

INTEGRATED NUMERICAL AND MIXED-DESCRIPTOR PHENOTYPING REVEALS COMPLEMENTARY STRUCTURE IN LENTIL (*LENS CULINARIS* MEDIK.) GERMPLASM

Chiranjit Mazumder^{1*}, Pankaj Das², Bikramjeet Ghose³, Pramit Pandit⁴, Mrinmoy Ray⁵

¹Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute, Pusa Campus, New Delhi 110012, India.

²ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, Pusa, New Delhi 110012, India.

³Department of Agricultural Economics and Statistics, M.S. Swaminathan School of Agriculture, Centurion University of Technology and Management, R. Sitapur, Odisha 761211, India.

⁴Central Silk Board-Central Sericultural Research and Training Institute, Berhampore 742101, Murshidabad, West Bengal, India.

⁵Agricultural Knowledge Management Unit, ICAR-Indian Agricultural Research Institute, Pusa Campus, New Delhi 110012, India.

*Corresponding author: Dr. Chiranjit Mazumder, E-mail: chiranjit@iari.res.in

ABSTRACT

Mixed-scale germplasm data can yield different biological groupings depending on whether continuous and coded descriptors are analyzed separately or jointly. We evaluated 518 lentil (*Lens culinaris* Medik.) accessions characterized by 10 continuous agro-morphological traits and 20 coded descriptors. Standardized continuous traits were analyzed by principal component analysis (PCA) followed by fuzzy c-means clustering, whereas continuous and coded descriptors were integrated through a factor-analysis-of-mixed-data-based embedding and a second fuzzy partition. Gower dissimilarity supported mixed-data validation, and the adjusted Rand index quantified agreement between pathways. Yield per plant was strongly related to pods per plant ($r = 0.807$) and biological yield per plant ($r = 0.681$). The first three principal components explained 67.35% of continuous-trait variance and resolved three profiles: a high-yield, high-pod group; a later, taller, lower-yield group; and an early, bold-seeded group. The mixed pathway resolved two broader descriptor-defined groups (fuzzy partition coefficient = 0.567; Xie-Beni index = 1.265; Gower silhouette = 0.087). Agreement between numerical and mixed allocations was moderate (adjusted Rand index = 0.325), showing that coded descriptor states altered accession grouping beyond the continuous-trait solution. High-membership accessions can represent group-specific phenotypes, whereas transitional accessions may provide useful recombination material. The framework offers a transparent way to combine performance, morphology, and assignment uncertainty in lentil germplasm characterization.

KEYWORDS: *Lens culinaris*; germplasm characterization; mixed-data analysis; principal component analysis; factor analysis of mixed data; fuzzy c-means

INTRODUCTION

Lentil (*Lens culinaris* Medik.) is a major cool-season pulse crop and a valuable source of dietary protein, micronutrients, and biological nitrogen fixation. Germplasm collections preserve variation in phenology, plant architecture, seed attributes, and yield components, but their value for crop improvement depends on systematic characterization and practical identification of contrasting parents. Descriptor systems therefore combine measured agronomic traits with categorical or ordinal states such as pigmentation, seed coat color, leaf pubescence, and growth habit (IBPGR and ICARDA, 1985; Mohammadi and Prasanna, 2003; Tripathi et al., 2022). These measurement scales are biologically complementary but cannot be treated as interchangeable without distorting distances, variances, and cluster structure.

Principal component analysis (PCA) is well suited to correlated continuous traits and can identify major productivity, phenology, and architecture axes (Jolliffe, 2002; Abdi and Williams, 2010; Jolliffe and Cadima, 2016). Categorical descriptors require a different treatment because arbitrary integer codes do not carry equal numerical distances. Factor analysis of mixed data (FAMD) balances standardized quantitative variables with frequency-scaled indicator variables and places both blocks in a shared latent space (Pagès, 2004; Lê et al., 2008; Husson et al., 2017). Gower dissimilarity provides a complementary validation measure because it combines range-scaled numerical differences and categorical mismatches without imposing a single measurement scale (Gower, 1971).

Hard clustering assigns every accession to one group even when its profile lies between biological types. Fuzzy c-means instead estimates a membership vector for each accession, allowing core and transitional materials to be distinguished (Dunn, 1973; Bezdek, 1981; Bezdek et al., 1984). This is particularly relevant to crop germplasm, where recombination, selection history, and quantitative inheritance frequently generate overlapping phenotypes. Cluster validity should nevertheless be evaluated with several criteria because compactness, separation, and membership certainty emphasize different properties (Rousseeuw, 1987; Xie and Beni, 1991).

Earlier analyses of the present 518-accession resource demonstrated the usefulness of PCA-based and nonlinear-PCA-based fuzzy classification (Jha et al., 2014; Mazumder et al., 2015). The current re-analysis asks a distinct question: how much does accession grouping change when continuous trait performance and the full coded descriptor representation are evaluated through parallel, explicitly comparable pathways? The objectives were to (i) quantify continuous and coded descriptor variation, (ii) identify major continuous-trait axes, (iii) derive and validate fuzzy partitions for numerical and mixed representations, (iv) quantify agreement between the two partitions, and (v) identify candidate accessions for further breeding evaluation while retaining membership uncertainty.

