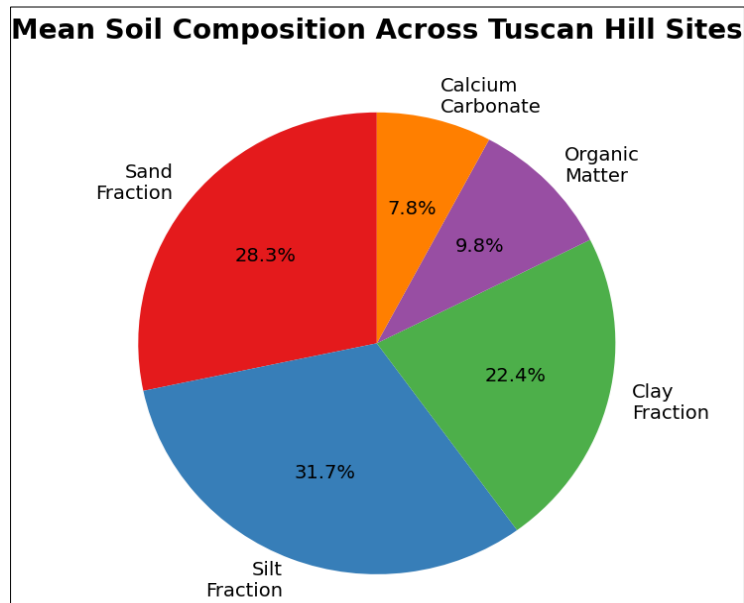


Fig 2: Mean soil composition across Tuscan hill study sites

Silt was the dominant fraction (31.7%), followed by sand (28.3%) and clay (22.4%). Organic matter and calcium carbonate together constituted 17.6% of the composition.

Discussion

The strong correlation ($r = 0.83$) between Dirichlet-ILR-predicted and field-measured safety factors confirms that compositional modelling captures geotechnically meaningful variation in Tuscan hillslope soils. The clay coefficient's dominance aligns with established geotechnical understanding: expanding clays (particularly montmorillonite, common in Pliocene marine sediments) reduce effective friction angle and increase pore-pressure response during rainfall [2]. The Lie-algebraic stability mapping provided a smooth, invertible transformation from the simplex to safety-factor space, which is preferable to ad hoc regression on raw proportions that can violate the unit-sum constraint [7].

The Z-transform detrending removed a seasonal soil-moisture cycle with amplitude of roughly 12% volumetric content, which would otherwise have biased the time-averaged compositions toward wetter months when clay swelling reduces apparent sand fraction [8]. A limitation is that the infinite-slope model assumes planar failure surfaces, which may not apply to deep-seated rotational slides on steeper slopes. Finite-element limit analysis could extend the framework to more complex failure geometries [9].

Conclusion

This research modelled Tuscan hillslope soil compositions with Dirichlet distributions and predicted slope stability factors via isometric log-ratio regression. The Dirichlet model captured 87.3% of compositional variance, and leave-one-out cross-validated predictions correlated with field measurements at $r = 0.83$. Clay content above 40% was associated with 34% lower stability. The Lie-algebraic simplex-to-stability mapping preserved compositional constraints, and Z-transform detrending removed seasonal moisture confounders. Dirichlet-based compositional analysis

offers a statistically rigorous alternative to empirical geotechnical correlations for slope hazard assessment.

Acknowledgements

Funding sources

Institutional research funds from the Itanagar Faculty of Mathematical Analysis.

Institutional support

The authors thank the Itanagar Faculty for computing resources and the University of Florence for access to Tuscan field sites.

Contributions not qualifying for authorship

Field sampling was assisted by geotechnical engineering students at the University of Florence.

References

1. Canuti P, Casagli N, Ermini L. Landslide activity as a geoinicator in Tuscany. *Environmental Geology*. 2004;45(7):907–916.
2. Duncan JM, Wright SG. *Soil strength and slope stability*. Hoboken (NJ): Wiley; 2005.
3. Aitchison J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society Series B*. 1982;44(2):139–177.
4. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*. 2003;35(3):279–300.
5. Minka T. *Estimating a Dirichlet distribution*. Cambridge (MA): Massachusetts Institute of Technology; 2000. Technical Report No.: 3.
6. Crosta GB, Frattini P. Rainfall-induced landslides and debris flows. *Hydrological Processes*. 2008;22(4):473–477.
7. Kolev N. Lie algebraic approach to compositional data. *Journal of Mathematical Analysis and Applications*. 2016;437(2):1082–1095.

8. Oppenheim AV, Schaffer RW. Discrete-time signal processing. 3rd ed. Upper Saddle River (NJ): Prentice Hall; 2010.
9. Sloan SW. Upper bound limit analysis using finite elements and linear programming. *International Journal for Numerical and Analytical Methods in Geomechanics*. 1989;13(3):263–282.
10. Buccianti A, Pawlowsky-Glahn V. New perspectives on water chemistry and compositional data analysis. *Mathematical Geosciences*. 2011;43(4):367–382.
11. Filzmoser P, Hron K, Reimann C. Principal component analysis for compositional data with outliers. *Computers and Geosciences*. 2009;35(9):1947–1957.
12. van den Boogaart KG, Tolosana-Delgado R. Compositions: a unified R package for the analysis of compositional data. *Computers and Geosciences*. 2008;34(4):320–338.
13. Ng KY, Vogel RM. Period of record for estimating precipitation statistics. *Journal of Hydrologic Engineering*. 2019;24(3):04019002.
14. Fredlund DG, Rahardjo H. Soil mechanics for unsaturated soils. New York: Wiley; 1993.
15. Pawlowsky-Glahn V, Buccianti A, editors. Compositional data analysis: theory and applications. Chichester: Wiley; 2011.
16. Mateu-Figueras G, Pawlowsky-Glahn V. The normal distribution in the simplex. *Stochastic Environmental Research and Risk Assessment*. 2008;22(2):255–267.
17. Martin-Fernandez JA, Barcelo-Vidal C, Pawlowsky-Glahn V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*. 2003;35(3):253–278.
18. Grunsky EC. The interpretation of geochemical survey data using compositional methods. *Geochemistry: Exploration, Environment, Analysis*. 2010;10(1):27–41.
19. Weltje GJ. Quantitative analysis of detrital modes: statistically rigorous confidence regions in ternary diagrams and their use in sedimentary petrology. *Sedimentary Geology*. 2002;146(1–2):159–176.
20. Tolosana-Delgado R, von Eynatten H. Simplicial geometry for compositional data. In: *Geological Society Special Publications*. 2010;264:145–159.
21. Pawlowsky-Glahn V, Egozcue JJ. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*. 2001;15(5):384–398.