MATERIAL AND METHODS

Plant material and experimental background

The dataset comprised 518 lentil accessions, including 206 exotic entries and 312 indigenous collections; 59 of the indigenous accessions were breeding lines. The accessions were evaluated during the rabi season at the Indian Institute of Pulses Research, Kanpur, India, in an augmented block design with five checks: DPL 62, JL 1, K 75, PL 406, and Ranjan. The trial contained 11 blocks. Each of the first 10 blocks contained 50 test accessions and five randomized checks, whereas the final block contained 18 test entries. Each accession was grown in a 3-m row under standard agronomic management.

Qualitative descriptors, days to 50% flowering, and days to maturity were recorded on a plot basis. Plant height, plant height at the lowest pod, primary branches, secondary branches, pods, seed yield, and biological yield were measured on five randomly selected plants per row and averaged. One hundred mature seeds were used to determine 100-seed weight. Descriptor recording followed biodiversity and national Distinctness, Uniformity and Stability guidance consistent with the lentil descriptor framework (IBPGR and ICARDA, 1985). The original observations were obtained through breeder communication; the present study re-analyzed the supplied accession-level working matrix.

Trait architecture and data audit

The continuous block contained days to 50% flowering (DF), plant height (Ph), days to maturity (DM), 100-seed weight (100SW), biological yield per plant (BYP), primary branches per plant (PB), secondary branches per plant (SB), pods per plant (PPP), yield per plant (YPP), and plant height at the lowest pod (PHLP). The coded block comprised 10 qualitative descriptors - stem pigmentation (SP), tendril length (TL), pod pigmentation (PPIG), cotyledon color (CC), seed coat color (SCC), seedling vigor (SV), leaf size (LS), leaf pubescence (LP), leaf color (LC), and plant growth habit (PGH) - plus categorized versions of the 10 continuous traits. Thus, the mixed pathway used 10 continuous variables and 20 coded descriptor representations. The two representations of the measured traits were block-scaled in the mixed embedding and should be interpreted as complementary encodings, not as 30 independent biological traits.

All 15,540 cells in the working matrix were complete, so no imputation was performed. Continuous traits were summarized by mean, standard deviation, coefficient of variation, range, skewness, and kurtosis. Pearson correlations described continuous-trait association. Descriptor balance was assessed using the Shannon diversity index, Shannon equitability, and Simpson diversity (Shannon and Weaver, 1949). Continuous variables were standardized to zero mean and unit variance before ordination and clustering.

Parallel numerical and mixed-data pathways

In the numerical pathway, PCA was applied to the correlation matrix of the 10 standardized continuous traits. Components were interpreted from their eigenvalues and loadings; the Kaiser criterion and cumulative variance supported retention (Kaiser, 1960). Fuzzy c-means was fitted to the retained first three component scores with fuzzifier $m = 2$. Candidate solutions from two to eight clusters were examined.

In the mixed pathway, each coded descriptor was expanded into indicator columns. Indicators were centered and scaled by category frequency, and each descriptor block was normalized so that variables with more levels did not dominate the latent space. The standardized continuous block and the balanced coded block were then combined in an FAMD-based embedding (Pagès, 2004; Lê et al., 2008). Fuzzy c-means was fitted to the first five mixed dimensions. Gower dissimilarities, calculated as range-scaled absolute differences for continuous traits and mismatch indicators for coded descriptors, provided an external mixed-scale silhouette assessment (Gower, 1971).

Both pathways were evaluated with the fuzzy partition coefficient (FPC), modified partition coefficient (MPC), Xie-Beni index, silhouette coefficient, and Davies-Bouldin index (Davies and Bouldin, 1979; Xie and Beni, 1991). Higher FPC, MPC, and silhouette values and lower Xie-Beni and Davies-Bouldin values indicate stronger partitions, although no single index is uniformly decisive. Final solutions were chosen by balancing validity statistics, stability, parsimony, and biological interpretability; alternative solutions are reported to make that decision transparent. The adjusted Rand index (ARI) quantified agreement between the defuzzified numerical and mixed allocations, with 0 indicating chance-level agreement and 1 indicating identical partitions (Hubert and Arabie, 1985). Maximum fuzzy membership below 0.60 was used descriptively to flag transitional accessions.

Candidate accessions were screened within each mixed cluster by YPP, while PPP, 100SW, and maximum fuzzy membership were retained to distinguish productivity from assignment certainty. Candidate ranking was

exploratory and does not substitute for adjusted-mean analysis, multi-environment testing, or molecular validation. Table 1 summarizes the analytical framework.

Table 1. Analytical framework and data architecture.

Component	Specification	Purpose or detail
Accessions	518	Complete accession-level matrix
Continuous traits	10	DF, Ph, DM, 100SW, BYP, PB, SB, PPP, YPP, PHLP
Coded descriptors	20	10 qualitative descriptors plus categorized versions of 10 continuous traits
Numerical pathway	PCA -> FCM	Three retained PCs; candidate clusters k = 2-8; fuzzifier m = 2
Mixed pathway	FAMD-based FCM ->	First five mixed dimensions; candidate clusters k = 2-8; fuzzifier m = 2
Validation	Multiple indices	FPC, MPC, Xie-Beni, silhouette, Davies-Bouldin, and Gower silhouette
Method agreement	Adjusted index Rand	Agreement between defuzzified numerical and mixed allocations

Note. FCM = fuzzy c-means; FPC = fuzzy partition coefficient; MPC = modified partition coefficient.

RESULTS

Continuous-trait variability and association

The continuous block showed substantial phenotypic dispersion (Table 2). DF averaged 78.69 days and DM 126.03 days. YPP averaged 3.72 g and ranged from 0.20 to 10.70 g; its coefficient of variation (CV) was 43.29%. PPP ranged from 3.70 to 309.30 and had a CV of 40.96%, while BYP had a CV of 28.03%. These distributions indicate that reproductive load, biomass, and plant yield supplied more relative differentiation than maturity duration.

Table 2. Descriptive statistics for continuous agro-morphological traits (n = 518).

Trait	Mean (SD)	CV (%)	Median	Range	Skewness	Kurtosis
DF	78.685 (10.760)	13.675	81.0	58.0-106.0	-0.502	-0.568
Ph	30.789 (4.795)	15.574	31.3	17.0-47.6	-0.240	0.260
DM	126.033 (4.703)	3.731	127.0	114.0-140.0	-0.744	0.362
100SW	2.429 (0.529)	21.768	2.3	1.2-4.1	0.443	-0.728
BYP	13.373 (3.749)	28.034	13.0	4.2-28.0	0.335	0.384
PB	3.761 (1.061)	28.223	3.6	2.0-9.0	1.270	2.799
SB	10.215 (2.393)	23.422	10.0	4.0-18.0	0.182	0.184
PPP	116.156 (47.574)	40.956	115.15	3.7-309.3	0.534	0.778
YPP	3.721 (1.611)	43.294	3.5	0.2-10.7	0.941	1.886
PHLP	10.738 (2.305)	21.465	10.7	1.0-19.0	-0.361	2.223

Note. CV = coefficient of variation. Trait abbreviations are defined in the Material and methods section.

The correlation structure supported a coherent productivity complex (Figure 1). YPP was most strongly associated with PPP ($r = 0.807$, $P < 0.001$) and BYP ($r = 0.681$, $P < 0.001$), followed by SB ($r = 0.511$). PB and SB were also closely related ($r = 0.74$). DF and 100SW were negatively associated ($r = -0.52$), suggesting that earliness and seed size contributed a second, partly contrasting dimension.

Correlation structure among continuous agro-morphological traits

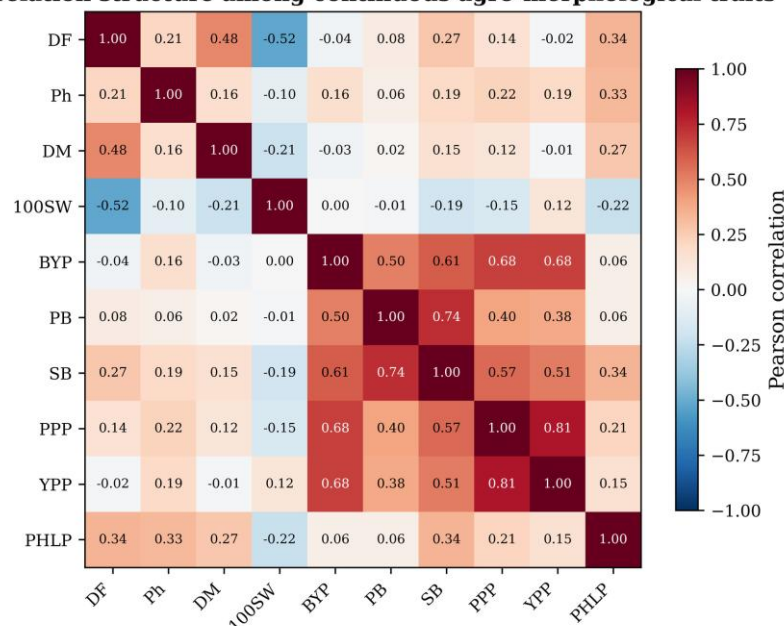


Figure 1. Pearson correlation structure among 10 continuous agro-morphological traits in 518 lentil accessions. Trait abbreviations are defined in the Material and methods section.

Diversity of coded descriptors

Descriptor balance differed widely (Table 3). Cotyledon color and stem pigmentation were highly dominated by one state (97.68% and 94.21% of accessions, respectively), giving low equitability. Tendril length, seedling vigor, leaf size, and leaf pubescence were more evenly distributed. The categorized continuous traits generally showed high equitability, particularly DF (0.938), 100SW (0.896), BYP (0.878), and DM (0.857). Consequently, rare qualitative states contributed distinctiveness, whereas the categorized agronomic traits contributed broad, balanced gradients.

Table 3. Diversity and balance of coded lentil descriptors.

Descriptor	No. classes	Shannon H	Equitability	Simpson 1-D	Dominant n	Dominant (%)
Stem pigmentation (SP)	2	0.221	0.319	0.109	488	94.21
Tendril length (TL)	2	0.668	0.964	0.475	317	61.20
Pod pigmentation (PPIG)	2	0.598	0.863	0.408	370	71.43
Cotyledon color (CC)	2	0.110	0.159	0.045	506	97.68
Seed coat color (SCC)	4	1.024	0.739	0.577	304	58.69
Seedling vigor (SV)	4	1.231	0.888	0.686	196	37.84
Leaf size (LS)	3	0.950	0.864	0.562	307	59.27
Leaf pubescence (LP)	3	1.015	0.924	0.618	231	44.59
Leaf color (LC)	4	0.976	0.704	0.562	290	55.98
Plant growth habit (PGH)	2	0.378	0.545	0.219	453	87.45
Days to flowering (DF)	8	1.951	0.938	0.845	112	21.62
Plant height (Ph)	8	1.763	0.848	0.799	151	29.15
Days to maturity (DM)	6	1.535	0.857	0.754	177	34.17
100-seed weight (100SW)	6	1.606	0.896	0.773	179	34.56
Biological yield (BYP)	6	1.572	0.878	0.765	173	33.40
Primary branches (PB)	7	1.552	0.797	0.744	194	37.45

Descriptor	No. classes	Shannon H	Equitability	Simpson 1-D	Dominant n	Dominant (%)
Secondary branches (SB)	8	1.810	0.870	0.809	155	29.92
Pods per plant (PPP)	9	1.806	0.822	0.808	140	27.03
Yield per plant (YPP)	6	1.349	0.753	0.691	198	38.22
Lowest pod height (PHLP)	7	1.546	0.795	0.745	186	35.91

Note. Dominant n and dominant (%) refer to the most frequent category in the source-coded descriptor. Categorized quantitative traits are included in the coded descriptor block.

PCA and numerical fuzzy clustering

The first three principal components explained 67.35% of standardized continuous-trait variance (Table 4). PC1 explained 35.76% and represented productivity: SB, PPP, BYP, and YPP had the largest positive loadings. PC2 explained 21.03% and separated later flowering and maturity from higher 100-seed weight. PC3 explained 10.56% and was dominated by plant height and branching architecture. The biplot therefore separated a pod-biomass-yield axis from a phenology-seed-size contrast (Figure 2).

Table 4. Eigenvalues and dominant loading structure for continuous-trait PCA.

PC	Eigenvalue	Variance (%)	Cumulative (%)	Largest absolute loadings
PC1	3.583	35.764	35.764	SB (+0.86), PPP (+0.85), BYP (+0.79), YPP (+0.78)
PC2	2.107	21.025	56.789	DF (+0.78), 100SW (-0.64), DM (+0.63), PHLP (+0.51)
PC3	1.058	10.562	67.352	Ph (+0.66), PB (-0.43), PHLP (+0.36), 100SW (+0.36)
PC4	0.810	8.083	75.434	PB (+0.46), PPP (-0.38), 100SW (+0.37), SB (+0.30)

Note. The first three components were retained for numerical fuzzy clustering.

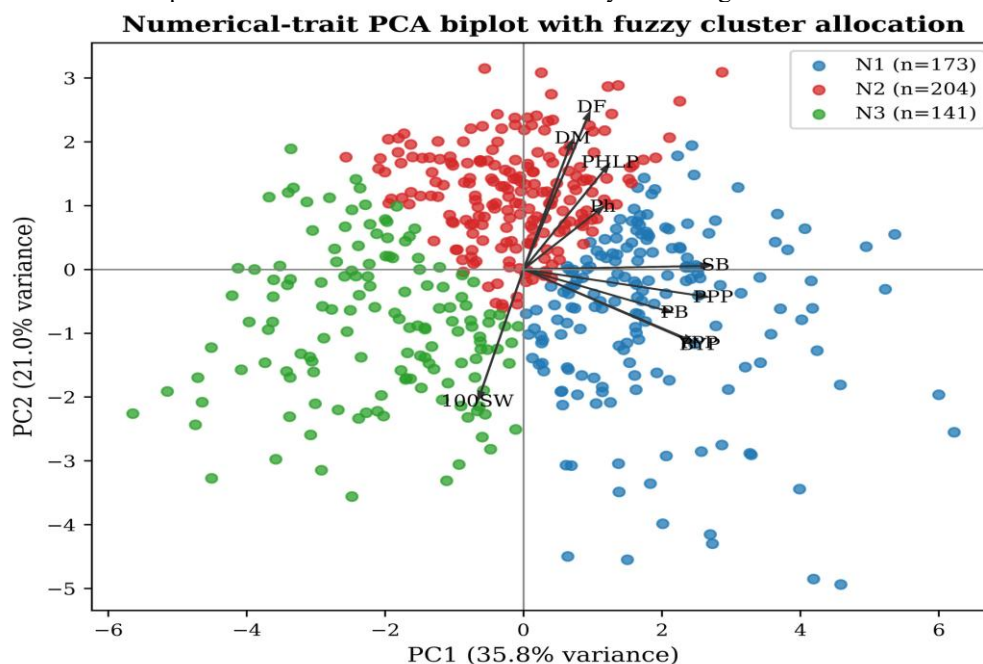


Figure 2. Continuous-trait PCA biplot with numerical fuzzy c-means allocation. Arrows show trait loadings and points show accession scores colored by the retained three-cluster solution.

Validity indices did not identify a single universally superior partition (Table 5; Figure 3). The two-cluster numerical solution had the highest silhouette (0.283), whereas the three-cluster solution had FPC = 0.525, MPC = 0.287, Xie-Beni = 0.429, silhouette = 0.266, and Davies-Bouldin = 1.191. The three-cluster solution was retained because it separated three agronomically distinct patterns that were conflated under two clusters. N1 (n = 173) had the highest BYP, PPP, and YPP; N2 (n = 204) was later, taller, and lower yielding; and N3 (n = 141)

was earliest and had the highest mean 100SW (Table 6). However, 213 accessions had maximum membership below 0.60, confirming broad transition zones.

Table 5. Selected cluster-validity results for numerical and mixed-data pathways.

Pathway		k	FPC	MPC	Xie-Beni	Silhouette	Davies-Bouldin
Numerical FCM	PCA-	2	0.645	0.289	0.487	0.283	1.299
Numerical FCM	PCA-	3	0.525	0.287	0.429	0.266	1.191
Numerical FCM	PCA-	4	0.447	0.262	0.430	0.258	1.213
Numerical FCM	PCA-	5	0.392	0.240	0.356	0.227	1.248
Mixed FCM	FAMD-	2	0.567	0.133	1.265	0.087	1.665
Mixed FCM	FAMD-	3	0.415	0.123	1.305	0.070	1.516
Mixed FCM	FAMD-	4	0.317	0.089	2.454	0.048	1.704
Mixed FCM	FAMD-	5	0.262	0.077	1.643	0.039	1.695

Note. For the mixed pathway, Silhouette is based on Gower dissimilarity. Higher FPC, MPC, and silhouette and lower Xie-Beni and Davies-Bouldin values indicate stronger partitions.

Table 6. Trait profiles of the retained numerical PCA-fuzzy c-means clusters.

Variable	N1 (n = 173)	N2 (n = 204)	N3 (n = 141)
Mean maximum membership	0.640	0.654	0.646
Membership <0.60 (n)	73	87	53
DF (days)	78.10	85.70	69.25
DM (days)	125.97	127.78	123.59
100SW (g)	2.50	2.13	2.78
BYP (g/plant)	16.80	12.19	10.88
PPP (pods/plant)	159.33	107.66	75.48
YPP (g/plant)	5.24	3.15	2.69
PHLP (cm)	10.97	11.60	9.21

Note. Values are cluster means except membership columns. N1 = high-yield/high-pod; N2 = late/tall lower-yield; N3 = early/bold-seeded.

Cluster-validity diagnostics for numerical and mixed-data pathways

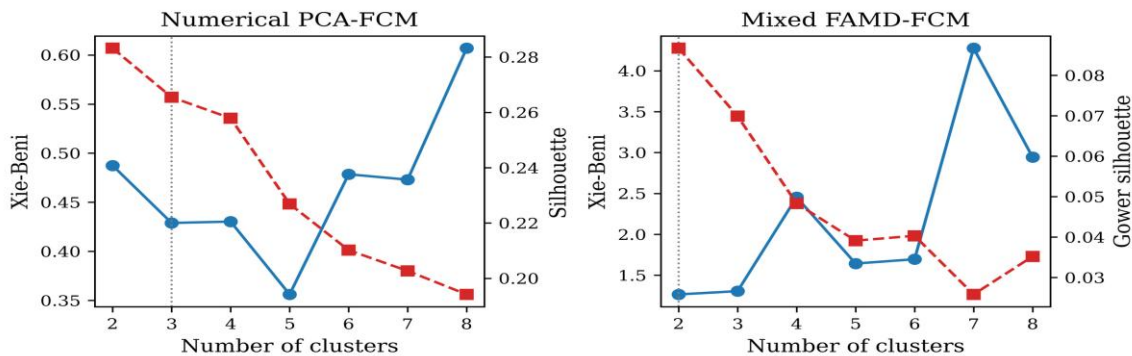


Figure 3. Cluster-validity diagnostics for numerical PCA-fuzzy c-means and mixed FAMD-fuzzy c-means pathways. Vertical dotted lines mark the retained cluster numbers.

Mixed-descriptor fuzzy clustering

The first five mixed dimensions explained 37.5% of the transformed inertia. The retained two-cluster solution had FPC = 0.567, MPC = 0.133, Xie-Beni = 1.265, Gower silhouette = 0.087, and Davies-Bouldin = 1.665 (Table 5). The low Gower silhouette indicates a continuum rather than sharply isolated descriptor groups, but the two broad meta-clusters remained interpretable in the ordination (Figure 4). M1 ($n = 293$) combined greater biomass, pod number, and yield; M2 ($n = 225$) was earlier, had higher mean 100SW, and had lower biomass and yield (Table 7). A total of 161 accessions had maximum mixed membership below 0.60.

Table 7. Trait profiles of the retained mixed FAMD-fuzzy c-means clusters.

Variable	M1 ($n = 293$)	M2 ($n = 225$)
Mean maximum membership	0.664	0.649
Membership <0.60 (n)	92	69
DF (days)	82.42	73.83
DM (days)	126.97	124.82
100SW (g)	2.28	2.62
BYP (g/plant)	15.06	11.17
PPP (pods/plant)	141.65	82.96
YPP (g/plant)	4.41	2.82
PHLP (cm)	11.52	9.72

Note. Values are cluster means except membership columns. M1 = higher biomass/pod/yield meta-cluster; M2 = earlier, bold-seeded, lower-pod meta-cluster.

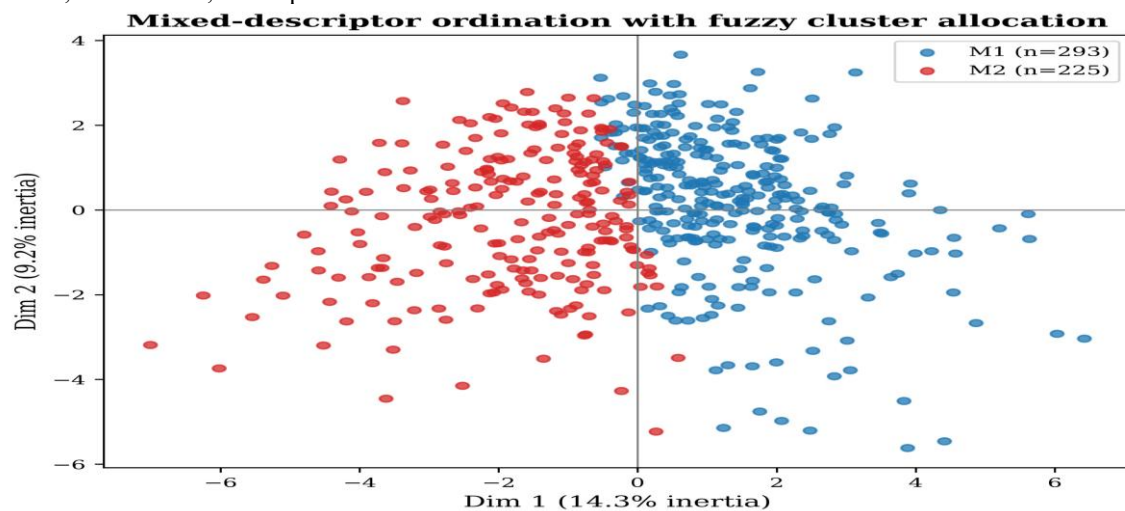


Figure 4. Mixed-descriptor ordination with fuzzy cluster allocation. Coordinates were obtained from the FAMD-based embedding of continuous and coded descriptor blocks.

Agreement between pathways and candidate accessions

The ARI between numerical and mixed hard allocations was 0.325, indicating moderate concordance. Cross-classification clarified the source of this difference (Table 8): all 141 N3 accessions entered M2, 161 of 173 N1 accessions entered M1, and the later N2 profile split between M1 (132 accessions) and M2 (72 accessions). The mixed descriptor representation therefore reproduced the early, bold-seeded pattern but reorganized accessions within the high-productivity and late/tall numerical groups. Standardized cluster profiles visualize these contrasts (Figure 5), while membership distributions show the substantial transition zones in both pathways (Figure 6).

Table 8. Cross-classification of numerical and mixed-data hard allocations.

Numerical cluster	Mixed M1	Mixed M2	Row total
N1	161	12	173
N2	132	72	204
N3	0	141	141
Column total	293	225	518

Note. Adjusted Rand index = 0.325.

Standardized cluster profiles across agronomic traits

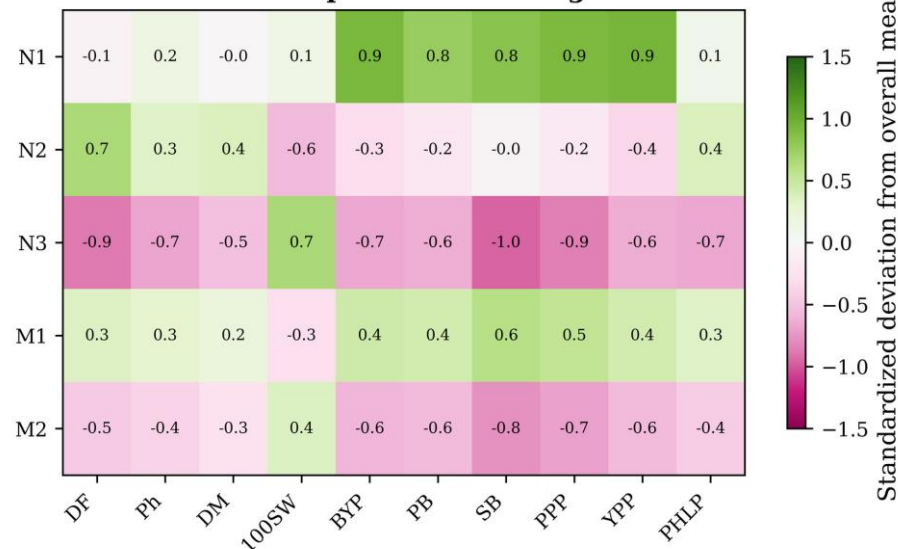


Figure 5. Standardized continuous-trait profiles of numerical (N1-N3) and mixed-data (M1-M2) clusters. Positive values indicate means above the full-collection mean.

Fuzzy membership certainty across accessions

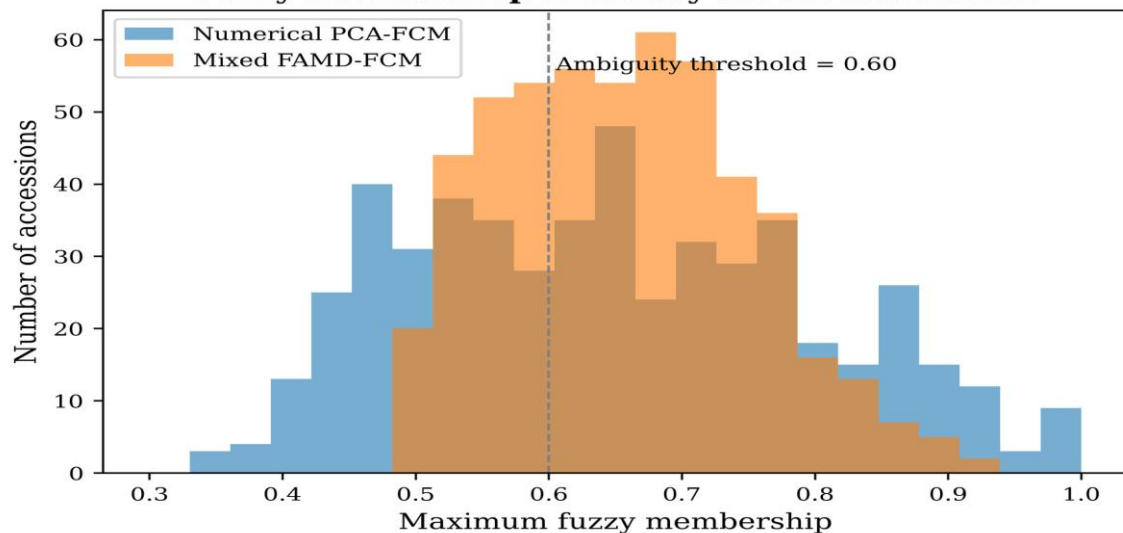


Figure 6. Distribution of maximum fuzzy membership in the numerical and mixed pathways. The dashed line at 0.60 marks the descriptive ambiguity threshold.

Within-cluster yield screening identified several accessions for further evaluation (Table 9). M1 included ILL 5958 (YPP = 10.7 g), LC 505 (10.3 g), ILL 6063 (9.7 g), LC 561 (9.7 g), and ILL 6090 (8.8 g). M2 included EC 382700 (10.3 g) and several accessions with moderate yield and relatively bold seed. Membership strength changed the interpretation: ILL 4344 was a strong M2 representative (0.756), whereas ILL 5958 (0.517) and IPL 2K-52 (0.500) were transitional despite their yield ranking.

Table 9. High-yield accessions identified within the mixed-data clusters.

Mixed cluster	Accession	YPP (g)	PPP	100SW (g)	Max. membership
M1	ILL 5958	10.7	135.0	2.9	0.517
M1	LC 505	10.3	265.0	3.7	0.562
M1	ILL 6063	9.7	309.3	3.1	0.604
M1	LC 561	9.7	252.6	3.1	0.564
M1	ILL 6090	8.8	224.0	2.9	0.636
M2	EC 382700	10.3	32.0	2.7	0.665
M2	IPL 2K-52	5.8	165.7	3.1	0.500
M2	IPKL2K-15	5.7	128.3	3.4	0.511
M2	ILL 4344	5.3	134.7	3.0	0.756
M2	ILL 6884	5.1	139.6	3.4	0.517

Note. Candidates are the five accessions with the highest observed YPP within each mixed cluster. They require confirmation with adjusted means and multi-environment evaluation.

DISCUSSION

The parallel analysis shows why continuous and coded descriptor pathways should be interpreted jointly. The numerical pathway isolated three agronomically recognizable profiles, whereas the mixed pathway compressed them into two broader descriptor-defined groups. Moderate ARI confirms that the two solutions were related but not interchangeable. This is the expected consequence of changing both the information content and its geometry: continuous PCA emphasizes covariance among measured traits, while a mixed embedding gives explicit weight to category prevalence and state mismatches (Gower, 1971; Pagès, 2004; Husson et al., 2017).

Pods per plant, biological yield, secondary branches, and seed yield formed the dominant productivity axis. The strong YPP-PPP and YPP-BYP correlations are biologically plausible because pod number reflects reproductive sink capacity and biomass reflects source development. In contrast, 100-seed weight loaded against later phenology on PC2 and was only weakly related to YPP. Selection for bold seed should therefore be combined with direct attention to pod-bearing ability and biomass rather than treated as a proxy for plant yield. Similar multi-trait considerations underpin germplasm characterization and core-set development in lentil (Mohammadi and Prasanna, 2003; Tripathi et al., 2022).

Fuzzy memberships supplied information that hard labels would hide. More than 40% of accessions were below the 0.60 threshold in the numerical solution and 31% in the mixed solution. High-membership accessions are useful as representative phenotypes for curation, comparison, and parental contrasts. Transitional accessions may be equally important for breeding because they combine features from neighboring groups. The low mixed Gower silhouette should not be interpreted as evidence of absent diversity; rather, it indicates that diversity is distributed along a continuum with overlapping descriptor combinations. Reporting uncertainty makes that structure explicit (Bezdek et al., 1984; Rousseeuw, 1987; Xie and Beni, 1991).

Relative to earlier PCA-fuzzy and nonlinear-PCA-fuzzy studies of the same broad germplasm resource (Jha et al., 2014; Mazumder et al., 2015), the present analysis contributes a directly comparable dual-pathway design, mixed-scale Gower validation, ARI-based method agreement, and explicit membership-aware candidate interpretation. These additions change the practical question from finding one definitive cluster solution to identifying which accession relationships persist, disappear, or emerge when the descriptor scale changes.

Several limitations define the appropriate scope of inference. First, the working matrix contained accession-level phenotypic values from one experimental setting; the re-analysis did not reconstruct block- and check-adjusted means. Second, the mixed pathway included both continuous traits and their categorized representations, so it should be viewed as a descriptor-augmented sensitivity analysis rather than a set of independent variables. Third, qualitative code definitions and frequency imbalance can influence latent coordinates. Fourth, candidate ranking was based on observed YPP within clusters and requires confirmation with adjusted means, replicated environments, stability analysis, and disease or stress-response traits. Finally, the results describe phenotypic

structure and do not estimate molecular genetic distance. These limitations do not negate the comparative finding, but they prevent direct cultivar recommendation.

CONCLUSIONS

Continuous and coded descriptors revealed complementary structure among 518 lentil accessions. Three numerical profiles separated high-productivity, late/tall, and early/bold-seeded materials, whereas the mixed representation produced two broader meta-clusters and only moderate agreement with the numerical solution. Pods per plant and biological yield were the strongest correlates of seed yield. Fuzzy membership further distinguished representative accessions from transitional material that crossed phenotypic boundaries. For germplasm management, the two pathways should be used together: continuous PCA-fuzzy clustering for agronomic performance profiles and mixed FAMD-fuzzy clustering for descriptor-integrated curation. Candidate accessions identified here should be treated as priorities for adjusted-mean re-analysis, multi-environment testing, and molecular validation rather than as immediately deployable varieties.

Data availability

The re-analysis used a supplied working matrix of 518 lentil accessions. The original observations were obtained through breeder communication. The analytical dataset may be made available by the corresponding author upon reasonable request, subject to permission from the data owner.

Funding

No specific funding supported this re-analysis.

Conflict of interest

The authors declare that they have no conflict of interest.

Author contributions

CM conceived and led the study, developed the analytical framework, interpreted the results, and drafted the manuscript. PD, BG, PP, and MR contributed to methodology, validation, interpretation, and critical revision. All authors reviewed and approved the final manuscript.

ACKNOWLEDGMENTS

The authors acknowledge the breeder and source scientist who generated and communicated the original lentil germplasm observations, and the experimental program at the Indian Institute of Pulses Research, Kanpur, where the accessions were evaluated.

REFERENCES

1. Abdi H and Williams LJ (2010). Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2: 433-459. <https://doi.org/10.1002/wics.101>
2. Bezdek JC (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.
3. Bezdek JC, Ehrlich R and Full W (1984). FCM: The fuzzy c-means clustering algorithm. Comput. Geosci. 10: 191-203. [https://doi.org/10.1016/0098-3004\(84\)90020-7](https://doi.org/10.1016/0098-3004(84)90020-7)
4. Davies DL and Bouldin DW (1979). A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1: 224-227. <https://doi.org/10.1109/TPAMI.1979.4766909>
5. Dunn JC (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybern. 3: 32-57. <https://doi.org/10.1080/01969727308546046>
6. Gower JC (1971). A general coefficient of similarity and some of its properties. Biometrics 27: 857-871. <https://doi.org/10.2307/2528823>
7. Hubert L and Arabie P (1985). Comparing partitions. J. Classif. 2: 193-218. <https://doi.org/10.1007/BF01908075>
8. Husson F, Lê S and Pagès J (2017). Exploratory Multivariate Analysis by Example Using R. 2nd edn. CRC Press, Boca Raton.
9. IBPGR and ICARDA (1985). Lentil Descriptors. International Board for Plant Genetic Resources, Rome.
10. Jha GK, Mazumder C, Kumari J and Singh G (2014). Nonlinear principal component based fuzzy clustering: A case study of lentil genotypes. Indian J. Genet. Plant Breed. 74: 189-196. <https://doi.org/10.5958/0975-6906.2014.00155.2>
11. Jolliffe IT (2002). Principal Component Analysis. 2nd edn. Springer, New York.
12. Jolliffe IT and Cadima J (2016). Principal component analysis: A review and recent developments. Philos. Trans. A Math. Phys. Eng. Sci. 374: 20150202. <https://doi.org/10.1098/rsta.2015.0202>
13. Kaiser HF (1960). The application of electronic computers to factor analysis. Educ. Psychol. Meas. 20: 141-151. <https://doi.org/10.1177/001316446002000116>
14. Lê S, Josse J and Husson F (2008). FactoMineR: An R package for multivariate analysis. J. Stat. Softw. 25: 1-18. <https://doi.org/10.18637/jss.v025.i01>
15. Mazumder C, Jha GK, Parsad R, Bharadwaj A, et al. (2015). Principal component based fuzzy c-means algorithm for clustering lentil germplasm. J. Indian Soc. Agric. Stat. 69: 307-314.
16. Mohammadi SA and Prasanna BM (2003). Analysis of genetic diversity in crop plants: Salient statistical tools and considerations. Crop Sci. 43: 1235-1248. <https://doi.org/10.2135/cropsci2003.1235>

17. Pagès J (2004). Analyse factorielle de données mixtes. *Rev. Stat. Appl.* 52: 93-111.
18. Rousseeuw PJ (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20: 53-65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
19. Shannon CE and Weaver W (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.
20. Tripathi K, Kumari J, Gore PG, Mishra DC, et al. (2022). Agro-morphological characterization of lentil germplasm of Indian National Genebank and development of a core set for efficient utilization in lentil improvement programs. *Front. Plant Sci.* 12: 751429. <https://doi.org/10.3389/fpls.2021.751429>
21. Xie XL and Beni G (1991). A validity measure for fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* 13: 841-847. <https://doi.org/10.1109/34.85677>